Universidad de Navarra | DATAI INSTITUTO DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

Abstracts' Book

VI CONGRESO CIENTÍFICO DE JÓVENES EN DISEÑO DE EXPERIMENTOS Y CIENCIA DE DATOS

Universidad de Navarra
5-7 June 2023

Editado por el Servicio de Publicaciones de la Universidad de Navarra.
DOI: https://doi.org/10.15581/028.00001
ISBN: 978-84-8081-767-7

Preface

The Institute of Data Science and Artificial Intelligence (DATAI) at the University of Navarra is organizing the “VI Scientific Congress of Young researchers in Experimental Design and Data Science (JEDE 6)” along with the III DATAI Scientific Conference, to be held on June 5th, 6th, and 7th, 2023 in Pamplona (University of Navarra).

The previous five JEDE meetings took place in Toledo in 2010, San Cristóbal de la Laguna in 2012, Pamplona in 2014, Salamanca in 2017, and Almería in 2021. Young Spanish and foreign researchers, many of them from Latin America, attended these meetings.

The main objective of this congress is the exchange of knowledge and experiences among young researchers in experimental design and data science from Spanish universities, as well as professionals in these fields of knowledge, through active participation in a gathering specifically tailored to them. Presentations will take place in a relaxed and receptive environment, with the presence of national and international experts who will encourage scientific debate.

Organizing Committee JEDE - 6

Organizing Committee

Rubén Armañanzas
Raúl Martín
Roberto Dorta
José A. Moler
Juan Manuel Rodríguez
Isabel Ortiz
Camelia Trandafir
Mabel Morales
Virginia Gracia
Carlos de la Calle
Edgar Benítez
Eduardo Iribas
Elena Martín
Horacio Grass
Alberto García
Álvaro Cía Mina
Marcos López
Ivan Cordón
Jose Luis Poveda
Juan Carlos Gamero
Leandro González
Pablo Urruchi
Montserrat Miranda
Rocío Santos
Stella Salvatierra

Scientific Committee

- Amparo Alonso Betanzos (Universidade da Coruña)
- Enrique del Castillo (The Pennsylvania State University)
- Nuria Oliver (Director - Data Scientist, Data Pop Alliance)
- John Stufken (George Mason University)
- Trevor Hastie (Stanford University)
- Chiara Tommasi (Universita degli Studi di Milano)
- Jesús López Fidalgo (President, Director - DATAI)

Contents

Conference Program

KL: Keynote Lecture, CT: Contributed Talk, FT: Flash Talk.

Monday, 05 of June

9:00–9:30	**Registration**		
9:30–10:00	**Welcome**		
10:00–11:00	KL	**Dr. Valerii Fedorov** Chairman: Juan M. Rodríguez	Best intention designs in dose-finding studies
11:00–11:30	**Coffee**		
11:30–13:30	CT	**SESSION I** Chairman: Stella Salvatierra	Artificial Intelligence and Applications
13:30-15:00	**Lunch**		
15:00–16:20	CT	**SESSION II** Chairman: Sergio Ardanza	Uncertainty Quantification
16:20–16:45	FT	**FLASH TALKS** Chairman: Rubén Armañanzas	Optimal design of experiments
16:45–17:15	**Coffee**		
17:15-18:15	CT	**SESSION III** Chairman: Montserrat-Ana Miranda	Optimization
19:00	**Reception**		

Tuesday, 06 of June

9:00–9:30	**Registration**		
09:30–11:00	CT	**SESSION IV** Chairman: Juan Carlos Gamero	Optimal design of experiments
11:00–11:30	**Coffee**		
11:30–12:30	KL	**Hugo Maruri-Aguilar** Chairman: Víctor Manuel Casero	Analysis of complex experiments
12:30–13:10	CT	**SESSION V** Chairman: Mabel Morales	Social Sciences
13:10-15:00	**Lunch**		
15:00–16:40	CT	**SESSION VI** Chairman: Carlos de la Calle Arroyo	Healthcare
16:40–17:00	FT	**FLASH TALKS** Chairman: Darian Horacio Grass	AI & Healthcare
19:00	**Guided tour**		
21:00	**Social dinner**		

Wednesday, 07 of June

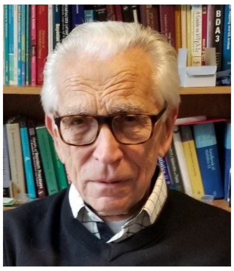
10:00–11:00	CT	**SESSION VII** Chairman: Edgar Benítez	Design of experiments
11:00–11:30	**Coffee**		
11:30–12:30	KL	**Mª Ángeles Gil Álvarez** Chairman: Licesio J. Rodríguez-Aragón	Análisis de cuestionarios con escala imprecisa
12:30–13:10	CT	**SESSION VIII** Chairman: Irene Mariñas	Longitudinal studies
13:10	**Lunch and Farewell**		

Valerii Fedorov

BEST INTENTION DESIGNS IN DOSE-FINDING STUDIES

Independent consultant involved in developing methods of optimal design of pharmaceutical research working. For decades statisticians applied optimal design methods in natural sciences and engineering. These methods worked almost flawlessly and led to impressive savings of time and other resources. However, their straightforward implementation in clinical trials may lead to unexpected situations and careful adjustment/modification is needed. Medical ethics, huge expenses, and interactions with regulatory agencies call for a more meticulous analysis of stochastic models and constraints. Often the initial enthusiasm about the (mathematical) efficiency of optimal designs quickly evaporates during the first discussion between a statistician and the medical team. I consider some promising compromises between often conflicting intentions to gain more information or to provide the best treatment to every subject participating in clinical trials.

Biography



During the last two decades, Valerii Fedorov was involved in developing methods of optimal design of pharmaceutical research working as an independent consultant; VP, Innovation Center, ICON plc; Quintiles, VP and the Head of Predictive Analytics; GlaxoSmithKline inc., the Head of the Research Statistics Unit. Most of the results developed during this period are presented in his monograph on Optimal Design for Nonlinear Response Models, 2014, CRC (coauthored with Dr. Sergei Leonov). Prior to that Valerii worked as the Senior Research Statistician at the Oak Ridge National Laboratory and lectured as a Visiting Professor of Statistics at the University of Minnesota. Before his USA career, Valerii Fedorov served as the Head of the Department of Mathematical Statistics at the Central Institute of Mathematical Economics of the Russian Academy of Sciences and as a senior researcher at the Laboratory of Mathematical Statistics of the Moscow State University. Valerii also lectured and conducted research as a visiting professor or as a visiting scholar at the Isaac Newton Institute for Mathematical Sciences, Cambridge; the Imperial College in London and the City University of

London, UK; Free University and Humboldt University in Berlin, Germany; Vienna University and the University of Economics in Austria and for five years at the International Institute of Applied System Analysis in Vienna, Austria. Professor Fedorov is the author of more than 200 publications including several books. His monograph on the Theory of Optimal Experiments, Academic Press, is one of the first monographs on optimal experimental design. He is an ASA Fellow, Honorary Professor of Cardiff University, UK, and Adjunct Scholar of the University of Pennsylvania, USA, elected member and former Council Member of the International Statistical Institute. In 2018 he initiated and chaired a Special Interest Group on Quantum Computing in Statistics and Machine Learning at American Statistical Association."

