L-Università ta' Malta
Faculty of Science

Department of
Statistics &
Operations Research

Final Year (4th Year) Dissertation Presentations

Presented by students of the

Bachelor of Science (Honours) in Statistics and Operations Research (Joint Area)

Class of 2025/6

Programme Schedule

Non-Stationary Extreme Value Analysis of Climate Extremes in the Netherlands

by

Mr. Gary Azzopardi

The term Extreme Value Theory (EVT) refers to the statistical framework utilised to analyse rare extreme events. Classically, EVT is formulated under the assumption that observations are independent and identically distributed (i.i.d.), and the Generalised Extreme Value (GEV) distribution arises as the standard limiting model for extremes under this framework. In this dissertation, the Block Maxima (BM) approach is adopted to extract annual extremes from the data, retaining the single most extreme observation within each block. Four climatic variables are examined at two weather stations in the Netherlands: De Bilt (inland) and De Kooy (coastal). These are annual maximum daily temperature, annual minimum daily temperature, annual maximum daily precipitation, and annual maximum hourly wind speed. The classical i.i.d. EVT framework is extended to stationary dependent sequences. The central question is whether stationary GEV models are sufficient to describe climate extremes in an inland station (De Bilt) and coastal station (De Kooy) in the Netherlands, or whether non-stationary extensions are required. Differences between inland and coastal results were also examined. To address this question, GEV models for the eight station/meteorological variable combinations were fitted by maximum likelihood and compared using information criteria and the likelihood ratio test, with return levels computed for selected return periods. A non-stationary specification was preferred for seven of the eight combinations analysed. Temperature and wind extremes at both stations were best modelled through a time-varying location parameter, with hot extremes intensifying and cold and wind extremes weakening. For precipitation, the best model for De Bilt station data consisted of a time-varying scale parameter, reflecting variation in dispersion over time, while precipitation at De Kooy was the only case where the stationary model was preferred. These results show that the assumption of stationarity should not be made automatically when modelling extrema of meteorological variables, and that the most suitable model may vary depending on the variable considered and the station's location.

A Hybrid ALNS-Tabu Search Algorithm for the Airport Stand and Gate Assignment Problem

by

Ms. Dania El Wohashi

This dissertation addresses the Airport Stand and Gate Assignment Problem (ASGAP), a combinatorial optimisation problem that deals with the joint assignment of parking stands and terminal departure gates to aircraft. While existing literature has treated stands and gates as the same physical resource, no prior work addresses the configuration in which all stands are remote, thereby requiring explicit coordination between stand and gate assignments under distinct feasibility constraints. Malta International Airport (MIA), which operates exclusively with remote stands, provides the motivating case study for this research.

A Binary Non-Linear Program formulation is developed to minimise a weighted objective of passenger walking distance, aircraft taxiing distance, and a robustness penalty penalising tight stand turnaround gaps. The ASGAP is proven NP-hard via a polynomial-time reduction from the Circular-Arc Graph Colouring Problem. Given this complexity, an Adaptive Large Neighbourhood Search with Tabu Search (ALNS-TS) metaheuristic is proposed, combining adaptive operator selection, an Exponential Monte Carlo acceptance criterion, and a tabu memory structure to diversify the search and avoid cycling.

Computational experiments are conducted on four instances derived from MIA flight schedules, ranging from 45 to 115 aircraft. The Branch and Cut method obtains optimal solutions for all instances and serves as a benchmark. The proposed ALNS-TS approach produces high-quality solutions within substantially shorter runtimes, achieving optimality gaps ranging from 9% to 17%. For the largest instance, ALNS-TS delivers assignments in under 15 minutes against the exact method's 4.58 hours to certify optimality. Overall, the proposed framework provides a practical and scalable framework for improving stand and gate coordination in airports operating under remote-stand configurations.

