PROCEEDINGS OF THE
INTERNATIONAL STATISTICS CONFERENCE 2024

Unleashing the Power of Data: Harnessing the Synergy Between Statistics and Data Science

Abstracts

28th and 29th of December 2024 at the Cinnamon Lakeside Colombo, Sri Lanka

Organized by

Institute of Applied Statistics, Sri Lanka

University of Sri Jayewardenepura, Sri Lanka

Simon Fraser University, Canada

ISSN: 3084-8288 **e-ISSN: 3084-8296**

Endorsed by

Gold Sponsor

Silver Sponsors

PROCEEDINGS OF THE
INTERNATIONAL STATISTICS CONFERENCE 2024
(ISC 2024)

Unleashing the Power of Data: Harnessing the Synergy Between Statistics and Data Science

The proceedings comprise the abstracts of the Keynote Speeches, Guest Speeches, Invited Talks, Invited Group Talks, Industry Invited Talks, Contributed Oral and Poster Sessions, presented at the International Statistics Conference 2024 (ISC 2024) organized by the Institute of Applied Statistics Sri Lanka (IASSL) together with the local collaborator University of Sri Jayewardenepura (USJ), Sri Lanka and international collaborator Simon Fraser University (SFU), Canada. The conference was held from 28th to 29th December 2024 at Cinnamon Lakeside, Colombo, Sri Lanka. The views and findings of the abstracts are those of the authors and are not meant to reflect those of the organizers.

Published by:
Institute of Applied Statistics, Sri Lanka
2nd Floor, OPA Building,
275/75, Prof. Stanley Wijesundera Mawatha,
Colombo 07.
Sri Lanka.

ISSN: 3084-8288

e-ISSN: 3084-8296

Responsibility for the content of these articles rests squarely upon their respective author(s) and not upon the editorial board or the members of the program committee. Data presented and conclusions reached by the author(s) are for information only and are not for use without independent substantiating investigation on the part of the user.

Preface

It is a profound honor to serve as the Chief-Editor of the Proceedings of the International Statistics Conference 2024 (ISC 2024) in Sri Lanka. ISC 2024 provides a forum for statisticians, decision-makers, and participants from diverse backgrounds, both local and international, to engage in essential discussions on statistical methods and data analysis to address real-world challenges. The conference welcomes contributions from professionals, undergraduates and postgraduates, offering an opportunity to present their research in various areas of statistics.

The submitted abstracts of the conference spanned a wide array of tracks within the field of Statistics/Data Science, including Probability and Statistical Inference, Time Series Analysis and Forecasting, Statistical Modeling, Sampling and Surveys, Bayesian Statistics, Machine Learning and Artificial Intelligence, Econometric Modeling, Biostatistics and Bioinformatics, Sports Analytics, Data Mining and Big Data Analytics, Applications in Agriculture, Financial Analytics and Marketing Data Analytics, etc. A total of 130 submissions were initially received for the conference. The editorial board meticulously managed the desk evaluation process and the allocation of reviewers. Each abstract underwent a rigorous double-blind peer review process, with at least two independent reviewers ensuring adherence to the highest academic standards. The review process involved 94 reviewers from diverse countries, culminating in a 32% rejection rate. The proceedings of ISC 2024 encompass an impressive collection of scholarly contributions, including abstracts from two keynote addresses and two guest speeches delivered by internationally renowned experts, 16 invited talks, 22 invited group talks, 6 invited industry talks, and 77 contributed abstract presentations, reflecting participation from several countries.

The conference is expected to facilitate productive discussions, knowledge sharing, and the exchange of modern statistical techniques while fostering international collaborations. It is hoped that ISC 2024 will play a pivotal role in advancing statistical research, education, and applications across diverse scientific fields, particularly in Sri Lanka.

I extend my sincere gratitude to all authors, members of the international and local program committees, the editorial board, organizing committee members, sponsors, and everyone who contributed to the success of this conference.

Prof. Vasana Chandrasekara
Chief-Editor of ISC 2024

Message from the President of IASSL

It is both a privilege and an honor to serve as the President of the Institute of Applied Statistics, Sri Lanka (IASSL), for the term 2023–2024. As the premier professional organization of statisticians in Sri Lanka, IASSL is dedicated to promoting and advancing the application of statistics to support research, development, education, training and extension services across the nation.

The triennial International Statistics Conference (ISC) is one of IASSL's flagship events and has been a cornerstone of our efforts since its inception. Unfortunately, due to the challenges posed by the COVID-19 pandemic and economic crises in Sri Lanka, we were unable to hold the conference in 2020. However, I am delighted to announce that we are hosting the 4th International Statistics Conference this year.

ISC 2024 brings together statisticians and data scientists from around the globe, providing a vibrant platform to showcase cutting-edge research in statistical theory and applications, as well as innovations in data science. This conference is also a unique opportunity for Sri Lankan statisticians and data scientists to connect with globally recognized experts, fostering collaborations that will contribute to the growth of Sri Lanka’s knowledge economy and drive national development.

I would like to extend my heartfelt gratitude to the co-chairs, co-secretaries, conference editor and all members of the conference committees for their tireless efforts and unwavering dedication in organizing this event.

As we embark on ISC 2024, I hope you find the conference both enriching and enjoyable. May it spark meaningful discussions, inspire groundbreaking research and forge enduring collaborations that advance the fields of statistics and data science.

Thank you for being a part of ISC 2024 and for contributing to its success.

Warm regards,
Dr. Niroshan Withaange
President, IASSL (2023–2024)

Message from the Co-Chairs of ISC 2024

It is with immense pride and pleasure that we warmly welcome you to the International Statistics Conference 2024 (ISC 2024), a flagship event organized by the Institute of Applied Statistics Sri Lanka (IASSL) in collaboration with the Department of Statistics, University of Sri Jayewardenepura, Sri Lanka and Simon Fraser University, Canada. With the theme *"Unleashing the Power of Data: Harnessing the Synergy Between Statistics and Data Science,"* ISC 2024 brings together statisticians, data scientists and researchers from across the globe to highlight the transformative role of data in research, innovation and decision-making.

The conference features an exciting program, including two keynote speeches by globally renowned statisticians—Professor Rob Hyndman of Monash University, Australia, who will also serve as the Guest of Honor and Professor Tim Swartz of Simon Fraser University, Canada. Additionally, two guest speeches will be delivered by distinguished scholars: Professor John W. Emerson of Yale University, USA and Professor Ji-Hyun Lee, President-Elect 2024 of the American Statistical Association and Professor of Biostatistics at the University of Florida College of Public Health. The program also includes 16 invited speeches, 7 invited sessions, 77 contributed presentations and 6 industrial invited talks, offering a diverse array of topics and insights. Adding to the academic excitement, a Case Study Competition featuring two challenging case studies has attracted participation from 28 teams, showcasing the analytical talents of young Statisticians/Data Scientists of the country. In line with the ISC 2024, a pre-conference workshop on Sports Analytics is conducted by a team from the Simon Fraser University.

We are deeply grateful for the collective efforts of our international collaborator, Simon Fraser University, Canada; our local collaborator, the University of Sri Jayewardenepura; the Chief Guest; keynote and guest speakers; invited session speakers; advisory board; local and international scientific committees; conference co-secretaries; the proceedings editor; and the entire ISC 2024 organizing team. Special thanks go to the American Statistical Association (ASA), the National Science Foundation (NSF) of Sri Lanka and all our sponsors for their generous contributions, as well as to the coordinator of the Case Study Competition and the companies that provided case studies and monetary support.

We sincerely hope ISC 2024 serves as a vibrant platform for exchanging ideas, establishing new research collaborations and inspiring impactful innovations, advancing the fields of statistics and data science to new heights. Thank you for joining us and we wish you an intellectually stimulating and enjoyable experience at ISC 2024!

Warm regards,
Prof. C. D. Tilakaratne and Dr. Niroshan Withanage
Conference Co-Chairs of ISC 2024

Message from the Guest of Honor / Keynote Speaker

I am delighted to be joining you at the 2024 International Statistics Conference. This will be my first visit to Sri Lanka and it feels particularly special to me as I have had the privilege of mentoring many talented students from Sri Lanka over the years. In fact, I have supervised more PhD students from Sri Lanka, than from any other country, including Australia! It is a pleasure to be able to enjoy and explore the environment, food and culture they have told me so much about.

Sri Lanka has a long history of data-driven planning. From the ancient kings who tracked seasonal rainfall patterns in order to plan irrigation systems, to the statistical analysis used in the 20th century to help move to an open market with efficient resource allocation. Today, sophisticated prediction models are used to help manage everything from tourism flows to agricultural yields. I've watched with admiration as many of my former students have returned home to contribute to this rich tradition, bringing cutting-edge methodologies while remaining grounded in local realities.

Throughout my career, there has been a great deal of development in the sophistication of the methods used. When I first started working as a statistician, nearly 40 years ago, forecasting was often seen as a simple application of time series analysis. Today, we are integrating machine learning, handling unprecedented amounts of real-time data and developing methods to quantify future uncertainties in ways we never imagined possible. The challenges are bigger, but so are the opportunities, whether we're predicting climate patterns, economic trends, or public health outcomes.

The timing of this conference couldn't be better. We're living in an age where everyone from business leaders to policymakers is hungry for better data analysis. Yet with this increased demand comes increased responsibility. How do we communicate uncertainty effectively? How can we combine statistical rigour with practical usability? How do we ensure our analysis helps deliver social good and does not just drive profits?

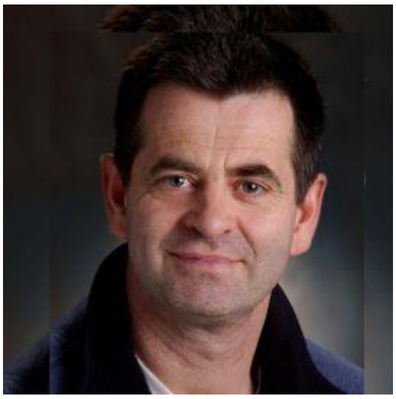
Let's make the most of our time together. Whether you're a seasoned statistician or just starting your journey, whether you work in government, academia, or industry, this is an opportunity to share insights and learn from each other. The breaks are just as important as the talks, as they provide an opportunity to connect with new people, discuss new developments and build relationships that can last a lifetime. Some of my best research ideas have grown out of casual conversations at conferences like this one!

I look forward to learning from all of you and to exploring how we can use statistics and data analysis to build a better future.

Here's to an engaging, thought-provoking and fun conference!

Best wishes,
Rob J Hyndman
Professor in Statistics
Monash University
Australia

Message from the Keynote Speaker



Welcome everyone to ISC 2024 in Colombo.

It is a privilege for me to attend the conference and serve as one of the keynote speakers. I will be speaking on sports analytics, an interesting field for many of us. I thank the Organizing Committee for the opportunity to speak on this growing area in applied Statistics. I hope to convey my excitement for this research area and I hope to encourage others to consider working in Sports Statistics. I also hope that some of you were able to attend the Workshop on Sports Analytics (held the day before the main conference) where spatio temporal data were used to illustrate the future of sports analytics. Imagine the problems that can be addressed when you have access the location of every play on the pitch measured 10 times per second.

It is also a privilege that my university (Simon Fraser University) is the foreign collaborator for the conference. SFU has a long history of collaboration in Sri Lankan Statistics. In fact, there have been more than 10 Sri Lankan students complete PhD's in our department. Some of these graduates have returned home to Sri Lanka and are now working as faculty members in Statistics Departments across Sri Lanka.

In 2004, I was involved in the organization of the first international meeting in Statistics in Sri Lanka: Visions of Future Statistical Methodology. This conference was held in Kandy and was spearheaded by our now departed colleague Professor Basil de Silva. This meeting helped the growth and profile of Sri Lankan Statistics. Since then, I believe that three subsequent international Statistics meetings have been held in Sri Lanka: 2011, 2014 and 2017. The 2024 meeting was originally intended to take place earlier; however, Covid had other plans for us.

I am confident that this meeting will be a great success. I hope that you return with new research ideas and are invigorated with energy to carry on work in the Statistical Sciences.

Best wishes,
Tim Swartz
Professor in Statistics
Simon Frazer University
Canada

Advisory Board

Editorial Board

International Program Committee

- Prof. J. Cao, Department of Statistics and Actuarial Science, Simon Fraser University, Canada.
- Prof. A. Chaturvedi, Department of Statistics, University of Allahabad, India.
- Prof. N. K. Haur, Institute of Mathematical Sciences, Universiti Malaya, Malaysia.
- Prof. K. Kumar, Department of Economics and Statistics, Bond University, Australia.
- Prof. A. Manatunga, Rollins School of Public Health, Emory University, USA.
- Prof. S. Muthukumarana, Department of Statistics, University of Manitoba, Canada.
- Prof. S. Peiris, School of Mathematics and Statistics, The University of Sydney, Australia.
- Prof. G. Perera, Department of Statistics and Actuarial Science, Simon Fraser University, Canada.
- Prof. U. Thayasivam, College of Science & Mathematics, Rowan University, USA.
- A/Prof. P. K. Don, Eberly College of Science, Pennsylvania State University, USA.
- Dr. A. De Silva, Melbourne School of Population and Global Health, The University of Melbourne, Australia.
- Dr. D. Jayasekara, Lee Kuan Yew Centre for Innovative Cities, Singapore University of Technology and Design, Singapore.
- Dr. J. M. T. Jayasinghe, Department of Mathematics, University of Dayton, USA.
- Dr. S. Kodikara, School of Mathematics and Statistics, The University of Melbourne, Australia.
- Dr. M. Mammmadov, School of Information Technology, Deakin University, Australia.
- Dr. L. Munasinghe, School of Computing, Robert Gordon University, UK.
- Dr. K. Nadarajah, School of Economics, University of Sheffield, UK.
- Dr. D. Rajapaksha, Department of Mathematics and Statistics, Texas Tech University, USA.
- Dr. M. Rohan, Wagga Wagga Agricultural Institute, Australia.
- Dr. S. S. M. Silva, Australian Institute of Health Innovation, Macquarie University, Australia.
- Dr. L. Wijesekara, School of Computing, Engineering and Mathematics, Western Sydney University, Australia.
- Ms. N. E. Dona, Department of Statistics and Actuarial Science, Simon Fraser University, Canada.

Local Program Committee

- Prof. S. Banneheka, Department of Statistics, University of Sri Jayewardenepura.
- Prof. N. V. Chandrasekara, Department of Statistics & Computer Science, University of Kelaniya.
- Prof. W. B. Daundasekera, Department of Mathematics, University of Peradeniya.
- Prof. D. D. M. Jayasundara, Department of Statistics & Computer Science, University of Kelaniya.
- Prof. S. S. N. Perera, Department of Mathematics, University of Colombo.
- Prof. A. J. Pinidiyaarachchi, Department of Statistics and Computer Science, University of Peradeniya.
- Prof. R. M. K. T. Rathnayaka, Department of Physical Sciences & Technology, Sabaragamuwa University of Sri Lanka.
- Prof. S. Samitha, Department of Crop Science, University of Peradeniya.
- Prof. B. M. L. D. B. Suriyagoda, Department of Crop Science, University of Peradeniya.
- Prof. C. D. Tilakaratne, Department of Statistics, University of Colombo.
- Prof. H. D. Weerasinghe, Department of Computer Systems Engineering, University of Kelaniya.
- Prof. P. Wijekoon, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. H. T. K. Abeysundara, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. S. P. Abeysundara, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. M. Atapattu, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. M. Dehideniya, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. K. A. D. Deshani, Department of Statistics, University of Colombo.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.
- Dr. H. A. S. G. Dharmarathne, Department of Statistics, University of Colombo.
- Dr. D. M. P. V. Dissanayaka, Department of Statistics & Computer Science, University of Kelaniya.

- Dr. L. Gamage, Department of Statistics, University of Colombo.
- Dr. A. Gunawardana, Department of Decision Sciences, University of Moratuwa.
- Dr. N. M. Hakmanage, Department of Computer Systems Engineering, University of Kelaniya.
- Dr. J. C. Hapugoda, Department of Organizational Studies, The Open University of Sri Lanka.
- Dr. I. U. Hewapathirana, Software Engineering Teaching Unit, University of Kelaniya.
- Dr. S. Jayalal, Department of Industrial Management, University of Kelaniya.
- Dr. I. T. Jayamanne, Department of Statistics, University of Colombo.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. R. Jayatillake, Department of Statistics, University of Colombo.
- Dr. H. S. Karunarathna, Department of Statistics, University of Colombo.
- Dr. H. W. B. Kavinga, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. B. M. T. Kumarika, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. P. Liyanaarachchi, Department of Statistics, University of Sri Jayewardenepura.
- Dr. R. Lokupitiya, Department of Statistics, University of Sri Jayewardenepura.
- Dr. C. H. Magalla, Department of Statistics, University of Colombo.
- Dr. T. Mahanama, Department of Industrial Management, University of Kelaniya.
- Dr. N. A. D. N. Napagoda, Department of Mathematical Sciences, Wayamba University of Sri Lanka.
- Dr. L. S. Nawarathna, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. R. D. Nawarathna, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. H. A. Pathberiya, Department of Statistics, University of Sri Jayewardenepura.
- Dr. M. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. S. P. Pitigala, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. G. A. C. N. Priyadarshani, Department of Statistics, University of Colombo.
- Dr. D. D. M. Ranasinghe, Department of Electrical and Computer Engineering, The Open University of Sri Lanka.

- Dr. J. Senarathne, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. R. M. Silva, Department of Statistics, University of Sri Jayewardenepura.
- Dr. A. A. Sunethra, Department of Statistics, University of Colombo.
- Dr. P. D. Talagala, Department of Computational Mathematics, University of Moratuwa.
- Dr. T. S. Talagala, Department of Statistics, University of Sri Jayewardenepura.
- Dr. U. Thayasivam, Department of Computer Science and Engineering, University of Moratuwa.
- Dr. H. A. Usoof, Department of Statistics and Computer Science, University of Peradeniya.
- Dr. S. Viswakula, Department of Statistics, University of Colombo.
- Dr. W. A. C. Weerakoon, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. C. Wickramarachchi, Department of Statistics, University of Sri Jayewardenepura.
- Mr. E. R. A. D. Bandara, Department of Statistics, University of Colombo.
- Mr. P. Dias, Department of Statistics, University of Sri Jayewardenepura.
- Mr. H. A. C. S. Hapuarachchi, Sports Sciences and Physical Education, Sabaragamuwa University of Sri Lanka.
- Mr. D. A. Rohana, Department of Mathematics, Kotelawala Defence University.
- Mr. O. Senaweera, Department of Statistics, University of Colombo.

Organizing Committee

Conference Co-Chairs:

- Prof. C. D. Tilakaratne, Department of Statistics, University of Colombo.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.

Conference Co-Secretaries:

- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.

Sponsorship Committee:

- Dr. C. Wickramarachchi, Department of Statistics, University of Sri Jayewardenepura.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.
- Prof. T. Sivananthawerl, Department of Crop Science, University of Peradeniya.
- Dr. T. S. Talagala, Department of Statistics, University of Sri Jayewardenepura.
- Dr. P. D. Talagala, Department of Computational Mathematics, University of Moratuwa.

Publicity Committee:

- Dr. R. M. Silva, Department of Statistics, University of Sri Jayewardenepura.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.
- Mr. T.R Withanage, Department of Statistics, University of Sri Jayewardenepura.
- Ms. G. A. T. Kaveesha, Department of Statistics, University of Sri Jayewardenepura.
- Ms. H. D. R. M. Fernando, IASSL.

Proceedings Committee:

- Prof. N. V. Chandrasekara, Department of Statistics & Computer Science, University of Kelaniya.
- Prof. C. D. Tilakaratne, Department of Statistics, University of Colombo.
- Dr. H. W. B. Kavinga, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. H. A. Pathberiya, Department of Statistics, University of Sri Jayewardenepura.
- Dr. K. A. D. Deshani, Department of Statistics, University of Colombo.
- Dr. I. U. Hewapathirana, Software Engineering Teaching Unit, University of Kelaniya.
- Dr. J. Senarathne, Department of Statistics and Computer Science, University of Peradeniya.
- Prof. S. Samitha, Department of Crop Science, University of Peradeniya.

CMT Management Committee

- Dr. H. W. B. Kavinga, Department of Statistics & Computer Science, University of Kelaniya.
- Prof. N. V. Chandrasekara, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. N. M. Hakmanage, Department of Computer Systems Engineering, University of Kelaniya.
- Ms. B. R. P. M. Basnayake, Department of Statistics & Computer Science, University of Kelaniya/ University of Peradeniya.

Pre-Conference Committee:

- Dr. M. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. P. Liyanaarachchi, Department of Statistics, University of Sri Jayewardenepura.
- Dr. R. M. Silva, Department of Statistics, University of Sri Jayewardenepura.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Ms. H. Nawagamuwa, Department of Statistics, University of Sri Jayewardenepura.
- Ms. I. M. G. U. K. Herath, Department of Statistics, University of Sri Jayewardenepura.

Venue Management Committee:

- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Prof. C. D. Tilakaratne, Department of Statistics, University of Colombo.
- Dr. R. M. Silva, Department of Statistics, University of Sri Jayewardenepura.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.
- Dr. H. A. Pathberiya, Department of Statistics, University of Sri Jayewardenepura.
- Prof. N. V. Chandrasekara, Department of Statistics & Computer Science, University of Kelaniya.
- Dr. I. U. Hewapathirana, Software Engineering Teaching Unit, University of Kelaniya.
- Mr. E. R. A. D. Bandara, Department of Statistics, University of Colombo.
- Dr. H. W. B. Kavinga, Department of Statistics & Computer Science, University of Kelaniya.
- Mr. H. A. C. S. Hapuarachchi, Department of Sports Sciences and Physical Education, Sabaragamuwa University of Sri Lanka.
- Ms. R. P. K.S. Nishani, IASSL.
- Ms. H. D. R. M. Fernando, IASSL.

ISC Case Study Competition Committee:

- Dr. T. S. Talagala, Department of Statistics, University of Sri Jayewardenepura.
- Dr. P. D. Talagala, Department of Computational Mathematics, University of Moratuwa.
- Dr. Harsha Perera, Department of Statistics and Actuarial Science, Simon Fraser University, Canada.

Finance Management, Souvenirs and Logistics Committee:

- Mr. J. De Silva, IASSL.
- Dr. P. De Silva, IASSL.
- Dr. C. L. Jayasinghe, Department of Statistics, University of Sri Jayewardenepura.
- Dr. W. N. N. K. Perera, Department of Statistics, University of Sri Jayewardenepura.
- Prof. C. D. Tilakaratne, Department of Statistics, University of Colombo.
- Dr. N. Devpura, Department of Statistics, University of Sri Jayewardenepura.
- Ms. R. P. K.S. Nishani, IASSL.

Contents

KEYNOTE SPEECHES

GUEST SPEECHES

INVITED TALKS

INVITED GROUP SESSIONS

Biostatistics and Bioinformatics

Time Series Analysis and Forecasting

Bayesian and Machine Learning Methods with Applications

Official Statistics and Data Ethics

Blockchain and Omics Data Analytics

Statistics in Sports

Data Science and Machine Learning Insights for Data Classification

INDUSTRY INVITED TALKS

CONTRIBUTED ABSTRACTS – ORAL PRESENTATIONS

Statistical Modeling

Machine Learning and AI

Biostatistics and Bioinformatics

Applications in Education

Biostatistics and Bioinformatics, Sample Surveys

<u>Applications in Agriculture, Manufacturing, Demography</u>

<u>Financial and Marketing Data Analytics, Econometric Modeling</u>

Sports Analytics

Data Mining and Big Data Analytics

Sports Analytics, Time Series Analysis and Forecasting

Bayesian Statistics and Applications, Statistical Modeling

Statistical Modeling, Data Mining and Big Data Analytics

CONTRIBUTED ABSTRACTS – POSTER PRESENTATIONS

KEYNOTE SPEECHES

Improving Forecasts via Subspace Projections

Hyndman R. J.*

Monash University, Victoria, Australia

Rob.Hyndman@monash.edu

0000-0002-2140-5352

Univariate, multivariate and hierarchical forecasts can all be improved using projections onto linear subspaces, regardless of what forecasting method is used. I will show some theoretical guarantees of this statement and demonstrate using empirical applications how linear projections can lead to (sometimes dramatic) improvements in forecast accuracy. The procedure involves creating new time series that are linear combinations of the observed time series. These are then forecast, along with the original series and all forecasts are adjusted via a projection matrix to ensure they satisfy the linear constraints. This procedure is now widely used for hierarchical forecasting. It has the potential to revolutionize multivariate forecasting as well.

Keywords: Forecast reconciliation, Hierarchical data, Multivariate forecasts, Time series

Problems that I have Enjoyed in Sports Analytics

Swartz T. B.*

Department of Statistics, Simon Fraser University, Canada

timothy_swartz@sfu.ca

0000-0001-6092-6727

This nontechnical presentation concerns three problems in sports analytics. The first involves an analysis of partnerships in cricket where it is suggested that the ``special'' relationship that exists between batting partners is not as great as many people believe. The second problem concerns an examination of the tactic of “parking the bus” in soccer where it is demonstrated that it is a counterproductive strategy to play extremely defensively when protecting a lead. The third problem concerns the analysis of aging curves in soccer. It is argued that aging curves have various deficiencies and that curves of maximum acceleration versus age may provide a proxy for aging curves. This may be useful in the context of player retention and acquisition.

Keywords: Kullback Leibler, Soccer styles, Sports analytics

GUEST SPEECHES

From Statistics to Data Science: Past, Present and Future

Emerson J. W.*

Department of Statistics & Data Science, Yale University, United States

john.emerson@yale.edu

0009-0004-3226-0023

We gather data. We estimate. We predict. We seek to understand and communicate. And we succeed and fail – outcomes observed with the benefit of hindsight. In 2015 I gave a Keynote Address (right here in Columbo). I didn't place sufficient emphasis on the importance of collaboration. I showed far too much code on a very technical topic. And I didn't predict the explosion of interest and activity in Data Science that we've witness in the last decade. My 2021 Keynote Address for ICDS 2021 (sadly, remotely via Zoom rather than in-person) attempted to acknowledge and correct these failures but missed out on predicting the AI revolution. Today, I'll try to correct that oversight as I briefly touch upon some of the topics covered previously using a new AI data analysis example along with an update on the Environmental Performance Index (with a special comment on Sri Lanka). Finally, I will describe my truncated efforts to study cricket data (surely many of you can do much better than I could) before concluding with a new analysis of a real-world problem from the realm of horse racing.

Keywords: Data analysis, Data science, Statistics

Everyday Statistician's Impact on Clinical Trials and Team Science: Unpacking My Own Questions

Lee J. H.*

University of Florida, Gainesville, FL, United States

jihyun.lee@ufl.edu

0000-0001-6420-5150

In a world that often prioritizes exceptionalism, the impactful role of everyday individuals can be easily overlooked. Terms like 'exceptional' and 'outstanding,' commonly highlighted in NIH grant proposal critiques, tend to emphasize high-impact contributions, potentially overshadowing the significance of the majority. Inspired by a poignant scene in the movie Barbie, where an ordinary person's voice resonates, I challenge the notion that only extraordinary individuals can effect change, underscoring the profound impact of many everyday contributors. I embrace my identity as an everyday biostatistician, collaborator and team scientist and will share personal anecdotes and professional insights to demonstrate how statistical thinking and leadership can significantly advance science and improve patient care through my current research and statistical practices. Furthermore, as the 2025 President of the American Statistical Association (ASA), I will outline my strategic initiatives aimed at building bridges within the ASA community and beyond. I'll discuss how my leadership journey supports the ASA’s strategic plans and mission to promote the practice and profession of statistics, emphasizing the vital role of 'everyday' professionals. This talk is based on my recent interview with *Significance* magazine, where I explored my personal and professional journey, along with my role in the ASA.

Keywords: ASA, Everyday impact, Professional journey, Statistical leadership

INVITED TALKS

Challenges and Strategies in Learning from Large Administrative Health Data

Hu X. J.*

Department of Statistics and Actuarial Science, Simon Fraser University, Canada

joanh@stat.sfu.ca

0000-0002-1970-2980

Administrative health data contain very rich information for investigating public health issues; however, many restrictions and regulations apply to their use. Moreover, the data are usually not in a conventional format since administrative databases are created and maintained to serve non-research purposes and only the information of people who seek health services is accessible. This presentation showcases challenges and strategies in statistical analysis of administrative health data using three public health research programs: (i) Cancer Survivorship Research Program, BC Cancer Agency, Canada; (ii) Pediatric Mental Health Care Program, University of Alberta; (iii) Clinical Management of Opioid Use Disorder, BC Centre for Excellence in HIV/AIDS, Canada.

Keywords: Doubly censoring, Event history data, Zero-truncation

Understanding of Grey Information and Grey System Modelling

Xie N.*

Institute of Grey System Studies, Nanjing University of Aeronautics and Astronautics, China

xienaiming@nuaa.edu.cn

0000-0002-7368-1746

The core of system analysis is to construct mathematical models or other models so as to better measure system status in different cases. While due to the presence of system disturbances and limitations of people's cognitive abilities, information that could be collected often carries some kinds of uncertainty. As a new uncertainty theory, grey system theory (GST) is founded by Chinese scholar Professor Deng Julong and GST is used to study systems with characteristics of " information partially known and partially unknown" or systems with "limited data" and "grey information". However, what is the framework of GST? What about the mechanism of grey system modelling? And how can we use grey information in the grey system modelling process? All these questions confused those who focus on study of grey system theory and applications. This paper tries to explain all these questions. This paper summarized two types of grey system modelling, i.e. grey models with limited data and grey models with grey numbers. For the grey model with limited data, the mechanism of grey forecasting was analyzed. It employs integral matching to elucidate the mechanism of generation accumulation within the modelling process. It utilizes increments and growth rates to explicate the overarching modelling process and mechanisms of grey forecasting and other grey models. Also, it comprehensively classifies grey forecasting models based on their differential equation structures. The modelling mechanism and relationships among diverse kinds of models are worth further exploration to better explain the advantages and effectiveness of grey forecasting modelling. For grey models with grey numbers, it conducts an in-depth analysis of the generation, computation and application of grey numbers within the grey system modelling process. It mainly discusses the definition and operations of grey numbers and tries to answer the following questions: (1) How to generate grey numbers? (2) How to calculate grey numbers? (3) How to apply grey numbers? Finally, a case of grey scheduling was adopted as an example to show how to collect, calculate and apply grey numbers in real applications. Results show that this paper explains why the accumulating operation is effective in the modelling process and why the grey forecasting model is suitable for limited data. It is effective to adopt grey numbers to measure flexible time quotas and make scheduling in production and service management.

Keywords: Grey forecasting, Grey number, Grey scheduling, Grey system modelling mechanism, Grey system theory

Automated Extraction of Acronym-expansion Pairs in Scientific Literature

Ali I.[1], Haileyesus M.[2], Hnatyshyn S.[3], Ott J. L.[4] and Hnatyshin V.[5*]

[1,2,5]Computer Science Department, Rowan University, Glassboro, United States, [3,4] Department of Bioanalytical Sciences, Bristol-Myers Squibb, Princeton, United States

[1]aliizh94@students.rowan.edu, [2]hailey74@students.rowan.edu, [3]serhiy.hnatyshyn@bms.com, [4]jan-lucas.ott@bms.com, [5]hnatyshin@rowan.edu

[3]0000-0003-0992-1634, [4]0000-0002-6345-1090, [5]0000-0001-9475-0467

This work is a part of a larger project which focuses on automatic analysis and summarization of technical literature. Specifically, this project addresses challenges posed by the widespread use of abbreviations and acronyms in digital texts. We propose a novel method that combines document preprocessing, customized regular expressions and a large language model, specifically GPT-4, to identify abbreviations and map them to their corresponding expansions. Our approach is particularly useful when regular expressions alone are insufficient to extract expansions, at which point our algorithm leverages GPT-4 to analyze the text surrounding the acronyms. By limiting the analysis to only a small portion of the surrounding text, we mitigate the risk of obtaining incorrect or multiple expansions for an acronym. The literature review highlights the challenges of processing text with lots of acronyms, including such problems as polysemous acronyms (multiple meanings), non-local acronyms (lacking explicit expansions nearby) and ambiguous acronyms (whose full forms do not correspond to the acronym letters). By addressing these issues with automated acronym identification and disambiguation, our approach enhances the precision and efficiency of NLP techniques. This study highlights the challenges of working with PDF files and the importance of document preprocessing which significantly improves the accuracy of our approaches. Furthermore, the results of this work showed that neither regular expressions nor GPT-4 alone can perform well. Regular expressions are good at identifying acronyms but have hard time finding their expansions within the paper due to the variety of formats for acronym definitions and often acronyms not being defined within the text. GPT-4, on the other hand, is very good at finding acronym expansions but it struggles with correctly identifying relevant acronyms. Furthermore, GPT-4 is challenging to work with and due to its probabilistic nature, often provides slightly different results for the same input. That is why we conclude that the use of preprocessing to eliminate irrelevant information from the text (i.e., author name, formulas, references, etc.), regular expressions for identifying acronyms and a large language model to help find acronym expansions provides the most accurate and consistent results. Overall, this work facilitates the creation of automated tools for extracting and expanding acronyms, thereby enhancing the readability and comprehension of scientific and technical documents.

Keywords: Acronym identification, Acronym expansion, Document preprocessing, GPT-4 NLP, Large language models, Regular expressions

Forecasting Trade Durations using Logarithmic Component ACD Model with Extended Generalized Inverse Gaussian Distribution: A Comparative Study with an Application

Tan Y. F.[1], Ng K. H.[2*], Koh Y. B.[3] and Shelton P.[4]

[1,2,3]Institute of Mathematical Sciences, Faculty of Science, University Malaya, Kuala Lumpur 50603, Malaysia, [2]Centre of Research for Statistical Modelling and Methodology, Faculty of Science, University Malaya, Malaysia, [4]School of Mathematics and Statistics, Faculty of Science, The University of Sydney, Sydney, Australia

[1]tanyiingfei@um.edu.my, [2]kokhaur@um.edu.my, [3]kohyoubeng@um.edu.my, [4]shelton.peiris@sydney.edu.au

[2]0000-0002-8763-7586, [3]0000-0002-8468-0757, [4]0000-0002-2612-0831

Autoregressive Conditional Duration (ACD) models are widely used in applied economics modelling of duration data. This talk discusses a logarithmic version of the two-component ACD (LogCACD) model with no restrictions on the sign of the model parameters while allowing the expected durations to be decomposed into the long- and short-run components to capture the dynamics of these durations. The extended generalized inverse Gaussian (EGIG) distribution is used for the error distribution as its hazard function consists of a roller-coaster shape for certain parameters' values. Empirical application is based on the trade durations of the International Business Machines stock index. Extensive comparisons are carried out to evaluate the modelling and forecasting performances of the proposed model with several benchmark models and different specifications of error distributions. The result reveals that the LogCACD$_{\text{EGIG}}$(1,1) model gives the best in-sample fit based on the Akaike information criterion and other criteria. Furthermore, the estimated parameters obtained confirm the existence of the roller-coaster shaped hazard function. The examination of LogCACD$_{\text{EGIG}}$(1,1) model also provides the best out-of-sample forecasts evaluated based on the mean square forecast error using the Hansen's model confidence set. Lastly, different levels of time-at-risk forecasts are provided and tested with Kupiec likelihood ratio test.