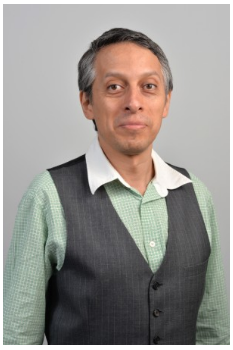
Hugo Maruri-Aguilar

ANALYSIS OF COMPLEX EXPERIMENTS

Computer simulations are widely used as substitute of experiments in situations where physical experimentation is costly or even impossible to do. My talk will describe through examples some of the challenges associated with analyzing data and creating models for simulation experiments. I will discuss through three examples: an infectious disease model, a model for a component of an engine and modelling motorway traffic data.

Biography

Hugo Maruri Aguilar is a professor in the Department of Statistics at the School of Mathematics of Queen Mary University in London. His areas of interest include computer experiments analysis, algebraic statistics, and penalized regularization in data modeling using likelihood.



María Ángeles Gil Álvarez

ANÁLISIS DE CUESTIONARIOS CON ESCALA IMPRECISA

One of the most appealing applications of the studies related to the Analysis of Imprecise Data (in particular, fuzzy-valued ones) is that of analyzing from a statistical pespective the responses to questionnaires where responses are based on fuzzy rating scales. The talk will succintly present the methodological developments that have been already made and the sound probabilistic bases for them. It will examine also in a rather comparative way the advantages of using these scales in contrast to some traditional ones (namely, Likert-type scales, visual analogue scales or fuzzy linguistic ones) and will illustrate their potentiality with a real-life example in which respondents are 9-year-old children and implications refer to education policies.

Biography

María Ángeles Gil Álvarez (BSc-Math and MSc-Math, University of Valladolid, Spain, PhD-Math University of Oviedo, Spain) first focused her research on Statistical Information and, for the last decades, she is involved in researching on the Statistical Analysis of Imprecise Data (more concretely and mostly, fuzzy-valued and interval-valued data) and its application, among others, to deal with questionnaires. Gil received the Silver Medal of Asturias in 2014, and the SEIO Medal of the Spanish Statistics and Operations Research Society in 2021. She was named a Fellow of the International Fuzzy Systems Association in 2015, elected to the Academia Asturiana de Ciencia e Ingeniería in 2021, and member of the Spanish Royal Academy of Sciences since January 2023.

SESSION I: Artificial Intelligence and Applications

Spectral-spatial classification of hyperspectral imagery using deep learning algorithms for metal sorting: Preliminary results

Pérez Reina, A., Borro Yagüez, D.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Hyperspectral Imagery (HSI) has emerged as a powerful tool for metal classification, offering unique potential for various applications in the mining and recycling industries. Despite the increasing interest in applying deep learning techniques to remote sensing and agro-food spectral data, one of the main challenges faced in this domain is the scarcity of labeled data. This issue becomes particularly pronounced in the context of metal sorting, where studies and available data are limited. To address this challenge, we leverage the hyperspectral properties of materials, as each substance exhibits a distinct spectral fingerprint that can be identified using HSI technology. By capitalizing on these unique characteristics, we construct a comprehensive database of metal spectral fingerprints. In this study, we employ state-of-the-art deep learning algorithms, such as 3D Convolutional Neural Networks (CNN) and transformers, to perform spectral-spatial metal sorting. This spectral-spatial classification approach harnesses both the spatial and spectral properties of the hypercube, resulting in improved performance. By integrating spectral information with spatial context, our method not only recognizes the unique spectral signatures of different metals but also accounts for the inherent spatial relationships among neighboring pixels, thereby enhancing classification accuracy. We demonstrate the effectiveness of our proposed approach by conducting extensive experiments on a range of metal samples. Furthermore, we highlight the benefits of incorporating both spectral and spatial information into the classification process, which ultimately leads to superior performance in metal sorting applications. This work contributes to the growing body of research on the application of deep learning algorithms for metal classification using HSI. Our findings underscore the potential of hyperspectral properties for addressing the lack of labeled data in this field and pave the way for future advancements in spectral-spatial classification techniques. The proposed methodology holds promise for various real-world applications, including mineral exploration, waste management, and recycling processes, ultimately fostering more efficient and sustainable practices within these industries.

Machine learning techniques for financial market risk measurement

García Muñoz, LM
BBVA Bank

In this talk, we will compare two neural network (NN) architectures: deep neural networks and differential machine learning, and their potential applications in reducing computational challenges in market risk under the Fundamental Review of the Trading Book (FRTB) regulatory framework. We will begin with a brief introduction to market risk, which refers to the risk of financial loss resulting from changes in market prices, such as fluctuations in interest rates, exchange rates, and stock prices. We will then provide an overview of the FRTB regulatory framework, which was introduced by the Basel Committee on Banking Supervision in 2016 to strengthen banks' market risk capital standards and address weaknesses identified during the 2008 financial crisis. We will explain how neural networks can help reduce the computational challenge of computing risk metrics under the FRTB framework by approximating payoffs for complex financial products. We will provide insights on how the NN works and demonstrate its potential effectiveness in reducing computation time. Next, we will discuss how metrics may behave for hedged portfolios and how NNs can help improve the accuracy of these metrics. We will also highlight potential areas for further research and emphasize the importance of continued development and improvement of NNs in the field of market risk management. Overall, our talk aims to provide an overview of the potential benefits of NNs in market risk management under the FRTB regulatory framework and to spark further discussion and exploration of these techniques in future research.

An approach to a cyberattacks management system

Gutiérrez Galeano, L., Domínguez Jiménez, J.J., Medina Bulo, I.
Universidad de Cádiz

Cybersecurity is a very necessary area due to the new types of cyberthreats that are constantly appearing. However, there are not enough tools and techniques to keep systems safe from cyber-attacks. Therefore, this work aims to develop a system that, based on network traffic, not only detects ongoing cyber-attacks in real time but is also able to predict them sufficiently in advance so that action can be taken in time to neutralise them. We hypothesize that such a system would make it very difficult for cybercriminals to achieve their goals and would reduce the risk of cyberattacks. This goal could be achieved by (1) selecting a sufficiently good dataset, which contains a wide variety of the most current and popular types of attacks, and contains enough data to be able to build a good system; and (2) using different artificial intelligence techniques. The use of these techniques requires sufficient data to train a successful model for the system to be able to predict well enough. Therefore, it will also be necessary the design of a network traffic simulator to inject both benign and malign network packets.

Wood heterogeneous grain classification for wine barrels: a cropping data augmentation approach

Ricardo, F.A., Eizaguirre, M., Borro, D.
CEIT and TECNUN School of Engineering, Universidad de Navarra

The technology for inspecting wood is essential in many facets of contemporary industry. Among other issues, the number of rings in a stave has a direct relationship with the wood quality. The appearance of sawn wood has many natural variations and distinct appearance that a human inspector can easily compensate when determining the type of each stave or board. However, for automatic wood inspection systems, these variations are a major source of complication. Several approaches to an automatic detection of tree-ring boundaries exist; however, they use basic image processing techniques. As a result, their accuracy is limited, and their application is restricted mainly to wood where the tree-ring boundaries are clearly defined. There also exists some works based on segmentation deep learning techniques but again, the wood processed has ring boundaries easily detectable. The aim of this paper is to deal with the problem of wood classification when there is noticeable heterogeneity in the texture of the samples. To the authors' best knowledge, this is the first approach to grain classification in such heterogeneous images. The solution uses an hybrid approach combining classic computer vision for pre-processing and deep learning-based algorithms in order to classify wood into three quality categories. Cropping data was used in order to augment the original dataset, to avoid intra-class problem that appears in single staves and to improve the performance and accuracy of final voting system.