Using the robust Cox regression model to analyze student dropouts in tertiary education

by

Ms. Mireille Fenech

This thesis is about the Cox Proportional Hazards (CPH) model using robust estimation techniques. Although CPH models have been used extensively to analyze time-to-event data in educational settings, they are vulnerable to outliers. These outliers can severely affect the model's parameter estimates (regression coefficients) and lead to incorrect inferences. To overcome this issue, this study employs the Robust Cox Regression Estimator, which is based on a smooth modification of the Partial Likelihood (PL) which is a reliable estimation technique originally proposed by Bednarski (1993). An advantage of this robust method is that it provides unbiased estimators of the hazard ratios in the presence of data contamination. This robust methodology will be used to investigate student dropouts in undergraduate courses at the Faculty of Science in the University of Malta (UOM).

Using the student cohort that started an undergraduate course at the Faculty of Science in 2022, the study examines the impact on dropout rates using a Cox regression model, where gender, admission score, and subject choices are the predictors. Moreover, the study compares results obtained from traditional Cox regression analysis with its robust alternative approach to assess the stability of Hazard Ratio (HR) estimators when potential outliers are downweighed. This thesis demonstrates that by reducing the impact of observations with high leverage in the PL function can yield more accurate parameter estimates without a notable loss of efficiency.

One of the findings shows that the number of dropouts peak at the end of the first year of tertiary education but then drop in subsequent years. Other findings show that pre-tertiary academic attainment and the selected specializing subjects have a huge impact on dropout rates. Moreover, while pointing out situations where classical estimates might be unduly impacted by atypical cases, the robust approach also identifies a number of covariates whose effects hold true across various estimation frameworks. These findings are essential to university administration to allocate resources and devise plans to reduce student dropouts from university courses.

A Game-Theoretic Bi-Level Optimisation Approach to the Competitive Facility Location Problem

by

Mr. Julian Formosa

For companies that service customers through physical facilities, the location of each facility is often a critical factor in determining the success of that company. The Competitive Facility Location Problem is concerned with optimising the locations of the facilities belonging to multiple competing companies that operate in the same market. This dissertation takes several aspects of existing models and combines them in a unique way to form a novel variant, which is modelled as a Bi-Level Optimisation program using a Game-Theoretic lens. In particular, this dissertation focuses on a Leader-Follower budgeted variant where the value of the Follower's budget is not known with certainty by the Leader. Further aspects included in the model are quality levels, elastic demands, pre-existing facilities, a maximum threshold on the number of facilities that can be owned, and customer behaviour based on the Huff gravitational rule.

This dissertation investigates the computational complexity of solving Bi-Level Optimisation programs, and shows that exact and approximation methods may become impractical for large-scale problem instances. To address this, a memetic algorithm composed of Genetic Algorithm and Tabu Search is formulated, and a novel heuristic specifically adapted to our problem is proposed. These two algorithms, together with an exhaustive search exact method, are used to solve problem instances constructed to model the locations and types of electric vehicle charging stations in Sliema, Malta. The results show that both inexact methods significantly outperform the exact method in terms of solving speed. Moreover, the memetic algorithm is shown to consistently produce near-optimal solutions, whereas the heuristic method is faster but yields solutions of comparatively lower quality.

Coffee Break

Distributed Lag Non-linear Models of Climatic and Tourism Effects on Campylobacter Incidence in Malta

by

Mr. Julian Nicholas Gatt

Campylobacter is the leading bacterial enteric pathogen in the European Union and ranks among Malta's most notified infections. Its incidence exhibits a pronounced seasonal cycle that is strongly associated with ambient temperature, yet the non-linear shape and lag structure of this relationship remain insufficiently characterised in the Mediterranean small-island context. This dissertation applies a Negative Binomial Distributed Lag Non-Linear Model (NB2-DLNM) to 509 weeks of national *Campylobacter* surveillance data from Malta (January 2016 – September 2025), jointly estimating the exposure-lag-response surface for weekly minimum temperature, the independent effect of weekly precipitation, and modification of the thermal association by tourism pressure. The outcome distribution accommodates overdispersion; a latent first-order autoregressive process captures residual temporal dependence; and natural cubic spline cross-bases simultaneously characterise non linearity and lag decay. Marginal likelihood estimation proceeds via the Laplace approximation. Monthly tourist arrivals were temporally disaggregated to weekly resolution using the Denton–Cholette procedure; subsequent cross-correlation analysis established tourism as a contextual effect modifier rather than a distributed lag predictor. Weekly minimum temperature was significantly associated with *Campylobacter* incidence in a non-linear, lagged pattern (joint cross-basis LRT: $p < 0.001$). Relative to an identified minimum-risk temperature threshold, interquartile warming was associated with a 2.44-fold increase in expected weekly incidence (95% CI: 1.26–4.74). Weekly precipitation showed no significant independent or modifying effect. The temperature exposure-lag-response surface differed significantly across tourism tertile strata ($p = 0.003$), with shoulder-season weeks exhibiting the greatest thermal sensitivity, suggesting that predicted incidence is highest when visitor pressure aligns with the steepest portion of the thermal exposure–response gradient. These findings have direct implications for climate-informed *Campylobacter* surveillance and proactive public-health response in Malta.

