SYSØR 2024

4th Spanish Young Statisticians and Operational Researchers Meeting

Conference proceedings

Santiago de Compostela, 19–21 June 2024

Seio
Sociedad de Estadística e Investigación Operativa

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DEPARTAMENTO DE ESTATÍSTICA, ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

Committees

Organizing Committee

Eduardo García Portugués (**Chair**) – Universidad Carlos III de Madrid
Jose Ameijeiras Alonso (**Co-chair**) – Universidade de Santiago de Compostela
Javier Álvarez Liébana – Universidad Complutense de Madrid
Diego Bolón Rodríguez – Universidade de Santiago de Compostela
María José Ginzo Villamayor – Universidade de Santiago de Compostela
Andrea Meilán Vila – Universidad Carlos III de Madrid
Iria Rodríguez Acevedo – Universidade de Santiago de Compostela
Alejandro Saavedra Nieves – Universidade de Santiago de Compostela

Scientific Committee

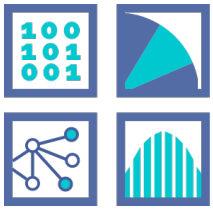
Beatriz Sinova (**Co-chair**) – Universidad de Oviedo
Vanesa Guerrero Lozano (**Co-chair**) – Universidad Carlos III de Madrid
Carmen Minuesa Abril – Universidad de Extremadura
Elena Castilla González – Universidad Rey Juan Carlos
Marta Baldomero Naranjo – Universidad de Cádiz

Editors: Organizing Committee

Conference website: http://eamo.usc.es/pub/sysorm4/

Organizers and sponsors

Organizers



Seio
Sociedad de Estadística e Investigación Operativa

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA, ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

Sponsors

EURO
THE ASSOCIATION OF EUROPEAN OPERATIONAL RESEARCH SOCIETIES

modestya
Modelos de Optimización, Decisión, Estadística y Aplicaciones

CITMAGA
CENTRO DE INVESTIGACIÓN E TECNOLOXÍA MATEMÁTICA DE GALICIA

uc3m | Universidad Carlos III de Madrid
Instituto Flores de Lemus

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE FILOLOXÍA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
VICERREITORÍA DE ESTUDANTES E CULTURA

Deputación DA CORUÑA

Contents

Foreword

We are honored to present the 4th Spanish Young Statisticians and Operational Researchers Meeting (SYSORM), the fourth edition of the conference series of the Spanish Society of Statistics and Operations Research (SEIO) that is organized for and by young researchers. In this edition, SYSORM makes a pilgrimage to Santiago de Compostela (19–21 June 2024), a historic city with a university that has a long tradition in Statistics and Operations Research. The present edition of SYSORM builds on the success of previous editions held in Granada (13–15 November 2017), El Escorial (5–7 June 2019), and, after a pandemic delay, Elche (21–23 September 2022).

The aim of the SYSORM conference series is to represent and give visibility to the newer generations of talented researchers in Statistics and Operations Research, with research topics ranging from methodological to applied, and to foster professional communication among them, both nationally and internationally. To this end, the 4th SYSORM brings together 40 young researchers (young: having less than three years of postdoctoral experience) presenting contributed talks in a single session. In keeping with its tradition, the meeting also features four renowned plenary speakers in Statistics and Operations Research: Christophe Croux (KU Leuven), Federico Perea Rojas-Marcos (Universidad de Sevilla), Maria Paola Scaparra (University of Kent), and Lola Ugarte (Universidad Pública de Navarra).

The scientific quality, diversity, and relevance of the contributions to the 4th SYSORM are a testament to the talent and dedication of the young researchers in Statistics and Operations Research in Spain. We express our gratitude to the participants for their scientific contributions, and to the members of the Scientific Committee, formed by five Ramiro Melendreras Award recipients, for their efforts in ensuring the scientific quality of the meeting. The 4th SYSORM is a reality thanks to the tireless work of the members of the Organizing Committee, who have worked to provide a memorable conference experience while adhering to the SYSORM organizational guidelines, and to whom we are very grateful.

We acknowledge the generous funding of the Spanish Society of Statistics and Operations Research (SEIO), The Association of European Operational Research Societies (EURO), the Modelos de Optimización, Decisión, Estadística y Aplicaciones (MODESTYA) research group, the Galician Centre for Mathematical Research and Technology (CITMAga), and the Flores de Lemus Institute (IFL), which made the meeting possible. We also thank the Facultade de Filoloxía and the Vicerreitoría de Estudantes e Cultura from the Universidade de Santiago de Compostela, and the Deputación de A Coruña, for their support.

We hope you enjoy your time at the 4th SYSORM and that you make the most of it scientifically and socially!

The Chairs of the Organizing Committee
Santiago de Compostela, 18 June 2024

Schedule

All sessions are held in the "Salón de Graos" of the Facultade de Filoloxía at Universidade de Santiago de Compostela.

Tuesday 18

- **19:30 – 21:30** Welcome cocktail at Fonseca

Wednesday 19

- **09:00 – 09:20** Registration
- **09:20 – 09:50** Opening session
- **09:50 – 10:50** **Plenary Session 1** STATS *Chair: Beatriz Sinova*
  - Robust correlation measures and maximum association estimators (Christophe Croux)
- **10:50 – 10:55** Short break
- **10:55 – 11:35** **Contributed Session 1** *Chair: Paula Segura*
  - 10:55 – 11:15. The 2D cutting stock problem with variable-sized stock (Paula Terán Viadero) OR
  - 11:15 – 11:35. Optimising Ryanair schedule (Sergio Vivó Sánchez) OR
- **11:35 – 12:00** Coffee break
- **12:00 – 13:20** **Contributed Session 2** *Chair: Albert Solà*
  - 12:00 – 12:20. Labour force indicators under bivariate Fay–Herriot model with correlated time and area effects (Esteban Cabello García) STATS
  - 12:20 – 12:40. Metaheuristics for the bi-objective RCPSP with time-dependent resource costs (Sofía Rodríguez Ballesteros) OR
  - 12:40 – 13:00. On the stationary distribution of non-local spatial branching processes with immigration (Pedro Martín Chávez) STATS
  - 13:00 – 13:20. The connection scheduling problem (Laura Davila Pena) OR
- **13:20 – 13:30** Short break
- **13:30 – 14:30** **Contributed Session 3** *Chair: Pedro Martín Chavez*
  - 13:30 – 13:50. Conditional expectations given the sum of independent heavy-tailed random variables (Patricia Ortega-Jiménez) STATS
  - 13:50 – 14:10. Impact of domain reduction techniques in polynomial optimization: a computational study (Ignacio Gómez-Casares) OR
  - 14:10 – 14:30. Generative invariance: causal extrapolation without exogeneity (Carlos García Meixide) STATS
- **14:30 – 16:00** Lunch at Auditorio de Galicia

- **16:00 – 17:20** **Contributed Session 4** *Chair: Laura Davila*
  - 16:00 – 16:20. Sequencing games with batch-ordered jobs (Iago Núñez Lugilde) **OR**
  - 16:20 – 16:40. Optimal subdata selection for prediction of streaming data in linear models (Álvaro Cia-Mina) **STATS**
  - 16:40 – 17:00. The Fesenthal power index for simple games with a priori unions (Alicia Mascareñas) **OR**
  - 17:00 – 17:20. Augmented designs for discrimination between constant absolute vs relative error (Carlos de la Calle-Arroyo) **STATS**
- **17:20 – 17:40** Water break
- **17:40 – 19:20** **Contributed Session 5** *Chair: Carlos de la Calle Arroyo*
  - 17:40 – 18:00. Balancing palatability, sustainability and cost in the menu planning problem using a Wierzbicki achievement function (Francisco Martos-Barrachina) **OR**
  - 18:00 – 18:20. Bayesian longitudinal beta regression for heart failure competing risk models (Jesús Gutiérrez Botella) **STATS**
  - 18:20 – 18:40. Data-driven models to support decisions in critical care (Daniel Garcia-Vicuña) **OR**
  - 18:40 – 19:00. A decision analysis model for colorectal cancer screening (Daniel Corrales) **STATS** **OR**
  - 19:00 – 19:20. Logistic regression and prediction methods with missing data (Susana Rafaela Guimarães Martins) **STATS**
- **19:20 – 21:00** Free time
- **21:00 –** Dinner at Auditorio de Galicia

Thursday 20

- **09:15 – 10:15** **Plenary Session 2** OR *Chair: Vanesa Guerrero*
  - Time indexes and other discretizations: are they worth it? (Federico Perea)
- **10:15 – 10:20** Short break
- **10:20 – 11:20** **Contributed Session 6** *Chair: Patricia Ortega*
  - 10:20 – 10:40. On uniqueness of the set of k-means (Luis Alberto Rodríguez Ramírez) STATS
  - 10:40 – 11:00. Optimal allocation of helicopters to firefighting operations: model and heuristic algorithm (Marta Rodríguez Barreiro) OR
  - 11:00 – 11:20. High density region estimation for manifold data (Diego Bolón) STATS
- **11:20 – 11:40** Coffee break
- **11:40 – 13:00** **Contributed Session 7** *Chair: Luis Alberto Rodríguez*
  - 11:40 – 12:00. Integrating shape-restricted multivariate adaptive regression splines and random forest: methodology and applications (Víctor Javier España Roch) STATS OR
  - 12:00 – 12:20. Survival analysis modelling for high- dimensional data (María del Pilar González Barquero) STATS
  - 12:20 – 12:40. A novel column generation framework for one- form-many counterfactual analysis (Jasone Ramírez-Ayerbe) OR
  - 12:40 – 13:00. A framework for adversarial regression (Pablo García Arce) STATS
- **13:00 – 13:10** Short break
- **13:10 – 14:30** **Contributed Session 8** *Chair: Iago Núñez*
  - 13:10 – 13:30. A distributionally robust optimisation approach to fair credit scoring (Pablo Casas Cendón) OR
  - 13:30 – 13:50. Temporal M-quantile models and robust bias-corrected small area predictors (María Bugallo Porto) STATS
  - 13:50 – 14:10. A profit-driven churn prevention approach within predict-and-optimize (Nuria Gómez-Vargas) STATS OR
  - 14:10 – 14:30. Digit analysis using Benford's law: a Bayesian approach (Pedro Fonseca) STATS
- **14:30 – 15:50** Lunch at Auditorio de Galicia
- **15:50 – 17:10** **Contributed Session 9** *Chair: Francisco Martos-Barrachina*
  - 15:50 – 16:10. Covariance dependence in mixture cure models using distance correlation (Blanca Estela Monroy Castillo) STATS
  - 16:10 – 16:30. The clustered-state Markovian arrival process in recurrent processes with terminal event (Álvaro Díaz Pérez) STATS
  - 16:30 – 16:50. A new formulation for the Chinese postman problem with load-dependent costs (Paula Segura Martínez) OR
  - 16:50 – 17:10. Density-based tests for the k-sample problem with left-truncated data (Adrián Lago) STATS
- **17:10 – 17:45** Free time
- **17:45 – 20:00** Old town tour
- **20:00 – 21:30** Free time
- **21:30 –** Dinner at A Horta d'Obradoiro

Friday 21

- **09:15 – 10:15** **Plenary Session 3** STATS *Chair: Elena Castilla*
  - Exploring disease mapping: models and applications (Lola Ugarte)
- **10:15 – 10:20** Short break
- **10:20 – 11:20** **Contributed Session 10** *Chair: Daniel García Vicuña*
  - 10:20 – 10:40. Optimal participation of energy communities in electricity markets under uncertainty. A multi-stage stochastic programming approach (Albert Solà Vilalta) OR
  - 10:40 – 11:00. Parameter estimation for a bivariate Wiener model subject to imperfect maintenance and varying observation strategies (Lucía Bautista Bárcena) STATS
  - 11:00 – 11:20. L1-approximation of supply curves (Zehang Li) OR
- **11:20 – 11:45** Coffee break
- **11:45 – 13:05** **Contributed Session 11** *Chair: Lucía Batista*
  - 11:45 – 12:05. Classifiers based on minimum spanning trees (Iria Rodríguez Acevedo) STATS
  - 12:05 – 12:25. Improving the estimation of production functions through machine learning: a gradient boosting approach (María D. Guillén García) STATS OR
  - 12:25 – 12:45. Statistical inference in random slope mixed models for small area estimation (Naomi Diz-Rosales) STATS
  - 12:45 – 13:05. Cost-sensitive semi-parametric classification model (Jorge C. Rella) STATS
- **13:05 – 13:10** Short break
- **13:10 – 14:10** **Plenary Session 4** OR *Chair: Marta Baldomero*
  - The role of OR in shaping resilient and sustainable communities: a case study and future prospects (Maria Paola Scaparra)
- **14:10 – 14:25** Closing session
- **14:25 – 15:00** Finger food lunch at Filoloxía
- **16:00 – 20:00** Visit to Ons Island
- **21:30 – 22:30** Free time
- **22:30 –** Dinner at Ferradura

Wednesday 19

SYSØR 2024
4TH SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Robust correlation measures and maximum association estimators

19th June
09:50–10:50
Plenary
Contributed
Session 1

Christophe Croux

Faculty of Economics and Business, KU Leuven, Belgium

Christophe Croux is a professor in Statistics and Econometrics at the KULeuven. My research expertise is on robust statistics, particularly multivariate methods resistant to outliers, and applied econometrics. My current research interest is twofold (i) fundamental research in robust statistics (ii) forecasting in high dimensions. I obtained a PhD in science in 1993, published more than 130 articles, and I'm very proud of my 19 PhD students.

The standard correlation measure, also known as the Pearson correlation coefficient, is not very robust. One single outlier can move its value towards the boundary, ie one. In the first part of this talk we will review some robust alternatives for the Pearson correlation, with emphasis on rank correlation measures. We will show that they are much more robust with respect to outliers, and the loss in statistical efficiency they pay for this is not large [2]. We call such alternatives for the Pearson correlation *association* measures.

In the second part of the talk we focus on the general regression model [3]. In this model the response is an unspecified non-decreasing function of a linear combination of predictor variables and an error term. The linear model, the binary choice (eg probit) and a censored regression model (eg Tobit) model are all special cases. The parameter of such a model may be estimated using maximum association estimators, who maximize an association measure between a linear combination of the predictors and a the response.

We study asymptotic properties of these maximum association estimators, extending results of [1] for the linear model. For the Spearman rank correlations mesure, results were obtained by [4]. We show that consistency, derive expressions for the influence function and compute asymptotic variances for several models and associations measures. Finite sample efficiencies are investigated by means of simulation.

We conclude that maximimum association estimators, which have been introduced long time ago, are very powerful: they combine robustness and efficiency. Developing fast algorithms to compute them remains a challenge.

Keywords: influence functions; rank correlation; robustness.

Bibliography

[1] Alfons, A., Croux, C. and Filzmoser, P. (2017). Robust maximum association estimators. *Journal of the American Statistical Association*, 112:436–445. doi:10.1080/01621459.2016.1148609.

[2] Croux, C. and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Stat Methods and Applications*, 19:497–515. doi:10.1007/s10260-010-0142-z.

[3] Han, A.K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics*, 35:303–316. doi:10.1016/0304-4076(87)90030-3.

[4] Sherman, R.P. (1993). The Limiting Distribution of the Maximum Rank Correlation Estimator. *Econometrica*, 61:123–137. doi:10.2307/2951780.

The 2D cutting stock problem with variable-sized stock

19th June
10:55–11:15
Contributed
Session 1

Paula Terán-Viadero[1], Antonio Alonso-Ayuso[2] and F. Javier Martín-Campo[1]

[1]Department of Statistics and Operations Research, Interdisciplinary Mathematics Institute, University Complutense of Madrid, Spain; [2]DSLAB–CETINIA, Rey Juan Carlos University, Madrid, Spain

Paula Terán-Viadero is a third-year PhD student in the PhD IMEIO at University Complutense of Madrid (UCM) under the grant for National Industrial PhD. Her research activity is focused on the development of mixed integer linear models to solve real-world problems arising from different companies. Before enrolling her PhD, she worked in the private sector as OR coordinator developing scheduling models for the hospitality sector. She did a MSc in Mathematical Engineering (UCM) and a BSc in Mathematics (UC).

