

Austrian Statistical Days 2025

02.-04.09.2025 | JKU Linz

Book of Abstracts

Nicolai David Amann

University of Vienna

03.09.2025 | Invited Talk

Uncertainty quantification via cross-validation and its variants under algorithmic stability

Recently, there has been substantial interest in statistical guarantees for cross-validation (CV) methods of uncertainty quantification in statistical learning (cf. Barber et al. 2021, Liang and Barber 2024, Steinberger and Leeb 2023). These guarantees should hold under minimal assumptions on the data generating process and conditional on the training data, because numerous predictions are usually computed based on one and the same training sample. We push this objective to the limit: We prove asymptotic conditional conservativeness of CV, that is, the probability of the actual coverage probability, conditional on the training data, undershooting its nominal level vanishes asymptotically, under minimal assumptions. In particular, we impose a stability condition, require that the prediction error is stochastically bounded, and show that neither condition can be dropped in general. By way of an asymptotic equivalence result, we also show that the closely related CV+ method of Barber et al. 2021 provides exactly the same conditional statistical guarantees as CV in large samples, thereby extending the range of applicability of CV+ to the high-dimensional regime. We conclude that, in view of its marginal coverage guarantee, CV+ does indeed improve over simple CV. For our proofs we introduce a new concept called Lévy gauge, which can be of independent interest.

References

- Barber, R. F., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2021). “Predictive inference with the jackknife+”. *The Annals of Statistics* 49.1, pp. 486–507.
- Liang, R. and R. F. Barber (2024). “Algorithmic stability implies training conditional coverage for distribution-free prediction methods”. arXiv: 2311.04295 [math.ST]. url: <https://arxiv.org/abs/2311.04295>.
- Steinberger, L. and H. Leeb (2023). “Conditional predictive inference for stable algorithms”. *The Annals of Statistics* 51.1, pp. 290–311.

Johann Bacher

JKU Linz

04.09.2025 | Invited Talk

Statistical Significance versus Substantive Relevance - The contribution of Andreas Quatember

Recently, the well-known problem of statistical significance and substantive relevance has once again received more attention in research. Several solutions have been proposed, including one by Andreas Quatember (2023). These proposals will be presented and discussed from an application perspective.

Quatember, A., 2023: Different Approaches to Incorporate the Aspect of Practical Relevance in the Statistical Inferential Process. *methods, data, analyses* | Vol. 17(1), 2023, pp. 121-130, DOI: 10.12758/mda.2022.07

Merle Behr

University of Regensburg

03.09.2025 | Keynote Talk

Interpretability from Random Forest tree ensembles

In many machine learning applications, especially in healthcare and the natural sciences, interpretability is as important as prediction accuracy. Random forests, renowned for their predictive strength and ability to handle high-dimensional data, are frequently used in such contexts. This talk delves into methods for systematically analyzing the tree structure of random forest ensembles to uncover influential features and their interactions. We present theoretical consistency results supporting this interpretability framework and demonstrate its application in identifying genetic interactions that illuminate genotype-phenotype relationships.

René Böheim

JKU Linz

03.09.2025 | Invited Talk

Labor Market Challenges and Data-Driven Insights

Labor markets in modern economies face a range of complex challenges, including skill shortages and rapid de-skilling; automation and de-industrialization; societal polarization and demographic aging. These challenges are deeply connected and require evidence-based approaches for effective policy-making and intervention. I sketch current challenges for the Austrian labor market and indicate areas where evidence-based policy relies on statistical data -- and where we might want more or better data.

Jolly Mae G. Catalan

Davao Center for Health Development, Philippines

02.09.2025 | Poster Presentation

Retrospective Spatial Analysis of the COVID-19 Threat to the High-risk Population in Davao Region, Philippines from March 2020 to August 2021

(Calo, Federico, Quinco, Mia Kristine, Lachica, Zython Paul, Logrosa, Gernelyn, Diche, Zarah Jean, Roxas, Pamela Grace, Ligue, Kim Dianne, Bompat, Joshua Mar, Catalan, Jolly Mae, Yumang, Annabelle, and Mata, May Anne (2023))

According to the World Health Organization (WHO), the elderly and people with comorbidities are most vulnerable to COVID-19 infection. With this, the threats posed to the vulnerable population require interventions. The implementation of COVID-19 spatiotemporal disease surveillance strategies specifically targeting the vulnerable population in the Davao Region had been unexplored. This paper investigated the COVID-19 incidence in the Davao Region from 03 Mar 2020, the earliest recorded date of onset, to 31 Aug 2021 using geospatial tools. The variables were visualized through choropleth maps and graduated symbols, and subsequently examined through spatial autocorrelation and hotspot analysis. Hotspots across the region were observed to be in high-density areas. These areas pose greater risks of infection due to the presence of a high concentration of cases. However, high case fatality rates were found in far-flung areas where access to COVID-19 healthcare facilities is a dilemma. In the COVID-19 setting and future disease outbreaks similar to COVID-19, results from this study may provide insights to government offices and other related agencies to improve healthcare systems and programs such as providing and initiating tailor-fitted isolation and consultation mechanisms appropriate to the vulnerable population in a community.

Haoyu Chen

JKU Linz

02.09.2025 | Poster Presentation

Testing data under overdispersion

To determine whether populations share the same distribution of categorical variables, tests for homogeneity are performed. Usually, common statistical tests such as chi-squared test or Fisher's exact test are used. However, in some cases overdispersion exists within the tested data, resulting in an inflated type I error if these tests are used directly. Previous work addressed this issue by performing modified tests that incorporate the additional variance components on 2x2 contingency tables, and combine the results of different categories using multiple testing corrections. This can however be suboptimal, since the performance of multiple testing approaches often rely on the underlying data structure, and information on covariance components is also lost during this process. We therefore propose a modified approach that utilises the full 2xk contingency table, which tests all k categories simultaneously. We show using simulations that our new approach enjoys higher power under various parameter configurations in comparison. The new method is also applied to a real dataset from evolutionary biology to illustrate its applicability.