Keywords: Autoregressive conditional duration, Extended generalized inverse gaussian, Hazard function, Time-at-risk

Unveiling Environmental Risks: A Statistical and Geospatial Perspective

Bomiriya R. P.[*]

R S Metrics Asia Holdings (Pvt) Ltd., Sri Lanka

rbomiriya@rsmetrics.com

Managing environmental risks effectively is essential for organizations striving to build resilience, comply with regulations and align with sustainability frameworks like CSRD, TCFD and TNFD. By integrating satellite data from sources like Sentinel Collections, MODIS and ERA alongside Geospatial Science, Physics and Statistical techniques, we at RS Metrics deliver precise asset-level insights into a variety of environmental risks. Our work spans physical climate risk assessments (e.g., heatwaves, cold waves, coastal and riverine inundation), methane emission detection and quantification, deforestation monitoring, biodiversity loss assessment and more. While these methods are technically complex, we provide corporations with easy-to-use tools to understand and mitigate risks, align with sustainability goals and help asset managers construct environmentally aligned portfolios, monitor them and engage more effectively via similar tools. For methane detection and quantification, we use Statistical anomaly detection to spot plumes in hyperspectral imagery and Bayesian methods in atmospheric transport modeling help us trace those emissions back to their sources. For wildfire risk assessment, spatial clustering techniques coupled with Local Moran's I enable the identification of hotspots and the assignment of scalable risk scores based on fire intensity and neighboring influences. This presentation will explore how Statistics intersects with Geospatial Science and environmental disciplines through real-world case studies, including methane emission detection and wildfire risk assessment. A deep dive into the wildfire case study will highlight the Statistical techniques driving these insights, while also demonstrating how they tackle pressing environmental challenges and close critical data gaps.

Keywords: Geospatial analytics, Methane emissions, Regulatory requirements, Spatial statistics, Wildfire risk

Recent Contributions to Long Memory Time Series Analysis and Applications: An Overview

Peiris S.[1*], Allen D.[2], Hunt R.[3] and Gadhi A.[4]

[1,2,3,4]School of Mathematics and Statistics, The University of Sydney, Australia

[1]Shelton.peiris@Sydney.edu.au, [2]profallen2007@gmail.com, [3]richard@huntemail.id.au, [4]agad0777@uni.sydney.edu.au

[1]0000-0002-2612-0831, [2]0000-0001-7782-0865, [3]0000-0002-5325-101X, [4]0000-0002-3324-909X

Analysis of long memory time series became very popular among the theoretical and applied researchers in the last 2-3 decades due to its flexibility in many applications in almost every field. In this paper, a particular attention has been paid to the development of Generalized Long Memory time series generated by Gegenbauer polynomials and Autoregressive Moving Average (ARMA) models. Several estimation methods will be discussed with applications in various fields will be presented. A multivariate or vector extension to GARMA family (ie. Vector GARMA or VEGARMA) will be introduced along with the relevant theoretical properties and applications.

Keywords: Forecasting, Generalized long memory, Hybrid, Long memory, Time series, Vector models

Compromise Designs Under Baseline Parameterization

Tang B.*

Department of Statistics and Actuarial Sciences, Simon Fraser University, Canada

boxint@sfu.ca

0000-0002-3586-6694.

We consider estimation of main effects using two-level fractional factorial designs under the baseline parameterization. Previous work in the area indicates that orthogonal arrays are more efficient than one-factor-at-a-time designs whereas the latter are better than the former in terms of minimizing the bias due to non-negligible interactions. Using efficiency criteria, this paper examines a class of compromise designs obtained by adding runs to one-factor-at-a-time designs. A theoretical result is established for the case of adding one run. For adding two or more runs, we develop a complete search algorithm to find optimal compromise designs.

Keywords: Efficiency criterion, Minimum aberration, One-factor-at-a-time design, Orthogonal array

Scalable and Non-iterative Graphical Model Estimation

Rajaratnam B.[1*], Khare K.[2], Rahman S.[3] and Zhou J.[4]

[1]Department of Statistics, University of California at Davis, Davis, CA, United States, [2]Dept. of Statistics, University of Florida, USA, [3]Apple Inc., USA, [4]Freddie Mac, United States

[1]brajaratnam01@gmail.com

Graphical models have found widespread applications in many areas of modern statistics and machine learning. Iterative Proportional Fitting (IPF) and its variants have become the default method for undirected graphical model estimation and are thus ubiquitous in the field. As the IPF is an iterative approach, it is not always readily scalable to modern high-dimensional data regimes. In this paper, we propose a novel and fast non-iterative method for positive definite graphical model estimation in high dimensions, one that directly addresses the shortcomings of IPF and its variants. In addition, the proposed method has a number of other attractive properties. First, we show formally that as the dimension p grows, the proportion of graphs for which the proposed method will outperform the state-of-the-art in terms of computational complexity and performance tends to 1, affirming its efficacy in modern settings. Second, the proposed approach can be readily combined with scalable non-iterative thresholding-based methods for high-dimensional sparsity selection. Third, the proposed method has high-dimensional statistical guarantees. Moreover, our numerical experiments also show that the proposed method achieves scalability without compromising on statistical precision. Fourth, unlike the IPF, which depends on the Gaussian likelihood, the proposed method is much more robust.

Keywords: Cholesky decomposition, Graphical models, Non-iterative estimation, Ultra-high dimensions

Ups and (Draw) Downs

Tommaso P.*

University of Rome “Tor Vergata” and CREATES, Aarhus, Italy

tommaso.proietti@uniroma2.it

0000-0001-5285-7522

The concept of drawdown quantifies the potential loss in the value of a financial asset when it deviates from its historical peak. It plays an important role in evaluating market risk, portfolio construction, assessing risk-adjusted performance and trading strategies. This paper introduces a novel measurement framework that produces, along with the drawdown and its dual (the drawup), two Markov chain processes representing the current lead time with respect to the running maximum and minimum, i.e., the number of time units elapsed from the most recent peak and trough. Under relatively unrestrictive assumptions regarding the returns process, the chains are homogeneous and ergodic. We show that, together with the distribution of asset returns, they determine the properties of the drawdown and drawup time series, in terms of size, serial correlation, persistence and duration. Furthermore, they form the foundation of a new algorithm for dating peaks and troughs of the price process delimiting bear and bull market phases. The other contributions of this paper deal with out-of-sample prediction and robust estimation of the drawdown.

Keywords: Dating bear and bull markets, Financial time series, Risk measures

Statistical Methods for Predicting Outcomes Based on Images and Other Covariate Information

Manathunga A.*

Emory University, Atlanta, Georgia, United States

amanatu@emory.edu

0000-0002-2289-563X

Our work is motivated by the need to develop non-invasive tools for monitoring anemia in very low birth weight (VLBW; birth weight < 1,500 grams) and reduce the number of routine painful, invasive blood sampling procedures (phlebotomy) that may alter infant neurodevelopment and behavior. Recently, a new smartphone application that collects and analyzes clinical pallor in patient-sourced fingernail photos and image metadata has been developed to predict hemoglobin levels among adults. We develop a new image analysis that analyzes full structural information of clinical pallor in fingernail photos to enable accurate, non-invasive prediction of blood hemoglobin level and anemia risk among VLBW infants. Our method is based on a novel functional principal component analysis method that provides a non-parametric and parsimonious means to jointly model high-dimensional photos and image metadata, while fully utilizing the other potential covariates. Furthermore, we discuss extensions of our methods that leverage longitudinal, patient-level clinical data and predictions to achieve the overarching clinical goal of minimizing the number of blood draws in VLBW infants throughout the care continuum. Our work uses the data of VLBW infants monitored at three neonatal intensive care units in Atlanta. The proposed methods are generally applicable to a wide variety of settings with diverse and complex modalities of data.

Keywords: Functional data, Prediction, Principal components

Can We Protect Time Series Data While Maintaining Accurate Forecasts?

Bale C. D.[1], Schneider M. J.[2*] and Lee J.[3]

[1]Marriott School of Business, Brigham Young University, United States, [2,3]Lebow College of Business, Drexel University, United States

[1]cameron.bale@byu.edu, [2]mjs624@drexel.edu, [3]jl3539@drexel.edu

[2]0000-0001-5667-4707

We evaluate the usefulness of protected time series by exploring how privacy protection affects forecast accuracy. Using both simulated and real-world time series data sets, we test various privacy methods, including a proposed swapping-based method (k-nTS+) designed to maintain time series features, a differentially private method and an approach based on sharing model weights trained on unprotected data. Based on forecasts from both simple and machine learning models, we find that none of the privacy methods can consistently maintain forecast accuracy at an acceptable level of privacy. We also show that sharing model weights enables accurate forecasts, but accurate forecasts can be used to uncover the identities of protected time series. To overcome these problems, we transform continuous time series into bounded rates to increase the similarities of features, values and forecasts across time series. This enables our proposed method to produce protected data with a reduction in average forecast accuracy of just 6%. Overall, we acknowledge that except under certain conditions, generating time series with acceptable privacy levels is incompatible with the goal of obtaining accurate forecasts.

Keywords: Data privacy, Differential privacy, Forecast accuracy, Machine learning, Time series features

Analytical Form of Influence Functions for Statistics

Maheswaran R.*

Wagga Wagga Agricultural Institute, DPIRD Wagga Wagga, NSW 2650, Australia

maheswaran.rohan@dpi.nsw.gov.au

0000-0001-7370-2431

The statistics for model parameters are often computed in either closed form or non-closed form. For example, this is a common practice when maximum likelihood (ML) or robust methods are used. The latter requires an iterative procedure to optimize the estimates of the parameters. One of the diagnostics tools for assessing statistics is the influence function, which measures the impact of small changes in the distribution on the value of an estimator. The computation of the influence function for closed-form estimates is relatively easy in comparison to that for non-closed-form estimates. However, obtaining the analytical form of the empirical influence functions of iteratively defined statistics for multiple parameters is absent in the current literature and not easy. In this talk, we use matrix algebra including matrix derivation to show how the influence functions can be derived analytically for M-estimators with multiple parameters in linear models.

Keywords: M-estimators, Iterative algorithm, One-step influence function, Jacobian matrix

Predicting Multilingual Student Performance with Llama 3

Ayesha B.[1] and Thayasivam U.[2*]

[1,2]Department of Mathematics, Rowan University, NJ, USA

[1]rathna55@rowan.edu, [2]thayasivam@rowan.edu

0000-0002-2093-8093

Predicting student performance is essential for enabling timely interventions and enhancing educational outcomes. Traditional models often rely on limited features and face challenges in processing multilingual and informal textual data. This study introduces a novel methodology that leverages Llama 3, an advanced large language model (LLM), to predict student grades in a multilingual educational context. Dataset comprising over 700 student records from diverse regions, featuring both quantitative data—such as historical examination scores across 12 subjects—and qualitative data, including personal ambitions, parental education levels, extracurricular activities and learning styles. It also includes textual responses in multiple languages with colloquial expressions. To preprocess this complex data, we employed the Llama 3.1 8B model for translating multilingual and informal textual entries into coherent English and for categorizing qualitative variables, thereby enhancing data consistency and quality. The Llama 3.1 model was fine-tuned using Transformer Reinforcement Learning (TRL) and Supervised Fine-Tuning (SFT) techniques, incorporating comprehensive student data to predict grade classes (A, B, C, S, F) for the subsequent term. Our approach explored both regression and classification methodologies, contrasting the performance of traditional models like XGBoost and a hybrid model with our fine-tuned Llama 3 model. Experimental results demonstrated that the fine-tuned Llama 3 model significantly outperformed traditional models across multiple subjects, achieving superior results across various evaluation metrics. Notably, it excelled in predicting extreme grades, effectively identifying both high-achieving and at-risk students with minimal misclassifications. The model's superior performance is attributed to its ability to handle complex feature interactions, process multilingual and colloquial data and mitigate the effects of imbalanced grade distributions. This study underscores the potential of advanced LLMs in predicting student performance, particularly in multilingual settings with diverse data types. By addressing the limitations of previous studies—such as handling informal textual data and relying on single predictive methodologies—our approach provides a robust framework for educational prediction tasks. Future research directions include validating the model in different educational contexts, enhancing the interpretability of predictions and exploring the prediction of continuous scores for more granular assessments.

Keywords: Large language models (LLMs), Llama 3, Multilingual data processing, Student performance prediction, Transformer reinforcement learning

Dynamic Data Science Applications in Algorithmic Trading

Thavaneswaran A.[1*] and Thulasiram R. K.[2]

[1]Department of Statistics, University of Manitoba, Canada, [2]Department of Computer Science, University of Manitoba, Canada

[1]aerambamoorthy.thavaneswaran@umanitoba.ca, [2]tulsi.thulasiram@umanitoba.ca

[1]0009-0007-7611-5081, [2]0000-0002-6519-3929

Recently, there has been a growing interest in using filtered estimates with data-driven innovation volatility for dynamic hedge ratios for pairs trading. Kalman filtering algorithms were successfully applied in pairs trading with only two co-integrated assets. Recently proposed non-Gaussian maximum informative filtering algorithms for dynamic state space models are used to obtain the filtered estimates of hedge ratios and applied in multiple trading. In this paper, ARIMA model based innovation correlation network and neuro correlation based dynamic networks are used to select the stocks for trading. Algorithmic trading profits made by the proposed neuro correlation based financial networks are compared with the trading profits made by the existing Pearson correlation based networks. Unlike the existing work, the novelty of the work is that it uses a data-driven neuro correlation based financial network to select the pairs for trading. Profit per transaction is used to compare the trading strategies.

Keywords: Financial networks, Innovations, Neuro correlation, Non-gaussian filtering algorithms, Pairs trading

Impact of Environmental Factors on two Prominent Diseases Prevalent in New South Wales, Australia: A Statistical Assessment

Dissanayake G. S.[1*] and Duck G.[2]

[1]NSW Ministry of Health, School of Mathematics and Statistics, University of Sydney, NSLHD,
[2]NSW Ministry of Health

[1]gnanadarshad@gmail.com, [2]Gerard.Duck@health.nsw.gov.au

[1]0000-0003-2765-5886

Changing disease prevalence over time is an important factor in projecting for future health system requirements. Case studies and analysis of the impact of environmental and climate change on diseases affecting humans have been done in various parts of the world, providing a range of data points and interpretations. A unique subset of the published literature are statistical causality studies that infer the impact of certain environmental factors on incidence frequencies of some diseases that afflict humans. Occurrence of such frequencies in a periodic manner generate time series that could be utilised to develop future forecast predictions known as projections. This study uses the Granger's statistical causality method to assess the relationship between environmental variables and the prevalence of two diseases (Asthma and Melanoma) and presents projections using appropriate time series methodology to inform the health service planning requirements of the state of New South Wales (NSW). Causality between Air quality and Asthma incidence as well as Solar irradiance and Melanoma incidence are established through statistical significance measures as a contribution. Thereafter time series based forecast projections are provided for both the diseases as another contribution. Statistical causality between environmental and health disease related linked variables have been established for Asthma and Melanoma and subsequently disease incidence frequency projections in the form of point and interval estimates for the NSW state in Australia have been provided.

Keywords: Asthma, ARIMA, Causality, Lags, Melanoma.

Bayesian State Space Models with Ecological Applications

Muthukumarana S.*

Department of Statistics, University of Manitoba, Canada

saman.muthukumarana@umanitoba.ca

0000-0001-8942-5352

Bayesian state-space models offer powerful tools for modelling individual-level animal movements in ecological studies. These models are particularly valuable because they can simultaneously account for process variation—the natural variability in the animal's movement patterns—and observational error, which represents the discrepancy between the observed and true positions of the animal. In this talk, I will explore various Bayesian state-space modelling techniques, focusing on different smoothing methods such as kernel smoothing and cross-validated local polynomial regression. Additionally, I will introduce a novel approach for reconstructing animal movement paths, providing a more accurate depiction of their trajectories. These methodologies will be illustrated with data collected from a telemetry receiver grid in Lake Winnipeg, demonstrating their practical application and effectiveness in ecological research.

Keywords: Bayesian, Ecology, MCMC

INVITED GROUP SESSIONS

Demystifying Estimands in Non-inferiority Trials in Anaesthesia: Insights from the CHEWY Trial

De Silva A. P.[1*], Darvall J. N.[2], Leslie K.[3] and Braat S.[4]

[1,4]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia, [1,4]Methods and Implementation Support for Clinical and Health (MISCH) Research Hub, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia, [2,3]Department of Critical Care, Melbourne Medical School, The University of Melbourne, Melbourne, Australia, [2,3]Department of Anaesthesia and Pain Management, The Royal Melbourne Hospital, Melbourne, Australia

[1]anurika.de@unimelb.edu.au, [2]jai.darvall@mh.org.au, [3]kate.leslie@mh.org.au, [4]s.braat@unimelb.edu.au

[1]0000-0003-0541-3202, [2]0000-0003-0579-8931, [3]0000-0001-8512-3517, [4]0000-0003-1997-3999

Non-inferiority trials in anaesthesia aim to demonstrate that a new anaesthetic agent or technique is not worse than an active control. These trials play a crucial role in anaesthesia research by allowing for the potential implementation of safer, more cost effective alternatives while ensuring that patient outcomes are not clinically worse than current standard of care. However, these trials pose several statistical challenges compared to superiority trials, such as reconciling competing non-inferiority and superiority objectives and defining a suitable non-inferiority margin. These challenges are further exacerbated by events occurring after randomisation, such as the use of rescue medication, disease recurrence, treatment discontinuation, or death, some of which are common in anaesthesia trials. These post-randomisation events, also known as intercurrent events, can affect the collection of the outcome of interest or its interpretation. However, it is often unclear how these events are handled, leading to poorly defined treatment effects. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9(R1) addendum introduced the estimand framework. The aim of this framework is to precisely define the treatment effect (i.e., the estimand) and align trial objectives with the study design, conduct, data collection, statistical analysis and interpretation of results. However, the ICH E9(R1) addendum primarily focuses on superiority trials and therefore provides limited guidance on how to construct estimands for non-inferiority trials. The chewing gum to treat postoperative nausea and emesis in female patients (CHEWY) trial was an international, multicentre, randomised controlled non-inferiority trial. It assessed the efficacy and safety of chewing gum, compared to the current standard, for treating postoperative nausea, retching and vomiting (PONV) in the postanaesthesia care unit (PACU). Female patients undergoing volatile anaesthetic-based general anaesthesia for laparoscopic or breast surgery, experiencing PONV in the PACU, were randomised to receive 15 min of chewing gum or 4 mg intravenous ondansetron. The primary outcome was defined as a composite of cessation of PONV, with no recurrence and no rescue medication for 2 h after treatment administration. The intercurrent events of use of rescue medication and recurrence of PONV were incorporated within the primary outcome definition. Sharing insights from the CHEWY trial, we examine how to clearly define research questions and treatment effects for non-inferiority trials in the presence of intercurrent events. We discuss how statistical considerations such as switching between non-inferiority and superiority objectives, the choice of non-inferiority margin and per protocol analyses could be aligned with the estimands. Finally, we establish the importance of clearly defining estimands when planning non- inferiority trials in anaesthesia, to improve the interpretability and clinical relevance of trials and ensure that trials answer their clinical questions of interest.

Keywords: Anaesthesia, Estimands, Noninferiority, Postoperative nausea and vomiting, Randomized controlled trial

Roadmap for Systematic Identification and Analysis of Multiple Biases in Causal Inference

Wijesuriya R.^{1*}, Carlin J. B.^{2}, Hughes R. A.^{3}, Peters R. L.^{4}, Koplin J. J.^{5} and Moreno-Betancur M.^{6}

[1,2,4,6] Murdoch Children's Research Institute and University of Melbourne, Australia [3]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; [3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK University of Bristol, UK; [5]Univesrity of Queensland and University of Melbourne, Australia

[1]rushani.wijesuriya@mcri.edu.au, [2]john.carlin@mcri.edu.au, [3,4]Rachael.Hughes@bristol.ac.uk, [5]rachel.peters@mcri.edu.au, [6]jennifer.koplin@mcri.edu.au, [7]margarita.moreno@mcri.edu.au

[1]0000-0003-1023-4065, [2]0000-0002-2694-9463, [3,4]0000-0003-0766-1410, [5]0000-0002-2411-6628 , [6]0000-0002-7576-5142, [7]0000-0002-8818-3125

Observational studies examining causal effects rely on unverifiable causal assumptions, the violation of which can induce multiple biases due to confounding, measurement and selection processes. Quantitative bias analysis (QBA) methods aim to examine the sensitivity of findings to such violations, generally by producing bias-adjusted estimates under alternative assumptions. Commonly, applications of QBA examine a single source of bias or multiple sources are considered but separately, thus not informing the overall impact of the potential biases. In this work we propose a roadmap for systematically identifying and analyzing multiple biases together. Briefly, (i) Articulate all assumptions made in the primary analysis using the target trial framework. This is achieved by specifying the ideal trial that defines the causal effect of interest and considering assumptions made to emulate it; (ii) Use causal diagrams to identify potential sources of bias by depicting plausible violations of the assumptions in the primary analysis; (iii) Obtain a single estimate adjusted for all potential sources of bias identified using a recent simultaneous bias adjustment approach. We illustrate the proposed roadmap in an investigation of the effect of breastfeeding on risk of childhood asthma. We further evaluate the simultaneous approach across a range of scenarios via simulations and illustrate the extent of remaining bias when performing single bias adjustments. Simulation results highlight the need for simultaneous adjustment to examine the overall impact of biases and the proposed roadmap facilitates the conduct of high-quality multiple bias analyses.

Keywords: Causal inference, Quantitative bias analysis, Target trial framework

Microbial Network Inference for Longitudinal Microbiome Studies with LUPINE

Kodikara, S.[1*] and Cao K. A. L.[2]

[1,2]School of Mathematics and Statistics, The University of Melbourne, Royal Parade, 3052, Victoria, Australia

[1]saritha.kodikara@unimelb.edu.au, [2]kimanh.lecao@unimelb.edu.au

[1]0000-0002-7039-8398, [2]0000-0003-3923-1116

The microbiome is a complex ecosystem of interdependent taxa that has traditionally been studied through cross-sectional studies. However, longitudinal microbiome studies are becoming increasingly popular. These studies enable researchers to infer taxa associations towards the understanding of coexistence, competition and collaboration between microbes across time. Traditional metrics for association analysis, such as correlation, are limited due to the data characteristics of microbiome data (sparse, compositional, multivariate). Several network inference methods have been proposed but have been largely unexplored in a longitudinal setting. We introduce LUPINE (LongitUdinal modelling with Partial least squares regression for NEtwork inference), a novel approach that leverages on conditional independence and low-dimensional data representation. This method is specifically designed to handle scenarios with small sample sizes and small number of time points. LUPINE is the first method of its kind to infer microbial networks across time, while considering information from all past time points. We validate LUPINE and its variant, LUPINE_single (for single time point analysis) in simulated data and three case studies, where we highlight LUPINE's ability to identify relevant taxa in each study context, across different experimental designs (mouse and human studies, with or without interventions, as short or long time courses). To detect changes in the networks across time, groups or in response to external disturbances, we used different metrics to compare the inferred networks. LUPINE is a simple yet innovative network inference methodology that is suitable for, but not limited to, analysing longitudinal microbiome data.

Keywords: 16S, Longitudinal, Network, Partial correlation

Forecasting Stock Volatility using a Robust GARCH-MIDAS Model

Choo W. C.^{1*}, Liu T.^{2} and Schneider M. J.^{3}

^{1,2}School of Business and Economics, Universiti Putra Malaysia, Seri Kembangan, Malaysia, ^{3}LeBow College of Business, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104

^{1}wcchoo@upm.edu.my, ^{2}dr.liuting01@gmail.com, ^{3}mattschneids@gmail.com

^{1}0009-0002-1106-5718, ^{3}0000-0001-5667-4707

This study proposes a robust generalized autoregressive conditional heteroskedasticity- Mixed Data Sampling (RGARCH-MIDAS) model to address the problem of volatility forecasting in the stock market, which is susceptible to outliers. Based on the traditional GARCH-MIDAS model, this model significantly improves forecasting performance in the presence of outliers by introducing an outlier correction mechanism. In the comprehensive evaluation, this study uses various loss functions and information criteria as assessment indicators and the findings indicate that the RGARCH-MIDAS model outperforms other models in the in-sample estimation. Particularly regarding forecasting ability, the RGARCH-MIDAS model achieves superior performance levels in several stock markets and significantly outperforms the competing models according to Theil's U statistic and MCS tests. Simulation experiments further validate the forecasting accuracy of the RGARCH-MIDAS model under outlier conditions. In conclusion, the RGARCH-MIDAS model proposed in this study provides a new and effective tool for stock market volatility prediction, especially in dealing with outlier.

Keywords: Forecasting, GARCH-MIDAS, Outliers, Stock market, Volatility

Modelling Water Quality in Small Scale Aquaculture Ponds using Machine Learning Techniques

Wickramaratne C.[1*], De Silva S.[2], Phayko U.[3], Palal M.[4], Leh M[5]. Mathesawaran K.[6] and McCartney M.[7]

[1,2,5]International Water Management Institute, 127, Sunil Mawatha, Pelawatte, Sri Lanka,
[3,4,6,7]International Water Management Institute Myanmar, c/o Irrigation Head Office Yangon Compound (Ministry of Agriculture, Livestock and Irrigation), Thit Sar Road, Yankin Township, Yangon, Myanmar

[1]c.wickramaratne@cgiar.org, [2]s.s.desilva@cgiar.org, [3]u.phayko@cgiar.org, [4]agripontpont@gmail.com, [5]m.leh@cgiar.org, [6]k.matheswaran@cgiar.org, [7]m.mccartney@cgiar.org

[1]0009-0004-7146-7860, [2]0000-0001-8852-800X, [5]0000-0001-8865-767X, [6]0000-0001-7377-0629, [7]0000-0001-6342-2815

Small-scale aquaculture provides a valuable source of protein-rich, nutrient-dense food and offers crucial livelihood opportunities for rural communities, particularly in developing countries like Myanmar. Water quality, especially Dissolved Oxygen (DO), is a key factor that influences fish growth and, consequently, the productivity and income of farmers. This study applied machine learning techniques, specifically Artificial Neural Networks (ANNs) and Random Forest (RF), to predict DO levels in fishponds based on field measurements from 15 ponds in Myanmar. Water quality influencing predictor variables consisting of pH, temperature electrical conductivity and nutrients (phosphate and nitrate) were used to simulate DO levels in the aquaculture ponds. Model performance was evaluated using correlation coefficient and root mean square error, with results indicating that both models are suitable for DO prediction. However, the performance of the RF model was superior to ANN. By examining the applicability of these techniques in a data limited environment, the study offers insights for improving pond management practices, enhancing water quality and encouraging farmers to adopt sustainable practices.

Keywords: Aquaculture, Artificial neural networks, Dissolved oxygen, Random forest

Feature-based Intermittent Demand Forecast Combinations: Accuracy and Inventory Implications

Li L.*

School of Economics and Management, University of Science & Technology Beijing, China

li.li@ustb.edu.cn

0000-0002-7922-1281

Intermittent demand forecasting is a ubiquitous and challenging problem in production systems and supply chain management. In recent years, there has been a growing focus on developing forecasting approaches for intermittent demand from academic and practical perspectives. However, limited attention has been given to forecast combination methods, which have achieved competitive performance in forecasting fast-moving time series. The current study examines the empirical outcomes of some existing forecast combination methods and proposes a generalized feature-based framework for intermittent demand forecasting. The proposed framework has been shown to improve the accuracy of point and quantile forecasts based on two real data sets. Further, some analysis of features, forecasting pools and computational efficiency is also provided. The findings indicate the intelligibility and flexibility of the proposed approach in intermittent demand forecasting and offer insights regarding inventory decisions.

Keywords: Diversity, Empirical evaluation, Forecast combinations, Intermittent demand forecasting, Time series features

A Bayesian Kalman Filter Model for Estimating the Actual Sea Surface Level using the Upward-looking Sonar Measurements

Karunanayake N.[1*] and Muthukumarana, S.[2]

[1,2]Department of Statistics, University of Manitoba, Winnipeg, MB, Canada

[1]karunann@myumanitoba.ca, [2]saman.muthukumarana@umanitoba.ca

[1]0009-0006-1724-9162, [2]0000-0001-8942-5352

The Arctic region has experienced significant and ongoing declines in sea ice extent and thickness, raising urgent concerns about global climate change and its implications for future Arctic navigation. Accurately identifying sea ice extent is critical for understanding these changes. One of the primary challenges in estimating sea ice extent using upward-looking sonar data is determining the sea surface level to calculate sea ice draft accurately. During the summer months, the sea surface level is relatively easy to identify. However, no reliable estimation methods have been established for determining the sea surface level during the winter months. This study proposes a method for estimating the true Arctic Sea surface level using Bayesian Kalman filter models. The Bayesian Kalman filter is employed to estimate the actual state of the sea surface by processing noisy acoustic range measurements. The model effectively handles both linear dynamics and Gaussian noise, refining its estimates by continuously updating posterior distributions as new data becomes available. The approach is demonstrated using long-term ice draft measurements collected via upward-looking sonar (ULS) instruments moored in the Beaufort Gyre between 2018 and 2020. These data and the results provide valuable insights into the dynamics of sea ice in this critical region, contributing to more reliable climate forecasts and improved navigation planning in ice-covered Arctic waters.

Keywords: Arctic sea ice, Bayesian filtering, Sea ice draft

Harnessing AI for Transformative Healthcare and Pharmaceutical Applications

Kaushik S.*

Data Science, Technology Solutions, Syneos Health, India

shrutikaushik.surenderpalkaushik@syneoshealth.com

0000-0001-9117-2868

Artificial intelligence (AI) and machine learning (ML) are revolutionizing predictive analytics in healthcare and pharmaceuticals. While statistical and ML methods have been widely used in areas such as forecasting healthcare expenditures, there is a significant gap in comparing these models, particularly when combined in ensemble approaches. This work examines the effectiveness of various statistical, neural network-based supervised learning and ensemble techniques in predicting patients' weekly expenditures on pain medications. We evaluated two statistical models persistence (baseline) and autoregressive integrated moving average (ARIMA) alongside a multilayer perceptron (MLP), a long short-term memory (LSTM) model and an ensemble model that integrates predictions from ARIMA, MLP and LSTM. Our results demonstrate that the ensemble model consistently outperforms individual models, underscoring the potential of hybrid approaches in healthcare time-series forecasting. Building on these findings, our current work extends AI methodologies to pragmatic and scalable opportunities for improving biopharma drug development. By leveraging ML models to predict enrollment rates and site activation times, we optimize clinical trial operations. This optimization involves identifying key factors such as high-performing sites, excluding slow enrollers and diagnosing recruitment challenges. These AI-driven insights enhance the efficiency of patient recruitment strategies, ultimately accelerating the clinical trial process. Our integrated approach highlights the transformative potential of AI in driving predictability across healthcare and biopharma, be it forecasting healthcare expenditures or identifying opportunities to accelerate clinical trials.

Keywords: Artificial intelligence, Biopharma innovation, Clinical trial optimization, Ensemble modeling, Predictive analytics

Uncovering Symptoms and Predicting Long COVID using Social Media Tweets and Clinical Notes Data: A Machine Learning Approach

Matharaarachchi S.[1*], Domaratzki M.[2], Katz, A.[3] and Muthukumarana S.[4]

[1,4]Department of Statistics, University of Manitoba, Winnipeg, MB, Canada,
[2]Department of Computer Science, Western University, London, ON, Canada,
[3]Department of Community Health Sciences and Family Medicine, University of Manitoba, Winnipeg, MB, Canada

[1]matharas@myumanitoba.ca, [2]mdomarat@uwo.ca, [3]alan.katz@umanitoba.ca,
[4]saman.muthukumarana@umanitoba.ca

[1]0000-0002-9490-4742, [2]0000-0001-9129-6676, [3]0000-0001-8280-7024, [4]0000-0001-8942-5352

The COVID-19 pandemic has posed a significant public health challenge, with long-term consequences such as Long COVID Syndrome (LCS) affecting patients' daily lives. Common symptoms like breathlessness, fatigue and brain fog persist for months after the initial infection, making it crucial to identify and analyze these symptoms to improve diagnosis and treatment strategies. This study aims to enhance our understanding of LCS by exploring patterns and behaviors of symptoms reported on Twitter and establishing a precise method for identifying LCS patients using actual clinical notes data using machine-learning classification models. Building on insights from online Twitter analysis, we developed machine-learning classification models incorporating patient attributes and pre- and post-COVID symptoms to distinguish LCS cases from other health conditions. The most effective predictive methodology combined logistic regression with elastic net regularization and Random Under-Sampling, achieving a sensitivity of 0.95, specificity of 0.81 and an AUC of 0.94. These findings highlight the potential of combining social media data with advanced analytical techniques to enhance our understanding of Long COVID Syndrome. By leveraging natural language processing and association rule mining, we have identified key symptoms and relationships within LCS. Furthermore, integrating machine learning models offers a robust approach to accurately identifying LCS patients, contributing to more precise diagnoses and improved treatment strategies.