A Bayesian Hierarchical Approach to Modelling Non-Residential Energy Consumption in the Maltese Islands

by

Ms. Christine Doris Grech

The principal goal of this dissertation is to determine and analyse the factors influencing electricity consumption across non-residential buildings in the Maltese islands, while explicitly modelling the hierarchical dependencies present within the data. To achieve this, a series of statistical models are developed within a Bayesian framework and applied to non-residential electricity consumption data. The models are constructed in an order of increasing complexity. Specifically, the framework progresses from the first model which includes NACE and locality level random intercepts to one that additionally includes covariates. This is followed by a model with added company-specific random intercepts and finally a formulation incorporating the pairwise interaction between NACE and locality. This incremental structure allows for an evaluation of how additional complexity improves model performance and interpretability. Inference is conducted using Bayesian methods, providing full posterior distributions for all parameters. Key summaries, including posterior means, medians and Highest Density Intervals, are used to quantify effects, associated uncertainty and practical significance. Model adequacy is assessed through convergence diagnostics and graphical Posterior Predictive Checks to ensure that the fitted models provide a reliable representation of the observed data. Through comparative analysis of the specified models, the one offering the best data description and most reliable inference is identified. The most complex model is identified to be the optimal specification, indicating that electricity consumption is not only influenced by sectoral and spatial factors, but also by their interaction. Substantial heterogeneity across sectors and regions, along with significant company-level effects, highlights the role of unobserved company-specific characteristics. Overall, the selected model quantifies the primary drivers of electricity consumption and explains the variance observed across different organisational levels. This dissertation highlights the value of Bayesian hierarchical modelling in addressing complex structured data and deepens understanding of electricity consumption patterns among companies in Malta.

Sentiment Classification of Football Match Reports using Transformer-Based Language Models

by

Mr. Aidan Scott Grima

Sentiment analysis has become an important tool for extracting meaning and emotional tone from large volumes of textual data. Within sports analytics, football match reports offer a structured and sentiment-rich source of language that reflects both factual outcomes and subjective framing. This dissertation investigates whether transformer based language models can classify match reports into win, tie, or loss categories based solely on linguistic cues, without relying on explicit indicators such as scorelines. Using the Multilingual Emotional Football Corpus (MEmoFC), which contains English, Dutch, and German reports from the 2015/2016 season, a supervised learning framework is implemented. Multilingual BERT (mBERT) and XLM-RoBERTa are fine-tuned and compared against more traditional models, including Support Vector Machines, Logistic Regression, and Naïve Bayes. Preprocessing techniques, such as masking scorelines and outcome indicative terms, are applied to ensure models learn from contextual information. Results demonstrated that transformer based models largely yield a statistically significant improvement in classification performance over traditional approaches, thereby meriting the trade-off of increased computational cost for enhanced predictive capability.

A Comparison of Chain Ladder and Over-dispersed Poisson Approaches for Forecasting Aggregate Motor Insurance Claims

by

Ms. Michela Lupi

In this dissertation, the performance of traditional reserving techniques - namely the Chain Ladder Method (CLM) and its extension, the Mack model - are compared to models that utilize a statistical framework. Specifically, we refer to models within an Over-dispersed Poisson (ODP) model framework, which is implemented using a quasi-Generalized Linear Model (GLM) structure. Most commonly, the ODP model is applied taking into account accident year and developing year, but this is also extended to include additional factors that may affect claim behaviour, with the final aim of assessing how the use of more granular data affects estimation. We refer to the latter as the extended ODP model.