Computer Vision and Deep Learning based road monitoring towards a Connected, Cooperative and Automated Mobility

Iparraguirre, O., Brazalez, A., Borro, D.
Ceit and Tecnun (University of Navarra)

The future of mobility will be connected, cooperative and autonomous. All vehicles on the road will be connected to each other as well as to the infrastructure. Traffic will be mixed and human-driven vehicles will coexist alongside self-driving vehicles of different levels of automation. This requires both the physical and digital infrastructure to be ready to embrace the new mobility paradigm. Perception plays an important role in this challenge, as dynamic information from the environment allows active decision-making to act on the control of a car in real-time along with an off-line analysis of the infrastructure conditions to identify possible improvements. In this thesis, an AI-based road monitoring system is designed that performs automatic auscultation of the status of the infrastructure and detects safety-critical events for driving. This system focuses mainly on the monitoring of road markings and road signs plus the level of visibility due to atmospheric phenomena, two fundamental elements to ensure safe and autonomous driving.

Econometrics meets AI: Predicting Airbnb prices using Machine Learning

Jaén Delgado, J.
Universidad Carlos III de Madrid

Predictive models for Airbnb rental prices are built using Data Mining techniques and Machine Learning & Deep Learning algorithms. The goal is to provide an accurate and scalable statistical toolkit for economic agents to make optimal decisions in the Big Data era. NLP and Computer Vision models are utilized to analyze unstructured information, making it possible to combine text and images with tabular data: face recognition and sentiment analysis tasks were performed. Algorithms such as MICE and k-NN are used to impute missing data and overcome the potential lack of information that would stem from discarding observations with missing values. Bayesian Inference techniques were applied as to obtain best performing hyperparameters and measure uncertainty around AI models' predictions. It is found that Artificial Intelligence models outperform traditional econometric methods at the expense of lower interpretability. This drawback is tackled by resorting to Explainable AI (XAI) methods, which break the black box problem related to AI algorithms. It is then posible to interpret feature contributions to the predicted value at both, global and local level, being able to recognize important patterns for selected observations. A web app was created as to simulate how this service would perform in production, allowing economic agents to access AI models' predictions in an user-friendly fashion.

SESSION II: Uncertainty Quantification

Aggregation of information and structures in fuzzy logic

Talavera, F.J., Elorza, J.
Universidad de Navarra

Aggregation of information is key in many areas of mathematics and artificial intelligence, such as decision making or approximate reasoning. In many cases, the data being aggregated has a certain structure that is useful to maintain. In particular, in fuzzy logic, many advances have been made in this field regarding different types of functions such as T-indistinguishability operators or fuzzy quasi-metrics. In particular, we study when an aggregation operator acting on n T-subgroups preserves the structure of T-subgroup. First it is important to consider that there are two known definitions applicable to the aggregation of fuzzy structures: on products and on sets. Because both definitions provide important but different information on the aggregation of T -subgroups, we perform a study on the preservation of T -subgroup properties in both situations. We deduce that in some cases, the operators that preserve T -subgroups are the same with both definitions and in other cases they are not. We also conclude that this difference is influenced by the structure of the subgroup lattice of the ambient group.

Uncertainty reduction and quantification through data, machine learning and mathematical modelling

López-De-Castro, M.
Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Effective management and quantification of uncertainty plays an essential role in the development of mathematical and predictive models. This role is highlighted when the predictions are performed in a high-risk domain such as medicine, wildfires or finance. In this short talk, we introduce two common sources of uncertainty commonly observed in real-world problems from a case-study perspective. The first source arises from the lack of data updating and data quality. The benefits from the application of machine learning-based procedures to overcome this problem will

be presented in a real-world case context. The second common source of uncertainty present in real world problems usually arises from the inherent complexities of the system of interest. This kind of uncertainty usually emerges when we are modelling highly non-linear and multi-scale problems. Accurate characterization of the uncertainty source, coupled with solid methods such as the ensemble forecasting technique, is crucial for realistic simulations of the system's behavior, the extraction of high-quality data and decision making.

Uncertainty Assessment of Algorithmic Prediction in Digital Medicine

Armañanzas Arnedillo, R., García Galindo, A., López De Castro, M.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Translational medicine seeks to combine diverse scientific disciplines to enhance prevention, diagnosis, and treatment of clinical conditions. It is by definition a highly interdisciplinary field whose main goal is to improve the global healthcare system. Although most biomedical disciplines have evolved to incorporate digital devices, the algorithms that collect, process, and analyze data are overwhelmingly unknown and/or inaccessible to most practitioners. These computer-based systems, often referred to as black box programs, are already providing physicians with risk measures for their patients developing illnesses. In the long run, these models will include more and more information and become highly complex. This complexity will make it increasingly critical to bring the methods closer to the professionals through the implementation of simple user interfaces and uncertainty quantifications of the predicted values coming from the algorithm.

Assessing response to neoadjuvant therapy in breast cancer using conformal prediction

García-Galindo, A., López-De-Castro, M., Armañanzas, R.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Neoadjuvant therapy (NAT) is recognized as the most effective preoperative treatment for decreasing tumor burden in breast cancer. However, the patient's condition and clinical factors greatly influence the tumor's pathological response. To improve personalized medical care plans, it is crucial to develop modelling tools that predict a patient's response to NAT. Recent studies have demonstrated the promising outcomes of machine learning (ML) techniques in breast cancer prognosis, combining imaging and molecular features obtained from biopsy analyses. Our study introduces an uncertainty-aware ML model to predict response to NAT in two sequential stages. Firstly, we trained a pre-treatment dynamic contrast-enhanced magnetic resonance imaging model. Secondly, a molecular biomarker-enriched dataset is used

to build a second model. To identify patients with large uncertainty in predicted responses, we propose integrating the Conformal Prediction (CP) framework in the first non-invasive model. These patients can then be referred to the second model that includes data from invasive tests, reducing the need for unnecessary biopsies. We explored various alternatives for the standard ML algorithms and CP methods on a publicly available clinical dataset. Our results demonstrate the potential of our uncertainty-aware clinical predictive tool in real-world scenarios.

FLASH TALKS I: Optimal design of experiments

Designing to discriminate between two random effect models

Casero-Alonso, V., López-Fidalgo, J., Pozuelo-Campos, S., Tommasi, C., Wong, WK.

Universidad de Castilla-La Mancha

Se está trabajando en el problema de obtención de diseños que permitan discriminar entre dos modelos de efectos aleatorios. La idea de la presentación es plantear el problema en general y particularizar a 2 modelos polinómicos fraccionales (FP) ilustrándolo con ejemplos y dejando claro el problema que resuelven los diseños para discriminar entre modelos.

Diseños para la estimación conjunta de la dosis objetivo y el riesgo de sesgo.

Flournoy, N., Moler, J., Plo, F., Hyum, W.

Universidad Pública de Navarra

En la fase I de los ensayos clínicos se busca estimar la dosis de un fármaco con la que se alcanza una toxicidad aceptable. La importancia de estimar conjuntamente la dosis objetivo junto con su pendiente en la función dosis-respuesta es bien reconocida en la literatura. La razón es que la pendiente actúa como un indicador del riesgo que conlleva el sesgo en la estimación puesto que pendientes altas indican un alto riesgo ético en la sobreestimación de la dosis y de eficiencia en la subestimación. Proponemos diversos modelos que se han estudiado en este contexto junto con sus ventajas y desventajas.

