Statistical Society of Slovenia

International Conference

21st Applied Statistics 2025

September 21–23, 2025

Koper, Slovenia

Book of Abstracts

as.mf.uni-lj.si | appl.stat@guest.arnes.si

International Conference

APPLIED STATISTICS 2025

Book of Abstracts

September 21–23, 2025

Koper, Slovenia

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 249623299
ISBN 978-961-94283-8-2 (PDF)

Edited by Andrej Kastrin and Lara Lusa
Published by Statistical Society of Slovenia
Litostrojska cesta 54
1000 Ljubljana, Slovenia
Publication year 2025
Publication format PDF
Electronic version: https://as.mf.uni-lj.si/uploads/pdf/as2025book.pdf

PROGRAM and ABSTRACTS

PROGRAM

Day	Time	Room 1	Room 2
	13:30–15.00	Registration	
Sunday	14.00–18.00	Workshop	
Monday	08.30–15.00	Registration	
	09.00–09.10	Opening	
	09.10–10.00	Invited Lecture	
	10.00–10.30	Coffee Break	
	10.30–12.00	Biostatistics I	Measurement and Modeling
	12.00–12.20	Coffee Break	
	12.20–13.35	Survey Design and Quality Control	Mathematical Statistics
	13.35–14.30	Lunch	
	14.30–16.00	Artificial Intelligence	Statistical Applications I
	16.00–16.20	Coffee Break	
	16.20–17.20	Statistical Literacy	Poster Presentations I
	18.00–20.00	Reception	
Tuesday	08.30–12.00	Registration	
	09.10–10.00	Invited Lecture	
	10.00–10.30	Coffee Break	
	10.30–12.00	Social Sciences and Humanities	Biostatistics II
	12.00–12.20	Coffee Break	
	12.20–13.20	Network Analysis	Poster Presentations II
	13.20–14.30	Lunch	
	14.30–16.00	Statistical Applications II	Data Science
	16.00–16.10	Closing	

14.00–18.00 **Workshop**
Room 1 *Chair: Organizer*

1. **Formulating Causal Questions and Providing Principled Statistical Answers: A Brave New World**
   Els Goetghebeur

9.00–9.10 **Opening**
Room 1 *Chair: Organizer*

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

1. **The Rise of Causal Inference in Observational Settings: Opportunities and Pitfalls**
   Els Goetghebeur

10.00–10.30 **Coffee Break**

10.30–12.00 **Biostatistics I**
Room 1 *Chair: Els Goetghebeur*

1. **Intelligent Insights on the Role of AI and Data Analytics in Medical Research**
   Jayawant Mandrekar
2. **The Lead Time in Cancer Screening Programmes: The Estimand and the Estimator**
   Maja Pohar Perme and Bor Vratanar
3. **Estimating Lead Time in Cancer Screening Programmes Based on Incidence Comparison: Application to Simulated and Real-World Data**
   Bor Vratanar and Maja Pohar Perme
4. **Optimal Treatment Strategies Following tsDMARD Discontinuation in Rheumatoid Arthritis**
   Diederik De Cock
5. **Operating Characteristics of Common Approaches to Showing the Similarity of Quality Attributes in Biosimilars Development**
   Maša Kušar
6. **Evaluation of Machine Learning Models for Biogeographical Ancestry Inference at Different Resolutions Using a Novel SNP Panel**
   Cosimo Grazzini, Michela Baccini, Daniele Castellana, Giulia Cereda, Giulia Cosenza, Stefania Morelli, Elena Pilli and Giorgia Spera

10.30–11.45 **Measurement and Modeling**
Room 2 *Chair: Irena Ograjenšek*

1. **Gridded Sampling Frames for Global Surveys: Methodology, Validation, and Real-World Applications**
   Arshad Aminu Yakasai and LinChiat Chang
2. **Was It Worth It: A Pooled Analysis of Participant Experience in Cancer Chemoprevention Trials**
   Sumithra Mandrekar and David Zahrieh
3. **Evaluating Web Survey Response Times: Determining the Actual Response Speed**
   Luka Štrlekar and Vasja Vehovar

4. **Why Researchers Should Not Ignore Skewness and Measurement Error in Scale Item Scores**
   Paul Lodder
5. **Agent-Based Modeling of Circular Packaging Systems for ESG Performance Optimization**
   Diana Bratić, Suzana Pasanec Preprotić, Denis Jurečić and Gorana Petković

12.00–12.20 **Coffee Break**

12.20–13.20 **Survey Design and Quality Control**
Room 1 *Chair: Ana Slavec*

1. **Evaluating Data Quality in Probability-Based Online Panels: Systematic Review and Meta-Analysis**
   Andrea Ivanovska, Michael Bosnjak and Vasja Vehovar
2. **A Novel Approach to Optimisation in Multivariate Stratified Sampling**
   Georgi Borros, Sebnem Er and Sulaiman Salau
3. **Descriptive Insights Into Business Survey Contact Data: A Four-Year Review**
   Tadej Prezelj
4. **Percentile-Based Control Charts for Moore and Bilikam Family of Lifetime Distributions Under Random and Progressive First-Failure Censoring With Applications**
   Neeraj Joshi, Aditya Mishra, Taru Singhal and Kashinath Chatterjee

12.20–13.35 **Mathematical Statistics**
Room 2 *Chair: Mihael Perman*

1. **Instrumental Variable Estimation in Compositional Regression**
   Andrej Srakar
2. **On the Improvement of Maximum Likelihood Estimation for the Ola Distribution Parameter with an Application to Medical Data**
   Wararit Panichkitkosolkul, Monthira Duangsaphon and Sudarat Nidsunkid
3. **Some Recent Developments in Change-Point Detection Using Integral Transforms**
   Žikica Lukić and Bojana Milošević
4. **Undominated Copulas With Given Diagonals**
   Damjan Škulj and Matjaž Omladič
5. **The Impact of Violation of Normality Assumption in Artificial Neural Network-Based Multivariate Shewhart Control Chart**
   Sudarat Nidsunkid, Kamon Budsaba and Wararit Panichkitkosolkul

13.35–14.30 **Lunch**

14.30–15.45 **Artificial Intelligence**
Room 1 *Chair: Jay Mandrekar*

1. **Forecasting Strong Subsequent Aftershocks in New Zealand: Preliminary Results**
   Letizia Caravella and Stefania Gentili
2. **Anomaly Identification of AML Cash Threshold-Based Communications**
   Michele Giammatteo and Pasquale Cariello
3. **Stochastic Gradient Langevin Dynamics With Non-Stationary Data**
   Attila Lovas
4. **A Practical Comparison of Variable Importance Techniques Across Modelling Frameworks**
   Markos A. Ktistakis, Andres Laverde Marin, Jaime Suarez, Leonidas Ntziachristos and Georgios Fontaras
5. **Identifying Non-Lexical Entities in Croatian Consumer Health Forums via Machine Learning and Large Language Models**
   Amila Kugic

14.30–15.45 **Statistical Applications I**
Room 2 *Chair: Rok Blagus*

1. **PIGNN-GPR: A Hybrid Machine Learning Framework for Spatio-Temporal PM2.5 Prediction**
   Reetha Thomas and Soudeep Deb
2. **Evaluating Statistical Methods for Estimating Behavioral Strategies From Indirect Reciprocity Experiments**
   Žiga Velkavrh, Aleksa Đorđević and Aljaž Ule
3. **Complex Network Science for the Study of Agricultural Ecosystems**
   Michele Bellingeri
4. **L2-Penalization of the Fixed Effects in Linear Mixed-Effects Models Using the Existing Maximum Likelihood Software**
   Lan Gerdej and Rok Blagus
5. **Limit of Detection in Biological Assays: A Comparison of Statistical Methods for Handling Missing Data**
   Zarja Fabjan, Nataša Kejžar and Stephen Nash

16.00–16.20 **Coffee Break**

16.20–17.20 **Statistical Literacy**
Room 1 *Chair: Vanja Erčulj*

1. **Data Visualization in Kindergarten and Early Primary School**
   Ana Zalokar and Lara Lusa
2. **Exploring Facts and Visuals: A Survey of Infographic Books for Children**
   Lara Lusa
3. **ISLP.SI 2025: The First Slovenian High School Data Visualization Competition**
   Irena Ograjenšek

4. **Visualizing Data in the Business World: Some Worst and Best Practices**
   Bruno Božičnik, Manica Erjavec, Lea Medveščck, Lara Prijatelj, Darja Števančec and Irena Ograjenšek

16.20–16.50 **Poster Presentations I**
Room 2 *Chair: Şebnem Er*

1. **Bangkok Population Forecasting: A Comparative Study of Time Series and Machine Learning Models**
   Kamon Budsaba, Benjamas Tulyanitikul, Siraprapa Manomat, Wikanda Phaphan and Nattanicha Tipwong
2. **Improved Efficiency of Bayesian Estimation in Presence of Laplace Prior Distribution for Type-I Right-Censored Discrete Weibull Regression Model**
   Monthira Duangsaphon, Wararit Panichkitkosolkul and Benjamas Tulyanitikul
3. **Electricity Price Volatility in Slovenia During Crises and Energy Transition: An EGARCH-X Approach**
   Ramanpreet Kaur and Dušan Gabrijelčič
4. **Applying Large Language Models for Structuring Pathology Reports in Slovenian Cancer Registry**
   Maja Jurtela, Tina Žagar, Miran Mlakar, Nika Bric, Mojca Birk and Vesna Zadnik
5. **Ensuring Excellence: How We Assess the Accuracy and Quality of Seasonally Adjusted Data**
   Nikolina Rizanovska
6. **Neuroevolution of Adapters for Large Language Models**
   Katarina Perman

18.00–20.00 **Reception**

9.10–10.00 **Invited Lecture**
Room 1 *Chair: Anuška Ferligoj*

1. **On the Analysis of Data From Sequential Experiments With an Unspecified Number of Observations**
   Tamás Rudas

10.00–10.30 **Coffee Break**

10.30–12.00 **Social Sciences and Humanities**
Room 1 *Chair: Tamás Rudas*

1. **Measuring and Preventing Bullying in Slovenian Elementary Schools: Insights From the kNOwBULLYING Project**
   Vanja Erčulj and Aleš Bučar Ručman
2. **The Role of Basic Psychological Needs and Motivation in PhD Mentoring Outcomes**
   Marjan Cugmas, Sara Atanasova and Luka Kronegger
3. **Insights From the Younger Generation Towards EU's Democracy**
   Andreea-Monica Munteanu and Andreea-Mihaela Niculae
4. **On the Determinants of Financial Literacy: Evidence From Italian Households**
   Gaetano Carmeci, Tommaso Cortivo, Alberto Dreassi and Giovanni Millo
5. **Design and Empirical Verification of a New Methodology for Managing a Retailer's Active Product Assortment**
   Domen Kozjek and Irena Ograjenšek
6. **House Prices and Income: An International Perspective**
   Giovanni Millo

10.30–12.00 **Biostatistics II**
Room 2 *Chair: Maja Pohar Perme*

1. **Net Survival of Colorectal Cancer by Stage in Chile: Addressing a Critical Evidence Gap Using Hospital-Based Cancer Registries**
   Felipe Andrés Medina Marín, Andrea Canals, Natalia Cuadros, Nicolás Silva and Tania Alfaro
2. **Comparison of Measurement Variability Between Groups With Repeated Observations**
   Nataša Kejžar
3. **Analyzing Mortality in Dutch Breast Cancer Patients Using an Extended Multi-State Model**
   Damjan Manevski
4. **The Importance of Using Appropriate Methodology in Interval-Censored Illness-Death Models**
   Gaber Kokovnik and Maja Pohar Perme

5. **The Aalen-Johansen Estimator as an Alternative to Kaplan-Meier in the Presence of Time-Dependent Covariates**
   Ema Požek, Damjan Manevski and Maja Pohar Perme
6. **Using Data Augmentation to Overcome Separation and Singular Random Effects Covariance Matrices in Logistic Mixed-Effects Models**
   Rok Blagus, Georg Heinze and Tina Košuta

12.00–12.20 **Coffee Break**

12.20–13.20 **Network Analysis**
Room 1 *Chair: Marjan Cugmas*

1. **Bibliometric Analysis of a Scientific Journal Based on OpenAlex Data**
   Vladimir Batagelj
2. **A Dynamic Stochastic Blockmodeling Approach: Simulations vs. Theory and Some Lessons About Optimization**
   Aleš Žiberna and Damjan Škulj
3. **Unveiling Complex Connectivity Patterns in Biomedical Systems Through Network Science**
   Marko Gosak
4. **Blockmodeling of the International Trade Network: Looking for a "Trump Effect"**
   Fabio Asthar Telarico and Aleš Žiberna

12.20–12.50 **Poster Presentations II**
Room 2 *Chair: Ana Zalokar*

1. **Some Modifications of INAR(1) Models for Under-Dispersed and Over-Dispersed Time Series of Counts**
   Predrag Popović, Zohreh Mohammadi and Hassan Bakouch
2. **Variable Selection via Fused Sparse-Group Lasso Penalized Multi-State Models Incorporating Clinical and Molecular Data**
   Kaya Miah, Jelle J. Goeman, Hein Putter, Axel Benner and Annette Kopp-Schneider
3. **An Interactive R Package for Bayesian Imputation of Censored Survival Data**
   Jamie Wilson, Shirin Moghaddam and Norma Bargary
4. **Mapping the Mind: Network Learning for Neurological Disorders**
   Bisera Nikoloska and Andrej Kastrin
5. **A Biopsychosocial Model for Stratifying Women and Predicting Outcomes in Women With Gestational Diabetes**
   Ana Munda, Draženka Pongrac Barlovič and Andrej Kastrin
6. **Permutation t-Test: A Simulation Study for Comparing Two Means Under Non-Ideal Conditions**
   Alja Nike Kastrin, Amer Mujagić and Maja Pohar Perme

13.20–14.30 **Lunch**

14.30–16.00 **Statistical Applications II**
Room 1 *Chair: Nataša Kejžar*

1. **Machine Learning Methods Applied in the Real Estate Market**
   Ion-Florin Răducu
2. **Profiling of Adolescents With Symbolic Data Analysis of Self-Assessments and App Usage**
   Simona Korenjak-Černe, Jasminka Dobša, Miranda Novak and Maja Buhin Pandur
3. **An Alternative to Classical Intention-to-Treat Analysis for Comparing a Time-to-Event Endpoint in Precision Oncology Trials**
   Marilena Müller
4. **Contrast Testing in Julia: Implementing GLHT Functionality**
   Jakob Peterlin
5. **Developing a Dashboard of Key Performance Indicators for Coronary Artery Disease Care Using Administrative Data in Slovenia**
   Janez Bijec, Petra Došenović Bonća and Irena Ograjenšek
6. **Kurtosis: Mainly Misunderstood and Practically Useless**
   Gaj Vidmar and Bor Vratanar

14.30–16.00 **Data Science**
Room 2 *Chair: Andrej Blejec*

1. **Permutation Entropy and Statistical Complexity Analysis of Sentinel-2 Time Series for Detecting Vegetation Pest Diseases: The Case of Toumeyella parvicornis**
   Luciano Telesca, Nicodemo Abate and Rosa Lasaponara
2. **Local Differential Privacy for Trajectory Anonymization With Map-Matching Techniques**
   Gabriele Gühring, Andreas Heinrich and Di Hu
3. **Forecasting Storms With Meteorological Variables and Digital Attention: A Multistage Statistical and Neural Hybrid Framework**
   Soudeep Deb and Lizan Meryl Pereira
4. **Monitoring FAIR Data Practices: Lessons From a Preliminary Study at the University of Primorska**
   Ana Slavec and Haris Zukić
5. **Statistical Analysis of Emotional Engagement in Interactive vs. Non-Interactive Documentaries**
   Matjaž Kljun, Una Vuletić and Klen Čopič Pucihar
6. **Application of Machine Learning to Fundamental Analysis of Securities**
   Aleksandr Panteleev

16.00–16.10 **Closing**
Room 1 *Chair: Organizer*

ABSTRACTS

Workshop

Formulating Causal Questions and Providing Principled Statistical Answers: A Brave New World

Els Goetghebeur

Ghent University, Gent, Belgium
els.goetghebeur@ugent.be

Causal inference has taken off over the past decade. A seemingly never ending stream of new and complex methods enters the literature allowing to draw causal conclusions from observational data, as long as one is willing to make causal assumptions in context. Many of these methods are derived from basic pillars. Using the potential outcomes framework, we describe principled definitions of causal effects and of estimation approaches classified according to whether they invoke the no unmeasured confounding assumption (including outcome regression and propensity score-based methods) or an instrumental variable (pseudo randomisation). Starting from (sequential) point exposures, we discuss interpretation, challenges and potential pitfalls. We illustrate application using a »simulation learner«, that mimics an existing study of the effect of various breastfeeding interventions on a child's later development. This involves a typical simulation component with generated exposure, covariate, and outcome data inspired by a randomised intervention study with intercurrent events. The simulation learner thus generates various (linked) exposure types with a set of possible treatment values per observation unit, from which observed as well as potential outcome data are generated. It thus provides true values of several causal effects. There will be several hands-on exercises allowing to also work with available R code for data analysis.