Keywords: Association rule mining, Class imbalance, Elastic net classification, Long covid, Natural language processing

Addressing Class Imbalance in Classification Tasks with Applications

Muthukumarana S.[*]

Department of Statistics, University of Manitoba, Canada

Saman.Muthukumarana@Umanitoba.ca

0000-0001-8942-5352

Class imbalance is a common challenge in classification tasks, particularly in real-world applications such as medical diagnosis, fraud detection and industrial anomaly detection, where minority classes are often underrepresented. Traditional classification models tend to be biased toward the majority class, leading to suboptimal performance in predicting the minority class. This talk explores this critical issue by introducing four novel Synthetic Minority Over-sampling Technique (SMOTE) extensions: Distance ExtSMOTE, Dirichlet ExtSMOTE, FCRP SMOTE and BGMM SMOTE. These methods leverage a weighted average of neighboring instances to enhance the quality of synthetic samples and mitigate the impact of outliers. Comprehensive experiments conducted on diverse simulated and real-world imbalanced datasets demonstrate that the proposed methods improve classification performance compared to the original SMOTE and its most competitive variants. We also demonstrate that Dirichlet ExtSMOTE outperforms most other proposed and existing SMOTE variants in terms of achieving better F1 score, MCC and PR-AUC.

Keywords: Bayesian gaussian mixture model, Chinese restaurant process, Dirichlet distribution

Democracy Dies in Darkness Without Official Statistics

Splendore L. D. G.[*]

University of Malta, funding Horizon Europe by Marie Skłodowska-Curie Actions, Malta

luca.s.digennaro@um.edu.mt

0009-0007-0860-8837

Is there a quantitative relationship between democracy and official statistics? Why do the National Statistics Office (NSO) disseminate their statistics free of charge? How can a democratic country organize an election without official statistics? How close are the NSO staff to politicians? In a world increasingly shaped by data, the importance of official statistics often remains in the shadows, even within NSOs. Yet, these statistics (like employment rates, gross domestic product and pandemic data) are the lifeblood of democracy, influencing policymaking, media narratives and electoral choices. In this work, the vital nexus between official statistics and democracy is highlighted by quantitative data and historical prospective (Keynes, Schumpeter and Popper). The correlation, I calculated, between the Democracy Index and the Statistical Performance index (SPI) for 2019 for 167 countries is 70.1%. High income countries have a high level of Democracy and SPI. This work not only underscores the global significance of official statistics, but also aligns with the United Nations Fundamental Principles of Official Statistics and the Sustainable Development Goals.

Keywords: Democracy, Fundamental principles of official statistics, Official statistics, National statistics office, Sustainable development goals

Unleashing the Power of Data: Harnessing the Synergy between Official Statistics and Data Science in Nigeria

Dzaan K. S.[1*] and Ojo E.[2]

[1]Central Bank of Nigeria, [2]National Bureau of Statistics

[1]ksdzaan@cbn.gov.ng, [2]eojo@nigerianstat.gov.ng

[1]0009-0007-6372-4785, [2]0009-0005-2910-5134

Nigeria, the most populous nation in Africa with an estimated population of well over 210 million and a rapidly growing economy, is increasingly recognizing the pivotal role of data in driving development and informed decision-making. The National Statistical System (NSS) coordinated by its national statistics office plays a crucial role in collecting, analyzing and disseminating accurate and reliable data. However, a synergistic approach that integrates statistical methodologies with advanced data science techniques is essential to harness the potential of data fully. The need for this synergy has been recognized by the Statistical System in Nigeria by joining the United Nations High Impact Initiative on Unlocking the Power of Data. This initiative seeks to harness the incredible potential of data and technology to accelerate progress towards attaining the Sustainable Development Goals (SDGs). Furthermore, the system has publicly expressed plans to develop a framework for establishing a Data Science Campus in the National Bureau of Statistics in the 2024 fiscal year. This paper explores the intersection of statistics and data science within the Nigerian context. It examines the contributions of the NSS in providing foundational statistical data and highlights the potential of data science to extract deeper insights and uncover hidden patterns. The paper discusses the challenges and opportunities faced by Nigeria in leveraging this synergy and proposes recommendations for strengthening the NSS and fostering a data-driven culture. By effectively combining statistical analysis with data science techniques, Nigeria can unlock the power of its data, drive innovation and improve the lives of its citizens.

Keywords: Data science, Official statistics, Synergy

Machine Learning for Blockchain Data Analysis: Progress and Opportunities

Akcora C. G.[*]

AI Initiative, University of Central Florida, USA

cuneyt.akcora@ucf.edu

0000-0002-2882-6950

Blockchain technology has rapidly emerged to mainstream attention. At the same time, its publicly accessible, heterogeneous, massive-volume and temporal data are reminiscent of the complex dynamics encountered during the last decade of big data. Unlike any prior data source, blockchain datasets encompass multiple layers of interactions across real-world entities, e.g., human users, autonomous programs and smart contracts. Furthermore, blockchain's integration with cryptocurrencies has introduced financial aspects of unprecedented scale and complexity, such as decentralized finance, stablecoins, non-fungible tokens and central bank digital currencies. These unique characteristics present opportunities and challenges for machine learning on blockchain data. On the one hand, we examine the state-of-the-art solutions, applications and future directions associated with leveraging machine learning for blockchain data analysis critical for improving blockchain technology, such as e-crime detection and trends prediction. On the other hand, we shed light on blockchain's pivotal role by outlining vast datasets and tools that can catalyze the growth of the evolving Machine Learning ecosystem. Our presentation will serve as a comprehensive resource for researchers, practitioners and policymakers, offering a roadmap for navigating this dynamic and transformative field.

Keywords: Bitcoin, Blockchain, Blockchain data analytics, Ethereum

Partitioning Variance in Generalized Linear Mixed Models

Rudra P.[1*] and Olaifa, J.[2]

[1,2]Department of Statistics, Oklahoma State University

[1]rudra@okstate.edu, [2]julius.olaifa@okstate.edu

[1]0000-0002-1089-7283, [2]0000-0002-3864-099X

Heritability, the proportion of variation in traits due to genetic factors within a population, serves as a fundamental metric in quantifying the genetic influences on these traits. Traditional approaches for estimating heritability, such as family-based designs or intra-class correlations in ANOVA, have provided foundational insights. However, when measuring heritability of intermediate traits from high-throughput sequencing count data, the need for more sophisticated models has become apparent. An earlier study introduced model-based heritability scores using generalized linear mixed models, focusing solely on genetic factors without considering covariates. This limitation presents a gap in understanding how different treatments, demographic factors and other covariates influence heritability estimates. We propose a more comprehensive framework for estimating conditional heritability that incorporates covariates, offering a nuanced view of heritability in high-throughput sequencing data. This approach not only adheres to the complexity of modern genetic data but also enhances the accuracy of heritability estimates by accounting for a broader range of influencing factors. Our approach involves partitioning the variability using a variance partition coefficient (VPC) using a generalized linear mixed model. We have then derived various asymptotic results on VPC that can help us conducting hypothesis tests and constructing confidence intervals in this context where we often have parameters on the boundary of the parameter space.

Keywords: Generalized linear mixed model, Heritability, Variance partition

What Does Rally Length Tell Us About Player Characteristics in Tennis?

Epasinghege D. N.[1*], Gill P. S.[2] and Swartz T. B.[3]

[1,3]Simon Fraser University, Canada, [2]University of British Columbia Okanagan, Canada

[1]nepasing@sfu.ca, [2]paramjit.gill@ubc.ca, [3]timothy_swartz@sfu.ca

[1]0000-0003-4239-7432, [2]0000-0003-2813-8127, [3]0000-0001-6092-6727

This project proposes increasingly complex models based on publicly available data involving rally length. The models provide insights regarding player characteristics involving the ability to extend rallies and relates these characteristics to performance measures. The analysis highlights some important features that make a difference between winning and losing and therefore provides feedback on how players may
improve.

Keywords: ATP and WTA tours, Bayesian inference, Rally length, Tennis analytics

Analysis of the Impact of Unforced Errors in Tennis

Peiris H.[1*], Epasinghege D. N.[2] and Swartz T. B.[3]

[1,2,3]Simon Fraser University, Canada
[1]hashan_peiris@sfu.ca, [2]nirodha_epasinghege_dona@sfu.ca, [3]timothy_swartz@sfu.ca

[1]0000-0002-4721-9881, [1]0000-0002-4721-9881, [2]0000-0003-4239-7432, [3]0000-0001-6092-6727

This study considers the impact of unforced errors in sport. Although the proposed methods are applicable to various sports, we demonstrate the approach in the context of professional tennis. The value of the approach is that we can provide estimates of the points lost, games lost, sets lost and matches lost due to unforced errors. The methods are based on a bootstrapping procedure which also yields standard errors for the estimates. The approach is valuable in terms of player evaluation and can also be used for training purposes where it is possible to assess the quantification of improvement based on fewer unforced errors.

Keywords: ATP and WTA tours, Bootstrapping, Tennis analytics

Identify Soccer Styles

Swartz T. B.*

Simon Fraser University, Canada

timothy_swartz@sfu.ca

0000-0001-6092-6727

This talk concerns a problem in soccer analytics that relies on tracking data. We develop a metric that identifies soccer players who have a similar style to a player of interest. Whereas performance variables have been widely studied, the same is not true of stylistic variables. Unlike assessments from scouting, the metric is automatic and objective. The metric is developed using a Bayesian framework.

Keywords: Bayesian Gaussian mixture model, Chinese restaurant process, Dirichlet distribution

Statistical Modeling for Estimating Baseball Batting Metrics

Muthukumarana S.*

Department of Statistics, University of Manitoba, Canada

Saman.Muthukumarana@Umanitoba.ca

0000-0001-8942-5352

This talk presents a statistical modeling approach to estimate key baseball batting metrics, offering a robust framework for analyzing player performance. By employing weighted likelihood and semi-parametric Bayesian approach based advanced modeling techniques, we address challenges such as variability in player statistics and sparse data. Our approach leverages a combination of historical data and modern statistical methods to provide more accurate and reliable estimates of batting metrics. The proposed models are demonstrated using Major League Baseball data, showcasing their potential to enhance the understanding of player performance and inform decision-making in the sport.

Keywords: Dirichlet process, Multinomial distribution, Weighted likelihood

Unveiling New Dimensions in Gait Dynamics Through an Advanced Self-similar Measure

Vimalajeewa H. D.[1*], Hinton Jr. R.[2], Ruggeri F.[3] and Vidakovic B[4]

[1]Department of Statistics, University of Nebraska Lincoln, USA, [2,4]Department of Statistics, Texas A&M University, USA, [3]Italian National Research Council, Italy

[1]hvimalajeewa2@unl.edu, [2]rhinton@tamu.edu, [3]fabrizio@mi.imati.cnr.it, [4]brani@stat.tamu.edu

[1]0000-0001-6794-4776, [2]0009-0000-1733-1613, [3]0000-0002-7655-6254, [4]0000-0001-9155-9325

The study introduces a novel method, Average of Level-Pairwise Hurst Exponent Estimates (ALPHEE), which refines the estimation of the Hurst exponent by integrating wavelet transforms and fractional Brownian motion models. This method focuses on self-similarity in natural processes, particularly using the Hurst exponent to quantify it. Wavelet-based techniques are typically used to estimate the Hurst exponent, but they can be affected by noise and outliers. The ALPHEE method accounts probabilistic properties of wavelet coefficients of signals to overcome these laminations and improve the Hurst exponent estimation performance. The method is applied to gait data from elderly adults to identify those who have experienced falls. By analyzing acceleration and angular velocity data from 147 subjects, the study shows that self-similarity features improve the detection of fallers. The proposed method achieves an accuracy 84%, outperforming the standard approach, which has a 79.55% accuracy. This improvement highlights the method's ability to better capture self-similar properties, leading to enhanced performance in gait analysis.

Keywords: Gait data analysis, Hurst exponent estimation, Self-similarity, Wavelet transform

Deep-ExtSMOTE: Integrating Autoencoders for Advanced Mitigation of Class Imbalance in High-dimensional and Big Data Classification

Matharaarachchi S.[1*], Domaratzki M.[2] and Muthukumarana S.[3]

[1,3]Department of Statistics, University of Manitoba, Winnipeg, MB, Canada,
[2]Department of Computer Science, Western University, London, ON, Canada

[1]matharas@myumanitoba.ca, [2]mdomarat@uwo.ca, [3]saman.muthukumarana@umanitoba.ca

[1]0000-0002-9490-4742, [2]0000-0001-9129-6676, [3]0000-0001-8942-5352

Class imbalance and the curse of dimensionality pose significant challenges in machine learning, particularly in high-dimensional classification tasks. This study introduces Deep-ExtSMOTE, a novel technique designed to tackle these challenges by integrating autoencoder-based dimensionality reduction with an extended version of SMOTE (Synthetic Minority Over-sampling Technique). Deep-ExtSMOTE leverages the power of autoencoders to reduce data dimensionality and capture complex, non-linear relationships, while the advanced resampling capabilities of Extended SMOTE generate synthetic samples through weighted averages of neighboring points. This integration aims to improve the representation of minority classes and enhance the overall robustness of classification models. Our empirical evaluation, conducted in both simulated and real-world scenarios, reveals that Deep-ExtSMOTE significantly outperforms traditional SMOTE and autoencoders, particularly in high-dimensional datasets. The method shows substantial improvements in classification performance, as indicated by higher F1 scores and effectively overcomes the limitations of existing techniques, highlighting its potential to boost model accuracy and reliability.

Keywords: Autoencoders, Class imbalance, Curse of dimensionality, Dirichlet ExtSMOTE, SMOTE

Single-cell Data Integration Using Optimal Transport

Demetci P.[1], Santorella R.[2], Tran Q. H.[3], Chakravarthy M.[4], Redko I.[5], Sandstede B.[6] and Singh R.[7*]

[1]Broad Institute of MIT and Harvard, USA, [2]OM1 Research, USA, [3]Université Bretagne Sud, France, [4,6,7]Brown University, USA, [5]Noah's Ark Lab Huawei, France

[7]ritambhara_singh@brown.edu

[1]0000-0002-5644-0326, [2]0000-0003-1432-3241, [3]0000-0002-2739-3217, [4]NA, [5]0000-0002-3860-5502, [6]0000-0002-5432-1235, [7]0000-0002-7523-160X

Recent advances in sequencing technologies have allowed us to capture various aspects of the genome at single-cell resolution. However, except for a few co-assaying technologies, it is not possible to simultaneously apply different sequencing assays on the same single cell. In this scenario, computational integration of multiple genomic measurements is crucial to enable joint analyses and drive discovery. This integration task is particularly challenging due to the lack of sample-wise or feature-wise correspondences across the single-cell measurements. Here, we present our methods - Single-Cell alignment with Optimal Transport (SCOT) and Augmented Gromov Wasserstein Optimal Transport (AGWOT) - unsupervised machine learning algorithms that use optimal transport to align single-cell multi-omics datasets. Our results demonstrate interesting properties of these methods, such as automatic hyperparameter selection, extensions to the integration setting where datasets can have disproportionate cell type representations and simultaneous alignment of cells and features for hypothesis generation in biology.

Keywords: Data integration, Genomics, Optimal transport, Single-cell sequencing

Bayes Estimation of Prevalence of AIH in females in the Las Vegas Metropolitan Area

Mukhopadhyay D.[1], Dalpatadu R. J.[2], Gewali L. P.[3] and Singh A. K.[4*]

[1]Department of Epidemiology and Biostatistics, [2]The Department of Mathematical Sciences, University of Nevada, Las Vegas, Nevada, [3]The Department of Computer Science, University of Nevada, Las Vegas, Nevada, [4]The William F. Harrah College of Hospitality, University of Nevada, Las Vegas, Nevada

[4]ashok.singh@unlv.edu

[4]0000-0001-8778-7958

Autoimmune hepatitis (AIH) is a liver disease which attacks the immune system of the body, causing swelling, irritation and damage to the liver. For patients who do not respond to immune-suppressants, a liver transplant is the only option. In this article, we compute Bayesian HPD credible sets for the prevalence of AIH in females in the Las Vegas Metropolitan Area.

Keywords: Bayes theorem, Frequentist approach, Bayesian updating, MCMC, R-package JAGS, prior distribution, posterior distribution, HPD Credible Set

INDUSTRY INVITED TALKS

Integration of Traditional and Telematics Data for Efficient Insurance Claims Prediction

Peiris H. [1*], Jeong H.[2], Kim J-K.[3] and Lee H.[4]

[1]Simon Fraser University, Canada, [2]Simon Fraser University, Canada, [3]Iowa State University, [4]Sungkyunkwan University, United States

[1]hashan_peiris@sfu.ca, [2]himchan_jeong@sfu.ca, [3]jkim@iastate.edu, [4]hangsuck@skku.edu

[1]0000-0002-4721-9881, [2]0000-0001-5695-9964, [3]0000-0002-0246-6029, [4]0000-0003-4749-803X

While driver telematics has gained attention for risk classification in auto insurance, scarcity of observations with telematics features has been problematic, which could be owing to either privacy concerns or favorable selection compared to the data points with traditional features. To handle this issue, we apply a data integration technique based on calibration weights for usage-based insurance with multiple sources of data. It is shown that the proposed framework can efficiently integrate traditional data and telematics data and can also deal with possible favorable selection issues related to telematics data availability. Our findings are supported by a simulation study and empirical analysis in a synthetic telematics dataset.

Keywords: Automobile insurance, Data integration, Driver telematics, Missing data analysis

From Historical Data to Future Insights: Actuarial Experience Studies in Life Insurance

Hansini R. A. M.*

Deloitte, Hong Kong

hmanosha@deloitte.com.hk

In today's dynamic insurance landscape, actuaries are crucial in forecasting future trends and guiding strategic decision-making. One of the critical tools at their disposal is the actuarial experience study, which focuses on analyzing historical data related to mortality, morbidity and policy lapses. By leveraging policy snapshots and transactional records, actuaries identify the critical drivers of actual outcomes compared to expected ones, allowing for calculating the actual-to-expected (A/E) ratio. This ratio serves as a benchmark, enabling actuaries to refine assumptions fundamental to financial planning, valuation, pricing and risk management. For insurers worldwide, including those in Sri Lanka, the importance of experience studies is only increasing. Companies seek more precise, company-specific assumptions to stay competitive in an evolving market. Experience studies provide a critical framework for understanding past behavior, which helps actuaries predict future trends more accurately. They use statistical analysis to derive and validate assumptions but must also weigh these findings against practical business realities. Professional actuarial judgment, therefore, becomes essential in balancing statistical rigor with the real-world implications of business decisions. This presentation will explore an experience study, how it is conducted and the professional judgments involved. First, we will focus on the methods used to perform an actuarial experience study, the pros and cons and the data collection and cleansing techniques. Further, we will examine how to calculate A/E ratios and derive assumptions through an illustrative example. Examining A/E trends can give us valuable insights into how past experiences have informed current actuarial practices. Beyond the technical aspects of data analysis, actuaries must apply professional judgment to interpret the results of experience studies within the broader context of business strategy to ensure that the assumptions derived from the data are statistically sound and aligned with the company's strategic goals. This presentation aims to provide an understanding of the indispensable value of actuarial experience studies. Attendees will learn how actuarial experience studies provide the foundation for deriving assumptions that influence future strategies. The talk will highlight how historical insights can be transformed into actionable recommendations, allowing insurers to navigate an ever-evolving landscape with greater confidence and preparedness.

Keywords: Actuarial science, Financial planning, Life insurance, Pricing, Risk management

Importance of Data – Sources and Accuracy

Gunatilleke S. K.*

Managing Director / Actuary – S G Actuarial Consultancy (Pvt) Ltd., Sri Lanka

saroja.gunatilleke@gmail.com

Actuarial calculations and projections performed for the insurance sector can be based on probabilities. And these probabilities can be based on past data. A previous session explored historical data being used for Experience Studies to come up with such probabilities. The data used for such assessments can be diverse. Collecting relevant and suitable data is as important as the methods used by actuaries for the above-mentioned calculations. This session will explore the various data sources available for the calculations for the insurance sector and how to select one. And then go on to discuss the paramount importance of checking the reasonableness and accuracy of the data as it can significantly impact, the results of the study.

Keywords: Accuracy, Data, Sources

Statistical Foundations of AI: How Data Science Builds on Classic Statistical Methods

Abeyratne D.*

MAS Capital Pvt Ltd, Sri Lanka

dhanushkaabeyr@masholdings.com

0009-0005-8599-2399

The synergy between classical statistics and artificial intelligence (AI) forms the backbone of data-driven decision-making in modern industries. This session delves into how foundational statistical methods provide the rigor and reliability necessary for building robust, interpretable and scalable AI systems. As AI evolves to tackle increasingly complex challenges, its reliance on statistical principles remains crucial, ensuring that models are both scientifically grounded and practically applicable. The session begins with an exploration of probability theory and its central role in AI algorithms, such as Bayesian networks and reinforcement learning. These techniques enable AI systems to handle uncertainty, quantify risks and make probabilistic predictions critical for dynamic decision-making environments. Core statistical models, such as linear regression and hypothesis testing, are highlighted as essential precursors to machine learning techniques, demonstrating how they inform algorithm design and model evaluation. A focus on Bayesian methods underscores their value in modern AI, particularly in applications requiring continuous learning and uncertainty quantification. From probabilistic programming to real-time updates in predictive models, Bayesian statistics exemplifies how traditional techniques have been reimagined to meet the demands of AI applications. The session also examines machine learning algorithms built on statistical principles, including logistic regression, clustering and dimensionality reduction techniques like Principal Component Analysis (PCA). These methods showcase how statistical concepts are operationalized to process and interpret complex datasets, enabling data scientists to uncover actionable insights in fields such as healthcare, finance and marketing. Model evaluation and validation are addressed through statistical inference techniques, such as cross-validation, confidence intervals and the bias-variance tradeoff. These methods ensure that AI systems generalize effectively, perform reliably under uncertainty and meet industry requirements for transparency and accountability. Finally, the session looks ahead to emerging trends, such as hybrid models that integrate statistical reasoning with neural networks and the growing demand for explainable AI (XAI). By leveraging statistical methods, XAI seeks to enhance the interpretability of AI models, ensuring they are transparent and trustworthy for high-stakes applications. This session will emphasize the enduring relevance of statistical methods in AI and data science. Participants will gain a deeper understanding of how statistical principles enhance AI's reliability and applicability, making them indispensable for solving complex, real-world problems.

Keywords: Artificial intelligence (AI), Bayesian statistics, Data science, Probability theory, Statistical foundations

The Intersection of Statistics and Actionable Insights in Competitive Market Research Analysis: Shaping Brand Strategies and Better Decision-Making

Amaratunga S.[1], Sundarampillai A.[2*] and Lasanthi R.[3]

Aura Insights (Pvt) Ltd, Sri Lanka

[1]subashan@aurainst.com, [2]abhishanya@aurainst.com, [3]rajika@aurainst.com

By bridging the gap between complex data and actionable strategies, we use data modeling techniques to identify and analyze the key factors that drive the performance of different brands and product categories. We primarily use two advanced statistical tools: Multivariate Analysis and Correspondence Mapping. Multivariate analysis helps us understand a brand's value and how it is perceived in the market by looking at key metrics like brand awareness, brand consideration, brand trial, repeat purchase and recommendation. These metrics are combined to give a single measure of brand equity, which helps us assess the strength and performance of the brand. We also use correspondence mapping to find unique differences among brands and how they are linked to different brands and service features. By using a contingency table with groups of variables in the rows and columns, we analyze the residuals in the cells, index them and find the coordinates for correspondence analysis. These coordinates are then used to create the correspondence map. These tools help to turn actionable insights into effective brand strategies and informed decision-making.

Keywords: Brand Equity, Correspondence Mapping, Multivariate Analysis

Analytics for All to Drive Business Value by Leveraging Data Science Skills

Perera A.[1*], Senevirathne T.[2*], Dissanayake D.[3*] and Navarathna R.[4*]

[1,3,4]OCTAVE John Keells Holdings Group, Sri Lanka, [2]Mas Holdings Group, Sri Lanka

amalipe.jkh@keells.com, thisuras@masholdings.com, dinushad.jkh@keells.com, rajitha.jkh@keells.com

In today’s data-driven world, businesses are increasingly relying on analytical solutions to drive decision-making and unlock growth opportunities. However, crafting effective analytical solutions requires a structured approach and a clear understanding of the essential data science skills. This workshop is designed to provide participants with a hands-on understanding of howto tackle a business problem analytically, from identifying the core issue to structuring a solutionthat delivers measurable business value. We will guide participants through the process ofaddressing a real-world business problem by structuring an analytical solution step by step, emphasizing essential data science skills along the way. Participants will learn to break down complex business challenges, identify relevant data sources and apply key data science techniques to generate actionable insights. The focus will be on the essential skills such as data exploration, statistical modeling and result interpretation, such that it accessible to both aspiring data scientists and those new to the field. A key highlight of the workshop is the exploration of areas where analytical solutions can lead tobusiness value loss. Through interactive discussions, we will identify common pitfalls, such as misaligned objectives, poor communication of insights and failure to act on data-driven recommendations. Participants will gain strategies to ensure that their analytical efforts are not only technically sound but also aligned with business goals to maximize impact. The latter part of the workshop is dedicated to career opportunities in the analytical domain, witha special focus on pathways for non-Math major participants. We will discuss how individuals from diverse educational backgrounds—such as Arts, Commerce, Agriculture and Bio-Science—can start successful careers in analytics by building essential skills in areas such as data visualization, storytelling and critical thinking. Participants will learn how individuals from diverse backgrounds can contribute to the field by acquiring essential analytical skills. By the end of this workshop, participants will have a comprehensive understanding of theanalytical problem-solving process, strategies to avoid business value loss and a clear roadmap for pursuing a career in analytics. Whether you’re a student, a working professional, or simply curious about the field, this workshop will empower you with practical insights and actionable takeaways to navigate the exciting world of data analytics.

Keywords: Business value loss, Data driven solution, Data science skills, Diverse educational backgrounds

CONTRIBUTED ABSTRACTS

ORAL PRESENTATIONS

Optimizing Inkjet Printing Parameters for High-resolution Microfluidic Device Fabrication

Andrews U. T.[1*], Jayatillake R. V.[2] and Weerawarne D. L.[3]

[1,2]Department of Statistics, University of Colombo, Sri Lanka, [3]Department of Physics, University of Colombo, Sri Lanka

[1]umeshaandrews@gmail.com, [2]rasika@stat.cmb.ac.lk, [3]dweerawa@phys.cmb.ac.lk

[1]0009-0005-4526-2180, [2]0000-0002-5414-999X, [3]0000-0002-2573-6298

Microfluidic devices, used in many fields ranging from medical diagnostics to biological analysis, heavily rely on precise fabrication methods. Inkjet printing has emerged as a promising approach due to its high-resolution capabilities and cost-effectiveness. However, optimizing key parameters such as droplet control, ink formulations and ink-substrate interactions is essential for achieving the necessary precision. Although ongoing research addresses some of these challenges, there is a lack of studies focused on optimizing inkjet printer factors specifically for microfluidic fabrication. This study fills that gap by exploring the impact of the four key factors of inkjet printers: printing quality, paper quality, paper type and color intensity, along with their interactions, on microfluidic test element fabrication, testing 81 combinations to identify optimal settings. The study investigates the 12 response variables, focusing on deviations from three required linewidths (100 µm, 150 µm and 200 µm) and gaps (250 µm, 500 µm and 750 µm) between lines for each linewidth. A full factorial completely randomized design (CRD) with three replicates per factor combination, was constructed utilizing Minitab software, resulting in 243 potential experimental runs. The run order was randomized to eliminate bias, ensuring the accuracy and reliability of the results. A test coupon was designed using Inkscape software to assess these parameters, allowing for the compact and efficient representation of all measuring elements. It was then printed using an Inkjet printer. Images of the printed lines and gaps were taken using a microscope and analyzed with Python code for automated measurement. Factorial Analysis of Variance (ANOVA) identified statistically significant factors affecting dimensional deviations. Results indicate the highest-order interactions are significant. Paper type and color intensity emerge as influential factors affecting linewidth and gap deviations, highlighting intricate interaction effects. The main assumptions of ANOVA were satisfied, while some violations of the normality assumption were noted. Statistical optimization through the Taguchi method was employed to determine optimal settings for minimizing deviations. The results from this study demonstrate that some lower-level factor combinations delivered more accurate results than using the highest levels of all factors. For 100 µm lines, optimal settings included standard printing quality and Epson premium quality with 100% color intensity on photo paper. For a 250 µm gap optimal settings were standard printing quality on PET paper with Epson matte quality and 90% color intensity. A 500 µm gap was best with photo paper and Epson matte quality at 80% color intensity and standard printing quality, while a 750 µm gap required high printing quality on PET paper, Epson matte quality and 100% color intensity. Similar results were obtained for other line widths (150 µm, 200 µm). Highlighting the importance of optimizing printing parameters to minimize deviations and improve the precision in inkjet printing for microfluidic test element fabrication.

Keywords: Full factorial, Inkjet printing, Microfluidic devices, Statistical optimization, Taguchi method

Joint Modeling of Survival and Count: A Review

Hapugoda J. C.[1*] and Sooriyarachchi M. R.[2]

[1]Department of Organizational Studies, The Open University of Sri Lanka, Sri Lanka, [2]Independent Researcher

[1]jchap@ou.ac.lk, [2]roshinis@hotmail.com

[1]0000-0001-7186-4488, [2]0000-0002-4551-0616

Recently, joint modeling of data with many response types has been a popular topic of research. When there is an association between two or more responses, a joint model may yield more interesting and improved findings than univariate models. These models will be relevant to several research/fields/studies. Medicine is one of the most common areas for such research, particularly in relation to survival analyses. Many researchers in the field of survival analysis have constructed joint models for longitudinal and survival outcomes. Furthermore, survival with non-longitudinal data, bivariate count data, bivariate binary data and bivariate binary and count data have recently been published, contributing to the advancement of this field of study. There are numerous ways of developing joint models. The objectives of this study are to carry out an extensive review of the literature to identify the various approaches that exist for modeling survival and count variables jointly and to suggest extensions to existing bivariate modeling techniques for survival and count variables. The methodology of the study is a comprehensive search that was conducted in the search engine using keywords. The findings include various scenarios where joint models of survival and count were developed and applied. Thus, many future potentials are available in this area to come up with different extensions.

Keywords: Approaches of joint modeling, Clustered data, Count, Joint modeling, Survival

Unveiling Elephant Mortality Patterns in Sri Lanka: Trends and Insights

Herath R. M.[1*], Dehideniya M. B[2], Senevirathne, T. M. D.[3] and Senavirathna R.[4]

[1]ravindumadushan279@gmail.com, [2]mahasen@sci.pdn.ac.lk, [3]thisura_mindula@hotmail.com, [4]rajitha.jkh@keells.com

[1,2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka, [3]MAS Holdings (PVT) LTD., Sri Lanka, [4]OCTAVE, John Keels Group, Sri Lanka

Sri Lanka was home to approximately 19,500 wild elephants at the start of the 19th century. However, due to rampant hunting and agricultural encroachment, this number plummeted to just 2,000 by the early 20th century. In recent years, the issue of elephant mortality has become more pronounced, with rising numbers of deaths, often in isolated incidents. This research examines elephant mortality data from 2010 to 2017, during which 1,894 elephants died across 18 districts, averaging 237 deaths annually. Deaths were at their lowest in June and peaked during February, March, September and October periods aligned with cultivation cycles: "MAHA KANNA" and "YALA KANNA". This highlights the strong link between human-elephant conflicts (HEC) and agriculture, especially in human-dominated landscapes outside protected areas. Gunshot wounds were identified as the leading cause of death, followed by hakkapatas (explosive devices) and electrocution. Adult and sub-adult elephants, primarily males, were most affected, with 67.50% of fatalities occurring among male elephants. Mortality patterns varied across districts, with the dry zone regions of Anuradhapura, Polonnaruwa and Ampara accounting for 43.18% of the deaths. This study underscores the urgent need to address the human-elephant conflict, particularly in Sri Lanka's dry zone, where agriculture and wildlife intersect.

Keywords: Asian Elephant, Data visualization, Human-elephant conflict

Development of Telematics Risk Scores in Accordance with Regulatory Compliance

Jeong H.[1], Zou B.[2] and Peiris H.[3*]

[1,3]Department of Statistics and Actuarial Science, Simon Fraser University, Canada, [2]Department of Mathematics, College of Liberal Arts and Science, University of Connecticut, United States

[1]himchan_jeong@sfu.ca, [2]bin.zou@uconn.edu, [3]hashan_peiris@sfu.ca

[1]0000-0001-5695-9964, [2]0000-0002-7390-7979, [3]0000-0002-4721-9881

The expansion of Usage Based Insurance (UBI) market allows the integration of telematics data into risk-scoring models. However, incorporating legal regulations into these models is still evolving. This study suggests a framework to provide an embedded risk score through a Feedforward neural network (FNN) and fit a model utilizing traditional data and the risk score as a generalized linear model (GLM), allowing regulatory constraint to reduce premium surcharges based on telematics data. Moreover, it discusses the impact of potential selection bias on premiums when policyholders with lower risk profiles are more likely to opt into UBI. An empirical analysis of a synthetic telematics data set supports the findings. Since telematics embedding models and their variants perform comparable to the raw telematics model in both in-sample estimation and out-of-sample validation, the proposed framework can be considered a regulatory-compliant alternative. Also, it implies that the discount-only regulation could provide no surcharge for those who choose telematics policies over traditional ones when favorable selection occurs by leading to market segmentation.

Keywords: Adverse selection, Automobile insurance, Driver telematics, Insurance regulation, Risk classification

Enhancing Predictive Performance of the Logistic Two-Parameter Estimator under Multicollinearity

Kayathiri T.[1*], Kayanan M.[2] and Wijekoon P.[3]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka, [2]Department of Physical Science, University of Vavuniya, Sri Lanka, [3]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

[1]skayathiri1994@gmail.com, [2]kayanan@vau.ac.lk , [3]pushpaw@sci.pdn.ac.lk

[1]0009-0008-0870-5722, [2]0000-0003-2662-4383, [3]0000-0003-4242-1017

This study aimed to propose a new estimator named Modified Logistic Two-Parameter Estimator (MLTPE) with the best predictive performance in logistic regression when multicollinearity exists. The performance of the proposed estimator was compared with existing estimators: Maximum Likelihood Estimator (MLE), Modified Almost Unbiased Ridge Logistic Estimator (MAURLE), Modified Almost Unbiased Ridge Logistic Liu Estimator (MAULLE) and Logistic Two-Parameter Estimator (LTPE), in terms of the scalar mean squared error (SMSE) and balanced accuracy. The performance of the estimators was evaluated using Monte Carlo simulations. In the design of the experiment, factors such as the degree of correlation and sample size were varied. The results showed that the performance of the estimators depended on these factors. Finally, the theoretical results were applied to a myopia real-world dataset and observed that the results agreed with the simulation study's results. It was noticed that the MAURLE performs well in terms of SMSE; however, the proposed estimator showed slightly better performance in terms of balanced accuracy.

