53rd Annual ORSSA Conference

25-28 August 2024
Skukuza

NWU
NORTH-WEST UNIVERSITY
NOORDWES-UNIVERSITEIT
UNIBESITI YA BOKONE-BOPHIRIMA

ORSSA

Contents

Welcome

Welcome to the 53rd annual conference of the Operations Research Society of South Africa (ORSSA), hosted against the magnificent backdrop of the Kruger National Park! We sincerely hope that you will enjoy both the stimulating academic programme of talks and discussion sessions as well as the true South Africa safari experience.

The scientific programme spans three days wherein we plan to present multiple plenaries, tutorial sessions, as well as panel discussions. We look forward to your engagement during these sessions as they promise to be both compelling and informative. The presentation sessions form the bulk of the programme and consists of just over 30 talks. The topics covered in these talks cover a wide variety of topics and we are sure that everyone will find something of interest.

The non-academic programme will consist of several early morning and afternoon planned trips out into the park. In addition, a bush braai (*i.e.*, 'barbeque' to our international guests) has been planned for the Tuesday evening; dining at night in the middle of the bush (potentially) surrounded by wild animals is an experience not to be missed (precautions are in place to ensure that it is perfectly safe though).

The conference is made possible by the generous partnership of our sponsors: The Department of Industrial Engineering at Stellenbosch University (sponsoring the social events), and the Department of Statistical Sciences and the Modelling and Simulation Hub, Africa (MASHA) at the University of Cape Town. The Centre for Business Mathematics and Informatics, North-West University again sponsors two student delegates of OR to attend the conference as part of the Society's development strategy. Finally, we would like to thank all of the delegates for attending this conference. Without your participation nothing would have realised. We hope you all have an amazing stay and that you enjoy every minute of it.

ORSSA2024 Local Organising Committee
Elzanie Bothma, James Allison, Leonard Santana, Lethani Ndwandwe, Marius Smuts

Scientific Programme Committee (SPC)

Elzanie Bothma, James Allison, Leonard Santana, Lethani Ndwandwe, Marius Smuts

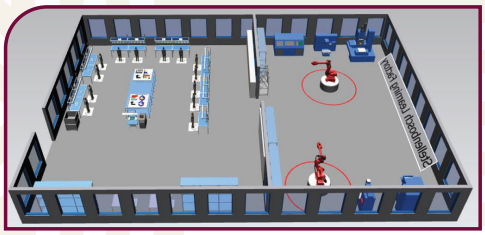
INDUSTRIAL ENGINEERING

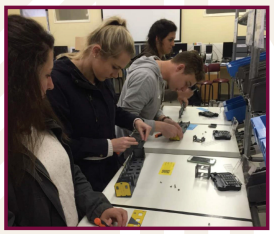
OPERATIONS AND SUPPLY CHAIN MANAGEMENT

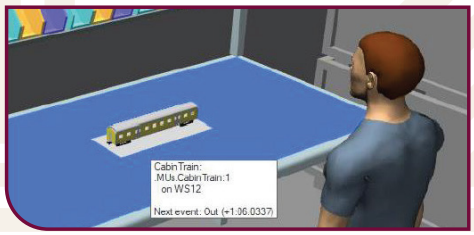
PROF JAN VAN VUUREN, DR STEPHAN NEL, PROF JAN HAVENGA AND DR ZANE SIMPSON

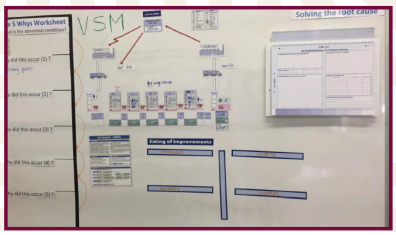
Digitilisation and Automation in Operations

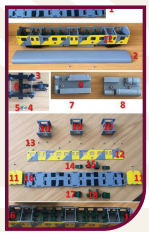
The work life and the organisational world is changing very fast. Digitilisation and the fourth industrial revolution are accelerating these changes. A key research focus is on the digitalisation of operations and value chains in manufacturing and service environments. Digitilisation has touched all aspects of businesses, including supply chains and operating models/ business models. Today, new technologies have enabled organizations to transform their existing hybrid structures (combination of paper-based and IT-supported processes) into more flexible, open, agile and collaborative digital models. Digital value chains enable business process automation, organizational flexibility and digital management of corporate assets.







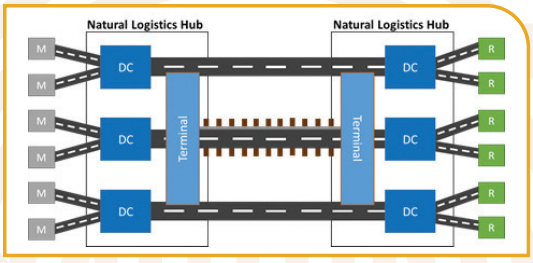


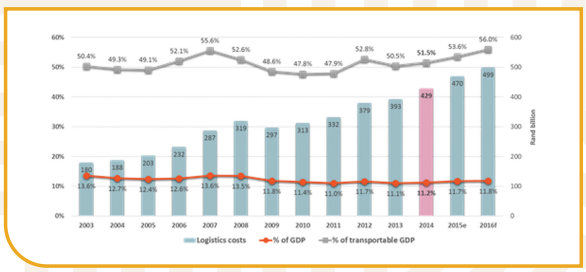


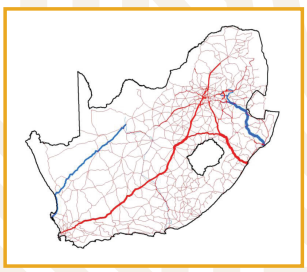
Macrologistics Management and their impacts

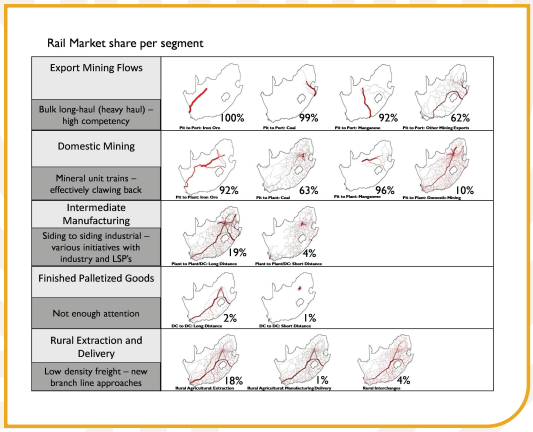
Macrologistics is the scoping and prioritizing of national policy, projects and macrologistics performance measurement. Macrologistics involves a complex group of interrelated economic, infrastructure and organizational systems that could be analyzed and improved using a systems engineering approach. This has been done to some extent within the Logistics Management domain through modelling sub-systems and research in micrologistics, but a vast number of research opportunities exist from the systems engineering perspective. This opportunity spans across engineering disciplines with applications on both detail level but also at strategic or even national level. These include but are not limited to the supply chain in general, logistics, systems modelling, operational research, financing and costing, infrastructure development, equipment design and policy environments with large scale benefits for a South Africa in need of growth and development.