Jiří Dvořák

Charles University Prague, Czech Republic

02.09.2025 | Keynote Talk

Bayesian inference for point processes and random sets - inhomogeneity, anisotropy and more

Point processes can be used for modelling random occurrence of objects or events in a spatial or spatio-temporal domain. The natural application areas include ecology, materials science, epidemiology, and many others. Point processes also serve as building blocks for popular models of random sets, often used for modelling materials, forest populations, and more. Some models of point processes and random sets are particularly suited for Bayesian analysis. The reason is that some unobserved information, which would greatly simplify the form of the likelihood, can be supplemented in the Bayesian MCMC approach. This applies e.g. to cluster point process models (where offspring points are clustered around unobserved parent points) or the Boolean models of random sets (where the union of particles is observed but not the individual particles or their centers).

In this talk, we discuss the benefits of the Bayesian approach in fitting of anisotropic cluster point processes, where inhomogeneity in the shape of the clusters and/or the orientation of the clusters can be considered. This allows much finer analysis compared to the classical methods. We also consider the case of the Boolean models, where again various types of inhomogeneity can be investigated easily in the Bayesian setting.

Florian Ertz

University of Trier

03.09.2025 | Keynote Talk

Undergraduate Students' Use of Digital Learning Materials in a Basic Statistics Course

Since the summer semester 2021, a research project at Trier University has been investigating the use of digital components (screencasts and electronic tutorials) in a large undergraduate statistics course by analysing process data gathered in teaching software. Our main research questions are how students engage in these components, how they affect students' performance in exams, and how digital components can successfully be implemented in undergraduate statistics courses at universities. Through several semesters, we collected data on screencast and electronic tutorial usage. They include, e.g., the total time that students watched a specific video or took to solve a specific electronic tutorial, the number of streams/repetitions, and the completion rates in electronic tutorials. From the raw data we derive interesting statistics that give an impression of the students' learning behaviour in the digital space. The data show that participation rates in digital components could be improved. While some students exhibit considerable tenacity, many either do not participate at all or only skim the contents. Students make ample use of the option of repetitions. Some students delay studying and then try to catch up fast relatively late. While the degree of engagement with digital components is usually positively correlated with exam performance, there are specific patterns. A constant engagement throughout the semester proves to be especially useful.

Kamila Fačevicová

Palacký-Universität Olmütz, Czech Republic

02.09.2025 | Invited Talk

Correspondence Analysis within the Framework of Compositional Tables

Kamila Fačevicová, Karel Hron¹, Peter Filzmoser

Correspondence analysis (CA) is a classical method for exploring associations between rows and columns in contingency tables. It has been linked to the log-ratio methodology used in compositional data analysis through the limiting case of the power transformation. The log-ratio approach emphasizes relative information and is invariant to rescaling of rows and columns—properties that also characterize the analysis of compositional tables, a two-factor generalisation of compositional data. In this framework, the table is decomposed into independent and interaction parts. Applying singular value decomposition (SVD) to the interaction part yields results equivalent to CA while also allowing for the assessment of variance contributions. Moreover, both approaches can incorporate weights to downplay the influence of selected cells. The equivalence between (weighted) CA and the (weighted) analysis of compositional tables enables the use of Bayes spaces—a mathematical foundation of compositional data—for a more comprehensive understanding of CA. This perspective also provides a natural path for extending CA to multi-factorial and continuous frameworks. Theoretical results are illustrated with numerical examples, highlighting how weighting can be used to handle uncertainty associated with small values in the data.

Gianluca Finocchio

University of Vienna

03.09.2025 | Invited Talk

Model-free identification in ill-posed regression

The problem of parsimonious parameter identification in possibly high-dimensional linear regression with highly correlated features is addressed. This problem is formalized as the estimation of the best, in a certain sense, linear combinations of the features that are relevant to the response variable. Importantly, the dependence between the features and the response is allowed to be arbitrary. Necessary and sufficient conditions for such parsimonious identification - referred to as statistical interpretability - are established for a broad class of linear dimensionality reduction algorithms. Sharp bounds on their estimation errors, with high probability, are derived. To our knowledge, this is the first formal framework that enables the definition and assessment of the interpretability of a broad class of algorithms. The results are specifically applied to methods based on sparse regression, unsupervised projection and sufficient reduction. The implications of employing such methods for prediction problems are discussed in the context of the prolific literature on overparametrized methods in the regime of benign overfitting.

Benedikt Fröhlich

University of Regensburg

02.09.2025 | Poster Presentation

De-correlated Feature Importance via Local Sample Weighting in Random Forests

Feature importance statistics are a prominent and valuable tool for gaining insight into the decision process of machine learning models, but their effectiveness has well-known limitations when correlation is present among the features in the training data. In this case, the feature importance tends to be distributed among all features which are in correlation with the response-generating signal features. We propose local sample weighting random forest (losawRF) which integrates a sample weighting approach locally within the split point selection process of random forest to obtain better estimates of the marginal effects of each feature. Our approach is motivated from inverse probability weighting in causal inference and uses inverse stabilized propensity score weights to de-correlate a target feature from the remaining features. This reduces model bias locally, whenever the effect of a potential single feature is evaluated and compared to others. Moreover, losawRF comes with a natural tuning parameter, the minimum effective sample size of the weighted population, which corresponds to an interpretation-prediction-tradeoff, analog to a bias-variance-tradeoff as for classical machine learning (ML) tuning parameters. In an extensive simulation study on settings with diverse correlation patterns and regression functions we found that losawRF improves feature importance consistently. Moreover, it also enhances prediction accuracy for out-of-distribution, while maintaining a similar accuracy for in-distribution test data.

Sylvia Frühwirth-Schnatter

WU Vienna

04.09.2025 | Keynote Talk

Bayesian Learning for State Space Models

State space models (SSM) have been a useful tool for flexible time series modelling for many decades. Estimation of the unknown states can be achieved for linear Gaussian SSM via the Kalman filter, if the model parameters and the model structure are known. The first part of this talk provides an overview of Bayesian techniques for learning unknown parameters and dealing with uncertainty in the model structure.