Keywords: Balanced accuracy, Logistic regression, Predictive performance, Scalar mean squared error

Study on the Efficiency of Fourier Shape Descriptors in Capturing Morphological Changes in Crushed Gneiss Aggregates Induced by Ball Milling

Mithulavan V.[1*], Sathiparan N.[2] and Subramaniam D. N.[3]

[1]Department of Electrical and Electronic Engineering, University of Jaffna, Sri Lanka,
[2,3]Department of Civil Engineering, University of Jaffna, Sri Lanka
[1]2020e197@eng.jfn.ac.lk, [2]sakthi@eng.jfn.ac.lk, [3]Daniel.subramaniam@gmail.com

[1]0009-0003-0703-2582, [2]0000-0001-8570-0580, [3]0000-0002-1023-1617

Material that constitutes aggregate depends on their packing efficiency that is defined by the shape distribution of aggregates. Geometrical shape descriptors would not capture morphological aspects of different dimensional scales and frequencies, failing to wholistically numerically represent shape in material prediction models. Fourier shape descriptors are used to analyze single particle morphology, but the efficiency to characterize the shape of lump samples has not been assessed. This study analyses aggregates milled for different number of revolutions (0 and 2000) in a Los Angeles Abrasion Value instrument, that contrives morphological alterations of different scales. The morphology based signal obtained was transformed into the frequency domain and the harmonics were split into three zones. Zone Form: first to sixth harmonics, zone angularity: seventh to sixteenth harmonics and zone texture: seventeenth to fifty third harmonics. All harmonics with a frequency higher than fifty third harmonics are considered as noise. The mean, median, variance and total of the zonal amplitudes were considered for subsequent analyses. A statistically significant difference with 95% confidence was observed for all twelve parameters, between 0 and 2000 Rev aggregates. The highest impact of milling was observed in the texture zone while the lowest was observed in the form zone, according to the F-score. This observation was also observed in feature ReliefF and Minimum redundancy maximum Relevance feature selection methods. 97%, 85% and 86% accuracy was obtained with Boosted-Forest, Support-Vector-Machine and Ensemble machine learning classification methods. The study indicated a significant difference in shape features of aggregate lump samples, based on Fourier Shape Descriptor Means.

Keywords: Angularity, Form, Fourier shape descriptor, Geometric shape descriptor texture, Supervised classification

Investigating the Impact of Censoring Proportion and Hazard Patterns on the Performance of Traditional Survival Analysis Approaches

Opatha O. M. D. N. W.[1*] and Jayasinghe C. L.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]nipuniopatha@gmail.com, [2]chathuri@sjp.ac.lk

[1]0009-0007-0456-5497, [2]0000-0003-2891-742X

Survival data analysis is distinct due to its potential of handling data with the characteristic of censoring. Censoring occurs when the event of interest has not yet occurred for some individuals or the exact event time is unknown. The proportion of censored observations in the data in most scenarios cannot be controlled but it can impact the accuracy and reliability of model estimates and predictions. The Cox Proportional Hazards model and the Accelerated Failure Time (AFT) model are two prominent methods used in survival analysis, each with its strengths and assumptions. The Cox model is a semi-parametric model, relying on the proportional hazards assumption and partial likelihood for parameter estimation, while the AFT model is fully parametric, assuming a linear relationship between the logarithm of survival time and covariates. This study aims to evaluate the impact of censoring proportion and hazard pattern on the performance of these models using simulated data of varying sample sizes with various hazard behaviors, identifying the best-performing model under different conditions. The simulation of the survival times was conducted using the 'simsurv' R package by using simulated covariate datasets with varying effects on the response. The event times were simulated under five sample sizes: 20, 100, 500, 1000 and 10000; five censoring proportions: 0%, 5%, 20%, 50% and 80%, across three different hazard behaviors (increasing, decreasing and bathtub-shaped) and two censoring types (Type 1 right censored and Type 3 right censored). One hundred datasets were generated for each combination to ensure robust results. The AFT model was fitted by assuming a Weibull distribution for the survival times. The models' performances were compared using Harrell's Concordance Index, calculating the mean index across the datasets and their variations were also compared. Results indicated that both models performed similarly overall but exhibited differences under specific conditions. The Cox model struggled with very small sample sizes, particularly with increasing hazard data, due to its complex baseline hazard estimation. Conversely, the AFT model, with its simpler assumptions, performed better in these scenarios. Both models demonstrated high performance with increasing hazard patterns but did not show significant changes in performance as the censoring proportion increased. For decreasing and bathtub-shaped hazard patterns, the performance remained relatively stable regardless of the censoring proportion in the data. These findings suggest that while the Cox model offers flexibility, the AFT model is more robust for small samples or increasing hazards. In conclusion, the choice between the Cox and AFT models depends on the hazard function of the data and sample size. The Cox model is flexible but less effective with small samples and increasing hazards, whereas the AFT model excels in these conditions. Both models' performance declines with higher sample sizes for decreasing and bathtub-shaped hazards. Future studies should consider additional hazard shapes such as bell-shaped and inverted bathtub and consider other predictive performance measures like the Integrated Brier Score to enhance the accuracy of conclusions.

Keywords: Accelerated failure time, Censoring proportion, Concordance index, Cox regression, Survival analysis

Checking for Collaboration in Online Multiple-choice Testing

Perera H.[1*], Silva R. M.[2] and Swartz T. B.[3]

[1,3]Department of Statistics and Actuarial Science, Simon Fraser University, Canada, [2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]gperera@sfu.ca, [2]rsilva@sjp.ac.lk, [3]tswartz@sfu.ca

[1]0009-0009-1647-807X, [2]0000-0002-4915-7979, [3]0000-0001-6092-6727

This paper, proposes a novel two-step procedure to detect potential collaboration among students during multiple-choice tests, with the aim of maintaining academic integrity in online education and testing settings. Cheating in online environments can involve much larger groups of students and therefore, traditional pairwise detection methods may not be effective. In the first step, suspicious common responses were identified using probabilistic reasoning and applied the UPGMA algorithm to cluster students who may have collaborated. In the second step, the probability of timing-related patterns among clustered students were calculated to provide additional evidence of collaboration. An example of how to implement these two steps and demonstrate their effectiveness in identifying potential cheating incidents is presented. Our proposed method offers a practical solution for maintaining academic integrity in online education and testing settings.

Keywords: Academic integrity, Clustering, Geometric mean, Probabilistic reasoning

SerpentSleuth: A Deep Ensemble Learning Approach for Venomous and Non-venomous Snake Identification

Abeysinghe C.[1*], Vidanagamachchi. S. M.[2] and Chathuranga L. L. G.[3]

[1]Department of Computer Science and Engineering, University of Westminster, London, UK,
[2,3]Department of Computer Science, University of Ruhuna, Sri Lanka

[1]cabeysinghe16@gmail.com, [2]smv@dcs.ruh.ac.lk, [3]gihan@dcs.ruh.ac.lk

[1]0009-0000-4121-9592, [2]0000-0002-2245-4527, [3]0000-0001-5484-6181

Snakes, with over 3900 species worldwide, cause about 5.4 million bites annually, with 100,000 resulting in death according to the World Health Organization. Killing snakes out of fear of being bitten is problematic. Snakes are important as predators and ecosystem engineers, giving people economic and medicinal benefits. Much research attention has been given to identifying snake species. At the same time, there is less research already conducted in computer vision and deep learning for classifying a snake as venomous or non-venomous. Although many deep learning models have been proposed for this purpose independently across different countries, none have been proposed for many other countries around the globe. Previous researchers have used Transfer Learning, Teachable Machines based on TensorFlow, the You Only Look Once (YOLO) algorithm, traditional machine learning models (K-Nearest Neighbors, Support Vector Machines and Logistic Regression), Vision Transformers and very few Ensemble models. This paper analyses snake classification utilizing a deep ensemble learning technique. Data collection, preprocessing, feature extraction, classification and finally, the analysis of the results are some of the steps involved in the methodology. First and foremost, datasets were collected from three Kaggle repositories, Facebook groups and the Department of National Zoological Gardens, Sri Lanka. Images with noise and duplicates were removed manually from the dataset. These were later resized for consistency in size and compatibility with the convolutional neural network. The capabilities of the ensemble model have been assessed by combining two transfer learning models along with a novel Convolutional Neural Network (CNN) approach. For the novel CNN model, MobileNetV2 and ResNet50, the test accuracies achieved are 64%, 82% and 82%, respectively. The proposed ensemble approach achieved 87% test accuracy and 0.94 Receiver Operating Characteristics and outperformed all the other models during the testing phase.

Keywords: Deep learning, Machine learning, MobileNet, ResNet50, Snake identification

Application of Transformer Models for Predicting Exchange Rates Related to Sri Lanka

Basnayake B. R. P. M.[1*] and Chandrasekara N. V.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka, [1]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

[1]pavithramalkibasnayake@gmail.com, [2]nvchandrasekara@kln.ac.lk

[1]0000-0002-0893-4524, [2]0000-0003-3232-837X

Forecasting exchange rate movements is vital for currency trading in financial markets, as fluctuations can greatly affect a country's economy. Currently, Sri Lanka faces economic challenges due to the depreciation of the LKR, leading to a trade deficit and rising inflation, impacting investor decisions. Nonetheless, the early identification of these effects is achievable through effective modeling and forecasting. The main objective of this study is to examine the effectiveness of utilizing time series transformer (TSF) models with various hyperparameter adjustments for forecasting exchange rates related to Sri Lanka. Based on the referred literature, this study is the first to explore the modeling of exchange rate movements using TSF models. Using daily exchange rate data for eight currency pairs against LKR from 2008 to 2022, this study compared the performance of TSF models with Seasonal Autoregressive Integrated Moving Average (SARIMA) and Double SARIMA (DSARIMA) models. Additionally, this study compared the performance of TSF models with neural network models, including Feedforward Neural Network (FFNN), Generalized Regression Neural Network, Time Delay Neural Network and Long Short Term Memory. The study used several combinations of historical lagged data and moving average indicators as input variables, with hyperparameters tuned through trial and error. Overall, the key findings of the fitted TSF models indicated that a larger size of embeddings (*d_model*) in the positional encoding layer is required to capture complex historical patterns. Additionally, the inner layer dimensionality of the feed-forward layer (d_{ff}) achieved optimal performance when set to four times the value of *d_model*, thereby maintaining an effective balance between model capacity and computational requirements. Further, increasing the number of encoder and decoder layers beyond 4 led to higher error values and when the learning rate exceeded 0.7, error values increased. A batch size of 16 produced the lowest error, as smaller batches enhance generalization by adding variability to training, helping the model learn more robust features. Furthermore, an increase in the dropout rate within the feed-forward layer corresponded with a rise in error values due to information loss, which reduces the model's effective capacity to learn and represent data during training. The comparison of actual and fitted graphs and error values indicated that TSF models outperformed SARIMA and DSARIMA models. However, TSF models had slightly higher error values than the better-performing NN models. In the case of AUD/LKR, the SARIMA model had a mean absolute error (MAE) of 0.9767 and mean absolute percentage error (MAPE) of 0.0069, while the FFNN achieved a MAE of 0.4356 and MAPE of 0.0031 and the TSF model indicated a MAE of 0.5113 and MAPE of 0.0035. This study provides insights into optimizing transformer architectures for univariate time series forecasting with a comprehensive evaluation of hyperparameter configurations including a comparative analysis with traditional statistical models and other deep-learning approaches.

Keywords: Deep learning, Exchange rates, Hyperparameter, Time series forecasting, Transformer models

Detection of Unusual Bike Trips in New York Bike-sharing System Using Extended Isolation Forest Algorithm

Herath H. M. O. D.[1*] and Dehideniya M. B.[2]

[1,2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

[1]omalih@sci.pdn.ac.lk, [2]mahasen@sci.pdn.ac.lk

[1]0009-0006-5307-2915, [2]0000-0002-1798-8591

Bike-sharing systems are gaining popularity worldwide. These systems are flooded with large volumes of trip data and they often include unusual trips that deviate significantly from the typical trips observed. Identifying such unusual bike trips is an important task in discovering the underlying pattern using machine learning methods, as these anomalies can significantly impact the accuracy of the finding. However, given the amount of data recorded in these systems, detecting unusual trips is a challenging task. Thus, it is necessary to have an efficient and effective system to identify such anomalies in large amounts of trip data. This study focused on detecting unusual bike trips in the New York bike-sharing system. To achieve this, a machine learning-based anomaly detection workflow was developed, utilizing the extended isolation forest algorithm, which is particularly effective in handling categorical variables. Further, a systematic approach was proposed to select the most suitable threshold instead of using a handpicked threshold value for the anomaly score that determines whether a trip was unusual. According to the results, trip duration, starting hour and day of the week were identified as highly relevant features in identifying unusual trips. Most unusual trips were associated with casual users and electric bikes and typically involved longer trip durations. In contrast, usual trips were commonly linked to classic bikes and subscribed members, who tend to use the system regularly and take shorter trips. The proposed machine learning workflow can be readily applied to detect unusual trips in any bike-sharing system in the world.

Keywords: Anomaly detection, Categorical variables, Threshold selection

An Approach for Predictive Modeling of Stock Price Movements Using Machine Learning Algorithms: A Comparative Study of Selected Equities in the Colombo Stock Exchange

Herath I. M. G. U. K.[1*], Pathberiya H. A.[2] and Devpura N.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]kalpniherath00@gmail.com, [2]hasanthi@sjp.ac.lk, [3]ndevpura@sci.sjp.ac.lk

[1]0009-0007-4187-7764, [2]0000-0002-0997-5897, [3]0000-0002-8299-0540

In stock trading, the ability to predict stock performance plays a critical role in decision making, helping investors and traders determine whether to buy, hold, or sell. In that case, identifying stock movement direction is significant rather than predicting stock prices. With the massive amount of daily data and high volatility in stock markets, the use of machine learning models has surged over the past decade, driven by advancements in technology. This study examines the shortcomings of current stock movement prediction methods, focusing on three companies A, B and C from the Colombo Stock Exchange, selected from the S&P SL 20 index. The study is conducted in two phases. In the first phase, the Daily Closing prices of stocks are predicted using previous lags of Daily Open, High, Low and Close Prices. K-Nearest Neighbors (KNN), Support Vector Regression (SVR) and Long Short Term Memory (LSTM) models are applied for comparison, along with feature and training set selection. The results show that the optimized feature set and training set length vary across companies and the machine learning model. This suggests that adopting a tailored approach for each stock and model, rather than relying on pre-defined models, is crucial for achieving the best results in stock price prediction. In the feature optimization process, these models are limited to using only the first two lags of Open, High, Low and Close prices as features. Including more lags leads to overfitting and reduces model performance. SVR outperforms the other models, achieving Mean Absolute Percentage Errors (MAPE) of 1.419833%, 0.6262777% and 0.9900023% for Companies A, B and C, respectively. Although these predictive models perform well in forecasting stock prices, they do not accurately identify stock movement direction. The accuracy of movement direction prediction is limited, with values of 0.333333, 0.4035 and 0.5364 for Companies A, B and C, respectively. The second phase of the study addresses this limitation by using the Daily Percentage Return of the Closing Price as the response variable, instead of the Daily Closing Price. Due to the higher volatility of Daily Percentage Return, more features are incorporated in the models. Various lags of technical indicators such as Simple Moving Average, Exponential Moving Average, Momentum and Rate of Change, etc. along with the S&P SL 20 Index and Consumer Price Index, are used. Furthermore, up to 30 lags of the Daily Percentage Returns for Open, High, Low and Close prices are included. SVR, the optimized machine learning model from the first phase, is applied in model development. While these models show poor Mean Absolute Error (MAE) values in predicting the Daily Percentage Return, they outperform the phase one models in identifying stock movement direction. Specifically, the accuracy values improve to 0.3684, 0.5874 and 0.59999 for Companies A, B and C, respectively. Compared to existing literature, this study demonstrates that identifying the stock direction by predicting Daily Percentage Returns is significantly more effective than using classification models for stock movement direction. Additionally, the results highlight the importance of incorporating additional features to enhance the accuracy of stock predictions.

Keywords: Colombo stock exchange, Data mining, Stock market predictions, Stock movement direction prediction, Tailored approach in stock prediction

Development of Sinhala Speech Emotion Recognition Models using Cross-lingual and Multi-lingual Approaches

Jayaweera Y. K.[1*] and Lakraj G. P.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]yasasj99@gmail.com, [2]pemantha@stat.cmb.ac.lk

[1]0009-0006-0532-7143, [2]0000-0003-3921-8552

Speech Emotion Recognition (SER) is a technology designed to identify and classify emotions conveyed in spoken language. By detecting emotional states such as happiness, sadness, anger and neutrality, SER systems enhance the naturalness and intuitiveness of human-computer interactions. SER's ability to understand and respond to emotional cues is vital for improving applications such as mental health monitoring, customer service and personalized user experiences. As the field of SER gains momentum, significant research has been conducted on resource-rich languages like English, French and Chinese. However, there is currently a lack of SER studies on the Sinhala language, which is a language that also faces the additional challenge of being under-resourced. SER in under-resourced languages in general is also an area that has not been adequately addressed. Therefore, in addressing this research gap, this study delves into the development and evaluation of SER models within diverse linguistic contexts, with the Sinhala language as the target language for the model. The study uses the Sinhala speech emotion dataset, SASER-D (Sinhala Automatic Speech Emotion Recognition Dataset), focused on the four emotions: happy, sad, angry and neutral, alongside datasets from other languages, including RAVDESS, CREMA-D, BASER-D and the Urdu Emotion dataset. The methodological framework comprises extensive data augmentation, feature extraction and the application of diverse approaches, employing both numerical features and visual representations for model development. Specifically, this study employs Random Forest and Support Vector Machine (SVM) for numerical features analysis and Convolutional Neural Network (CNN) for emotion classification using visual representations. A significant aspect of this research is the investigation of the linguistic impact on SER, examining cross-lingual and multi-lingual models, particularly in the context of under-resourced languages. The experimental results reveal that employing numerical features for model development yields better results compared to utilizing visual representations, with a notable accuracy gap of 11% observed between them. Specifically, eGeMAPSv01b feature set was revealed to be the most effective numerical feature set. While the Random Forest model achieved the highest accuracy of 67%, the SVM model obtained an accuracy of 63%. Cross-lingual models performed poorly compared to mono-lingual models, highlighting the significant impact of linguistic differences on SER effectiveness. However, multi-lingual models demonstrated greater efficacy, achieving an accuracy of 66% with Random Forest, which is close to that of mono-lingual models. The results also reveal that linguistic proximity among the languages used in cross-lingual and multi-lingual models is a crucial factor affecting performance. While mono-lingual models perform better within a single language, multi-lingual models are more suitable for under-resourced contexts, addressing data scarcity and improving generalizability.

Keywords: Cross-lingual SER, Multi-lingual SER, Sinhala SER, Under-resourced languages SER

A Clustering-based Machine Learning Approach to Forecast Cryptocurrency Prices

Kalutharage S. N.[1*] and Pathberiya H. A.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]sachinisansala2@gmail.com, [2]hasanthi@sjp.ac.lk

[1]0009-0001-8598-5018, [2]0000-0002-0997-5897

Cryptocurrencies have gained popularity as newly emerged assets offer new avenues for the financial market. However, predicting cryptocurrency prices remains challenging due to their volatile nature and complex dynamics. The study seeks to develop a forecasting model capable of predicting cryptocurrency prices with enhanced accuracy and reliability at high frequency by incorporating a changepoint analysis and clustering-based approach to machine learning algorithms. The analysis begins with the application of the pruned exact linear time (PELT) method for changepoint analysis, which was used to segment high-frequency intraday data from Bitcoin (BTC), Ethereum (ETH) and Cardano (ADA). This segmentation identifies distinct regions within the data where shifts in volatility occur, effectively capturing different price behavior patterns. Once the volatility regions were identified, regions were then used as inputs for clustering. K-means clustering is applied to group these regions based on similar volatility characteristics, such as high price variance, low price variance and volume variance. Machine learning algorithms, namely support vector regression (SVR), long short-term memory (LSTM) and random forest (RF), were trained on both non-clustered and clustered data to create predictive models for forecasting cryptocurrency closing prices at high frequency. SVR's flexible hyperparameter tuning allows for a balance of precision and generalization, making it ideal for managing the noise and volatility in cryptocurrency data. RF excels at capturing complex relationships in cryptocurrency data due to its ensemble learning approach and LSTM is well-suited for capturing temporal patterns in volatile time series data. By incorporating clustering, these models better target volatility patterns, improving forecasting accuracy. Error measures and R-squared values were derived from test data, demonstrating that clustered data based on the properties of time regions with different levels of volatility significantly improved model performance. For ADA, the SVR model's Mean Absolute Error (MAE) improved from 0.0018 to 0.0003, while RF model improvements were noted, with MAE reducing from 0.0010 to 0.0004. However, the LSTM model's results indicated no significant difference for ADA. For the BTC, the SVR model's best-clustered subset showed an MAE reduction from 27.0674 to 12.8647 and the LSTM model's MAE dropped from 29.2393 to 10.7100, while RF model improvements for BTC were less consistent. For ETH, the SVR model's MAE improved from 2.2837 to 0.3720 and the LSTM model's MAE decreased from 3.0901 to 0.3532. RF model's metrics showed similar improvements with the reduction of MAE from 1.4566 to 0.7359. These results highlight the enhanced predictive accuracy achieved through this approach, which could better inform trading strategies in high-frequency cryptocurrency markets. It is significant to consider different market conditions concerning changes in volatility when deriving forecasts, as this study demonstrates that clustering volatility patterns lead to more reliable price predictions. Overall, the study showed that data clustered based on the properties of time regions with different levels of volatility significantly improved the performance of machine learning models in forecasting cryptocurrency prices.

Keywords: Clustering, Cryptocurrency, Machine learning

Advanced Real-time Sign Language Conversion to Text and Audio Using YOLOv8, Transfer Learning and Data Augmentation

Kumari H. M. L. S.[1*] and Gammulle H. M.[2]

[1]Computer Center, University of Peradeniya, Sri Lanka, [2]Department of Mechanical Engineering, University of Peradeniya, Sri Lanka

[1]lihinisangeetha99@gmail.com, [2]harindugml@gmail.com

[1]0009-0001-1410-7840, [2]0009-0005-8044-1360

Real-time hand gesture recognition can connect people without knowledge of sign language with those who have disabilities, using a computer, as this approach is user-friendly and versatile. A real-time hand gesture system should focus on building a reliable interface that works well for every user. Currently, transfer learning models and You Only Look Once (YOLO) models, specifically the YOLOv8 model are used to train models that can predict different classifications. However, the unavailability of extensive labeled data poses challenges for these models, leading to difficulties in achieving satisfactory accuracy. As a result, data augmentation techniques are employed. The resulting models can be connected to a webcam to create a real-time sign language system, providing output in both text and audio formats. The proposed model achieved an accuracy of 99.6% with YOLOv8m and 98.62% with MobileNet. This research reveals that transfer models, specifically MobileNet, InceptionV3, DenseNet121 and DenseNet169, performed better than what was observed in recent studies. In this case study, American sign language has been highlighted, which has been collected in the form of three image datasets, namely RGB, gray and binary and all three datasets were used. The YOLOv8m model was on top of the MobileNet model indicating that it was the best model of choice for real-time hand gesture recognition. These findings fill a crucial research gap and present concrete ways of enhancing the accessibility of communication. Future work may improve the accuracy of models through better handling of images, more complex data augmentation and fine-tuning of the transfer learning models.

Keywords: Data Augmentation, Hand Gesture Recognition, RGB Datasets, Transfer learning, YOLOv8m

Application of Adaptive Catmull-Rom Spline Activation Function in Feedforward Neural Networks for Forecasting Noisy Time Series Data

Lakshmi M. P. D. S.[1*] and Chandrasekara N. V.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[1]mpdslakshmi@gmail.com, [2]nvchandrasekara@kln.ac.lk

[1]0009-0004-3563-1897, [2]0000-0003-3232-837X

Artificial Neural Networks (ANNs) have been applied to the real-world time series data, which often contains noise and nonlinear patterns. The linear techniques like exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) will not yield accurate outcomes for nonlinear noisy data. Although there are many other nonlinear approaches, they all require model assumptions. ANNs can be used to model complex, nonlinear patterns without being limited by linearity or predetermined assumptions. Hence, time series analysis can produce more accurate solutions by utilizing artificial neural networks (ANN). However, existence of noise can lead to inconsistencies even with the robustness of ANN. In order to improve performance of ANN in forecasting noisy, nonlinear time series data, this study presents an adaptive spline activation function (ASAF) based on the adaptive Catmull-Rom cubic spline. The objective of the study is to improve the performance of forecasting noisy time series data and evaluate the results by comparing with inbuilt activation functions that were already in the literature using simulated ARIMA based models. The primary concept is to implement the activation function using adaptive Catmull-Rom cubic spline in the output layer using feedforward neural networks. Hundred (100) non-linear time series under each of the models AR(1), MA(1), ARMA(1,1), ARIMA(1,1,1), ARIMA(1,1,0), ARIMA(0,1,1), AR(2) and MA(2) were simulated with thousand (1000) observations as an input for the NN to evaluate the performance of activation functions. A zero-mean noise with and finite variance were added to the simulated time series data and the gamma test was used to quantify finite variance. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were computed using testing data to measure the performance of models. The results of the study revealed that the linear activation function in hidden layer and Catmull-Rom activation function in output layer produced the lowest RMSE and MAE values among all models. The RMSE improvement percentages of 4.06%, 0.15%, 1.73%, 0.60%, 5.32%, 7.22%, 0.05% and 0.19%, along with MAE improvement percentages of 2.75%, 0.46%, 1.85%, 7.98%, 5.10%, 11.18%, 0.08% and 0.87% indicate that Linear-Catmull combination reduces overall prediction error when compared to the baseline Sigmoid-Linear combination across the above mentioned simulated time series models respectively. The findings of the study suggest that the use of linear function with Catmull-Rom activation function will increase the accuracy of forecasting noisy time series data using feedforward neural networks. The results of this study provide a foundation for applying the ASAF to a variety of neural networks, offering improved accurate forecasting in ARIMA models and facilitating the analysis of real-world datasets in related fields.

Keywords: Activation function, Catmull-Rom cubic spline, Neural Network, Noisy time series

Leveraging Generative Adversarial Networks to Improve LSTM and GRU Models Performances for Stock Price Prediction

Maduranga G. D. L.[1*] and Dharmarathne H. A. S. G.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]lasithamaduranga99@gmail.com, [2]sameera@stat.cmb.ac.lk

[1]0009-0000-4504-5921, [2]0000-0001-8544-0379

Accurate stock price prediction is essential for investment decisions, portfolio optimization and risk management. Deep learning models, particularly Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are widely used to achieve this. Generative Adversarial Networks (GANs), a unique deep learning technique involving a generator and discriminator, have demonstrated significant potential for modeling complex distributions in other areas. Building upon this success, this research investigates leveraging GANs to potentially enhance LSTM and GRU performance specifically for one-step-ahead prediction of the closing stock price of LOLC Holdings, listed on the Colombo Stock Exchange (CSE). Importantly, the study incorporates features such as the All-Share Price Index, Diversified Financial index, foreign market indices and economic indicators. The dataset includes ten years of daily historical data from June 2010 to March 2020, containing 2588 observations and 29 features. The dataset was subsequently divided into a training set and a test set using an 80/20 split. The training set encompasses the period from June 1st, 2010, to April 3rd, 2018 (2070 observations), while the test set includes data from April 4th, 2018, to March 13th, 2020 (518 observations). Two GAN models were implemented, each with a GRU/LSTM as a generator and a Convolutional Neural Network (CNN) as the discriminator. Findings from this study revealed significant improvements in Test Root Mean Squared Error (RMSE) results. The LSTM-GAN model demonstrated a nearly 14% improvement over the baseline LSTM model (LSTM RMSE: 3.79, LSTM-GAN RMSE: 3.27). Similarly, the GRU-GAN model exhibited an almost 8% increase in performance compared to the baseline GRU model (GRU RMSE: 3.6, GRU-GAN RMSE: 3.32). Furthermore, GAN-based models exhibited improved generalization and robustness, particularly when encountering unforeseen market shifts like the COVID-19 pandemic, highlighting the potential of GANs as a valuable tool for stock price prediction. This improvement is likely due to the adversarial process in GANs, where the generator and discriminator compete against each other, forcing the generator (in this case, an LSTM or GRU) to become better at modeling realistic sequences. This continuous adversarial training helps refine the LSTM/GRU's ability to capture subtle and complex temporal dependencies in stock price data. Finally, by focusing on LOLC Holdings on the Colombo Stock Exchange, this study contributes to the growing body of research in emerging markets, which often behave differently from more developed markets, where the majority of previous studies have been conducted.

Keywords: Deep learning, Generative adversarial networks, Stock price prediction

LSTM Based Unidirectional Sinhala Sign Language Translator for Enhanced Healthcare Accessibility of Deaf and Hard-of-hearing

Rajamanthri A. R. P. M.[1*], Senaweera O.[2] and Deshani K. A. D.[3]

[1,2,3]Department of Statistics, University of Colombo, Sri Lanka

[1]pramudimadhurya@gmail.com, [2]oshada@stat.cmb.ac.lk, [3]deshani@stat.cmb.ac.lk

[1]0009-0001-1912-5044, [2]0000-0001-8523-7012, [3]0000-0002-5489-3436

The global deaf and hard-of-hearing community faces significant communication challenges in the matter of conversing with the speaking community. Thus, proper translation is essential. Current translation methods, relying on human professionals, are impractical due to limited availability and high costs, especially in healthcare settings. To address this gap, this research focuses on integrating deep learning for Sinhala sign language (SSL) translation, targeting communication in the healthcare sector. The study begins with a comprehensive exploration of global sign languages, emphasizing the shortfall of research on SSL. The primary objective of this research is to create a model capable of translating SSL phrases used in healthcare interactions into Sinhala text. In order to achieve this, following a thorough analysis, a novel text dataset of 400 Sinhala phrases was created by collecting well-designed commonly used phrases by patients. Subsequently, their corresponding SSL interpretations were recorded using a mobile phone camera under the optimum camera setting. To process the recorded videos, the 'MediaPipe' framework was incorporated due to its lightweight nature, high processing speed and optimized performance. Using the 'MediaPipe' framework, numerical key points of human pose and hand landmarks were extracted from the collected SSL videos. Simultaneously, the text dataset was preprocessed, creating a medical vocabulary. Afterward, the extracted numerical key points were converted into sequences of words using deep learning models. The model fitting phase involved training several Long Short Term Memory (LSTM) models, including a Stacked LSTM model, Encoder-Decoder model with LSTM layers and Encoder-Decoder model with LSTM layers and Attention. The LSTM based models were employed due to their high performance in processing sequential data and their promising nature demonstrated in past global research. To ensure the model's performance and linguistic accuracy, the evaluation metrics Accuracy, Loss, Bilingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores were utilized. Evaluation metrics revealed that the encoder-decoder models, particularly the one that incorporates LSTM and attention mechanism, outperformed other models. The selected model achieved training and testing METEOR scores of 0.2410 and 0.2349, respectively. Considering factors such as the size of the dataset, model complexity, computational advancement and BLUE and METEOR scores of existing global research on sign language translation, the obtained results and the model's performance of the best model of this study was found to be satisfactory. By increasing the size of the dataset and using such a large dataset to train more advanced and sophisticated models, the obtained BLUE and METEOR scores can be further improved. Summarizing the study, the created deep learning pipeline with the best model managed to achieve the capability of converting Sinhala sign language phrases in video format into Sinhala phrases in text format with adequate accuracy. The research provides a viable alternative to human mediators, addressing a critical need in healthcare settings and potentially serving as a foundation for future advancements in sign language translation research.

Keywords: LSTM, Neural networks, Sign language translation, Sinhala sign language

Deep Learning-based Mobile Network Coverage Prediction

Ranasinghearachchi D.[1*] and Senanayake N.[2]

[1]Dialog Axiata PLC, Sri Lanka, [2]Informatics Institute of Technology, Sri Lanka

[1]dumindu.ranasinghearachchi@dialog.lk, [2]nipuna.s@iit.ac.lk

[1]0009-0009-4060-5980, [2]0000-0003-0233-4461

Mobile network coverage prediction is crucial for optimizing the placement of network towers, especially in dynamic urban environments. Traditional coverage planning tools rely on free space propagation models, which lack a feedback loop to incorporate real-world data. This study introduces a novel approach that leverages crowdsourced Minimization of Drive Test (MDT) data from a major mobile network operator in Sri Lanka as the target dataset. Public geospatial data from sources like Google Open Buildings, OpenStreetMap and NASA Shuttle Radar Topography were used as input data for various deep learning models, including Multi-Input Artificial Neural Network (ANN), Siamese and U-Net models. Previous research and models primarily focused on omnidirectional predictions. In contrast, this research considers directional beamforming into consideration using a propagation boundary, which resulted in predictions that are more closely aligned with real-world scenarios. Among the developed models, the Multi-Input and Siamese models demonstrated high accuracy in predicting Reference Signal Received Power (RSRP) levels. However, to address the issue of unverified predictions in areas lacking MDT data, a filtering mechanism using the U-Net model was introduced. This approach identified regions with the highest confidence in prediction, significantly improving the model's reliability. Further, experimental results show that after applying the U-Net filter, the Mean Absolute Error (MAE) of the predictions dropped from 4.255 to 1.452 for the Multi-Input model and from 4.354 to 1.462 for the Siamese model, indicating a substantial improvement. The findings suggest that integrating real-world data into deep learning models, taking directional beamforming into consideration, can significantly enhance the accuracy of network coverage predictions, offering mobile operators a powerful tool for more efficient network deployment.