Invited Lecture

The Rise of Causal Inference in Observational Settings: Opportunities and Pitfalls

Els Goetghebeur

Ghent University, Gent, Belgium
els.goetghebeur@ugent.be

Causal inference has taken off over the past decades. New and ever more powerful statistical methods have been proposed with well developed properties under well-defined but sometimes complex and typically untestable assumptions. There is now a wealth of promising methods (with software) to choose from. Few are easy to justify and implement in observational settings, however. Fortunately, the new emphasis on the estimand framework helps steer statistical research towards more direct real world relevance and transparency. Challenges arise, for instance, when estimating treatment effects on repeated outcome measures in a mortal population. Policy oriented estimands are then designed to involve continued outcome measures after non-terminal intercurrent events occur. Popular alternative estimands yield different answers for a different purpose. We consider limitations of the (time-varying) survivor average causal effect and of alternative principal stratum like estimands. Given their restrictions and those of other known estimands, such as the hypothetical estimand often presented following mixed models, we present the two-dimensional outcome (survival time and disease outcome while alive) as a basis for causal estimands with prime interpretation and relevance. For estimation, we compare »double inverse probability weighting« with »outcome regression followed by adapted standardization« which handle censoring and death in a distinct manner. As an important secondary problem we discuss the often joint occurrence of missing data and intercurrent events which may lead to non-positivity and/or missingness dependent on underlying values. On the more technical side, we point to naïve combinations of Inverse Probability Weighting and outcome regression which do not generally lead to double robust estimation, especially for right censored survival outcomes. With a case study on the evaluation of Patient Reported Outcomes in late stage oncology, we will highlight sensible estimands, solutions currently available for estimation and suggestions for further research.

Biostatistics I

Intelligent Insights on the Role of AI and Data Analytics in Medical Research

Jayawant Mandrekar

Mayo Clinic, Rochester, MN, United States
mandrekar.jay@mayo.edu

The integration of Artificial Intelligence (AI) and data analytics is reshaping clinical trials, providing innovative solutions to persistent challenges in drug development and patient care. This talk highlights the transformative role of AI in improving trial design, streamlining patient recruitment, enabling real-time monitoring, and enhancing predictive analytics. By processing vast amounts of clinical data, AI-powered algorithms can identify patterns, optimize trial operations, and support more informed decision-making. We will explore how these technologies contribute to more adaptive and efficient clinical trials, ultimately accelerating the path from discovery to regulatory approval. Additionally, the session will address key ethical and regulatory considerations, including algorithmic bias, data privacy, and the need for transparent, trustworthy AI systems. As the industry evolves, the use of AI and data-driven approaches promises to make trials more personalized, inclusive, and cost-effective. By embracing these tools, clinical research can achieve faster, more reliable outcomes, offering new hope for patients and advancing the future of medicine.

The Lead Time in Cancer Screening Programmes: The Estimand and the Estimator

Maja Pohar Perme and Bor Vratanar

University of Ljubljana, Ljubljana, Slovenia
maja.pohar@mf.uni-lj.si

In cancer screening programmes, participants are regularly screened every few years using medical tests to detect early signs of cancer. Without screening, cancer would likely progress undetected until symptoms appear. The interval between early detection based on screening test and the eventual onset of symptoms, had screening not been conducted, is known as lead time, which is the estimand of interest. Estimating lead time is challenging for two reasons. First, it is hypothetical—there is no direct way to measure when symptoms »would have« occurred for screen-detected cancers as treatment starts right after screen diagnosis. Second, some screen-detected cancers are overdiagnosed: they would have died due to other causes before hypothetical symptoms take place. As a result, lead time is not only hypothetical but also subject to competing risk, even if it could have been observed. We propose a new method, called the MOCCI method (Minimizing Observed and Counterfactual Cancer Incidence) that uses cancer incidence data to estimate lead time. When a new screening programme is introduced, incidence rates typically shift to younger ages, since cancers are diagnosed at an earlier age due to screening. This shift in cancer incidence, stratified by age groups and calendar years, carries vital information about lead time. The MOCCI method aims to find a distribution of lead time distribution that best explains the observed shift while accounting for overdiagnosis, using maximum likelihood estimation principles. This talk introduces the core ideas of the proposed method for estimating lead time. We explain the source of information used for estimation, outline the required datasets, and describe the construction of the likelihood function.

Estimating Lead Time in Cancer Screening Programmes Based on Incidence Comparison: Application to Simulated and Real-World Data

Bor Vratanar and Maja Pohar Perme

University of Ljubljana, Ljubljana, Slovenia
bor.vratanar@mf.uni-lj.si

We evaluate the performance of the MOCCI method (Minimizing Observed and Counterfactual Cancer Incidence) through simulations and illustrate its practical application to data from the Slovenian breast cancer screening programme. Simulation study: In each run, we generated two datasets: in one, subjects were assumed to be invited to the screening programme and screened biennially for breast cancer; in the other, subjects were assumed not to be invited, and no screening was performed. The date and age at cancer diagnosis were in the second dataset shifted by the lead time distribution, assumed to follow exponential distribution. The first aim of the simulation study was to show that the MOCCI method satisfies the standard maximum likelihood estimator properties (i.e. consistency, asymptotic normality) in a simple setting. The second aim was to assess the consistency of the MOCCI method in a more complex setting where some cancers are non-progressive. The results showed that the MOCCI method is consistent in both settings; the rate of convergence, however, drops dramatically with the inclusion of non-progressive cancers. Practical application: The Slovenian breast cancer screening programme began in 2008 and invites women aged 50–69 to attend mammography screening every two years. Initially, the programme was limited to the central region of the country, but was gradually expanded to cover all regions by 2018. This gradual rollout created a natural experiment: women who were invited to screening form the experimental group, while those who had not yet been invited serve as the control group. The MOCCI method was applied to these data to estimate lead time and the proportion of non-progressive cancers. The estimated mean lead time was 1.67 years (95 % CI: 1.2–2.6), and the estimated proportion of non-progressive cancers was 0.34 (95 % CI: 0.26–0.41).

Optimal Treatment Strategies Following tsDMARD Discontinuation in Rheumatoid Arthritis

Diederik De Cock

Vrije Universiteit Brussel, Brussels, Belgium
diederik.de.cock@vub.be

tsDMARDs represent an important class in Rheumatoid Arthritis (RA) treatments. It is unknown which specific bDMARD or mode-of-action should be selected after stopping a JAKi. We aimed to examine if clinical response differs between advanced therapies that are initiated after stopping a JAKi. Patients were included from the electronic platform »Tool for Administrative Reimbursement Drug Information Sharing« (TARDIS). Patients were selected for analysis if they had stopped JAKi therapy and initiated a subsequent therapy. Patients were grouped by TNFi, T/B cell therapy, IL6-inhibition or JAKi therapy. The DAS28 response and proportion of patients in remission at the first 3 follow-up (after 3–6 months, after 15–18 months, and 27–30 months) moments were compared between groups. Remission was defined as DAS28 < 2.6. Sensitivity analyses took treatment line into account. Mixed-effects models were applied to assess the association between time, treatment groups, and their interaction on clinical outcomes, with random intercepts and slopes included to account for patient-level variability and repeated measures over time. In total, 2389 RA patients who had stopped JAKi therapy could be included. Of these, 1274(53 %) patients had follow-up data. Patients on rituximab were excluded as these were retreated and their data collected only in case of flare, following Belgian reimbursement criteria. Hence, the B/T cell group was now reduced to only Abatacept. An unadjusted mixed-effects model revealed that Abatacept had a lower DAS28 response (−0.151, 95 % CI: −0.280 to −0.022) compared to the tsDMARD reference group ($p = 0.021$). An adjusted mixed-effects model adding baseline disease duration, ESR, CRP, PGA, SJC28 and TJC28 as random effects and treatment line as fixed effect showed again that Abatacept had a lower DAS28 response (−0.299, 95 % CI: −0.499 to −0.003) compared to the tsDMARD reference group ($p = 0.021$). This real-world study highlights that abatacept appears to be a less favourable choice after JAKi cessation.

Operating Characteristics of Common Approaches to Showing the Similarity of Quality Attributes in Biosimilars Development

Maša Kušar

University of Ljubljana, Ljubljana, Slovenia
masa.kusar@mf.uni-lj.si

Biosimilar drugs are assumed to be different, but sufficiently similar to the original approved drug to allow for marketing authorisation. In these cases, bioequivalence is neither presumed nor required, because the manufacturing process doesn't allow for the production of an identical drug substance. The current FDA and EMA guidelines require various approaches to demonstrate similarity without prescribing a single unifying criterion. Among the important parts of a marketing authorisation application are clinical trials confirming safety and efficacy of the candidate molecule as well as a physicochemical comparison of a wide array of quality attributes (QAs) of the originator and candidate drugs. According to proposed change in EMA guidelines, it may now be possible to waive or significantly reduce clinical trials and base the biosimilarity evaluation in a much larger part on the demonstration of physicochemical difference. To ensure this meets the standard of showing that the substances are »highly similar«, the analytical and statistical methods used in the process need to be sufficiently sensitive to identify a significant difference, if such a difference in present. While analytical methods are constantly evolving, the statistical methods used are rather rudimentary and further limited by the limited sample sizes used. We conducted a simulation study to estimate the operating characteristics (FPR and FNR) of the most commonly used approached to biosimilarity assessment. We then extended this simulation to estimate the substantive equivalents of the positive and negative predictive value of the results obtained in such a comparability exercise. Our results show that some of the most commonly used approaches are anti-conservative and too often lead to a conclusion of similarity even in QAs that vary widely between the originator and the candidate drugs.

Evaluation of Machine Learning Models for Biogeographical Ancestry Inference at Different Resolutions Using a Novel SNP Panel

Cosimo Grazzini, Michela Baccini, Daniele Castellana, Giulia Cereda, Giulia Cosenza, Stefania Morelli, Elena Pilli and Giorgia Spera

University of Florence, Florence, Italy
cosimo.grazzini@unifi.it

BioGeographical Ancestry (BGA) refers to an individual's biologically inherited ethnic component and can be inferred from DNA, particularly through Single Nucleotide Polymorphism (SNP) markers. BGA plays a vital role in various fields, such as population studies, medicine, epidemiology, and forensics. Recent advances in gene sequencing technologies have significantly expanded access to high-resolution genomic data, revolutionizing BGA inference and necessitating the adoption of appropriate methods for accurate prediction. This study investigates the evaluation of a novel SNP panel combined with supervised Machine Learning (ML) algorithms to infer BGA in a single-step approach and on a global scale, both at inter-continental and more detailed levels. Several ML models were applied, including Categorical Naive Bayes, Penalized Multinomial Logistic Regression, Linear Support Vector Machines, Random Forests, and tree-based Gradient Boosting. Promising results were obtained at both levels of BGA resolution, highlighting the effectiveness of the SNP panel coupled with the ML approach. Analysis of misclassification patterns revealed interesting aspects, strongly suggesting that the observed inaccuracies stem from the combination of the complexity of inferring BGA and its high-dimensional uncertainty, encompassing spatio-temporal, genetic, and information availability factors. These findings support the potential for future research aiming to infer BGA at finer levels of detail. In conclusion, this study highlights the proposed SNP panel and ML methods as valuable tools for experts in several applied contexts, while laying the groundwork for future work on significantly larger SNP datasets.

Measurement and Modeling

Gridded Sampling Frames for Global Surveys: Methodology, Validation, and Real-World Applications

Arshad Aminu Yakasai and LinChiat Chang

Independent consultant, Cape Town, South Africa
contact@linchiat.com

Sample designs for probability-based national surveys require valid sampling frames that provide complete and unbiased coverage of the entire country. Such national frames are often lacking in countries with outdated, incomplete or unavailable census data. We present the development of a globally consistent gridded area sampling frame, built to support multi-stage random cluster sampling with probability proportional to size (PPS). Leveraging a high-resolution micro area dataset with 1 km^2 granularity, we create a mega sampling frame that integrates geospatial and demographic attributes including population density, urbanicity, administrative markers, proximity to key infrastructure such as health facilities, schools, markets, and essential utilities such as water and power plants. Further, precipitation and temperature data is incorporated to enable stratification by climate elements, or identify areas where human settlements have been especially hard hit and possibly displaced by floods or drought. Beyond sampling, multiple layers of geo-markers can support survey logistics and planning from the outset, as well as quality assurance and monitoring during data collection. By anchoring survey samples within spatially precise, verifiable, and transparent sampling frames, our approach can enhance equity and accuracy in conducting population surveys across diverse domains including health, education, climate change and more. We demonstrate the validity and limitations of this approach using a few case studies. First, we present contrasting case studies in Nigeria where no recent census data is available vs. South Africa where micro data from the 2022 census is available, to evaluate the veracity of PPS samples drawn from this approach using available benchmarks. Further, we visit primary sampling units (PSUs) from a 2024 survey in Kenya and Côte d'Ivoire to tackle limitations of this approach that can arise in practice. These findings underscore the transformative potential of geo-spatially integrated sampling frames—with proven benefits for survey sampling, logistics, and quality assurance.