The Cutting Stock Problem (CSP) is a well-known combinatorial problem that arises in many real-world applications and belongs to the family of Cutting and Packing problems. The classical version of the CSP consists of cutting a set of small items, given by their length, width and demand, from a set of larger objects, while minimising the amount of raw material used. This work is being developed in collaboration with a Spanish company in the honeycomb cardboard industry sector. It is commonly assumed that the CSP has predetermined and known stock dimensions. However, the company we are working with produces its own panels and can therefore decide the stock dimensions based on the orders received. This leads us to a new variant of the CSP, known as the 2-dimensional CSP with Variable-Sized Stock (2DVSCSP), where the stock dimensions (width and length) are to be decided.

The literature on the 2DVSCSP is limited, the problem was introduced in K. Hadj Salem et al. (2023) [1] applied to the textile sector. Two models are presented but are useful only with cases where the demand of items is very low. Later, P. Terán-Viadero et al. (2024) [2] presented a model for a simplified version where only 1-item patterns were considered. In this work we present a mixed-integer linear optimisation model for 2-stage exact guillotine cutting patterns based on the idea presented in [2] able to generate n-item patterns allowing us to tackle much more complex problems. In addition, a pre-processing analysis is conducted to significantly reduce the model dimensions, enabling the solution of large instances, which is a key challenge when using exact approaches to solve Cutting and Packing problems. The model has been validated using real company data rich in variability, resulting in a reduction of the company's current waste by almost 50%.

Keywords: cutting; mixed integer linear optimisation; variable-sized stock; 2-stage guillotine.

Bibliography

[1] Hadj-Salem, K., Silva, E., Oliveira, J.F., and Carravilla, M.A. (2023). Mathematical models for the two-dimensional variable-sized cutting stock problem in the home textile industry. *European Journal of Operational Research*, 306(2):549–566. doi:10.1016/j.ejor.2022.08.018.

[2] Terán-Viadero, P., Alonso-Ayuso, A., and Martín-Campo, F.J. (2024). A 2-dimensional guillotine cutting stock problem with variable-sized stock for the honeycomb cardboard industry. *International Journal of Production Research*, 1:483–500. doi:10.1080/00207543.2023.2279129.

SYSØIR 2024
4TH SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

Optimising Ryanair schedule

19th June
11:15–11:35
Contributed
Session 1

Sergio Vivó Sánchez[1] and Federico Perea Rojas-Marcos[2]

[1]Optimization Engineering Team, Ryanair, Spain; [2]Department of Applied Mathematics II, University of Seville, Spain

Sergio Vivó is an Optimization Engineer at Ryanair and second-year Industrial PhD student specializing in Operations Research at the University of Seville. In Ryanair, he is engaged in crafting solutions towards optimizing airline operations, with a primary focus on schedule optimization, central theme of his PhD. Prior to joining Ryanair, he contributed as a Data Scientist at Prodevelop. He holds a MSc in Data Science and Analytics at University of Cardiff and a BSc in Industrial Engineering at Polytechnic University of Valencia.

The airline industry plays a key role allowing economic growth, cultural exchange and global connectivity. Among all the airlines, Ryanair, stands out as Europe's leading low-cost carrier, focused on cost efficiency and accessibility. Ryanair is operating more than 21000 flights per week using more than 500 aircraft that results in 185 million passengers moving through 2300 routes between 225 airports in 36 countries [2]. The scheduling process within the industry plays a pivotal role as it involves the coordination of aircraft, crew and airport resources. In order to enhance scheduling efficiency and maximize resource utilization, optimizations models are used in the process of building the schedule. Operations research has significantly influenced the scheduling processes of airlines since a few decades, as is was shown by Barnhart et al. [1] in 2003. This research pushes for innovative approaches to address the complexities of scheduling in the dynamic airline environment.

This presentation focuses on the core aspect of the PhD research, building the Ryanair schedule. From a conventional linear programming approach where commercial solvers only handle a limited portion, to more crafted solutions employing matheursitics based on the initial formulation. Additionally, a column generation solution will be introduced like Xin, W. (2022) [3] but tailored to accommodate the operational constraints inherent in Ryanair's modus operandi. The research also includes a model designed not to construct schedules from scratch, but to enhance the resilience of an existing schedule. This involves producing larger turnarounds between flights, reinforcing the schedule's ability to deal with disruptions and delays and promoting recoverability. This post-processing tool aims also to reduce existing violations in the schedule.

Keywords: airline; column generation; linear programming; matheursitics; schedule.

Bibliography

[1] Barnhart, C., Cohn, A., Johnson, E., Klabjan, D., Nemhauser, G., and Vance, P. (2003). Airline Crew Scheduling. In *Handbook of Transportation Science*, pp. 517–560. doi:10.1007/0-306-48058-1_14.

[2] Ryanair Group Annual Report 2023. https://investor.ryanair.com/wp-content/uploads/2023/07/Ryanair-2023-Annual-Report.pdf.

[3] Xin, W., Xuting, S., Hoi-Lam, M., and Yige, S. (2022). A column generation approach for operational flight scheduling and aircraft maintenance routing. *Journal of Air Transport Management*, 105:102270. doi:10.1016/j.jairtraman.2022.102270.

Labour force indicators under bivariate Fay–Herriot model with correlated time and area effects

19th June
12:00–12:20
Contributed
Session 2

Esteban Cabello[1], María Dolores Esteban[1], Domingo Morales[1] and Agustín Pérez[2]

[1]Center of Operations Research Institute, Miguel Hérnandez University of Elche, Spain;
[2]Department of Economic and Financial Studies, Miguel Hernández University of Elche, Spain.

Esteban Cabello is a second-year PhD student in the PhD in Statistics, Optimization and Applied Mathematics at Miguel Hérnandez University of Elche. His research activity is focused on temporal and spatial linear mixed models applicable to the estimation of small-area multivariable indicators. Before enrolling his PhD, he did a MSc in Applied Statistics at Granada University and a BSc in Mathematics at Málaga University.

Small area estimation (SAE) involves the estimation of parameters in small subsets (called small areas) of an original population, it improves the efficiency of direct estimators by combining methodologies from survey sampling and finite population inference with statistical models. The basic area-level model is the Fay–Herriot (FH) model [3]. Benavent and Morales [1] introduced a class of multivariate Fay–Herriot (MFH) models with one random effect per component of the target vector and different covariance patterns between the components of the vector of random effects. After that, they extended the bivariate Fay–Herriot model to the temporal setup [2] but only considering correlation from the sampling errors, not from random errors.

Our work continues the investigation of Benavent and Morales [2] and develops an area-level temporal bivariate linear mixed model that integrates correlated time effects for estimating socio-economic indicators in small areas. We also provide an approximation for the matrix of mean squared errors (MSE) and propose four MSE estimators, one of which uses a plug-in approach while the others rely on parametric bootstrap procedures. Three simulation experiments are used to evaluate the performance of the fitting algorithm, the predictors and the MSE estimators. Finally, an application to real data from the Spanish Living Conditions Survey 2016 to 2022 is carried out.

Keywords: bootstrap; Fay-Herriot models; linear mixed models; SAE; SLCS.

Bibliography

[1] Benavent, R., and Morales, D. (2016). Multivariate Fay-Herriot models for Small Area Estimation. *Computational Statistics and Data Analysis*, 94:372–390. doi:10.1016/j.csda.2015.07.013.

[2] Benavent, R., and Morales, D. (2021). Small Area Estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods and Applications*, 30(1):195–222. doi:10.1007/s10260-020-00521-x.

[3] Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277. doi:10.2307/2286322.

Metaheuristics for the bi-objective resource-constrained project scheduling problem with time-dependent resource costs

19th June
12:20–12:40
Contributed
Session 2

Sofía Rodríguez-Ballesteros[1], Javier Alcaraz[1] and Laura Anton-Sanchez[1]

[1]Center of Operations Research, Miguel Herández University of Elche, Spain.

Sofía Rodríguez-Ballesteros is a third-year PhD student in Statistics, Optimization and Applied Mathematics at Miguel Hernández University of Elche. Her research activity focuses on new extensions for the basic Project Scheduling Problem, and the development of solving methods adapted to specific contexts. Before enrolling her PhD, she worked on her Master's thesis at the Technological Institute for Industrial Mathematics, and did a MSc in Statistics and a BSc in Mathematics, both at University of Santiago de Compostela.

The bi-objective resource-constrained project scheduling problem (RCPSP) with time-dependent resource costs was recently introduced in [1] and consists of scheduling a set of activities subject to precedence and resource constraints, minimizing the makespan and the total cost for resource usage. Precisely, costs are determined by the resource being considered together with the time it is used. In such a multi-objective context, solving the aforementioned problem poses a challenge, as both objectives conflict with each other, giving rise to a set of trade-off optimal solutions, commonly known as the Pareto front (PF). Given that many medium or large-sized instances of this problem cannot be solved by exact methods, the development of metaheuristics to find the PF is necessary. So far, only one metaheuristic has been developed to solve this problem.

In [2], we have implemented six additional multi-objective evolutionary algorithms (MOEAs), representing different paradigms, and subsequently, an exhaustive comparison of their performance has been carried out. In particular, all the compared MOEAs share the same encoding and main operators, focusing the comparison on the general algorithm framework rather than specific versions. Metaheuristic algorithms typically yield an approximation of the optimal PF, prompting the question of how to assess the quality of the obtained approximations. To this end, a computational and statistically supported study is conducted, choosing a benchmark of bi-criteria resource-constrained project scheduling problems and applying a set of performance measures to the solution sets obtained by each methodology. The results show that there are significant differences among the performance of the metaheuristics evaluated.

Keywords: metaheuristic; multi-objective; Pareto front; performance indicator; resource-constrained project scheduling.

Bibliography

[1] Alcaraz, J., Anton-Sanchez, L., and Saldanha-da-Gama, F. (2022). Bi-objective resource-constrained project scheduling problem with time-dependent resource costs. *Journal of Manufacturing Systems*, 63:506–523. doi:10.1016/j.jmsy.2022.05.002.

[2] Rodríguez-Ballesteros, S., Alcaraz, J., and Anton-Sanchez. L. (2024). Metaheuristics for the bi-objective resource-constrained project scheduling problem with time-dependent resource costs: An experimental comparison. *Computers & Operations Research*, 163:106489. doi:10.1016/j.cor.2023.106489.

On the stationary distribution of non-local spatial branching processes with immigration

19th June
12:40–13:00
Contributed
Session 2

Pedro Martín-Chávez[1], Andreas Kyprianou[2], Miguel González[1] and Inés del Puerto[1]

[1]Department of Mathematics, University of Extremadura, Spain; [2]Department of Statistics, University of Warwick, United Kingdom

Pedro Martín-Chávez is a third-year PhD student in the PhD program Modeling and Experimentation for Science and Technology at University of Extremadura (UEx). His research activity is focused on stochastic modeling through branching processes, and he is funded by fellowship FPU20/06588 of the Spanish Ministry of Universities. Before enrolling his PhD, he worked at UEx as Substitute Lecturer and Research Assistant, and did a MSc in Mathematics at UEx and a BSc in Mathematics and Physics at University of Sevilla.

Branching processes are stochastic models used to understand complex systems. They focus on determining how the total number of elements evolves, taking into account that each of them branches following certain rules. For instance, population dynamics studies the size of populations (systems) in which their individuals (elements) reproduce (branch) giving rise to offspring following a certain probability distribution (rules). Within the large family of branching models, our focus is on non-local branching Markov processes. In this setting, we have a particle system where particles move through a space while they origin new particles located not necessarily in the same position of the parent. Applications may be found for neutron chain reactions in nuclear fission reactors. In addition, we deal with superprocesses that can be seen as limits of the previous model in which particles split and die at infinite rates, and the population is rescaled obtaining a measure ("population density"). Notably, recent works such as [1] and [2] have respectively provided comprehensive treatments of these models, laying the foundation for our research.

Within this context, our contribution extends both models by introducing immigration. We establish conditions under which these processes exhibit limiting distributional stability, assuming a Perron–Frobenius type behavior for the immigrated mass mean semigroup, alongside the existence of second moments. These findings not only advance the understanding of stochastic models with immigration but also complement classical branching theorems concerning Galton–Watson processes and CBI-processes, cf. [3, 4].

Keywords: branching Markov process; superprocess; immigration; distributional stability.

Bibliography

[1] Li, Z. (2022). *Measure-Valued Branching Markov Processes*. Springer Berlin, Heidelberg. doi:10.1007/978-3-662-66910-5.

[2] Horton, E., and Kyprianou, A.E. (2023). *Stochastic Neutron Transport And Non-Local Branching Markov Processes*. Birkhauser. doi:10.1007/978-3-031-39546-8.

[3] Yang, Y.S. (1972). On Branching Processes Allowing Immigration. *Journal of Applied Probability*, 9(1):24–31. doi:10.2307/3212633.

[4] Pinsky, M.A. (1972). Limit theorems for continuous state branching processes with immigration. *Bulletin of the American Mathematical Society*, 78(2):242–244. doi:10.1090/S0002-9904-1972-12938-0.

The connection scheduling problem

19th June
13:00–13:20
Contributed Session 2

Laura Davila-Pena[1,2], Peter Borm[3], Ignacio García-Jurado[4] and Jop Schouten[5]

[1]Department of Analytics, Operations and Systems, Kent Business School, University of Kent, UK; [2]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Spain; [3]Department of Econometrics & Operations Research, Tilburg University, the Netherlands; [4]CITMAga, Department of Mathematics, Universidade da Coruña, Spain; [5]Amsterdam School of Economics, University of Amsterdam, the Netherlands

Laura Davila-Pena is a Postdoctoral Research Associate at Kent Business School, University of Kent. Her research activity involves various operations research problems with applications in areas such as healthcare or logistics, as well as cooperative game theory. She got her PhD in Statistics and Operations Research from the Universities of Santiago de Compostela, A Coruña and Vigo. Before moving to the UK, she worked as a lecturer at University of Santiago de Compostela, from which she also obtained a MSc in Statistical Techniques and a BSc in Mathematics.

Several real-life scenarios, such as urban water supply infrastructure, require efficiently connecting agents to a central source, minimizing construction costs [1]. The conventional approach typically involves utilizing the minimum cost spanning tree framework. However, in critical facilities like hospitals, where continuous service provision is essential, the ordering in which agents are connected must be considered, resulting in a sequencing problem.

This work analyzes so-called connection scheduling problems (CSPs), a type of interactive operations research problem that combines elements from minimum cost spanning tree problems and sequencing problems [2]. Given a graph, our objective is twofold: firstly, to establish an optimal connection order among agents to minimize the overall cost of connecting them to a source, and secondly, to develop a cost allocation strategy for this optimal order among the involved agents. Our investigation specifically focuses on CSPs with treelike precedence relations, for which we propose a recursive solution method integrated with an allocation approach.

Keywords: connection scheduling problems; sequencing problems; precedence relations; cost allocation.

Bibliography

[1] Bergantiños, G., Gómez-Rúa, M., Llorca, N., Pulido, M., and Sánchez-Soriano, J. (2014). A new rule for source connection problems. *European Journal of Operational Research*, 234(3):780–788. doi:10.1016/j.ejor.2013.09.047.

[2] Davila-Pena, L., Borm, P., García-Jurado, I., and Schouten, J. (2023). An allocation rule for graph machine scheduling problems. (CentER Discussion Paper; Vol. 2023-009). CentER, Center for Economic Research. https://research.tilburguniversity.edu/en/publications/an-allocation-rule-for-graph-machine-scheduling-problems.

Conditional expectations given the sum of independent heavy-tailed random variables

19th June
13:30–13:50
Contributed
Session 3

Michel Denuit[1], Patricia Ortega-Jiménez[1] and Christian Y. Robert[2]

[1]Institute of Statistics, Biostatistics and Actuarial Science - ISBA (LIDAM), UCLouvain, Belgium;
[2]Laboratory of Actuarial and Financial Science - LSAF, Université Lyon 1, France.

Patricia Ortega-Jiménez is a postdoctoral researcher in UCLouvain, Belgium. She obtained her PhD in mathematics at the statistics department of the University of Cádiz in 2022 under the supervision of Miguel A. Sordo and Alfonso Suárez-Llorens. Her main interests are stochastic orders, dependence modelling and its applications to risk-sharing schemes.

Given independent random variables X and Y, the study of the monotonicity of the conditional expectation of X given the sum S, defined by $m_X(s) = \mathrm{E}[X|S = s]$, plays an important role in various contexts, such as signal processing or risk sharing schemes. For instance, let X and Y model the insurance losses of two economic agents who decide to form a pool to share the total loss $S = X + Y$. Then, as suggested by [1], the conditional mean risk allocation rule determine that $m_X(\cdot)$ and $m_Y(\cdot)$ allocate the total loss among the risk holders. Under this context, non-decreasingness of $m_X(\cdot)$ is referred to as the no-sabotage condition.