The Mack, ODP and extended ODP models are applied to a synthetic dataset consisting of motor insurance claims data. Their performance and predictive ability are assessed and compared through a combination of fitted claim estimates, confidence intervals and multiple error measures such as the MAE, RMSE and MAPE. Moreover, further evaluation of the goodness of fit of the models, particularly of those that utilize the GLM structure, is carried out through the analysis of residual, leverage and Cook's distance plots.

The results of this comparison show that the Mack and ODP models produce identical point estimates for ultimate claims, differing only in confidence intervals, in which the ODP yielded narrower intervals. The extended ODP provides additional insight into how additional factors affect claim severity. It is found that age range and car brand have a significant influence on claims behaviour, however no substantial improvement in predictive accuracy is seen over that of simpler models.

Overall, the study highlights that while traditional reserving techniques remain reliable and simple to implement, the use of statistical frameworks, such as a GLM structure, allows for flexibility and further inference. Furthermore, the inclusion of additional factors may allow for a deeper understanding on what affects claims, however it may not always provide significant improvements in predictive performance.

Lunch Break

Modelling and Solving the University Examination Timetabling Problem

by

Mr. Jean Carl Muscat

Constructing an effective timetable is a complex optimisation problem involving many constraints and limited resources. This is particularly evident in the University Examination Timetabling Problem (UETP), where examinations must be scheduled without student conflicts, while also respecting venue capacities and institutional rules. As problem size increases, the number of feasible solutions grows rapidly, making exact optimisation methods difficult to apply.

This dissertation investigates both exact and metaheuristic approaches to the UETP, with the aim of producing feasible and high-quality timetables within reasonable computational time. The problem is formulated as a mathematical model with hard constraints, such as conflict avoidance, and soft constraints, such as efficient venue usage. The dataset is designed to reflect the structure of examination timetabling within the Faculty of Science at the University of Malta. A Branch-and-Cut approach is used to obtain benchmark solutions for smaller instances. Metaheuristic methods, including a Genetic Algorithm, an adaptive Evolutionary Algorithm, and Tabu Search, are developed for larger in stances. The results show that exact methods perform well on small problems, while metaheuristics scale more effectively as complexity increases.

Examining Ordinal Categorical Data Using Multilevel Modelling: Estimation Methods and an Application to Dance Competition Data

by

Ms. Milena Pellicano

Ordinal categorical outcomes frequently arise in applied research, particularly in settings where responses are recorded using rating scores rather than continuous measurements. In many practical applications, the data may be nested in higher level structures, with observations clustered within higher-level units, leading to dependence that cannot be adequately handled by standard regression models. This dissertation analyses ordinal categorical responses within a multilevel modelling framework, with particular emphasis on likelihood-based estimation methods for generalized mixed-effects models and their practical application to dance competition data.

The study first develops the theoretical foundation of two-level and three-level random effects models, beginning with continuous-response formulations and extending these to cumulative logit models for ordinal outcomes. Particular attention is given to random intercept and random coefficient structures, the interpretation of between-cluster variation, and the role of intra-class correlation in assessing dependence within hierarchical data. The dissertation then discusses the estimation and inferential techniques used for multilevel ordinal models and explores procedures to overcome difficulties that arise when the marginal likelihood involves integrals over random effects that do not have closed-form solutions.

The study examines various numerical integration methods used in marginal likelihood estimation, with emphasis on Gaussian quadrature and Gauss-Hermite quadrature. The construction of quadrature rules and their role in approximating intractable integrals are discussed in detail, together with the use of modified Newton-Raphson procedures for maximising the approximated likelihood. In addition, Bayesian ideas are introduced in the context of predicting random effects, with empirical Bayes posterior means used to obtain cluster-specific predictions within the fitted models.

These modelling methods are applied to the StarDance competition dataset. Since the dancing performance scores awarded by judges had a left skewed non-normal distribution, it was decided to categorise these scores to ordinal responses and analysed using multilevel cumulative logit models. The hierarchical structure is represented by individual performances (level-1 units) nested within dance types (level-2 units) and competition seasons (level-3 units), allowing both two-level and three-level models to be fitted. The models are implemented in Stata using the gllamm software, which provides flexible likelihood-based estimation for multilevel ordinal responses using numerical integration. The analysis demonstrates how multilevel ordinal models can be used to account for clustering, estimate random-effects variability, compare alternative model structures, and interpret the effects of explanatory variables in a practically meaningful way.