Was It Worth It: A Pooled Analysis of Participant Experience in Cancer Chemoprevention Trials

Sumithra Mandrekar and David Zahrieh

Mayo Clinic, Rochester, MN, United States
mandrekar.sumithra@mayo.edu

Individual participant-level data from 13 early-phase chemoprevention trials targeting four disease sites were included in the current study. The 5-item »Was It Worth It?« (WIWI) questionnaire was administered at the end of each trial's intervention period or at the time of early termination for participants who ended the intervention early. The binary outcome, satisfied overall, was defined as a participant who answered yes to the first three questions on the WIWI questionnaire: (Q1) »Was it worthwhile for you to participate in this research study«; (Q2) »If you had to do it over, would you participate in this research study again«; and (Q3) »Would you recommend participating in this research study to others«. Seventeen factors covering trial-design, baseline, and on-study features were identified based on subject matter knowledge and were interrogated with the random forests algorithm. A hierarchy of features based on quantification of the importance of their effects on being satisfied overall was constructed, and a multiple logistic regression model was used to understand the impact of these features on participant satisfaction. 652 (94.4 %) completed the WIWI questionnaire, of whom, 493 (75.6 %) were White, non-Hispanic or Latino; 193 females (29.6 %), 121 (17.5 %) were $\geq$ 65 years, and 517 (79.3 %) participated in a placebo-controlled trial. One-third of these participants were enrolled outside the US. 85 % indicated that they were satisfied overall. After controlling for age, sex, and intervention duration, the odds of not satisfied overall was higher for (i) participants who terminated the intervention early, (ii) spent > 5 % of the intervention duration experiencing adverse events, (iii) had cumulative number of preintervention AEs experienced $\geq$ 1, (iv) Black/Asian/> 1 race, non-Hispanic or Latino. Knowing the set of features associated with satisfaction from our large series of geographically and demographically diverse participants, can inform design of subsequent trials and develop strategies to improve accrual, retention, adherence, and diversity.

Evaluating Web Survey Response Times: Determining the Actual Response Speed

Luka Štrlekar and Vasja Vehovar

University of Ljubljana, Ljubljana, Slovenia
luka.strlekar@fdv.uni-lj.si

Web surveys can capture digital traces, known as paradata, which record respondents' activities when completing questionnaires and provide insights into their behavior. Among the various types of paradata, response times (RTs)—the time it takes to complete a question, a page or the entire survey—are the most used in practice. RTs are particularly often studied in relation to response quality. To accurately assess the relationship between RTs and response quality, it is important to properly analyse RTs. Although web survey tools can usually measure RTs accurately at the page level, the main dilemma is whether these default RTs can be unreservedly used in further analyses. This issue is inadequately addressed in the literature, which generally assumes simple surveys and engaged respondents. The RTs and related response speeds should only reflect respondents' cognitive processes and questionnaire characteristics, and exclude confounding factors that could affect comparability, as analyses based on RTs are only meaningful under the assumption that all respondents answer the same questionnaire. These factors include: (i) pauses, multitasking behavior, and (ii) backtracking cause recorded RTs to be misestimated; (iii) answering open-ended questions lengthens RTs and consequently gives the appearance of a slower response speed; while (iv) not answering questions; and (v) not being exposed to questions due to branching give the appearance of a faster response speed. By removing the confounding effects of these factors (e.g., by subtracting pause duration), we introduce the concept of »actual response speed«—the speed when respondents are engaged in the response process as their primary cognitive activity and are exposed to standardized cognitive tasks (i.e., questions). This approach ensures comparable survey conditions for all respondents and enables the correct use of the adjusted RTs in further analyses. We also develop standardized practical solutions for addressing confounding factors in R.

Why Researchers Should Not Ignore Skewness and Measurement Error in Scale Item Scores

Paul Lodder

Tilburg University, Tilburg, The Netherlands
paultwinlodder@gmail.com

In the medical and social sciences, researchers commonly study associations between latent variables measured with multi-item scales. Examples are scales measuring depressive symptoms or quality of life. Such items contain measurement error and often show skewed ordinal score distributions. However, researchers commonly ignore these characteristics by applying statistical analyses to total scale scores. We used computer simulations to investigate the extent to which ignoring measurement error and skewness introduces bias in estimated regression coefficients for the main and interaction effects of two latent variables. We simulated data on two independent latent variables, each associated with a dependent latent variable through both main and interaction effects. We simulated four levels of skewness in the ordinal item score distributions. Main and interaction effects on the dependent variable were estimated using OLS regression on the total scale scores and structural equation models for continuous (SEM) or categorical (catSEM) item scores. In addition to modeling relations between latent variables, both SEM approaches model the relation between each latent variable and the individual item scores measuring it. The relative bias in the estimated effects was assessed across levels of item score skewness. When ordinal item scores were normally distributed, both SEMs yielded unbiased estimates, while OLS regression underestimated both the main and interaction effects. When item scores were skewed, OLS regression underestimated both effects even more. Although linear SEM prevented bias due to measurement error, it still produced biased estimates because the skewed ordinal items were treated as continuous items. In contrast, categorical SEM yielded relatively unbiased estimates of both main and interaction effects. Despite the common use of total scale scores in statistical modeling, we illustrate the importance of using categorical SEM when estimating associations between latent variables measured with skewed ordinal item scores.

Agent-Based Modeling of Circular Packaging Systems for ESG Performance Optimization

Diana Bratić, Suzana Pasanec Preprotić, Denis Jurečić and Gorana Petković

University of Zagreb, Zagreb, Croatia
diana.bratic@grf.unizg.hr

This study presents an agent-based simulation model of the packaging life cycle in the graphic industry, designed to support the technical evaluation of ESG (Environmental, Social, Governance) performance. The model includes four types of agents, producers, users, recyclers, and logistics operators, who interact over 1,000 time steps, simulating the flow of packaging through production, use, return, and recycling phases. Each simulation run involves 500 agents, and 300 iterations are conducted with varying input conditions to account for system variability. The agents operate according to predefined rules and probabilistic parameters, based on empirically observed behavior in closed-loop systems. These behavioral rules can be adapted to reflect different regulatory or operational contexts. The model tracks indicators such as the amount of material recovered, total waste generated, average duration of packaging use, and estimated carbon emissions. Across all scenarios, the average material recovery rate was 42.8 %, with the highest performance reaching 55% in situations where return rates exceeded 60% and contamination levels remained low (below 18 %). A moderate negative correlation was observed between contamination and carbon footprint, confirming the importance of effective sorting and clean return flows. Sensitivity analysis revealed that increasing return rates from 45 % to 60 % resulted in an 11–14 % reduction in total waste compared to the baseline scenario. The results also indicated that combining improvements in collection logistics and contamination control provided stronger ESG benefits than focusing on either factor alone. Different packaging scenarios were tested under controlled conditions, allowing for a comparison of outcomes using descriptive statistics and correlation analysis. This simulation framework supports informed, sustainability-oriented decision-making and offers a flexible tool for evaluating circular packaging strategies under realistic uncertainty. It is adaptable to various packaging types and production environments, making it relevant for companies aiming to implement circular economic principles in their production and supply chains.

Survey Design and Quality Control

Evaluating Data Quality in Probability-Based Online Panels: Systematic Review and Meta-Analysis

Andrea Ivanovska, Michael Bosnjak and Vasja Vehovar

University of Ljubljana, Ljubljana, Slovenia
ai67517@student.uni-lj.si

Probability-based online panels (PBOPs) are increasingly adopted by academic and official statistics communities as a cost-effective alternative to traditional face-to-face and telephone surveys. While PBOPs use random sampling, their reliance on online data collection raises concerns about estimate accuracy. This study aimed to evaluate the quality of survey estimates from PBOPs by quantifying their deviation from established benchmarks. We conducted a systematic review and meta-analysis of 44 studies comprising 1,897 effect sizes comparing PBOP estimates to external benchmarks. The primary outcome was the absolute value of the relative bias (RB), which standardizes error relative to benchmark values. A three-level meta-analytic model was used to account for study-level, item-level, and sampling variance. Moderator analyses examined the effects of item sensitivity, measurement level, country, and topic. The pooled absolute RB across all estimates was 23.14 %, indicating a substantial level of bias. Most heterogeneity was due to within-study item-level differences, rather than study-level differences. Questions with a lot of sensitivity showed significantly higher RB, while items measured on ordinal scales showed significantly lower RB than those with nominal response formats. Country and topic did not moderate RB levels significantly. Sensitivity analyses excluding the top 5 % of extreme RB estimates produced lower overall RB values, indicating that a few highly influential cases highly contributed to the overall bias level. These findings provide critical insights into the limitations of PBOPs for certain survey items, particularly those involving sensitive or low-frequency behaviours. Item construction, especially regarding sensitivity and measurement scale, plays a vital role in data quality. Understanding the methodological risks of PBOPs and developing strategies to improve online survey design are essential steps toward enhancing the accuracy and reliability of online data collection.

A Novel Approach to Optimisation in Multivariate Stratified Sampling

Georgi Borros, Sebnem Er and Sulaiman Salau

University of Cape Town, Cape Town, South Africa
sebnem.er@uct.ac.za

Stratified sampling is a popular survey sampling method that can enhance efficiency in estimation and survey administration. As survey research often consists of multiple variables of interest, optimal stratified sampling subsequently becomes a complex multivariate grouping and sample allocation problem. In stratified sampling, the procedure involves partitioning a heterogeneous population into more homogeneous subgroups and then allocating the sample size across strata with the ultimate aim of estimating population parameters for one or more study variables. In the multivariate setting, where we have several variables of interest, survey statisticians still continue to offer new insights to both problems of strata formation and sample allocation either separately or jointly. Finding the best sample allocation becomes problematic for a simple reason that the best allocation for one characteristic will not in general be best for another. Besides, the size of each stratum, that determines the sample sizes to be allocated, heavily depends on the strata formation. There are numerous studies that deal with the allocation problem using compromise optimisation solutions given the strata. However, there are very few papers that deal with solving the two problems simultaneously. Our research offers a novel method to tackle these two problems simultaneously in the multivariate context, while considering various objective functions to best capture the multivariate structure of the data. The work aims to be a resource for implementation by sampling practitioners, while more broadly contributing to research in multivariate optimisation.

Descriptive Insights Into Business Survey Contact Data: A Four-Year Review

Tadej Prezelj

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
tadej.prezelj@gov.si

As part of the Annual Statistical Programme, the Statistical Office of the Republic of Slovenia (SURS) conducts approximately 70 statistical surveys annually, inviting businesses to provide data. For the majority of them, data collection is supported by the Data Collection Division. Particularly the Enterprise Cooperation Section plays a pivotal role in this process, ensuring timely acquisition and data accuracy. By systematically recording and monitoring both outgoing and incoming contacts, the contact center enables operational assessment of the communication load associated in business surveys. These records offer valuable insight into the intensity and efficiency of the communication process. In 2024, the contact center provided substantive and technical support to 12,382 businesses for 46 surveys, recording 24,169 incoming contacts. Additionally, it initiated 68,249 outgoing contacts with 12,421 businesses. The aim of this article is to analyse patterns in outgoing and incoming contacts related to business surveys, to indetify practical opportunities to optimise work allocation and improve communication strategies. Findings indicate that, after a noticeable decrease in communication volume in the post-pandemic period (2023), the total number of recorded contacts increased slightly again in 2024. However, considering larger sample sizes in recent surveys, this increase is less significant in proportional terms. Compared to 2023, there was a modest rise in the average number of contacts per reporting unit in 2024, reaching 5.1 contacts among contacted units and 2.0 contacts within the whole sample. In addition, the distribution of contacts among reporting units was analysed to determine its consistency with the Pareto principle, which posits that a small share of units accounts for a large share of outcomes. Preliminary results revealed a skewed distribution, with a minority of businesses responsible for most recorded contacts, mirroring patterns commonly observed in broader economic-statistical contexts.

Percentile-Based Control Charts for Moore and Bilikam Family of Lifetime Distributions Under Random and Progressive First-Failure Censoring With Applications

Neeraj Joshi, Aditya Mishra[1], Taru Singhal[2] and Kashinath Chatterjee

Indian Institute of Technology Delhi, New Delhi, India
[1]mt1221271@iitd.ac.in, [2]mt1221922@iitd.ac.in

We propose a new class of Shewhart-type control charts for monitoring percentiles in lifetime distributions from the Moore and Bilikam family, under both random and progressive first-failure censoring. These charts are constructed using maximum likelihood estimators (MLEs) of the percentile function, and their asymptotic properties are derived to ensure statistical validity under censored conditions. A comprehensive simulation study, implemented in Python, evaluates the in-control (IC) performance of the proposed charts across various percentile levels, false-alarm rates (FARs), and sample sizes. Sensitivity analyses confirm robustness across starting values, sample sizes, and censoring levels. We further assess the out-of-control (OOC) detection capability under distributional shifts. To demonstrate practical relevance, we apply the methodology to two real-world survival datasets—one on healthcare survival and another on industrial reliability—using bootstrapped subgroups and the two separate censoring schemes. The results show that the proposed charts effectively identify OOC signals even under substantial censoring, establishing them as promising tools for lifetime monitoring under censoring.

Mathematical Statistics

Instrumental Variable Estimation in Compositional Regression

Andrej Srakar

University of Ljubljana, Ljubljana, Slovenia
andrej.srakar@ier.si

Time use surveys are used in many areas of economics, including economics of health and long-term care. If analyzed in a regression context, time use survey data suffer from the problem of spurious correlation noted in early works of Aitchison (1986). This problem leads to a need for compositional regression perspective on a geometric simplex. We develop an instrumental variable compositional regression model, building on two strands of literature with applications for health economics and economics of long-term care. We extend Florens and Van Bellegem (2015) functional instrumental variables model to compositional data setting where either or both independent and dependent variables are of compositional nature. We show there exist two ways of deriving compositional IV's, one using isometric log-ratio transform and Chesher et al. (2013)'s IV model of multiple discrete choice; and another deriving from the recent literature on compositional functional data in Bayes spaces (Machalova et al., 2021). We follow the latter and show that estimation, similar to the one of Florens and Van Bellegem leads to an ill-posed inverse problem with known but data-dependent operator. We resolve this in a context of multiplication by an instrument-dependent operator and by a penalized least squares estimation, and we also extend the notion of instrument strength to compositional setting. We establish appropriate central limit theorem in the context of Bayes spaces instead of more conventional Hilbert spaces and study the finite sample performance in a Monte Carlo simulation setting. Our application studies relationship between long term care for older people and paid work, using Slovenian time use survey from Survey of Health, Ageing and Retirement in Europe (SHARE).

On the Improvement of Maximum Likelihood Estimation for the Ola Distribution Parameter with an Application to Medical Data

Wararit Panichkitkosolkul, Monthira Duangsaphon and Sudarat Nidsunkid

Thammasat University, Pathumthani, Thailand
wararit@mathstat.sci.tu.ac.th

Maximum Likelihood Estimation (MLE) is a fundamental method in statistical inference that aims to estimate the parameters of a statistical model by finding the values that maximize the likelihood function, which measures the probability of observing the given data under different parameter values. While the method is widely used, its reliability decreases when applied to small or moderate sample sizes, where biased estimates are more likely to occur. To improve the accuracy of the maximum likelihood estimator for the Ola distribution, this study employs two bias-correction methods—namely, the Cox-Snell correction and the parametric bootstrap technique. Monte Carlo simulation was conducted to evaluate the estimators in terms of their average bias and root mean square error (RMSE). The findings demonstrate that the proposed bias-corrected estimators effectively reduce both bias and RMSE, leading to more accurate parameter estimates. The parametric bootstrap method proved superior across all scenarios, including small and moderate sample sizes. The bias-corrected estimators were also applied to medical data, highlighting their real-world applicability.

Some Recent Developments in Change-Point Detection Using Integral Transforms

Žikica Lukić and Bojana Milošević

Parexel Serbia, Belgrade, Serbia • University of Belgrade, Belgrade, Serbia
zikicamaster@gmail.com

Change-point inference has many real-world applications, including finance, medicine, algorithmic trading, and other domains. The integral transform method has recently gained popularity, particularly in the context of analyzing complex data structures. In this study, we introduce novel statistical tests for detecting change-points in the distribution of a sequence of independent observations of various types. These new tests are based on integral transforms and provide a practical and consistent way of identifying distributional changes. We focus on recommending tailored solutions for real-world use with the aim of making the methods practical and accessible. In addition, we discuss the possible limitations of the proposed method, especially in comparison to some other existing methodologies. This includes considerations of performance, applicability, and potential constraints in specific scenarios.