Stochastic monotonicity of X and Y given the value of their sum $S = X + Y$ has been linked to log-concave densities since [2]. However, the log-concavity assumption is not realistic in some applications because it excludes heavy-tailed distributions. In this work, considering random variables with regularly varying densities, we delve on how heavy tails affect the monotonicity of $m_X(\cdot)$.

Firstly, we identify situations where a non-monotonic behavior appears according to the difference in the tail-heaviness of X and Y. In addition, we study the asymptotic behavior of $m_X(s)$ as the value s of the sum gets large, considering the different situations that can be encountered depending on the tail indices. The analysis is then extended to zero-augmented probability distributions, commonly encountered in applications to insurance and to sums of more than two random variables. Many numerical examples illustrate the results.

Keywords: regular variation; risk sharing; stochastic monotonicity.

Bibliography

[1] Denuit, M., and Dhaene, J. (2012). Convex order and comonotonic conditional mean risk sharing. *Insurance: Mathematics and Economics*, 51:265–270. doi:10.1016/j.insmatheco.2012.04.005.

[2] Efron, B. (1965). Increasing properties of Polya frequency function. *The Annals of Mathematical Statistics*, 36:272–279. doi:10.1214/aoms/1177700288.

Impact of domain reduction techniques in polynomial optimization: a computational study

19th June
13:50–14:10
Contributed
Session 3

Ignacio Gómez-Casares[1,2], Brais González-Rodríguez[1,2], Julio González-Díaz[1,2] and Pablo Rodríguez-Fernández[2]

[1]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Spain; [2]CITMAga (Galician Center for Mathematical Research and Technology), Spain

Ignacio Gómez-Casares is a third-year PhD student in the PhD in Statistics and Operations Research at the University of Santiago de Compostela (USC). His research activity is focused on the development of global non-linear optimization solvers. Before enrolling in his PhD, he did a MSc in Statistics and a BSc in Mathematics at the USC.

The design and implementation of global optimization algorithms for general mixed-integer non-linear programming (MINLP) problems is a very active field of research, and the number of available solvers has been steadily increasing over the past years. Recently, companies behind state-of-the-art mixed-integer linear programming (MILP) solvers such as Xpress [3] and Gurobi [2] have announced the release of new versions capable of solving general MINLP problems to certified global optimality. Convexifications and domain reduction techniques are probably the two most important elements behind the efficiency of the spatial branching required to handle the nonconvexities as can be seen, for instance, in [4], [5], and [1].

The focus of this work is precisely on the joint assessment of the individual impact of different aspects of domain reduction on the performance of a global optimization algorithm. We hope our analysis can guide the efforts in the development and implementation of this type of enhancements in present and future global optimization solvers for nonconvex problems. The two aspects covered are both related to domain reduction and present in any spatial branching algorithm: the selection of the branching variable and the selection of the branching point, both of them done at each and every node of the branch-and-bound tree. We will study the computational performance of several approaches to this two aspects of the algorithms.

Keywords: domain reduction; global optimization; polynomial optimization; reformulation-linearization technique; machine learning.

Bibliography

[1] Belotti, P., Lee, J., Liberti, L., Margot, F., and Wächter, A. (2009). Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4-5):597–634. doi:10.1080/10556780903087124.

[2] Gurobi Optimization (2023). Gurobi Optimizer Reference Manual. http://www.gurobi.com.

[3] FICO (2023). FICO Xpress optimization. https://community.fico.com/s/optimization.

[4] Ryoo, H.S., and Sahinidis, N.V. (1996). A branch-and-reduce approach to global optimization. *Journal of Global Optimization*, 8(2):107–38. doi:10.1007/bf00138689.

[5] Tawarmalani, M., and Sahinidis, N.V. (2004). Global optimization of mixed-integer nonlinear programs: A theoretical and computational study. *Mathematical Programming*, 99(3):563–591. doi:10.1007/s10107-003-0467-6.

Generative invariance: causal extrapolation without exogeneity

19th June
14:10–14:30
Contributed
Session 3

Carlos García Meixide[1] and David Ríos Insua[1]

[1]Instituto de Ciencias Matemáticas, Madrid, Spain

Carlos García Meixide is a first-year PhD student at Instituto de Ciencias Matemáticas in Madrid. His research activity is focused on causal inference, and the development of statistical methodology for medical data-driven decision making. Before enrolling his PhD, he worked at Hoffmann-La Roche as Biostatistician and did a MSc in Statistics at ETH Zürich and a BSc in Mathematics and Physics at Universidade de Santiago de Compostela.

The challenge of accurately predicting outcomes in situations where the distribution of data changes from training to test phases is a pressing issue in the realm of statistics and machine learning, a phenomenon known as distribution shift [2]. This problem is critical in fields such as healthcare, economics, and social sciences, where models trained on historical data must be applied to future or unseen conditions. The state-of-the-art methods, before the present work, largely depended on the availability of proxies that, being correlated with the explanatory variable but not directly with the outcome's error term, allow for the estimation of causal effects by circumventing the confounding factors [1].

We present a new estimator for predicting outcomes in different distributional settings under hidden confounding without relying on instruments or exogenous variables. The population definition of our estimator identifies causal parameters, whose empirical version is plugged into a generative model capable of replicating the conditional law within a test environment. We check that the probabilistic affinity between our proposal and test distributions is invariant across interventions. This work enhances the current statistical comprehension of causality by demonstrating that predictions in a test environment can be made without the need for exogenous variables and without specific assumptions regarding the strength of perturbations or the overlap of distributions. This context is crucial for statisticians and researchers looking to develop models that are not only accurate in controlled conditions but also adaptable to real-world heterogeneity.

Keywords: generalization; invariance; probabilistic predictions; structural causal models.

Bibliography

[1] Angrist, J. D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455. doi:10.1080/01621459.1996.10476902.

[2] Rothenhäusler, D., and Bühlmann, P. (2023). Distributionally robust and generalizable inference. *Statistical Science*, 38(4):527–542. doi:10.1214/23-STS902.

Sequencing games with batch-ordered jobs

19th June
16:00–16:20
Contributed
Session 4

Iago Núñez Lugilde[1], Arantza Estévez Fernández[2] and Estela Sánchez Rodríguez[1,3]

[1]Department of Statistics and Operations Research. SiDOR. University of Vigo, Spain. [3]CITMAga;
[2]Department of Operations Analytics. Vrije Universiteit Amsterdam, The Netherlands

Iago Núñez Lugilde has defended his PhD thesis, entitled "Computation and comparison of division rules to adjudicate conflicting claims", on 9 January 2024. His research is focused on game theory. Before enrolling his PhD, he obtained the degree in Mathematics in the University of Santiago de Compostela in 2018 and the inter university master's degree in Statistical Techniques and Operations Research at the universities of A Coruña, Santiago de Compostela and Vigo in 2019. He currently holds a one-year postdoctoral contract funded by the Xunta de Galicia.

The problem of sequencing a set of jobs waiting to be processed by a single machine at a minimum cost is a well-known topic in operations research. The problem of how to fairly distribute the total cost among the jobs' owners is modelled using cooperative game theory. [2] started this line of research by assuming the existence of a predetermined order between jobs. In many situations, there is no (clear) initial order because the arrival pattern can be stochastic or in batches. One-machine sequencing situations in which no initial order is specified are studied in [4] and [1]. Subsequently, [3] assume that, given an initial order, jobs arrive in batches.

Real-world scenarios sometimes introduce uncertainty in sequencing situations when jobs arrive in batches and the order of jobs within a batch is unknown. In this talk, we propose two new approaches to address such situations and discuss their practical applications. We define two cooperative cost games where the worst-case scenario for each coalition is considered. First, we assume that jobs in a coalition are placed last in their batch. Second, we assume that, once jobs are placed last in their batch, they can swap positions if they belong to a connected coalition and the change reduces the cost. Then, we define and characterize rules to distribute the total cost which provide core elements of the corresponding games. It is noteworthy that our model is a generalization of two classic situations: if there is one agent in each batch, it corresponds to the sequencing situation with an initial order; and if there is only one batch, it corresponds to the sequencing situation without an initial order.

Keywords: batchs; cooperative games; rules; sequencing.

Bibliography

[1] Chun, Y. (2006). A pessimistic approach to the queueing problem. *Mathematical Social Sciences*, 51(2):171–181. doi:10.1016/j.mathsocsci.2005.08.002.

[2] Curiel, I., Pederzoli, G., and Tijs, S. (1989). Sequencing games. *European Journal of Operations Research*, 40(3):344–351. doi:10.1016/0377-2217(89)90427-X.

[3] Gerichhausen, M., and Hamers, H. (2009). Partitioning sequencing situations and games. *European Journal of Operational Research*, 196(1):207–216. doi:10.1016/j.ejor.2008.03.003.

[4] Klijn, F., and Sánchez Rodríguez, E. (2006). Sequencing games without initial order. *Mathematical Methods of Operations Research*, 63(1):53–62. doi:10.1007/s00186-005-0012-x.

Optimal subdata selection for prediction of streaming data in linear models

19th June
16:20–16:30
Contributed
Session 4

Álvaro Cia-Mina[1,2] and Jesús López-Fidalgo[1,2]

[1]Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Pamplona, Spain.

[2]Tecnun Escuela de Ingeniería, Universidad de Navarra, San Sebastián, Spain.

Álvaro Cia-Mina is a third-year PhD student in the PhD in Statistics for Data Science at University of Navarra. His research activity is focused on prediction models with random covariates, using optimal subdata selection and active learning methods to reduce prediction error. Before enrolling his PhD, he did a MSc in Big Data Science at University of Navarra and a BSc in Mathematics and Physics at University of Valladolid.

In response to big data challenges, the field of optimal design of experiments has seen significant advancements in active learning and subsampling techniques, aiming to refine computational efficiency and labeling processes. Subsampling is widely used to downsize the data volume and allows computing estimators efficiently in regression models. While most of the existing methods focus on reducing the estimation error of the parameters, usually the practical goal of statistical models is to minimize the prediction error. Key contributions [1] and [2] also show the high influence of the distribution of explanatory variables—the "Random-X" versus "Fixed-X paradigms"—in both the estimation and subsampling strategies.

We propose a new subdata selection method for linear models based on the distribution of the covariates. The case of a big sample where the labels of the response variable are expensive to obtain is considered. The criterion provided is based on the Random-X prediction error. Theoretical results are provided to justify the criterion as well as an interpretation from usual linear optimality criteria. A sequential selection algorithm is proposed to extend the applicability of the method to streaming data. As expected by the theory it shows a reduction in the prediction mean squared error compared to other existing methods. The performance of the new approach is illustrated with simulations.

Keywords: subsampling; active learning; streaming data; optimal design of experiments.

Bibliography

[1] Pronzato, L., and Wang, H.Y. (2021). Sequential online subsampling for thinning experimental designs. *Journal of Statistical Planning and Inference*, 212:169–193. doi:10.1016/j.jspi.2020.08.001.

[2] Rosset, S., and Tibshirani, R.J. (2020). From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation. *Journal of the American Statistical Association*, 529:138–151. doi:10.1080/01621459.2018.1424632.

The Fesenthal power index for simple games with a priori unions

19th June
16:40–17:00
Contributed
Session 4

Silvia Lorenzo-Freire[1], Jose María Alonso-Meijide[2] and Alicia Mascareñas[1]

[1]Department of Mathematics, Faculty of Computer Science and CITIC, University of A Coruña, Spain; [2]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela and CITMAga, Spain

Alicia Mascareñas Pazos is a first-year PhD student in the PhD in Operations Research and Game theory at University of A Coruña. Her research activity is focused on cooperative game theory. Before enrolling her PhD, she did a MSc in Mathematics at University of Santiago de Compostela and a BSc in Mathematics at University of Santiago de Compostela.

Within the scope of game theory, simple games are tipically used to model situations of collective decision-making. The interest in these games stems from the desire to understand the power or influence that a player has on the final result. Power indices appear as a quantitative gauge of the ability of the different players to turn a losing coalition into a winning one. Literature on the subject focuses on proposing new power indexes and characterizing them through different axioms. Deegan and Packel presented and characterized a power index based on the formation of minimal winning coalitions [2]. Fesenthal introduced a new index, focusing on a subset of the latter, composed of the winning coalitions of smaller size [3].

In decision-making situations, certain pre-existing groupings among participants naturally emerge. These groups are motivated by a certain affinity on some topic and are reflected in a partition of the set of players. Player's power is influenced by these coalitions, and adequate power indices must be proposed for these new situations. The Deegan-Packel index has been extended to simple games with a priori unions, in the sense that it corresponds to the original index when each group is formed by only one player or there is only one group that includes all players [1].

In this context, we generalize the definition of the Fesenthal index taking into account the coalition structure, and obtain and axiomatic characterization for this new power index with a priori unions.

Keywords: decision and negotiation; power indexes; simple game.

Bibliography

[1] Alonso-Meijide, J.M., Casas-Méndez, B, Fiestras-Janeiro, M.G., and Holler, M.J. (2011). The Deegan–Packel index for simple games with a priori unions. *Quality & Quantity*, 45:425–439. doi:10.1007/s11135-009-9306-z.

[2] Deegan, J., Packel, E. (1978). A new index of power for simple n-person games. *International Journal of Game Theory*, 7:113–123. doi:10.1007/BF01753239.

[3] Felsenthal, D.S. (2016). A well-behaved index of a priori P-Power for simple n-person games. *Homo Oeconomicus*, 33:367–381. doi:10.1007/s41412-016-0031-2.

Augmented designs for discrimination between constant absolute vs relative error

19th June
17:00–17:20
Contributed
Session 4

Carlos de la Calle-Arroyo[1], Samantha Leorato[2], Licesio J. Rodríguez-Aragón[3] and Chiara Tommasi[2]

[1]Departamento de Estadística, Investigación Operativa y Didáctica de las Matemáticas, Universidad de Oviedo, Spain; [2]Facoltà di Economia e Scienze Politiche, Università degli Studi di Milano, Italy; [3]Escuela de Ingeniería Industrial y Aeroespacial de Toledo, Universidad de Castilla-La Mancha, Spain

Carlos de la Calle-Arroyo is Assistant Professor of Stastistics in the University of Oviedo. His research activity is focused on methodological and software solutions in optimal experimental design. Before starting his current role this year, he finished his PhD in University of Castilla-La Mancha in 2022 and did a postdoc in the University of Navarra.

In the realm of experimental sciences, particularly in disciplines like chemistry, the presence of measurement errors can manifest in either homoscedastic or heteroscedastic forms. The acquisition of data necessitates a deliberate effort to discern the appropriate error-variance structure, as a misjudged model could yield erroneous conclusions. One design criterion aimed at achieving this objective, based on the Kulback-Leibler divergence, is KL-optimality[2], which aims at maximizing the power of the hypothesis test between two probability distributions. [3] studied designs to detect heteroscedasticity in the case that the variance structure is nested. Designs optimized for KL-criterion, however, often exhibit inefficiencies when utilized for other inferential objectives, such as precise estimation. In such scenarios, the incorporation of additional experimental points becomes advantageous.

This study is focused on enhancing design methodologies by introducing supplementary support points, with the explicit aim of ensuring a minimum level of KL-efficiency for optimal selection among various variance specifications. The methodology is based on the extension of the D-augmented design methodology[1] for the KL-optimality criterion. Furthermore, this strategy proves beneficial in modifying existing designs, such as D-optimal designs, to address the challenge of accurate error-variance specification. These two approaches are compared to the 'gold standard' in multiobjective designs, which is the compound criteria, which in this problem require the implementation of more complex algorithmical techiques, such as metaheuristics.

Keywords: D-optimality; design augmentation; hypothesis testing; KL-optimality.

Bibliography

[1] de la Calle-Arroyo, C., Amo-Salas, M., López-Fidalgo, J., Rodríguez-Aragón, L.J., and Wong, W.K. (2023). A methodology to D-augment experimental designs. *Chemometrics and Intelligent Laboratory Systems*, 237:104822. doi:10.1016/j.chemolab.2023.104822.

[2] Fidalgo, J., Tommasi, C., and Trandafir, P.C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B*, 69(2):231–242. doi:10.1111/j.1467-9868.2007.00586.x.

[3] Lanteri, A., Leorato, S., López-Fidalgo, J., and Tommasi, C. (2023). Designing to detect heteroscedasticity in a regression model. *Journal of the Royal Statistical Society: Series B*, 85(2):315–326. doi:10.1093/jrsssb/qkad004.