The second part of this talk addresses recent advances in shrinkage estimation for SSM with a focus on time-varying parameter (TVP) models which are a popular tool for handling data with parameters that change over time. The horseshoe, the lasso and related shrinkage priors are a very popular tool for efficient Bayesian estimation in situations where signals need to be separated from noise. While these priors mostly have been used in applications to regression analysis, they are also very useful for SSM and TVP models by shrinking the innovation variances toward zero. The most elegant way to introduce shrinkage in TVP model is to define a so-called dynamic shrinkage process prior within a multi-layer hierarchical structure that imposes strong group sparsity within each state and simultaneously allows for conditional heteroskedasticity and occasional large jumps in turbulent periods, while most of the innovations are close to zero for most of the time.

A novel dynamic shrinkage process is discussed which has several attractive features. For one thing, this process has a well-known stationary distribution which allows insight into the mathematical properties of this shrinkage process. An important special case is a dynamic horseshoe process prior, where the stationary distribution equals a horseshoe prior. Second, efficient algorithms for Bayesian can be designed. Finally, applications to macroeconomic as well as financial time series clearly indicate that this novel dynamic shrinkage process prior reduces the risk of overfitting and increases statistical efficiency in a TVP modelling framework compared to other dynamic shrinkage process priors.

Michal Fusek

Brno University of Technology, Czech Republic

02.09.2025 | Invited Talk

Extremal index and its estimators

Michal Fusek, Brno University of Technology, Czech Republic

Jan Holešovský, Brno University of Technology, Czech Republic

The extremal index is an important parameter in the extreme value theory, since it measures short-range dependence of extreme values, and governs clustering of extremes in a stationary series. Several new estimators of the extremal index that are mostly based on interexceedance times within the peaks-over-threshold model have emerged in recent years. There are, for example, the censored estimator based on artificial censoring of interexceedance times, the truncated estimator based on truncation of interexceedance times, etc. In many cases, these estimators rely on suitable choice of auxiliary parameters and/or are derived under assumptions that are related to validity of the local dependence condition $D^{\{k\}}(u_n)$ which can have a major effect on the extremal index estimates. The aim of this contribution is to present selected estimators of the extremal index and various approaches to assessing validity of the $D^{\{k\}}(u_n)$ condition.

Zuzana Greganova

Dachverband der Sozialversicherungsträger

03.09.2025 | Keynote Talk

Individual-Based Statistics in Austria's Social Security System

This keynote provides an overview of the Personenbezogene Statistik (“Individual-Based Statistics in Austria’s Social Security System”) — a set of aggregated statistics based on microdata derived from administrative records. This statistical framework draws on comprehensive records that document every instance in which an individual is registered with any of Austria’s social insurance institutions—such as employment, pensions, or the receipt of unemployment benefits or childcare allowance. The data originate at the individual level and are then aggregated into meaningful official statistics. In particular, the Personenbezogene Statistik focuses on overlapping statuses: How many pensioners are also actively employed and/or self-employed, and in which industries? How many people receive multiple (different) pensions? How many people work multiple jobs, or are both employed and self-employed? We have used these statistics to address policy-relevant questions in light of current economic challenges. For example, we examined trends in employment among retired people—by age, gender, employment type and NACE section—to assess the impact of a new initiative in which the federal government covered part of the pension insurance contributions for working pensioners.

Florian Gundl

Statistik Oberösterreich

03.09.2025 | Invited Talk

Supporting Economic Policy with Regional Data: Experiences from Upper Austria

In times of economic uncertainty, data is essential for understanding challenges and supporting decisions. In this talk, I will give a brief overview of the types of economic data we work with at Statistics Upper Austria, the official statistical office for the region.

I will share how we gather and prepare statistical information and how we communicate it to decision-makers and the public. My goal is to show how even basic statistical indicators can help paint a clearer picture of regional developments—and how statistics play a role in responding to economic challenges.

Markus Hainy

JKU Linz

02.09.2025 | Poster presentation

Practical aspects of the virtual noise convex optimum design approach for correlated responses

We present several practically-oriented extensions and considerations for the virtual noise method in optimal design under correlation. First we introduce a slightly modified virtual noise representation which further illuminates the parallels to the classical design approach for uncorrelated observations. We suggest more efficient algorithms to obtain the design measures. Furthermore, we show that various convex relaxation methods used for sensor selection are special cases of our approach and can be solved within our framework. Finally, we provide practical guidelines on how to generally approach a design problem with correlated observations and demonstrate how to utilize the virtual noise method in this context in a meaningful way.

Anders Bredahl Kock

University of Oxford

03.09.2025 | Keynote Talk

High-dimensional Gaussian and bootstrap approximations for robust means

Anders Bredahl Kock, David Preinerstorfer

Recent years have witnessed much progress on Gaussian and bootstrap approximations to the distribution of max-type statistics of sums of independent random vectors with dimension d large relative to the sample size n . However, for any number of moments $m > 2$ that the summands may possess, there exist distributions such that these approximations break down if d grows faster than $n^{\frac{m}{2}-1}$. In this paper, we establish Gaussian and bootstrap approximations to the distributions of winsorized and trimmed means that allow d to grow at an exponential rate in n as long as $m > 2$ moments exist. The approximations remain valid under some amount of adversarial contamination. Our implementations of the winsorized and trimmed means are fully data-driven and do not depend on any unknown population quantities. As a consequence, the performance of the approximation guarantees "adapts" to m .

Arnošt Komárek

Charles University Prague, Czech Republic

04.09.2025 | Invited Talk

Statistical modelling of multivariate mixed type panel data

In longitudinal studies, multiple outcomes are routinely gathered, both continuous and discrete at each subject's follow-up visit. This leads to so called multivariate longitudinal data of a mixed type while statistical modelling methods are needed to tackle the respective datasets. For example, Wagner and Tüchler (CSDA, 2010) present Bayesian estimation of random effects models to analyze such datasets. Several other models and approaches, partly related to that of Wagner and Tüchler and also other Austrian statisticians will be discussed along with their software implementations in R.