Macrologistics improvement is possible through a multidisciplinary hierarchical system, i.e. on the first level by avoiding unnecessary inputs (macroeconomics), on the second level by avoiding unnecessary activity (industrial engineering and logistics) and on the third level by improving the efficiency of infrastructure (civil engineering) and equipment (mechanical engineering) used in logistics.









Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

ENGINEERING
EYOBUNJINELI
INGENIEURSWESE

elytica

The world's first Cross-Solver Collaborative Optimization Platform

Optimization, simulation and predictive modeling in a single integrated environment.

1. Capture models using LaTeX

```
bin x, forall n in N;
bin y, forall c in C, forall n in N;

max sum_{i in N}{ v_{i}*x_{i} };

constr x_{i} <= 0, forall i in E;
constr x_{i} <= x_{j}, forall i in L, forall j in A(i);
constr x_{i} >= sum_{j in J(i)}{ x_{j} } - 3, forall i in NH;
constr x_{i} <= x_{j}, forall i in I, forall j in B(i);
constr x_{i} <= sum_{j in D(i)}{ x_{j} }, forall i in S;
```

2. Simulate/Predict using Python

```
import elytica
import simpy
import numpy
import math
import pandas as pd
import networkit as nk

SIMULATION_RUN_HOURS=100
network_nodes = json.loads((pd.read_excel("simulation
node_xyz = {i["id"]:[ float(i["x"]), float(i["y"]), f
network_arcs = json.loads((pd.read_excel("simulation.
arc_distances = { i["from_id"]: { j["to_id"]: float(j
trucks = json.loads((pd.read_excel("simulation.xlsx",
truck_type_info = json.loads((pd.read_excel("simulati
truck_type_efficiency = json.loads((pd.read_excel("si
truck_type = {i["id"]: i["type_id"] for i in trucks}
```

3. Inspect models visually

$$\text{Maximize} \sum_{i \in N} v_i x_i \quad : 25$$

$$x_i \le 0, \quad \forall i \in E \quad : 27$$

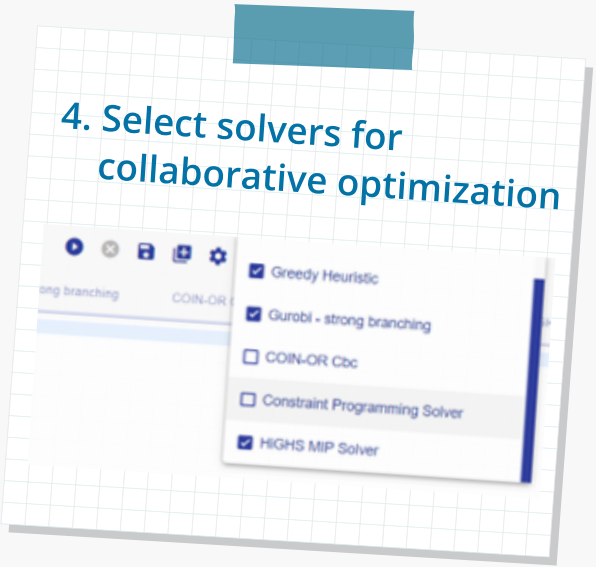
$$x_i \le x_j, \quad \forall i \in L, j \in A(i) \quad : 28$$

$$x_i \ge \sum_{j \in J(i)} x_j - 3, \quad \forall i \in NH \quad : 29$$

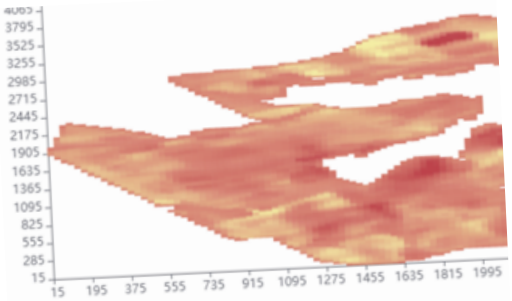
$$x_i \le x_j, \quad \forall i \in I, j \in B(i) \quad : 30$$

$$\sigma_0 x_i \le \sum_{j \in D(i)} x_j, \quad \forall i \in S \quad : 31$$

4. Select solvers for collaborative optimization



5. Graph & export results



SCAN ME

Sign up for a FREE trial account:

compute.elytica.com

RESEARCH:

OPERATIONAL RESEARCH

This research group focuses on the development of decision modelling and support tools, including both "hard" (e.g. optimization and simulation) and "soft" (e.g. problem structuring and systems modelling) approaches, relevant to critical national decision and policy making.

DATA SCIENCE

Data science is an exciting new field that uses computer-intensive statistical methods to identify patterns and make predictions using large volumes of data. The applications of data science are diverse, ranging from predicting fraudulent transactions before they occur to extracting marketing insights from unstructured social media data.

BIOMEDICAL STATISTICS

This interdisciplinary group brings together researchers involved in the analysis of medical and biomedical data with a focus on advanced methodology applicable to bioinformatics, infectious disease modelling, longitudinal and time-to-event modelling, growth curve modelling, causal modelling, methods for incidence estimation and multivariate analysis.

STATISTICS IN ECOLOGY, ENVIRONMENT & CONSERVATION

The Centre for Statistics in Ecology, Environment and Conservation focuses on conservation, animal and plant demography, climate change, understanding of biodiversity patterns, evolutionary ecology and macroecology, with an overarching theme of structured decision support.

STATISTICAL FINANCE

This is a transdisciplinary research group working on the intersection of the statistical analysis in data-intensive finance and mathematical statistics that considers various data-informed approaches to both practical and theoretical problems in low and high-frequency finance.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

STATISTICAL SCIENCES

WHO WE ARE:

The Department of Statistical Sciences at the University of Cape Town is committed to the development of the statistical sciences within and beyond the university. In its teaching, the department recognises that there are needs both to train professionals in the statistical sciences, and to provide quantitative and decision-making skills to students in other disciplines. In its research, the department seeks to maintain a balance between the development of theory and applications of that theory with a focus on supporting policy and decision-making.

The Department of Statistical Sciences was established in 1965 as the Department of Mathematical Statistics under the headship of the late Professor Cas Troskie. It assumed its current name at the beginning of 1991 to reflect its broader teaching and research activities. Located in the south wing of the PD Hahn building, the department is situated at the heart of the beautiful UCT upper campus at the foot of Table Mountain.