Undominated Copulas With Given Diagonals

Damjan Škulj and Matjaž Omladič

University of Ljubljana, Ljubljana, Slovenia
damjan.skulj@fdv.uni-lj.si

Given a strictly increasing track $B_\varphi = \{(x, \varphi(x)) \mid x \in [0,1]\}$ with track section $\delta(x) = C(x, \varphi(x))$, we search for undominated copulas corresponding to this track section. We can express each of these copulas as C_ψ in terms of an increasing function ψ as a parameter. There exists a region $\mathcal{R}_\psi$ bounded above and below by two increasing functions such that outside this region, C_ψ equals the classical Fréchet–Hoeffding upper bound M, while inside the region, C_ψ is expressed in terms of ψ. In the case of the diagonal track $\varphi(x) = x$, we can also prove that *all* copulas of the form C_ψ for some ψ are undominated with the same diagonal section.

The Impact of Violation of Normality Assumption in Artificial Neural Network-Based Multivariate Shewhart Control Chart

Sudarat Nidsunkid, Kamon Budsaba and Wararit Panichkitkosolkul

Kasetsart University, Bangkok, Thailand
sudarat.n@ku.th

The most familiar multivariate control chart is a multivariate Shewhart control chart for monitoring the mean vector of the process where p related quality characteristics are controlled jointly. A multivariate normal distribution of observations is an important assumption which is often made before applying multivariate control charts. In recent years, machine learning (ML) techniques have been developed to be used in control charts for many reasons, such as dimension reduction, change point estimation, signal detection, or identification. An artificial neural network (ANN) is one of the most popular ML techniques that mimics human brain activity. This research study the impact of violations in multivariate normal assumptions in ANN based multivariate Shewhart control chart, various multivariate non-normal distributions are used to generate random vectors of quality characteristics and getting ANN structure's output. An average run length (ARL) is derived to investigate how robust or sensitive ANN multivariate Shewhart control chart are to violations of the multivariate normal assumption. The results show that the departure from normality can affect the statistical performance of the ANN multivariate Shewhart control chart in different ways. If the violation occurs when sampling from a more heavy-tailed distribution, then the process may have a biased underestimate of ARLs. If the violation occurs when data are sampled from skewed-right distributions, the process may also have a biased underestimate of ARLs. However, if the violation occurs when data are sampled from skewed-left distributions, there is a significant increase in ARLs.

Artificial Intelligence

Forecasting Strong Subsequent Aftershocks in New Zealand: Preliminary Results

Letizia Caravella and Stefania Gentili

National Institute of Oceanography and Applied Geophysics, Udine, Italy
letizia.caravella@outlook.com

NESTORE (NExt STRong Related Earthquake; Gentili et al., 2023) is an algorithm for the probabilistic forecasting of strong aftershocks after a major seismic event. The algorithm evaluates nine features that characterize the seismic activity at increasing time intervals after the mainshock. These features are then analyzed using a combination of supervised machine learning and statistical validation to estimate the probability that an initial strong event of magnitude Mm will be followed by another event of magnitude $\geq$ Mm−1 within a given space-time window associated with the corresponding seismic cluster. If such an aftershock occurs, the cluster is classified as »Type A« (indicating a higher potential risk). The algorithm outputs the probability that the cluster is of type A. New Zealand is one of the most seismically active regions in the world, located on the boundary between the Australian and Pacific tectonic plates. The complex country's seismicity has generated so far earthquakes up to magnitude 7.8. Understanding and forecasting seismic activity is therefore critical for risk assessment and mitigation efforts. We present here the preliminary results of the applications of NESTORE to the seismicity of New Zealand. We split the dataset of the clusters in the area between training (1988–2015) and testing (2016–2025) for a retrospective forecasting. We refined the training dataset using the outlier detection method REPENESE (RElevant features, PErcentage class weighting, NEighborhood detection and SElection), which was developed for skewed distributions of feature values (Gentili et al., 2025). We found that twelve hours after the first earthquake 88 % of the clusters were correctly classified. Funded by the RETURN project European Union Next-GenerationEU (National Recovery and Resilience Plan, PE0000005).

Anomaly Identification of AML Cash Threshold-Based Communications

Michele Giammatteo and Pasquale Cariello

Bank of Italy, Rome, Italy
michele.giammatteo@bancaditalia.it

This project addresses the challenge of identifying cash transaction anomalies – such as unusually frequent or large cash deposits and withdrawals – that could indicate potential money laundering activities in the Italian financial sector. Specifically, it focuses on the monthly Cash Threshold-Based Communications submitted by banks to the Financial Intelligence Unit of Italy (UIF), using them as a primary data source for detecting unusual transactions. To enhance detection accuracy, unsupervised machine learning techniques are employed, with a particular emphasis on the Isolation Forest algorithm, which is well-suited for identifying outliers in complex and large datasets. The model integrates several key information provided by the database of cash threshold-based reports (historical transaction records of relevant subjects, transaction types, client characteristics, branch information, etc.) with other financial indicators derivable from the UIF's archive of aggregated money laundering reports (SARA), as well as external variables such as crime rates, regional income statistics, and data related to the underground economy. In addition to the Isolation Forest algorithm, other unsupervised algorithms were applied to cross-validate the main findings obtained, ensuring their reliability and effectiveness in identifying true instances of illicit financial activity. The developed methodology should enhance model's accuracy in detecting fraudulent activities and reduce false positives, making it a valuable operational tool to uncover and prevent money laundering.

Stochastic Gradient Langevin Dynamics With Non-Stationary Data

Attila Lovas

HUN-REN Alfréd Rényi Institute of Mathematics, Budapest, Hungary
attila.lovas@gmail.com

The Stochastic Gradient Langevin Dynamics (SGLD) algorithm and its variants have gained significant popularity in machine learning, particularly for training deep neural networks, due to their demonstrated effectiveness in finding global minima of complex, high-dimensional objective functions—assuming certain regularity conditions hold for the gradient. We investigate the SGLD algorithm with a fixed step size. While most existing studies on SGLD assume an i.i.d. data stream, this assumption is often unrealistic in practical applications, such as financial time series analysis, natural language processing, and sensor data processing. In such settings, the sequence of iterates no longer forms a Markov chain, significantly complicating the mathematical analysis. To address this challenge, we model the iterates as a Markov chain in a random environment (see Lovas & Rásonyi, 2021, 2023, and Rásonyi & Tikosi 2022). Under standard dissipativity and Lipschitz conditions, we establish the transfer of α-mixing properties from the data stream to the sequence of iterates (Lovas, 2024). This enables us to derive key theoretical results, including the law of large numbers, the central limit theorem, and concentration inequalities for SGLD in the non-convex setting. Our findings provide theoretical guarantees for SGLD in a more realistic scenario where the data merely weakly dependent.

A Practical Comparison of Variable Importance Techniques Across Modelling Frameworks

Markos A. Ktistakis, Andres Laverde Marin, Jaime Suarez, Leonidas Ntziachristos and Georgios Fontaras

Aristotle University of Thessaloniki, Thessaloniki, Greece
markos.ktistakis@ext.ec.europa.eu

Understanding the relative importance of input features is critical in applied modelling contexts and to inform policy. However, variable importance (VarImp) techniques can yield inconsistent results depending on model assumptions, data characteristics, and the notion of importance captured, such as explanatory power, predictive relevance or internal model contribution. This study presents a comparative evaluation of six widely used VarImp methods, using a large real-world dataset of energy and emissions from millions of vehicle-level records collected across Europe. Three modelling frameworks are evaluated, each paired with a representative importance technique: (i) linear regression: LMG (variance decomposition), random forests: permutation importance, and (iii) XGBoost: gain-based importance. Additionally, SHAP (SHapley Additive exPlanations) values are computed for all models to provide a unified, model-agnostic benchmark. The techniques are assessed across four dimensions: (i) stability of feature rankings under bootstrap resampling, (ii) agreement between methods, (iii) predictive performance based on cross-validation (R^2, MAE, MSE), and (iv) computational efficiency. The main analysis focuses on a continuous response variable (real-world CO_2 emissions) and continuous predictors related to vehicle and usage characteristics. Results show that while predictive accuracy is often comparable, importance rankings can diverge significantly, particularly across different model families. LMG and permutation importance yielded the most stable rankings, while SHAP scores were more sensitive to resampling, especially in XGBoost. To assess robustness and broader applicability, additional analyses explore predictors with varying collinearity, categorical inputs, and classification targets. These results offer practical guidance for researchers and analysts seeking interpretable and reliable importance estimates. No single technique emerges as optimal in all aspects, reinforcing the need to balance consistency, interpretability, and performance depending on modelling objectives.

Identifying Non-Lexical Entities in Croatian Consumer Health Forums via Machine Learning and Large Language Models

Amila Kugic

Medical University of Graz, Graz, Austria
amila.kugic@medunigraz.at

Non-lexical entities (NLEs) are commonly found within clinical narratives, such as short forms or jargon expressions. The identification of NLEs is important to ideally assign the type of NLE to the text passage in question, so that specialized methods for the expansion or disambiguation of NLEs can be applied. Especially in lower resource languages, such as Croatian, semantically mapping to international standards, e.g., HL7 FHIR or SNOMED CT, would be beneficial, although complicated in the realization by the various NLE types and low amounts of natural language processing resources. Four types of NLEs were highlighted in past investigations for the identification in consumer health forums, i.e., short forms, lexical variations, brand names and proper names. The dataset consisted of 12,023 sentences annotated in the beginning-inside-outside labeling format split into a 80 % training, 10 % validation, 10 % test set. Detection of NLE types with machine learning (ML) methods via two different language models (BERT, ELECTRA) for named entity recognition showed high performance results in the range of 0.89–0.91 F1-measure. The application of large language models (LLMs) in a zero-shot approach (gpt-3.5-turbo) did not achieve satisfactory results with an F1-measure of 0.45, while fine-tuning with even a third of the dataset improved the results to comparable levels as with ML methods, i.e., 0.90 F1-measure. Following up those investigations, advanced prompt guidelines with the same LLM model did not offer any performance benefits. The application of a more advanced LLM model (gpt-4-turbo) did improve the performance to 0.71 in F1-measure. The results showcase that gold standard annotated datasets in combination with fine-tuning LLMs offer better results in comparison to a zero-shot approach. However, if no gold standard dataset is available, the performance can be boosted by using more advanced LLM models.

Statistical Applications I

PIGNN-GPR: A Hybrid Machine Learning Framework for Spatio-Temporal PM2.5 Prediction

Reetha Thomas and Soudeep Deb

Indian Institute of Management Bangalore, Bangalore, India
reethathomas19@gmail.com

Accurate prediction of pollutant concentration is critical for both environmental sustainability and public health. While numerous studies have integrated the physics-based and data-driven approaches in deep learning for air quality forecasting, neural networks still face challenges with interpretability and explainability due to their hidden nature. Among the various air pollutants, PM2.5 stands out as one of the most hazardous, making its precise monitoring and forecasting a complex task. In this study, we present an innovative and efficient approach for forecasting hourly PM2.5 concentrations using a hybrid Physics-Informed Graph Neural Network–Gaussian Process Regression (PIGNN-GPR) model. The physics-informed component integrates fundamental principles from the reaction-diffusion-advection equation, ensuring adherence to physical constraints, while the Graph Neural Network (GNN) captures complex spatial dependencies by leveraging wind speed and wind direction across multiple locations. To improve forecast reliability, we incorporate Gaussian Process Regression to refine the PM2.5 predictions from the Physics-Informed Graph Neural Network (PIGNN), offering confidence intervals that quantify prediction uncertainty. Additionally, we apply the Inverse Distance Weighting (IDW) spatial interpolation method to estimate PM2.5 concentrations at unmonitored sites, further enhancing predictive accuracy. For better model interpretability, we use SHAP (SHapley Additive exPlanations) to assess the contribution of key input varivariables, latitude,ngitude, wind speed, and wind direction, to predictions. The model is evaluated using real-world air quality data from multiple locations in the Delhi region, India. This hybrid PIGNN-GPR framework marks a significant step forward in improving both the explainability and accuracy of air quality predictions, providing a powerful tool for environmental monitoring and decision-making.

Evaluating Statistical Methods for Estimating Behavioral Strategies From Indirect Reciprocity Experiments

Žiga Velkavrh, Aleksa Đorđević and Aljaž Ule

University of Primorska, Koper, Slovenia
ziga.velkavrh@upr.si

Various statistical estimation methods exist to uncover behavioral strategies that humans use in experimental games. In this study, we evaluate the performance of four methods from the experimental literature, using data from an indirect reciprocity experiment. Based on our experimental data alone only the relative performance of the estimation methods can be assessed, because true strategies are unknown. A more rigorous approach is to run agent-based simulations in which simulated players are assigned one of several different strategies uncovered in experiments, estimate their strategy from the simulated data, and compare the estimations to the true strategies. Using simulations, we show that the method based on finite mixture models outperforms the other three estimation methods. This result holds for different levels of noise in the data.

Complex Network Science for the Study of Agricultural Ecosystems

Michele Bellingeri

University of Parma, Parma, Italy
michele.bellingeri@unipr.it

The science of complex networks provides a powerful framework for analyzing the structure and dynamics of ecological systems, particularly food webs in agricultural landscapes. In this context, agricultural ecosystems are modeled as networks of biological species (nodes) connected by links representing their trophic interactions. Understanding how biodiversity loss propagates through these networks is crucial for sustaining ecosystem services. Traditional approaches based on binary topology often oversimplify extinction dynamics by ignoring species' energetic dependencies. In this study, we employ complex network analysis enriched with energetic thresholds to model secondary extinctions across hundreds of empirical food webs derived from agricultural ecosystems. Our findings reveal that food web robustness is highly sensitive to energy-based interactions with different crop management regimes, such as conventional and genetically modified herbicide-tolerant (GMHT) systems, showing distinct patterns of network fragility. We identify shifts in keystone species and highlight the stabilizing role of omnivorous interactions, particularly among invertebrate taxa. This work demonstrates the limitations of purely topological analyses and emphasizes the need to integrate energetic and functional traits into ecological network models. Ultimately, complex network science offers valuable tools for improving biodiversity conservation and designing more resilient agroecosystems in the face of environmental change.

L2-Penalization of the Fixed Effects in Linear Mixed-Effects Models Using the Existing Maximum Likelihood Software

Lan Gerdej and Rok Blagus

University of Ljubljana, Ljubljana, Slovenia
lg12410@student.uni-lj.si

We present a simulation study of L2-penalized fixed effect estimation in linear mixed models using a pseudo-observation approach. By augmenting the data with appropriately constructed pseudo-observations, the penalty term is incorporated directly into the likelihood, allowing implementation with standard mixed model software such as lme4 and glmmTMB in R. This approach enables regularized inference in high-dimensional settings, where standard estimation techniques fail due to identifiability and matrix singularity issues. The main focus of the study is predictive performance in clustered high-dimensional data, where fixed effects are sparse and random intercepts induce within-cluster correlation. We generate synthetic data under varying signal-to-noise ratios, correlation structures, and dimensionalities, and evaluate model accuracy on large, independently generated test sets. Predictive performance is evaluated using mean squared error. We also examine strategies for selecting the ridge regularization parameter, including leave-one-cluster-out cross-validation.