Overall, the dissertation provides a methodological and applied examination of multilevel modelling for ordinal categorical data. It shows that when ordinal responses are analysed within a hierarchical framework, likelihood-based estimation supported by quadrature methods offers a rigorous and practical approach for modelling complex dependence structures and obtaining interpretable statistical inferences.

An Individual-Based Approach to Modified Exponentially Weighted Moving Average Control Charts for Monitoring Skewed Financial Time Series

by

Mr. Adam Perici Ferrante

This dissertation investigates the performance of modified Exponentially Weighted Moving Average (EWMA) control charts for monitoring financial time series data under non-normal conditions. This study presents the theoretical groundwork on three approaches, the Weighted Variance (WV-EWMA), Weighted Standard Deviation (WSD EWMA) and Skewness Correction (SC-EWMA).

These charts are implemented on the bases that the underlying data forms rational subgroups, however in many financial time series this is not the case. These charts were extended on an individuals framework, appropriate for financial data observed sequentially, which do not form rational subgroups, common in daily exchange rate returns. This adaptation is treated as a contribution to this study. These control charts are applied on real-life financial data and on simulated data to compare their performance.

An empirical analysis is conducted using daily USD/EUR exchange rate returns, transformed into log-return to achieve stationarity. Diagnostic testing confirms no significant autocorrelation in the transformed data, while descriptive statistics and distribution fitting indicate the presence of excess kurtosis and mild skewness. Control limits are thus estimated under unknown parameters using an individual chart framework with smoothing parameter λ=0.10 and control limit multiplier L=2.6952 in order to reach target Average Run Length of 370.

The simulation study approach is used evaluate performance, focusing on out-of-control ARL under varying scenarios of effects of skewness and effects of excess kurtosis. The results show clear differences across methods. In particular, the WSD EWMA produces the best performance compare to the other two, while SC-EWMA performed the worst. WV-EWMA provided a balance between detection and false alarm rates. Overall, the results show a trade-off between detection speed and false alarm control.

Risk Modelling of Azathioprine-Associated Pancreatitis in IBD via Firth's Penalised Logistic Regression

by

Mr. Matthew Portelli

Azathioprine-associated acute pancreatitis is a rare but clinically significant ad verse drug reaction in patients with inflammatory bowel disease (IBD), posing challenges for conventional statistical analysis. This dissertation investigates risk factors for pancreatic enzyme abnormalities following thiopurine initiation in a retrospective cohort of 91 IBD patients treated at Mater Dei Hospital, Malta. A three category outcome framework was adopted (normal amylase, asymptomatic hyperamy lasaemia, and symptomatic pancreatitis), with Firth's penalised logistic regression used to address rare event bias and separation. Given the limited number of pancreatitis cases (n = 5), a binary outcome combining hyperamylasaemia and pancreatitis (n = 26) was selected as the primary model. The main effects model identified male sex as the only statistically significant predictor (OR = 2.74, 95% CI: 1.03–8.07, p = 0.043), corresponding to an approximate 2.7-fold increase in odds of an adverse outcome. However, this finding should be interpreted with caution given the limited sample size and evidence of moderate overfitting from bootstrap validation. Younger age and absence of steroid use showed consistent but non-significant associations. Classical multinomial regression produced unstable estimates due to separation, underscoring the importance of penalised methods. This work demonstrates a robust framework for modelling rare adverse events in clinical epidemiology, with implications for thiopurine risk stratification in IBD.