Diseños óptimos para el modelo Gamma simplificado

Santos Martín, MT, Rodríguez Díaz, JM, Mariñas del Collado, I
Universidad de Salamanca

El modelo más usado para describir la eliminación de la concentración de alcohol en sangre asume una cinética de orden cero, es decir, que el alcohol se elimina a una tasa constante. Sin embargo, este modelo no considera la fase de absorción en la que la concentración de alcohol aumenta hasta alcanzar un pico máximo. Existen modelos alternativos que incluyen ambas fases como son los compartimentales que elevan la complejidad y dificultan su uso práctico. En este trabajo, se propone un modelo Gamma simplificado que ajusta las diferentes fases y que disminuye la complejidad de los anteriores modelos. Se han calculado los diseños D-óptimos para este modelo y se han comparado con diseños empleados en otros trabajos. Además, se han estudiado diseños equiespaciados, al ser de uso más común entre los científicos que prefieren tomar muestras que cubran todo el intervalo del diseño en lugar de restringirse a los habitualmente pocos puntos contenidos en los diseños óptimos. También se han calculado las eficiencias de estos diseños en comparación con los óptimos. En todos los casos, será necesario asumir una estructura de covarianza entre las respuestas, aunque se ha estudiado además el caso de observaciones independientes para poder comparar los resultados con los trabajos existentes en la literatura.

Maximum and Bayesian optimal designs

Tommasi, C., Rodríguez-Díaz J.M., López-Fidalgo J.
Universidad de Salamanca

In optimal experimental design setup maxi-min efficiency criteria are sometimes the best option, since enable us to take into consideration several tasks expressed by different component-wise criteria. However, they are difficult to manage because of their lack of differentiability. The connection between maxi-min efficiency and Bayesian optimality (which is differentiable) has been already explored in literature, but always center on specific problems and/or optimality criteria. In this study, we prove a more general version of the equivalence theorem, because it covers any multi-objective problem that can be expressed as a minimum design efficiency (for any component-wise criteria). Furthermore, a method to determine the prior probability that matches the maxi-min efficiency criterion and the Bayesian optimality is provided, which allows the application of the equivalence theorem.

Diseño óptimo de la supervivencia celular a la radioterapia

Rivas-López, MJ y Rodríguez-Díaz, JM
Universidad de Salamanca

El crecimiento de las células orgánicas se ve alterado al aplicarse una radiación ionizante. En esto se basa la radioterapia que, junto con la quimioterapia y la cirugía, es fundamental en los tratamientos oncológicos. Tras irradiar un tejido, sus células pueden sufrir ciertos daños en el ADN y si no consigue arreglarlos morirá. A la hora de modelar la supervivencia celular ante una radiación ionizante se utiliza un modelo exponencial con exponente lineal-cuadrático. Se muestra porqué se utiliza este modelo y el diseño óptimo para la estimación de sus parámetros.

SESSION III: Optimization

Online data repair towards demographic parity implemented in Python

M. de Diego, E., Gordaliza Pastor, P., López Fidalgo, J.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Automated decision-making systems are increasingly used in various domains such as healthcare, recruitment, and justice, which has made the intersection between AI and ethics a crucial issue in recent years. Fair learning has established itself as a very active area of research which tries to ensure that predictive algorithms are not discriminatory towards any individual at individual or group level, based on personal characteristics such as race, gender, disabilities, sexual orientation, or political affiliation. Recent statistical approaches have focused on data repairing methodologies that map conditional distributions of each sensitive group towards their Wasserstein barycenter. While these pre-processing methods are effective, they are data-dependent bias mitigation mechanisms imposing a limitation in a production ML process. As time progresses, additional data will become available, potentially acquired within a different socio-economic context and therefore there exists the opportunity of generate predictions based on this new information by retraining the AI model. This work proposes a novel pipeline which integrates an efficient algorithm for repairing the incoming data. The purpose is to achieve fairness by using an extension of the empirical optimal transport map to new data. The procedure interpolates the repaired values for new data without recomputing the discrete optimal transport map for each new observation, achieving a computational complexity of order $\mathcal{O}(mn\epsilon^{-1})$ if the new set of points is an m-sample, n is the number of samples of each group and ϵ is a parameter of the interpolation map. An efficient open source implementation in Python language of the algorithm is provided, and several experiments show that the proposed method is promising in bridging the gap between continuous and empirical transport.

An Experimental Comparison of Metaheuristics for the Bi-objective Resource-Constrained Project Scheduling Problem with Time-Dependent Resource Costs

Rodríguez Ballesteros, S., Alcaraz, J., Anton Sanchez, L.
Universidad Miguel Hernández de Elche

The bi-objective resource-constrained project scheduling problem with time-dependent resource costs is to schedule a set of activities subject to precedence and resource constraints such that the make span and the total cost for resource usage is minimize. Precisely, costs are determined by the resource being considered together with the time it is used. This generalization of the traditional resource-constrained project scheduling problem has garnered significant interest as it succeeds in meeting a wide range of real-world demands. In such a multi-objective context, solving the aforementioned problem imply some challenge, as both objectives are of conflict to each other, giving rise to a set of trade-off optimal solutions, commonly known as the Pareto front. Given that many medium or large-sized instances of this problem cannot be solved by exact methods, it makes necessary the development of metaheuristics for approximating the Pareto front. To attain this issue, seven multi-objective evolutionary algorithms (MOEAs) have been implemented to solve this problem and then, an exhaustive comparison of their performance has been carried out. Metaheuristic algorithms typically yield an approximation of the Pareto optimal front, prompting the question of how to assess the quality of the obtained approximate fronts. 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 important differences among the performance of the metaheuristics evaluated.

Ranking rules for Pareto pruning in multi-objective problems

Suárez Dosantos, P., Mariñas-Collado, I., Bouchet, A., Montes, S.
Universidad de Oviedo

Multi-objective problems (MOP) have gained importance during the past few decades. The creation of Pareto sets is an often-used approach for solving these types of problems. However, since the size of MOP grows exponentially in relation to the amount of input, there is a need to decrease, or at least order, Pareto sets. In this work, the main goal is to identify preferences among the best non-dominated options by using ranking methodologies.

SESSION IV: Optimal design of experiments

Optimal subdata selection for prediction based on the distribution of the covariates

Cia-Mina, A., López-Fidalgo, J.
Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

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. 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. Theoretical results are provided to justify the criterion as well as an interpretation from usual linear optimality criteria. As expected by the theory it shows a reduction in the prediction MSE compared to other existing methods. The performance of the new approach is illustrated with simulations.

Optimal Experimental Design Applied to Predictive Microbiology

Muñoz del Río, A., Casero-Alonso, V., Amo-Salas, M.
Universidad de Castilla-La Mancha

Predictive microbiology studies, through mathematical modelling, the behaviour of microorganisms growing in food. This work applies Optimal Experimental Design theory to Baranyi model, one of the most common models used in predictive microbiology. D-optimal designs are obtained and, considering the complexity of the model, a parameter sensitivity analysis is conducted, showing that D-optimal designs are sensitive to two of the parameters. In view of these results, a methodology is proposed to augment the D-optimal design to improve its robustness.