Limit of Detection in Biological Assays: A Comparison of Statistical Methods for Handling Missing Data

Zarja Fabjan, Nataša Kejžar and Stephen Nash

University of Ljubljana, Ljubljana, Slovenia
zarja.fabjan@gmail.com

This work is motivated by secondary hormone data from a randomized controlled trial, which investigated effects of six different doses of tamoxifen in women at high-risk for breast cancer. For two of the hormones, nearly half of the baseline and follow-up values are below the limit of detection (LOD). In the absence of LOD-related missingness, a log-transformed linear regression model would be used to asses dose-dependent changes in hormone levels. Methods recommended in two simulation studies are adapted to our specific scenario. Complete case analysis (CCA), single imputation using LOD/sqrt(2), Tobit regression, and an accelerated failure time (AFT) model are compared for handling LOD in dependent variables. For independent variables, CCA, single imputation, and the inclusion of a missingness indicator are assessed. Single value imputation introduces bias when used either for independent or dependent variables. CCA also leads to biased estimates when used for dependent variables. For independent variables, CCA leads to unbiased, but less efficient estimates. Including a missingness indicator improves statistical power without introducing bias. Tobit and AFT models are based on the same underlying theory and yield similar, unbiased results in this simulation setting. However, AFT provides more flexibility in defining which of the observations are censored, which can make a difference in real-world applications. Based on these simulation results, the combination of an AFT (or Tobit) model for handling LOD in the dependent variable and a missingness indicator for LOD in the independent variable offers a practical and effective approach for managing data with values below LOD in regression models.

Statistical Literacy

Data Visualization in Kindergarten and Early Primary School

Ana Zalokar and Lara Lusa

University of Primorska, Koper, Slovenia
ana.zalokar@famnit.upr.si

This presentation explores how young children can develop foundational data literacy through hands-on visualization activities, with a particular emphasis on the use of object graphs—physical representations of data using real-world items. Object graphs, also known as concrete or manipulative graphs, allow children to visually compare and interpret data without needing to count or calculate, making abstract statistical concepts tangible and accessible. Drawing from classroom-tested activities, we demonstrate how children can classify, sort, and summarize data using everyday materials such as coloured blocks, paper strips, and toy objects. These activities encourage children to ask investigative questions, make predictions, collect data, and interpret results, all through visual and tactile engagement. For example, stacking blocks by colour helps children identify the most common category, while arranging name strips by length reveals patterns in discrete numerical data. By embedding data visualization into early education, we lay the groundwork for lifelong statistical literacy and empower children to make sense of the world through evidence and inquiry.

Exploring Facts and Visuals: A Survey of Infographic Books for Children

Lara Lusa

University of Primorska, Koper, Slovenia
lara.lusa@famnit.upr.si

Infographic books for children combine data and visuals such as charts, diagrams, maps, illustrations, icons, and brief explanations to present information in an engaging and accessible way. These books simplify complex topics, making them easier for young readers to understand and enjoy. While infographics can cover any subject, children's books often use them to explore science and nature themes like animals and habitats, the human body, space and astronomy, and landforms. They also explain processes, timelines, and present interesting facts and records. Infographics also appear in various non-fiction titles to enhance comprehension. In this talk, I will survey the types of facts commonly covered in infographic books and explore how these books use different graphical displays, including pie charts, bar graphs, timelines, maps, and icon arrays to communicate information clearly. I will provide examples of well-designed visuals that aid understanding, alongside examples where poor design can confuse readers. I will also discuss how relatable examples and visual comparisons explain abstract concepts like time or speed by connecting them to everyday experiences. Additionally, I will highlight a popular concept in many books: presenting global statistics through simple analogies. For example, some books imagine the world population as a village of 100 people, breaking down language, religion, access to resources, and living conditions into proportional segments. This method transforms vast numbers into concrete, visual terms that children can easily grasp.

ISLP.SI 2025: The First Slovenian High School Data Visualization Competition

Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia
irena.ograjensek@ef.uni-lj.si

The goal of the International Statistical Literacy Project (ISLP), which celebrated its 30th anniversary in 2024, is to spread quantitative skills around the world, especially in developing countries and among the young. To this end, as one of its key activities, the ISLP has run several international poster competitions targeted at elementary school pupils, as well as high school and university (bachelor level) students since 2007. The posters submitted for the competition are supposed to reflect or illustrate usage analysis, interpretation and communication of statistics or statistical information at the knowledge level appropriate for each competition category. In this academic year, Statistical Society of Slovenia, with the invaluable support of the Statistical Office of the Republic of Slovenia and Petrol d.d., decided to run its first ever national poster competition labelled ISLP.SI 2025. Keeping it small in order to learn the ropes, the competition was only open to the high (secondary) school students also participating in the 8th Edition of the European Statistics Competition (ESC). In the framework of this presentation, we will take a closer look both at some of the winning submissions from the past ISLP annual international competitions and at the ISLP.SI 2025 poster competition submissions. We will also discuss challenges faced by the jury members when evaluating the ISLP.SI 2025 poster competition submissions.

Visualizing Data in the Business World: Some Worst and Best Practices

Bruno Božičnik, Manica Erjavec[1], Lea Medveščček[2], Lara Prijatelj[3], Darja Števančec[4] and Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia
[1]me22752@student.uni-lj.si, [2]lea.medve27@gmail.com, [2]lp94554@student.uni-lj.si, [4]ds11529@student.uni-lj.si

This presentation focuses on the perennial questions whether or not the business world is (i) capable and (ii) willing to communicate statistical data and information through graphical displays clearly, effectively, and without manipulation. In order to provide answers to these questions, we first present some of the most useful frameworks for quality evaluation of graphical displays. We then proceed to identify the most common types of graphical displays used in the business world, along with the contextual settings in which they usually appear, and related sets of common messages they are supposed to convey to various groups of stakeholders. Some of the worst and best practices pertaining to data visualisation in the business world are showcased next. We conclude the presentation by discussing how could formal educational programmes, informal (business and civic) initiatives, as well as the (business and general) press help to improve data visualisation practices both in the business world and beyond it.

Poster Presentations I

Bangkok Population Forecasting: A Comparative Study of Time Series and Machine Learning Models

Kamon Budsaba, Benjamas Tulyanitikul, Siraprapa Manomat, Wikanda Phaphan and Nattanicha Tipwong

Thammasat University, Pathumthani, Thailand
kamon@mathstat.sci.tu.ac.th

This study presents a comparative analysis of population forecasting models for Bangkok, utilizing both traditional time series methods and machine learning approaches, including hybrid models. Monthly population data from 2002 to 2022 were analyzed, and model performance was evaluated using Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The results indicate that the Long Short-Term Memory (LSTM) model outperformed all other models, including Autoregressive Integrated Moving Averag (ARIMA), Multilayer Perceptron (MLP), and Support Vector Regression (SVR), as well as hybrid models such as ARIMA-MLP, ARIMA-SVR, and ARIMA-LSTM. These findings suggest that LSTM is the most effective model for forecasting Bangkok's population, providing valuable insights for urban planning and resource allocation.

Improved Efficiency of Bayesian Estimation in Presence of Laplace Prior Distribution for Type-I Right-Censored Discrete Weibull Regression Model

Monthira Duangsaphon, Wararit Panichkitkosolkul and Benjamas Tulyanitikul

Thammasat University, Pathumthani, Thailand
monthira@mathstat.sci.tu.ac.th

Regression models have been shown to count response variables in experimental and observational research throughout a wide range of disciplines, including the social sciences, industry, economy, and public health. Both over-dispersion and under-dispersion count data can be fitted to the discrete Weibull distribution. A standard model may not be as effective if it focuses on over-dispersion data, where the variance is greater than the mean. The term censored data describes an observation or measurement whose value is only partially known. In some cases, the response variable may occasionally take on large values or outliers that alter its mean and variance, leading to over-dispersion that could impair a regression model's effectiveness. Therefore, this over-dispersion can be controlled by censoring the large values of this response. This study uses the Laplace prior distribution to offer an efficiency Bayesian estimate for the type-I right censored discrete Weibull regression model. The random walk Metropolis-Hastings algorithm was used to execute Bayes estimators, and the credible intervals were examined. A Monte Carlo simulation study was used to compare the Bayes estimators' performance with that of the maximum likelihood estimators, with an emphasis on mean square error. The coverage probability and average length were used to assess the intervals' performance. To illustrate how the proposed model and approach can be used in practice, an actual dataset was analyzed. The simulation and application results show that the Bayesian approach using a Laplace prior distribution outperforms other approaches.

Electricity Price Volatility in Slovenia During Crises and Energy Transition: An EGARCH-X Approach

Ramanpreet Kaur and Dušan Gabrijelčič

Jožef Stefan Institute, Ljubljana, Slovenia
ramanpreet.kaur@ijs.si

This study investigates the volatility dynamics of Slovenia's day-ahead electricity prices using an exponential GARCH (EGARCH) model with exogenous variables (EGARCH-X). The analysis spans multiple predefined event windows from 2019 to 2024, covering major structural disruptions such as COVID-19 pandemic, the Russia-Ukraine war, and a growing share of renewable energy (RE) in the electricity mix. The model incorporates key fossil fuel prices (natural gas, coal, and brent oil) and the percentage share of renewable electricity generation as external drivers of volatility. The results indicate persistent volatility clustering and asymmetric responses, with negative price shocks amplifying volatility more than the positive ones. Exogenous variables, including natural gas, coal, and the share of renewable energy have a positive and significant impact on electricity price volatility, especially during the post-war and renewable integration phases. These effects highlight how fuel price uncertainty and RE intermittency contribute to short-term market fluctuations. Overall, the findings reflect the increasing complexity and volatility sensitivity of electricity markets during periods of global disruption and structural transition.

Applying Large Language Models for Structuring Pathology Reports in Slovenian Cancer Registry

Maja Jurtela, Tina Žagar, Miran Mlakar, Nika Bric, Mojca Birk and Vesna Zadnik

Institute of Oncology, Ljubljana, Slovenia
maja.jurtela@gmail.com

As cancer diagnoses rise, manual extraction of structured variables from free-text medical records is becoming unsustainable. Slovenia is among the few European countries developing, in addition to population-based and hospital-based cancer registries, dedicated clinical registries for detailed data collection on common cancers. To support and assist trained medical personnel responsible for coding these data, we are exploring the partial integration of large language models (LLMs). Early testing on melanoma pathology reports suggests that prompt-based LLM approaches may outperform conventional machine learning models trained on labeled data. This insight emerged from a national student competition on developing new analytical methods in medicine (RIS 2025). Building on this, we are developing a locally hosted, semi-automated system that processes free-text pathology reports and outputs structured tables with proposed variable values. Each value will be accompanied by a confidence score derived from logit probabilities and linked to the corresponding text span in the report, allowing coders to efficiently verify or edit the value. We plan to evaluate LLM outputs against manually coded historical data to assess accuracy and iteratively improve prompts. This presentation will outline early findings as well as technical and methodological challenges. It will offer a practical example of how registries can begin integrating LLMs into data workflows, improving efficiency and consistency even with limited infrastructure or technical staff.

Ensuring Excellence: How We Assess the Accuracy and Quality of Seasonally Adjusted Data

Nikolina Rizanovska

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
nikolina.rizanovska@gov.si

Eurostat anticipates countries to validate time series data before submission. This validation process is essential to ensure the quality, consistency, and comparability of statistical information across EU member states. National Statistical Institutes (NSIs) perform this validation through a combination of manual quality assessments using JDemetra+, predefined rules, and recently, automated tools. At the Statistical Office of Slovenia, we have started using automated tools that quickly perform a comprehensive set of validation checks on multiple series simultaneously. We applied this automated validation to assess the accuracy and quality of seasonally adjusted (SA) data for 22 time series within the non-financial sector accounts. The validation follows a structured, multi-level approach. Level 1 Validation involves six core checks. First, it identifies seasonality in the non-adjusted (NA) series and evaluates the SA series accordingly. Second, it ensures that if the NA series is non-seasonal, the SA series matches it exactly, with any discrepancy suggesting an adjustment issue. Third, it tests for residual seasonality in the SA data. Fourth, it compares annual totals between the SA and NA series. Fifth, it identifies unexpected negative values. Finally, it detects signs of over-adjustment. Level 2 Validation provides a comprehensive five-page dashboard featuring a broad range of indicators and visualizations to facilitate a deeper understanding of the seasonal adjustment results. Level 3 Validation builds upon the previous levels by supplying even more detailed diagnostics for in-depth analysis. All validation steps are performed using a custom R-based tool, which employs the functions level1_validation(), level2_function(), and level3_function() from the Savalidation package, ensuring a systematic and rigorous assessment of seasonal adjustment quality across all monitored time series. These automated tools significantly speed up the validation process and generate warnings for potential issues that require further detailed investigation and resolution.

Neuroevolution of Adapters for Large Language Models

Katarina Perman

University of Ljubljana, Ljubljana, Slovenia
kp73570@student.uni-lj.si

This work touches upon the problem of where to put adapters in large language models to achieve the best performance possible. Large language models are nowadays a key tool for natural language processing. We use them for an array of tasks, but we must first fine-tune them for these tasks. Training the whole model, which is often very large, is computationally expensive and time consuming, so we use adapters instead, but we must place them in the right places in the large language model architecture and pick the right hyperparameters for them to be successful. We approach this problem by making use of genetic algorithms, as checking every possible configuration would take way too long. We tested this on the BERT model with the LoRA adapter, on emotion classficiation. That way we get not only concrete solutions but also insight into the patterns that appear in favorable solutions. We find that increasing the rank of the adapters is more effective than changing their placement, but a higher rank means more parameters, so care must be taken. Our results also show that the biggest sensible amount of parameters is up to 5 % the amount of parameters of the original model, as more parameters don't contribute to higher accuracy, and that the best possible accuracy for a certain adapter size has an upper bound.

Invited Lecture

On the Analysis of Data From Sequential Experiments With an Unspecified Number of Observations

Tamás Rudas

Eötvös Loránd University, Budapest, Hungary
trudas@elte.hu

In many sequential experiments whether or not a further observation is made depends on the outcome of a previous observation. For example, one may wish to estimate the efficacy of a vaccine based on data collected from an experiment in which those not reacting to the first vaccination receive a booster shot after some time, and to those who still do not show a reaction, a second booster shot is administered. In such a design, the total number of observations is not known in advance and this fact needs to be taken into account in the analysis of the observed data. There are several statistical problems with this structure, including test-retest problems, data about offsprings, and event history (survival) analysis. Such designs have a tree structure and the resulting data, in the discrete case, may be represented in an incomplete contingency table. The talk discusses the correct distributional assumptions, the maximum likelihood estimates, and their properties. The data generating process implies a multiplicative structure of the parameters which generalizes log-linear models. The sampling probabilities are as in multinomial sampling, but the total number of observations is random. This introduces an adjustment factor in the maximum likelihood estimates and also in their covariances. The results are illustrated on several data sets. This is joint work with Anna Klimova.

Social Sciences and Humanities

Measuring and Preventing Bullying in Slovenian Elementary Schools: Insights From the kNOwBULLYING Project

Vanja Erčulj and Aleš Bučar Ručman

University of Maribor, Ljubljana, Slovenia
vanja.erculj@um.si

Bullying is commonly defined as the intentional and often repeated use of power or popularity to harm, threaten, or humiliate another peer. It differs from conflicts between peers of equal power, such as mutual arguments or fights. Bullying can take various forms—including physical, verbal, social, and cyberbullying. These forms may overlap or be used interchangeably. Further, pupils exposed to bullying can on the other side also bully other peers. Existing research indicates that bullying is most prevalent during the final years of elementary school. In Slovenia, comprehensive national data on this issue has been lacking; prior studies have typically relied on small, non-representative samples, with the exception of a nationwide study focusing solely on cyberbullying. The primary aim of our research was to measure the prevalence and characteristics of bullying in Slovenian elementary schools and to develop an evidence-based prevention program. These goals are pursued within the framework of the international kNOwBULLYING project. The process of developing the survey instrument involved multiple stages, including a review of existing tools and the design of a final questionnaire for use in the nationwide study. Particular attention was given to resolving methodological challenges, ensuring a balance between scientific rigor and practical feasibility.