Balancing palatability, sustainability and cost in the menu planning problem using a Wierzbicki achievement function

19th June
17:40–18:00
Contributed
Session 5

Francisco Martos-Barrachina[1], Laura Delgado-Antequera[2]and Mónica Hernández[2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Málaga, Spain; [2]Departamento de Economía Aplicada (Matemáticas), Universidad de Málaga, Spain

Francisco Martos-Barrachina is a recently graduated PhD in Economics and an assistant lecturer in the Area of Statistics and Operations Research. His research activity is focused on modelling and solving a holistic instance of the Menu Planning Problem. Before enrolling his PhD, he worked for a bank as a Data Analyst, and did a MSc in Health Economics and a BSc in Actuarial Science in the University of Málaga.

The current Menu Planning Problem is approached within the SHARP framework, which prioritizes sustainability, healthiness, affordability, reliability, and palatability. This holistic approach within the SHARP framework aims to address the multifaceted challenges of menu planning, promoting a well-rounded and responsible approach to food choices. In addressing sustainability, menus are designed to minimize environmental impact by incorporating locally sourced, seasonal, and eco-friendly ingredients [1]. The focus on health involves creating balanced and nutritious meals that meet dietary guidelines. Affordability is considered by optimizing cost while designing the menu plan. Reliability ensures consistent access to a diverse range of food options, taking into account factors such as supply chain resilience and availability. Lastly, palatability remains a key element, ensuring that meals are enjoyable and culturally appropriate to encourage adherence to healthy and sustainable eating patterns.

We have worked towards incorporating all five dimensions into the problem. In our work, we have incorporated a novel *Similarity Function*, in order to enhance palatability. We developed a Multi Objective Combinatorial Optimization model that has 3 objectives, sustainability, palatability and cost and a set of constraints that consider nutrition, Mediterranean diet standards [2], repetition, balance and schedule. In order to consider all the objective simultaneously we use an Extended Wierzbicki Achievement Function that allows us to explore different regions of the Pareto Front [3].

Keywords: combinatorics; MPP; multi objective optimization; SHARP diets.

Bibliography

[1] Benvenuti, L., De Santis, A., Di Sero, A., and Franco, N. (2019). An optimal plan for food consumption with minimal environmental impact: The case of school lunch menus. *Journal of Cleaner Production*, 129:704–713. doi:10.1016/j.jclepro.2019.117645.

[2] Hernández, M., Gómez, T., Delgado-Antequera, L., and Caballero, R. (2019). Using multiobjective optimization models to establish healthy diets in Spain following Mediterranean standards. *Operational Research: An International Journal*, 21(3):1927–1961. doi:10.1007/s12351-019-00499-9.

[3] Martos-Barrachina, F., Delgado-Antequera, L., and Hernández, M. (2024). A novel cost-palatability bi-objective approach to the menu planning problem with an innovative similarity metric using a path relinking algorithm. *Journal of the Operational Research Society*, 1–13. doi:10.1080/01605682.2024.2326188.

Bayesian longitudinal beta regression for competing risk models

19th June
18:00–18:20
Contributed
Session 5

Jesús Gutiérrez-Botella[1], Carmen Armero[2], Thomas Kneib[3], María Pata[4] and Javier García-Seara[5]

[1]Biostatech, Advice, Training and Innovation in Biostatistics SL; GRID-BDS, University of Santiago de Compostela, Spain; [2]Department of Statistics and OR, Universitat de València, Spain; [3]Georg-August-Universität Göttingen, Germany; [4]Biostatech, Advice, Training and Innovation in Biostatistics SL; [5]Arrhythmia Unit. University Hospital of Santiago de Compostela, Spain

Jesús Gutiérrez-Botella is a third-year PhD student in the PhD in Statistics and Operational Research at the University of Santiago de Compostela. His research is focused on Bayesian joint modeling of survival and longitudinal data, specifically on competing risks and multistate models. He has studied Biotechnology and Biostatistics, and he is now doing an Industrial PhD at Biostatech, Advice, Training & Innovation in Biostatistics SL.

Joint modeling of longitudinal and survival data (JM-LS) deals with statistical models that allow to combine longitudinal and survival information [2]. These models can be expressed as a joint probability distribution $p(y, s, \boldsymbol{\theta}, \boldsymbol{b})$, where y and s represent the longitudinal and the survival model, respectively, $\boldsymbol{\theta}$ is a parametric vector, and $\boldsymbol{b}$ a set of random effects. JM-LS are complex and flexible models that allow introducing internal temporal covariates in the survival processes as well as providing inferential frameworks for the inclusion of non-ignorable dropout mechanisms through survival tools.

We propose a joint modelling of a competing risks survival model fed by a time covariate with unit interval support that we model using a mixed linear Beta regression model [3]. The cause-specific hazard function for each relevant event is defined in terms of the Cox proportional hazards structure with Weibull baseline hazards, baseline covariates and a final term including longitudinal information. We discuss different proposals for the latter term, such as the mean of the time-dependent covariate or the specific random effects associated to individuals. This modelling is used in a Heart Failure study where patients underwent Cardiac Resynchronization Therapy [1].

Keywords: posterior predictive distribution; shared-parameter models; time to death.

Bibliography

[1] Álvarez-Álvarez, B., García-Seara, J., Martínez-Sande, J. L., Rodríguez-Mañero, M., Fernández López, X. A., González-Melchor, L., Iglesias-Alvarez, D., Gude, F., Díaz-Louzao, C. and González-Juanatey, J. R. (2021). Long-term cardiac reverse remodeling after cardiac resynchronization therapy. *Journal of Arrhythmia*, 37(3):653–659. doi:10.1002/joa3.12527.

[2] Armero, C. (2021). Bayesian Joint Models for Longitudinal and Survival Data. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08129.

[3] Cribari-Neto, F. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2):1–24. doi:10.18637/jss.v034.i02.

Data-driven models to support decisions in critical care

19th June
18:20–18:40
Contributed
Session 5

Daniel Garcia-Vicuña[1], Ana María Anaya-Arenas[2], Janosch Ortmann[2] and Angel Ruiz[3]

[1]Department of Statistic, IT and Mathematics, Public University of Navarre, Spain;[2]Department of Analysis, Operations, and Information Technology, Université du Québec à Montréal, Canada;
[3]Department of Operations and Decision Systems, Université Laval, Canada

Daniel Garcia-Vicuña has a Ph.D. in Industrial Engineering at the Public University of Navarre, Spain. His research interests are focused on the field of simulation modeling of complex real-world problems. The main projects in which he has participated are framed within the scope of societal challenges and the scientific response to the health emergency situation posed by COVID-19. Currently, he holds a position as an assistant professor at the Public University of Navarre and is a member of the q-UPHS (Quantitative Methods for Uplifting the Performance of Health Services) research group at the same university.

The dilemma of the last bed [1, 2] underscores the shortage of intensive care unit (ICU) beds and the tough decisions hospitals must make when allocating limited resources. Faced with high demand, healthcare systems are pressured to prioritize patients, often resulting in canceled surgeries, early discharges, or denied admissions. This presents logistical and operational challenges to ensure adequate medical care. Efficient bed management becomes a strategic priority to optimize patient treatment and maximize available resources during healthcare crises.

The main purpose of this work is to develop new methodologies using real patient data to assist healthcare professionals in decision-making regarding resource management and capacity planning in healthcare services. Specifically, the methodology focuses on predicting patients' lengths of stay in the ICU, using a large number of variables such as patient characteristics, medical history, and variables related to the patient's health status. These parameters define the patient's status at a given time, and using distance measures, the most similar state of each historical patient is sought to determine remaining time intervals in the ICU. Based on these intervals and the degree of similarity with each patient, a distribution function of the patient's remaining length of stay is obtained. The results of the methodology are illustrated with real data, and simulations demonstrate that these techniques can be effective in predicting short-term patient demand and facilitating decision-making in resource management in healthcare services.

Keywords: decision-making; ICU; prediction; resource planning; simulation.

Bibliography

[1] Azcarate, C., Esparza, L., and Mallor, M. (2020). The problem of the last bed: Contextualization and a new simulation framework for analyzing physician decisions. *Omega*, 96:102120. doi:10.1016/j.omega.2019.102120.

[2] Teres, D. (1993). Civilian Triage in the Intensive Care Unit: the Ritual of the Last Bed. *Critical Care Medicine*, 21(4):598–606. doi:10.1097/00003246-199304000-00022.

A decision analysis model for colorectal cancer screening

19th June
18:40–19:00
Contributed
Session 5

Daniel Corrales[1] and David Ríos Insua[1]

[1]Institute of Mathematical Sciences, ICMAT-CSIC, Spain

Daniel Corrales is a first-year PhD student at the Institute of Mathematical Sciences (ICMAT-CSIC). His research is focused on the development of Bayesian Networks and Influence Diagrams for medical applications. Before joining ICMAT, he carried out a research internship at the Basque Center for Applied Mathematics (BCAM), and did a MSc in Machine Learning for Health at Universidad Carlos III de Madrid and a BSc in Mathematical Engineering at Universidad Complutense de Madrid.

Colorectal cancer (CRC) is the third most common type of cancer worldwide, making up for about 10% of all cases and being accountable for around 12% of all deaths due to cancer. Despite this, as an example, only about 14% of susceptible European Union citizens participate in screening programmes. Hence, there is an urgent need for accurate, non-invasive, cost-effective screening tests based on novel technologies and raise further awareness on the disease and its detection. Moreover, personalized screening approaches are required to consider socioeconomic variables as well as environmental stressors that can lead to different onsets of the disease.

This work outlines one such approach within the ONCOSCREEN Horizon Europe project. First, we develop a Bayesian network model to facilitate CRC predictions drawing on expert judgement and a large database from an observational study. This network is used to map CRC risks depending on numerous factors. We then embed such network in an influence diagram model aimed at advising about personalised screening strategies depending on patient information and cost-effectiveness of the methods. We also discuss incentives in relation to promoting the uptake of screening in the population.

Keywords: Bayesian inference; decision analysis; healthcare; influence diagrams; value of information.

Bibliography

[1] Ankan, A. and Panda, A. (2015). pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. `https://pgmpy.org/`.

[2] Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. `https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/`.

[3] Ordovas, J.M., Ríos-Insua, D., Santos-Lozano, A., Lucía, A., Torres, A., Kosgodagan, A., and Camacho, J.M. (2023). A Bayesian network model for predicting cardiovascular risk. *Computer Methods and Programs in Biomedicine*, 231:107405. `doi:10.1016/j.cmpb.2023.107405`.

[4] Spiegelhalter, D., Dawid, P., Lauritzen, S., and Cowell, R. (1993). Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–247. `doi:10.1214/ss/1177010888`.

Logistic regression and prediction methods with missing data

19th June
19:00–19:20
Contributed
Session 5

Susana Rafaela Guimarães Martins[1], María del Carmen Iglesias-Pérez[2] and Jacobo de Uña-Álvarez[2]

[1]Escola Superior de Desporto e Lazer, Instituto Politécnico de Viana do Castelo, Portugal;
[2]Department of Statistics and OR, Universidade de Vigo, Spain

Susana Rafaela Martins is a sixth-year PhD student in Statistics and Operations Research at the University of Vigo. Her research is focused on logistic regression and prediction methods with missing data. She is master in Systems Statistics as well as a Bachelor's degree in Educational Mathematics. Since 2011, she has been teaching Statistics and Linear Algebra at the Polytechnic Institute of Viana do Castelo, Portugal.

In some studies the goal is to relate a main outcome of interest to a number of variables, the so-called regression setup. Missing information complicates the analysis. Our research focuses on logistic regression with missing data. In particular, we are interested in the performance of regression coefficient estimates as well as in measuring the predictive capacity of the logistic regression model, where the Area Under the ROC Curve (AUC) plays an outstanding role.

In the first phase of our work logistic regression was considered when both the response variable and the predictive variable may be missing. Several existing approaches were reviewed, including complete case analysis, inverse probability weighting, multiple imputation and maximum likelihood. The methods were compared in a simulation study and the maximum likelihood was the one that presents the best results, followed by multiple imputation [3]. Currently we focus on studying the predictive capacity of the logistic regression model. The issue of estimating the AUC in the presence of missing data has been investigated through complete case analysis, inverse probability weighting, multiple imputation. Traditionally, the apparent AUC overestimates the true AUC, so we considered several approaches to correct this overestimation: Split-Sample, K-fold and Leave-one-out, adapted to missing data [1, 2]. We also conducted a simulation study to evaluate performance of the correction methods in the presence of missing data.

Keywords: apparent AUC; missing data; Monte Carlo simulation; ROC curve.

Bibliography

[1] Iparragirre, A., Barrio, I., Rodríguez-Álvarez, M.X. (2019). On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. *SORT-Statistics and Operations Research Transactions*, 43(1):145–162. `https://raco.cat/index.php/SORT/article/view/356185`.

[2] Li, P., Taylor, J.M., Spratt, D.E., Karnes, R.J., and Schipper, M.J. (2021). Evaluation of predictive model performance of an existing model in the presence of missing data. *Statistics in Medicine*, 40(15):3477–3498. `doi:10.1002/sim.8978`.

[3] Martins, S.R., Uña-Álvarez, J., and Iglesias-Pérez, M.C. (2023). Logistic regression with missing responses and predictors: a review of existing approaches and a case study. `arXiv:2302.03435`.

Thursday 20

Time indexes and other discretizations: are they worth it?

20th June
09:15–10:15
Plenary
Contributed
Session 2

Federico Perea

Department of Applied Mathematics II and Institute of Mathematics, University of Sevilla, Spain

Federico Perea is an associate professor at the University of Seville since 2020, where he earned his degree in Mathematics in 2001 and received his Ph.D. in 2007. Afterwards, his academic career continued at the University of Seville, University of Zaragoza and, between 2010 and 2020, at the Department of Applied Statistics, Operation Research and Quality at the Technical University of Valencia. He has also held a research position at THALES company (The Netherlands) within the Marie Curie European Programme for 18 months between 2005-2006. His main research lines are the analysis of optimization problems both in academia and in industry, within topics that cover transportation, scheduling, location, game theory, and aerospace.

Arguably, Mixed Integer Linear Programming (MILP) is one of the most commonly used techniques to solve optimization problems subject to constraints. MILPs give much flexibility and allow to model many realistic characteristics. And maybe because of this *integer popularity*, many researchers and practitioners tend to discretize continuous aspects of reality such as time or space (consciously or even unconsciously). These aspects are often parts of the so called *sets* of the arising mathematical programming models, which are the elements that do participate in the optimization problem. Therefore, and in order to avoid having to deal with *infinite programming*, it is usual to discretize time, space, and other continuous characteristics. This leads to *natural* formulations which are easy to understand, explain, modify, etc. However, is this the only option?

In this talk, we will review some related literature. In [1], an Unrelated Parallel Machine scheduling problem with additional resources during the processing of the jobs is treated. The authors compare MILP formulations with and without a time index. In a more recent paper, [2] also propose mathematical programming models with and without time indexes for another machine scheduling problem with additional resources during setups. As for space discretization, [3] propose a mathematical programming formulation without space indexes, which improves a previous formulation which does rely on such discretization.

Keywords: mixed integer linear programming; operations research; optimization.

Bibliography

[1] Fanjul, L., Perea, F., and Ruiz, R. (2017) Models and matheuristics for scheduling problems with additional resources. *European Journal of Operational Research*, 260(2):482–493. doi:10.1016/j.ejor.2017.01.002.

[2] Yepes-Borrero, J. C., Perea, F., Villa, F., and Vallada, E. (2023) Flowshop with additional resources during setups: mathematical models and a GRASP algorithm. *Computers and Operations Research*, 154:106192. doi:10.1016/j.cor.2023.106192.

[3] Linares, L., Vazquez, R., Perea, F., and Galán-Vioque, J. (2024) A Mixed Integer Linear Programming Model for Resolution Of the Antenna-Satellite Scheduling Problem. *IEEE Transactions on Aerospace and Electronic Systems*, 60(1):463–473, doi:10.1109/TAES.2023.3326422.

On uniqueness of the set of k-means

20th June
10:20–10:40
Contributed
Session 6

Javier Cárcamo[1], Antonio Cuevas[2] and Luis-Alberto Rodríguez[3]

[1]Department of Mathematics, University of Basque Country, Spain; [2]Department of Mathematics, Autónoma de Madrid University, Spain, [3]Institut für Mathematische Stochastik, Georg August University of Göttingen, Germany

Luis Alberto Rodríguez Ramírez is a postdoctoral researcher at Institut für Mathematische Stochastik, Georg August University of Göttingen, Germany. His main interests are empirical process theory and functional data analysis. He received his BSc from Autonomous University of Madrid in 2017 in mathematics. He earned his MSc in mathematics and applications from Autonomous University of Madrid one year later. In 2018, he enrolled in the doctoral program in mathematics at Autonomous University of Madrid under the supervision of Javier Cárcamo and Antonio Cuevas.