Keywords: ANN, MDT, RSRP, Siamese Model, U-Net

A Machine Learning Algorithm to Enhance the Performance of Selected Technical Indicators: Evidence from Forex and Cryptocurrency Markets

Sandeepanee H. P. S.[1*] and Pathberiya H. A.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]sumedhasandeepanee@gmail.com, [2]hasanthi@sjp.ac.lk

[1]0009-0006-9095-4212, [2]0000-0002-0997-5897

Successful trading and investing in financial markets strongly rely on accurate prediction of price movement trends in the securities by employing a thorough market evaluation and a reliable decision-making procedure without any subjective judgment. In recent years, the analysis of foreign exchange (forex) and cryptocurrency markets has been recognized within the financial world for maximum profits regardless of the highly volatility nature. With the development of market assessment approaches due to the increased liquidity within those markets, technical analysis through indicators has shown reliable predictions of currency price fluctuating directions by incorporating price and volume data. However, despite their popularity, there is still a debate over the productivity of using technical indicators to predict trading signals. Aiming at this concern, this study provides an overview of the performance of the selected momentum indicators on forex and cryptocurrency markets while proposing a method to enhance their performance through the integration of machine learning with technical indicators. Within the initial phase of the study, the performance of the three most popular momentum indicators; Moment Average Convergence Divergence (MACD), Relative Strength Index (RSI) and Stochastic Oscillator within the forex and cryptocurrency markets are investigated by incorporating a preliminary analysis which compares the trading signals from existing rules with proposed rules, providing the accuracy rates to determine the correct trading signals for each indicator. Then the subsequent phase of the study explored the effectiveness of machine learning algorithms, including Long Short Term Memory (LSTM), Temporal Convolution Network (TCN), Random Forest Regression and Support Vector Regression (SVR). They were examined using historical trading records of two forex rates: EUR/USD, USD/JPY and two cryptocurrencies: BTC/USD, ETH/USD as well as technical indicator calculations under the lookback period of 5 trading hours as the predictors. The goal is to predict the one-hour ahead closing price and closing price movement direction of the currencies, alongside generating trading signals. The optimal hyperparameters are obtained by implementing a tuning process which integrates the Bayesian optimization through the Optuna technique with time series cross-validation. The accuracy rates for predicting trading signals are determined by comparing forecasted signals with actual signals based on proposed rules focusing on trend reversals. The study results illustrate that the existing momentum indicator rules only produce 68% - 74% accurate signals. Random Forest Regression and SVR models predict the currency prices with a lower loss whereas TCN and LSTM models predict the trading signals with a higher accuracy. Overall, the proposed algorithms can predict trading signals which exceed the 80% accuracy rate. The moderately accurate single-indicator results in the preliminary analysis triggered the development of machine learning models that combine multiple indicators together. This emphasizes the importance of integration of several indicators with the historical price data, proposed algorithms and trading rules for more informed and reliable decisions.

Keywords: Cryptocurrency market, Forex market, Machine learning algorithms, Technical indicators

Feature Selection-based Rainfall Classification: A Machine Learning Approach to Forecast Rainfall Occurrences at Ratnapura

Saubhagya K. S.[1*], Tilakaratne C. D.[2], Mammadov M. A.[3] and Lakraj G. P.[4]

[1,2,4]Department of Statistics, University of Colombo, Sri Lanka, [3]School of Information Technology, Geelong Warun Ponds Campus, Deakin University, Australia

[1]shanthi@stu.cmb.ac.lk, [2]cdt@stat.cmb.ac.lk, [3]musa.mammadov@deakin.edu.au, [4]pemantha@stat.cmb.ac.lk

[1]0000-0002-9239-8022, [2]0000-0003-0330-845X, [3]0000-0002-2600-3379, [4]0000-0003-3921-8552

Early prediction of occurrences of rain is imperative as it drives important decisions made in day-to-day life and even in many fields such as transportation, agriculture and supply chain management. Out of rainfall and no rainfall events, obtaining accurate predictions on rainfall events are given prime attention since its influence in bringing forth disastrous consequences. Ratnapura area, situated in wet zone in Sri Lanka, is identified as a highly vulnerable area to frequent flood events. Therefore, the accurate forecasting of rainfall events in Ratnapura is taken as the key priority of this study. Daily values of meteorological variables from 1st January 2015 to 31st December 2019 were used. Two rainfall classes (0 – No or Minor Rainfall (≤12.5mm), 1 - Moderate or Extreme Rainfall (>12.5mm)) were determined and considered for further analysis. Twenty-four weather and meteorological variables were identified from the literature as predictors. Since our preliminary study suggested that the past three lags of these variables were important for rainfall occurrence of the next day, finally, 72 (24x3) potential predictor variables were considered. Thereafter, prior to modeling rainfall occurrences, to obtain a fair decision, the feature ranking was conducted by building 25 Multiple Classification Trees which were trained to solve the binary classification problem. Training of these 25 trees was done through 3 methods: firstly, using the whole data set to train the trees, secondly, with randomly selected 25 balanced samples of size 50 and finally, with randomly selected 25 independent imbalanced samples of size 50. The final feature ranking was obtained by averaging the feature scores produced by these 3 methods. Based on the final feature ranking, rainfall occurrence was then modeled using the Random Forest method. Through the backward elimination approach (i.e. starting with the model with all variables and iteratively removing the least significant predictor at a time, evaluating the model performance until an optimal set of predictors is achieved), 27 models (with different subsets of predictor variables) which increased overall model performance and class-wise performance were extracted. The sizes of class 1 (24%) and class 0 (76%) show a significant class imbalance in both training and test data. Therefore, Synthetic Minority Oversampling Technique-Edited Nearest Neighbours (SMOTE-ENN) was employed prior to performing binary classification. This class imbalance appeared in extracted 27 models was further addressed by filling the minority class using actual data records taken out of the study period. Based on the overall and class-wise performances on test data (1st January to 31st December 2019), the model with the subset of 57 predictor variables which showed optimal performance and considerably low information loss was chosen for forecasting one-day ahead rainfall occurrence at Ratnapura. The results for the optimal model showed overall training Accuracy, Precision, Recall and F1-Score of over 0.89. The test set showed an overall model Accuracy of 0.71, Precision of 0.77, Recall and F1-Score of 0.71. Most importantly, the minority rainfall class, which represents moderate and extreme rainfall, generated a class-wise Accuracy of 0.88, while this value for the other class is 0.60.

Keywords: Binary rainfall classification, Class imbalance, Feature selection, Machine learning, SMOTE-ENN

Computational Finance Applications of Combined Estimating Functions

Saumyamala M. G. A.[1*], Thavaneswaran A.[2] and Thulasiram R. K.[3]

[1,2]Department of Statistics, University of Manitoba, Canada, [3]Department of Computer Science, Manitoba, Canada

[1]moragama@myumanitoba.ca, [2]Aerambamoorthy.Thavaneswaran@umanitoba.ca, [3]tulsi.thulasiram@umanitoba.ca

[1]0000-0001-6571-6868, [2]0009-0007-7611-5081, [3]0000-0002-6519-3929

Portfolio optimization and risk forecasting are two major areas discussed in computational finance. Recently, there has been a growing interest in risk forecasting and model risk forecasting, including volatility, value-at-risk (VaR) and expected shortfall (ES). VaR and ES are functions of volatility; therefore, volatility forecasting plays a crucial role in making better decisions and avoiding major losses. In portfolio optimization, it is shown that portfolio returns have significant skewness and kurtosis. However, the risk measures involved in commonly used minimum variance and maximum Sharpe ratio portfolios do not include skewness or kurtosis. The theorem of combining different estimating functions based on portfolio returns is given and it is shown that the variance of the combined estimating function (which depends on skewness and kurtosis of portfolio returns) is smaller than the variance of the components. Traditionally, in minimum variance portfolio optimization, portfolio variance is minimized, but in this study, the variance of the combined estimating function based on portfolio returns is minimized. Sign correlation is defined as the correlation between the least squares estimating function and the least absolute deviation estimating function, while volatility correlation is defined as the correlation between the estimating functions for variance and volatility. Sign correlation enables us to identify the underlying distribution and obtain probability forecasts. In this paper, volatility correlation is used to define volatility networks. Three applications discussed in this study are: first, using 30 asset returns, fuzzy volatility networks based on four different correlations (empirical, data-driven, partial, volatility) are introduced and portfolios are formed by combining community detection and page ranking methods. The cumulative returns of portfolios based on volatility correlation suggest the superiority of the proposed approach over commonly used diversification and portfolio optimization methods. Secondly, using cryptocurrency returns during pre and post COVID periods, generalized VaR, ES and model risk forecasting by applying regularization techniques are studied. Finally, recently proposed neuro volatility forecasting models and other forecasting models such as Prophet and long short term memory (LSTM) models with model risk are studied in detail. Portfolio optimization models were validated using an 80:20 train-test split, while time series forecasting models were validated using time series cross-validation methods. Unlike the existing work, the novelty of this study is suggesting a new portfolio diversification strategy by combining fuzzy correlation networks, network-based community detection and Page ranking methods and applying recently proposed risk measures in the context of portfolio optimization.

Keywords: Combined estimating functions, Financial networks, Neuro volatility models, Portfolio optimization, Volatility forecasting

Advancements in Cross-modality Imaging: A Survey of CNN and GAN Applications in Medical Imaging

Thahir K.[1*] and Meyen N.[2]

[1]Department of Physics, University of Sri Jayewardenepura, Sri Lanka, [2]Cogniata (Pvt) Ltd., Sri Lanka

[1]khadeeja20184@gmail.com, [2]nmeyen@gmail.com

[1]0009-0003-3900-1684, [2]0009-0009-8653-6300

This paper presents a comprehensive survey of recent advancements in cross-modality imaging techniques. It focuses on the application of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) in medical imaging. Cross-modality imaging involves translating images from one modality to another such as MRI to CT scans and can assist medical experts in accurate diagnostics and patient care. CNNs and GANs have been quite popular in various fields and their abilities to work with images and generate high quality outputs are quite an advantage for a situation like this. Our survey highlights several studies utilizing CNN methodologies for various applications, including synthetic MRI generation, MRI to CT translations for different anatomical regions and bone structure identification. GANs, on the other hand, excel in generative modeling by training two neural networks: the generator and the discriminator in a competitive setting. Their application in generating missing MRI modalities, creating 3D medical images and producing synthetic CT scans for radiotherapy are discussed. The integration of CNN and GAN models in cross-modality imaging shows significant promise in improving image quality, reducing noise and enhancing the accuracy of diagnostic tools. This survey underscores the importance of advanced deep learning techniques in medical imaging and sets the stage for future research in this rapidly evolving field.

Keywords: Image generation; Image synthesis, Machine learning, Medical imaging, Synthetic CT/MRI

Deep Q-learning Based Adaptive Traffic Light Control System using Google Traffic Data for Urban Traffic Optimization in Sri Lanka

Wickramasinghe U. S.[1*] and Lakraj G. P.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]sanjaniwickramasinghe@gmail.com, [2]pemantha@stat.cmb.ac.lk

[1]0009-0002-6615-3387, [2]0000-0003-3921-8552

Urban areas globally face persistent traffic congestion, leading to wasted time, increased fuel consumption and environmental damage. Common strategies to address this issue include traffic signal control, road pricing and intelligent transportation systems. Traffic signal control is essential for ensuring smooth vehicle movement at intersections and Adaptive Traffic Signal Control Systems (ATSCS) utilize real-time data to dynamically adjust signals based on current conditions, significantly improving traffic flow and reducing congestion. Traffic congestion in Sri Lanka (SL) is worsened by rapid urban growth, outdated signal control systems and the lack of scalable, cost-effective traffic management solutions. Current fixed-time signal systems in SL cannot adapt to fluctuating traffic conditions. Existing studies done in the Sri Lankan context on ATSCS mainly rely on CCTV data, which is costly to deploy and maintain at scale. For a developing nation, the logistical and financial challenges of installing cameras at every junction are significant, making a comprehensive CCTV network prohibitively expensive. While CCTV footage offers detailed insights into traffic patterns, its practicality for widespread traffic management in SL is limited. This study explored the use of Google Maps data in ATSCS, enabling real-time monitoring of congestion levels across a wider area for more accurate traffic flow assessments. This approach is also highly cost-effective compared to installing and maintaining CCTV cameras. This research developed a reinforcement learning (RL) algorithm, specifically deep Q-learning, leveraging real-time Google traffic data parameters to enhance traffic signal control in dynamic urban settings within SL. The algorithm was trained to adapt signal control based on lane speed averages, a state representation chosen for its ability to capture traffic flow effectively, aiming to minimize vehicle waiting times and reduce queue lengths. The model was developed to operate with eight signal phases, as four phases were too simple to handle complex traffic patterns, while six phases did not yield significant improvements. Given safety constraints preventing real-world implementation, the study adopted a simulation framework utilizing the Simulation of Urban MObility (SUMO) platform to develop and assess the RL-based traffic signal controller. Focused on Gamsaba Junction in Colombo for simulation, the study's findings underscore the efficacy of the proposed approach, demonstrating notable enhancements in traffic flow metrics over traditional fixed-time signal control systems. Notably, the RL-based traffic signal controller achieved a roughly 35% reduction in total vehicle waiting time, an 11% increase in mean vehicle speeds and a 15% decrease in CO_2 emissions compared to the fixed-time method. In essence, this research contributes to advancing traffic engineering practices by showcasing the effectiveness of RL algorithms and Google traffic data in mitigating traffic congestion challenges in resource-constrained settings like Sri Lanka. While the proposed approach shows significant improvements, limitations exist. The SUMO had constraints, including simplified traffic dynamics and a lack of unpredictable factors like pedestrian movement, impacting real-world applicability. Additionally, focusing on a single junction limits the generalizability of the findings.

Keywords: Adaptive traffic signal controller, Deep Q-learning, Google traffic data, Reinforcement learning, SUMO

Attention Mechanism Based Deep Learning Approaches in Forecasting the Colombo Consumer Price Index (CCPI): A Comparative Study

Wijamunige C. E.[1*], Deshani K. A. D.[2] and Arachchige C. N. P. G.[3]

[1,2,3]Department of Statistics, University of Colombo, Sri Lanka

[1]chalani.wijamunige99@gmail.com, [2]deshani@stat.cmb.ac.lk, [3]chandi@stat.cmb.ac.lk

[1]0009-0001-4250-9900, [2]0000-0002-5489-3436, [3]0000-0002-9084-0905

The Consumer Price Index (CPI) measures changes in the prices of goods and services purchased by households and is widely regarded as a proxy for inflation. This paper focuses on forecasting the Colombo Consumer Price Index (CCPI) in Sri Lanka, highlighting its importance for monetary policy planning. Designed to reflect price fluctuations for households in urban areas of the Colombo district, the CCPI provides essential insights into market trends. However, classical time series forecasting methods utilize conventional techniques, which may not fully capture the complexities of consumer price inflation dynamics. Therefore, this study explores advanced methodologies, particularly deep learning approaches incorporating attention mechanisms, to enhance the accuracy of CCPI forecasts. These techniques provide a more nuanced understanding of temporal dependencies, especially in complex economic conditions, such as the significant economic shifts caused by the COVID-19 pandemic, underscoring the need for a deeper exploration of CCPI dynamics in this evolving context. Moreover, the study employs a range of forecasting techniques, from classical machine learning approaches to advanced Sequence-to-Sequence (Seq2Seq) models. Support Vector Regression (SVR) with a polynomial kernel was effective in capturing short-term CCPI fluctuations. However, Long Short Term Memory (LSTM) models outperformed SVRs, especially showing significant improvements, for both one-step and six-step ahead forecasts, despite the absence of attention mechanisms. Models augmented with attention layers marked superior performances compared to those without attention while the Seq2Seq models incorporating attention mechanisms, offered only marginal gains in forecasting accuracy. Notably, the model comprising a single unidirectional LSTM layer with self-attention mechanism achieved a Root Mean Squared Error (RMSE) of 2.7026 and a Mean Absolute Percentage Error (MAPE) of 2.209 for one-step ahead forecasting and a RMSE of 2.919 and a MAPE of 1.996 for six-step ahead forecasting, proving its potential for accurate and interpretable forecasting, despite their complexity. Overall, the attention mechanism proved effective by prioritizing recent time steps in forecasting, distinguishing these models from classical approaches, which assign equal weight to all time steps and may overlook recent fluctuations.

Keywords: Attention mechanism, Colombo Consumer Price Index, Forecasting

Statistical Analysis of Episiotomy Practice and Associated Factors in Sammanthurai Base Hospital

Dharmasena D. M. P .G. I. S. S[1], Alibhutto M. C.[2], Azaath M. H. M[3] and Perera A. S. A.[4*]

[1,2,4]Department of Mathematical Sciences, Southern Eastern University of Sri Lanka, Sri Lanka,
[3]District General Hospital, Mannar, Sri Lanka

[1]isuri.dharmasena10@gmail.com, [2]mcabuhtto@seu.ac.lk, [3]azaath18@gmail.com,
[4]anupamapereraas@gmail.com

[1]0009-0006-2392-1269, [2]0000-0002-3926-6119, [3]0009-0004-8682-4453, [4]0009-0006-5839-1771

Episiotomy is the most common surgical procedure performed in obstetrics to enlarge the vaginal opening during childbirth. The aim of this study is to evaluate the prevalence of episiotomy and identify the factors associated with higher rates among women delivering at Sammanthurai Base Hospital, in line with World Health Organization (WHO) recommendations to minimize unnecessary episiotomies. A retrospective analysis was conducted on 139 obstetric records over a three-month period at Sammanthurai Base Hospital. Data were analyzed using chi-square tests and binary logistic regression to identify significant associations and predictors of episiotomy. The results indicate that the episiotomy rate was found to be 78%, significantly higher than WHO recommendations. The chi-square test revealed a significant association between episiotomy and factors such as parity (p-value<0.01), baby weight (p-value<0.05) and maternal Body Mass Index (BMI) (p-value<0.05). Binary logistic regression identified parity (OR=2.5, 95% CI=1.2-5.3), baby weight (OR=1.8, 95% CI=1.1-3.0) and maternal BMI (OR=1.7, 95% CI=1.1-2.9) as important predictors of episiotomy. The standard episiotomy rate recommended by WHO is 25%. However, the observed rate of 78% indicates that episiotomy is practiced routinely rather than selectively, contrary to WHO guidelines, which recommend that episiotomy should only be performed when medically necessary. This study provides valuable insight into the factors contributing to high rates of episiotomy in some hospitals, highlighting the need for changes in practice to align with global standards and improve maternal health outcomes. The findings offer crucial insights into the high episiotomy rates at Sammanthurai Base Hospital and underscore the importance of adopting targeted strategies to bring practices in line with global standards. Such efforts are vital to reduce maternal morbidity, including pain, infection and prolonged recovery times, associated with unnecessary episiotomies.

Keywords: Episiotomy, Healthcare interventions, Maternal factors, WHO guidelines

MetaHD: An R Package for Meta-analyzing High-dimensional Data

Liyanage J. C.[1*], Prendergast L.[2], Staudte R.[3] and De Livera A. M.[4]

[1,2,3,4]Mathematics and Statistics, School of Computing, Engineering and Mathematical Sciences, La Trobe University, Australia

[1]j.liyanage@latrobe.edu.au, [2]luke.prendergast@latrobe.edu.au, [3]R.Staudte@latrobe.edu.au, [4]a.delivera@latrobe.edu.au

[1]0009-0004-8276-8191, [2]0000-0002-9122-5429, [3]0000-0003-1281-7387, [4]0000-0003-4981-4155

Modern studies in evidence synthesis, such as those in large-scale biological datasets, have focused on combining results from studies that have measured multiple effect sizes associated with multiple correlated outcomes, necessitating multivariate approaches to meta-analysis. In this context, usually, there are many more variables of interest than there are studies and so the meta-analysis can be considered high-dimensional. Metabolomics was considered as an example of high-dimensional meta-analysis. Since summary measures on the same set of metabolites are often not available, resulting in a reduced number of studies for individual variable inference, multivariate meta-analysis techniques can be more useful. Meta-analysis methods that are currently available for combining such high-dimensional biological data, however, often overlook considerations such as the correlation between the metabolites, missing values and within- and between-study variability. These can lead to false identification of biomarkers or missing out on true biomarkers. The R package MetaHD performs a multivariate meta-analysis for high-dimensional data, which can be used to integrate and collectively analyze both individual-level data as well as summary statistics from multiple studies. This approach accounts for the correlation between metabolites, considers variability within and between studies, handles missing values and uses shrinkage estimation to allow for high-dimensional data analysis. This study exhibits that our approach leads to lower root mean squared error compared to existing methods used for metabolomics data and is particularly useful in the presence of missing data, as it exploits the borrowing strength between metabolites. MetaHD will serve as a valuable tool for integrating and collectively analyzing biological data generated from multiple independent studies, facilitating the accurate identification of biomarkers.

Keywords: High-dimensional data, Meta-analysis, Metabolomics, R package

A Novel User-friendly Weighted Hybrid Method (WHM) for Multiple Hypothesis Testing in Large Scale Genomic Data

Ouchithya R. A. S.[1*], Hettiarachchi N .N.[2], Dharmarathne H. A .S. G.[3] and Attygalle M. D. T.[4]

[1,3,4]Department of Statistics, University of Colombo, Sri Lanka, [2]Tokyo Metropolitan Institute of Medical Science, Tokyo, Japan

[1]susara.ouchithya@gmail.com, [2]nilminihett@gmail.com, [3]sameera@stat.cmb.ac.lk, [4]dilhari@stat.cmb.ac.lk

[1]0009-0006-3673-7022, [2]0000-0002-2314-9874, [3]0000-0001-8544-0379 , [4]0009-0006-2145-3975

Across many scientific disciplines, researchers conduct statistical tests in unison to explore diverse hypotheses. However, Multiple Hypothesis Testing (MHT) presents inherent challenges, primarily due to the heightened risk of producing false inferences. Despite numerous discussions and solutions for nearly a century, a comprehensive framework for MHT remains lacking. Even with flawless experiments and upstream analyses, incorrect inferences without proper multiple corrections can be detrimental, especially in fields related to genomics. This study addresses the challenges of MHT in genomic data by enhancing and consolidating existing methods into a robust framework. This approach is user-friendly, allowing even novices to understand their data and guiding them toward intuitive, nearly perfect inferences. A comparative simulation study was conducted to evaluate six methods commonly used for experimental genomic data: Bonferroni, Holm, Sequential Goodness of Fit (SGoF), Benjamini-Hochberg, Benjamini-Yekutieli and Storey's Q-value across different scenarios related to comparing two independent groups. Our results suggest that Storey's Q-value has higher statistical power compared to other methods in scenarios with large effect sizes, while the meta-test based SGoF demonstrates more power in cases with lower effects. Methods such as Bonferroni and Holm, with high precision and specificity, show reliability when strict control of false positives is required. This illustrates the impracticality of relying solely on one method. Therefore, an exploratory hybrid technique is presented to determine significant hypotheses under most methods by sequentially testing and removing hypotheses until a satisfactory result is obtained. Here, the researcher is given the choice to traverse between Family-Wise Error Rate (FWER) and False Discovery Rate (FDR) techniques, hence the name "Hybrid". Furthermore, an innovative "Significant Index Plot (SIP)" is introduced to aid in detecting significant hypotheses across the six correction procedures. SIP visually represents rejections with vertical lines for significant hypotheses and white spaces for the absence of a signal, thus allowing an intuitive understanding of how each hypothesis is rejected. The methodology was applied to an experimental dataset of 10,724 p-values for bona fide enhancer elements in the human genome obtained via Massively Parallel Reporter Assay (MPRA) and Self-Transcribing Active Regulatory Region sequencing (STARR-seq). Storey's Q-value identified 1,944 significant hypotheses for MPRA and 871 for STARR, while the Hypothesis Weighting method confirmed 1,902 significant hypotheses for MPRA and 839 for STARR, showcasing practical adaptability. Additionally, a novel Python library named "MultiDST" is introduced to facilitate the usage of the framework. MultiDST allows users to implement and compare the results of different correction methods using minimal lines of code. By establishing a robust framework and a user-friendly platform, this study eradicates the "blind" application of multiple correction techniques, empowering the scientific community to engage in more rigorous and reliable hypothesis testing within genomics and medical studies where correct inferences are crucial.

Keywords: Gene expression, Multiple hypothesis testing, Significant index plot

Integrating Machine Learning and Statistical Analysis for Enhanced Understanding of Wound Microbial Communities

Peiris T. H. K.[1*], Jayatillake R. V.[2], Jayathilaka N.[3] and Garrison J. A.[4]

[1,2]Department of Statistics, University of Colombo, Sri Lanka, [3]Department of Chemistry, University of Kelaniya, Sri Lanka, [4]Allied Wound Consultants

[1]heshikavindya99@gmail.com, [2]rasika@stat.cmb.ac.lk, [3]njayathi@kln.ac.lk, [4]do15.john.garrison@nv.touro.edu

[1]0009-0003-8996-1309, [2]0000-0002-5414-999X, [3]0000-0001-8741-3075, [4]0009-0006-5307-2456

Wound infections remain a significant challenge in healthcare, often leading to prolonged healing times, increased morbidity and elevated healthcare costs. Understanding microbial communities in various wound types is essential for targeted treatments. This study was conducted with the aim to identify microorganisms associated with specific wound types and examine microbial co-occurrence patterns in the wound microbiome. The dataset of the study consists of microbial DNA copy numbers for 22 reported microorganisms related to 234 wound samples along with patient level information such as age, gender and location of the wound. Ethical clearance for the study was not required since no personal identification information was retrieved for the study. Data was collected and compiled by a wound care center in USA, from January 2023 to February 2024. Wound types in the study, included non-pressure chronic ulcers, wounds associated with Type 2 diabetes mellitus, pressure ulcers and surgical wounds. Preliminary analysis revealed Corynebacterium spp. as the most prevalent microbe in all wound types, with the highest prevalence in Type 2 diabetes mellitus wounds (59.09%). Staphylococcus aureus was also frequently detected, particularly in non-pressure chronic ulcers (41.53%). Enterococcus spp. and Peptostreptococcus spp. were common in Type 2 diabetes mellitus and surgical wounds (31.82% and 33.90%, respectively). These findings underscore the diverse microbial profiles associated with various wound types. Machine learning models, including multinomial logistic lasso regression, XGBoost, random forest and SVM, were used to classify wound types based on microorganism copy numbers and patient demographics. Model fitting prioritized both explainability and performance. The multinomial logistic lasso regression model, using data resampled with the SMOTE technique, achieved the highest accuracy (0.6761) and F1-score (0.5932). This model revealed microorganisms that are highly associated (p-value < 0.05) with each wound type compared to the baseline Type 2 diabetes mellitus wounds. It was found that pressure ulcers are significantly associated with Corynebacterium spp., Escherichia coli and Pseudomonas aeruginosa; non-pressure chronic ulcers are linked to Acinetobacter baumannii, Enterobacter spp., Escherichia coli, Pseudomonas aeruginosa, Trichophyton spp., Staphylococcus aureus and Staphylococcus spp.; and surgical wounds are primarily associated with Bacteroides spp. Further analysis using the Apriori algorithm (support = 0.05, confidence = 0.7) revealed distinct microbial associations in each wound type. For instance, when both Staphylococcus aureus and Staphylococcus spp. are present in Type 2 diabetes mellitus wounds, Corynebacterium spp. also tends to appear, with a relatively high confidence of 90%. These findings highlight the importance of considering wound etiology when improving wound care outcomes. However, future research needs to address limitations such as the lack of detailed wound characteristics, patient comorbidities and prior antibiotic use to enhance the understanding of microbial dynamics in wound healing.

Keywords: Association rule mining, Predictive modeling, Wound infections

Statistical Analysis on Length of Stay for Pediatric Hospitalizations: A Case Study

Weerakoon W. M. A. S.[1*], Gunawardhana G. M. M. S.[2] and Abeysundara S. P.[3]

[1,2,3]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

[1]weerakoonsandunika@gmail.com, [2]malkisuhadani@gmail.com, [3]sachitha@sci.pdn.ac.lk

[1]0009-0000-5047-9759, [2]0009-0008-8345-9648, [3]0000-0002-1228-4771

Children are the leaders and the future of our country. Therefore, it is a duty to raise every child with safety and provide healthy individuals for the future. In this process, understanding child diseases and injuries is very important. Literature indicated that limited studies have been conducted in Sri Lanka on child diseases and injuries with respect to different types of diseases and injuries. The main objective of our study is to analyze and identify the patterns of all types of diseases and injuries across all the age groups and fit a statistical distribution for Length of Stay (LOS) of pediatric admissions. The dataset was obtained from Sirimavo Bandaranayake Specialized Children's Hospital Peradeniya, Sri Lanka for the period 2019-2023. It contains variables such as Date of admission and discharge, Age, Gender, Mode of discharge, ICD code and injury related data. Preliminary Analysis revealed that the average stay per admission is 1.52 days with the standard deviation of 1.92 days for injuries while average stay per admission is 3.06 days with the standard deviation of 4.37 days for diseases. The maximum length of stay due to injuries was 62 days and due to diseases was 346 days during this period. Time series plot indicates that during the COVID-19 period number of child admissions was decreased. In child admissions due to injuries, the head is the most injured body part and 79% of them were superficial injuries while upper limbs are the second most injured body part and more than half of them were fractures. The average LOS in hospital due to head injury (1.30 days) is significantly different from the average LOS due to upper limbs (1.53 days) injuries. The most common disease type among children admitted was respiratory system (14.9%). Other top disease categories were certain infections and parasitic diseases, factors influencing health status and digestive system problems. Mean comparison tests indicate that the LOS was significant with respect to the gender and the type of disease. Throughout the last 5 years, the number of injured children who have left against medical advice has decreased while the number of injured children who treated and sent home has increased. Three distinct distributions were identified for LOS, which is a heavily right skewed distribution. One corresponding to injuries, one corresponding to diseases in the respiratory system and to certain infections and parasitic diseases. The distribution obtained for LOS due to injury follows a 2-component Exponential Mixture Model. Both other types of diseases follow a Zero-truncated Negative Binomial model (p-value<0.01). Age is significant in determining LOS in hospital for child admission. The findings of this study lead a healthcare professional to look at specific diseases and injury types and associated variables which will lead developing more efficient methods for treatments and future research will be carried out while looking at distribution patterns of LOS among injuries and different types of diseases that lead to more accurate predictive models on LOS.

Keywords: Child disease, Child injury, Hospitalizations, Length of stay (LOS)

Assessing the Impact of Library Website Quality on User Satisfaction: A Structural Equation Modeling Approach at the University of Ruhuna

Britto K. K. S. N.[1*] and Jayasekara L. A. L. W.[2]

[1,2]Department of Mathematics, University of Ruhuna, Sri Lanka

[1]nadishanishalidi97@gmail.com, [2]leslie@maths.ruh.ac.lk

[1]0009-0005-0124-129X, [2]0009-0009-1795-6571

Academic libraries are essential in the modern technology environment as they provide a wide range of information and facilities through their websites, contributing to the achievement of educational goals. The quality of these digital platforms is growing in importance as they influence users' experiences and academic performance. This study aims to identify the factors that are connected with library website quality and develop a model that evaluates the effect of the quality of the library website (QoW) on user satisfaction for academic achievement (AA) by comparing service quality within the library (QoLP) and non-library website quality (QoNLW) at the University of Ruhuna using Structural Equation Modeling (SEM). For this purpose, a survey was carried out among 591 undergraduates at the University of Ruhuna through a structured questionnaire. This study revealed that information quality (QoI), usability quality (QoU) and service interaction quality (QoSI) are the three main factors influencing library website quality throughout factor analysis. The SEM model showed that the library website quality explained 79.9% of the variation in user satisfaction. The study found that QoSI, QoU and QoI significantly impacted user satisfaction. Subsequently, QoW and QoLP significantly influenced academic achievement, while QoNLW had a strong impact. These results can serve as a valuable reference for university library administrators to improve the services provided on the library website and ensure students' academic achievements by helping them develop and utilize necessary skills.

Keywords: Academic achievement, Library website quality, Structural equation modeling, User satisfaction

Digital Literacy and Academic Success: A Case Study of Undergraduates in the Faculty of Business at the University of Moratuwa, Sri Lanka

Karan S.[1*], Gunawardana A.[2] and de Silva T. S.[3]

[1,2,3]Department of Decision Sciences, University of Moratuwa, Sri Lanka

[1]karansubramaniam8@gmail.com, [2]asankag@uom.lk, [3]tilokad@uom.lk

[1]0009-0005-2074-5792, [2]0000-0003-3205-6589, [3]0000-0002-0543-2182

In the digital age, proficiency in using and understanding digital platforms has been essential for academic success. As technology becomes increasingly integrated into higher education, exploring the relationship between digital literacy and academic performance is both timely and necessary. This study aims to investigate the role of digital literacy in influencing academic performance among higher education students. Specifically, the study examines the correlation between digital literacy levels and academic performance, while also identifying potential disparities among students and suggesting strategies to address these differences. This study addresses a significant gap in the existing literature, which has produced conflicting conclusions regarding the relationship between digital literacy and academic achievement. A quantitative approach was employed, with data gathered through a structured survey and the reliability of the survey was confirmed through a pilot study. The research focused on students from the Faculty of Business at the University of Moratuwa, Sri Lanka, with a sample of 208 students selected using a stratified random sampling technique. This sample appropriately represents all academic years and departments within the faculty. To analyze the data, Kendall's Tau correlation, a proportional odds logistic regression model along with Kruskal-Wallis rank sum and Dunn's tests were used. The findings revealed a moderate positive linear relationship (tau = 0.47, p-value < 0.05) between digital literacy and academic performance, indicating that students with higher digital literacy tend to achieve better grades. Although digital literacy levels did not significantly differ among students in adjacent academic grade groups, notable differences were observed between more distant grade groups. This study contributes to higher education by highlighting the importance of digital literacy in enhancing academic performance, offering insights that could inform strategies to improve student outcomes and the educational sector as a whole. However, the present study's conclusions are based on a relatively medium sample size and are limited to a specific geographic area. Future research will aim to expand the sample size and include a more diverse population to validate and extend these findings.