Aneta Kostárová

Charles University Prague, Czech Republic

02.09.2025 | Poster Presentation

Specification Tests for Integer-valued Time Series

This contribution addresses models for time series of integer-valued variables. Such series arise in various applications, often as increment series for counts of interest. A model with a GARCH-type structure with the Skellam conditional distribution is considered. We propose a novel testing procedure to assess the null hypothesis that a set of integer-valued observations follows such model.

Alexander Kowarik

Statistics Austria

04.09.2025 | Invited Talk

Experience with Tailored Designs at Statistics Austria

Declining response rates and increasing costs are key drivers for enhancing the efficiency of survey designs - both in terms of cost and respondent engagement. Drawing inspiration from the tailored design principles introduced by Dillman in the 1970s, Statistics Austria developed an experimental sampling approach that optimizes mode assignment between web and personal interviews based on modelled response propensities. Logistic regression models were estimated using data from previous surveys to predict the likelihood of response in Computer-Assisted Personal Interviews (CAPI) and Computer-Assisted Web Interviews (CAWI) for all households in the sample frame. These predicted probabilities were the input for the allocation of households to a specific mode. The results of an experiment during the pilot study for the Austrian Socio-Economic Panel (ASEP) and experience from the application in other surveys are presented.

Dillman, Don A. Mail and Internet surveys: The tailored design method--2007 Update with new Internet, visual, and mixed-mode guide. John Wiley & Sons, 2011.

Tatyana Krivobokova

University of Vienna

03.09.2025 | Invited Talk

Machine learning and statistical methods for high-throughput experimental data.

Machine learning (ML) and artificial intelligence (AI) techniques are transforming the way chemical reactions are studied today. Valuable datasets from high-throughput experimentation (HTE) are increasingly being generated to better understand reaction conditions that are crucial for outcomes such as yields and selectivities. However, it is often overlooked that data from such designed experiments possess a very specific structure, which can be captured by appropriate statistical models. Ignoring these underlying data structures when applying ML/AI algorithms can result in completely misleading conclusions. In contrast, leveraging knowledge about the data-generating process together with suitable estimation approaches yields reliable, interpretable, and comprehensive insights into the chemical reaction mechanisms. A particularly complex dataset is available for the Buchwald-Hartwig amination. Using this dataset, an appropriate statistical model for HTE-generated chemical data is introduced, and a suitable parameter estimation algorithm is developed. Based on the estimated model, new insights into the Buchwald-Hartwig amination are thoroughly discussed. Our approach is directly applicable to a wide range of HTE-generated data for chemical reactions and beyond.

Ursula Laa

BOKU Vienna

03.09.2025 | Talk

Frame to frame interpolation for high-dimensional data visualisation using the woylier package

The woylier package implements tour interpolation paths between frames using Givens rotations. This provides an alternative to the geodesic interpolation between planes currently available in the tourr package. Tours are used to visualise high-dimensional data and models, to detect clustering, anomalies and non-linear relationships. Frame-to-frame interpolation can be useful for projection pursuit guided tours when the index is not rotationally invariant. It also provides a way to specifically reach a given target frame. We demonstrate the method for exploring non-linear relationships between currency cross-rates.

Patrick Langthaler

University of Salzburg

03.09.2025 | Talk

A new coefficient of separation

S. FUCHS, C. Limbach, and P. B. Langthaler

A coefficient is introduced that quantifies the extent of separation of a random variable Y relative to a number of variables $X = (X_1, \dots, X_p)$ by skillfully assessing the sensitivity of the relative effects of the conditional distributions. The coefficient is as simple as classical dependence coefficients such as Kendall's tau, also requires no distributional assumptions, and consistently estimates an intuitive and easily interpretable measure, which is 0 if and only if Y is stochastically comparable relative to X , that is, the values of Y show no location effect relative to X , and 1 if and only if Y is completely separated relative to X . As a true generalization of the classical relative effect, in applications such as medicine and the social sciences the coefficient facilitates comparing the distributions of any number of treatment groups or categories. It hence avoids the sometimes artificial grouping of variable values such as patient's age into just a few categories, which is known to cause inaccuracy and bias in the data analysis. The mentioned benefits are exemplified using synthetic and real data sets.

Gertraud Malsiner-Walli

WU Vienna

04.09.2025 | Invited Talk

Effect fusion using model-based clustering

In social and economic studies many of the collected variables are measured on a nominal scale, often with a large number of categories. The definition of categories can be ambiguous and different classification schemes using either a finer or a coarser grid are possible. Categorization has an impact when such a variable is included as covariate in a regression model: a too fine grid will result in imprecise estimates of the corresponding effects, whereas with a too coarse grid important effects will be missed, resulting in biased effect estimates and poor predictive performance.

To achieve an automatic grouping of the levels of a categorical covariate with essentially the same effect, we adopt a Bayesian approach and specify the prior on the level effects as a location mixture of spiky normal components. Model-based clustering of the effects during MCMC sampling allows to simultaneously detect categories which have essentially the same effect size and identify variables with no effect at all. Fusion of level effects is induced by a prior on the mixture weights which encourages empty components. The properties of this approach are investigated in simulation studies. Finally, the method is applied to analyse effects of high-dimensional categorical predictors on income in Austria.

Benjamin Müller

University of Innsbruck

02.09.2025 | Keynote Talk

Simultaneous estimation and model choice for big discrete time-to-event data with additive predictors

This work introduces an extension of the novel batchwise backfitting algorithm for the simultaneous estimation and model choice for big discrete time-to-event data with additive predictors. The algorithm is designed to handle large datasets efficiently while incorporating a boosting-type approach for automated variable selection. An extensive simulation study is conducted to evaluate the performance of the algorithm against established methods. The effectiveness is further demonstrated by modelling infant mortality in ten Eastern sub-Saharan African countries. The results highlight the algorithm's strong estimation performance, excellent variable selection capabilities and scalability for large datasets.

Ralf Münnich

University of Trier

04.09.2025 | Keynote Talk

Data quality and representativity – excuse or concept?