Optimal designs for detecting and characterizing hormesis in toxicological tests

Pozuelo Campos, S., Casero Amo Salas, M Alonso, V.
Universidad de Castilla-La Mancha

Toxicological tests are experiments that show the effects of a toxic on organisms, ecosystems, etc. This study focuses on tests in the aquatic environment, in which the test involving Ceriodaphnia Dubia organism stands out. The literature indicates that in two out of every three experiments carried out with this organism, there is hormesis. This study applies optimal experimental design theory to a linear quadratic model with a Poisson distribution for the response, in order to obtain designs that allow efficient detection and characterization of hormesis. To this end, a variety of utility functions are used, including the dose for the zero-equivalent point, the area under the curve, the dose at which maximum response is reached or the dose at which there is a given relative inhibition with respect to the control or the maximum. A study of cross efficiencies of the calculated designs shows the importance of correctly defining the goal of the experiment, in order to obtain the most appropriate design.

Optimal experimental design with optedr

de la Calle-Arroyo, C., López-Fidalgo J., Rodríguez-Aragón, L. J.
Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

Experimental Design is an essential stage of experimentation, prior to obtaining information about a phenomenon, which indicates the best way to carry out observations based on the objective of the experiment. In optimal experimental design, we start with a model for the relationship between the variables that we want to measure. Finding optimal designs is generally not a simple task. Most of the time, it is not possible to find analytical results, so algorithmic techniques are often used. This is the main motivation behind the optedr package, which allows for the calculation of optimal designs for nonlinear models with one independent variable. It is also possible to specify the probability distribution of the response variable within the exponential family and calculate optimal designs for different optimality criteria. This package has been developed with an applied approach, providing a straightforward handling to design generation. The package also allows for comparing user-proposed designs with other optimal designs based on their optimality criterion. In addition to calculating optimal designs, the functionality of augmenting designs has been added. This procedure addresses the issue that often occurs in practice where a design is optimal according to a criterion, but it is still not suitable for real-world use. Experimenters may have specific needs or constraints, preferences for certain experimental points, statistical requirements, etc. In such cases, when it is not possible to directly implement the optimal design, one can choose to use the optimal design as a benchmark or modify the design to fit the preferences of the user. Finally,

since the package works with approximate designs, a rounding algorithm has been implemented to transform the approximate optimal designs and augmented designs into exact designs ready for use by the experimenter.

SESSION V: Social Sciences

Model-based estimation of small area dissimilarity indexes: An application to sex occupational segregation in Spain

Bugallo, M., Esteban, M.D., Morales, D., Pagliarella, M.C.
Universidad Miguel Hernández de Elche

This paper introduces a new statistical methodology for estimating Duncan dissimilarity indexes of occupational segregation by sex in administrative areas and time periods. Given that direct estimators of the proportion of men (or women) in the group of employed people for each occupational sector are not accurate enough in the considered estimation domains, we fit to them a three-fold Fay-Herriot model with random effects at three hierarchical levels. Based on the fitted area-level model, empirical best predictors of the cited proportions and Duncan segregation indexes are derived. A parametric bootstrap algorithm is implemented to estimate the mean squared error. Some simulation studies are included to show how the proposed predictors have a good balance between bias and mean squared error. Data from the Spanish Labour Force Survey are used to illustrate the performance of the new statistical methodology and to give some light about the current state of sex occupational segregation by province in Spain. Research claims that there is a sex gap that persists despite advances in the inclusion of women in the labour market in recent years and that is related to the unequal sharing of family responsibilities and the stigmas still present in modern societies.

Why we socially interact? The need, the ability, and the desire

Fernández, A., Sádaba, C., García-Manglano, J., Vanden Abeele, M.
Universidad de Navarra - Ghent University

Each interpersonal relationship is formed and maintained through continuous social interactions in everyday life, which are the key unit of interpersonal communication. Their occurrence is the result of a complex balance between satisfying the fundamental need to belong and managing human social energy. In simple words, we interact based on three factors: the need, the ability, and the desire to interact. The need. Social interactions have a core relationship constituting function, making it present in everyday life while "talking" with others in normal conversations. We interact in response to our need for frequent, affectively pleasant interactions

with a few other people. The ability. We continuously engage in social interactions, being willing to expend the social energy needed to develop our relationships and feel belonging; trying to use as little as possible and to engage in those interactions that give us the greatest return on our expenditure. The desire. There is a homeostatic principle between interaction and non-interaction moments: as long as the momentary need to belong arises, the desire to be alone decreases, impacting on interactions occurring. Preregistered hypotheses In our preregistered study (https://osf.io/z374v), we explored the previous three dimensions that explains why we interact using multilevel time series models. Within-individual associations to explain why we interact considering the in-the-moment role of the need (closeness and loneliness), the ability (energy level), and the desire (to be alone), our preregistered hypotheses were mainly within-individual, analyzing mainly the average within-individual associations. Between-individual differences in within-individual associations When looking at between-individual differences, we suggested that the associations between the occurrence of social interactions and the predictors, as described in our within-individuals hypotheses, vary significantly between individuals. Sample and study design One hundred and thirty-one young adults make up the sample (18-25 years old, Mage = 20.69, SDage = 2.09, women = 62.6%). For 14 days, participants answered 7397 momentary questionnaires, registering 4616 social interactions (62.4%) and 2781 solitude moments. The project codebook is available at OSF (https://osf.io/kawhn). To test the effect of the time interval between questionnaires we used a second, larger dataset with similar characteristics (257 participants, 21158 momentary questionnaires answered, and 12937 social interactions registered). Preregistered analysis plan We analyzed concurrent and lagged within-individual associations between social interaction occurrence and closeness, loneliness, social energy and desire to be alone through four bivariate multilevel vector autoregressive models using Dynamic Structural Equation Modeling in Mplus (version 8.9). To test the hypothesis, variables were decomposed in their within- and between-individual components. In the preregistration we included all the specifications of the models and the actions to be taken in different scenarios; we applied some preregistered actions to converge the models. Main results Considering the in-the-moment need to belong, the average within-individuals associations show a different direction than expected: closeness is positively associated with interaction occurrence (95% CI: [0,250; 0,357]) whereas loneliness is negatively associated (95% CI: [-0,242; -0,127]). Conversely, energy (95% CI: [0,062; 0,150]) and desire to be alone (95% CI: [-0,336; -0,232]) are, as expected, positively and negatively associated with social interaction occurrence.

SESSION VI: Healthcare

Assessment of weights matrix alternatives in the spatial conditional autoregressive modelling of COVID-19 data

Morales-Otero, M., Faes, C., Núñez-Antón, V.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

In this work, we study the geographical spread of COVID-19 cases in the municipalities of the Flanders region in Belgium during the period going from September 2020 to January 2021, focusing on the search of the spatial structure which best accommodates the spatial correlation in this dataset. In order to be able to fit these data, we consider the spatial conditional overdispersion models, which assume the incidence of cases is conditional on the incidence of cases in the other neighboring regions. Furthermore, we also propose an extension of these models based on the geometric mean of the incidence rates. These models offer great flexibility, allowing us to incorporate any spatial structure in a very simple and direct way, also providing the possibility to obtain a straightforward interpretation within the context of the specific dataset under analysis. Results suggest the presence of a strong spatial correlation in the data, which is best explained by the distance band spatial weights matrix. This implies that, for the data under study, the underlying spatial process is well explained and modelled by this spatial structure. Nevertheless, in order to further investigate the performance of the proposed methods when the correlation among the regions is given by another connectivity pattern, such as the mobility of the individuals among regions in a given time period, a simulation study was carried out, where we induce correlation in the response variable based on the mobility matrix, and we have been able to appropriately verify that the models are able to identify the correct spatial structure for most of the cases under study.