The Role of Basic Psychological Needs and Motivation in PhD Mentoring Outcomes

Marjan Cugmas, Sara Atanasova and Luka Kronegger

University of Ljubljana, Ljubljana, Slovenia
marjan.cugmas@fdv.uni-lj.si

A mentoring relationship between a PhD student and their mentor is a complex process that is deeply embedded in a broader organisational and academic context. Yet much of mentoring takes place at the individual level, within the dyadic relationship between mentor and mentee. This presentation examines the interplay between autonomy-related psychological characteristics, relatedness, and perceived competence, and their associations with key PhD study outcomes. To this end, the survey with a sample of 241 PhD students who completed their doctoral studies in various scientific fields between 2001 and 2020 in Slovenia was conducted was conducted in June and July 2024. Structural equation modelling confirmed that psychological traits—specifically, resilience and self-initiative—positively influence perceived competence during PhD study. This perceived competence, in turn, fosters greater academic and research group inclusion, which subsequently enhances satisfaction with the mentoring relationship, PhD programme, and work. Perceived competence and intrinsic motivation also positively affect attitude toward publishing. Notably, the presence of a co-mentor is associated with greater satisfaction in the mentoring relationship. Overall, the results emphasise the complex interplay between psychological traits and motivation in doctoral supervision and illustrate their influence on both subjective satisfaction and tangible results such as competence development and academic commitment.

Insights From the Younger Generation Towards EU's Democracy

Andreea-Monica Munteanu and Andreea-Mihaela Niculae

Bucharest University of Economic Studies, Bucharest, Romania
munteanuandreea18@stud.ase.ro

In today's context of political turmoil across many European countries, the perceptions and attitudes of young people aged 16 to 30 towards democracy offer a glimpse into the future of democratic stability and legitimacy. This research investigates the associations between the EU's impact on younger people's lives, their satisfaction with how democracy functions in the EU, and several demographic factors. We identify key characteristics of the younger population towards their satisfaction with the EU democratic system and the EU's impact on their lives by age group, residency country, and type of community using data from Flash Eurobarometer 556 and the Correspondence Analysis technique. The analysis shows that young people who perceive the EU as positively impacting society tend to also view democracy as highly beneficial for their well-being. Additionally, younger age groups are generally more satisfied with democracy, while older age groups tend to express some dissatisfaction. Urban youth living in big cities are more optimistic about democratic processes, whereas rural populations are more skeptical. Regionally, the younger population living in developed and Western European countries exhibits higher dissatisfaction with the EU's democratic system, while those in Eastern European and developing countries tend to be more satisfied with how democracy functions within the EU. These findings suggest that perceptions of democracy among young Europeans are diverse and mainly influenced by demographic and regional factors.

On the Determinants of Financial Literacy: Evidence From Italian Households

Gaetano Carmeci, Tommaso Cortivo, Alberto Dreassi and Giovanni Millo

University of Trieste, Trieste, Italy
tommaso.cortivo@phd.units.it

This study investigates the determinants of financial literacy among Italian households using data from the SHIW survey. We estimate the relationship between financial knowledge and a set of demographic, socioeconomic, behavioral, and contextual variables. Results from an OLS regression suggest that education, income, age, gender, and risk tolerance are strongly associated with financial literacy. Regional disparities and intra-household dynamics also play a role. Our findings confirm most theoretical expectations, highlight persistent gender and territorial gaps, and point to the relevance of both cognitive traits and economic autonomy. The study contributes to the debate on effective financial education and inclusion policies.

Design and Empirical Verification of a New Methodology for Managing a Retailer's Active Product Assortment

Domen Kozjek and Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia
domen.kozjek@windowslive.com

The active product assortment of a retailer is determined by the set of stock keeping units (SKUs) available at any time at different retailer's outlets. The goal of designing and managing an active product assortment is to try and simultaneously meet expectations of customers (focused on price) on the one, and expectations of the retailer (focused on sales efficiency) on the other hand. In attempting to reach the optimal balance, retailers face many limitations. Most often, the constraints are related to available financial resources, such as the amount of funds a retailer can invest in the SKU stock, employee training, advertising, promotion, and other relevant activities. Therefore, compromises are almost often a necessity, with the optimal point perpetually elusive: once achieved, it may very quickly change again due to factors such as the seasonal nature of SKUs, innovations in the field of packaging, new scientific findings on ingredients, changes in consumer tastes, and similar. In this paper, we propose a new methodology for managing a retailer's active product assortment. The methodology differs from the existing ones (such as Data Envelopment Analysis and Stochastic Frontier Analysis) not only because of a different decision-making unit, but also because of the assumption, that deviations from output limit values are not just the consequence of the sales process inefficiency, but can also be the result of other factors. Our three-step approach, which we test on real-life data, makes it possible both to differentiate the effects of managerial inefficiency from statistical noise and other influential factors, and to remove managerial inefficiency from the model in a controlled manner. This greatly increases sales efficiency's measurement reliability in comparison with the existing approaches.

House Prices and Income: An International Perspective

Giovanni Millo

University of Trieste, Trieste, Italy
giovanni.millo@deams.units.it

This paper is a robustness test (scientific replication) of the US housing market analysis by Holly et al. (2010), drawing on an international sample of developed countries; it is also an extension, updating the methods according to the more recent literature. The importance of controlling for common factors is confirmed, but the effect is reversed: the income elasticity of house prices is here significantly decreased. Spatial effects, significant in Holly et al., are less relevant in this context. The role of population growth and net financing cost is highlighted beyond what originally found by Holly et al. This last result bears special relevance in the current times of rising interest rates after a decade-long stagnation.

Biostatistics II

Net Survival of Colorectal Cancer by Stage in Chile: Addressing a Critical Evidence Gap Using Hospital-Based Cancer Registries

Felipe Andrés Medina Marín, Andrea Canals, Natalia Cuadros, Nicolás Silva and Tania Alfaro

University of Chile, Santiago, Chile
f.medina@uchile.cl

Chile has shown some of the lowest colorectal cancer net survival estimates among high Human Development Index countries in the CONCORD studies. This alarming disparity could be attributed to a later stage at diagnosis, a lower net survival within stages, or both. However, neither hypothesis has been thoroughly evaluated due to lack of population-level data. In this study, we leveraged hospital-based cancer registry data, including stage of diagnosis and follow-up data among others, from 16 catchment areas across Chile encompassing 11,597 colorectal cancer cases diagnosed between 2011 and 2022. Individual records of this data were linked with vital statistics and catchment area specific mortality rate, stratified by age, sex, and calendar year were used to estimate population hazard. We estimated net survival using both non-parametric Pohar-Perme estimator via the R package relsurv and flexible parametric excess hazard models via the R package mexhaz. To handle missing values in key variables, such as stage of diagnosis or treatment intention, we conducted multiple imputation using chained equations using R package mice. Our findings indicate that 70 % of patients were diagnosed at advanced stages (III-IV). Five-year net survival was 55.1 % overall, with substantial variation by stage: 89.3 % for stage I, 73.1 % for stage II, 51.5 % for stage III, and 24.3 % for stage IV. Flexible excess hazard models confirmed significant effects of stage, age at diagnosis (modeled with natural splines), and treatment intention on net survival. These results offer the first robust estimates of stage-specific net survival for colorectal cancer in Chile and reveal the importance for improving early detection. The approach also demonstrates how hospital registries, often overlooked in international comparisons, can contribute valuable insights when combined with appropriate statistical methods.

Comparison of Measurement Variability Between Groups With Repeated Observations

Nataša Kejžar

University of Ljubljana, Ljubljana, Slovenia
natasa.kejzar@mf.uni-lj.si

In physiotherapy, many measurement tools are not fully automated, and certain steps within their protocols depend on the physiotherapist conducting the assessment—for example, adjusting the device to match the size and musculature of the examined limb. As a result, evaluating both intra- and inter-rater reliability is essential to ensure that the measurement tool is suitable for clinical or research use. To this end, Bland–Altman plots are typically examined, and the minimal detectable change (MDC) is calculated based on repeated measurements from the same individuals. When a measurement tool is applied across different subpopulations (e.g., healthy versus clinical populations), it is important to assess whether reliability remains consistent. In such cases, the differences between repeated measurements are calculated within each group. If there is no evidence of systematic bias, the next step involves comparing the variability of these differences between groups. Tests for equality of variance, such as Levene's test, may be employed for this purpose. However, when data are collected from both limbs of the same individual, independence assumptions underlying these tests may be violated. To mitigate this, one limb is often selected at random per subject, although this approach is suboptimal. One solution is to repeat the random selection process multiple times, performing the statistical test in each iteration to generate a distribution of p-values. Alternatively, a more sophisticated statistical framework—such as Generalized Additive Models for Location, Scale and Shape (GAMLSS)—can be considered. GAMLSS allows modeling not only the mean (location) but also the variance (scale), skewness, and kurtosis of the outcome distribution, and can be extended to incorporate random effects. A simulation study comparing these approaches will be presented, and the statistical properties, including type I error rates and power, will be discussed.

Analyzing Mortality in Dutch Breast Cancer Patients Using an Extended Multi-State Model

Damjan Manevski

University of Ljubljana, Ljubljana, Slovenia
damjan.manevski@mf.uni-lj.si

The Dutch population-based cohort of breast cancer patients provides rich information on the patients' survival through time, together with any possible adverse events such as locoregional or distant recurrence. When focusing on older patients, one is not interested solely in the overall mortality, but also in the disease-specific mortality which provides further insight into the severity of the disease. Note that the two mortalities differ in this case since many patients die due to other (population) causes in older age groups. The goal of this study is to consider both the adverse events and the causes of death through survival analysis. Multi-state models have been developed as a statistical approach for incorporating additional events (apart from death) in the survival analysis. We apply this framework on the Dutch breast cancer data set. Furthermore, since cause of death is not given in the data, an extended multi-state model can be used which considers cause of death. This approach is based on relative survival, a subfield of survival analysis, which incorporates external mortality tables to distinguish between disease-specific and other-cause mortality. Previous work has applied a non-parametric extended multi-state model to these data (de Boer et al., 2022). Building on this, our study incorporates regression modeling using both the multiplicative Cox model and the additive Aalen model, which have recently been extended to the multi-state setting. These models allow us to assess covariate effects and gain deeper insight into the progression and mortality dynamics in older breast cancer patients.

The Importance of Using Appropriate Methodology in Interval-Censored Illness-Death Models

Gaber Kokovnik and Maja Pohar Perme

University of Ljubljana, Ljubljana, Slovenia
gk25294@student.uni-lj.si

In many studies, disease onset times are not known exactly. For example, when a patient's disease status can only be assessed during scheduled follow-up visits. In such cases, the onset time is interval-censored, meaning it is only known to have occurred within a specific time interval. This poses a critical challenge in survival analysis. The irreversible illness-death model describes a subject's progression from an initial state to an absorbing state, either directly or through an intermediate state. It is commonly applied in medical research, where the intermediate state represents illness and the absorbing state represents death. Despite the prevalence of interval-censored data, it is often ignored in practice. Instead, naive imputation strategies, such as midpoint or right-end imputations, are frequently used as substitutes for appropriate statistical modeling. These approaches can introduce substantial bias into the estimation of transition intensities and covariate effects. This study emphasizes the importance of properly handling interval-censored data, combining theoretical motivation with empirical evidence. We utilize the `SmoothHazard` R package, which supports semi-parametric estimation of illness-death models using M-spline-based smoothing for baseline transition intensities. Through simulations, we try to demonstrate that ignoring interval censoring can lead to considerable bias. We also apply these methods to a real-world dataset with interval-censored disease times to illustrate the practical consequences of appropriate modeling.

The Aalen-Johansen Estimator as an Alternative to Kaplan-Meier in the Presence of Time-Dependent Covariates

Ema Požek, Damjan Manevski and Maja Pohar Perme

University of Ljubljana, Ljubljana, Slovenia
ema.pozek@mf.uni-lj.si

In survival analysis, time-dependent covariates are frequently incorporated via Cox regression. However, the standard Kaplan-Meier survival estimator, as the most basic presentation of the data, does not make use of the longitudinal information available. As an alternative, we study the idea of a multi-state modeling approach in which states represent possible values of the time-dependent categorical variable. Within this framework, the Aalen-Johansen estimator serves as a natural extension of the Kaplan-Meier estimator for multi-state models. Under certain assumptions, Aalen-Johansen estimator provides identical estimates as Kaplan-Meier, while additionally taking into account changes of the time-dependent covariate as transitions in the model. Consequently, it can substantially reduce the variance of the survival estimates when compared to Kaplan-Meier estimator. On the other hand, its equivalence to the Kaplan-Meier estimator relies critically on the validity of the Markov assumption of the underlying multi-state model. Through theoretical considerations and simulation studies, we investigate the mechanisms of variance reduction in the Markov model setting and assess the robustness of the approach when the Markov assumption is violated, particularly in scenarios where a truly continuous variable is artificially discretized.

Using Data Augmentation to Overcome Separation and Singular Random Effects Covariance Matrices in Logistic Mixed-Effects Models

Rok Blagus, Georg Heinze and Tina Košuta

University of Ljubljana, Ljubljana, Slovenia
rok.blagus@mf.uni-lj.si

The logistic mixed-effects model is commonly used to relate binary outcomes to covariates while accounting for dependencies among observations. When using the maximum likelihood (ML) method to estimate the model's parameters, the estimates are commonly on the boundary of the parameter space, causing numerical instability and adversely affecting statistical inference. We investigates penalized ML estimation to avoid boundary estimates, applying Jeffreys' prior to the fixed-effects coefficients and an inverse Wishart prior to the random effects covariance matrix. While these priors have been used separately, this is the first time they are combined in this setting. We show how both priors can be implemented through data augmentation, simplifying penalized estimation. We also introduce a computationally efficient single-step procedure for approximating Jeffreys' prior. We show the superiority of the proposed approach over the ML estimator and some other solutions that were proposed in the literature by performing a large Monte-Carlo simulation. The method is also illustrated on a real dataset.

Network Analysis

Bibliometric Analysis of a Scientific Journal Based on OpenAlex Data

Vladimir Batagelj

Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia • University of Primorska, Koper, Slovenia • University of Ljubljana, Ljubljana, Slovenia
vladimir.batagelj@fmf.uni-lj.si

We are developing an R package OpenAlex2Pajek (https://github.com/bavla/OpenAlex) for creating bibliographic networks from the OpenAlex database. The basic package version supports the collection of data on selected topics. In this contribution, we present an extension, the function OpenAlexSources, that creates networks related to a selected journal (all papers published by the journal chosen and all works citing/cited by these papers). Since in networks the units (works, authors, sources, keywords, etc.) are identified by their OpenAlex IDs, another function, unitsInfo, provides the user with additional information about the units appearing in the results of analyses. We applied the new functions to create bibliographic networks for the journals Metodološki zvezki—Advances in Methodology and Statistics (https://openalex.org/S4210169332) and Ars Mathematica Contemporanea (https://openalex.org/S61442588). The analyses' results will be presented at the conference.