Keywords: Academic performance, Digital literacy, Higher education, Quantitative approach

An Investigation on Unemployment Duration of Science Graduates of Faculty of Science, University of Ruhuna

Kulasekara S. R. T.[1*] and Jayasinghe C. L.[2]

[1]Department of Mathematics, University of Ruhuna, Sri Lanka, [2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]ruchirakulasekara@gmail.com, [2]chathuri@sjp.ac.lk

[1]0009-0007-1613-9946, [2]0000-0003-2891-742X

High unemployment rates have had an adverse effect for decades on Sri Lankan youth, especially those aged 15-29 with high educational qualifications, as reported by Sri Lankan labor force surveys over the years. This study investigates the duration of unemployment after graduation of graduates from the Faculty of Science, University of Ruhuna. Another aim was to identify the factors that are associated with the time to first job after graduation. An online questionnaire was shared among the 2018 graduates of the Faculty of Science, University of Ruhuna, to capture responses and the analysis was conducted based on the sample of graduates who responded, with a response rate of 47%. To fulfill the objectives, the data were analyzed using survival analysis techniques. Descriptive analysis was conducted and then the Kaplan-Meier estimator was used to estimate the survival function of unemployment duration, while the log-rank test evaluated the significance of differences in the survival function of unemployment duration between different groups of graduates. Semiparametric and parametric models, including Accelerated Failure Time (AFT) models, were fitted and the best model was selected using the Akaike Information Criterion (AIC). Among the four models considered, namely, Cox proportional hazards, Exponential regression model, Weibull regression model, lognormal regression model and log-logistic regression model, the Weibull model was selected as the best model. Wald test p-values indicated that being a player in a university team, holding positions in clubs or societies during university also significantly reduces the hazard rate. Using personal networks only for job search significantly increases unemployment duration; directly contacting employers was also associated with a significantly longer duration of unemployment. Having completed a longer internship program has also led to shorten the unemployment duration. According to conclusions made based on data relevant to science graduates of the University of Ruhuna, it can be recommended that shorter durations of unemployment can be achieved for science graduates by being/having engaged in extracurricular activities, promoting graduates to follow professional qualification programs that have been aligned with market demand to ensure graduates' skills are relevant and in demand, by educating students on effective job search strategies and by enhancing internship opportunities and industry collaborations.

Keywords: Employability of graduates, Science graduates, Survival analysis, Unemployment duration

Modeling Migration Intentions of Sri Lankan Youth Using Logistic Regression Analysis: Evidence from Colombo District Universities

Palihawadana P. H.[1*] and Dharmarathne H. A. S. G.[2]

[1]Department of Finance, University of Sri Jayewardenepura, Sri Lanka, [2]Department of Statistics, University of Colombo, Sri Lanka

[1]phansini@sjp.ac.lk, [2]sameera@stat.cmb.ac.lk

[1]0009-0000-0161-9081, [2]0000-0001-8544-0379

Against the backdrop of escalating international migration rates among Sri Lankan youth, this research aims to investigate the determinants of migration intentions among young adults in Sri Lanka. The study has opted for fresh graduates and final-year undergraduates in the Colombo district, as a representative sub-segment of the target variable, mainly due to ease of identifying and approaching the units and their significant impact on the long-term economic prosperity of the country. Utilizing a quantitative approach, data were collected through a structured questionnaire administered to a sample of 432 participants from universities in the Colombo district. Multistage cluster sampling method was used as the sampling technique to collect the primary data needed for the study. Each university in the Colombo district (both public and private) was recognized as a cluster and then the questionnaire was distributed among selected clusters (universities) in the sampling frame using convenient sampling and lastly, snowball sampling method was used to select sample units (final-year undergraduates and fresh graduates) from each cluster. Examining the migration intentions among Sri Lankan youth, the study has utilized a multi-level perspective, considering both micro (individual) level and meso (familial-social) level factors. This analytical framework was adopted based on a comprehensive review of existing literature. Accordingly, 'education and career' and 'preservation' (personal aspirations, psychical and psychological security) factors are used to elaborate the impact of micro-level factors on youth migration intentions; while 'family' related and 'social network' factors were identified under the meso-level. The analysis employed binary logistic regression (to examine the dichotomous dependent variable) and multinomial logistic regression (to examine ordinal scale dependent variable) models to examine the migration intentions of sample units. The findings reveal that a significant proportion (67%) of respondents exhibit strong intentions to migrate, driven primarily by educational and career aspirations, as well as concerns for physical and psychological security. These micro-level factors were consistently identified as key predictors across different models, aligning with established theories such as Human Capital and Rational Choice. Conversely, meso-level factors, particularly family ties, were found to play a crucial role in shaping migration intentions, highlighting the cultural significance of familial connections in the Sri Lankan context. Interestingly, the impact of social networks on migration intentions was inconsistent, challenging existing theories of Social Capital and Network. Furthermore, the study found no significant difference in migration intentions between undergraduates and graduates, suggesting that educational attainment alone does not fully account for the desire to migrate. These results underscore the complexity of migration decisions among Sri Lankan youth, influenced by a combination of personal aspirations and socio-cultural factors. The study concludes with recommendations for policy interventions that address both macroeconomic challenges and the individual circumstances driving youth migration. Such interventions are crucial to mitigate brain drain and support the long-term economic development of Sri Lanka.

Keywords: Brain-drain, Binary logistic regression, Determinants, Multinomial logistic regression, Youth migration

Investigating Factors Associated with Level II Undergraduates' Grade Point Average in the Faculty of Science, University of Ruhuna, Sri Lanka

Perera W. T. C.[1*] and Jayasekara L. A. L. W.[2]

[1,2]Department of Mathematics, University of Ruhuna, Sri Lanka

[1]thejanicperera@gmail.com, [2]leslie@maths.ruh.ac.lk

[1]0009-0001-9935-5818, [2]0009-0009-1795-6571

Academic performance at the university level is assessed through grades and the Grade Point Average (GPA). This study aims to investigate the factors associated with the GPA of the level II undergraduates of the Faculty of Science, University of Ruhuna. Specifically, it examines the association of Z-Score, Level I English results, Lecture attendance, Degree Program, Gender, Hostel residency and General Certificate of Education Advanced Level (G.C.E. A/L) attempt on the GPA of these students. The study contains secondary data of 619 Level II undergraduates. Descriptive and inferential statistical analyses using both bootstrap and non-parametric approaches were conducted. The regression models were constructed using the Ordinary Least Squares method and bootstrap regression methods. According to the results, a higher proportion of female students got admission to the university and their academic performance was better than that of male students. The performance of male students in the Physical Science Degree program was lower than the other groups. Students who were admitted to the university on their first attempt performed better compared to the others. The results of the non-parametric tests and Bootstrap approaches concluded that the Z-Score, Level I English results, Lecture attendance, degree program, gender and G.C.E. A/L attempts significantly affected the GPA while hostel residency did not affect the GPA. The paired bootstrap regression method yielded a model with better accuracy. These results will further assist students, lecturers, administrators and policymakers in observing and taking appropriate action to improve academic performance.

Keywords: Academic performance, Bootstrap, Grade point average, Non-parametric methods, Regression

\

Deciphering the Influence of Metals on Zebrafish Behavior through GAMM and WCCNA Network

Hewawasam D. M.[1*] and Withanage N.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]dinithimuthuwanthih@gmail.com, [2]niroshan@sjp.ac.lk

[1]0009-0004-7480-1457, [2]0000-0001-8905-3878

Consumption of well water has been linked to significant health issues in humans. Contaminated water directly causes severe diseases. "Recognizing the similar effects of drinking contaminated water on humans is crucial in the fight against chronic diseases and for protecting human health. Zebrafish, known as the high-throughput detectives of the aquatic world, serve as efficient investigators of chemical effects in water. A common challenge in zebrafish experiments is the nested data structure; linear analyses are not well-suited to detect the complexity of behavioral responses across chemicals over time, nor to capture the actual effect rather than just the significant effect. This study aims to determine the individual and joint contributions of metals by examining the non-linear behavior of zebrafish over time and to develop a network analysis method for identifying highly interconnected metal groups that influence zebrafish behavior. Zebrafish behavior was analyzed by measuring swimming distance within a 25 minute period, with dark and light conditions changing every 5 minutes, starting with dark. The three-level nested dataset from well waters in Maine and New Hampshire, USA, reported by Babich et al. (2021), was employed to obtain results. Zebrafish were nested within the sample and time points were nested within the zebrafish. Visualizations highlighted a sophisticated pattern over time in all 92 well water samples, each consisting of approximately 24 zebrafish, which included the concentration levels of 14 metals. The study was divided into two sections for analysis: one examines the effects of metals on zebrafish, assuming that observations are independent based on aggregated data for dark and light conditions separately and the other assesses the impact of metals on the temporal behavior of zebrafish with random effects. In contrast to the original research, this study employed a generalized additive model and weighted chemical co-expression network analysis (WCCNA) to detect complex patterns through independent observations and a generalized additive mixed model to capture sample effects using a random intercept and zebrafish effects using both a random intercept and random slope. The results show that significant factors, such as time, copper and lead, have the potential to exhibit non-linear effects on zebrafish behavior, while nickel, cadmium and barium had a significant linear impact on zebrafish behavior. Moreover, these metal mixtures are capable of altering zebrafish behavior. Using a mixture-based approach, three metal groups were identified through Weighted Correlation Network Analysis (WCCNA), which was employed to create a weighted correlation matrix, identify highly interconnected metals and examine the relationship between metal groups and zebrafish behavior. Notably, the nickel, copper and cadmium metals group demonstrated a higher negative relationship with zebrafish behavioral responses in dark than in light conditions. In summary, this study provides guidance on how to detect individual, joint potential drivers and mixture drivers of complex biological responses.

Keywords: Generalized additive mixed models, Generalized additive models, Nested, Significant effect, Weighted chemicals correlation network analysis

A Simulation Study to Examine the Performance of a Joint Piecewise Hazard Model and a Normal Model for Two Responses Survival and Transformation of Count

Jisam F. S.[1*] and Sooriyarachchi M. R.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]sarahjisam@gmail.com, [2]roshini@stat.cmb.ac.lk

[1]0009-0009-6586-4567, [2]0000-0002-4551-0616

The frequent association of survival and count responses in medical research provides remarkable attention in the field and methodological literature of joint modeling. This research is a simulation based study that is motivated by the joint piecewise hazard model and normal model for the correlated outcomes of survival times and transformation of counts with shared random effects. The primary objective of this study was to examine the properties of the above joint model namely, type I error and power for simple randomized data. It was identified that the type I error holds well for the above joint model only when the constant hazard assumption is met and high power is yielded with the increase in sample size. In addition, as a sub objective of this study, the above joint model was further illustrated on a lymphoma patient dataset and this showed superior results to the respective univariate models.

Keywords: Joint modeling, Joint piecewise hazard model, Random effects, Simulation

Statistical Analyses of Linear Regression and Contingency Tables for Differentially-private Matrix Masking Data

Nghiem L. H.[1*], Ding A. A.[2] and Wu S.[3]

[1]School of Mathematics and Statistics, University of Sydney, Australia, [2]Department of Mathematics, Northeastern University, United States, [3]Department of Biostatistics, University of Florida, United States

[1]linh.nghiem@sydney.edu.au, [2]samwu@biostat.ufl.edu, [3]a.ding@northwestern.edu

[1]0000-0003-2874-9067, [2]0000-0003-1397-2442, [3]0000-0003-2684-436X

Balancing privacy preservation with data utility poses a significant challenge in big data analytics. While there is a wealth of literature on designing privacy protection schemes, statistical inference methods for analyzing privacy-protected data remain relatively unexplored. This paper presents statistical methods tailored for analyzing datasets collected with a privacy protection scheme involving matrix masking and noise addition. The initial analysis focuses on the linear regression model, aiming to elucidate the connections and distinctions between analysis methods for data collected under this privacy protection scheme and those for conventional measurement error settings. Subsequently, an analysis was conducted on a 2×2 contingency table collected with this scheme, proposing valid statistical inference procedures for comparing the difference in two proportions. Our theoretical investigations unveil the trade-off between privacy protection and statistical precision. Finally, these methods were illustrated on a dataset concerning hypertension prevalence in the United States, comparing analysis of the original data with those from the privacy-protected release.

Keywords: Contingency table, Linear model, Measurement errors, Privacy

Predicting the Professional Quality of Life among Tertiary Healthcare Nurses in Major Teaching Hospitals of Colombo, Sri Lanka: A Cross-sectional Study

Perera K. S. P.[1*] and Jayamanne I. T.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]senuriperera553@gmail.com, [2]imali@stat.cmb.ac.lk

[1]0009-0009-4747-0185, [2]0009-0002-0314-5021

Tertiary healthcare typically involves highly specialized treatments, such as advanced diagnostic services and complex surgeries. Nurses in tertiary healthcare often deal with difficult patients and stressful situations, affecting their overall well-being. Studies indicate nurses experience the highest stress levels among healthcare professionals, putting them at high risk for a low Professional Quality of Life (ProQOL). Hence, understanding and addressing factors influencing their ProQOL is very crucial. ProQOL is a measure that describes how one's work as a helper affects one's well-being. It encompasses three distinct aspects: compassion satisfaction (CS), burnout (BO) and secondary traumatic stress (STS). The ProQOL-V Scale is a widely used tool for measuring the positive and negative effects of helping others who experience suffering and trauma. No prior studies specifically address ProQOL among Sri Lankan nurses. This study investigates the ProQOL among tertiary healthcare nurses across four teaching hospitals in Colombo, Sri Lanka: National Hospital of Sri Lanka (NHSL), Lady Ridgeway Hospital (LRH), De Soysa Maternity Hospital (DMH) and Apeksha Hospital. It aims to identify associated factors for ProQOL among tertiary healthcare nurses and predict ProQOL. The choice of the four largest and busiest hospitals in the Colombo district adds significant value to the research, considering the highly stressful work settings. From the initial population of 4558 nurses, a sample size of 368 nurses was selected for this study. Utilizing the ProQOL-V scale and surveys, data collection proceeded through a two-stage stratified sampling design, distinguishing hospitals as strata and main units within hospitals as sub-strata. Preliminary analysis revealed significant differences in ProQOL scores between hospitals and main units by getting the population estimates of score values for each strata combination. The mean ProQOL scores of Sri Lankan nurses in the target group are CS at 37.2 ± 4.3, BO at 36 ± 3.8 and STS at 35.5 ± 3.2 which highlights that tertiary healthcare nurses working in teaching hospitals experience moderate CS, high BO and high STS, which significantly decreases the ProQOL. Out of four hospitals LRH exhibits significantly a low ProQOL of nurses by displaying a high BO and STS among its nursing staff. In addition, inward care nurses are highly experiencing STS compared to other main units. Most nurses in this study are aged 25 to 35, indicating a decrease in senior staff. Although STS is positively associated with mental health, this finding suggests that good mental health alone does not guarantee a high ProQOL. This study pioneers the use of machine learning over traditional methods to improve prediction accuracy by identifying complex patterns. By fitting the Decision Tree Regressor, Random Forest Regressor, Extreme Gradient Boosting and Support Vector Regressor, the Random Forest Regressor demonstrated superior performance in predicting the scores of ProQOL among nurses by giving the Mean Absolute Percentage Error (MAPE) values for the scores of CS, BO and STS as 7.97, 9.21 and 7.30 respectively. Notably, patient-to-nurse ratio, Body Mass Index and age emerged as crucial variables in ProQOL prediction.

Keywords: Cross-sectional survey, Machine learning, Professional quality of life, Tertiary healthcare nurses, Two-stage stratified sampling

A Decision Support System for Sustainability Planning of an Agri-food Supply Chain

Cardoso M. G.[1*] and Lopes M. P.[2]

[1,2]ISEP - School of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, [1]Associate Laboratory for Energy, Transports and Aerospace (LAETA-INEGI), Rua Dr. Roberto Frias 400, 4200-465 Porto, Portugal, [2]INESC TEC, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

[1]mon@isep.ipp.pt, [2]mpl@isep.ipp.pt

[1]0000-0001-9920-4975, [2]0000-0002-1146-7226

The main objective of this work was to develop a decision support system for sustainability planning of a pig production supply chain, with particular attention paid to the negative impacts of noise and greenhouse gas emissions on local communities and protected wildlife areas. The proposed decision support system integrates a mathematical programming model with a discrete simulation model. The mathematical programming model optimizes the supply planning process using a weighted sustainability function of economic, environmental and social impacts, managed by the decision maker. The scenario approved by the decision maker feeds a discrete simulation model of the agri-food supply chain enabling dynamic analysis and assessment of the impact of variability. The results show that it is possible to have alternative scenarios with significant reductions in environmental and social impacts with marginal impacts on the economic dimension (4.7% improvement in social sustainability and a 2.4% improvement in environmental outcomes with only a 0.051% increase in economic costs) and highlight the importance of detailed transport planning for operational optimization.

Keywords: Decision support system, Green agri-food supply chain, Triple bottom line

A Case Study on Daily Production Efficiency Prediction in Apparel Manufacturing: Integrating Learning Curve Theory with Multifactorial Optimization Strategies

Madurangi M. A. A. U[1*] and Wickramarachchi D. C.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]ayomiupekkha@gmail.com, [2]chitraka@sjp.ac.lk

[1]0009-0007-1809-7372, [2]0000-0001-6958-9667

The apparel industry in Sri Lanka is renowned for its labor-intensive processes and skilled workforce. This research centers on modeling the learning process of operators within sewing lines to improve the accuracy of forecasting production efficiency and planning. The study considered twelve months of daily production data from a major apparel manufacturer, focusing on long-run apparel styles across multiple workstations. Average optimal sewing efficiencies were calculated, accounting for downtime hours to capture lost production hours on both a daily and workstation basis. Initially, one style was selected and tested with seven traditional manufacturing learning curve models: Wright's model, Stanford-B model, 2-parameter and 3-parameter exponential models and 2-parameter and 3-parameter hyperbolic models. Models' accuracy was evaluated using Mean Absolute Percentage Error (MAPE), Sum of Squared Error (SSE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, with the models exhibiting lower values being selected. The exponential models were rejected due to their tendency to reach early steady-state performance. Hyperbolic models were selected to further validate the process using five different styles. The results indicated that the 2-parameter hyperbolic model was best suited for this apparel industry. As a further enhancement, the model was developed to incorporate three style categories. The final selected model for daily efficiency is expressed as: daily efficiency$_i$ = [(0.48*day$_i$)/(0.42*day$_i$)] +0.17D$_1$ +0.04D$_2$, where i represents the day (1, 2, …) and D$_1$, D$_2$ are dummy variables representing style categories. This model yielded a lower Mean Squared Error (MSE) value of 0.006, indicating a good fit for the data. Nonlinear regression techniques with Levenberg–Marquardt optimization were used to build the model, which effectively captures operator learning and transition to steady-state performance. The study also addresses real-time manufacturing challenges, including machine, material downtime and quality defects, which negatively impact sewing efficiency. Specifically, a 10% increase in the previous day's efficiency results in a 0.82% efficiency boost on the current day Machine, material downtime and quality defects decrease efficiency by 0.63%, 0.51% and 0.67%, respectively, for each additional 10 minutes of downtime or each defect. Time dummy variables reveal daily efficiency fluctuations, with notable increases of 0.52% on Day 3 and 1.2% on Day 4 compared to Day 1, continuing similarly through Day 30. To analyze these daily efficiency influencers, a dynamic panel study using the system generalized method of moments was employed, offering a more robust approach than traditional methods. Overall, the findings underscore the importance of operator learning in production efficiency. Accurate daily efficiency predictions are vital for setting production targets. Identifying downtime and quality defect impacts helps mitigate risk factors, offering valuable guidance for decision makers to optimize processes and manage disruptions, thereby fostering industry growth and competitiveness.

Keywords: Dynamic panel regression, Learning curve, Manufacturing, Non-linear regression

Reliability Analysis of the Akash Distribution: A Case Study with Progressively First-failure Censoring

Wijekularathna D. K.*

Department of Mathematics and Statistics, Troy University, United States

dwijekularathna@troy.edu

0000-0002-4222-960X

The progressive censoring of first-failure data has become increasingly popular over the past decade due to its usefulness in life-testing experiments and reliability theory. The number of failures is recorded by this method of data collection, which leads to improved accuracy of estimators. The Akash distribution, a relatively new probability model, can be utilized to model lifetime data, particularly in situations with varying hazard rates, whether increasing, decreasing, or constant. Although previous studies have explored the Akash distribution under complete and traditional censoring schemes, it is applied in the context of progressively first-failure censored data for the first time in this paper. A gap in the literature is filled by this novel application, providing new insights into the utility of the Akash distribution for reliability estimation. Model parameters and key reliability characteristics, including the reliability function, hazard rate function and mean residual life function, are estimated using both classical and Bayesian approaches. Maximum likelihood estimation (MLE) is included among the classical estimation methods, along with the construction of asymptotic and bootstrap confidence intervals. Bayes estimators are evaluated using a two-parameter gamma informative prior and Tierney-Kadane approximations, importance sampling and Metropolis-Hastings (M-H) algorithms are applied under the squared error loss function. The highest posterior density (HPD) intervals are constructed using an M-H algorithm. The practical applicability of the Akash distribution model is demonstrated using real-world data. The evaluation of these estimation procedures is conducted through numerical computations using a simulation study. In each scheme, the HPD credible intervals are consistently the shortest, followed by the asymptotic CI, bootstrap-t CI and bootstrap-p CI. As a result, the HPD approach appears to be the most efficient for this data set. Additionally, the estimates for reliability, hazard rate functions and mean residual life function are found to be highly similar across the different estimation methods. Overall, this research contributes to the field of reliability estimation by providing new methodologies, comparative insights and practical implications that can enhance both theoretical understanding and real-world applications.

Keywords: Akash distribution, Approximation methods, Bayesian estimation, Bootstrap confidence intervals, Maximum likelihood estimation

Comprehensive Analysis of Age and Gender-specific Mortality Rates in Sri Lanka

Yasara M. G. N.[1*], Jayasinghe C. L.[2] and Silva R. M.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]yasaramgn@gmail.com, [2]chathuri@sjp.ac.lk, [3]rsilva@sjp.ac.lk

[1]0009-0003-7241-9011, [2]0000-0003-2891-742X, [3]0000-0002-4915-7979

This study examined historical mortality patterns in Sri Lanka from 1937 to 2019, with a focus on age and gender-specific variations. The analysis is based on mortality data obtained from secondary sources published by the Department of Census and Statistics, Sri Lanka. The primary objective of the study was to analyze these mortality patterns for different age groups and genders. This study provides valuable insights into long-term health trends in Sri Lanka. It highlights the impact of social, economic and healthcare changes over time, offering crucial information for policymakers to optimize resources allocation and public health strategies. The analysis is conducted separately for periods 1937 to 1951 and 1950 to 2019 due to the variation in age categories recorded in the data source. No missing data was observed for the considered periods regarding age categories and gender. The analysis included examining death rates over time, conducting demographic analysis of mid-year population data, constructing population pyramids to visualize demographic shifts and performing disaggregated mortality rate analysis by age and gender categories to discern disparities in mortality. Findings indicate a general decline in mortality rates from 1937 to 1951. The most notable change occurs between 1946 and 1947, with a corresponding percentage decrease of approximately 32.25%. Females experienced higher death rates than males, as evidenced by an average Male to Female Mortality Rate Ratio (M: F MRR) of approximately 0.928 during the 1937-1951 period. In the Post-1950, mortality rates showed significant declines and stabilization in later decades. Notably, in the mid-2000s, there was a sudden uptick in death rates, followed by a gradual decline in subsequent years. The most recent years depict a relatively stable pattern, indicating a stabilization of mortality rates over this period. Gender disparities shifted after 1950, with males exhibiting higher mortality rates than females. The Average Male to Female Mortality Ratio (M: F MRR) for the 1950-2019 period is approximately 1.229. Age-specific trends showed that mortality rates were highest in certain age groups, mainly in older age groups for both genders across all years, with a general decline in death rates over time for most age groups. The observed patterns in mortality rates over the decades in Sri Lanka reflect a confluence of public health improvements, socioeconomic development and changing lifestyle factors. Population growth and demographic shifts were analyzed, revealing that in 1950, the population pyramid had a classic triangular shape, but by 2019, it had evolved into a more columnar structure, reflecting improved life expectancy and an aging population. This transition reflects improvements in healthcare, education and socio-economic development, resulting in longer life expectancy. The findings highlight the importance of investing in health research to develop new strategies for reducing mortality rates and improving public health outcomes.

Keywords: Age-specific mortality, Gender-specific mortality, Historical mortality patterns, Male to Female Mortality Rate Ratio (M: F MRR), Mortality rate

Yield Curve Modeling: Applicability of the Traditional Factor Models for Sri Lanka Government Bonds

Dayarathne K. P. N. S.[1*] and Thayasiwam U.[2]

[1,2]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

[1]sanjeewa.22@cse.mrt.lk, [2]rtuthaya@cse.mrt.ac.lk

[1]0000-0003-2365-2565, [2]0000-0002-3936-8174

Modeling the yield curve has resurfaced in the global capital market due to its capacity to forecast recessions and future inflation. Despite there are many attempts to uncover the yield curve models, globally it is a novel domain in Sri Lanka. Nelson-Siegel (NS-Model) has been identified as the state-of-the-art Model for understanding the term structure of the interest rate. The primary goal of this research is to evaluate the applicability of the most often discussed yield curve models, the NS-Model and Nelson-Siegel-Svensson models (NSS-Model), to Sri Lankan government bond yields. For this study, the secondary market data from 2010 to 2022 were used and focused on extracting the T-Bond yield data for 91 days, 182 days, 364 days and 2, 3, 4, 5, 6, 7and 10 years. The zero-coupon bond yields were extracted from the discounted bond rates for yield curve modeling. The NS and NSS models accounted for an average R-squared value of 96.25% and 98.75%, respectively for the sample period. Despite the relatively high R-squared value recorded as the average for the entire sample period, further investigation revealed, that R-squared values tend to vary across the entire sample period and fall sharply during some periods. Further, the above-average R-squared value is relatively lower than the results extracted from the literature for other developed markets, signaling a higher stable R-squared value is required for the full sample period. Throughout the data period, the NSS model outperformed the NS model in terms of R-squared value consistency. The additional parameter in the NSS model could capture sudden ups and downs in the yield curve. However, the R-squared value for both models fell in 2022 when compared to the preceding era, as the yield curve coincided with more hums and bums due to unpredictable economic conditions. Out of the three proxy parameters of the Nelson-Siegel model, namely the “level” factor (β_0), the “slope” factor (β_1) and the “curvature” factor (β_2), β_0 was highly correlated with 10 year Bond yield and β_1 was also highly correlated with yield difference between six months and 10 years. This was the standard explanation of the NS model parameters. These findings indicate that there is more room for constructing a better representative yield curve model or upgrading the existing traditional model by incorporating a more influential feature for the Sri Lankan capital market. In the future, the country's monetary authorities and investment banks will pay more attention to data-driven decision-making in interest rate behavior when setting economic and monetary aims. Investors also need to be well aware of the interest rate movements to achieve the hurdle rates for client portfolios. An accurate yield curve model would play a significant role in this regard.

Keywords: Government bond, Nelson-Siegel, Sri Lanka, Svensson, Yield curve

Forecasting Sri Lankan Gold Prices through Macroeconomic Variables

Devasurendra S. G.[1*], Tilakaratne C. D.[2] and Karunarathna G. H. S.[3]

[1,2,3]Department of Statistics, University of Colombo, Sri Lanka

[1]sanujidevasurendra@gmail.com, [2]cdt@stat.cmb.ac.lk, [3]hasani@stat.cmb.ac.lk

[1]0009-0008-1332-5943, [2]0000-0003-0330-845X, [3]0000-0003-3790-4903

In today's economic landscape around the globe, gold has become important to a country as it is scarce and a precious metal. Hence, forecasting the gold price has become very important to a country like Sri Lanka, which depends on gold imports for its necessities. Despite the critical importance of forecasting, there is a notable gap in an up-to-date study to predict the Sri Lankan gold price post-COVID-19 pandemic, as a rapid increment is observed afterward and existing literature has mostly paid attention to developing conventional time series models for the objective of forecasting. Therefore, this study aims to forecast the Sri Lankan gold price and identify the most influential macroeconomic indicators for the dynamics of the gold price, which includes 288 observations in the span of January 2000 to December 2023, where the data were primarily sourced from the Central Bank of Sri Lanka, the Department of Census and Statistics in Sri Lanka and the Colombo Stock Exchange. The study's descriptive analysis revealed that the financial crisis in 2022 had an impact on gold price and other economic indicators. The study concludes that US dollar exchange rates, narrow money supply, broad money supply and Colombo Consumer Price Index (CCPI) play a pivotal role in forecasting gold prices. Furthermore, the study offers a forecasting model comparison between conventional time series models and machine learning approaches under the objective of forecasting the gold price. The comparison includes Autoregressive Integrated Moving Average (ARIMA), Vector Autoregressive (VAR), Random Forest, XGBoost and two Long Short Term Memory (LSTM) approaches, which are one-step ahead of LSTM and feature-based LSTM. One-step ahead LSTM performed comparably well in forecasting for both the 12-month (MAPE = 0.03400 and RMSE = 26672.51) and 6-month (MAPE = 0.01422 and RMSE = 9986.36) trajectories incorporated in the study and the pre-crisis model forecasting (MAPE = 0.01364 and RMSE = 5799.14) also revealed the same conclusion. Furthermore, to address overfitting issues in the LSTMs, dropout layers and early stopping criteria were employed. Hence, by bridging the gap between traditional forecasting models and machine learning approaches based on their performance with an up-to-date study, this research contributes valuable insights into forecasting the complex fluctuations in gold prices in Sri Lanka.

Keywords: Gold price, LSTM, Macroeconomic indicators, Time series forecasting

Modeling Inflation, Exchange Rate and Interest Rate of Sri Lanka: Time Series Approach with Political, Natural and Health Crises

Iroshani W.[1*] and Mathugama S. C.[2]

[1,2]Institute of Technology, University of Moratuwa, Sri Lanka

[1]iroshanie@itum.mrt.ac.lk, [2]mathugamas@itum.mrt.ac.lk

[1]0009-0005-4279-8529, [2]0000-0003-0571-1680

In time series analysis, external shocks are essential for accurately capturing the effects of sudden, non-recurring events, enabling more precise estimation and forecasting. This study examines Sri Lankan economic variables including monthly year-on-year inflation rates, exchange rates and interest rates using monthly data from 2003 to 2023 obtained from the International Monetary Fund (IMF) database and the Central Bank of Sri Lanka. Dummy variables for natural disasters, health crises and political events were included to account for external shocks. Time series plots of inflation, exchange rates and interest rates were generated to identify trends, seasonal patterns and anomalies. Cointegration of the residuals of the fitted models was tested, revealing the presence of cointegration. Three Vector Error Correction (VEC) models were estimated, with one model selected based on information criteria and model diagnostic techniques. Model diagnostics ensured the reliability of the VEC models and the VEC model with log transformed time series variables with dummy variables was chosen due to its lowest Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Diagnostic tests, including the Portmanteau test for serial correlation and stability checks, were conducted. The Portmanteau test revealed significant autocorrelation in the residuals, indicating that the VEC model may not fully capture the data’s dynamics and might need modifications. However, stability analysis confirmed that all eigenvalues of the selected VEC model are within the unit circle, ensuring reliable forecasts. The selected VEC model provides valuable insights into Sri Lanka's economic variables. Nonetheless, the findings suggest that further adjustments may be needed to address residual autocorrelation.

Keywords: Exchange rate, Inflation, Interest rate, VEC

Investigating the Economic Convergence in Terms of Economic Policies: Evidence from East and South Asia

Kaveesha G. A. T.[1*], Devpura N.[2] and Karunasena K. A. D. B. S.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]tharushikaveesha000@gmail.com, [2]ndevpura@sci.sjp.ac.lk, [3]buddhik@sjp.ac.lk

[1]0009-0007-7476-9962, [2]0000-0002-8299-0540, [3]0009-0009-3381-192X

This study examines the economic convergence among 18 countries in the SAARC and ASEAN regions, as well as additional nations from East and South Asia, over the period 2000 to 2022. Given the significant disparities in per capita income across these countries, the study focuses on the role of economic policies in driving convergence, using variables such as trade, foreign direct investment, gross fixed capital formation, final consumption expenditure, broad money, inflation and population growth. By applying absolute, conditional β convergence and σ convergence models, the study investigates whether poor countries are catching up to richer countries in terms of per capita income. By applying initial growth regressions to check absolute β convergence and both static and dynamic panel regressions to assess conditional β convergence, the study investigates whether poor countries are catching up to richer countries in terms of per capita income. Static panel models were validated using the Breusch-Pagan test and dynamic panel data models using Hansen's J test and the Arellano-Bond test for serial correlation. A significantly negative β coefficient indicates that as the level of log initial income increases, the GDP growth rate declines, resulting in convergence. Additionally, standard deviations and the coefficient of variation are used to measure σ convergence, illustrating the degree of income dispersion among countries over time. The results indicate evidence of absolute and conditional β convergence, with poor countries growing faster than richer ones, thereby moving towards similar income levels. However, the findings on σ convergence are mixed, showing weak overall convergence with periods of divergence, particularly in the early years of the study. These mixed results in σ convergence could be the result of external shocks and structural changes in the region, such as the global financial crisis and the COVID-19 pandemic. These findings imply that although economic policies and structural factors play a role in promoting convergence, disparities in per capita income are likely to continue, driven by varying economic conditions. This study highlights the need for targeted economic strategies and improved regional collaboration to promote more equitable and sustainable growth across these Asian nations.

Keywords: Economic convergence, Economic policies, Growth disparities, Panel data analysis

Portfolio Formation Framework for Cryptocurrency Investment

Maleesha M. A. N.[1*] and Liyanage U. P.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[1]nipunimaleesha8@gmail.com, [2]prabhathuoc@gmail.com

[1]0009-0005-3586-1438, [2]0000-0003-1201-3418

Cryptocurrencies are digital money secured by cryptography and those have decentralized transactions over blockchain technology. There are over 10,000 cryptocurrencies up to now. Some of them are Bitcoin, Ethereum and Ripple. Bitcoin was the first invented cryptocurrency in 2009 by an anonymous entity known as Satoshi Nakamoto and it is often seen as "digital gold" due to its limited supply and store of values properties. These digital currencies can be owned by purchasing and alternatively, through mining, staking, or receiving it as payments for goods and services. In this study, the trading part of the cryptocurrency investment was addressed. As assets cryptocurrencies have high volatility and potential for significant return. This is the main reason for addressing the cryptocurrency market for this study. In Sri Lanka, a user base over 320,000 works with cryptocurrency transactions. The main objective of this study is to find optimal portfolios by introducing strategies for investing based on risk, return and price classifications, ensuring the maximum return with minimum risk. Portfolio formation is carried out randomly and every cryptocurrency has its own price profile having diverse volatility. To have balanced portfolios, it is essential to identify the classes in which the risk, return, volatility and prices belong. The K-means clustering technique for single-value clustering was used to identify the stable risk and return levels and their associated classes. Risk and Return levels are defined as High, Moderate and Low respectively. Price categories are defined by using behaviors of the prices of the currencies. Based on the risk aversion behaviors of the investors, 6 levels are formed as High risk- High return, Low risk - High return, Moderate risk - High return, Low risk- Low return, High risk - Moderate return, High risk - Low return. 100 random portfolios for each strategy were designed according to the above classifications. The optimal portfolio for each strategy was identified based on Modern Portfolio Theory (MPT). The outcome of this study is the valid framework for cryptocurrency investment. After gathering data, identify the stable classes for risk and return levels. Then based on those details according to price movements portfolio building task can be carried out. The portfolio formation framework for cryptocurrency enhances risk management through diversification, optimizes returns using MPT and provides a structured, data-driven approach to investment. Based on the investigation, optimal portfolios are justified using the formed simulation platform.