Coffee Break

A Bayesian Hierarchical Model of Individual Player Performance in the 2022 FIFA World Cup

by

Mr. Andreas Psaila

This dissertation examines a Bayesian hierarchical model to assess individual player ability in the 2022 FIFA World Cup. This is a challenging task since player performance is affected not only by individual skill, but also by external factors such as team quality, opponent strength, weather conditions, and match context. Moreover, the tournament consists of a small number of matches, especially for players whose teams are eliminated at earlier stages. Using player data from the entire tournament, the analysis focuses on three main performance indicators: interceptions, shots and passes, which are modelled using either a Poisson or a negative binomial distribution. A negative binomial is more adequate when overdispersion is present. The intensity of an event includes player ability, team effects, and opponent strength in order to evaluate the contribution made by each player in a given match. The two models obtained for each indicator were compared, namely using Bayesian RMSE and MAE. When fitting the interceptions and shots models, convergence was achieved. Since fewer players were included for each of these models, and since interception and shot counts are relatively low, the models perform as expected and produce acceptable values of the Bayesian RMSE and MAE. This suggests that the modelling framework utilised in this dissertation is well suited to these two performance metrics and it provides reliable and stable estimates. However, fitting the passing model proved more complicated due to convergence issues encountered and higher than desirable Bayesian RMSE and MAE. This indicates that the proposed model is more suited for events where individual player ability is more easy to separate from the corresponding team ability (like interceptions and shots) and less suited for events (like passing) where individual player ability and corresponding team ability are more easily conflated.

An Adaptive Large Neighbourhood Search Framework for Vehicle Routing under Stochastic Travel Times

by

Ms. Julia Maria Sciberras

The Vehicle Routing Problem (VRP) is a fundamental optimization problem in logistics, where efficient routing decisions are required to reduce operational costs. In practice, travel times in VRPs are often uncertain due to congestion, accidents, and varying traffic conditions, which limits the realism of deterministic routing models. Therefore, this dissertation studies the Capacitated VRP with Stochastic Travel Times (CVRPSTT) and proposes a solution framework that combines data-driven scenario generation with metaheuristic optimization.

Historical travel time matrices across 150 locations in Malta were first transformed using Principal Component Analysis to reduce dimensionality, while preserving the dominant variation of the data. The reduced data was then clustered using the K-means clustering technique, to generate representative traffic scenarios with associated probabilities. The CVRPSTT was then represented as a mixed integer model with an expected objective function and solved using a hybridization of an Adaptive Large Neighbourhood Search (ALNS) and Simulated Annealing.

The computational experiments on three instances of varying sizes show that the proposed framework consistently produces high-quality solutions within practical computational time. Among the different initialization methods considered, the Savings heuristic yields the strongest overall performance. Sensitivity analysis demonstrates that most of the selected parameters are robust, while moderate destruction levels and stronger search intensification generally improve solution quality. Comparisons with the Branch and Cut exact method on the smaller instances show that ALNS produces highly competitive solutions, while larger instances highlight the practical limitations of exact approaches. Overall, the results demonstrate that integrating statistical learning techniques with adaptive metaheuristic search provides an affective and scalable approach for solving the CVPRSTT.

Synthetic Realism: GAN-Based Data Augmentation for Lung Cancer Imaging

by

Ms. Rebecca Josephine Zammit

The ability to generate realistic synthetic data stands as one of the most profound challenges in machine learning, a challenge that Generative Adversarial Networks (GANs) are able to address. GANs learn to generate synthetic data through an adversarial process formalised as a minimax game between two neural networks: a generator and a discriminator. GANs have been applied across various fields, including healthcare. In medical imaging, particularly in chest CT scan analysis for lung tumor detection, the complexity and variability of tumor presentation, combined with the high cost of data acquisition and annotation, present significant challenges for the development of deep learning models. This dissertation explores the use of GANs for data augmentation in this domain, giving particular focus to Deep Convolutional GANs (DCGANs) and Wasserstein GANs with gradient penalty (WGAN-GP) to generate synthetic chest CT images containing lung tumors. The aim is to assess the different capabilities of these models when applied to the IQ OTH/NCCD lung cancer dataset, which comprises CT scans from 110 subjects and 1,190 annotated CT slices collected from two specialised oncology centres in Iraq. The analysis is restricted to a subset of malignant cases within the dataset. To identify the optimal configuration, extensive parameter tuning was conducted for both models, supported by quantitative and qualitative evaluations to determine which model performs better on this dataset. This dissertation, therefore, provides a comparative and practical evaluation of GAN-based approaches, highlighting their potential for use in the medical field.

The End

We hope you found today's presentations enjoyable.

Thank you for joining us!

Department of Statistics and Operations Research

University of Malta