A Dynamic Stochastic Blockmodeling Approach: Simulations vs. Theory and Some Lessons About Optimization

Aleš Žiberna and Damjan Škulj

University of Ljubljana, Ljubljana, Slovenia
ales.ziberna@fdv.uni-lj.si

A new stochastic blockmodeling approach for dynamic »snapshot« networks is presented. The aim of the approach is to be especially useful for cases where the blockmodel changes in time, as previous work has shown that in such cases existing method for blockmodeling of dynamic snapshot networks are often surpassed by more general blockmodeling approaches (e.g. those for blockmodeling linked/multipartite networks). The approach was evaluated and compared with competing approaches in a variety of settings in simulated dynamic networks with known partitions. The proposed approach was the best approach overall. Only for stable blockmodels did more restrictive approaches perform noticeably better. In addition, it was applied to a real dynamic co-authorship network. This presentation will mainly focus on some mismatch of results based on simulations with theory and some lessons about optimization.

Unveiling Complex Connectivity Patterns in Biomedical Systems Through Network Science

Marko Gosak

University of Maribor, Maribor, Slovenia
marko.gosak@um.si

Over the past 25 years, we have witnessed the coming of age of network science as a central paradigm shaping some of the most remarkable scientific advances of the 21st century. In parallel with the ongoing data deluge, network science has not only offered new metaphors for understanding complexity but has also provided robust algorithms and statistical tools that have transformed entire fields across natural, social, and biomedical sciences. One particularly powerful concept is functional connectivity, which enables the quantification of coordinated activity patterns between components of a system, even when the underlying physical connections remain unknown. This approach is highly compatible with the statistical characterization of high-dimensional systems, offering an elegant bridge between modern data analytics and the modeling of collective behavior. Yet, many biological systems are too intricate to be fully captured by classical network models. The presence of multiple types of interactions, temporally evolving relationships, and interdependencies between different subsystems calls for more sophisticated representations. In this context, the multilayer network formalism has emerged as a powerful framework to assess such multi-dimensional nature systems. In this contribution, I will first provide an accessible overview of the core concepts in network theory, including commonly used network metrics and analysis techniques. This will set the stage for two specific applications of network science in biomedicine. First, I will introduce the extraction and analysis of functional beta-cell networks from calcium imaging data in pancreatic islets, highlighting metrics that describe intercellular coordination and their implications for metabolic health and diabetes. Second, I will explore the domain of network neuroscience, with emphasis on EEG-derived brain networks, discussing their utility in seizure prediction and clinical decision support for epilepsy. In both cases, I will address the challenges and promises of multilayer approaches for capturing collective dynamics and temporal evolution of functional patterns.

Blockmodeling of the International Trade Network: Looking for a "Trump Effect"

Fabio Asthar Telarico and Aleš Žiberna

University of Ljubljana, Ljubljana, Slovenia
fabio-ashtar.telarico@fdv.uni-lj.si

The first Trump administration's (2017–2021) trade-policy agenda was defined by tariff battles, rewritten agreements and pronounced economic nationalism, a dramatic turn that reinvigorated arguments about globalisation, protectionism and trade. This paper asks whether, and to what extent, those initiatives reshaped the International Trade Network (ITN). Adopting blockmodelling, a network-clustering method long associated with world-systems research but now seldom used, we outline practical steps for applying it to weighted and directed trade networks. The findings analyse the extent to which blockmodelling can be effective at revealing patterns of trade in terms of partner selection and shifts in the ITN's topology. Several existing blockmodeling approaches are benchmarked, gauging how well each captures patterns in partner selection and trade volumes amid external shocks such as tariffs and sanctions. Acknowledging the ITN's scale-free nature and uneven trading capacity, we confront key methodological hurdles including data standardisation and the integration of external covariates (e.g., tariff schedules, exchange-rate movements). By marrying social-network analysis with international-trade theory, the study not only offers an empirical assessment of a possible »Trump effect« on global commerce but also repositions blockmodelling as a valuable instrument for international-trade analysis using network methods.

Poster Presentations II

Some Modifications of INAR(1) Models for Under-Dispersed and Over-Dispersed Time Series of Counts

Predrag Popović, Zohreh Mohammadi and Hassan Bakouch

University of Niš, Niš, Serbia
popovicpredrag@yahoo.com

As time series of counts appear in many scientific fields, there is a growing need for a better understanding of their evolution and statistical properties. One approach to modeling these series involves the use of integer-valued autoregressive (INAR) models. This research focuses on adapting first-order INAR models to make them suitable for modeling both under-dispersed and over-dispersed integer-valued time series. Since INAR models consist of a survival component and an innovation component, we concentrate on adjusting the distribution of the innovation component to better reflect the characteristics of the observed series. For this purpose, we employ the weighted negative binomial Lindley distribution. By appropriately selecting the parameters of this distribution, we can capture a wide range of data series. Furthermore, the survival component is defined using various thinning operators. The effectiveness of these modifications is demonstrated through the modeling of real-world time series data.

Variable Selection via Fused Sparse-Group Lasso Penalized Multi-State Models Incorporating Clinical and Molecular Data

Kaya Miah, Jelle J. Goeman, Hein Putter, Axel Benner and Annette Kopp-Schneider

German Cancer Research Center, Heidelberg, Germany
k.miah@dkfz.de

In the era of precision medicine with increasing molecular information, the use of a multi-state model is required to capture the individual disease pathway along with underlying etiologies with greater precision. Especially the availability of big data with numerous covariates induces several statistical challenges for model building. For multi-state models based on high-dimensional data, effective modeling strategies are required to determine an optimal, ideally parsimonious model. Standard methods integrate regularization into the fitting procedure to conduct variable selection. In the multi-state framework, linking covariate effects across transitions is needed to conduct joint variable selection. A useful technique to reduce model complexity is to address homogeneous covariate effects for distinct transitions. We integrate this approach to data-driven variable selection by extended regularization methods within multi-state model building. We propose the fused sparse-group lasso (FSGL) penalized Cox-type regression in the framework of multi-state models combining the penalization concepts of pairwise differences of covariate effects along with transition grouping. For optimization, we adapt the alternating direction method of multipliers (ADMM) algorithm to transition-specific hazards regression in the multi-state setting. In a simulation study and application to acute myeloid leukemia (AML) data, we evaluate the algorithm's ability to select a sparse model incorporating relevant transition-specific effects and similar cross-transition effects of biomarkers. We investigate settings in which the combined penalty is beneficial compared to global lasso regularization. Thus, effective model selection strategies in multi-state survival analysis are required for enhancing comprehension and interpretation of individual disease pathways, distinct oncological entities and tailored precision therapies, leading to improved personalized prognoses.

An Interactive R Package for Bayesian Imputation of Censored Survival Data

Jamie Wilson, Shirin Moghaddam and Norma Bargary

University of Limerick, Limerick, Ireland
jamie.wilson@ul.ie

The presence of censored observations in survival data leads to unique challenges for patient communication. While traditional methods such as Kaplan-Meier estimation handle censoring appropriately, the resulting visualisations and median survival times are often difficult for patients to understand and can be misinterpreted. Furthermore, these methods do not provide a robust framework for quantifying individual-level variability and uncertainty, limiting their utility. Moghaddam et al. (2022) proposed treating censored observations as a form of missing data and developed a Bayesian imputation framework to produce completed datasets by sampling from the posterior predictive distribution. This approach enables the use of standard descriptive statistics and familiar graphical displays including histograms, boxplots, and density plots, and complements traditional survival analysis visualisations. Building on this work, we are developing an interactive R package that implements this methodology using Stan-based Hamiltonian Monte Carlo. Our package is capable of handling both parametric and non-parametric Bayesian approaches for modelling specified cohorts. The parametric framework uses typical survival distributions such as Weibull, exponential and log-normal, while the non-parametric approach employs Dirichlet Process mixture models to avoid restrictive distributional assumptions and to help mitigate model mis-specification. Our package offers an intuitive and tiered workflow based on user experience, with a single command capable of running the end-to-end procedure using sensible defaults and data-adaptive priors. More advanced features allow the selection of specific distributions, custom priors and MCMC settings. Ultimately, this package will equip practitioners with both a principled statistical foundation and a flexible and practical software tool for extracting imputed datasets for further analysis.

Mapping the Mind: Network Learning for Neurological Disorders

Bisera Nikoloska and Andrej Kastrin

University of Ljubljana, Ljubljana, Slovenia
bisera.nikoloska@hotmail.co.uk

Functional MRI (fMRI) provides rich insights into brain activity but generates complex, high-dimensional data that is difficult to interpret. To address this issue, brain networks can be built from fMRI scans and analyzed with network representation learning, transforming complex networks into low-dimensional embeddings. These embeddings are then used as features for machine learning classifiers, including random forests, logistic regression, and support vector machines. Among the approaches tested, node2vec combined with random forest achieved the highest accuracy at 90 %, enabling effective differentiation between patients with neurological disorders and typically developing individuals.

A Biopsychosocial Model for Stratifying Women and Predicting Outcomes in Women With Gestational Diabetes

Ana Munda, Draženka Pongrac Barlovič and Andrej Kastrin

Ljubljana University Medical Centre, Ljubljana, Slovenia • University of Ljubljana, Ljubljana, Slovenia
ana.munda251@gmail.com

The rising incidence of gestational diabetes (GDM) burdens women and healthcare systems. This study aimed to develop a biopsychosocial model to stratify women after GDM diagnosis by target glucose levels and pregnancy-related outcomes. Outcomes were (i) glucose levels within the target range below 80 % during pregnancy, and (ii) at least one GDM-related outcome (large-for-gestational age, neonatal hypoglycemia, jaundice, clavicle fracture, stillbirth, neonatal death). Predictors included biological, social, and psychological factors. Models were built using logistic regression, random forest, support vector machine, and XGBoost. 470 and 477 participants (median age 31 [28–35 years]; BMI 24.7 [21.6–28.7]) were included in the models focused on target glucose levels and GDM pregnancy-related outcomes, respectively. The XGBoost model demonstrated the highest predictive accuracy for the target glucose levels model (Precision–Recall AUC = 0.87, concordance $c = 0.93$, Brier score = 0.11) and the pregnancy outcomes model (PR AUC = 0.81, concordance $c = 0.91$, Brier score = 0.14). XGBoost performed best (glucose: PR AUC = 0.87, concordance $c = 0.93$; outcomes: PR AUC = 0.81, concordance $c = 0.91$). Top predictors were BMI, gestational age at diagnosis, and empowerment (glucose model), and empowerment, gestational age, age, impact of GDM, and self-efficacy (pregnancy—related outcome model). Biopsychosocial predictors enable early stratification of women with GDM, representing a step toward personalized treatment and tailored interventions.

Permutation t-Test: A Simulation Study for Comparing Two Means Under Non-Ideal Conditions

Alja Nike Kastrin[1], Amer Mujagić[2] and Maja Pohar Perme

University of Ljubljana, Ljubljana, Slovenia
[1]ak06297@student.uni-lj.si, [2]am34483@student.uni-lj.si

Statistical hypothesis testing often relies on assumptions that are not always satisfied in practice. A common task is to compare the means of two independent groups. The Student's t-test is widely used but assumes normality and equal variances. When these assumptions—particularly homogeneity of variances—are violated, results can be misleading. The Welch t-test addresses variance inequality but still relies on distributional assumptions. Permutation tests provide a flexible, non-parametric alternative, relying only on the assumption of exchangeability under the null hypothesis. We systematically compare the performance of the Student's t-test, the Welch test, and the permutation test using the difference in means as the test statistic. Through simulations, we evaluate test size and power under various conditions, including different distributions (normal, exponential, uniform), equal and unequal variances, and balanced versus unbalanced sample sizes. Our results confirm that the Welch test controls Type I error more effectively than the Student's t-test when variances are unequal. The permutation test, while free of parametric assumptions, is sensitive to asymmetry and variance heterogeneity when the sample mean is used as the test statistic. It performs well in symmetric settings or when sufficiently large samples are employed. Based on our findings, we offer the following practical guidelines: when comparing small samples from asymmetric distributions with approximately equal variances, the permutation test is most appropriate. In other scenarios where assumptions are reasonably met, the Welch test is generally preferred, though achieving adequate power may require larger effect sizes when variances differ substantially.

Statistical Applications II

Machine Learning Methods Applied in the Real Estate Market

Ion-Florin Răducu

University of Economic Studies, Bucharest, Romania
raducuion18@stud.ase.ro

The aim of this paper is to develop machine learning models that use Lasso regression and XGBoost (Extreme Gradient Boosting) to estimate prices in the US real estate market. Lasso regression and XGBoost are more advanced machine learning models that can be successfully applied to a wide range of data sets in various fields. Lasso regression, also known as L1-penalized regression, is a linear regression method that penalizes the sum of absolute coefficient values. XGBoost, on the other hand, is a decision tree-based machine learning algorithm that employs boosting techniques to build robust models by combining multiple weak trees. The research findings provide an overview of estimated prices in the US real estate market, which were captured using advanced models with relatively high accuracy and performance. Estimating real estate market prices using machine learning methods is critical for optimizing property valuation and prediction processes, as well as decision-making behavior. Given the impressive evolution of artificial intelligence and machine learning, these techniques enable the analysis of large amounts of data, such as property characteristics, market trends, economic conditions, and demographic data, resulting in more accurate estimates tailored to current socioeconomic realities. Machine learning can detect hidden patterns and correlations that traditional analysis might miss, thereby improving decision-making for both investors and buyers. Estimating prices in the real estate market has numerous practical implications. These estimates can help investors and developers make informed purchasing and selling decisions, identifying profitable opportunities and reducing financial risks. In terms of public policy, accurate price estimates can have an impact on urban development, making it easier to build sustainable communities. They can also assist financial institutions in determining the risks associated with mortgage loans, thereby contributing to financial market stability.

Profiling of Adolescents With Symbolic Data Analysis of Self-Assessments and App Usage

Simona Korenjak-Černe, Jasminka Dobša, Miranda Novak and Maja Buhin Pandur

University of Ljubljana, Ljubljana, Slovenia
simona.cerne@ef.uni-lj.si

Positive youth development and the quality of friendships are important factors in promoting the mental health of adolescents. Nowadays, digital media strongly influence adolescents' self-image. Exploratory data analysis can help us to better understand the complexity of relationships. The aim of our broader research is therefore to develop a methodology for exploratory analysis of symbolic data to assess the positive youth development framework using traditional and digital mobile assessments. The present study focuses on the identification of profiles of adolescents based on ecological momentary assessment and passive data using clustering methods of symbolic data analysis. The dataset under consideration consists of reports submitted by one hundred and thirty Croatian high school students, each of whom evaluated the quality of their close friendships and their affect seven times a day for one week (i.e. 49 ratings). The present study was conducted using the Effortless Assessment Research System (EARS) from Ksana Health, University of Oregon. This system enables the collection of passive data on the use of the mobile applets in addition to the collection of ecological momentary assessment data (i.e., reported self-assessments). During the study 927 applications used by adolescents were identified. For the purpose of data analysis, the applications were semi-automatically categorised into 16 groups by using generative artificial intelligence (ChatGPT, Google Bard). As each student answered the ecological momentary assessment questions multiple times, this data can be viewed as symbolic showing the variability in each student's ratings throughout the day and over the entire observation period. This variability (which would be lost in the conventional representation with the mean value) can then be incorporated into the clustering process. Based on the results obtained, the advantages and disadvantages of the symbolic clustering approach, which takes intrinsic variability into account, are examined.