Evidence-based policies shall be based on high-quality data and statistical methods. However, in practice, the use of the word 'representativity' often induces high-quality output disregarding any possible limitations. In survey and official statistics, the European Statistics Code of Practice provides recommendations and concepts for quality in statistics. Nonetheless, the term 'representativity' in the context of quality and the corresponding application is hardly defined

The aim of this presentation is to shed light on boon or bane of the term 'representativity' in the context of surveys and statistics. Starting with an example, an attempt of a concept for representativity will be provided. It will be shown that different use cases require highly specialized settings.

Additionally, the presentation will cover classical survey statistical methods and their developments from past to present. It will also include modern methods of multi-source estimation and synthetic data generative methods including an attempt to measure uncertainty, as well as an outlook on future challenges of integrating non-sampling and big data. The presentation closes with communication of quality and representativity.

Petr Musil

University of Prague

03.09.2025 | Talk

Regional macroeconomic statistics – pitfalls of interpretation

Jana Fischerová, Petr Musil, Jaroslav Kahoun

Statistical data users are very often interested in regional breakdowns of indicators as they can either identify themselves with the results or put the results into questions. Naturally, macroeconomic statistics focuses on main economic trends with limited breakdowns of indicators. National accounts offer a number of various indicators describing national economy. All indicators are fully consistent and balanced. Regional accounts offer selected indicators only while not being in fact 'accounts'. A set of these indicators is limited to a production or income approach to GDP in basic breakdowns, household incomes, investment and employment. It is well-known that price levels among regions differ in many countries, however the official statistics does not publish regional prices levels. Therefore, regional indicators are expressed in an average price level of a given country. We, as a research team, estimated regional price levels for Czechia in the last decade. Since we observed changes in economy and consumption habits, the results have been recalculated for the reference year 2020. It is not just update because several data sources and methods have been changed. The estimated regional price levels make regional comparison of incomes and other indicators more relevant. Unsurprisingly, the highest price level is registered in the capital city Prague, while many other regions are below national average. Additionally, expenditure approach to GDP in Czech regions has been estimated, that allows analysis of differences in regional structure and the level of consumption and savings. Obviously, consumption expenditures can be adjusted to local regional price differences.

Nina Niederhametner

Statistics Austria

04.09.2025 | Invited Talk

Municipal-Level Estimation of Tourism Perception: A Machine Learning-Based Approach to Small Area Estimation

Understanding the perception and acceptance of tourism among local residents is crucial for sustainable tourism management. Tourism intensity and its impact on the local population varies in different regions of Austria. Consequently, there is a need to monitor tourism acceptance and perception on a smaller regional level than the federal province. However, interviewing a representative sample in each region or municipality in Austria is neither affordable nor feasible. Therefore, we have developed a machine-learning based Small Area Estimation (SAE) model, which builds on possible tourism indicators that can influence tourism acceptance in the municipalities.

To estimate the perception of the entire population, we propose to employ machine learning models, leveraging auxiliary data to impute the response variable for non-surveyed citizens on a unit level. We do this by linking each respondent's answer to auxiliary administrative data, including demographic information (age, sex, income), employment sector (NACE classification), and municipal-level data (tourism-related profits, number of overnight stays per capita). These auxiliary variables are available for the entire country's population. An XGBoost model is then trained on this data to predict a person's answer for the question "How do you perceive the number of tourists in your place of residence?". We predict the answers of residents not included in the survey. This allows us to aggregate both the predicted estimates and, if applicable, the actual survey responses of all residents within the same municipality, providing a single acceptance estimate per municipality. Additionally, we use a bootstrap approach to estimate the errors, and construct confidence intervals.

This approach parallels methods used in Small Area Estimation, extending the analysis beyond the sample to provide comprehensive estimates.

Lena Ortega Menjivar

BOKU Vienna

02.09.2025 | Talk

Improving rainfall distribution fits by finding rainfall event types

Lena Ortega Menjivar, Nur Banu Özcelik, Johannes Laimighofer, Svenja Fischer, Gregor Laaha

In hydrological statistics, it is a well discussed concept that there is often a lack of fit between empirical rainfall distributions and supposed theoretical (Gamma) distributions, partly due to meteorological and regional heterogeneities (f.i. Laaha 2023).

We propose to (1) improve rainfall distribution fit and (2) gain new insights into types of rainfall events by: Separating rainfall time series into rainfall events and gathering information on event characteristics such as volume, duration, and binary lightning occurrence; Clustering the event into rainfall type groups; And fitting groupwise Gamma distributions to rainfall which are then aggregated into an additive overall rainfall distribution. We apply this strategy to rainfall and lightning data from 2 exemplary Austrian weather stations.

To obtain rainfall event groups, we compare two clustering strategies: (i) PAM clustering with Gower's distance (with Euclidean and Simple Matching Distance in our case), where cluster number is determined by maximum average silhouette width; and (ii) model-based clusterwise regression guided by minimum ICL.

PAM clustering yields interpretable and well-structured event groups aligned with theoretical rainfall concepts, but offers only marginal improvement in aggregated overall rainfall distribution fit. Conversely, clusterwise regression produces less interpretable clusterings, yet leads to substantial improvements in the fit of the aggregated rainfall distributions.

In conclusion, we are not only able to significantly improve rainfall distribution fit, but also to comparatively showcase the strengths and weaknesses of very different clustering approaches in this field.

References:

Laaha G (2023). "A Mixed Distribution Approach for Low-Flow Frequency Analysis – Part 1: Concept, Performance, and Effect of Seasonality." *Hydrology and Earth System Sciences*, 27(3), 689–701. doi:10.5194/hess-27-689-2023.

Yevjevich, V (1967). "An objective approach to definitions and investigations of continental hydrologic droughts." Yevjevich, Vujica M. An objective approach to definitions and investigations of continental hydrologic droughts. Vol. 23. Fort Collins, CO, USA: Colorado State University.