The k-means procedure is one of the most commonly used techniques for finding a given number of groups in a data set. It plays also a central role in localization problems in operations research. The notion of k-means is a natural, almost elementary, idea with a clear interpretation and a great number of relevant applications. However, despite its simplicity, the underlying methodology still has some extraordinarily complex challenges associated with it (such as the choice of the parameter k) and many intriguing theoretical and computational aspects. First, the effective calculation of the sample k-means is a formidable computational (NP-hard) problem: algorithms have to cope with a non-convex optimization problem in a possibly high-dimensional (or even infinite-dimensional) space. Moreover, the usual algorithms do not guarantee to reach a global optimum, but rather a local one; see, for example, [1] for an overview of various relevant clustering algorithms. Choosing of a good value for k still receives considerable attention in this field and constitutes an area of active research.

In this talk, conditions on the uniqueness (and non-uniqueness) of the sets of k-means are given in the context of cluster analysis. This problem is associated with the choice of k: depending on the underlying distribution, some values of this parameter could lead to multiple sets of k-means, which hampers the interpretation of the results and/or the stability of the algorithms (see [2], [3] and references therein). We give a general assessment on consistency of the empirical k-means, adapted to the setting of non-uniqueness. We also provide a statistical characterization of k-means uniqueness.

Keywords: clustering; empirical process; Gromov-Hausdorff distance; k-means consistency; non-supervised classification.

Bibliography

[1] Morissette, L., and Chartier, S. (2013). The k-means clustering technique: general considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15–24. doi:10.20982/tqmp.09.1.p015.

[2] Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 9:135–140. doi:10.1214/aos/1176345339.

[3] Pollard, D. (1982). A central limit theorem of k-means clustering. *The Annals of Statistics*, 10:919–926. doi:10.1214/aop/1176993713.

Optimal allocation of helicopters to firefighting operations: model and heuristic algorithm

20th June
10:40–11:00
Contributed
Session 6

Marta Rodríguez Barreiro[1,2], María José Ginzo Villamayor[2,4], Fernando Pérez Porras[5], María Luisa Carpente Rodríguez[1] and Silvia María Lorenzo Freire[1,3]

[1]MODES Group, Department of Mathematics, University of A Coruña, Spain; [2]Galician Centre of Mathematical Research and Technology (CITMAga), Spain; [3]Centre for Information and Communications Technology Research (CITIC), Spain; [4]Modestya Group, Department of Statistics, Mathematical Analysis and Optimisation, University of Santiago de Compostela, Spain; [5]Department of Graphic and Geomatic Engineering, University of Córdoba, Spain

M. Rodríguez Barreiro is a PhD student at the University of A Coruña. Her main interests are the application of operations research (OR) methods for logistic optimization. She received her BSc from University of Santiago de Compostela in mathematics. She earned hers master's degree in statistics and operational research from the Polytechnic University of Catalonia. She worked as researcher in the field of operations research in different research centres such as the Catalonia Energy Research Institute (Barcelona). She is currently a researcher at the Galician Centre for Mathematical Research and Technology (CITMAga).

Wildfires are one of the biggest problems faced by humanity in this century. In recent years, they have become more severe and devastating, requiring more resources and personnel to fight them. Many operational research models in the literature focus on this problem. However, no work has been found to plan the work of firefighting helicopters in their entirety, from their rest base to the water loading and dropping points, taking into account the flight regulations and working methods of these aircraft.

This work presents a model developed to plan the work of helicopters during the extinguishing of a large wildfire. The developed model takes into account the main decisions to be made by fire managers: which helicopters will be involved in extinguishing the fire, at what time intervals they should work, and which water load points and rest bases they will use. The model also determines where helicopters should drop water, considering the efficiency associated with each zone at each time interval, which is set by the fire manager. It also considers the way in which helicopters operate, in elliptical circuits sharing loading and dropping points, as well as Spanish aviation legislation. This is modelled in a time-extended graph, which makes it possible to reflect the expected evolution of the wildfire obtained from a fire simulator. Given the complexity of the model, it takes a long time to solve it using a commercial solver, such as Gurobi. Therefore, since the model will be used in an emergency situation, a heuristic was implemented to reduce the time required to solve the model. The heuristic is based on the Simulated Annealing [1] technique and yields satisfactory results when tested on different data instances.

Keywords: helicopters; optimization; planning; simulated annealing; wildfire.

Bibliography

[1] Kirkpatrick, S., Gelatt Jr, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680. doi:10.1126/science.220.4598.671.

High density region estimation for manifold data

20th June
11:00–11:20
Contributed
Session 6

Diego Bolón[1,2], Rosa M. Crujeiras[1,2] and Alberto Rodríguez-Casal[1,2]

[1]CITMAga, Santiago de Compostela, Spain;
[2]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain.

Diego Bolón is a third-year PhD student in the PhD in Statistics at Universidade de Santiago de Compostela. His research activity is focused on developing inference procedures for data supported on Riemannian manifolds. Before enrolling his PhD, he did a MSc in Statistics and a BSc in Mathematics, both at Universidade de Santiago de Compostela.

High density regions (in the sequel, HDRs) are the sets where the density function of the data exceeds a given (and usually high) level. HDR estimation has been shown to have plenty of practical applications, like data clustering [3] or the analysis of seismic data [2]. Most of the existing contributions just focus on HDR estimation of densities supported on an Euclidean space, while HDR estimation in other domains has received very little attention in comparison. Up to our knowledge, [4] and [1] provide the only two proposals for HDR estimation for manifold data, which both rely on the plug-in approach. However, this perspective ignores the inner geometry of the problem: if one knows that the HDRs of the population of study fulfill some geometric property (e.g. some shape condition), there is no guarantee that the plug-in estimator will satisfy it too.

Trying to overcome these issues, a new HDR estimator for manifold data under smoothing conditions is introduced. Specifically, the new proposal can be viewed as a generalization of the euclidean HDRs estimation technique introduced by [5] to Riemannian manifolds. The consistency of the new proposal will be shown, and its convergence rate will be derived. Finally, the performance in practice of the new HDR estimator is illustrated with a real data example.

Keywords: non-parametric statistics; set estimation; high density regions; manifold data.

Bibliography

[1] Cholaquidis, A., Fraiman, R., and Moreno, L. (2022). Level set and density estimation on manifolds. *Journal of Multivariate Analysis*, 189:104925. doi:10.1016/j.jmva.2021.104925.

[2] Huo, X., and Lu, J.C. (2004). A network flow approach in finding maximum likelihood estimate of high concentration regions. *Computational Statistics & Data Analysis*, 46(1):33–56. doi:10.1016/S0167-9473(03)00134-8.

[3] Rinaldo, A., and Wasserman, L. (2010). Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722. doi:10.1214/10-AOS797.

[4] Saavedra-Nieves, P., and Crujeiras, R.M. (2021). Nonparametric estimation of directional highest density regions. *Advances in Data Analysis and Classification*, 16(3):761–796. doi:10.1007/s11634-021-00457-4.

[5] Walther, G. (1997). Granulometric smoothing. *The Annals of Statistics*, 25(6):2273–2299. doi:10.1214/aos/1030741072.

Integrating shape-restricted multivariate adaptive regression splines and random forest: methodology and applications

20th June
11:40–12:00
Contributed
Session 7

Víctor J. España[1], Juan Aparicio[1] and Xavier Barber[1]

[1]Center of Operations Research (CIO), Miguel Hernandez University of Elche

Víctor J. España is a third-year PhD student in Statistics, Optimization and Applied Mathematics at the Miguel Hernández University of Elche. His research is centered on developing shape-restricted adaptation of Machine Learning algorithms for applications in the efficiency analysis field. Before enrolling his PhD, he earned a degree in Business Statistics at the Miguel Hernández University of Elche, and he received his master's degree in Management and Analysis of Big Data from the Miguel de Cervantes European University.

In many research fields, understanding the relationship between a set of predictors and a response variable often necessitates the estimation of curves that adhere to specific properties. For instance, when estimating mortality rates, isotonic regression can be used, which involves a monotonically increasing relationship between older age groups and mortality risk. Similarly, the field of efficiency analysis can be reinterpreted as a shape-restricted regression problem, aiming to estimate a non-decreasing and concave function that envelopes the observed data points. In this context, Data Envelopment Analysis (DEA) is frequently employed in operations research to provide a non-parametric estimate of production frontiers. However, DEA is prone to overfitting, leading to overly optimistic and potentially inaccurate efficiency estimates.

In [1], an adaptation of the Multivariate Adaptive Regression Splines (MARS) algorithm is introduced to estimate production functions, specifically designed to mitigate overfitting issues. Our contribution expands upon the methodology presented in [1], with three primary objectives. Firstly, we propose a procedure that allows for the integration of variable interaction in the model-fitting process while preserving the required shape constraints of a production function. This enhancement enables the identification of non-additive relationships among the variable of the problem, thereby improving the model's predictive capacity. Secondly, we increase the technique's robustness by incorporating data and input variables randomization during the model construction process, drawing inspiration from the Random Forest methodology. Finally, under the Random Forest paradigm, we can identify the most relevant inputs in relation to the output production.

Keywords: DEA; efficiency analysis; Machine Learning.

Bibliography

[1] España, V. J., Aparicio, J., Barber, X., and Esteve, M. (2024). Estimating production functions through additive models based on regression splines. *European Journal of Operational Research*, 312(2):684–699. doi:10.1287/mnsc.30.9.1078.

Survival analysis modelling for high-dimensional data

20th June
12:00–12:20
Contributed
Session 7

Pilar González-Barquero[1], Rosa E. Lillo[2] and Álvaro Méndez-Civieta[3]

[1]uc3m-Santander Big Data Institute, University Carlos III of Madrid, Spain; [2]uc3m-Santander Big Data Institute, University Carlos III of Madrid, Spain; [3]Department of Biostatistics, Columbia University, New York, U.S.A.

Pilar González-Barquero is a PhD student at the University Carlos III of Madrid enrolled in the doctoral program in Statistics for Data Science, supervised by Rosa E. Lillo and Álvaro Méndez-Civieta. Her PhD thesis is centered in survival analysis in high-dimensional scenarios. She received her BSc in Mathematics from Autonomous University of Madrid in 2021 and earned her master's degree in Statistics for Data Science from the University Carlos III of Madrid in 2023. She was also working at BBVA as Data Analyst and at the University of Extremadura as Research Assistant.

Survival analysis has been proven to have many applications, particularly in the medical field. It plays a crucial role in understanding and predicting survival times of patients with a specific disease under varying conditions enabling medical professionals to make decisions and develop treatment strategies.

Nowadays, one of the most significant characteristics of the Big Data era we are in is the appearance of datasets with a high number of covariates. These high volume datasets are commonly referred to as high-dimensional data. Although this seems beneficial due to the great amount of information provided, the treatment of such huge volumes of data adds significant complexity and difficulty to the decision making process. In this context, traditional models are unfeasible and variable selection methods need to be applied.

This work examines the efficacy of Cox regression models [1] in the context of high-dimensional data ($p \gg n$) with a significant proportion of censored observations. In scenarios where dimensionality reduction is essential for improving model interpretability and predictive accuracy, two regularization techniques-Lasso [2] and Adaptive Lasso [3] penalties-are evaluated. This research involves the exploration of various weight calculation procedures for the Adaptive Lasso. These proposed weightings are derived from Principal Component Analysis, ridge regression, univariate Cox regressions, and the Random Survival Forest (RSF) algorithm. The main contribution of this work lays in the proposal and evaluation of different weight calculation procedures for Adaptive Lasso and their comparison with Lasso in high censoring scenarios, along with a procedure for selecting the best model or variable selection for Cox regression.

Keywords: Cox regression; Lasso; Adaptive Lasso; Survival analysis; High dimension.

Bibliography

[1] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.

[2] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

[3] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429. doi:10.1198/016214506000000735.

A novel column generation framework for one-for-many counterfactual analysis

20th June
12:20–12:40
Contributed
Session 7

Jasone Ramírez-Ayerbe[1] and Andrea Lodi[2]

[1]Instituto de Matemáticas de la Universidad de Sevilla, Spain; [2]Cornell Tech, Cornell University, USA

Jasone Ramírez-Ayerbe is a fourth-year PhD student in the Department of Statistics and Operations Research at University of Seville. Her research focuses on Data Science and Optimization. She investigates interpretability, such as counterfactual analysis by means of mathematical optimization and applied to Machine Learning models. She has a double BSc in Mathematics and Physics and a MSc in Mathematics from Universidad de Sevilla.

In this talk, we consider the problem of generating a set of counterfactual explanations, i.e., the changes needed to be made in the features of instances to change the prediction made by a given classifier [2]. We study the problem for a group of instances, with the one-for-many allocation rule, where one explanation is allocated to a subgroup of the instances [1]. By computing global explanations, one can find patterns in the explanations [3] as well as detect key features that need to change, globally, for all the instances to change their prediction. Moreover, with the one-for-many allocation rule we search for just a few counterfactual instances for a group to be seen as benchmarks of the class.

For the first time, we solve the problem of minimizing the number of explanations needed to explain all the instances, while considering sparsity by limiting the number of features allowed to be changed collectively in each explanation [5]. A novel column generation framework is developed to efficiently search for the explanations. Our framework can be applied to any black-box classifier, like neural networks. Compared with a simple adaptation of a mixed-integer programming formulation from the literature, the column generation framework dominates in terms of scalability, computational performance and quality of the solutions.

Keywords: column generation; counterfactual analysis; explainability; integer programming.

Bibliography

[1] Carrizosa, E., Ramírez-Ayerbe, J., Romero Morales, D. (2024). Mathematical optimization modelling for group counterfactual explanations. Forthcoming in *European Journal of Operational Research*. doi:10.1016/j.ejor.2024.01.002.

[2] Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. Forthcoming in *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-022-00831-6.

[3] Keane, M., and Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *International Conference on Case-Based Reasoning*, pp. 163–178. https://doi.org/10.1007/978-3-030-58342-2_11.

[4] Liu, J., Zhu, X., and Ohannessian, H. (2016). The teaching dimension of linear learners. In *Proceedings of The 33rd International Conference on Machine Learning*:117–126. https://proceedings.mlr.press/v48/liua16.html.

[5] Lodi, A., Ramírez-Ayerbe, J. (2024). One-for-many Counterfactual Explanations by Column Generation. arXiv:2402.09473.

A framework for adversarial regression

20th June
12:40–13:00
Contributed
Session 7

Pablo G. Arce[1], Cristina Lopez Amado[2], Roi Naveiro Flores[3] and David Ríos Insúa[1]

[1]ICMAT, Madrid, España; [2]CITUS, Universidad de Santiago de Compostela, España; [3]CUNEF, Madrid, España

Pablo García Arce is a first-year PhD student in the PhD in Computer Science. His research activity is focused on the robustification of machine learning models against adversarial attacks. Before enrolling in his PhD, he worked at Predictia Intelligent Data Solutions as Data Scientist and did a MSc in Statistics and Computation at Universidad Complutense de Madrid and a BSc in Mathematics and Physics at Universidad de Cantabria.

In addition to the positive developments attributable to machine learning (ML), important misuses have also been reported. Many of these issues arise from attempts by adversaries to fool ML algorithms to attain a benefit and have led to the emergence of the relatively recent field of *adversarial machine learning* (AML) [3, 4, 2]. Conventional ML models rely on the assumption of independent and identically distributed (iid) data during training and operations. However, in a wide range of applications, it is important to consider that an adversary may modify model inputs altering thereby the incumbent distributions. The final aim of AML is to provide algorithms that are more robust against adversarial manipulations and has mainly focused on three issues: studying attacks to ML algorithms to understand their vulnerabilities; designing defenses against attacks to better protect the algorithms; and providing frameworks that determine best defenses against potential attacks.

This paper introduces a pipeline to robustify linear regression models against adversarial attacks. The pipeline integrates procedures to protect models both during training and operations, following the work done in [1]. Protection during training is based on robustly training the model to face adversarial manipulation at inference, while protection during operations is based on adjusting the process at inference time to take into account the presence of an attacker. The pipeline also integrates techniques to detect attacks and changes in attack patterns. A change in attack patterns showcases the need to adapt the pipeline, either by retraining (in a robust manner) or adjusting inference. Several examples illustrate the role of the pipeline.