DEPARTMENT OF STATISTICAL SCIENCES
CONTACT DETAILS: Tel: +27 (0)21 650 3219
www.facebook.com/statistics.uct
https://twitter.com/StatsUct
www.stats.uct.ac.za

DEPARTMENT OF STATISTICAL SCIENCES

PHYSICAL ADDRESS: PD Hahn Building (South Entrance)
Level 5 below the Science Faculty Office, Upper Campus, UCT
CONTACT DETAILS: Tel: +27 (0)21 650 3219 Fax: +27 (0)21 650 4773
www.facebook.com/statistics.uct https://twitter.com/StatsUct
www.stats.uct.ac.za

Monday, 26 September 2024			
09:30		**Welcoming**	
10:00		**Plenary session: Rohit Kumar**	
11:00		**Tea Break**	
11:30		**Panel discussion: Prof Rènette Blignaut**	
13:00		**Lunch**	
		Public health modelling and simulation	
14:20	Jared Norman	Modelling and Simulation Hub, Africa (MASHA), University of Cape Town	Rapid modelling to inform diphtheria outbreak response vaccination in Nigeria - A case study
14:40	Sheetal Silal	Modelling and Simulation Hub, Africa (MASHA), University of Cape Town	Reflections on COVID-19 modelling in South Africa: Lessons learned and looking forward
15:00	Retselisitsoe Monyake	Modelling and Simulation Hub, Africa (MASHA), University of Cape Town	DTP Boost: A novel decision-making tool for DTPCV booster introduction in LMICs
15:20	Lené Nortje	Institute of Biomedical Engineering, University of Stellenbosch	Using machine learning to predict the risk of late effects of a South African childhood cancer survivor cohort
15:40		**Tea Break**	
		Dynamic complex systems analysis	
16:00	Marlize Visser	Department of Industrial Engineering, Stellenbosch University	A new inventory routing problem with route and schedule unpredictability
16:20	Lieschen Venter	Department of Logistics, Stellenbosch University	Predicting learner progress for the South African public high school system
16:40	Robert Bennetto	Icepack	A tale of two grids: Chasing the sun
18:00		**End of sessions**	
18:30		**Dinner at the Kruger Shalati**	

Tuesday, 27 September 2024			
09:30	**Plenary session: Prof Jan van Vuuren**		
10:30	**Tea break**		
	Finance, infrastructure, and energy		
11:00	Johan du Pisanie	Viewwave	On the impact of distributional changes on the probability of default in credit risk scoring
11:20	Roelinde Bester	Centre for Business Mathematics and Informatics, North-West University	The TruEnd-procedure: Treating trailing zero-valued balances in credit data
11:40	Christa Searle	Centre for Logistics & Sustainability, Edinburgh Business School, Heriot-Watt University	A Framework for multimodal hydrogen refuelling infrastructure design
12:00	Andrea Plumbley	Department of Statistical Sciences, University of Cape Town	Multi-objective optimisation of the generalised bin packing problem
12:30	**Lunch**		
13:30	**Tutorial session: Elytica**		
15:30	**Tea break**		
16:30	**Bush Braai**		

Wednesday, 28 September 2024			
		National Student Competition	
09:00	Himil Parshotam	Department of Industrial Engineering, Stellenbosch University	A framework for evaluating clinical data using graph-based link prediction
09:20	Hannah Callaghan	Department of Logistics, Stellenbosch University	Mobile clinic deployment in the Witzenberg region of South Africa
09:40	Thomas Simon	Department of Industrial Engineering, Stellenbosch University	A practical method for joint replenishment under supplier constraints
10:00	Sha-abaan Slamang	Department of Logistics, Stellenbosch University	Analysing student progression by means of a hybrid simulation modelling paradigm
10:20		**ORSSA 2024 Awards Ceremony**	
10:35		**Tea Break**	
10:45		**Panel discussion: Prof Sheetal Silal**	
12:00		**Lunch**	
		Optimisation and security	
13:00	Helena Ramalhinho	Department of Economics and Business, University Pompeu Fabra	The rich heterogeneous dial-a-ride problem with trip time prediction
13:20	Nathan Morgenrood	North-West University	Enhancing IoT security: Integrating AutoML and SHAP for improved anomaly detection and interpretability
13:40	Winnie Pelser	Council for Scientific and Industrial Research	Decision making in the current military environment
14:00	Magda Dettlaff	University of Gdansk	A new approach to b-coloring of regular graphs - for the purpose of organizing trainings
14:20	Funanani Netshitungulu	Department of Statistical Sciences, University of Cape Town	An overview and comparison of heuristics for the generalized bin packing problem
14:40		**Tea Break**	
		Optimisation and security	
15:00	Kit Searle	School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh	A branch-and-cut algorithm for the κ-connected forest-star problem
15:20	Joerg Kalcsics	University of Edinburgh	The distance-limited firefighter problem
15:40	Sarah-lee de Greeff	Industrial Engineering, Stellenbosch University	A computer vision framework towards automated scene reconstruction analysis for the use of energy auditing
16:00	Hano Strydom	North-West University	Using automated supervised autoencoders to detect IoT botnet attacks
16:20		**Close of Conference**	
18:00		**Dinner at Lodge**	

Abstracts

A tale of two grids: Chasing the sun

Monday, 26 Aug, 16h40

Robert Bennetto

Icepack

This talk builds on the topic the author presented last year. The problem of going off-grid with solar generation at scale in South Africa, with a limited budget, multiple stakeholders, and a high-interest rate environment. This talk will begin with a brief review of the content previously presented. This will be followed by a dive into the practical challenges of the implementation itself, a chronological review of the performance, and the importance of the data. Lastly, we'll review what we would have done differently in hindsight.

I have read that Eskom is fixed (May 2024) so this talk may be less relevant by the time it is presented, in which case, I apologise.

Keywords: *IoT, Solar, Combinatorial Optimisation, Data streaming*

The TruEnd-procedure: Treating trailing zero-valued balances in credit data

Tuesday, 27 Aug, 11h20

Roelinde Bester[1], Tanja Verster[1], Arno Botha[2]

[1]Centre for Business Mathematics and Informatics, North-West University
[2]National Institute for Theoretical and Computational Sciences (NITheCS)

A novel procedure is presented for finding the true but latent endpoints within the repayment histories of individual loans. The monthly observations beyond these true endpoints are false, largely due to operational failures that delay account closure, thereby corrupting some loans in the dataset with 'false' observations. Detecting these false observations is difficult at scale since each affected loan history might have a different sequence of zero (or very small) month-end balances that persist towards the end. Identifying these trails of diminutive balances would require an exact definition of a small balance , which can be found using our so-called TruEnd-procedure. We demonstrate this procedure and isolate the ideal small-balance definition using residential mortgages from a large South African bank. Evidently, corrupted loans are remarkably prevalent and have excess histories that are surprisingly long, which ruin the timing of certain risk events and compromise any subsequent time-to-event model such as survival analysis. Excess histories can be discarded using the ideal small-balance definition, which demonstrably improves the accuracy of both the predicted timing and severity of risk events, without materially impacting the monetary value of the portfolio. The resulting estimates of credit losses are lower and less biased, which augurs well for raising accurate credit impairments under the IFRS 9 accounting standard. Our work therefore addresses a pernicious data error, which highlights the pivotal role of data preparation in producing credible forecasts of credit risk.