Vladimir Pastukhov

University of Vienna

03.09.2025 | Invited Talk

Fused lasso nearly-isotonic signal approximation in general dimensions

We introduce and study fused lasso nearly-isotonic signal approximation, which is a combination of fused lasso and generalized nearly-isotonic regression. We show how these three estimators relate to each other and derive solution to a general problem. Our estimator is computationally feasible and provides a trade-off between monotonicity, block sparsity, and goodness-of-fit. Next, we prove that fusion and near-isotonisation in a one-dimensional case can be applied interchangeably, and this step-wise procedure gives the solution to the original optimization problem. This property of the estimator is very important, because it provides a direct way to construct a path solution when one of the penalization parameters is fixed. Also, we derive an unbiased estimator of degrees of freedom of the estimator.

Simon Pauli

JKU Linz

02.09.2025 | Poster Presentation

Estimating Identity by Descent and the Inbreeding Coefficient using Phase-Type Distributions

Identity by descent and the inbreeding coefficient are foundational metrics in population genetics, yet their accurate estimation frequently depends on detailed pedigree records — an impractical requirement for many wild or understudied populations. We introduce a novel framework that infers key mating parameters directly from genotype data, by quantifying pairwise mutational differences between alleles on an individual's two haplotypes. Central to our approach is a tailored discrete phase-type (DPH) distribution in which state-visits contribute to the distribution with mutation probability μ . Here, each state encodes the generational relation between allele lineages, until they coalesce in their most recent common ancestor τ generations in the past. The model parameters reflect species-specific mating dynamics — monogamy versus polygamy — and degrees of inbreeding. We derive a closed-form likelihood for these parameters, enabling maximum-likelihood estimation without pedigrees, and present an explicit expression for the inbreeding coefficient as a function of the estimated parameters. Building on earlier work by [Campbell](#) and by [Severson 19](#), we extend their monogamy-based frameworks to explicitly incorporate non-monogamous mating systems. Through extensive simulations, we demonstrate that our estimators of the coefficient of inbreeding are unbiased and robust: they reliably recover true values, even under complex mating behaviors not explicitly parameterized. Finally, we construct both Wald-statistic and bootstrap confidence intervals for the inbreeding coefficient; both intervals achieve desirable coverage and practical widths across simulation scenarios. This method opens the door to pedigree-free inference of inbreeding and mating structure in natural and managed populations alike.

Corinna Perchtold

JKU Linz

02.09.2025 | Talk

Temperature-humidity-related mortality in Austria: A spatio-temporal model with fine-grained climate and mortality data.

We analyse the weekly spatio-temporal distribution of mortality in Austria over the period 2002–2019 and assess the extent to which this distribution can be explained by meteorological factors. Specifically, we look at the combination temperature/humidity, as well as at the demographic variables such as age, gender, and population size. Our approach is region-specific and incorporates interactions across space-time, space-age, and age-time dimensions. The analysis is based on the fine-grained observed mortality and weather data in Austria and seeks to identify patterns and correlations between mortality and age, temperature, and gender. The results clearly show a different picture for female and male, identifying females as generally more affected by environmental factors and temporal fluctuations.

Roman Pfeiler

JKU Linz

04.09.2025 | Talk

Shrinkage in a Bayesian Zero-Inflated Negative Binomial Panel Data Model with Time-Varying Coefficients

Roman Pfeiler / Helga Wagner

We propose a flexible Bayesian panel data model for analyzing count data subject to overdispersion and zero-inflation. For this, we consider a Bayesian Zero-Inflated Negative Binomial (ZINB) regression model, which is a latent class model that partitions zeros into “structural zero” and “at-risk zero”. A “structural zero” means that the corresponding subject is not at risk of experiencing an event and, thus, will never have a positive count. An “at-risk zero”, on the other hand, implies that the subject is at risk of experiencing an event, but for some reason has reported a zero. In practice, a Logit model is used to predict whether a subject is “at-risk” and then a Negative Binomial model is fit for only those subjects who are considered “at-risk”. Bayesian inference for the ZINB model can be accomplished via MCMC sampling techniques and by using data augmentation based on Pólya-Gamma random variables, which makes Gibbs-Sampling of the parameters in Logit and Negative Binomial regression models feasible. Our Bayesian ZINB model is flexible in the sense that we allow for both time-varying regression effects and time-varying random effects. Hierarchical shrinkage priors are placed on the model parameters to avoid overfitting and to identify whether the effects are time-varying, time-invariant or zero. We evaluate our model in a simulation study and apply it to a real-world dataset.

Martin Romaňák

Charles University Prague, Czech Republic

02.09.2025 | Invited Talk

Tensor changepoint detection and eigenbootstrap

Tensor data consisting of multivariate outcomes over the items and across the subjects with longitudinal and cross-sectional dependence are considered. A completely distribution-free and tweaking-parameter-free detection procedure for changepoints at different locations is designed, which does not require training data. A CUSUM type test statistic is employed and its asymptotic properties are derived for a large number of available individual profiles. We introduce an eigenbootstrap superstructure to address the computational challenges of high dimensionality without information loss, while preserving all dependencies within and between panels. The validity of this novel and efficient resampling technique is established in this general setting. The empirical performance of the detection algorithm is evaluated through a simulation study. Finally, the fully data-driven test is applied to real-world data from EEG and psychometrics.

Theresa Scharl-Hirsch

BOKU Vienna

02.09.2025 | Talk

Clustering in Forest Recreation Monitoring

Forest areas, especially those located close to urban amenities, play an important role in Central Europe. They provide crucial cultural ecosystem services to city dwellers, contributing significantly to their psychological and physical health and well-being. The ForRest research project aims to explore current international data resources and investigate their potential and limitations in studying the social demand for nature and the supply function of forest ecosystems in a recreational context. Multiple data sources are integrated and analysed together to determine how Viennese residents typically use forest areas. The data consists of open spatial data along with socio-empirical observations based on online panel survey accompanied by Public Participation GIS. Different approaches for clustering the mixed-with-ordinal data are investigated including partitioning and model-based methods implemented in R package flexord as well as R package DBSCAN for density-based spatial clustering. The goal of the cluster analysis is to identify prototypical forest visitors in the Viennese area which will ultimately provide the city of Vienna with recommendations regarding social demand for nature and the supply function of forest ecosystems in a recreational context.