Keywords: adversarial risk analysis; Bayesia; Machine Learning; regression.

Bibliography

[1] Gallego, V., Naveiro, R., Redondo, A., Rios-Insua, D., and Ruggeri, F. (2024). Protecting Classifiers From Attacks. A Bayesian Approach. `arXiv:2004.08705`.

[2] Huang, L., Joseph, A., Nelson, B., and Rubisntein, B., and Tygar, J.D. (2011). Adversarial Machine Learning. In *Proceedings of 4th ACM Workshop on Artificial Intelligence and Security*:43–58. `doi:10.1145/2046684.2046692`.

[3] Rios-Insua, D., Naveiro, R., Gallego, R., and Poulos, J. (2023). Adversarial Machine Learning: Bayesian Perspectives. *Journal of the American Statistical Association*, 118(53):2195–2206. `doi:10.1080/01621459.2023.2183129`.

[4] Vorobeichyk, Y., and Kantarcioglu, M. (2018). *Adversarial Machine Learning*. Springer. `doi:10.1007/978-3-031-01580-9`.

A distributionally robust optimisation approach to fair credit scoring

20th June
13:10–13:30
Contributed
Session 8

Pablo Casas[1], Christophe Mues[1] and Huan Yu[1]

[1]Business School, University Of Southampton

Pablo Casas is a third-year PhD student at the University of Southampton. His research focuses on fair machine learning in Credit Scoring under uncertainty and data shifts. Before enrolling on his PhD, he did an MSc in Intelligent and Adaptive Systems at the University of Sussex, where he focused on complex systems and long-range temporal correlation in brain oscillations and a BSc in Economics at the Complutense University of Madrid.

Optimization under uncertainty has been a topic of large debate in Operations research (OR) and Machine Learning (ML) communities for a long time. Recently, many authors have turned to distributional robust optimization (DRO) in search of a solution [4]; however, fewer studies have focused on enhancing fairness under uncertainty [5], note that the topic of fairness in credit scoring (not taking into consideration uncertainty) is a well-studied issue [3]. This is of great interest in those applications of ML that can have a direct damaging impact on the population, credit scoring (CS) being one of the most potentially damaging fields according to regulatory bodies [2].

The paper presented [1] explores the effects of using a DRO-based logistic regression (LR), one of the most commonly used classifiers in credit scoring, across multiple CS datasets. This study will show how robustness has a greater impact on fairness than the fairness constraint, and that the impact on performance is negligible and, in some datasets, positive. We will also provide an empirical analysis of the effect of the different hyperparameters that are unique to DRO-based LR. Furthermore, we will argue that the level of robustness narrows the dispersion of the probability of default distribution and that the parameters in charge of the ground metric have an unnoticeable impact. On a side note, we suggest traditional fairness metrics used in credit scoring are not best suited for the task.

Keywords: credit scoring; fairness; Machine Learning; robust optimization.

Bibliography

[1] Casas, P., Mues, C., and Yu, H. (2024). A Distributionally Robust Optimisation Approach to Fair Credit Scoring. `arXiv:2402.01811`.

[2] Falque-Pierrotin, I. (2017). Guidelines on data protection officers ('dpos'). `https://ec.europa.eu/newsroom/article29/items/612048/en`.

[3] Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094. `doi:10.1016/j.ejor.2021.06.023`.

[4] Long, D.Z., Sim, M., and Zhou, M. (2022). Robust Satisficing. *Operations Research*, 71(1):61–82. `doi:10.1287/opre.2021.2238`.

[5] Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. (2020). A distributionally robust approach to fair classification. `arXiv:2007.09530`.

Temporal M-quantile models and robust bias-corrected small area predictors

20th June
13:30–13:50
Contributed
Session 8

María Bugallo[1], Domingo Morales[1], Nicola Salvati[2] and Francesco Schirripa[2]

[1]Center of Operations Research, Miguel Hernández University of Elche, Spain;
[2]Department of Economics and Management, University of Pisa, Italy.

María Bugallo is a third-year PhD student in Statistics at Miguel Hernández University of Elche. Her research is focused on finite population inference and statistical modelling for Small Area Estimation. Before enrolling her PhD, she was beneficiary of a 10-month collaboration grant at University of Santiago de Compostela, and then worked for more than a year at the Singular Research Centre on Intelligent Technologies. She completed a MSc in Statistics and a BSc in Mathematics, both at University of Santiago de Compostela.

Sampling is done according to cost-effectiveness principles, although disaggregated statistics facilitate more effective targeting of decision-making. Furthermore, outliers are a common occurrence in survey data, and even more of a concern in small area estimation (SAE), where model-based inference is the standard. Smart solutions to deal with small area sample sizes are based on data measured over time. Existing outlier robust small area estimators can be either substantially biased or have uncontrollably increased variance.

Our contribution extends the M-quantile (MQ) modelling approach to SAE [1] to temporal data, softening the widely imposed assumptions of unit-level independence. Under the new model, robust bias-corrected predictors of small area linear indicators are derived. In addition, the optimal selection of the robustness parameters for bias correction in MQ models is a theoretical strength of the current research, exploring its applicability as a diagnostic tool for outlier detection. Research on optimality criteria is based on the fact that the bias correction does not come at the price of increased variability, reaching a trade-off. Therefore, we propose data-driven, small-area specific criteria and demonstrate the existence and uniqueness of a solution. As far as the estimation of the mean squared error (MSE) is concerned, we have also obtained, under general conditions, a first-order approximation and proposed several analytical estimators.

Numerous simulation experiments are carried out to test the performance of the new predictors and ensuing MSE estimators, as well as the optimal selection of the robustness parameters. We have found that properly bias-corrected MQ predictors, both derived from temporal and non-temporal models, outperform empirical best linear unbiased predictors based on linear mixed models, even in the most favourable scenarios for the latter. Finally, an application to the 2013-2022 Spanish Living Conditions Survey data is included to illustrate the usefulness of what has been done. Specifically, we analyze changes in the average level of income in the provinces of *Empty Spain*.

Extending the methodology to spatio-temporal data is an avenue for future research.

Keywords: SAE; robust inference; MQ models; optimal bias-correction.

Bibliography

[1] Chambers, R., Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2):255–268. doi:10.1093/biomet/93.2.255.

SYSØR 2024
4TH SPANISH YOUNG STATISTICIANS AND OPERATIONAL RESEARCHERS MEETING

A profit-driven churn prevention approach within predict-and-optimize

20th June
13:50–14:10
Contributed
Session 8

Nuria Gómez-Vargas[1], Sebastián Maldonado[2] and Carla Vairetti[3]

[1]Department of Statistics and Operations Research, University of Seville, Spain; [2]Department of Management Control and Information Systems, University of Chile, Chile; [3]Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Chile

Nuria Gómez-Vargas is a third-year PhD student in the PhD in Mathematics in the department of Statistics and Operations Research at University of Seville (US). Her research activity falls into the area of Prescriptive Analytics, where Supervised Machine Learning is used to address the uncertainty in the parameters of Operational Research problems. Before enrolling and during her PhD, she has worked in research projects in collaboration with business, and did a MSc in Mathematics and a BSc in Mathematics at US.

Churn prediction [3] has become one of the main marketing analytics applications. The goal is to identify the customers that are more likely to leave a company via predictive models. One approach is to use profit metrics for model evaluation [2], computing the total or the average profit of retention campaigns instead of utilizing statistical measures. Traditional profit metrics usually consider an average customer lifetime value (CLV) for all customers to ease the analysis. This simplification, however, can lead to suboptimal decisions in case the CLVs are very heterogeneous, which is common in several industries [2].

In our paper [1], we introduce a novel predict-and-optimize method for profit-driven churn prevention. We frame the task of targeting customers for a retention campaign as a regret minimization problem. The main objective is to leverage individual CLVs to ensure that only the most valuable customers are targeted. In contrast, many profit-driven strategies focus on churn probabilities while considering average CLVs. This often results in significant information loss due to data aggregation. Our proposed model aligns with the guidelines of Predict-and-Optimize (PnO) frameworks and can be efficiently solved using stochastic gradient descent methods. Results from 12 churn prediction datasets underscore the effectiveness of our approach, which achieves the best average performance compared to other well-established strategies in terms of average profit.

Keywords: Profit Metrics; Churn Prediction; Predict-and-optimize; Business Analytics; Machine Learning.

Bibliography

[1] Gómez-Vargas, N., Maldonado, S., and Vairetti, C. (2023). A predict-and-optimize approach to profit-driven churn prevention. `arXiv:2310.07047`.

[2] Maldonado, S., Domínguez, G., Olaya, D., and Verbeke, W. (2021). Profit-driven churn prediction for the mutual fund industry: a multisegment approach. *Omega*, 100:102380. `doi:10.1016/j.omega.2020.102380`.

[3] Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., and Mason, C.H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models *Journal of Marketing Research*, 43(2):204–211. `doi:10.1509/jmkr.43.2.204`.

Digit analysis using Benford's law: a Bayesian approach

20th June
14:10–14:30
Contributed
Session 8

Pedro Fonseca[1] and Rui Paulo[1]

[1]Department of Mathematics, Lisbon School of Economics and Management (University of Lisbon), Portugal

Pedro Fonseca is a PhD candidate in Applied Mathematics for economics and Management at the University of Lisbon. His research activity is focused on Bayesian model selection and hypothesis testing. He is also a visiting assistant professor in Statistics at ISEG Lisbon School of Economics and Management.

Naturally-occurring collections of numbers are known to often exhibit a logarithmically decaying pattern in the frequencies of leading digits, known as Benford's law, which can be used to screen datasets for anomalies like erroneous or fraudulent data. However, assessing conformance to Benford's law usually requires testing point null hypotheses, and classical significance tests of fixed dimension are known to over-reject point null hypotheses in large samples due to the high levels of power they attain, as the acceptance region shrinks with sample size, hence being prone to high false-positive rates. This can result in suspicions being unduly raised in a large proportion of datasets.

As an alternative, we address digit analysis within the Bayesian hypothesis testing framework. Since digit analysis can be seen either as a Multinomial or a Binomial goodness-of-fit problem, we developed methodologies based on the Multinomial $\wedge$ Dirichlet and Binomial $\wedge$ Beta Bayesian models with prior distributions and hyperparameter specifications tailored for digit analysis that allow us to obtain closed form solutions.

An empirical application with macroeconomic statistics from Eurozone countries demonstrates the applicability of the suggested methodology and explores the conflict between classical and Bayesian measures of evidence in the context of digit analysis. We found that classical tests often reject conformance to Benford's law in situations in which the Bayesian measures of evidence suggest otherwise, and that even lower bounds on the Bayesian measures often provide more evidence in favour of Benford's law than what p-values on classical test statistics seem to suggest.

Keywords: Bayes factor; fraud detection; hypothesis testing; P–value calibration; point null hypothesis.

Bibliography

[1] Berger, J.O. and Sellke, T. (1989). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, 82(397):112–122. doi:10.2307/2289139.

[2] Hill, T.P. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, 10(4):354–363. doi:10.1214/ss/1177009869.

[3] Pericchi, L., and Torres, D. (2011). Quick anomaly detection by the Newcomb—Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Statistical Science*, 26(4):502–516. doi:10.1214/09-STS296.

[4] Sellke, T., Bayarri, M.J., and Berger, J.O. (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71. https://www.jstor.org/stable/2685531.

[5] Wasserstein, R.L., and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133. doi:10.1080/00031305.2016.1154108.

Covariance dependence in mixture cure models using distance correlation

20th June
15:50–16:10
Contributed
Session 9

Blanca Monroy[1], Amalia Jácome[1] and Ricardo Cao[1]

[1]Department of Mathematics, Modes Group, Universidade da Coruña, Spain

Blanca Monroy is a second-year PhD student in the Doctoral Programme in Statistics and Operational Research at Universidade da Coruña. Her research activity is focused on cure models in large data sets, and the development of new methods for censored and genomic data. Before enrolling her PhD, she did a MSc in Statistics at Postgraduate College and a BSc in Mathematics at University of Guadalajara, México.

Survival models, widely employed in time-to-event analysis, often assume that all subjects within a study population are susceptible to the event of interest. Nevertheless, real-world data may include individuals who will never experience the event, resulting in infinite event times and their categorization as cured. The mixture cure model, notably introduced by Boag [1], is a prominent solution in the literature for handling such scenarios. This model enables the estimation of cure probabilities and survival functions for the uncured population, considering diverse covariates. An ongoing challenge involves determining whether specific covariates influence the cure rate or survival time of susceptible patients. Over the years, significance tests, proposed in both parametric and nonparametric contexts [2, 3, 4], have been instrumental in addressing this challenge.

On the other hand, in recent years, there has been a focus on a novel measure of dependence known as distance correlation, introduced by Szekely et al. [5]. This study delves into the behavior of the distance correlation ($\mathcal{R}(X, \nu)$) between covariates X and the cure indicator ν to assess the impact of covariates on the likelihood of cure. Additionally, the analysis explores the distance correlation ($\mathcal{R}(X, Y|\nu = 0)$) between covariates and survival times of uncured subjects, evaluating the influence of covariates on the survival times of the uncured through a simulation study.

Keywords: cure model; distance correlation; simulation.

Bibliography

[1] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53. `https://www.jstor.org/stable/2983694`.

[2] López-Cheda, A., Jácome, M.A., Van Keilegom, I., and Cao, R. (2020). Nonparametric covariate hypothesis tests for the cure rate in mixture cure models. *Statistics in Medicine*, 39(17):2291–2307. `doi:10.1002/sim.8530`.

[3] Müller, U.U., and Van Keilegom, I. (2018). Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika*, 106(1):211–227. `doi:10.1093/biomet/asy058`.

[4] Peng, Y., and Yu, B. (2021). *Cure Models: Methods, Applications and Implementation*. CRC Press. `doi:10.1111/biom.13671`.

[5] Székely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring ans testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794. `doi:10.1214/009053607000000505`.

The clustered-state Markovian arrival process in recurrent processes with terminal event

20th June
16:10–16:30
Contributed
Session 9

Álvaro Díaz[1], Rosa E. Lillo[12] and Pepa Ramírez-Cobo[3]

[1]UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain; [2]Department of Statistics, Universidad Carlos III de Madrid, Spain; [3]Department of Statistics and Operations Research, Universidad de Cádiz, Spain

Álvaro Díaz Pérez is a first-year PhD student in the PhD in Statistics for Data Science at Universidad Carlos III de Madrid. His research activity is focused on the Clustered-state Markovian Arrival process, a model derived from the Markovian Arrival Process. Before enrolling his PhD, he did a MSc in Statistics for Data Science at Universidad Carlos III de Madrid and a BSc in Mathematics at Universidad Autónoma de Madrid.

The Markovian Arrival Process (MAP), introduced by M. F. Neuts in 1979 (see [1]), is a versatile stochastic process widely used for modeling point processes with dependent and non-exponentially distributed inter-event times. This model and its derivatives have been extensively studied in numerous publications by P. Ramírez-Cobo and R. E. Lillo (see, for example, [2]).

In this study, we propose an extension of the MAP, which we call the Clustered-state Markovian Arrival Process (CS-MAP), tailored for modeling marked point processes with finite mark space. This new model is constructed by grouping the states of the MAP into clusters, one of which corresponds to a different mark, allowing dependent and non-identically distributed inter-event times conditioned to the marks. Our research focuses on the application of this novel stochastic process to model recurrent processes with terminal event, particularly prevalent in survival analysis contexts. Specifically, we present the application of this new model to real data involving patients with oncological diseases. Each patient exhibits not only a survival time but also a temporal sequence of recurrences, capturing instances of relapse or the emergence of new tumors. Given that right-censoring is prevalent within these contexts, it is imperative that we integrate it into our analysis. In this research, we present explicit expressions for the joint and marginal density functions of the inter-event times within the CS-MAP, as well as for the moments and correlations, focusing on the context of recurrent processes with terminal event. Additionally, we provide an explicit expression for the likelihood function, which leads to proposing a maximum likelihood approach for parameter inference. The efficiency of the proposed inference approach is demonstrated through its application to simulated data and the method is finally applied in real data.

Keywords: Markovian arrival process; survival analysis; stochastic processes.

Bibliography

[1] Neuts, M.F. (1979). A Versatile Markovian Point Process. *Journal of Applied Probability*, 16(4):764–779. doi:10.2307/3213143.