Keywords: *Data errors, Optimisation, Decision analysis, Credit risk modelling, Loss Given Default, Survival analysis*

Mobile clinic deployment in the Witzenberg region of South Africa

Wednesday, 28 Aug, 09h20

Hannah Callaghan, Linke Potgieter
Department of Logistics, Stellenbosch University

In South Africa, the Department of Health has introduced mobile clinics to improve access to healthcare for rural communities. In this study, we consider the Witzenberg region, where six mobile clinics have been deployed. The aim was to determine routes and schedules to improve the workload balance, fairness, and transportation cost, while ensuring patients get satisfactory care. The problem is modelled in three phases using both primary qualitative data and secondary quantitative data. In phase 1, a multi-vehicle routing problem is formulated to construct feasible daily routes for the mobiles. Phase 2 distributes the daily routes fairly between the mobile clinics to ensure fairness. Finally, in phase 3, a vehicle routing formulation is used to determine a 4-week schedule for each mobile clinic, by using the daily routes obtained during phase 2 as input. Four different service time estimations are used as input, resulting in four different schedules, each with their own advantages and disadvantages, including cost-effectiveness, robustness, fairness, and continuity of care. AHP was then performed with main decision makers to determine their preferred schedule. Final routes and schedules were determined based on model results, AHP results, and final practical input from the decision makers, resulting in an improvement in workload balance, a 23% reduction in total distance travelled and willingness by decision makers to implement the changes.

Keywords: *Vehicle routing, decision analysis, practice of OR*

On the impact of distributional changes on the probability of default in credit risk scoring

Tuesday, 27 Aug, 11h00

Johan du Pisanie

Viewwave

Credit risk scorecards are essential tools used to assess the probability of a customer defaulting on their payments. These models play an important role in determining whether to grant credit to an applicant as well as the cost of this credit. To ensure that these models remain accurate, it is crucial to periodically test the hypothesis of population stability. This hypothesis states that the distribution of the attributes of customers when the scorecard is reviewed is the same as was the case when the scorecard was developed. If a substantial change in the distribution of one or more attributes is observed, then the scorecard may need to be redeveloped; a process that is both costly and time-consuming. Various measures of population stability are employed in practice, with several new metrics being proposed in recent literature. Although the existing metrics are certainly useful in determining whether the distribution of the attributes has changed over time, no metric currently in use quantifies the impact of the change in the distribution of a given attribute on the overall probability of default of the population. That is, no metric is available for assessing the impact of a distributional change on the overall risk level of the population. This presentation is based on a recent paper and introduces a new population stability metric, aimed specifically at quantifying this impact.

Keywords: *credit scoring, population stability*

A new approach to b-coloring of regular graphs – for the purpose of organizing trainings

Wednesday, 28 Aug, 14h00

Magda Dettlaff[1], Hanna Furmanczyk[1], Iztok Peterin[2], Adriana Roux[3], Radoslaw Ziemann[1]

[1]University of Gdansk
[2]University of Maribor
[3]Stellenbosch University

Let G be a graph and $c : V(G) \to \{1, \ldots, k\}$ a proper k-coloring of G, i.e., $c(u) \neq c(v)$ for every edge uv from G. A proper k-coloring is a b-coloring if there exists a vertex in every color class that contains all the colors in its closed neighborhood. The maximum number of colors k admitting b-coloring of G is the b-chromatic number $\chi_b(G)$. We present two separate approaches to the conjecture posed by Blidia et. al [Discrete Appl. Math., 157(8):1787–1793, 2009] that $\chi_b(G) = d + 1$ for every d-regular graph of girth at least five except the Petersen graph. For an application of the model of b-coloring consider the following example. Assume that we want to organize an engaging training session in company X. We want to divide people into groups where no two persons know each other or work in the same department. But we also want to ensure that in each such a group there is at least one person who knows at least one person in each other group. Such a situation will help to ensure a good information exchange between groups and, as a result, better integration of all teams. Because we do not want individual groups to be too large, we want to divide employees into the maximum number of groups that provide the above-described property. Note that such a problem can be modeled by an appropriate graph G and its b-coloring with $\chi_b(G)$ colors. Vertices of the graph G correspond to workers and two vertices are adjacent if and only if the corresponding employees work in the same department or know each other. The condition that there is a person in each group who will ensure good communication with the other teams is implemented by b-vertices in each of the color classes in a b-coloring of such a graph. We are interested in maximizing the number of groups for training, so we are interested in a b-coloring with $\chi_b(G)$ colors of the obtained graph. We are attacking the above mentioned conjecture by two different approaches. First we deal with the case when there exists a vertex x that does not belong to too many six cycles which have all vertices on a distance at most two from x. In particular, we prove conjecture when there is such a vertex with at most five mentioned six cycles. The second approach assumes that all the neighbors of two neighbors of x are at distance at most two to x. With this we add a new brick to the confirmation of Conjecture

Keywords: *b-coloring, regular graph*

A computer vision framework towards automated scene reconstruction analysis for the use of energy auditing

Wednesday, 28 Aug, 15h40

Sarah-lee de Greeff
Department of Industrial Engineering, Stellenbosch University

In this paper, multiple object detection algorithms were investigated to determine their suitability for the application within the energy audit domain. A comparative analysis was conducted using a custom dataset composed of energy appliances typically found in school buildings, obtained from previous energy audits. To ensure fairness and optimal performance, the dataset was augmented and balanced appropriately. Six state-of-the-art algorithms were evaluated, namely: Faster RCNN, DeTR, YOLOS, YOLOv7, YOLOv8, and YOLOv9. These algorithms were assessed based on their Average Precision and their proficiency in identifying small, medium, and large objects within image frames. The results indicated that the YOLOv8 model outperformed the others with a narrow margin, including the more recently developed YOLOv9, in this specific application. This research lays the foundation for automating the energy auditing process using computer vision, with the subsequent integration of an object tracking model depending heavily on the accuracy of the selected object detection algorithm.

Keywords: *computer vision, object detection, object tracking, energy auditing*

The distance-limited Firefighter problem

Wednesday, 28 Aug, 15h20

Joerg Kalcsics[1], Marta Baldomero-Naranjo[2], Antonio Rodriguez-Chia[2], Cat Wedderburn[1]

[1]University of Edinburgh
[2]Universidad de Cadiz

Let G be an undirected graph, F be a subset of f vertices, and d be a given number of firefighters. In the Firefighter game, a fire breaks out on all vertices in F at time zero. In each subsequent time step, first the d firefighters can defend one vertex each from catching fire. Afterwards, the fire spreads from each burning vertex to every adjacent vertex that is neither burning nor defended. Both burning and defending are permanent, i.e., a vertex remains burning or defended for the rest of the game, and the game ends when the fire can not spread any further. The goal of the Firefighter game is to find a defence strategy that minimizes the number of burning vertices.

In the classical version of the Firefighter game, each firefighter can move arbitrarily around the graph. In this talk, we modify this assumption and allow each firefighter to move only a certain distance between time steps. As a result, defence strategies for the classical Firefighter game are likely no longer feasible for this distance-limited version; moreover, a defender may not even be able to defend vertices in each time step.