Alexandra Stadler

JKU Linz

02.09.2025 | Poster Presentation

Green LIME: Improving AI Explainability through Design of Experiments

In artificial intelligence (AI), the complexity of many models and processes often surpasses human interpretability, making it challenging to understand why a specific prediction is made. This lack of transparency is particularly problematic in critical fields like healthcare, where trust in a model's predictions is paramount. As a result, the explainability of machine learning (ML) and other complex models has become a key area of focus.

Efforts to improve model interpretability often involve experimenting with AI systems and approximating their behavior through simpler mechanisms. However, these procedures can be resource-intensive. Optimal design of experiments, which seeks to maximize the information obtained from a limited number of observations, offers promising methods for improving the efficiency of these explainability techniques.

To demonstrate this potential, we explore Local Interpretable Model-agnostic Explanations (LIME), a widely used method introduced by Ribeiro, Singh, and Guestrin (2016). LIME provides explanations by generating new data points near the instance of interest and passing them through the model. While effective, this process can be computationally expensive, especially when predictions are costly or require many samples. LIME is highly versatile and can be applied to a wide range of models and datasets. We focus on models involving tabular data, regression tasks, and linear models as interpretable local approximations.

By utilizing optimal design of experiments' techniques, we reduce the number of function evaluations of the complex model, thereby reducing the computational effort of LIME by a significant amount. We consider this modified version of LIME to be energy-efficient or "green".

Robert Stehrer

WIIW

03.09.2025 | Invited Talk

New technologies, employment and wages – Selected evidence based on Austrian microdata

The impact of new technologies on firm performance, employment and wages is debated given the rise of robots or artificial intelligence. This presentation first overviews available microdata in Austria to address such questions. Further, selected evidence on the impact of new technologies on labour market outcomes is presented based on such data. Overall, results based on various approaches suggest only a limited impact on employment or wages, though there are some differences by gender or educational attainment.

Keywords: ICT, new technologies, employment, wages

JEL codes: O33, O5

Gerhard Svolba

SAS

02.09.2025 | Talk

Feature Engineering in Predictive Modeling – And how do you feed your machine learning models?

Derived variables and data aggregations (?!?) Do we even need to talk about this topic in 2025? Modern Machine Learning and AI methods should be able to do it all automatically anyway. Or is analytic data preparation and feature engineering one of the last analytic domains where we, the data scientists, can make the difference? The presentation shows feature engineering examples from successful real-world projects. Methods are presented for both major scenarios: creating a one-row-per subject data mart for machine learning and data preparation for time series analysis. You will learn how you can convert specifics in the behavior of your analysis subjects into columns to be used in machine learning.

Problem

Recent enhancements like autotuning and advanced algorithms have increased the quality of analytical models. However, the tuning of the parameters of your machine learning models and time series forecasting models is just one way to increase the predictive power and accuracy of these models. Even these advanced models are bound by the availability of relevant features that describe the content of the data. To increase model quality, data scientists should focus on both areas, model tuning but also the preparation of relevant features in the data.

Solution

Feature engineering is a creative task. The data scientist wants to find out which methods provide the best picture of the customer behavior in the data. And creativity requires flexibility. With the analytic offering in many software packages data scientists get a rich and flexible set of methods to condense the data into actionable knowledge and input variables for the analysis.

Matthias Templ

FHNW Switzerland

04.09.2025 | Invited Talk

Advances in Robust Multiple Imputation for Complex and Non-linear Data

Missing data imputation is a fundamental task in statistical data analysis and of particular importance in survey statistics. Classical imputation methods based on are sensitive to outliers, both representative and non-representative, which can severely distort results. A robust imputation framework is presented. Moreover, when relationships between variables are highly nonlinear, flexible models such as Random Forest, XGBoost, Deep Learning Approaches, or General Additive Models provide a powerful framework for imputing missing values. While tree-based methods are robust against outliers in explanatory variables, they are not robust for all kinds of outliers, and other methods mentioned can be severely influenced by any outliers.

In this contribution, I present a robust imputation algorithm that addresses three key challenges: addressing non-linear relationships, model and imputation uncertainty, and outlier sensitivity. The approach combines a robust bootstrap scheme with stable model fitting of a generalized additive model using a Bacon algorithm and explicitly accounts for imputation uncertainty in a robust framework. The proposed method offers increased stability and reliability, especially in the presence of outlying observations.

This work aims to contribute to a long-standing tradition of methodological development in the field of survey statistics, which Andreas Quatember has significantly shaped. I will also briefly outline current challenges in robust imputation.

Irene Tubikanec

JKU Linz

04.09.2025 | Talk

Network inference via approximate Bayesian computation. Illustration on a stochastic multi-population neural mass model

In this talk, we propose an adapted sequential Monte Carlo approximate Bayesian computation (SMC-ABC) algorithm for network inference in coupled stochastic differential equations (SDEs) used for multivariate time series modeling. Our approach is motivated by neuroscience, specifically the challenge of estimating brain connectivity before and during epileptic seizures. To this end, we make four key contributions. First, we introduce a $6N$ -dimensional SDE to model the activity of N coupled neuronal populations, extending the (single-population) stochastic Jansen and Rit neural mass model used to describe human electroencephalography (EEG) rhythms, particularly epileptic activity. Second, we construct a reliable and efficient numerical splitting scheme for the model simulation. Third, we apply the proposed adapted SMC-ABC algorithm to the neural mass model and validate it on different types of simulated data. Compared to standard SMC-ABC, our approach significantly reduces computational cost by requiring fewer model simulations to reach the desired posterior region, thanks to the inclusion of binary parameters describing the presence or absence of coupling directions. Finally, we apply our method to real multi-channel EEG data, uncovering potential similarities in patients' brain activities across different epileptic seizures, as well as differences between pre-seizure and seizure periods.