A physiologically-based platform to predict organ exposure for colorectal cancer drugs in clinical settings

Peribáñez – Domínguez, S., Parra-Guillen, ZP., Pascoal, S., Díez Punzano, R., Fernández Troconiz, I.

Universidad de Navarra

Introduction: Systemic treatment with chemotherapy is usually chosen in cases of metastatic cancer without primary tumor surgery. Irinotecan, 5-fluorouracil, oxaliplatin, and the co-adjuvant, leucovorin, are drugs used in the treatment of cancer diseases. Several studies have described the pharmacokinetics of these drugs however, less is known about their disposition in different organs. Understanding the distribution of antineoplastic drugs in the body is of great relevance, particularly in those organs where the tumor is located and grows. Physiologically-based pharmacokinetic models (PBPK) aim to integrate mechanistically body physiology and drug physicochemical attributes to describe/project both systemic and tissue drug longitudinal exposure based on the treatment specificities. These models represent an opportunity to personalize the treatment based on the individual patient phenotypes minimizing the risk/benefit therapeutic ratio. This project aims to describe the systemic exposure and generate tissue exposure of three different anti-cancer drugs (irinotecan, 5-fluorouracil, oxaliplatin) and one co-adjuvant (leucovorin) through the building of physiological-based pharmacokinetic models and using literature data. Materials and methods: A literature search was performed to collect pre- and clinical data on plasma concentration-time profiles of each drug. Data were scanned using the tool WebplotDigitalizer. Each dataset was built and explored using R version 4.0.5 through RStudio interface version 1.4.1106. Physicochemical features were obtained from the literature for each drug as well as parameters associated with metabolizing and elimination processes. PBPK models were built in PK-Sim®(Open Systems Pharmacology Suite 11) software by including the parameters learned previously. Models were validated by contrasting model predictions with observations from datasets. Model-based simulations were performed to evaluate the magnitude of drug exposure under different treatment scenarios. Results: Data from 17 clinical studies (including 5 studies of oxaliplatin, 4 of 5-fluorouracil, 5 of leucovorin, and 3 of irinotecan) were combined in one unique dataset for each drug and normalized by dose administration. Irinotecan PBPK model involved a metabolizing process carried out by CYP3A4. Additionally, this model included renal and biliary clearance elimination routes. For the case of 5-fluorouracil, the model comprised a metabolism elimination governed by the enzyme dihydropyrimidine-dehydrogenase. Oxaliplatin model incorporated an unspecific hepatic clearance elimination process. Finally, the model developed for leucovorin included a metabolizing process elicited by the enzyme 5,10-methylenetetrahydrofolate-reductase and renal excretion. All models developed described successfully the exposure vs time profiles with respect to both, the typical tendency and dispersion shown by the literature data. Simulations with different scenarios and dosing regimens were carried out to evaluate drug exposure both in plasma and in the different organs of the body. Conclusions: The bottom-up approach has been used to describe successfully the exposure vs time profiles of three anti-cancer drugs and one co-adjuvant.

GeNNius: An ultrafast drug-target interaction inference method based on graph neural networks

Veleiro, U., de la Fuente, J., Serrano, G., Pizurica, M., Pineda-Lucena, A., Vicent, S., Ochoa, I., Gevaert, O., Hernaez, M.
CIMA Universidad de Navarra

Drug-target interaction (DTI) prediction is a relevant but challenging task in the field of drug discovery. In-silico approaches have drawn special attention as they can reduce associated costs and time commitment of traditional methodologies. Yet, current state-of-the-art methods present limitations; existing DTI prediction approaches are computationally expensive, thereby hindering the ability to use large networks and fully exploit available datasets. Further, the generalization to unseen DTI datasets of DTI prediction methods has not yet been explored. We introduce GeNNius (Graph Embedding Neural Network Interaction Uncovering System), a Graph Neural Network-based method that outperforms state-of-the-art models in both accuracy and time efficiency. GeNNius consists of a Graph Neural Network (GNN) followed by a 2-layer Neural Network (NN)-classifier. The GNN (i.e., the encoder) consists of four SAGEConv layers, responsible for generating node embeddings by aggregating information from each node's neighborhood. Afterward, an NN-based classifier aims to learn the existence of an edge given a set of concatenations of drug and protein node embeddings; at this step, negative edges, i.e., non-interacting pairs, are generated. GeNNius yields an outstanding performance in terms of AUC and AUPR while demonstrating ultrafast training and testing and showing stable results in independent runs. Its performance was compared to state-of-the-art DTI models and off-the-shelf machine learning baselines. GeNNius outperformed all benchmarked models, not only in the evaluating metrics but also with a significant decrease in execution time. Specifically, the running time was hundreds of minutes faster than most models when employing DrugBank, the largest dataset. In addition, we demonstrated the generalization capability of GeNNius to unseen datasets, preventing an oversimplification of the task by excluding from the training procedure those DTIs shared to both the train and test datasets. Also, our results showed that the presented methodology improves the DTI prediction task, i.e., when training in larger datasets and testing in a smaller one, the AUC increased considerably. Furthermore, we confirmed the prediction power of uncovering new interactions by evaluating unannotated DTIs for each dataset. DTI datasets have been released in different years, incorporating data from various sources. Therefore, some drug-target pairs have been recently classified as DTIs, while they remain non-interacting in others. We used those edges to assess the predictive power of our model. GeNNius displayed compelling inference abilities by detecting more than 80% of DTIs in the largest datasets (more than 90% for DrugBank). Finally, we investigated qualitatively the embeddings generated by GeNNius, revealing that the encoder maintains biological information while diffusing this information through nodes, eventually distinguishing node embedding by protein families. A better understanding of this capability may promote the reliability of machine learning approaches for guiding experimental validations. In summary, we present GeNNius, a

DTI prediction model that outperforms state-of-the-art models while being several orders of magnitude faster. GeNNius's ability to generalize and predict novel DTIs reveals its suitability for drug repurposing. However, its remarkable speed is the main contribution to its usability by enabling the validation of several drug-target pairs in less than an hour, achieving a significant milestone in the field.

Voltage Mapping Study of Patients with Atrial Fibrillation: Regional Statistics

Moriones L., Ravassa S, Bragard J.
Universidad de Navarra

Atrial fibrillation (AF) is one of the most prevalent cardiac pathologies, associated with ageing, a high risk of morbidity and mortality, and high social and health care costs due to clinical complications and treatment costs. Therefore, there is a medical need to accurately phenotype the electroanatomical alterations of the left atrium present in patients with AF. This work proposes to characterise atrial electrical remodelling using high-density voltage mapping (HDvM). Using a recent technique, the left atrium of each of the 122 patients participating in the study was divided into 24 ”standard” regions. The proposed methodology takes the regionalisation into 24 regions and includes it in the original geometry of the atrium. Possible controversies that may exist are studied, such as the error produced in the translation of regions from 2D to 3D. The method provides a clean-up to correct this anomaly based on neighbourhood rings. These rings are the determining point to establish the correction criteria in those anomalous geometries. Finally, to determine that the proposed method is correct in its use and the extracted data are satisfactory for further regional statistical analysis.