The classical Firefighter game with $f = d = 1$ is NP-hard even on trees of degree at most three when the fire starts on a vertex of degree three. In this talk we present some polynomial-time algorithms for the distance-limited firefighter game on paths.

Keywords: *Graph, Firefighter Game, Polynomial-time Algorithms*

DTP Boost: A novel decision-making tool for DTPCV booster introduction in LMICs

Monday, 26 Aug, 15h00

Retselisitsoe Monyake

Modelling and Simulation Hub, Africa (MASHA), University of Cape Town

Low- and middle-income countries (LMICs) face resource constraints when considering the introduction of booster doses for diphtheria-tetanus-pertussis containing vaccines (DTPCV). Despite recommendations for these boosters, there is a lack of country-specific information on health and economic benefits to guide decision-making.

This presentation introduces DTP Boost, a web-based decision-making tool developed through a collaboration between the Modelling and Simulation Hub, Africa (MASHA) at the University of Cape Town, Centers for Disease Control and Prevention (CDC), African Field Epidemiology Network (AFENET), and the Uganda National Expanded Program on Immunisation (UNEPI). DTP Boost combines epidemiological and economic models to generate tailored projections for selected countries. The tool allows users to design and compare multiple vaccination strategies, exploring the health impact, budget impact, and cost-effectiveness of introducing DTPCV booster doses. It accounts for country-specific factors such as existing vaccination programmes, current disease burden, and health system characteristics.

The project's collaborative nature ensured that end-user needs and perspectives guided tool development. As part of the development process, we conducted a series of in-country workshops with UNEPI to pilot the tool, held in August 2022, May 2023, and November 2023. These workshops, and regular engagement in between, facilitated collaborative refinement of the tool based on real-world application and user feedback. The resulting DTP Boost tool provides a user-friendly, interactive platform to support policymakers to make informed decisions about DTPCV booster dose implementation, tailored to their country's unique context and resource constraints. I will demonstrate the DTP Boost tool then discuss the design and technical aspects of the tool.

Keywords: *Web-based decision-making tool, Health impact, Budget impact, Policymakers*

Enhancing IoT security: Integrating AutoML and SHAP for improved anomaly detection and interpretability

Wednesday, 28 Aug, 13h20

Nathan Morgenrood, Tiny du Toit, Linda Redelinghuys
North-West University

The Internet of Things (IoT) enhances efficiency and automation through the interconnection of diverse devices. However, significant concerns regarding security and privacy, often overlooked, compromise data confidentiality and integrity. Traditional IoT security measures may also lack transparency, undermining trust. This study addresses these issues by enhancing the accuracy and interpretability of anomaly detection models in IoT security through an evolutionary algorithm combining automated machine learning (AutoML) and SHapley Additive exPlanations (SHAP) within multilayer perceptron architectures. The CICIoT2023 dataset utilised consists of 33 distinct attacks executed on an IoT topology comprising 105 devices. This study focuses on eight of these attacks. SHAP was first employed for feature selection, where the top five influential features out of 46 were chosen, resulting in a significant reduction in dataset size. On the full dataset, three models were produced, with the best achieving an accuracy of 97.55%. The reduced dataset resulted in the production of eight models, with the best achieving 99.85% accuracy, surpassing the compared study's best model at 93.13%. SHAP was lastly employed to gain further insight into the models, which is useful for model transparency and future model diagnostics and improvements. The significant reduction in dataset size and the superior accuracy of these models highlight the efficacy of combining AutoML with SHAP for feature selection and model interpretability, contributing to more resilient and intelligent anomaly detection models and IoT security solutions.

Keywords: *Anomaly Detection, Automated Machine Learning, SHapley Additive exPlanations, Internet of Things, Model Interpretability, Multilayer Perceptron*

Rapid modelling to inform diphtheria outbreak response vaccination in Nigeria – A case study

Monday, 26 Aug, 14h20

Jared Norman, Rachel Hounsell, Sheetal Silal
Modelling and Simulation Hub, Africa (MASHA), University of Cape Town

Though vaccination coverage for the third dose of diphtheria-tetanus-pertussis-containing vaccine (DTP3) has improved from 33% in 2017 to 57% in 2021, the 2022 diphtheria outbreak in Nigeria was underpinned by the disproportionately large numbers of un- and under vaccinated children and adolescents. Nigeria accounts for the highest burden of zero dose children (as measured by those who have not received any dose of DTP1) with more than 2.3 million children in 2023. Diphtheria is not routinely reported through the Integrated Disease Surveillance and Response (IDSR) system in Nigeria. This makes it significantly harder to gather estimates of disease burden and background protection rates.

The Government of Nigeria (GoN) requested modelling support from the US CDC to guide prioritisation of reactive vaccination efforts with inadequate numbers of DTP-Hepatitis B-Haemophiles influenzae Type B (Hib) (Pentavalent) and tetanus-diphtheria (Td) vaccine, which was due to arrive in-country in batches in late 2023. The incoming vaccine was intended to support three rounds of outbreak response vaccination campaigns to reach targeted age groups and geographies with three doses of age-appropriate vaccine.

Given our ongoing collaboration with the US CDC, MASHA was asked if a model we had previously developed could be used to support decision making. The specific policy question was how to allocate the vaccines to the states in a way which prevents deaths.

We were able to work with speed and precision to provide an answer to this question using a data-driven approach. Our initial model involved using the available relevant data to categorise each LGA according to a data-driven decision tree with custom decision rules. We built some tooling to facilitate communication with stakeholders and guide the discussion during meetings. Another modelling group answered the question using a different technique and it was useful to compare the two models.

In this talk I will give a detailed account of the full process we undertook to answer the policy question, how we implemented this solution and how we communicated our results.

Keywords: *Modelling, Policy, Health*

Using machine learning to predict the risk of late effects of a South African childhood cancer survivor cohort

Monday, 26 Aug, 15h20

Lenè Nortje[1], Jacomine Grobler[2], Anel van Zyl[3], Mariana Kruger[3]

[1]Institute of Biomedical Engineering, University of Stellenbosch
[2]Department of Industrial Engineering, University of Stellenbosch
[3]Department of Paediatrics and Child Health, University of Stellenbosch

Over the past 50 years, significant improvement in childhood cancer treatment has led to an increasing survivor population. However, childhood cancer survivors (CCSs) have a substantial risk of developing potentially serious and life-threatening late effects. Late effects are physical or psychosocial complications developing due to the previous cancer diagnosis or the treatment received. They may develop during treatment or up to decades after treatment. Research is currently focusing more on the CCSs' late effects risk, including the impact of these complications on their health-related quality of life (HRQoL). Identifying the risk of developing late effects is important for planning long-term follow-up (LTFU) care, maximising the chances of survival, and potentially minimising morbidity and mortality. Applying machine learning algorithms to predict CCSs' late effects risk could aid in appropriate LTFU care planning, early intervention, and preventative care.