Ondřej Vencálek

Palacky University in Olomouc, Czech Republic

02.09.2025 | Poster Presentation

Problem of division of the stakes from traditional and bayesian point of view

The problem of division of the stakes has an important place in the history of the development of probability theory. Its classic form and solution was suggested by Pascal and Fermat in the 17th century. Our contribution deals with both this classic form and the situation when the probability of victory of individual players in sub-games is unknown. In this situation, we propose a Bayesian solution, as Laplace suggested in the 18th century.

Nils von Norsinski

Statistics Austria

04.09.2025 | Invited Talk

Usage of Remote Sensing data at Statistics Austria – Creation of a land cover Map

At Statistics Austria we use remote sensing data for several tasks, which are land cover mapping, age of grassland, flood mapping, PV mapping and crop type mapping.

The creation of the landcover map is described in more detail.

The objective was to create a timeseries of land cover classification maps of Austria, with a resolution of 10 m and an annual interval. The resulting map should have 10 classes. The resulting map is used for different use cases like Ecosystem Accounting, SDG monitoring like Mountain Green Cover Index or land price statistics.

The land cover map is created since 2019. Time series of Sentinel-1 gamma-nought corrected monthly mean and Sentinel-2 data and a digital elevation model are used as remote sensing data. As reference data the digital cadastre data and land parcel identification system data are used. Within the polygons of these reference data 4000 points are randomly distributed per class. The classes of these data are reclassified according to the 10 classes which the model should discriminate. Certain ambiguous classes are removed. To test the quality of the training data we used Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP).

With this training dataset a Histogram-Based Gradient Boosting Model is trained. The model achieves good classification accuracy for most classes, but certain ambiguous classes such as sparsely vegetated land have expectedly to produced weaker results. When predicting the model with remote sensing data from a year the model is not trained on, it needs the ability to generalise. To do so we also trained a model based on reference data from the years 2019 to 2022 and predicted on 2023 data.

Gregor Zens

IIASA

04.09.2025 | Keynote Talk

Bayesian Matrix Factor Models for Demographic Analysis Across Age and Time

Analyzing demographic data collected across multiple populations, time periods, and age groups is challenging due to the interplay of high dimensionality, demographic heterogeneity among groups, and stochastic variability within smaller groups. This paper proposes a Bayesian matrix factor model to address these challenges. By factorizing count data matrices as the product of low-dimensional latent age and time factors, the model achieves a parsimonious representation that mitigates overfitting and remains computationally feasible even when hundreds of subpopulations are involved. Smoothness in age factors and a dynamic evolution of time factors are achieved through informative priors, and an efficient Markov chain Monte Carlo algorithm is developed for posterior inference. Applying the model to Austrian district-level emigration data from 2002 to 2023 demonstrates its ability to reconstruct demographic processes using only a fraction of the parameters required by conventional factor models. Extensive cross-validation and out-of-sample forecasting exercises show that the proposed matrix factor model consistently outperforms standard benchmarks. Beyond statistical demography, the framework holds promise for a wide range of applications involving noisy, heterogeneous, and high-dimensional non-Gaussian matrix-valued data.

Hao Zhu

JKU Linz

04.09.2025 | Talk

Bayesian optimal experimental prediction design for physical experiments informed by computer experiments

We consider Bayesian optimal experimental design for optimal prediction at specified design configurations for physical systems of interest. We assume that a computer model describing the physical system is available to make predictions of the output of interest depending on the design variables as well as some additional calibration parameters. However, the computer experiment is not a perfect description of the physical system, there is model error. In addition, the values of the calibration parameters giving the best representation of the real physical system by the computer model are also unknown. In order to take account of these uncertainties, we employ the calibration model of Kennedy and O'Hagan (2001), who use Gaussian processes to model the computer experiments as well as the model discrepancy. We propose to use the mutual information between the observations at the design and the prediction points as our design criterion and employ Bayesian experimental design to account for parameter uncertainty. This approach is compared to other approaches where other design criteria like the integrated mean squared prediction error are used or parameter uncertainty is neglected. In our examples, we consider batch design, sequential designs, as well as myopic adaptive designs.

Georg Zimmermann

University of Salzburg

02.09.2025 | Poster Presentation

A systematic empirical comparison of different statistical analysis approaches for a multi-aspect analysis of clinical trial data in Epidermolysis Bullosa

Georg Zimmermann, Martin Geroldinger, Linda Hajdari, Wanda Lauth, Martin Laimer, Verena Wally, Johann W. Bauer

The servEB project (WISS 2025, federal state of Salzburg, 20102/F2300645-FPR) combines clinical expertise, advanced statistical analyses, and AI-driven image classification technology to advance Epidermolysis Bullosa (EB) research. By fostering collaboration between clinicians, statisticians, IT experts and patient representatives, the project aims to improve the design and analysis of clinical trials in several ways.

In particular when defining meaningful endpoints, multiple aspects of the disease have to be considered, including quantitative, validated outcomes (e.g., number of lesions) as well as patient-relevant outcomes (e.g., quality of life, pain, pruritus). Accordingly, the statistical analysis approach should appropriately account for the multi-faceted characteristics of those outcomes. Therefore, we address this challenge by systematically comparing a range of uni- and multivariate statistical methods with respect to both their empirical performance (i.e., type I error rates and power) as well as the interpretation and the properties of the respective estimands. Based on the results, we discuss some recommendations that are informed by statistical as well as clinical expertise.

Alonso Zuniga Irigoien

WU Vienna

03.09.2025 | Invited Talk

Targeted Sensitive Covariate Privatization in Optimal Policy Learning

We consider a treatment assignment problem where a decision-maker (DM) aims to maximize a target functional while respecting privacy constraints on “sensitive” covariates (e.g., race, gender, religion). To address this, we introduce an econometric framework in which a DM jointly learns a privatization mechanism, i.e., a transformation of the non-sensitive covariates that breaks their dependence to the sensitive ones, and a treatment-assignment rule. We propose a type of empirical success policy and show that its expected regret decays at the optimal \sqrt{n} -rate. We validate our theoretical results through different simulations and showcase the advantages of our method with two empirical applications. This talk is based on joint work with David Preinerstorfer (WU Vienna).