Keywords: Cryptocurrencies, Framework, K-Means, Portfolios, Risk

Forecasting Global Crude Oil Prices using Multivariate Time Series and Machine Learning Techniques

Wijesekera G. M. P.[1*] and Tilakaratne C. D.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]madushiwijesekera@gmail.com, [2]cdt@stat.cmb.ac.lk

[1]0009-0001-7789-874X, [2]0000-0003-0330-845X

Crude oil, which is considered as the life blood of the global economy plays a major role across many industries. An oil benchmark is a specific type of crude oil that is used as a reference price for buyers and sellers of crude oil and the three major oil benchmarks are the Brent, WTI and Dubai. Forecasting oil benchmark prices is challenging due to market volatility, yet essential for industry stakeholders to make informed decisions and manage risks. Many studies have forecasted crude oil prices using time series techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Generalized Autoregressive Conditional Heteroskedastic (GARCH) and machine learning methods such as Random Forest, XGBoost and Long Short Term Memory (LSTM). However, only a few have combined all three major benchmarks in forecasting the oil prices. This study addresses the above mentioned gap, by incorporating all three major benchmark oil prices for forecasting, using multivariate time series and machine learning techniques. Monthly close prices of three benchmarks from 2002 to 2023 were used, with the dataset comprising 16 variables; 3 representing oil benchmarks and 13 related to demand, supply and financial factors. Cointegration among the benchmarks, identified through the Engle-Granger and Johansen tests, led to the fitting of a Vector Error Correction Model (VECM). Granger causality analysis revealed causal relationships between oil prices and financial factors, namely the US Dollar Index (USDX), U.S. Gross Domestic Product (GDP) and the S&P500 index. Random Forest feature selection was also employed to identify the most important factors associated with the three oil benchmark prices, revealing that lagged values of the U.S. Dollar Index are highly associated with oil prices. Subsequently, three LSTM models were fitted with three different feature subsets for two forecast horizons: 6 months and 1 year. The LSTM model incorporating features determined by the Granger Causality Test showed the best performance among all the candidate models for both the short-term and long-term horizons. Moreover, the performance of the long-term forecasting was even better than the shorter horizon forecasts, with lower values of 0.663, 0.468 and 0.524 for test Root Mean Squared Error (RMSE) and 3.856, 3.712 and 2.555 for test Mean Absolute Percentage Error (MAPE) for the Brent, WTI and Dubai benchmarks, respectively. These high accuracies and better performances of the best model, will enable stakeholders in the oil industry to make better strategic decisions and improve risk management.

Keywords: Crude oil, Granger causality, Forecasting, Machine learning, VECM

An Application of Autoregressive Distributed Lag (ARDL) Model Approach to Forecast the Index of Industrial Production of Sri Lanka

Withanage T. R.[1*], Perera K. M. P[2] and Devpura N.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]thilinawithanage1@gmail.com, [2]manjula@sjp.ac.lk , [3]ndevpura@sci.sjp.ac.lk

[1]0009-0000-1529-6412, [2]0000-0001-5288-669X, [3]0000-0002-8299-0540

This study evaluates the effectiveness of multivariate Autoregressive Distributed Lag (ARDL) models in forecasting Sri Lanka's Index of Industrial Production (IIP), a key short-term indicator that measures volume changes in industrial output. It also addresses a gap in the literature by applying ARDL models, which have not been previously used to forecast IIP or investigate the lag effect of significant factors influencing IIP in Sri Lanka. For the ARDL models, explanatory variables Colombo Consumer Price Index (CCPI), Import Trade Index (ITI) – Volume, Export Trade Index – Volume (ETI), Monthly Average Exchange rate in USD (MAER), Electrical Power Consumption in Gigawatt hours (EPC), Average Weighted Prime Lending rate (AWPR) were selected based on the literature. The Augmented Dickey-Fuller (ADF) test and the Hylleberg, Engle, Granger and Yoo (HEGY) test were conducted to identify non-seasonal and seasonal unit roots respectively. The ADF test indicated that all variables are non-stationary, while the HEGY test revealed the presence of seasonal unit roots in the ETI and the IIP. Variables with seasonal unit roots were seasonally differenced and other variables were non-seasonally differenced before modeling. In the ARDL approach, six models were developed by progressively adding explanatory variables until the sixth model, which included all the variables. The selection of variables for each model was based on the coefficient of determination (R^2) and the lag length for each explanatory variable was determined using the Bayesian Information Criterion (BIC). Additionally, the Seasonal Autoregressive Integrated Moving Average-Autoregressive Conditional Heteroskedastic (SARIMA-ARCH) model was built using the same period of data to serve as the base model for comparing the performance of the ARDL models. Data from January 2000 to December 2021 was used to develop the models and forecasting accuracy was evaluated from January 2022 to March 2022 using static forecasting. Mean Absolute Percent Error (MAPE) was used to assess the accuracy of the forecasts. Model validation was carried out using the Ljung-Box, Breusch-Pagan and Jarque-Bera (JB) tests. The forecasting accuracies were evaluated after back-transforming the results to the original scale and accounting for seasonal effects for the 1-month and 3-month forecast horizons. Among the ARDL models, the one that included seasonally differenced IIP as the response variable, along with seasonally differenced ETI and non-seasonally differenced EPC, CCPI and ITI as explanatory variables, demonstrated the best forecasting performance for both the 1-month and 3-month forecast horizons, with overall superiority in the 1-month forecast horizon. The SARIMA-ARCH model showed the best overall performance in the 3-month forecast horizon. Furthermore, this study identified electricity power consumption, exports, imports and inflation as significant factors influencing Sri Lanka's industrial production, based on the significance of variables EPC, ETI, ITI and CCPI.

Keywords: Autoregressive conditional heteroskedastic (ARCH), Autoregressive distributed lag (ARDL) models, Index of Industrial Production (IIP), Seasonal autoregressive integrated moving average (SARIMA)

A Statistical Perspective on Record-breaking Batting Performances in ODI Cricket

Kumarage K. D. Y. R[1*] and Abeysundara S. P.[2]

[1,2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

[1]yasindrark@gmail.com, [2]sachitha@sci.pdn.ac.lk

[1]0009-0003-0761-294X, [2]0000-0002-1228-4771

Extreme Value Theory (EVT) offers a powerful framework for predicting and analyzing rare but significant performance outcomes, such as record-breaking achievements in sports. Cricket, one of the world's most popular sports, places a strong emphasis on batting and the pursuit of breaking records significantly enhances its appeal. This study aims to model the distribution of highest individual scores in One Day International (ODI) cricket using Extreme Value Theory, identify significant factors influencing exceptional batting performances and estimate the return levels and probabilities of surpassing current record scores. The Generalized Extreme Value (GEV) distribution was utilized to model outstanding batting scores from ODI matches, using data sourced from ESPNcricinfo, which includes the highest individual batting score of a player in a match recorded annually from 1971 to 2024, totaling 54 scores. Preliminary analysis reveals that the highest recorded score is 264, the lowest is 82 and the average highest score is 168. The analysis indicates that top-order and opening batters tend to achieve higher ODI scores compared to other positions. Multiple Regression with Box-Cox transformation on the highest individual score followed by a stepwise method revealed that batting average, number of career centuries, number of fours and sixes are statistically significant in achieving the highest individual batting score. The adjusted R-squared value for the model is 89.59%. The block maxima method was employed to fit the GEV model and stationarity was confirmed using the Augmented Dickey-Fuller test. Model fit was further validated with diagnostic plots and the Kolmogorov-Smirnov test indicating that the GEV model accurately represents the distribution of highest individual scores in ODI matches. The study analyzes return levels for highest scores, projecting the expected maximum scores for various return periods. The current highest individual ODI score of 264 is expected to be surpassed, as the estimated highest score for a 100-year return period is 271, with a 95% confidence interval of (244.19, 298.59). The probability of breaking the current record is approximately 0.25% within the next year and 0.74% over the next three years. These findings provide valuable insights for sports analysts and coaches seeking to understand and optimize extreme batting performances in cricket through the application of Extreme Value Theory.

Keywords: Cricket, Extreme value theory, GEV model, Highest scores

An Introduction to Pickleball and Some Analytics

Muthukumarana M.[1*] and Swartz T. B.[2]

[1,2]Department of Statistics, Simon Fraser University, Canada

[1]mahen_muthukumarana@sfu.ca, [2]timothy_swartz@sfu.ca

[1]0009-0001-5979-9991, [2]0000-0001-6092-6727

The sport of pickleball is the fastest growing sport in the United States and is catching on across the world. As a relatively new sport, pickleball analytics is underdeveloped. This research examines two pickleball problems that provide insights as to how the game ought to be played optimally. It will be seen that analytics sometimes challenges conventional pickleball wisdom. The first example concerns defensive positioning and how using geometrical reasoning can lead to alternative alignments. The second example concerns an examination of the merits of the third shot drop versus the defensive lob. This comparison is made using probabilistic reasoning involving personal probabilities of executing various shots.

Keywords: Analytics, Lob, Pickleball

Applying a Markov Chain Model for Ranking and Analyzing Performance of Cricket Teams in T20 International Matches

Thuvaragan C. [1*] and Arivalzahan S.[2]

[1,2]Department of Mathematics and Statistics, University of Jaffna, Sri Lanka

[1]thuvaragan980@gmail.com , [2]arivu@univ.jfn.ac.lk

The primary objective of ranking teams in T20 International Cricket is to identify the most challenging opponents for each team. While the International Cricket Council (ICC) uses a ranking system as standard practice, it may not effectively capture certain head-to-head match results. This study introduces a stochastic Markov chain model for ranking teams, using data obtained from ESPNcricinfo.com on T20 international matches played between 2005 and 2021. The study focuses on ten ICC full-member countries: Sri Lanka, India, Pakistan, Australia, England, South Africa, West Indies, New Zealand, Bangladesh and Zimbabwe. The frequency approach was used to determine transition probabilities, classifying outcomes as either Win or Not Win (including Loss, Tie, or No Result). Team rankings were derived from steady-state probabilities, with India emerging as the top-ranked team according to our model. Additionally, separate steady-state winning probabilities were calculated for each team against all their respective opponents. Based on these steady-state probabilities, opponents were ranked for each team. The effectiveness of the proposed head-to-head prediction method was evaluated through cross-validation with the ICC rankings, using a test sample of matches from January 2022 to December 2022. The results show that the proposed model outperforms the ICC T20 rankings in predicting head-to-head outcomes in T20 cricket.

Keywords: Cricket ranking, Markov chain model, Steady-state probability, Stochastic process, Transition probability

Analyzing the Impact of Game-time Weather on One-day International (ODI) Cricket: Repository of Cricket and Weather Data for Economic Evaluations and Research

Walpitage D. L.[1*], Wickramasinghe R. I. P.[2] and Sanjeewa. S.[3]

[1]Department of Enterprise Analytics, University of Kansas Medical Center, United States ,
[2]Department of Mathematics, Prairie View A&M University, United States, [3]Citizens Development Business Finance PLC, Sri Lanka

[1]dwalpitage@kumc.edu, [2]iprathnathungalage@pvamu.edu, [3]sajith.sanjeewa@cdb.lk

[1]0009-0003-5721-4560, [2]0009-0002-8594-4742, [3]0009-0007-9985-7889

Previous research on cricket and weather has primarily addressed the effects of rain-shortened games and the resulting implications for game outcomes and financial aspects. Research on the impact of weather-related conditions on player and team performance, including conditions such as cloud cover, humidity and temperature has produced mixed results. A major limitation of these studies is that they often consider weather at a macro level, without accounting for changes that occur within individual innings. This study fills the gap by combining ball-by-ball cricket performance data with real-time weather data to create a comprehensive repository. A Power BI® dashboard was developed to allow users to visually analyze data from 1,232 One Day International (ODI) games played over the last decade, enabling users to explore how weather-affected ground conditions influence game interruptions and player performance. The dashboard offers multiple geo-analytic views, interactive graphs with pop-ups and detailed tooltips and drillable tables. One functionality focuses on rain-affected games where the Duckworth-Lewis (D/L) method determined the outcome, highlighting that Sri Lanka, Scotland and the West Indies have the highest percentages of such matches. A detailed analysis reveals that rain interruptions significantly impact win rates, particularly in Sri Lanka. Another functionality, the game-time weather view, analyzes how playing conditions impact performance, showing that most non-D/L games occurred under sunny or clear skies. Regression analysis examined the impact of weather conditions; Sunny or Clear, Cloudy, Overcast and Rain Possible and Some Rain Around, on batting performance. The over grouping (e.g., 0.1-5, 5.1-10 overs) explained 20.04% of the variance in runs scored per five-over period. Including the effect of wickets that fell in the previous five overs added another 4.02% (p-value < 0.05) and batting team and their recent batting form contributed an additional 1.92% (p-value < 0.05). Venue-based weather conditions explained additional 2.67% (p-value < 0.05) variance in runs scored per five-over period. The analysis showed that under overcast weather, with high humidity ($M = 27.86$, $SD = 11.66$, p-value < 0.05) fewer runs were scored compared to other humidity levels. The fewest runs were scored when conditions are cloudy and high humid ($M = 27.66$, $SD = 11.34$). No significant differences were found in runs scored under sunny and clear conditions across varying temperature categories. The study findings and interactive dashboard offer insights for cricket planning and performance optimization.

Keywords: Climate, Cricket, Data Integration, Visualizations, Weather

A Versatile Framework for Pruning False Alarms in Anomaly Detection Systems with an Emphasis on Interpretability

Dissanayake D. M.[1*], Navarathna R.[2] and Viswakula S. D.[3]

[1,3]Department of Statistics, University of Colombo, Sri Lanka, [2]OCTAVE, John Keels Group

[1]dindisn.stat@stu.cmb.ac.lk, [2]rajitha.jkh@keells.com, [3]sam@stat.cmb.ac.lk

[1]0009-0000-9267-5053, [2]0009-0001-3674-4146, [3]0000-0002-6961-898X

Anomaly detection plays a critical role in diverse domains such as manufacturing, cybersecurity, finance, retail, telecommunications and healthcare. Despite its significance, traditional anomaly detection methods often struggle to accurately distinguish between genuine and normal variations in data, resulting in a considerable number of false positives. Genuine anomalies refer to instances that deviate substantially from expected patterns. These erroneous alerts not only disrupt operational workflows but also burden resources with unnecessary investigations into non-existent issues. Further, it leads to suboptimal resource allocation and increased operational costs. In response to this issue, this study proposes an innovative approach aimed at mitigating false positives in anomaly detection by integrating association rule mining (ARM) with the Isolation Forest (ISF) algorithm. ARM is a data mining technique that identifies interesting relationships or patterns within datasets. By uncovering meaningful associations between variables, ARM can provide valuable insights into the underlying structure of the data. In the context of anomaly detection, ARM offers the potential to identify strong patterns indicative of normal behavior, which can then be used to differentiate anomalies from legitimate data points. Complementing ARM, the ISF algorithm is a state-of-the-art machine learning-based anomaly detection technique that operates by isolating anomalies in a dataset. Unlike traditional methods that rely on distance or density-based metrics, ISF employs a tree-based approach to identify anomalies more efficiently. By isolating anomalies in relatively few steps compared to normal data points, ISF can effectively detect outliers or anomalies present in the dataset. By leveraging the combined strengths of ARM and ISF, our approach enhances anomaly detection reliability while reducing false positives. The target anomalies were flagged by ISF with borderline probabilities and assessed whether they violated significant ARM patterns. Empirical evaluations on real-world tabular numerical datasets, including Shuttle, Annthyroid, Mammography and Statlog from the UCI repository, demonstrate that our methodology improves detection performance and reduces false alarms, outperforming the baseline ISF approach. The area under the receiver operating characteristic curve (AUC) increased by 4.3% each for Shuttle and Mammography data, 5.9% for Stat-log data and 0.4% for Annthyroid data, while FPR reduction was 6.6% for Mammography data, 4.5% for Shuttle data, 12.4% for Stat-log data and 5.3% for Annthyroid data, underscoring the effectiveness of our method. To ensure robustness, each dataset was run multiple times to account for the inherent randomness of the ISF algorithm, with results averaged and standard deviations calculated. Additionally, explanations are provided for any anomalies by highlighting the strong rules violated by each instance, facilitating the interpretability of detected anomalies. While integrating ARM with ISF improves anomaly detection and reduces false positives, it also presents potential challenges. The computational complexity of ARM can affect scalability, especially with large datasets or numerous variables. Nevertheless, the proposed approach significantly enhances anomaly detection performance and provides a valuable framework for future applications.

Keywords: Anomaly detection, Association rule mining, False positive pruning, Isolation forest

A Novel Hybrid Algorithm for Binary Classification Tasks: Addressing the Class Imbalance Problem in 2D and 3D Feature Spaces

Madhuwanthi U. S. P.[1*] and Chandrasekara N. V.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[1]uduwagepm@gmail.com, [2]nvchandrasekara@kln.ac.lk

[1]0009-0000-2013-2316, [2]0000-0003-3232-837X

Class imbalance is prevalent in many binary classification tasks, where the uneven distribution of class instances leads to biased models and reduced prediction accuracy. Addressing this challenge requires careful balancing techniques that prevent overfitting or losing valuable data. Classification is a task that involves categorizing data into predefined classes or categories based on their features. The class imbalance problem (CIP) in which the number of instances within the classes is unevenly distributed, is crucial in many real-world datasets when classifying the instances into class labels or categories. That is the number of minority class instances (positive class) which is the class interested in, in many cases is significantly less than the number of majority class instances (negative class). Different techniques such as oversampling, under-sampling and hybrid techniques can be used to address the CIP. Oversampling involves increasing the number of instances in the minority class by either duplicating existing instances or generating synthetic examples while under-sampling involves lowering the number of instances in the majority class. Applying oversampling alone causes data replication while under-sampling causes losing valuable information. Based on the number of classes in the target variable, classification tasks can be defined into two; Binary classification problems that relate to cases where the target variable has only two classes and multi-class classification problems that relate to more than two classes in the target variable. This study has mainly focused on the datasets that include only two classes in the target variable within a limited number of features that is two-dimensional and three-dimensional spaces. The algorithm proposed in this study is a novel and unique algorithm that can be used to balance the instances in a dataset within its classes. The proposed algorithm to overcome the limitations of existing techniques leverages a quartile-based approach, balancing class distributions by both oversampling the minority class and under-sampling the majority class, thereby maintaining a balanced dataset size. The proposed hybrid resampling technique is evaluated with selected most effecting two and three features using Logistic Regression, of the PIMA Indian Diabetes dataset with an imbalanced class distribution, ‘DiabetesPedigreeFunction’, ‘Pregnancies’ and ‘BMI’ for 2D and 3D spaces respectively. Performance evaluation is carried out using average accuracy for 100 iterations to assess the effectiveness of the approach. For classification, a Support Vector Machine (SVM) with one of the simplest kernel functions, the polynomial kernel function has been applied as the classifier. The performance of the algorithm is compared with existing oversampling techniques; ROS, SMOTE and ADASYN and under-sampling techniques; RUS and Tomek Links and a hybrid technique; SMOTETomek. The experimental results demonstrated that the proposed algorithm had gained 83.56% and 68.77% accuracy for two-dimensional and three-dimensional spaces respectively while all the above-mentioned oversampling, under-sampling and hybrid techniques gained less accuracy than the novel algorithm. This algorithm would provide the advantage of effectively balancing the datasets allowing them to make more accurate predictions toward the minority class. The novel hybrid approach effectively balances the dataset and improves classification performance on imbalanced data. Future work will extend the algorithm to higher-dimensional spaces and multi-class tasks.

Keywords: Binary classification, Class imbalance problem, Hybrid techniques, Resampling, SVM

Evaluation of Selected Spatiotemporal Missing Value Imputation Techniques: An Application to Estimate Sequential Missing Values of Rainfall Data at Ratnapura Area

Heendeniya H. R. N. C.[1*] and Tilakaratne C. D.[2]

[1,2]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[1]nayanachathuranga2@gmail.com, [2]cdt@stat.cmb.ac.lk

[1]0009-0003-8006-6674, [2]0000-0003-0330-845X

Ratnapura area experiences floods and landslides because it receives heavy rains, especially during the southwest monsoon. Therefore, it is important to build forecasting models to forecast rainfall values in this area. However, modeling the variability in rainfall can be frequently hindered by the missing values, especially the sequential missing values. When estimating missing rainfall data, it is worth considering not only the temporal dependencies but also the spatial dependencies. Currently, there are various spatiotemporal imputation techniques used to fill in the gaps in the rainfall data. However, studies comparing the spatiotemporal missing value techniques in estimating missing values in rainfall data are extremely rare. To address this gap, this study compared a few chosen techniques, namely K-Nearest Neighbors (KNN), MissForest and Denoising Autoencoder in estimating missing values. In this approach, Ratnapura station and nearby rain gauging stations with complete daily rainfall data sets (i.e. without missing values), were selected. The selected stations are Moralioya, Detanagala, Keragala, Ilubbuluwa, Ratnapura, Balangoda and Elston. To test the accuracy of imputation methods, daily rainfall data of the six stations from 2015 to 2019 were used. This period was selected as there were no missing values in the series corresponding to selected stations. Subsequently, the daily rainfall data of the target station (Ratnapura) was estimated using the data of surrounding stations based on the selected techniques, so that actual data and the estimated data could be compared. Each estimated series was compared with the actual data series using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Imputation was done in various missing data scenarios. Sequential missing values were imputed focusing on wet and dry seasons. For each season, missing periods of varying durations (week, two weeks and month) were examined. Additionally, for random individual missing values, different missing rates ranging from 5% to 40% were considered. Furthermore, when performing the MissForest technique, a seasonal component was added using time series decomposition. According to the literature, such application was not done in previous studies in the context of spatiotemporal missing data imputation. However, the addition of the seasonal component did not get the expected higher imputation accuracy compared to that without the seasonal component. The results of the study show that the KNN method is the most suitable method for all the scenarios showing lower RMSE and MAE values in all the cases. Specifically, for random individual missing values with missing rates 5%, 10%, 20% and 40%, the KNN method showed RMSE values 12.9, 11.2, 11.7, 12.8 and MAE values 8.3, 6.9, 7.1, 7.4 respectively. For sequential missing values with missing periods of lengths one week, two weeks and a month, the KNN method achieved MAE values of 1.9, 4.6 and 6.2 for the wet season and 0.2, 3.2 and 1.9 for the dry season, respectively.

Keywords: Deep learning, Machine learning, Missing value imputation, Rainfall data, Spatiotemporal data

Periodic Time Series Models for Multi-station Temperature Analysis in Nuwara Eliya District

Jayathilake W. M. Y. L[1*] and Hewaarachchi A. P.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[1]yasanthalakmal36@gmail.com, [2]anuradhah@kln.ac.lk

[1]0009-0006-4286-0852, [2]0000-0002-3608-5881

Temperature data analysis is significant in various fields, climatology is an important area where accurate forecasts play a key role. Temperature analysis has focused mainly on trends and seasonality, often overlooking the potential information offered by periodic correlation. Periodicity, characterized by patterns recurring at regular intervals, is a crucial aspect that can have a significant impact on the accuracy of temperature models. The temperature in Nuwara Eliya district in Sri Lanka displays seasonal and periodic behaviors. However, research on modeling the periodic correlations is limited. The key objective of this study was to model the temperature data using the periodic autoregressive model to predict under the periodic correlation. For this study, monthly mean minimum and maximum temperatures for the period 1997 to 2023 from four weather stations, namely Kotmale, Nuwara Eliya, Seetha Eliya & Talawakelle were considered. Fisher's g test is used to identify periodic correlation and AIC & BIC criteria for select model order for these series. Periodic autoregressive (PAR) models are fitted to the temperature which displays the periodic correlations. The mean minimum temperature series of Nuwara Eliya station model as PAR(2), the mean minimum temperature series of Seetha Eliya model as PAR(1) and the mean minimum temperature series of Kotmale model as PAR(3). Further ARIMA(1,0,1)(1,1,1)[12] model for Nuwara Eliya monthly mean maximum temperature data, ARIMA(1,0,1)(1,1,1)[12] model for Seetha Eliya monthly mean maximum temperature data, ARIMA(1,0,1)(2,1,2)[12] model for Kotmale monthly maximum temperature data, ARIMA (1,1,2)(1,1,1)[12] model for Talawakelle monthly mean maximum temperature data and ARIMA (2,1,1)(1,1,1)[12] model are fitted. These fitted models have high accuracy in forecasting temperature data and the models capture the periodic correlations. Further multivariate periodic time series models can be built to improve forecast accuracy.

Keywords: Periodic, Seasonal, Temperature, Time Series, Weather

Investigating Basketball Players' Rebounding Ability Using GLMM and BGLMM

Kulasekara O. V.[1*], Silva R. M.[2] and Withanage N.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]oshinikulasekara123@gmail.com, [2]rsilva@sjp.ac.lk, [3]niroshan@sjp.ac.lk

[1]0009-0000-9666-3551, [2]0000-0002-4915-7979, [3]0000-0001-8905-3878

Examining a player's defensive rebounding ability the skill of recovering the ball after an opponent's missed shot is a key factor in determining the optimal starting lineup. Defensive rebounding is essential in basketball, as it creates additional scoring opportunities while limiting the opponent's chances to score. Over the years, numerous approaches have been proposed to assess players' rebounding ability. However, key player-specific factors have often been overlooked. This study aims to introduce a more effective framework for evaluating rebounding ability by utilizing Generalized Linear Mixed Models (GLMM) and Bayesian Generalized Linear Mixed Models (BGLMM). For the analysis 2022-23 National Basketball Association (NBA) season data were used. Findings of the study show that player specific factors such as height, boxout, period have a significant impact on defensive rebounding ability and also in periods 2, 4 with periods 1, 3 have significantly different impact on defensive rebounding ability. Furthermore, these study results suggest a slightly different top 10 rebounding players order with respect to the usual order. This study concludes valuable insights to coaches and managers to select the best starting lineup from a pool of players for a game and about alternative player substitutions.

Keywords: BGLMM, Defensive rebounding ability, GLMM, NBA, Sport analytics

LSTM-based Predictive Modeling for Weekly Tea Auction Prices in Sri Lanka with Economic Sentiment Analysis

Perera W. L. S. D. D.[1*], Meyen N.[2] and Lakraj G. P.[3]

[1,3]Department of Statistics, University of Colombo, Sri Lanka, [2]Cogniata (Pvt) Ltd., Sri Lanka

[1]perera.dehara@gmail.com, [2]Nuzhi@cogniata.com, [3]pemantha@stat.cmb.ac.lk

[1]0009-0002-5654-5600, [2]0009-0009-8653-6300, [3]0000-0003-3921-8552

Prediction of the outcomes of Sri Lankan tea auctions heavily relies on the subjective judgment of tea brokers and might not account for global economic fluctuations. Therefore, a solid tea price prediction system is important for the country's economy. This research investigates the effectiveness of Long Short Term Memory (LSTM) networks and the use of economic sentiment analysis in predicting weekly tea auction prices for Sri Lanka's diverse tea categories (Low, Medium and High Grown). LSTM networks will be explored here to stand out from previous studies and economic sentiment data to understand how global market sentiment affects auction prices. The study explores the influence of various factors on price fluctuations by developing two LSTM models; a baseline LSTM model associating tea auction quantity, weather data, crude oil price and USD exchange rate with the target tea auction price for each tea elevation category and a combined model integrating economic sentiment analysis derived from news articles related to Sri Lanka and its top tea importers (Iraq, Turkey, Russia and UAE). Hyperparameter tuning is employed to optimize each model's performance. While the inclusion of economic sentiment improved predictive performance for Low and Total (general) elevation types, it did not outperform the baseline model across all categories. The Total elevation model achieved a Root Mean Squared Error (RMSE) of 0.0893 with economic sentiment data, compared to 0.1054 without it, indicating the possible influence of market sentiment on auction prices. These findings offer valuable insights for stakeholders in the Sri Lankan tea industry, showing the potential of data-driven predictions for informed decision-making.

Keywords: LSTM, Sentiment analysis, Tea auction price, Time series prediction

Assessing the Risk of Giving Advance Payments in the Tea Brokerage Industry: A Case Study

Rajapaksha R. M. S. D.[1*] and Viswakula S. D.[2]

[1,2]Department of Statistics, University of Colombo, Sri Lanka

[1]sajini.rajapakshasd@gmail.com, [2]sam@stat.cmb.ac.lk

[1]0009-0003-5922-9109, [2]0000-0002-6961-89X

The Sri Lankan tea industry boasts a rich history. Today it has evolved into a prominent global player. It is currently the highest net foreign exchange generator. In 1963 tea brokers were introduced in Sri Lanka to address the growing complexity of the tea industry and facilitate fair and efficient transactions between tea producers and buyers. Tea brokers are increasingly involved in the financial sector, one of the industry's most critical areas. The advance payments are typically extended as weekly loans. The amount can vary based on the agreement between the tea producers and the tea broker where stock works as an asset. This study assesses the risk of giving advance payments in the tea brokerage industry. An indicator to measure the risk of giving advance payments is derived by using factor analysis. Linear and non-linear relationships between the variables were analyzed using the Pearson correlation coefficient and the Spearman Rank correlation coefficient. Two factors were derived from factor analysis and those two were combined using sigmoid function considering the factor scores. The sigmoid function was used to combine factor scores to get a probability-like score. The risk indicator was finalized as a value between 0 and 1 based on its effectiveness. The main variables used in this study to develop the indicator to assess the risk of giving advance payments are Stock Quantity, Proceeds, 70% of estimated Stock Value, Advance Balance, Advance Interest and Over Advance Amount. Autoregressive Integrated Moving Average (ARIMA) and Long-short Term Memory (LSTM) models were used to forecast the proposed risk indicator individually for the factory as these models can be used as precious and accurate methods of predicting the value for the next time lag based on historical data. The models were developed for a factory identified by the tea broker as high-risk. For the selected factory, the LSTM model with hyperparameter tuning performs better compared to other models with a lower Mean Absolute Percentage Error (MAPE) value of 17.22% and a lower Root Mean Squared Error (RMSE) of 0.06). The proposed model can be used as a prototype to test the risk of giving advance payments to other factories. The main limitation of this study is the absence of a base reference level for the risk assessment which leads to a challenge in evaluating the risk measure that is developed under this case study. However, in this study, the performance of the indicator was evaluated by seeking the domain expertise knowledge. This study can be further developed by customizing risk assessment methodologies to suit the specific characteristics of individual tea factories. Additionally, expanding the scope of variables through the collection of more comprehensive data could significantly enhance the accuracy and relevance of the risk assessment. This study can be used as a case study to assess the risk of advance payments in the tea brokerage industry.

Keywords: Advance payments, ARIMA, Ceylon tea, Factor analysis, LSTM

Utilizing Bayesian Regression Analysis and Optimization Approaches to Identify Main Drivers in Inventory Management

Arachchi K. A. I. H. K.[1*] and Thilan A.W. L. P.[2]

[1,2]Department of Mathematics, University of Ruhuna, Sri Lanka

[1]ishanihimaya@gmail.com, [2]pubudu@maths.ruh.ac.lk

[1]0009-0008-6940-5651, [2]0000-0001-7708-6349

Inventory management is a crucial research area focusing on stock optimization, cost reduction and supply chain efficiency, with particular attention given to optimizing practices and identifying key factors. If inventory management fails to maintain a sustainable level, it directly hampers the company. To address this problem, a new methodology is proposed that combines Bayesian regression analysis with optimization techniques. The historical sales data was obtained for this study from an online database platform called "Kaggle". A Bayesian logistic regression approach was utilized to incorporate prior knowledge from historical data into the Bayesian design. The Bayesian optimal experimental design was obtained by maximizing the expected utility function. The Kullback-Leibler divergence was selected as the utility function in our study, facilitating precise estimation of the model parameters. The Approximate Coordinate Exchange optimization approach was employed to determine the optimal design, starting with three initial random designs. This study demonstrates the efficacy of Bayesian regression analysis and optimization techniques in identifying and prioritizing key drivers impacting inventory management. By incorporating prior knowledge and accounting for uncertainty, businesses can achieve more accurate demand forecasting, improved inventory turnover and reduced holding costs, ultimately driving superior supply chain efficiency and organizational performance.

Keywords: Approximate coordinate exchange, Kullback-Leibler divergence utility, Logistic regression approach, Optimal design

Mixture Modeling to Explore the Relationship between Chemicals in Drinking Water and Zebrafish Behavioral Responses

Dilrukshi R. K. K.[1*], Niroshan W.[2], Nishad J.[3] and Fernando P. W.[4]

[1]Department of Spatial Sciences, General Sir John Kotelawala Defence University - Southern Campus, Sri Lanka, [2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka, [3]The Nicholas School of the Environment, Duke University, North Carolina 27708, United States ,
[4]Department of Mathematics, University of Sri Jayewardenepura, Sri Lanka

[1]kdilrukshi@kdu.ac.lk, [2]niroshan@sjp.ac.lk, [3]nj58@duke.edu, [4]panif71@gmail.com

[1]0009-0006-8799-5873, [2]0000-0001-8905-3878, [3]0000-0003-2485-6893, [4]0009-0002-0578-2454

Mixtures of chemical contaminants can pose significant health risks to humans and wildlife, even at levels deemed safe for individuals. Zebrafish (Danio rerio) offer a high-throughput exposure model with the complexity of a vertebrate organism, making them well-suited for evaluating mixture toxicity. However, substantial gaps exist in statistical methods for assessing the associations between chemical mixtures and phenotypic endpoints in toxicity testing. Here, a Weighted Gene Co-expression Network Analysis (WGCNA) was developed to address this challenge, focusing on a larval behavioral assay and leveraging data from 92 well-water samples from Maine and New Hampshire, USA. Our study aims to implement mixture-relevant statistical approaches to elucidate the relationships between chemicals in drinking water and the behavioral responses observed in zebrafish. A WGCNA was employed to uncover gene expression patterns that mediate behavioral effects induced by these chemical mixtures. Our quantile g-computation analysis identifies associations between the grey, brown and turquoise modules and neutrophil responses. Individual chemicals within these mixture models exhibit both positive and negative partial effects. For instance, within the turquoise module, Cadmium (Cd) is positively weighted, while Copper (Cu), Nickel (Ni) and Lead (Pb) are negatively associated with zebrafish behavior. Similarly, within the grey module, Arsenic (As), Uranium (U) and Chromium (Cr) show positive effects, whereas Selenium (Se) and Antimony (Sb) exhibit negative effects. These findings underscore the importance of evaluating overall mixture exposure effects and highlight the critical need to consider complex interactions within chemical mixtures in environmental toxicology.