This study used a dataset comprising 160 South African CCSs whose late effects were graded according to the Common Terminology Criteria for Adverse Events (CTCAE) classification. Following data analysis, clustering- and classification algorithms were applied to select the best strategy for predicting the risk and severity of late effects. Three clustering algorithms, namely, k-means, agglomerative hierarchical clustering (AHC) and density-based spatial clustering of applications with noise (DBSCAN) were independently applied to a data subset to cluster the CCSs according to the severity of the late effects. The clustering was necessary because the target feature required for the supervised machine learning algorithms was not a single obvious feature already present in the dataset. Therefore, to comprehensively report the extent to which late effects manifested among CCSs, the features related to the grade and number of late effects were used during the clustering. These newly created clusters formed the target feature for the supervised machine learning algorithms.

The results revealed the k-means model outperformed the AHC and DBSCAN. Following the clustering, the performance of six classification algorithms, including a decision tree, random forest, logistic regression, support vector machine, gradient boosting and mutilayer-perceptron, were compared to predict the late effects risk of the CCSs. Five performance metrics were measured to evaluate the respective classification models. Support vector machine and random forest performed the best in terms of balanced accuracy (0.7199 ± 0.033). The random forest and gradient boosting performed best in terms of precision (0.8302 ± 0.01). The gradient boosting performed best in terms of accuracy (0.8490 ± 0.019), recall (0.8490 ± 0.019), and $F1$ (0.8405 ± 0.024). The gradient boosting model ranked anthracycline dose, radiotherapy dose, age at study visit, treatment modalities, and body mass index as the top five important features for predictions. A domain expert (a paediatric oncologist) validated the results and findings. This domain expert was involved in feature se-

lection to ensure relevant features were selected, as well as confirmation that the results were valid from a medical perspective.

This research is significant because it is the first application of machine learning algorithms to predict the risk of late effects of a South African CCS cohort.

Keywords: *Machine learning, childhood cancer survivors, late effects*

An overview and comparison of heuristics for the generalized bin packing problem

Wednesday, 28 Aug, 14h20

Funanani Netshitungulu, Rosephine Georgina Rakotonirainy
Department of Statistical Sciences, University of Cape Town

The Generalized Bin Packing Problem consists of loading a set of profitable non-compulsory and compulsory items into bins so that the total cost, given by the difference between the total cost of selected bins and the total profit of the loaded items, is a minimum. Such a problem is mainly encountered in the transportation industry which has grown massively over the years. In this presentation, an overview is provided with respect to a number of heuristic solution approaches for this problem. The relative performance of these algorithms in respect to a large set of benchmark instances are compared in terms of the solution quality. An instance space analysis is also performed to test the effectiveness of the heuristics with regards to the problem instance characteristics. The relative power of each algorithm is analysed and insights are drawn from the analysis on their strengths and weaknesses.

Keywords: *Bin-packing problem, combinatorial optimization, heuristics, instance space analysis*

Multi-objective optimization of the generalized bin packing problem

Tuesday, 27 Aug, 12h00

Andrea Plumbley, Rosephine Georgina Rakotonirainy
Department of Statistical Sciences, University of Cape Town

The generalized bin packing problem (GBPP) involves the allocation of compulsory and non-compulsory items into a set of bins. The items have characteristics such as weight, width, length and due date, while the bins have characteristics such as capacity and cost. Compulsory items must be packed while non-compulsory items add additional profit if allocated. The main objective of this problem is to minimize cost which usually corresponds to minimizing the number of bins used. However, in many real-world applications there may be multiple objectives that are trying to be met and these may be competing. For example, one may wish to minimize the load imbalance or lateness of the items in addition to minimizing overall cost.

Classical methods for solving such problems involve combining the objectives into a single objective or making some of the objectives into constraints with associated goals. Both approaches require one to have prior knowledge of the decision makers preferences in terms of the trade-off between the different objectives. A better approach would be to present a decision maker with a set of solutions, without requiring prior information, from which they can then make a trade-off decision.

Multi-objective evolutionary algorithms (MOEAs) have shown promise when it comes to multi-objective optimization. These algorithms can produce solution sets and handle large scale problems, finding solutions faster than exact methods. MOEAs simulate the process of evolution and make use of populations which iteratively create new generations using selection, recombination, mutation and replacement methods. These steps of the MOEA ensure that the fittest individuals are used to produce subsequent solution populations. As the generations continue, the candidate solutions move towards the optimal solution. An MOEA has been applied to the multi-objective GBPP in order to find near optimal allocation of the items into the available bins. The algorithm implemented makes use of non-dominated sorting and crowding distance measures to decide which solutions should be selected for the next generation. This presentation will outline the multi-objective GBPP formulation, the steps of the implemented MOEA and the preliminary results of the research.

Keywords: *Multi-objective optimization, generalized bin packing problem, evolutionary algorithms, non-dominated sorting, crowding distance*

A framework for evaluating clinical data using graph-based link prediction

Wednesday, 28 Aug, 09h00

Himil Parshotam, Stephan Nel
Department of Industrial Engineering, Stellenbosch University

Clinical data repositories, the most prevalent of which are electronic health records, typically comprise a broad range of information pertaining to various fields within the medical domain. Clinical data fields encapsulate distinct notions and are typically expressed in various formats — indicative of the specialised nature of the clinical domain. Data range from complex medical imaging data employed for diagnosis, therapy planning, intraoperative navigation, and post- operative monitoring, to clinical summaries derived from information collected during doctor- patient consultations, such as conditions, medications, and patient demographics. These data sources may be characterised by a notable degree of interconnectedness due to complex interactions and relationships that are embedded in respect of the various medical concepts.

In this thesis, a generic framework is proposed for constructing and analysing a knowledge graph derived from clinical data. The proposed framework conceptually represents a unified pipeline (or architecture) that may be employed towards transforming raw clinical data into an appropriate knowledge graph representation, and subsequently performing graph-based analysis, i.e. link prediction. The proposed framework comprises three functional components, each of which addresses an integral step in the overarching process. Two main types of use cases may be realised by means of the framework's implementation, the first of which relates to the prediction of medical conditions in order to identify misdiagnosed conditions, while the second use case pertains to the prediction (*i.e.* suggestion) of medication for prescription purposes. Due to the challenges associated with access to real-world clinical data (attributable to privacy and confidentiality), reputable synthetic data are considered in order to demonstrate the methodological utility of the proposed framework. More specifically, two different data sets are considered, each of which differs in respect of the underlying clinical context and in terms of complexity. Three computerised instantiations are carried out, each of which varies in respect of different data sets that are subjected to the modelling pipeline and/or the clinical use case under consideration.

An algorithmic verification study is carried out prior to these instantiations in order to verify the functional correctness of the sixteen link prediction algorithms considered. The three main paradigms of link prediction algorithms considered are common neighbour -based algorithms, machine learning classifiers, and graph neural networks. Hyperparameter tuning is also carried out in order to identify suitable algorithmic configurations in respect of the different link prediction algorithms. During each computerised instantiation, algorithmic performance is evaluated both quantitatively (including statistical analyses) and qualitatively by means of appropriate visualisations and contextual reflections. The algorithmic output is also further contextualised in respect of practical insight pertaining to the aim of the respective use cases. Furthermore, the framework is validated by means of a subject matter expert, during which the methodological utility of the proposed framework

and its applicability to real-world operations is corroborated.