Bioinformatic tools to analyze cell infiltration within the tumor microenvironment along with the clinical outcome of patients

Alonso-Moreda, N., González-Velasco, O., Berral-González, A., Sánchez-Santos, J.M., De Las Rivas, J.
Universidad de Salamanca

Tumors are composed not only of malignant cells but are also involved in a complex Tumor Microenvironment (TME), a dynamic of interactions between malignant and normal cells. The TME includes immune cells (T and B lymphocytes, NK cells, monocytes, neutrophils, and DCs), that can recognize, control, and stop tumor development, as well as stromal cells (fibroblasts, adipocytes, and endothelial cells), which provide support and structure to organs, glands, and tissues of the organism. It seems that the level of immune and stromal cell infiltration within the TME has an impact on tumor progression and response to immunotherapy treatment. However, this effect is not the same for all tumor types. Specifically, immune, and stromal cells can create an immunosuppressive environment in colon cancer, hindering the anti-tumor response. In contrast, most breast cancers contain many immune cells in the TME, which is a good prognostic factor. For these reasons, understanding the infiltration of non-malignant cells within the TME according to the genes that characterized them (biomarkers or gene signatures) is also essential to predict the individual immune response to such antitumor treatments based on the immunological features of the patients. In our work, we focused on the analysis of immune and stromal cells present in the TME and their gene signatures by studying their transcriptomic profiles. For this purpose, we collected RNA expression data from biological samples of Breast and Colorectal Cancer (BRCA and CRC) and we applied deconvolutional algorithms, mathematical and bioinformatics tools to decompose complex cell mixtures and identify the primary components (cell populations) contained inside them. Firstly, we used CIBERSORT and quanTIseq to study the variability of immune cells present in the TME of BRCA luminal samples. Then, immune and stromal cell infiltration were estimated by xCell and ESTIMATE methods to reveal the dynamic cell changes between the different tumor samples. Finally, a Kaplan-Meier model was fitted to compare patients with high and low levels of immune and stromal cell infiltration, to analyze their impact on BRCA and CRC patient survival. As a result, we obtained that CIBERSORT estimated a higher number of T lymphocytes, DCs, and monocytes, whereas quanTIseq more easily recognized macrophages and neutrophils. As for the gene markers characterizing the cell types analyzed, they are more specific for identifying lymphoid cells than myeloid cells. Moreover, Kaplan-Meier curves showed higher Overall Survival (OS) in CRC patients with less stromal cell infiltration. By contrast, in BRCA the OS was significantly higher in subjects with a higher number of immune cells found in the TME, although the results also suggested a slight increase in OS related to higher levels of stromal cells.

FLASH TALKS II: AI & Healthcare

Sistemas de apoyo a la decisión para el diseño de un servicio de urgencias hospitalarias

Artigues Femenia, Raúl

Universidad de Navarra

La crisis de la COVID-19 tuvo un impacto negativo directo en el ámbito de gestión sanitaria, aumentando significativamente los tiempos de espera y la saturación en los servicios de urgencias hospitalarias tanto públicas como privadas. Fundamentando en este problema, se ha procedido a realizar un algoritmo matemático en R software capaz de analizar por simulación un servicio de urgencias en un hospital. Para ello se desarrolla una herramienta para la toma de decisiones que permite analizar los tiempos de servicio de los pacientes que acuden a urgencias dependiendo de los parámetros que definan dicho servicio. Con esta herramienta se puede optimizar el diseño y dimensionamiento de las unidades que componen el servicio de urgencias en un hospital. La interpretación de los resultados obtenidos del simulador se basa en los modelos de colas, concretamente, en la teoría de colas, es por eso, que es de vital importancia tener claros los conceptos que se desarrollan en el presente proyecto. Tanto las medidas de eficacia de la teoría de colas como las trazabilidades de los distintos pacientes del sistema se observan en una interfaz básica que ofrece la posibilidad de realizar consultas y sus respectivas conclusiones.

Chipless RFID Tag Implementation and Machine Learning Workflow for Robust Identification

Fodop Sokoudjou, J.J, Villa-Gonzalez, F., García-Cardarelli, P., Díaz Dorronsoro, J., Valderas Gázquez, D., Ochoa Álvarez, I.

Tecnun School of Engineering - University of Navarra

The integration of chipless RFID with internet of Things (IoT) has attracted attention in many domains, such as agriculture or retail, due to its low-cost, flexibility, versatility and sensing capabilities; and it is expected to eventually replace barcode technology. Due to the lack of electronic in chipless tags, their response are strongly affected by the environment, making the signal processing difficult. In this work, we describe step by step a complete workflow to apply machine learning (ML) classification for chipless RFID tag identification, covering: i) the tag implementation criteria for circular ring resonator (CRR) and square ring resonator (SRR) arrays for ML interoperability; ii) the data collection procedure to get a sufficiently representative dataset of real measurements; iii) the ML techniques to visualize the data and reduce its dimensionality; iv) the evaluation of the ML classifier to ensure high accuracy predictions on new measurements; and v) a thresholding scheme to increase the certainty of the predictions. The differences in the tags' frequency responses are maximized by optimizing the Hamming distance between the tag identifiers and by controlling each resonator array's radar cross section (RCS) level. We show that the proposed workflow achieves perfect accuracy for the identification of 4 tags at a fixed distance of 160 cm. We also evaluate the performance of the proposed workflow to identify up to 16 tags within a flexible range (up to 140 cm), showcasing the trade-off between the number of tags that can be correctly classified based on the reading range.

Hierarchical clustering and forecasting of the spanish day-ahead electricity supply curves

Li, Z., Alonso, A., Elías, A., Morales, J.
Universidad Carlos III de Madrid

In this paper, we perform a hierarchical clustering procedure on the day-ahead supply curves of the Spanish electricity market. The dissimilarity of curves is measured by weighted Euclidean distances. Once the cluster labels have been obtained, they are used in a supervised classification procedure, which allows us to characterize the obtained clusters. Additionally, a distance-based learning procedure is proposed to forecast the day-ahead curves. The procedure combines the idea of nearest neighbours with a machine learning procedure. The performance of our proposal was evaluated in an extensive prediction exercise and compared against two nearest neighbour benchmarks.

Segmentation CNN to evaluate street assets and traffic signs occlusion

Argüelles, P., Iparraguirre, O., Borro, D.
CEIT and TECNUN (University of Navarra)

Smart maintenance of roads is crucial for ensuring road safety and minimizing the costs of road infrastructure investments. One of the most common obstacles that interrupts road signage and requires more maintenance is partial occlusion by vegetation. Early detection of this occlusion is essential to ensure road safety and reduce the risk of traffic accidents. Nevertheless, this can be a tedious task if done manually. An automatization of this procedure could speed up the process while freeing up the technician's time for other, more complex, tasks and reducing the risk of errors or oversights. In this work, we present our first steps for detecting partial occlusion of traffic signals by vegetation through convolutional neural networks (CNN) and 3D scene reconstruction from videos taken from onboard RGB camera. The proposed method is based on the use of CNNs trained for object detection and semantic segmentation of road scenes. The neural networks can identify and separate traffic signals from their environment, allowing for the detection of partial occlusion by vegetation. Additionally, a 3D scene reconstruction of the road is built to evaluate the predictions (multiple signs, optimized georeferencing, distance calculation etc.) study the correctness of the predictions and the importance of such occlusion over road safety. In conclusion, this work presents the advances in an innovative method for detecting partial occlusion of traffic signals by vegetation through semantic segmentation with CNN. Early and accurate detection of partial occlusion is essential to ensure safety on the roads and minimize road maintenance costs. The proposed method can be efficiently and scalable implemented, making it suitable for use on roads of all types and sizes.