An Alternative to Classical Intention-to-Treat Analysis for Comparing a Time-to-Event Endpoint in Precision Oncology Trials

Marilena Müller

German Cancer Research Center, Heidelberg, Germany
marilena.mueller@dkfz-heidelberg.de

We consider a two-arm randomized clinical trial in precision oncology with time-to-event endpoint. The control arm consists of standard of care (SOC) whereas patients in the treatment arm are offered personalized treatment, when available. Patients in the personalized treatment arm, for which no personalized treatment is available or who do not consent to treatment also receive SOC. Intention-to-treat analysis hence involves comparing the outcomes of a group of patients receiving either personalized treatment or SOC to patients receiving exclusively SOC. This does not lead to an unbiased estimator of the treatment effect for those eligible for personalized treatment in the classical intention-to-treat approach. We investigate the performance of intention-to-treat and per-protocol analyses and develop more appropriate alternative analysis schemes. For this purpose, the patients are divided into groups based on whether they receive their intended treatment or not. An extension of the Cox proportional hazards model is proposed for estimating the conditional intensities in each group simultaneously via Maximum Likelihood estimation on the partial likelihoods. Counting process theory as well as martingale theory is used to develop suitable test statistics for various settings of interest. Both groups can be evaluated distinctly, thus enabling comparison between groups. This includes the investigation of a possible selection effect via the groups' respective regression coefficients. An in-depth simulation study and a real data example complement the theoretical results. A novel more rigorous model for the analysis of the treatment effect in the presence of mixtures or asymmetric trials is proposed. Guidelines are provided to identify scenarios where this model is necessary or appropriate, and when a classical intention-to-treat analysis remains preferable.

Contrast Testing in Julia: Implementing GLHT Functionality

Jakob Peterlin

University of Ljubljana, Ljubljana, Slovenia
jakob.peterlin92@gmail.com

I aim to introduce a small Julia package I've developed for generalized linear hypothesis testing. Its core purpose is to let users specify and evaluate arbitrary contrasts derived from a single statistical model. Unlike approaches that treat contrasts as independent, this package respects whatever dependence structure the underlying model imposes. The key challenge—one that this implementation tackles—is conducting tests that fully exploit that dependence, using simulation to derive accurate p-values and confidence intervals. In essence, my goal was to recreate in Julia the functionality of R's glht from the multcomp package. My version works seamlessly with generalized linear models, generalized linear mixed-effects models, and several other model types. In my talk, I'll walk through the package's interface, demonstrate its use, and offer a brief comparison against R's glht results.

Developing a Dashboard of Key Performance Indicators for Coronary Artery Disease Care Using Administrative Data in Slovenia

Janez Bijec, Petra Došenović Bonća and Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia
janez.bijec@gmail.com

This presentation introduces a prototype dashboard that visualizes key performance indicators (KPIs) of quality and efficiency of care for patients with coronary artery disease (CAD) in the Slovenian healthcare system. Based on administrative reimbursement data collected by the Health Insurance Institute of Slovenia (ZZZS), this work applies the Value-Based Healthcare framework and Donabedian's model to identify, calculate, and present 13 KPIs across five stages of care: mortality, re-hospitalizations, process of care, post-discharge care, and economic aspects. The dashboard was developed using the Shiny package in R, allowing for interactive, transparent, and user-friendly visualization. It compares 14 hospitals and displays regional variation across 212 municipalities. KPI values are standardized using predictions from generalized linear mixed models, which control for patient, hospital, and municipality-level risk factors. This standardization enables fair comparison between providers. The dashboard consists of three pages: a population overview with filters by date, age, sex, and CAD type; comparative scatterplots for hospital benchmarking with 95 % confidence intervals; and a regional view comparing observed vs. model-predicted values by municipality. Key findings show significant differences between providers in both quality and efficiency, even after risk adjustment. Most variability in outcomes is explained by patient characteristics, with minimal impact from hospital and municipality contexts. Data quality was assessed using structured criteria and deemed fit for purpose, despite known limitations of administrative data and small hospital sample size. This presentation will demonstrate the dashboard, explain the methodology used to develop it, and discuss its application for supporting evidence-based healthcare quality monitoring in Slovenia. Participants will gain insights into leveraging routine administrative data and advanced statistical modeling to build interactive, decision-support tools for healthcare systems.

Kurtosis: Mainly Misunderstood and Practically Useless

Gaj Vidmar and Bor Vratanar

University Rehabilitation Institute, Ljubljana, Slovenia
gaj.vidmar@ir-rs.si

Kurtosis is routinely taught in introductory statistics courses, and applied statistics textbooks continue to present ill-founded cut-off values for kurtosis (e.g., +2 and −2) as guides for judging »normality« of empirical distributions. Furthermore, the internet abounds with wrong illustrations of kurtosis (e.g., two normal distributions with the same mean and different variance presented as one leptokurtic and one platykurtic). Hence, we review the definition and true meaning of kurtosis, and present extensive simulations demonstrating extremely instability of kurtosis estimates, especially when viewed jointly with skewness. In addition, our literature search found only one single valid actual practical application of the kurtosis statistic, which is completely negligible given the unfathomable amount and breadth of online publications. The simulations were conducted with 100,000 draws of samples of size 30, 100 and 1000 from standard normal, standard uniform, lognormal, arcsine (i.e., beta with alpha and beta parameters equal to 0.5) and standard triangular distribution. Univariate strip-plots and bivariate density contours were produced for separate and joint skewness and kurtosis estimates. With the lognormal distribution, which is highly skewed, the vast majority of skewness-kurtosis pairs fell far from the theoretical population values even in samples of size 1000. In addition, kurtosis estimates of the arcsine distribution, for which the population value is −1.5, practically never fell below −2 even in samples of size 1000. We made two further sets of didactic simulations with data from standard normal distribution to illustrate the extreme sampling variability of kurtosis estimates as compared to the mean estimates. We can therefore conclude that kurtosis can safely be avoided in introductory statistics courses, it need not be routinely calculated as part of numerical data description, and it should not be presented as a criterion for assessing appropriateness of using the normal distribution as the model for an empirical dataset.

Data Science

Permutation Entropy and Statistical Complexity Analysis of Sentinel-2 Time Series for Detecting Vegetation Pest Diseases: The Case of Toumeyella parvicornis

Luciano Telesca, Nicodemo Abate and Rosa Lasaponara

Institute of Methodologies for Environmental Analysis, Tito Scalo, Italy
luciano.telesca@cnr.it

In this study, we examine Sentinel 2 (S2) time series to detect and assess pest-affected vegetation anomalies. The S2 time series was analysed using multiple vegetation indices. The analyses were performed on a case study located in Castel Porziano (central Italy), chosen due to its significant impact from Toumeyella Parvicornis (TP) in recent years. The area of Follonica, which is not yet affected by TP, was used as a comparison. Our goal is to identify patterns associated with TP in the statistical features of S2 data. The analysis employs permutation entropy and statistical complexity, which together form the basis for constructing the so-called complexity–entropy causality plane (CECP). Permutation entropy and statistical complexity provide insight into two different aspects of a dataset. Permutation entropy measures the level of intrinsic randomness: signals that are more predictable and tend to repeat a limited number of ordinal patterns exhibit lower permutation entropy, whereas signals with a greater variety of patterns and less predictability show higher values. For a fixed value of permutation entropy, statistical complexity indicates the extent to which certain ordinal patterns are favored over others. In other words, higher complexity—at a given entropy level—reflects a greater deviation from a uniform distribution, suggesting that some ordinal patterns occur more frequently than others. By computing both measures for a time series, one can simultaneously assess the randomness of the data and the degree of structural or correlational organization within its fluctuations. Analysis of the Receiver Operating Characteristic (ROC) curve indicates that these two measures are highly effective in distinguishing between infected and healthy sites. This work was supported by the project COELUM (Spies of Climate change and tools for mitigating the effects: EO and AI based methodological approach for Urban Park Management).

Local Differential Privacy for Trajectory Anonymization With Map-Matching Techniques

Gabriele Gühring, Andreas Heinrich and Di Hu

Esslingen University of Applied Sciences, Esslingen am Neckar, Germany
di.hu@hs-esslingen.de

To avoid CO2 emissions in the transport sector, it is necessary to understand, predict and precisely plan traffic flows. Trajectories of various mobility users can be collected and analyzed in order to plan future demand for public transport services and to measure emission distributions. One of the challenges hereby are the privacy concerns during the collection of mobility trajectories since they may contain sensible identity information and might be abused as the data is stored with a centralized data curator. Local differential privacy is a well-known privacy model that aims to provide privacy guarantees depending on a privacy budget for each user while collecting and analyzing data. It might solve the problem of trajectory anonymization by sending only certain features of a trajectory to a central data curator and synthesizing trajectories out of probability distributions estimated with the data curator for each feature. On the other hand, this kind of anonymization technique may lead to utility loss for each trajectory, depending on the utility measure used. In this paper, we show how local differential privacy can be used together with map matching algorithms in order to combine the anonymization of individual trajectories with different still useful utility measures i.e. Jensen-Shannon divergence.

Forecasting Storms With Meteorological Variables and Digital Attention: A Multistage Statistical and Neural Hybrid Framework

Soudeep Deb and Lizan Meryl Pereira

Indian Institute of Management Bangalore, Bangalore, India
lizan.pereira24@iimb.ac.in

Classification of storm occurrences is a significant challenge due to their infrequent occurrence and the resulting class imbalance in available datasets. Despite the growing use of machine learning and deep learning models in this domain, their performance often suffers due to data imbalance and the rarity of storm events, leading to biased or unreliable forecasts. However, accurate predictions are critical for mitigating the environmental and social damage caused by these extreme weather events. To address the limitations of existing methods, we propose a multistage ensemble approach in this paper. In the first stage, we develop a suitable algorithm (leveraging the concepts of conventional time series models as well as that of neural networks) to forecast key meteorological variables such as wind-speed, temperature, humidity, precipitation etc. Then, in the second stage, noting that the environmental features alone are not enough to predict the extreme events, we integrate digital indicators, namely information obtained from news article trends and Google Trends. These additional variables can effectively incorporate real-time signals and public attention towards storm activity. Our proposed methodology combines these inputs with the meteorological variables to classify the occurrences of storm events. The efficacy of our methodology is demonstrated using a daily dataset from Washington, D.C., spanning the years 2020 to 2024. The model's performance is evaluated using suitable metrics such as F1-score, precision-recall, and Brier score, and is shown to be superior to other techniques. Specifically, the proposed method is able to improve storm prediction accuracy in imbalanced data scenarios and contribute to the development of more effective early warning systems, which could reduce the social and environmental impact of extreme weather events.

Monitoring FAIR Data Practices: Lessons From a Preliminary Study at the University of Primorska

Ana Slavec and Haris Zukić

University of Primorska, Koper, Slovenia
ana.slavec@famnit.upr.si

In response to Slovenia's new Scientific Research and Innovation Activities Act (2021) and the accompanying Decree on the Implementation of Scientific Research Work in Accordance with the Principles of Open Science (2023), research institutions are now required to ensure open access not only to scientific publications but also to research data when projects are at least 50 % publicly funded. The national Action Plan for Open Science further mandates that research organizations increase the share of research data published in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable). However, many institutions, including the University of Primorska, currently lack the internal mechanisms to systematically collect and monitor data necessary for tracking compliance with these requirements. To address this gap, we conducted a baseline study to assess the current state of open science practices. We analyzed a simple random sample of 120 original scientific articles (bibliographic category 1.01) published by university-affiliated researchers in 2024, drawn from a total of 548 publications from all six faculties and one research institute. Each article was manually reviewed and coded for key open science indicators: open access status, use of primary or secondary data, presence of data availability statements, availability of data (e.g., in supplements or repositories), and references to data sources. The findings were triangulated with results from a researcher survey, internal reporting data, and repository records. Based on this comprehensive analysis, we developed a strategic plan to increase the proportion of publications with openly available research data, aligned with national policy goals and institutional capabilities. This study provides a replicable model for other institutions seeking to operationalize open science mandates and highlights the importance of structured monitoring in achieving systemic change.

Statistical Analysis of Emotional Engagement in Interactive vs. Non-Interactive Documentaries

Matjaž Kljun, Una Vuletić and Klen Čopič Pucihar

University of Primorska, Koper, Slovenia
89232042@student.upr.si

Documentaries have long been used to raise awareness about important topics and foster emotional connection to social and cultural issues. However, traditional documentaries rely on passive storytelling, which can limit the emotional involvement. To allow viewers to feel like active participants in the story they are exploring, interactive documentaries allow viewers to choose their own path and pace of exploration. This study explores how interactivity in documentaries influences emotional engagement by analyzing facial expression data from 44 participants who watched either an interactive or non-interactive version of a documentary about the last residents of a small Croatian island. Emotion duration was extracted for six basic emotions, identified by Ekman, using Affectiva's facial analysis software and analyzed using a non-parametric statistical test. The results show that there is no significant difference in overall emotion expression duration during video segments between the two participant groups ($p = 0.059$), while participants in the interactive group displayed significantly longer times expressing emotions during interactive content ($p = 0.020$). We also found a statistically significant monotonic relationship ($\rho \approx 0.65$, $p < 0.001$) between emotion expression across both formats. When it comes to specific emotions, we discovered that disgust ($p = 0.025$) and fear ($p = 0.046$) had significantly higher durations in the interactive group, while no significant differences were found for the remaining four basic emotions. However, we also found a moderate correlation across content types for anger, joy, and surprise, which suggests the existence of stable individual emotional responsiveness. These findings indicate that interactivity in documentaries can meaningfully enhance emotional involvement with storytelling content, particularly for certain emotions.

Application of Machine Learning to Fundamental Analysis of Securities

Aleksandr Panteleev

University of Ljubljana, Ljubljana, Slovenia
pantel2212@gmail.com

This research compares the effectiveness of linear models and more advanced machine learning techniques in automating the value investing approach articulated by Benjamin Graham and later popularized by investors like Warren Buffett. Unlike technical analysis, which relies on price and volume data, this strategy is rooted in fundamental analysis. It seeks to determine a company's fair value by evaluating its assets, historical and projected earnings, and the broader market context. The study's findings suggest that both linear models and, particularly, machine learning algorithms are highly effective at identifying undervalued securities. Notably, both linear and non-linear models demonstrated exceptional performance compared to the broader market, especially during periods of economic instability and in sideways-moving markets. However, the research also confirms that as algorithmic trading becomes more prevalent, opportunities arising from market inefficiencies are diminishing, particularly within the large-cap segment. Another critical observation from this study is the profound impact of data quality on backtesting results. The research highlights specific instances where using standard, »out-of-the-box« data from major data providers can be misleading. This can compromise the objectivity and replicability of the findings, underscoring the importance of meticulous data curation and validation in quantitative financial modeling.

INDEX

Index

```
// Backpropagation algorithm

#include <stdio.h>
#include <afx.h>
#include <math.h>
#include <stdlib.h>

#define InputN 64   // No. of neurons in the input layer
#define HN 25       // No. of neurons in the hidden layer
#define OutN 64     // No. of neurons in the output layer
#define datanum 500 // No. of training samples

void main(){
  double sigmoid(double);
  CString result = "";
  char buffer[200];
  double x_out[InputN]; // Input layer
  double hn_out[HN];    // Hidden layer
  double y_out[OutN];   // Output layer
  double y[OutN];       // Expected output layer
  double w[InputN][HN]; // Weights from input layer to hidden layer
  double v[HN][OutN];   // Weights from hidden layer to output layer

  double deltaw[InputN][HN];
  double deltav[HN][OutN];

  double hn_delta[HN];  // Delta of hidden layer
  double y_delta[OutN]; // Delta of output layer
  double error;
  double errlimit = 0.001;
  double alpha = 0.1, beta = 0.1;
  int loop = 0;
  int times = 50000;
  int i, j, m;
  double max, min;
  double sumtemp;
  double errtemp;

  // Training set
  struct{
    double input[InputN];
    double teach[OutN];
  }data[datanum];

  // Generate data samples
```

https://as.mf.uni-lj.si