Keywords: *Knowledge graphs, diagnosis prediction, graph-based link prediction, deep learning*

Decision making in the current military environment

Wednesday, 28 Aug, 13h40

Winnie Pelser
Council for Scientific and Industrial Research (CSIR)

Today's worldwide environment includes political turmoil, humanitarian crises, human rights violations, significant protest activities, natural disasters, climate change, economic woe and recession, terrorist activities, immigration and refugee issues, civil wars, military conflicts, and even full- scale war. The operational situation for the military within this bigger picture, require decision makers to embrace improvisation, reflection and proper training. Effective decisions must be based on a rational analysis of a specific scenario. Regardless of good intensions, things often go very wrong. Many military operations in the past failed due to poor decisions. The bad part is that if a military or security operation fail, lives are often lost. This places a high premium on sound decision making. The underlying cause of failed military or security operations have been attributed to a variety of explanations, but a failure in human thinking or human cognitive processes may be the underlying cause. The effect of stress adds to the difficulty of making good and ethical decisions. What can be done to assist soldiers to make sound and ethical decisions under uncertain and stressful circumstances as well as to avoid decision failures that stem from stressful situations.

Keywords: *Decision making, military environment, stressors in uncertain making*

The rich heterogeneous dial-a-ride problem with trip time prediction

Wednesday, 28 Aug, 13h00

Helena Ramalhinho[1], Laura Portell[1,2]
[1]University Pompeu Fabra
[2]Universitat Politecnica de Cataluny

The dial-a-ride problem (DARP) involves designing vehicle routes to fulfill the door-to-door transportation requests of users where the goal is to minimize costs while satisfying transportation requests. In this paper, we introduce the rich heterogeneous DARP, which extends the generalized heterogeneous DARP to consider a fleet of buses and taxis, multiple depots, time windows at pickup and delivery locations, maximum ride and waiting times, and the possibility of an accompanying person. Our approach is based on a real service in Barcelona, and we also consider the variation in trip duration based on the time of day and day of the week. A predictive model is developed using machine learning techniques to estimate trip durations accurately. We apply our proposal to the daily door-to-door transportation of people with reduced mobility in Barcelona and demonstrate its superiority in terms of costs and quality of service by using the Gurobi optimizer. Additionally, we provide an analysis of the consequences of varying certain features on the costs and quality of service.

Keywords: *dial-a-ride problem; heterogeneous vehicle; people transportation; vehicle routing*

Reflections on COVID-19 modelling in South Africa: Lessons learned and looking forward

Monday, 26 Aug, 14h40

Sheetal Silal

Modelling and Simulation Hub, Africa (MASHA), University of Cape Town

SARS-CoV-2 stands as the most disruptive virus encountered by the world in the past hundred years. Without prior understanding the behaviour of this disease and the population impact of various pharmaceutical and non-pharmaceutical control measures, mathematical modelling played an important role chart the course of the disease and provide decision-making support to our nation's leaders and healthcare workforce. Similarly, mathematical modelling was crucial in shaping the global response to the pandemic. Given the swiftly evolving nature of the outbreak both worldwide and specifically in South Africa, it was necessary to continuously update mathematical models as new information emerged. Factors such as changes in testing policies, contact tracing efforts, and hospitalization criteria significantly impacted the number of detected cases, the hospitalisations, and the fatalities recorded.

This presentation will trace the development of the National COVID-19 Epi Model (NCEM) over the course of the pandemic, highlighting how assumptions have evolved and what lessons have been learned. It will conclude by discussing future considerations for modelling and policy-making in the context of COVID-19 and other diseases, emphasizing the importance of flexibility and adaptability in the face of ongoing and new challenges.

Keywords: *mathematical modelling, COVID-19, evidence-based decision-making*

A framework for multimodal hydrogen refuelling infrastructure design

Tuesday, 27 Aug, 11h40

Christa Searle[1], Kit Searle[2], Joerg Kalcsics[2], Philip Greening[1]

[1]Centre for Logistics & Sustainability, Edinburgh Business School, Heriot-Watt University

[2]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh

With the pressured timescale to determine effective and viable net zero solutions within the transport sector, it is important to understand the extent of implementing a new refuelling infrastructure for an alternative fuel, such as hydrogen. Research questions arising involve estimating the minimum number of refuelling stations required, determining the location and size of stations, while also configuring the hydrogen production supply and delivery to stations. The Analysis of a Strategic Hydrogen Refuelling Infrastructure (ATHENA) framework has been developed to determine and recommend a hydrogen refuelling infrastructure for effective deployment. The ATHENA framework involves a three-phase approach which includes demand data analysis, an optimisation model and analysis, and finally, an agent-based model and analysis.

In this research, we extend the data analysis and optimisation model within the ATHENA framework to consider an integrated multimodal strategy that allows for the delivery of hydrogen by pipelines. The demand data analysis phase provides as output a spatio-temporal mapping of the demand across the different transport modes. This forms part of the input data to the optimisation model which determines the locations and capacities of hydrogen refuelling stations as well as the configuration of hydrogen production supply and delivery to each station, while minimising the overall set-up and operational cost. For each station, hydrogen production supply can be localised or centralised and, in the case of centralised production, the hydrogen can be delivered by tube trailers or a hydrogen pipeline. The design of extensions to an existing hydrogen pipeline to serve the refuelling network is incorporated within the optimisation model.

A case study is performed for the North of England in designing a multimodal hydrogen refuelling infrastructure serving heavy goods vehicles, buses, refuse collection vehicles, and port handling equipment. The case study considers a set of scenarios to analyse the effect of the different hydrogen production supply and delivery options.

Keywords: *Hydrogen refuelling infrastructure; Data analysis; Discrete optimisation; Agent-based simulation modelling*

A practical method for joint replenishment under supplier constraints

Wednesday, 28 Aug, 09h40

Thomas Simon, Jan van Vuuren
Department of Industrial Engineering, Stellenbosch University

In order to maintain their competitive advantage, retailers are regularly required to review their inventory management strategies with a view to save time, cut costs, and enhance labour efficiency geared towards faster delivery. A challenging aspect of inventory management emanates from the interaction between upstream distributors and retail suppliers, and relates to suppliers' requirements for justifiable order volumes or acceptable container utilisation during inventory replenishment. In this presentation, we propose a simulation-optimisation modelling approach in support of the complicated replenishment decisions faced by inventory managers who are subjected to these requirements. The modelling approach is aimed at avoiding excess inventory by consolidating inventory replenishment orders across products experiencing different demand profiles. Inventory replenishment cycles of products supplied by a single supplier are simulated with a view to determine appropriate safety stock durations for each product. The simulation model incorporates product-specific data, supplier parameters, justifiable order volume and acceptable container utilisation constraints, as well as a specified overall target service level. An optimisation capability is embedded over the simulation model in order to optimise safety stock durations heuristically, while ensuring the achievement of the specified target service level and minimising capital tied up in inventory with high statistical confidence. It is finally demonstrated in a real-world case study that the modelling approach is capable of substantial cost savings during inventory replenishment.