SESSION VII: Design of experiments

Computational design for thermal comfort and daylight optimisation in indoor environments

Gamero-Salinas, J., López-Fidalgo, J.
Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

This study presents a computational design approach for optimizing thermal comfort and daylight in indoor environments. The study focuses on two metrics: indoor overheating hours (IOH) for thermal comfort evaluation and spatial daylight autonomy (sDA) for daylight optimization. To efficiently explore the design space, a fractional factorial design methodology is employed. The optimization of both responses is then studied using response surface modelling and desirability functions. By combining these methods, the research aims to achieve a dual response optimization, ensuring occupants' thermal comfort while maximizing the utilization of natural daylight. The findings from this study contribute to the development of efficient computational design strategies for indoor environments.

Advances in Orthogonal Minimally Aliased Response Surface (OMARS) Designs

Núñez Ares, J., Schoen, E., Goos, P.
KU Leuven (Belgium)

The family of orthogonal minimally aliased response surface designs or OMARS designs bridges the gap between the small definitive screening designs and classical response surface designs, such as central composite designs and Box-Behnken designs. The initial OMARS designs involve three levels per factor and allow large numbers of quantitative factors to be studied efficiently using limited numbers of experimental tests. Many of the OMARS design possess good projection properties and offer better powers for quadratic effects than definitive screening designs with similar numbers of runs. Therefore, OMARS designs offer the possibility to perform a screening experiment and a response surface experiment in a single step, and thereby offer the opportunity to speed up innovation and process improvement. A technical feature of the initial OMARS designs is that they study every quantitative factor at its middle level the same number of times. As a result, every main effect can be estimated with the same precision using the initial OMARS designs, the power is the same for every main effect, and the quadratic effect of every factor has the same probability of being detected. In this talk, we show how to create OMARS designs in which the main effects of some factors are emphasized at the expense of their quadratic effects, or vice versa. We call the new OMARS designs non-uniform-precision OMARS designs, and show that relaxing the uniform-precision requirement opens a new large can of useful three-level experimental designs. The non-uniform-precision OMARS designs form a natural connection between the initial OMARS design, involving three levels for every factor and corresponding to one end of the OMARS spectrum, and the mixed-level OMARS designs, which involve three levels for some factors and two levels for other factors and correspond to another end of the OMARS spectrum.

Diseñando Experimentos para enseñar Diseño de Experimentos

López Redondo, D., Muñoz Calcerrada, J. S., Rodríguez-Aragón, L. J.
Universidad de Castilla-La Mancha

En este trabajo presentamos ejemplos reales desarrollados por alumnos de grado en ingeniería industrial con los que profundizar en el ámbito del diseño de experimentos. Nuestro trabajo se inspira en el Toy Experiment: los helicópteros de papel de Box, que hemos modificado y para el que hemos construido un mecanismo experimental avanzado. Además, hemos explorado el fenómeno físico de la autorrotación basándonos en un prototipo de cápsula espacial. Por último, nos hemos adentrado en los modelos de respuesta no lineal y los diseños de mezclas a través de recetas de cocina. Estos ejemplos permiten estudiar una disciplina práctica como el Diseño de Experimentos de forma práctica con un presupuesto reducido y una gran facilidad para replicar en diversas condiciones. Exploramos los diseños Factoriales Completos, 2k, las fracciones de diseños factoriales completos, 2k-p, las metodologías de superficie de respuesta, los modelos de Sheffé con su representación gráfica sobre el simplex y los modelos no lineales. Estos métodos se han aplicado a través de paquetes de R y los trabajos se han preparado teniendo en cuenta la reproducibilidad con Quarto markdown. Esta experiencia práctica puede servir como germen para nuevas ideas y métodos para presentar y profundizar en la docencia de la metodología del diseño de experimento.

SESSION VIII: Longitudinal studies

Robust interval-censored inference for step-stress accelerated-life tests under Weibull lifetimes

Jaenada, M., Balakrishnan, N. y Pardo, L.

Universidad Complutense de Madrid

Inferential methods in reliability and survival analysis investigate lifetime distributions from a parametric approach. Interval censored data arises in reliability tests when exact failure times cannot be observed, but are known to lie within an interval. Additionally, some products are highly reliable with large lifetimes under normal operating conditions, which makes experimentation quite difficult in terms of cost and time. In those cases, accelerated life-testing (ALT), wherein the experimental units are subjected to higher stress levels than normal conditions, may be considered. In particular, the step-stress ALT model increases the stress level to all units under test inducing the failure of more devices at higher stress levels. The Weibull distribution is particularly popular in survival analysis, as it can accurately model the time-to-failure of real-world events and is sufficiently flexible despite having only two parameters. On the other hand, recent works on one-shot devices have shown the advantage of using divergence-based methods in terms of robustness, with an unavoidable (but not significant) loss of efficiency, under censored data. In this work we develop robust inferential procedures based on the density power divergence (DPD) for interval-censored data under the step-stress accelerated life test model and Weibull lifetime distribution.

Efecto de la correlación de las observaciones en la selección de un diseño óptimo: Caso modelo de Gompertz simplificado

Benitez, E., Aznárez-Sanado, M., López-Fidalgo, J., Trocóniz, I., Parra-Guillén, Z.

Institute of Data Science and Artificial Intelligence (DATAI), TECNUN School of Engineering, Universidad de Navarra

En el presente trabajo, se evaluaron las implicaciones del uso del enfoque OED en el diseño experimental de crecimiento tumoral en términos de su impacto en la definición del número y ubicación de los tiempos de medición. Se comparó con un ejemplo común de ese tipo de diseño, y considerando diferentes escenarios de estructura de correlación entre observaciones tomadas a lo largo del tiempo. Para ello se seleccionó el modelo Gompertz-Laird, por su robustez y sencillez. En este análisis se encontró que la correlación tiene efectos positivos en todas las medidas de eficiencia de los diseños seleccionados, y que este efecto aumenta a medida que los niveles de correlación son mayores. Como aspecto negativo se encontró que los diseños bajo altos niveles de correlación amplían el área de exploración, e incluso, si superan cierto nivel, el espacio de evaluación presenta un "salto" que extiende la región de medición hacia los límites más extremos. Sin embargo, los análisis de eficiencia muestran que estos puntos no tienen un impacto mayor, y que incluso pueden descartarse. Al comparar los OED con el diseño real en el que se realizó el experimento, se identificó que se podría haber ahorrado un 25% de los recursos utilizando el enfoque OED. Además, los tiempos de duración requeridos para una estimación óptima de los parámetros fueron más cortos que los utilizados en el experimento real. En conjunto, OED se presenta como una metodología valiosa, no solo por su ahorro en costo económico, sino también porque permite reducir el número de observaciones o el número de sujetos.

Author Index

ORGANIZED BY

Universidad
de Navarra

DATAI
INSTITUTO DE CIENCIA DE LOS
DATOS E INTELIGENCIA ARTIFICIAL

IN COLLABORATION WITH

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

SPONSORS

(na)stat
Instituto de Estadística de **Navarra**
Nafarroako Estatistika Erakundea

Gobierno de Navarra
Nafarroako Gobernua
Departamento de Universidad,
Innovación y Transformación Digital
Unibertsitateko, Berrikuntzako eta
Eraldaketa Digitaleko Departamentua

Seio
Sociedad
de Estadística
e Investigación
Operativa