[2] Rodríguez, J.V., Lillo, R.E., and Ramírez-Cobo, P. (2015). Failure modeling of an electrical N-component framework by the non-stationary Markovian arrival process. *Reliability Engineering and System Safety*, 134:126–133. doi:10.1016/j.amc.2020.125869.

A new formulation for the Chinese Postman Problem with load–dependent costs

20th June
16:30–16:50
Contributed
Session 9

Isaac Plana[1], José María Sanchis[2] and Paula Segura[2]

[1]Department of Mathematics for Economics and Business, University of Valencia, Spain; [2]Department of Applied Mathematics, Polytechnic University of Valencia, Spain

Paula Segura is a postdoctoral researcher at the Polytechnic University of Valencia. She received her PhD in Statistics and Optimization from the University of Valencia in 2023. Her research activity focuses on the design and development of mathematical models for location and routing problems. Before enrolling her PhD, she did a MSc in Advanced Mathematics (with Operations Research specialization) at the University of Murcia, and a BSc in Mathematics at the University of Alicante.

The Chinese Postman Problem (CPP) is to find a least cost tour on a connected undirected graph that traverses each edge at least once (see [1]). This important arc routing problem models several real life situations, such as meter reading or waste management, and many variants of it have been studied in the scientific literature. The Chinese Postman Problem with load–dependent costs (CPP–LC) was introduced in [2] as an extension of the CPP in which the cost of traversing an edge depends on its length and also on the weight of the vehicle at the moment the edge is traversed. It arises from the desire to reduce pollution in transportation, since the level of emissions from a vehicle is influenced by factors beyond the distance traveled, such as its load. The authors provide in [2] two mathematical programming formulations for the problem, and propose two metaheuristic algorithms for its solution.

In this work, we present a new mixed–integer linear programming formulation for the CPP–LC, and propose some valid inequalities to reinforce it. Furthermore, we develop a study of the incompatibilities of a subset of variables from the formulation, and describe all the cliques of the underlying intersection graph, categorizing them into five families of inequalities that are valid for the problem. We design a branch–and–cut algorithm for the CPP–LC solution based on this formulation that incorporates the separation of all the valid inequalities proposed. Several computational results obtained with our new exact procedure are compared with those provided in [2].

Keywords: arc routing problems; branch and cut; clique constraints; combinatorial optimization.

Bibliography

[1] Laporte, G. (2014). The undirected Chinese postman problem. *Arc Routing: Problems, Methods, and Applications*. MOS–SIAM Series Optimization, 20:53–64. doi:10.1137/1.9781611973679.ch3.

[2] Corberán, A., Erdoğan, G., Laporte, G., Plana, I., and Sanchís, J.M. (2018). The Chinese Postman Problem with Load–Dependent Costs. *Transportation Science*, 52(2):370–385. doi:10.1287/trsc.2017.0774.

Density-based tests for the k-sample problem with left-truncated data

20th June
16:50–17:10
Contributed
Session 9

Adrián Lago[1], Jacobo de Uña-Álvarez[1], Juan Carlos Pardo-Fernández[1] and Ingrid Van Keilegom[2]

[1]Department of Statistics and Operations Research, Universidade de Vigo, Spain
[2]Department of Decision Sciences and Information Management, KU Leuven, Belgium

Adrián Lago is a third-year PhD student at the Universidade de Vigo. His research activity is focused on the comparison of populations with left-truncated data. Before enrolling his PhD, he got a MSc in Statistics at the Universidade de Vigo and a BSc in Mathematics at the Universidade de Santiago de Compostela.

The comparison of distributions has been addressed for more than a century, aiming to determine whether a target variable is equally distributed in two or more populations. One can, for example, use estimators for the density to address this task, as proposed first for the two-sample problem in [1]. Moreover, in many practical situations, data is subject to one-sided truncation (famous datasets are studied in [5], in Astronomy, and [3], in Medicine). A Kolmogorov-Smirnov goodness-of-fit test for left-truncated data was studied in [4] but it has not been adapted for the two-sample problem until very recently in [2]. It represents an alternative to the well-known log-rank test, which may lead to a low statistical power when the proportional hazards assumption is violated.

In this work, two density-based tests for the k-sample problem with left-truncated data are proposed. It is shown that they both follow a normal distribution asymptotically under the null hypothesis. As the mean and variance cannot be either computed exactly or estimated, two bootstrap resampling plans are suggested to approximate the null distribution of the two test statistics. The performance of such a method and the effect of the smoothing parameter is studied via Monte Carlo simulations. The proposed tests are compared with other tests for left-truncated data, such as the Kolmogorov-Smirnov and the log-rank, to determine which is the most powerful one in different scenarios. The performance of the tests is exemplified with two real datasets regarding pregnancy and unemployment times.

Keywords: Density-based test; kernel smoothing; log-rank; truncation; Survival Analysis

Bibliography

[1] Anderson, N. H., Hall, P., and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54. doi:10.1006/jmva.1994.1033.

[2] Lago, A., de Uña-Álvarez, J., and Pardo-Fernández, J. C. (2024). A Kolmogorov-Smirnov-type test for the two-sample problem with left-truncated data. Submitted for publication.

[3] Lagakos, S. W., Barraj, L. M., and Gruttola, V. D. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523. doi:10.2307/2336602.

[4] Guilbaud, O. (1988). Exact Kolmogorov-type tests for left-truncated and/or right-censored data. *Journal of the American Statistical Association*, 83(401), 213–221. doi:10.2307/2288943.

[5] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1):95–118. doi:10.1093/mnras/155.1.95.

Friday 21

Exploring disease mapping: models and applications

21st June
09:15–10:15
Plenary
Contributed
Session 3

Lola Ugarte[1,2]

[1]Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain;
[2]INAMAT, Public University of Navarre, Spain

Lola Ugarte is full professor in the Department of Statistics, Computer Science, and Mathematics at the Public University of Navarre and director of INAMAT[2] (Institute of Advanced Materials and Mathematics) at the same university. Her research interests lie in the general field of statistical modelling and spatial and temporal-spatial statistics with applications in many fields. She is currently the President of the Federation of European National Statistical Societies (FENStatS).

In this presentation, we will explore one of the most compelling applications of areal data: disease mapping. Following a brief historical overview, we will introduce the most widely used univariate space-time models in this domain [2, 5]. Subsequently, we will delve into recently developed multivariate models, showcasing an application in gender-based violence [6]. Additionally, we will discuss a straightforward alternative for analyzing large datasets, encompassing both univariate [3] and multivariate models [4]. Methods and algorithms are implemented in the R package bigDM [1]. Finally, we will touch upon other research topics within this domain.

Keywords: Bayesian inference; crimes against women; multivariate models; spatio-temporal models.

Bibliography

[1] Adin, A., Orozco-Acosta, E., and Ugarte, M.D. (2023). bigDM: scalable Bayesian disease mapping models for high-dimensional data. R package version 0.5.3 `https://github.com/spatialstatisticsupna/bigDM`.

[2] Goicoa, T., Adin, A., Ugarte, M.D., and Hodges, J. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch Environ Res Risk Assess*, 32(3):749–770. `doi:10.1007/s00477-017-1405-0`.

[3] Orozco-Acosta, E., Adin. A. and Ugarte, M.D. (2021). Scalable Bayesian modeling for smoothing disease risks in large spatial data sets using INLA. *Spatial Statistics*, 41:100496. `doi:10.1016/j.spasta.2021.100496`.

[4] Vicente, G., Adin, A., Goicoa, T., and Ugarte, M.D. (2023). High-dimensional order-free multivariate spatial disease mapping. *Statistics and Computing*, 33:104. `doi:10.1007/s11222-023-10263-x`.

[5] Vicente, G., Goicoa, T., Fernandez-Rasines, P. and Ugarte, MD. (2020). Crime against women in India: unveiling spatial patterns and temporal trends of dowry deaths in the districts of Uttar Pradesh. *Journal of the Royal Statistical Society. Series A*, 183:655–679. `doi:10.1111/rssa.12545`.

[6] Vicente, G., Goicoa, T., and Ugarte, M.D. (2023). Multivariate Bayesian spatio-temporal P-spline models to analyze crimes against women *Biostatistics*, 24(3):562–584. `doi:10.1093/biostatistics/kxab042`.

Optimal participation of energy communities in electricity markets under uncertainty. A multi-stage stochastic programming approach

21st June
10:20–10:40
Contributed
Session 10

Albert Solà Vilalta[1], Marlyn D. Cuadrado[2], Ignasi Mañé Bosch[1] and F.- Javier Heredia[1]

[1]Department of Statistics and Operations Research, Polytechnic University of Catalonia, Spain;
[2]Department of Product, Hoop Carpool, Spain;

Albert Solà Vilalta is a postdoctoral researcher in Optimisation at Polytechnic University of Catalonia. His research interests are decomposition techniques in Optimisation, and applications of Operations Research in the energy transition and sustainable development. Before joining the Polytechnic University of Catalonia, he was a Maxwell Institute Postdoctoral Research Fellow at the University of Edinburgh (UK), obtained a PhD in Optimisation and Operations Research from the University of Edinburgh (UK), a MSc in Mathematics from the University of Bonn (Germany), and a BSc in Mathematics from the University of Barcelona.

An energy community is a new legal figure in the European Union that creates a framework to encourage active participation of citizens and local entities in the energy transition to net-zero [1]. In this work, we study the optimal participation of energy communities in day-ahead, reserve, and intraday electricity markets. The motivation to do so is that there are time periods where energy communities cannot meet their internal demand and periods where they generate excess electricity because most of the electricity they generate comes from variable renewable resources like solar and wind. Electricity market participation is a natural way to ensure they meet their internal demand at all times, and, simultaneously, make the most of the excess electricity.

Following [2] and [3], we propose a multi-stage stochastic programming model that captures variable renewable and electricity price uncertainty. The multi-stage aspect models the different times at which variable renewable generation and electricity prices from different markets are revealed. This results in a scenario tree with 34 stages, and hence a large optimization problem. Scenario reduction techniques are applied to make the problem tractable. Case studies with real data are discussed, considering different energy community configurations, to analyse proposed regulatory frameworks in Europe. The added value of considering uncertainty is also analysed.

Keywords: energy communities; electricity markets; multi-stage stochastic programming; operative research; stochastic optimization.

Bibliography

[1] Boulanger, S.O.M., Massari, M., Longo, D., Turillazzi, B., and Nucci, C.A. (2021). Designing collaborative energy communities: a European overview. *Energies*, 14:8226. doi:10.3390/en14248226.

[2] Corchero, C., Mijangos, E., and Heredia, F.J. (2013). A new optimal electricity market bid model solved through perspective cuts. *TOP*, 21:84–108. doi:10.1007/s11750-011-0240-6.

[3] Heredia, F.J., Cuadrado, M.D., and Corchero, C. (2018). On optimal participation in the electricity markets of wind power plants with battery energy storage systems. *Computers & Operations Research*, 96:316–329. doi:10.1016/j.cor.2018.03.004.

Parameter estimation for a bivariate Wiener model subject to imperfect maintenance and varying observation strategies

21st June
10:40–11:00
Contributed
Session 10

Lucía Bautista Bárcena[1], Inma T. Castro[1], Christophe Bérenguer[2], Olivier Gaudoin[3] and Laurent Doyen[3]

[1]Department of Mathematics, University of Extremadura, Spain; [2]GIPSA-lab, University Grenoble-Alpes, France; [3]Laboratoire Jean Kuntzmann, University Grenoble-Alpes, France

Lucía Bautista Bárcena is a Substitute Professor in the Department of Mathematics at the University of Extremadura. She received her Ph.D. degree from the University of Extremadura in 2023. Her research activity is focused on developing stochastic models in reliability engineering and system maintenance, for which she has carried out a 3-month research stay at the University Grenoble-Alpes, in France.

Research in degradation modelling has traditionally focused on univariate stochastic processes. In practice, industrial systems have a more complex structure with different interrelated parts or components that influence the system performance. For instance, in lighting systems composed of many LED lamps, which present a likely dependence because of the common usage history. For these systems, a multivariate model to describe the degradation evolution is needed. The *trivariate reduction method*, which consists of constructing two dependent stochastic processes by the superposition of three independent univariate processes, has recently attracted a lot of attention to model this multiple degradation [1]. To reduce the degradation effect on system components, imperfect maintenance actions are periodically performed on it. The effect of such maintenance is to reduce the level of degradation by an amount proportional to the degradation accumulated in the system [3].

The aim of this work is to estimate the parameters of a bivariate Wiener model considering different observation strategies and imperfect maintenance. In previous studies, the degradation levels are generally observed just before maintenance actions [2]. Here, four different observation strategies are considered, so that degradation levels can be observed between maintenance actions, as well as just before or just after maintenance times. In each of them, degradation levels are observed between successive maintenance actions. Within this framework, the maximum likelihood function is obtained for each observation strategy.

Keywords: multivariate degradation modelling; Wiener process; preventive maintenance; maximum likelihood estimation.

Bibliography

[1] Lai, C.D. (1995). Construction of bivariate distributions by a generalised trivariate reduction technique. *Statistics & Probability Letters*, 25(3):265–270. doi:10.1016/0167-7152(94)00230-6.

[2] Leroy, M., Bérenguer, C., Doyen, L., and Gaudoin, O. (2023). Statistical inference for a Wiener-based degradation model with imperfect maintenance actions under different observation schemes. *Applied Stochastic Models in Business and Industry*, 39(3):352–371. doi:10.1002/asmb.2742.

[3] Mercier, S. and Castro, I.T. (2019). Stochastic comparisons of imperfect maintenance models for a gamma deteriorating system. *European Journal of Operational Research*, 273:237–248. doi:10.1016/j.ejor.2018.06.020.

L1-Approximation of supply curves

21st June
11:00–11:20
Contributed
Session 10

Zehang Li[1] and Andrés M. Alonso[2]

[1]Department of Statistics, University Carlos III of Madrid, Spain; [2]Department of Statistics and Institute Flores de Lemus, University Carlos III of Madrid, Spain

Zehang Li is a four-year PhD student in the PhD in Mathematical Engineering at University Carlos III of Madrid. His research activity is focused on the modeling of the electricity market and, in particular, on the daily market offer curves. Before enrolling his PhD, he worked at Company Hangzhou DATA-WTC Technology Co., Ltd, China as Data Analyst, and did a MSc in in Statistics for Data Science at University Carlos III of Madrid and a BSc in Telecommunications engineering at University of Hohai, China.

A supply curve is a non-decreasing step function having steps in the positions corresponding to the ordered prices of the offers. A "natural" representation is to consider the set of prices of all bids, but the size of that set makes this approach intractable. In [3], a mesh-free interpolation technique was proposed for supply and demand curves. However, their approach fails because the resulting approximation is a smooth function. The encoded offer curves (EOC) was proposed by [2] which consists on a continuous piecewise version of the true offer curve. The EOC has the disadvantage of increasing the dimension of representation. In [1], a functional data nonparametric techniques have been used to model residual demand curves. They assume that the functions are sufficiently smooth with up to two derivatives, something that is not appropriate in step functions.

In this paper, we illustrate the computation of the approximation of the supply curves using a one-step basis. We study the L1 approximation and propose two procedures for the selection of nodes of the approximation. The L1 approximation is formulated as a linear programming problem and the selection of nodes is formulated as a mixed-integer programming problem. We provide a simple procedure to solve this linear programming problem that, as a byproduct, allows us to propose a dyadic search for nodes. We illustrate the procedures using the supply curves of the Spanish daily electricity market and show the proposed approach obtains a better and more parsimonious approximation than other alternatives that assume smoothness of the curves.

Keywords: electricity market; linear programming problem; mixed-integer programming problem; supply curves.

Bibliography

[1] Aneiros, G., Vilar, J.M., Cao, R., and Muñoz San Roque, A. (2013). Functional prediction for the residual demand in electricity spot markets. *IEEE Transactions on Power Systems*, 28(4):4201–4208. doi:10.1109/TPWRS.2013.2258690.

[2] Mestre, G. Sánchez-Úbeda, E.F., Muñoz San Roque, A. and Alonso, E. (2022). The arithmetic of stepwise offer curves. *Energy*, 239 (Part E):122444. doi:10.1016/j.energy.2021.122444.

[3] Soloviova, M., and Vargiolu, T. (2022). Efficient representation of supply and demand curves on day-ahead electricity markets. *Journal of Energy Markets*, 14(1). https://ssrn.com/abstract=3836018.