Keywords: *Simulation-optimisation, Safety stock, Multi-product, Joint replenishment, Container filling, Minimum order value*

Analysing student progression by means of a hybrid simulation modelling paradigm

Wednesday, 28 Aug, 10h00

Sha-abaan Slamang, Lieschen Venter
Department of Logistics, Stellenbosch University

Stellenbosch University (SU) aims to be one of the leading research institutions on the African continent. High school matriculants (both locally and internationally) can apply to SU with the hope of graduating through one of SU's ten faculties. The selection of students for enrolment into a degree programme in the Economic Management Sciences (EMS) Faculty is based on academic merit.

For a student to progress to the each academic year, they must obtain the minimum academic credits required to continue. This study provides a simulation model of SU's EMS Faculty. A hybrid paradigm brings together the micro-analysis functionality of agent based simulation modelling and the macro-analysis functionality of system dynamics. This model can assist in setting policies and intervention strategies to better understand the challenges students face and how can the university can assist students so that the maximum number of students graduate in minimum time.

Keywords: *Simulation modelling, education systems, agent based, system dynamics, policy setting*

A branch-and-cut algorithm for the κ-connected forest-star problem

Wednesday, 28 Aug, 15h00

Kit Searle, Joerg Kalcsics

School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh

Given a set of vertices, denoted by $\mathcal{V}$, a complete digraph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ and a set of roots, denoted by $\mathcal{R} \subset \mathcal{V}$, the κ-connected forest-star problem is defined as follows: Find a set of directed trees each rooted at an element in $\mathcal{R}$ such that every vertex in $\mathcal{V} \setminus \mathcal{R}$ is either connected to exactly κ directed trees each with a unique root in $\mathcal{R}$ or is assigned to another vertex which is on a tree, resembling a star. The objective of the problem is to minimise the cost of the directed trees and assignment costs. This problem has applications in the design of a telecommunications infrastructure where the vertices in $\mathcal{V} \setminus \mathcal{R}$ can be viewed as customers and roots in $\mathcal{R}$ can be viewed as data centres. The forest topology can therefore be viewed as the backbone of the network while the assignments can be viewed as wireless transmission. In such an application it is important that the customers are connected to exactly κ data centres to ensure robustness.

In this presentation, I will present a two-index formulation for solving this problem, based on network connectivity constraints which are exponential in the cardinality of $\mathcal{V}$. I will then present some valid inequalities and a branch-and-cut algorithm for solving this problem followed by some preliminary results.

Keywords: *Network design, Connected facilities, Branch-and-cut*

Using automated supervised autoencoders to detect IoT botnet attacks

Wednesday, 28 Aug, 16h00

Hano Strydom, Tiny du Toit, Henri van Rensburg
North-West University

Internet of Things (IoT) devices, such as smart home devices and industrial sensors, are interconnected through telecommunication. Radio, satellite and optical fiber technologies facilitate data transmission over long distances. Unfortunately, this introduced new challenges, such as IoT botnet attacks, when hackers use infected computers and other devices connected to the Internet to carry out harmful attacks. Traditional detection techniques, such as artificial neural networks (ANN) and logistic regression, were used identify these IoT botnet attacks. Still, it is challenging due to the large amount of data that IoT devices generate, making it more difficult to train models and capture the underlying patterns. This study proposes a supervised autoencoder (SAE) using Bayesian hyperparameter optimisation, feature selection and automated machine learning (AutoML) to predict these botnet attacks. The N-BaIoT dataset, which includes data from various IoT devices under different botnet attacks, is used during this study. It is a relatively large dataset, and therefore, a Random Forest Classifier is used for feature selection, ensuring that only the most relevant features are selected. Stratified sampling decreases the amount of data used to train the model. The optimised SAE model achieved an accuracy of 99.98%, outperforming traditional methods from other studies which used ANNs, and logistic regression performed on the same dataset. This high accuracy shows that the model, created using an SAE combined with Bayesian hyperparameter optimisation, feature selection and AutoML, offers a promising solution for detecting IoT botnet attacks.

Keywords: *Artificial Intelligence, Botnet Attack, IoT, Network-based IoT botnet attack, Supervised Autoencoder*

A new inventory routing problem with route and schedule unpredictability

Monday, 26 Aug, 16h00

Marlize Helene Visser, Jan van Vuuren
Department of Industrial Engineering, Stellenbosch University

A new inventory routing problem (IRP) variant, called the heterogeneous fixed fleet IRP with time-windows (HeFIRPTW) and with route and schedule unpredictability, is introduced in this presentation. The application of an IRP in a retail supply chain allows for the simultaneous optimisation of delivery schedules, the set of delivery vehicle routes to be followed, and the delivery quantities to be delivered to customers. The adoption of a vendor-managed inventory strategy has the potential not only to improve upon the transportation costs incurred in a supply chain, but may also reduce the costs associated with inventory storage and stock-outs. We propose the incorporation of simultaneous route and schedule unpredictability to mitigate the inherent safety and security threats associated with the transportation of valuable goods. A novel mixed-integer linear programming formulation of a bi-objective IRP is presented, explicitly integrating both route and schedule unpredictability within routing solutions. The HeFIRPTW with route and schedule unpredictability addresses the challenge of distributing valuable goods in a retail supply chain by optimising delivery routes and schedules with a view to minimise the operational costs incurred, whilst also ensuring that routes are not traverses regularly within the same time intervals and that customers are not visited during overlapping time intervals. The feasibility of adopting an exact approach, utilising the ϵ-constrained method, is demonstrated in respect of a set of small, adapted benchmark instances.

Keywords: *Combinatorial optimisation, Inventory routing problem, Mixed-integer linear programming, Multi-objective optimisation, Safety and security constraints*

Predicting learner progress for the South African public high school system

Monday, 26 Aug, 16h20

Lieschen Venter, Maymarie van den Heever, Esranel Becker, James Bekker

Departments of Logistics and Industrial Engineering, Stellenbosch University

The South African education system offers diverse paths to the National Senior Certificate through both public and private schooling. This study employs an integrated approach that merges machine learning and agent-based modelling to simulate learner progression in public high schools. Using data from the 2019 General Household Survey in South Africa, factor analysis is first conducted to identify and quantify the principal characteristics defining learners. These factor scores then train an XGBoost machine learning model, which is integrated within an agent-based framework to simulate learner progression from Grades 8 to 12. The model's accuracy is validated against the Learner Unit Record Information and Tracking System (LURITS) dataset, demonstrating its efficacy in predicting educational outcomes, particularly in lower grades. Demonstrating its efficacy, it could subsequently serve as a strategic sandbox for evaluating the effectiveness of educational interventions.

Keywords: *Agent-based modelling, machine learning, factor analysis, learner success*

Author Index