Classifiers based on minimum spanning trees

21st June
11:45–12:05
Contributed
Session 11

Julio González-Díaz[1,2], Beatriz Pateiro-López[1,2] and Iria Rodríguez-Acevedo[1]

[1]Department of Statistics, Mathematical Analysis and Optimization and MODESTYA Research Group, University of Santiago de Compostela;
[2]CITMAga (Galician Center for Mathematical Research and Technology)

Iria Rodríguez Acevedo is a second-year PhD student in the PhD in Statistics and Operations Research at University of Santiago de Compostela. In her research activity there are two lines of study: on the one hand, the development of a new classification technique based on minimum spanning trees and, on the other hand, the adaptation of the RLT technique to mixed-integer polynomial optimization problems.

The classification of observations into specific categories based on their observable characteristics is a fundamental task in data analysis. Various classification techniques, such as k-nearest neighbours [2], support vector machines [1], and more recently, those based on neural networks [4], have made it applicable in numerous fields. On the other hand, minimum spanning trees are widely used due to their ease of construction using greedy algorithms like Prim's algorithm [3]. However, we could not find any literature that uses minimum spanning trees to define classification rules.

This work presents a new methodological contribution in the form of a classification rule. The method is based on constructing minimum spanning trees for each class using a training set. To assign a new observation to a class, the method selects the class whose spanning tree is least affected by the new observation. Conformity is measured by the separation of a network, which is the ratio of the cost of the minimum spanning trees to the number of observations. An improvement of the method has also been studied to make it more robust to outlier observations and, at the same time, more computationally efficient. The computational study concludes that the robust method performs adequately and is particularly competitive with k-NN when applied to datasets that naturally define graphs in each class. Furthermore, the study concludes that the robust version is not very sensitive to changes in its parameters, which contrasts with the observed behaviour of k-NN when the number of neighbours is varied.

Keywords: classifier; k-NN; minimum spanning tree; network; robust.

Bibliography

[1] Boser, B.E., Guyon, I.M., and Vapnik V.N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*:144–152. doi:10.1145/130385.130401.

[2] Fix, E., and Hodges, J. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

[3] Prim, R.C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401. doi:10.1002/j.1538-7305.1957.tb01515.x.

[4] Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. doi:10.1038/323533a0.

Improving the estimation of production functions through machine learning: a gradient boosting approach

21st June
12:05–12:25
Contributed
Session 11

María D. Guillén[1] and Juan Aparicio[1]

[1]Center of Operations Research (CIO). Miguel Hernandez University of Elche (UMH)

María D. Guillén is a third-year PhD student in the PhD in Statistics, Optimization and Applied Mathematics at the University Miguel Hernández of Elche. Her research activity focuses on the intersection of Data Envelopment Analysis and machine learning. Before enrolling on her PhD, she did a MSc in Big Data at the University of Santiago de Compostela and the University of Murcia and a BSc in Computer Engineering at the University of Murcia.

In microeconomics and production engineering, a topic of interest is to assess the performance of firms from the estimation of an unknown production function. The production function represents the maximum output that can be obtained from an input profile. Technical inefficiency is determined as the distance of each firm to the production function. These production functions are monotonic nondecreasing functions that envelop the data from above. Therefore, a valid estimator must satisfy the same shape constraints. Standard nonparametric methods for efficiency measurement such as Data Envelopment Analysis (DEA) [1] provide approximations of the production functions that suffer from overfitting, systematically underestimating firms' inefficiency and yielding inaccurate predictions of the output (i.e., the response variable).

In this work, we improve existing nonparametric methods for efficiency measurement by proposing a new approach that, following the machine learning paradigm, provides a more accurate prediction of the underlying true production function by adapting the Gradient Tree Boosting algorithm to the production context. Through simulations, we prove that the new technique outperforms the standard techniques in terms of bias and mean squared error [2]. Moreover, we show how to calculate different efficiency measures using the estimator determined through the new algorithm. Nevertheless, from a computational point of view, the new approach presents thousands of decision variables, making it complex to solve. To tackle this problem, we also propose and check a heuristic approximation for the exact measures [3].

Keywords: Data Envelopment Analysis; Machine Learning; Gradient Tree Boosting; Shape Constraints.

Bibliography

[1] Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2, 429–444. doi:10.1016/0377-2217(78)90138-8.

[2] Guillen, M. D., Aparicio, J., and Esteve, M. (2023). Gradient tree boosting and the estimation of production frontiers. *Expert Systems with Applications*, 214, 119134. doi:10.1016/j.eswa.2022.119134.

[3] Guillen, M. D., Aparicio, J., and Esteve, M. (2023). Performance Evaluation of Decision-Making Units Through Boosting Methods in the Context of Free Disposal Hull: Some Exact and Heuristic Algorithms. *International Journal of Information Technology & Decision Making*, 1–30. doi:10.1142/S0219622023500050.

Statistical inference in random slope mixed models for small area estimation

21st June
12:25–12:45
Contributed
Session 11

Naomi Diz-Rosales[1], María José Lombardía[1] and Domingo Morales[2]

[1]CITIC, Universidade da Coruña, Spain; [2]IUICIO, Universidad Miguel Hernández de Elche, Spain

Naomi Diz-Rosales is a second year PhD student in the PhD in Statistics and Operations Research at the Universidade da Coruña (UDC). Her research is focused on the development of random slope mixed models to solve small area estimation problems. She obtained the BSc in Biology and the MSc in Bioinformatics, both at the UDC.

Small area estimation (SAE) is a multidisciplinary branch of statistics that aims to obtain precise estimates in areas where the number of available observations is very small or even zero. It has its origins in the studies of Fay and Herriot [4] and continues to evolve constantly, with random intercept mixed models (MMs) standing out. However, when the relationship between the target variable and the auxiliary variables is not constant in all domains, it is necessary to make the modelling more flexible by also incorporating random slopes. Dempster et al. [1] first defined random slope MMs, but since then, despite the good results obtained, their application to SAE problems has been virtually unexplored, and they have never been defined for generalised mixed models.

In this context, we have developed an area-level random regression coefficient Poisson model, which we call ARRCP model. In parallel to this mathematical development, programming has already been carried out in the statistical software R. Finally, two applications to real data are being developed with very good results. First, we have started to study the modelling of the poverty ratio by province and sex using data from the Survey of Living Conditions in Spain with the ARRCP model [2]. Subsequently, we have started to apply the ARRCP model to the estimation and prediction of the future occupancy ratio of Intensive Care Units by COVID-19 patients in Spain [3].

Keywords: COVID-19; mixed models; poverty; random slope; small area estimation.

Bibliography

[1] Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374):341–353. doi:10.1080/01621459.1981.10477653.

[2] Diz-Rosales, N., Lombardía, M.J., and Morales, D. (2023). Poverty Mapping Under Area-Level Random Regression Coefficient Poisson Models. *Journal of Survey Statistics and Methodology*, 12(2):404–434. doi:10.1093/jssam/smad036.

[3] Diz-Rosales, N., Lombardía, M.J., and Morales, D. (2023). Predicting intensive care unit bed occupancy under random regression coefficient Poisson models: Application to the COVID-19 pandemic in Galicia. In *Proceedings of The XVI Congreso Galego de Estatística e Investigación de Operacións*, pp. 148–159. http://xvicongreso.sgapeio.es/informacion/Libro_actas_XVI_congreso_SGAPEIO.pdf.

[4] Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277. doi:10.1080/01621459.1979.10482505.

Cost-sensitive semi-parametric classification model

21st June
12:45–13:05
Contributed
Session 11

Jorge C. Rella[1,2], Ricardo Cao[1] and Juan M. Vilar[1]

[1]Department of Mathematics, Research Group MODES, CITIC, University of A Coruña, Spain;
[2]Department of Risks, ABANCA Financial Services, Spain

Jorge C. Rella is a third-year PhD student in the PhD in Statistics for Data Science at University of A Coruña. His research activity is focused on cost-sensitive classification techniques for credit risk. Before enrolling his PhD, he worked at Minsait and Abanca as Data Analyst, and did a MSc in Statistics at University of Santiago de Compostela and a BSc in Mathematics at University of Santiago de Compostela.

In classification tasks, the relationship between a binary variable, $Y \in \{0,1\}$, and a d-dimensional covariate, $\boldsymbol{X} \in \mathbb{R}^d$, is modeled as:

$$Y = \mathbb{I}\left(\theta'\mathbf{X} - U \geq 0\right) \tag{1}$$

where $\theta \in \mathbb{R}^d$ is a parameter vector and U is an unobserved random error term. Denoting s as the conditional distribution function of U, equation (1) can be expressed as:

$$P(Y = 1 \mid \mathbf{X}) = \mathbb{E}\left(\mathbb{I}\left(U \leq \theta'\mathbf{X}\right) \mid \mathbf{X}\right) = s(\theta'\mathbf{X}) \tag{2}$$

Assuming that U follows a logistic distribution, (2) becomes the logistic model. Single index models (SIMs) assume that s is completely unknown, but still make some parametric restrictions assuming that the distribution of U depends on $\mathbf{X}$ only through an index $\theta'\mathbf{X}$ [3, 4]. Thus, offering both interpretability and flexibility for data modeling.

Classical classification algorithms are usually trained maximizing accuracy. However, decision-making based solely on minimizing the probability of incorrect classification can lead to poor performance in problems with different missclasification error costs [1, 2].

We propose a cost-sensitive SIM by estimating the parameter θ and the link function s in a two-step iterative process minimizing the expectation of losses. Training the model with a cost-sensitive approach combined with the flexibility of SIMs, leads to proficient results. This is demonstrated through an extensive simulation study and the analysis of three real data sets, where the proposed approach outperforms both cost-sensitive, parametric and semi-parametric previous approaches.

Keywords: Cost-sensitive classification; semi-parametric modeling; link function.

Bibliography

[1] Bahnsen, A.C., Stojanovic, A., Aouada, D., and Ottersten, B., (2013). Cost-sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) 2013, pp. 333–338. doi:10.1109/ICMLA.2013.68.

[2] Elkan, C., (2001). The foundations of cost-sensitive learning. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01). https://cs.fit.edu/~pkc/ml/related/elkan-ijcai01.pdf.

[3] Klein, R., and Spady, R., (1993). An efficient semiparametric estimator of the binary response models. *Econometrica*, 61:387–421. doi:10.2307/2951556.

[4] Ker, A.P., and Sam, A.G., (2018). Semiparametric estimation of the link function in binary-choice single-index models. *Computational Statistics*, 33:1429–1455. doi:10.1007/s00180-017-0779-2.

The role of OR in shaping resilient and sustainable communities: A case study and future prospects

21st June
13:15–14:15
Plenary
Contributed
Session 4

María Paola Scaparra

Kent Business School, University of Kent, UK

Maria Paola Scaparra is a Professor of Management Science at Kent Business School. Her research activity is focused on developing optimization models for solving intractable challenges in developing countries and contributing to the achievement of the UN Sustainable Development Goals. She has extensive experience leading international, multi-disciplinary and consultancy projects, including two recent Global Challenges Research Fund projects to support communities in Southeast Asia.

Established by the United Nations in 2015, the Sustainable Development Goals (SDGs) represent a shared commitment to ending poverty, protecting the planet, and ensuring that all people enjoy peace and prosperity by 2030. They address various interconnected global challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. Achieving these goals requires innovative approaches and collaboration across disciplines. Operational Research (OR) offers a unique set of tools to streamline decision-making processes and resource allocation, thereby holding an enormous potential to contribute to the attainment of the SDGs.

The first part of this talk will use a real case study to exemplify the impactful role of OR in sustainable development and derive insights about key ingredients to deliver successful OR projects in developing countries. The case study is based on the OSIRIS project, a project funded by the UK Global Challenges Research Fund (GCRF) to identify optimal flood mitigation strategies in Vietnam cities. By integrating data-driven models, inter-disciplinary approaches and stakeholder engagement, OSIRIS demonstrates how OR can address a complex environmental challenge, ultimately contributing to SDG 11 (Sustainable Cities and Communities) and SDG 13 (Climate Action).

The second part of the talk will outline other urgent development needs in Southeast Asia and beyond, which can be tackled using OR methodologies within an interdisciplinary framework. These include for example challenges in the areas of transport infrastructure development, waste management, food security and healthcare.

Overall, this talk is a call for action to the young OR community to take an active role in developing novel OR tools which can tackle pressing challenges in developing countries. By harnessing interdisciplinary collaboration and innovative methodologies, OR can catalyse progress towards achieving the Sustainable Development Goals and building a more equitable and resilient world.

Keywords: flood mitigation; optimization; SDGs; sustainable development.

Bibliography

[1] Optimal Investment Strategies to Minimize Flood Impact on Road Infrastructure Systems in Vietnam (OSIRIS). `https://research.kent.ac.uk/gcrf-osiris/`.

List of all attendees

1. Ameijeiras Alonso, Jose (Universidade de Santiago de Compostela)
2. Baldomero Naranjo, Marta (Universidad de Cádiz)
3. Bautista Bárcena, Lucía (Universidad de Extremadura)
4. Bolón, Diego (Universidade de Santiago de Compostela)
5. Bugallo Porto, María (Universidad Miguel Hernández de Elche)
6. Cabello García, Esteban (Universidad Miguel Hernández de Elche)
7. Casas, Pablo (University of Southampton)
8. Castilla González, Elena (Universidad Rey Juan Carlos)
9. Cía Mina, Álvaro (Universidad de Navarra)
10. Corrales Alonso, Daniel (CSIC-ICMAT)
11. Croux, Christophe (KU Leuven)
12. Davila Pena, Laura (University of Kent)
13. de la Calle Arroyo, Carlos (Universidad de Oviedo)
14. Díaz Pérez, Álvaro (Universidad Carlos III de Madrid)
15. Diz Rosales, Naomi (Universidade da Coruña)
16. España Roch, Victor Javier (Universidad Miguel Hernández de Elche)
17. Fonseca, Pedro (ISEG Lisbon School of Economics and Management)
18. García Arce, Pablo (CSIC-ICMAT)
19. García Meixide, Carlos (CSIC-ICMAT)
20. García Portugués, Eduardo (Universidad Carlos III de Madrid)
21. García Vicuña, Daniel (Universidad Pública de Navarra)
22. Ginzo Villamayor, María (Universidade de Santiago de Compostela)
23. Gómez Casares, Ignacio (Universidade de Santiago de Compostela)
24. Gómez Vargas, Nuria (Universidad de Sevilla)
25. González Barquero, María del Pilar (Universidad Carlos III de Madrid)
26. Guerrero Lozano, Vanesa (Universidad Carlos III de Madrid)

27. Guillén, María D. (Universidad Miguel Hernández de Elche)
28. Guimarães Martins, Susana Rafaela (Universidade de Vigo)
29. Gutiérrez Botella, Jesús (Biostatech – Universidade de Santiago de Compostela)
30. Lago, Adrián (Universidade de Vigo)
31. Li, Zehang (Universidad Carlos III de Madrid)
32. Martín Chávez, Pedro (Universidad de Extremadura)
33. Martos Barrachina, Francisco (Universidad de Málaga)
34. Mascareñas, Alicia (Universidade da Coruña)
35. Minuesa Abril, Carmen (Universidad de Extremadura)
36. Monroy Castillo, Blanca Estela (Universidade da Coruña)
37. Núñez Lugilde, Iago (Universidade de Vigo)
38. Ortega Jiménez, Patricia (UC Louvain)
39. Perea Rojas-Marcos, Federico (Universidad de Sevilla)
40. Ramírez Ayerbe, Jasone (Universidad de Sevilla)
41. Rella, Jorge C. (Universidade da Coruña)
42. Rodríguez Acevedo, Iria (Universidade de Santiago de Compostela)
43. Rodríguez Ballesteros, Sofía (Universidad Miguel Hernández de Elche)
44. Rodríguez Barreiro, Marta (Universidade da Coruña)
45. Rodríguez Ramírez, Luis Alberto (Georg August University of Göttingen)
46. Saavedra Nieves, Alejandro (Universidade de Santiago de Compostela)
47. Scaparra, María Paola (University of Kent)
48. Segura Martínez, Paula (Universitat Politècnica de València)
49. Sinova, Beatriz (Universidad de Oviedo)
50. Solà Vilalta, Albert (Universitat Politècnica de Catalunya)
51. Terán Viadero, Paula (Universidad Complutense de Madrid)
52. Ugarte, Lola (Universidad Pública de Navarra)
53. Vitoriano, Begoña (SEIO, Universidad Complutense de Madrid)
54. Vivó Sánchez, Sergio (Ryanair)

Index of authors


SYSØIRM
2024