Keywords: Chemical mixtures, Quantile g-computation analysis, Weighted gene co-expression network analysis

Test for Normality Based on an Edgeworth Expansion

Hapuarachchi K. P.^{1}, Wickramarachchi D. C.^{2} and Peiris K. G. H. S.^{3*}

^{1,2,3}Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

^{1}kphapu18@gmail.com, ^{2}chitraka@sjp.ac.lk, ^{3}hashan_peiris@sfu.ca

^{1}009-0002-1717-6806, ^{2}0000-0001-6958-9667, ^{3}0000-0002-4721-9881

Most statistical tests require the assumption that the observations in a random sample are drawn from a normal distribution. Several statistical techniques, mostly based on either population moments or empirical distribution functions, are currently available to test whether the observations in a random sample are normally distributed. In this study, Edgeworth series expansion was used to model deviations from normality, depending on the skewness and excess kurtosis of a normal population. A test statistic was developed based on these two measures to test for normality. However, the sampling distribution of the test statistic is not mathematically tractable. Therefore, a simulation study was conducted by generating sampling distribution of the proposed test statistic for different sample sizes when the data are normally distributed. The critical values were calculated at different levels of significance as well. The power of the proposed test was empirically compared with the Shapiro-Wilk, Shapiro-Francia, Jarque-Bera, Cramer-von Mises, Lilliefors, Kolmogorov-Smirnov and Anderson-Darling tests. The power is better than that of the Kolmogorov-Smirnov test and compares well against others for small samples with low skewness.

Keywords: Edgeworth expansion, Kurtosis, Normality tests, Power comparison, Skewness

Application of Bayesian Conditional Autoregressive Modeling to Investigate Area-level Behavioral Risk Factors of Hypertension Prevalence

Koralegedara S. H[1*], Chandrabose M.[2] and Dehideniya M. B.[3]

[1,3]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka, [2]Centre for Urban Transitions, Swinburne University of Technology, Melbourne, Australia

[1]s18424@sci.pdn.ac.lk, [2]mchandrabose@swin.edu.au, [3]mahasen@sci.pdn.ac.lk

[1]0009-0002-4281-9666, [2]0000-0002-5311-3020, [3]0000-0002-1798-8591

The prevalence of chronic diseases often shows spatial patterns, with some areas having higher rates of diseases while others have lower rates. It is important to accurately identify the area-level risk factors that explain these spatial patterns. Traditional statistical methods, such as ordinary least square (OLS) regression analysis, are not suitable for these types of data because they may not accurately capture spatial dependence structures. When data involve area units, spatial autocorrelation can occur, meaning nearer observations can be highly correlated than those that are far apart. In this research, an empirical investigation was conducted to identify the area-level behavioral risk factors (rates of alcohol consumption, physical inactivity and smoking) of hypertension prevalence in Sydney, Australia. Secondary data from the Australian National Survey conducted from July 2017 to June 2018 were used. The spatial unit of analysis for this study is the Population Health Area. A Bayesian Conditional Autoregressive (CAR) model was used to estimate the relationships between hypertension prevalence and area-level risk factors, incorporating the spatial dependence structures present in the data. This model was constructed using Gaussian likelihood with non-informative priors for regression coefficients and CAR prior for the spatial random effects, which accounts for spatial dependencies between areas. Estimation of model parameters was done using Markov Chain Monte Carlo (MCMC) simulation. The Geweke diagnostic test (the acceptable range of the Geweke test is -1.96 to 1.96 and all coefficients are within this range) and trace plots were used to check for the convergence of MCMC simulation. For comparison, an OLS regression model was fitted but found violations of key assumptions such as homoscedasticity and independence of residuals. The *CARBayes* (6.2.3) package was used for Bayesian CAR models and *sf* (1.1.8) package was used for additional spatial analyses. Since the estimated regression coefficients (posterior mean) are small in magnitude, they are described for a 10% change in the corresponding predictor variable, while holding all other variables constant. According to the results, on average, when areas have a 10% higher rate of alcohol consumption, the prevalence of hypertension is lower by 1.1% (95% credible interval: [-1.2, -0.6]), assuming all other predictors remain constant. The estimated regression coefficient for the rate of physical inactivity is 0.1 (95% credible interval: [0.1, 0.8]). This means that, on average, when areas have a 10% higher rate of physical inactivity, the prevalence of hypertension increases by 0.1%, assuming all other predictors remain constant. For the rate of smoking, the estimated regression coefficient is not statistically significant, indicating that smoking rates are not associated with hypertension prevalence. Our study shows that Bayesian CAR models better capture spatial complexities, offering a more accurate understanding of relationships in spatially dependent data than traditional statistical models.

Keywords: Bayesian hierarchical model, CARBayes, Chronic diseases, Risk mapping, Spatial regression analysis

A Case Study on the Impacts of Climate Variables on Road Accidental Casualties

Weerasekara S. N.[1*] and Devpura N.[2]

[1]School of Engineering, University of Southern Queensland, Australia, [2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]u1150337@umail.usq.edu.au, [2]ndevpura@sci.sjp.ac.lk

[1]0009-0004-6713-9047, [2]0000-0002-8299-0540

This paper examines the impact of climate variables including average temperature, average minimum temperature, average maximum temperature, rainfall and the Southern Oscillation Index (SOI) on road accident casualties. While extensive research has examined road safety factors, limited studies focus on how specific climatic conditions influence accident rates, especially in sub-tropical climates like Queensland, Australia. Using monthly time series data from January 2001 to December 2022 for Brisbane (postcode 4000), we applied time series regression models to assess climate effects on road casualties. The data were sourced from the Australian government's road crash records and the Bureau of Meteorology. Our findings reveal that colder conditions correlate with an increase in road casualties, with lagged casualty counts emerging as a significant predictor. However, evidence linking other climate variables, such as SOI or rainfall, to road accidents was inconclusive. This study provides insights for civil engineers and policy makers to consider climate-responsive interventions, especially during colder months, to enhance road safety and mitigate accident casualties. This research fills a gap in understanding how climate conditions, beyond typical road safety variables, impact accident rates. Its novel approach lies in integrating climate indices with road accident data, potentially guiding new strategies in road safety under varying climate conditions. This research uniquely identifies July as a peak month for road casualties, highlighting a seasonal factor influenced by colder climate conditions. The findings suggest that lower temperatures and seasonal weather patterns may significantly impact accident rates, likely due to changes in road conditions and driver behavior.

Keywords: Casualties, Climate, Rainfall, Southern oscillation index, Temperature

Comparative Analysis of Predictive Models and Factors Associated with Sleep Quality among Undergraduates of the University of Sri Jayewardenepura

Fernando C. M. T.[1*], Dias P.[2] and Wickramarachchi D. C.[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]thimalifernando11@gmail.com, [2]dias@sjp.ac.lk, [3]chitraka@sjp.ac.lk

[1]0009-0003-3677-3961, [2]0000-0002-5584-2526, [3]0000-0001-6958-9667

Everyone needs sleep. Sleep is the reset switch that refreshes the body and mind of all human beings alive. The quality of one's sleep significantly affects the overall person, impacting not only health and well-being but also the quality of life. Undergraduates, in particular, can be identified as a vulnerable group concerning sleep quality, given the demands of academic work and social activities. Despite its significance, research on undergraduate sleep quality in Sri Lanka has been limited, often focusing on sociological and descriptive aspects. This study aims to address this gap by applying advanced statistical methodologies to identify key factors associated with sleep quality among undergraduates. Further, this study aims to develop a predictive model that can provide insights into the sleep quality of this group, helping to improve their overall well-being. A cross-sectional survey using both web-based and in-person data collection methods has been conducted. Quota sampling was applied to select participants from the chosen faculties of the University of Sri Jayewardenepura. Data on socio-demographic, behavioral, sleep-related and academic-related factors of undergraduates were collected through a questionnaire. The Pittsburgh Sleep Quality Index (PSQI) was utilized to evaluate sleep quality, while the Epworth Sleepiness Scale (ESS) was utilized to evaluate daytime sleepiness. This study employed logistic regression to identify significant factors associated with sleep quality. Following this, logistic regression, random forest, support vector machine (SVM) and ensemble techniques such as bagging and AdaBoost were applied to predict the sleep quality of undergraduates. A comparison of these models was conducted across a range of performance metrics to determine the best predictive model. The results from 526 participants showed that the prevalence of poor sleep quality in the sample was 63%. Nine significant predictors of poor sleep quality were identified from the fitted logistic regression model. Among these, three were protective factors: a higher Current Grade Point Average (CGPA), better concentration and a longer time gap between dinner and sleep. Conversely, six risk factors were identified: being in a higher academic year, being employed compared to having no job, identifying as an evening person rather than a morning person, having a heavier academic workload and experiencing a higher frequency of headaches and neck pain. In predicting sleep quality, the random forest model was selected as the most accurate and balanced, with 83.4% accuracy and 81% precision. The nine identified predictors effectively assess sleep quality among undergraduates, making the model a practical tool for routine assessments due to its brevity. This eliminates the need for completing the lengthy and more complex PSQI questionnaire. To address the high prevalence of poor sleep quality, efforts should focus on promoting protective factors and mitigating risk factors identified in the study. To enhance its impact, implementing this questionnaire is recommended in campus health services to monitor and address sleep quality issues among students.

Keywords: AdaBoost, Bagging, Logistic regression, Random forest, Support vector machine

Consumer Profiling Based on Preferences and Concerns Towards Fresh-cut Vegetables: A Cross-sectional Exploratory Study of Sri Lankan Urban Community

Randika P. G. S.[1], Dharmapriya U. S. S.[2*], Kulathunga A. K.[3], Premarathne H. D. P.[4] and Daundasekara S. S.[5]

[1]Department of Mathematics, University of Peradeniya, Sri Lanka, [2,3]Department of Manufacturing and Industrial Engineering, University of Peradeniya Sri Lanka, [4]Department of Sociology, University of Peradeniya, Sri Lanka, [5]Department of Food Science and Technology, University of Peradeniya, Sri Lanka

[1]sandunir@sci.pdn.ac.lk, [2]subodhad@eng.pdn.ac.lk, [3]aselakk@eng.pdn.ac.lk, [4]pri.hapuarachchi@arts.pdn.ac.lk, [5]saumalid@agri.pdn.ac.lk

[1]0009-0003-1024-5677, [2]0000-0001-6471-0042, [3]0000-0002-9241-3149, [4]0000-0002-7710-986X, [5]0000-0002-9913-1937

Fresh-cut vegetables (FCVs) offer convenience and nutritional benefits, aligning well with the demands of modern, busy lifestyles. While the Sri Lankan market is still emerging, many developed countries have embraced FCVs. This study aimed to identify consumer profiles in urban communities in Sri Lanka based on their preferences and purchasing concerns. A cross-sectional survey was conducted using both face-to-face interviews and online questionnaires, resulting in 1,722 responses. Statistical analysis revealed that convenience is the key driver for purchasing FCVs and the most important purchasing concerns are expiry date, freshness and healthiness. Those purchasing concerns varied across demographic factors of monthly income and education levels, with price being a consistent concern across all income categories. Additionally, there was a strong preference for fruit-type vegetables, such as jackfruit, baby jackfruit and breadfruit, in fresh-cut form. Consumers preferred FCVs treated with natural disinfectants and packaged in environmentally friendly materials. Cluster analysis was performed using K-Means clustering and Hierarchical clustering, resulting in three consumer segments: less concerned, moderately concerned (quality-driven) and highly concerned (safety-driven). Quality-driven consumers are concerned more about information such as shelf life and quality certifications on product labels. In contrast, safety-driven consumers are concerned about details regarding disinfectants and preservation methods. These findings provide valuable insights to promote FCVs as a commercial product in Sri Lanka. In future studies, non-urban communities in Sri Lanka will be explored to generalize the findings.

Keywords: Correspondence analysis, Factor analysis, Fresh-cut vegetables, Hierarchical clustering, K-means clustering

Unlocking the Colombo Stock Exchange: Leveraging Sentiment Analysis, Technical Indicators and Behavioral Economics in Computerized Trading

Senarathna B. S.[1*], Abeynayake A. D. L.[2] and Deshani K. A. D.[3]

[1,3]Department of Statistics, University of Colombo, Sri Lanka, [2]Acuity Knowledge Partners, Sri Lanka

[1]buddhimasenarathna@gmail.com, [2]devmindaabey@gmail.com, [3]deshani@stat.cmb.ac.lk

[1]0009-0009-3585-8270, [2]0009-0005-1454-8952, [3]0000-0002-5489-3436

Emotions can be considered an integral part of behavior and thus in investment decisions, human emotions and sentiments often hold considerable correlation. Despite the impact emotions hold on investor behavior, most of the current automated trading strategies do not accommodate the market sentiment for the decision-making process. In the context of the Colombo Stock Exchange, the lack of research available suggests the integration of sentiments for the rules when making trading decisions on stocks justifies the above claim. This study endeavors to bridge this gap by initiating the development of sentiment-integrated strategies specifically for practical application in algorithmic trading. For the study, historical price and sentiment data were precisely collected for the selected four S&P SL20 companies for a period of 20 years. Price data was extracted directly from the Colombo Stock Exchange (CSE), while news articles were scraped from two major financial websites using Selenium. Sentiment polarity was determined using transformer NLP models. During fundamental analysis, 4 companies were selected to proceed with the study. After testing diverse statistical and machine learning models including Random Forest, Autoregressive Integrated Moving Average (ARIMA) and XGBoost, a Long Short Term Memory (LSTM) architecture proved most effective in forecasting closing prices with nearly a 100% accuracy, utilizing sentiment scores and historical prices as predictors. Leveraging insights from sentiment, technical and forecasting models, three bespoke trading strategies were formulated and backtested. Results revealed superior returns compared to conventional benchmarks after parameter (Equity Final) optimization, highlighting sentiment's significant impact on price dynamics. The implementation of sentiment-infused strategies revealed the potential for strong trading outcome improvements in Sri Lanka's equity market.

Keywords: Algorithmic trading strategies, Colombo stock exchange, LSTM model, Market sentiment-technical integration

CONTRIBUTED ABSTRACTS

POSTER PRESENTATIONS

Time Series Modeling and Forecasting Daily Gold Prices During Sri Lanka's Economic Crisis Period

Bandara S. M. G. S.[1*] and Sudarshani E. G. D.[2]

[1,2]Department of Mathematics, University of Ruhuna, Sri Lanka

[1]gayan@maths.ruh.ac.lk, [2]egdsudarshani@gmail.com

[1]0000-0002-0598-6206, [2]0009-0008-0949-4883

Gold is recognized as one of the most precious commodities globally and serves as a safe-haven asset during periods of economic uncertainty or instability. Further, gold serves as a hedge against asset fluctuations during inflation, deflation, exchange rate changes and interest rate changes and can significantly increase the time value of money. It is evident that investors are more interested in buying gold as an investment, especially during such an economic crisis. Sri Lanka is also facing an economic crisis from 2022 and investors tend to buy gold more than other assets during this period. Therefore, it is better to approach a new model for gold price forecasting based on daily gold price data in this economic crisis period. This study aims to develop the optimal Autoregressive Integrated Moving Average (ARIMA) model for forecasting the daily gold prices in Sri Lanka under certain circumstances following the start of the economic crisis and analyze the relationship between the gold price and USD/LKR exchange rate. In order to construct the model and the analysis, daily data of the gold prices per troy ounce and USD/LKR exchange rate data from April 2022 to October 2023 were employed. Various ARIMA models were taken into consideration. Based on the minimum value of the corrected Akaike Information Criteria (AICc), the ARIMA(5, 1, 4) is chosen as the best model for forecasting gold demand under volatility during the economic crisis era. The fitted model was validated using the Mean Absolute Percentage Error (MAPE) value, which came out at 3.45% (< 10%). During this period of economic crisis, there is a huge depreciation and slight appreciation of the rupee against the US dollar. The correlation coefficient value between the gold price and the USD/LKR exchange rate was 0.55 (p-value = 2.2e-16 < 0.05). Accordingly, a moderately positive linear relationship is shown between the price of gold and the USD/LKR exchange rate.

Keywords: Autoregressive integrated moving average model, Economic crisis, Forecasting, Gold price, Mean absolute percentage error

Statistical Investigation of the Impact of Aggregate Shape, Size and Aggregate-to-cement Ratio on Porosity of Pervious Concrete

Dilaxsan S.[1*], Sathiparan N.[2] and Subramaniam D. N.[3]

[1,2,3]Department of Civil Engineering, University of Jaffna, Sri Lanka

[1]2020e032@eng.jfn.ac.lk, [2]sakthi@eng.jfn.ac.lk, [3]Daniel.subramaniam@gmail.com

[1]0009-0008-4997-6588, [2]0000-0001-8570-0580, [3]0000-0002-1023-1617

Pervious concrete is primarily designed for use as a porous paving material in urban areas, offering additional benefits like noise reduction and heat insulation. It mainly consists of coarse aggregates and a binder (cement), with a small proportion of fine aggregates (around 5%). The induced pores enhance porosity and permeability but reduce overall strength. Key casting conditions for previous concrete include zero slumps and zero compaction (no workability) to prevent binder migration that could block the surface. Unlike conventional concrete, which is continuous, pervious concrete is modeled as a configuration of binder-coated aggregates. This unique packing leads to variability in porosity. While compaction can improve the homogeneity of properties, the effects of aggregate size on porosity have been studied, but the impact of aggregate shape remains unexplored. This study intends to evaluate how aggregate shape, altered through milling, affects the properties of pervious concrete. The investigation into pervious concrete samples utilized ordinary Portland cement, crushed aggregates and water sourced locally. Aggregates were tested for various properties, which ensured they met the minimum standards for concrete production. The mixing process combined aggregates, cement and water in an electric drum, varying aggregate-to-cement (AC) ratios of 3, 4 and 5, with a constant water-to-cement ratio of 0.3. Three aggregate sizes (5 – 12 mm, 12 – 18 mm and 18 – 25 mm) and five milling degrees (0, 50, 100, 200 and 1000 revolutions) were tested. A total of 810 specimens were produced across 135 mix designs, each with six replicates. After casting, the cubes were air-dried for 24 hours and then immersed in water for 28 days of curing. The porosity of each specimen was calculated based on the constituent ratios and the weight of the fresh concrete. The study examined the effects of aggregate characteristics—aggregate-to-cement ratio, aggregate size (ASize) and aggregate shape (AShape)—on the theoretical porosity (T.Poro) of previous concrete. Theoretical porosity showed significant skewness and kurtosis, indicating the influence of these parameters, with distributions symmetric around the mean for all groups. The mean T.Poro for samples with 50 and 1000 revolutions of milling differed notably from others, suggesting a potential impact of aggregate shape. Three-way ANOVA revealed a significant effect of AC, aggregate size and aggregate shape on T.Poro, with an F-test value of 298.73 and individual F-test values were 298.793, 39.226 and 250.824 with a significance below 0.001 and effect sizes of 0.726, 0.259 and 0.817 respectively, indicating strong statistical confidence. The impact of interaction of aggregate size and aggregate size shape on T.Poro yielded an F-value of 55.945 and effect size of 0.657 (highest of the interaction) with a statistical 95% confidence. The post hoc test revealed that there is a statistically significant increment in porosity with an increase in AC ratio and the porosity decreased significantly from small to medium aggregate sizes, but no significant difference was observed between medium and large aggregate sizes.

Keywords: Aggregate shape, Milling, Pervious concrete, Theoretical porosity

Non-parametric Evaluation of Assessment Formats and Their Impact on Academic Achievement: A Study on University Students Sample

Kapuliyadda K. T. M. A. N.[1*], Peiris W. B. T.[2], Amarasinghe S. L. T.[3], Suvineetha K. K. S.[4] and Samaraweera N. P. I. H.[5]

[1,2,3,4,5] Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Sri Lanka

[1]amandakapuliyadda@gmail.com, [2]buddhinitharushika412@gmail.com, [3]lakshikatharushi98@gmail.com, [4]sasanisuvineetha623@gmail.com, [5]Ishaniharshika9907@gmail.com

[1]0009-0000-4176-0055, [2]0009-0001-7897-4759, [3]0009-0009-1405-9234, [4]0009-0007-9574-7228, [5]0009-0006-7861-3749

This research seeks to establish the impact of assessment types on university students' academic performance in a single university. The purpose of the study is to investigate whether there is a meaningful difference between three different types of assessments, using non-parametric techniques. The study is equally significant to point out that, rather than expecting an overall disparity in the performances which should reflect the delivery formats, none was noticeable in the results. Nevertheless, using estimates from descriptive analyses, mean values indicate that there are formats that are more productive when it comes to learning. To begin with, the introduction section will present the key issues on the effectiveness of the mainstream and non-mainstream techniques of assessment in higher learning institutions. Thus, to improve student outcomes, the new curriculum gives information about the correlation between assessment practices and learning results. However, the generalizability of the results is restricted to the participants and cannot be extended to all the Undergraduates in the country; but it is not devoid of utility to the educators and policy makers. The findings indicate that the method of instruction might not affect academic achievement significantly, but some forms may benefit student learning in a better way. Therefore, this research sought to establish whether the variety of assessment methods has a statistically significant effect on the enhancement of university students' learning and performance regarding the research questions employing non-parametric and descriptive analysis. Importantly, the role of finding proper forms of assessment for enhancing educational outcomes is depicted in the study.

Keywords: Academic performance, Evaluation types, Learners in universities, Non-parametric tests

Impact Characterization of Promotional Leverage in Forex Trading using Ornstein - Uhlenbeck (OU) Process and Monte-carlo Simulation

Kaushalya H. A. T.[1*] and Liyanage U. P.[2]

[1,2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[1]thilinihettiarachchi9@gmail.com, [2]prabhath.uoc@gmail.com

[1]0009-0002-7154-6150, [2]0000-0003-1201-3418

Promotional leverage is a standard practice exercised by many brokers in forex trading. However, determining the optimal use of leverage for traders or brokers is challenging due to market uncertainties. Trade leverage offers small-scale traders opportunities for good returns when used strategically. Forex trading benefits from market expertise and advanced strategies using one-to-many leverage scales, enhancing the potential for positive returns. This study aims to analyze the impact of promotional leverage on trading outcomes from both brokers' and traders' perspectives. The Ornstein-Uhlenbeck (OU) process, commonly used for exchange rate modeling, was recommended in some literature and it was applied to simulate leverage effects in forex trading. Globally, the most active currency pair, EUR/USD, was used to assess leverage impacts the simulation was based on the mean and standard deviation of hourly market close prices over a 763-hour period. This study selected four broker companies based on the criteria: best overall, best for professional traders, best for low spreads and best for beginners. From that, six leverage levels - 1:20, 1:500, 1:1000and 1:2000, were selected for the analysis. Monte Carlo simulations, repeated 1000 times, were used to analyze the impact of leverage on trader returns. The results demonstrated that increasing leverage from 1:20 to 1:2000 can significantly increase possible profits for traders. According to the simulation results, at the leverage of 1:20, the expected profit increases to 3.1940×10^{-1} EUR (Min = 4.7741×10^{-2} EUR, Max = 1.4056 EUR). At the leverage of 1:500, the profit increases to 7.9851 EUR (Min = 9.8986×10^{-1} EUR, Max = 2.4402×10^{1} EUR), whereas, at the leverage of 1:1000, the profit increases to 1.5970×10^{1} EUR (Min =1.9797 EUR, Max = 7.0278×10^{1} EUR). Level of leverage 1:2000 results in 3.1940×10^{1} EUR of expected profit (Min = 3.9594 EUR, Max = 1.4056×10^{2} EUR). However, in the absence of any leverage the expected profit would be 1.5970×10^{-1} EUR. Leverage can significantly increase trade profits, but it also increases the risk of large losses. This analysis focused on trading values and excluded additional costs such as commissions, swap fees and withdrawal fees which can be integrated for a more accurate profit estimate.

Keywords: Monte-carlo simulation, OU process, Profit, Promotional leverage, Trading

A Novel Bagging Ensemble of Oblique Trees for Data Classification

Karunasena K. A. D. B. S.[1*], Wickramarachchi D. C.[2], Robertson B.[3], Reale M.[4]
and Price. C.[5]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka, [3,4,5]Faculty of Engineering, University of Canterbury, New Zealand

[1]buddhik@sjp.ac.lk , [2]chitraka@sjp.ac.lk , [3]blair.robertson@canterbury.ac.nz,
[4]marco.reale@canterbury.ac.nz , [5]chrisj.price@canterbury.ac.nz

[1]0009-0009-3381-192X, [2]0000-0001-6958-9667, [3]0000-0002-2077-907X, [4]0000-0002-4831-0337,
[5]0000-0001-6776-0037

Ensemble learning is a powerful data mining approach that combines multiple base learners to create an optimal predictive model. Improved accuracy, reduced overfitting and robustness to noise are some of the advantages highlighted in this method. Among the popular ensemble learning methods such as bagging, random forest and stacking, this study proposes a novel bagging approach named HHBag. A bagging model is expected to perform efficiently when the base learner within the ensemble is also efficient. The HouseHolder Classification and Regression Tree (HHCART) oblique decision tree serves as the base learner incorporated in HHBag model. HHCART is based on the Householder reflection method where axis parallel splits are searched on reflected space. Hence, it is an efficient heuristic algorithm for data classification when the class boundaries are oblique to the original feature axes. Consequently, the proposed HHBag algorithm employs bagging to improve prediction accuracy by generating multiple HHCART oblique decision trees from bootstrap samples. In the proposed algorithm, bootstrap samples are drawn from the training set and unpruned HHCART decision trees are generated from each sample. Then the final prediction is made by aggregating the predictions of these decision trees through majority voting. The size of the ensemble was determined by hyperparameter tuning, with the optimal number of trees chosen based on where the prediction error stabilizes. Five-fold cross-validation is employed to assess the predictive performance of HHBag. The proposed HHBag algorithm is tested for data classification on a few datasets in the UCI Machine Learning Repository (Iris, Balance Scale and BUPA datasets), all of which contain numerical features. The results are compared with those of traditional bagging (CARTBag), single CART and single HHCART models. The results indicate that HHCART outperforms CART on these datasets, demonstrating that oblique decision trees are more effective than axis-parallel decision trees for this type of data. Additionally, the ensemble methods (HHBag and CARTBag) achieved higher average accuracies across all datasets compared to individual decision trees. This indicates that ensemble methods can outperform a single decision tree model. Furthermore, when comparing the two bagging methods, the proposed HHBag algorithm yields more accurate results. In comparison with CARTBag, HHBag also requires a small ensemble size to stabilize the prediction error. Considering all these factors, it can be concluded that the proposed bagging approach is a promising ensemble method for classification problems, offering improved predictive performance.

Keywords: Classification, Ensemble learning, HHBag, HHCART

Multi factorial Survival Analysis on the Effect of Lymphovascular Invasion in Colorectal Cancer

Nanayakkara N. G. P. M[1], Rahumath M. C. R[2*], Ediriweera D. S[3], Mahendra B. A. G. G.[4], Hewavisenthi S. J. De S.[5], Kumarage S. K.[6] and Chandrasinghe P. C.[7]

[1,2]Postgraduate Institute of Medicine, University of Colombo, Sri Lanka, [3]Health Data Science Unit, University of Kelaniya, Sri Lanka, [4,5]Department of Pathology, University of Kelaniya, Sri Lanka, [6,7]Department of Surgery, University of Kelaniya, Sri Lanka

[1]m41740@pgim.cmb.ac.lk, [2]m38883@pgim.cmb.ac.lk, [3]dileepa@kln.ac.lk, [4]baggmahendra@kln.ac.lk, [5]janaki13@kln.ac.lk, [6]sumuduk@kln.ac.lk, [7]pramodh@kln.ac.lk

[1]0009-0007-9075-0241, [2]0009-0000-7333-2290, [3]0000-0001-5679-2893, [4]0000-0001-9724-6944, [5]0000-0002-5890-1288, [6]0000-0002-5919-4087, [7]0000-0002-3485-961X

Lymphovascular invasion (LVI) in colorectal cancer (CRC) denotes the systemic spread of cancer and is crucial for determining adjuvant chemotherapy for node-negative disease, highlighting its significance in patient care. This study aimed to assess the effect of LVI as an independent factor on survival in CRC, addressing the scarcity of literature on this subject in South Asian populations. A retrospective cohort study was conducted on CRC patients who underwent surgical resection with curative intent between January 2013 to May 2018, with a minimum follow-up of 5 years. Data on patients' LVI status, age, gender, family history of cancer and follow-up time from the date of surgery to the last follow-up date or date of death were collected. Pearson's Chi-square test was used to compare the LVI-positive and LVI-negative groups in terms of gender, family history of CRC and family history of other cancers. Right-censored survival data were analyzed. Survival rates were estimated using Kaplan-Meier curves and compared with the Log-rank test. Cox proportional hazards models were employed to assess factors associated with survival, rather than for prediction purposes. Forward variable selection and log-likelihood ratio tests were used to refine the model. The final model was tested for proportional hazard assumptions using the Schoenfeld test. Outliers were assessed using deviance residuals and beta values. The results showed that 62 (42.76%) of the 145 participants were male. The median follow-up time was 5 years and 4 months (IQR: 1 year 7 months - 7 years 10 months). There were 36 (52.17%) patients having LVI and 69 (47.59%) died during follow-up. LVI positive and negative groups were similar in terms of gender (p-value = 0.7278), Family history of CRC (p-value = 0.1947) and Family history of other cancers (p-value = 0.3493). The overall 5-year survival rate was 57.9% (95% CI=50.2%-66.8%). LVI significantly impacted overall survival (Log-rank test, p-value < 0.0001). Univariate analysis reported a hazard ratio (HR) of 2.937 (95% CI =1.816-4.75, p-value < 0.0001) for LVI. In the adjusted final model, LVI was associated with poor survival (HR = 2.786; 95% CI=1.719-4.514, p-value < 0.0001), alongside increasing age (HR = 1.031 per year increase, 95% CI =1.009-1.054, p-value = 0.0064). The concordance of the model was 0.672 with a standard error of 0.031. The proportional hazard assumption was not violated and no outliers were identified. No underlying non-linear patterns were noted. In conclusion, LVI is an independent risk factor for survival in CRC, with the hazard ratio underscoring its clinical importance, consistent with observations in Western literature. This highlights the need for a thorough histological assessment of LVI in CRC patients to inform treatment decisions. Promoting data awareness in the medical community and encouraging practices that ensure data quality should be prioritized for better evidence-based care.

Keywords: Colorectal cancer, Cox proportional hazard model, Lymphovascular invasion, Survival analysis

Comparative Analysis of Oblique Decision Trees and Variants of Linear Discriminant Function: A Study Based on Accuracy and Weighted F-Score

Nawagamuwa H.[1*] and Wickramarachchi D. C.[2]

[1,2]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[1]nawagamuwahansani@gmail.com, [2]chitraka@sjp.ac.lk

[1]0009-0006-0290-5657, [2]0000-0001-6958-9667

Classification is a key supervised machine learning method, where the model is trained on the observed data to predict the correct class label of a given input. While numerous classification methods exist, their effectiveness often varies across different contexts, highlighting the need for comparative studies to guide algorithm selection in diverse applications. Among various classification methods, Linear Discriminant Analysis (LDA), stands out for its simplicity and effectiveness, providing a closed-form solution under the assumptions of multivariate normality and equal covariance matrices. Recent advancements have introduced generalized versions of LDA offer enhanced flexibility in model assumptions. On the other hand, decision trees employ iterative splitting based on features, allowing them to capture complex patterns in the data without relying on specific assumptions. Oblique decision trees, in particular, produce more compact trees with improved accuracy compared to axis-parallel trees. Previous comparative studies have primarily examined the performance of generalized LDA forms and older oblique decision tree algorithms, such as OC1, SADT, FACTand QUEST, revealing a research gap regarding newer oblique decision tree algorithms, including HHCART and Geometric Decision Tree. The inclusion of these newer algorithms could advance comparative study in classification methods. This study aimed to conduct a comparative study between LDA and its generalized forms and Oblique Decision Trees across a range of distinct datasets and identify the most suitable classification methods based on varying data characteristics. Flexible Discriminant Analysis (FDA), Penalized Discriminant Analysis (PDA) and Mixture Discriminant Analysis (MDA) were considered under the variants of generalized LDA. HHCART(A), OC1, QUESTand Geometric DT were considered under the Oblique Decision Tree algorithms. The objective involved examining algorithm performance across two distinct data scenarios: only quantitative features and both qualitative and quantitative features. Datasets were sourced from the UCI and KEEL datasets repository, ensuring a diverse dataset varying in the number of classes, observations and attributes. The initial dataset evaluation involved the simple measurements of the datasets, including the number of observations, attributes and classes, alongside statistical assessments such as covariance homogeneity and the assessment of multivariate normality for each class. The performance of the classifiers was evaluated using repeated random sub-sampling validation for each dataset with accuracy and weighted F-score due to class imbalance. The weights for each class in calculating the weighted F-score were determined based on the class distribution within the dataset. To assess which classification method performs better in general across various datasets, the Friedman test was employed. The forms of FDA achieved high performance across the datasets containing only quantitative features, irrespective of the number of observations in the dataset. Conversely, the QUEST algorithm outperformed the generalized forms of LDA on most datasets containing both qualitative and quantitative attributes. Furthermore, the study’s findings reveal that GDT and OC1 have weak performances across the binary class datasets having only quantitative features. The key limitation of the study is the prolonged computational time required by the HHCART algorithm when working with large datasets due to suboptimal code optimization.

Keywords: Classification, Generalized linear discriminant analysis, Linear discriminant analysis, Oblique decision trees, Weighted F-score

A Novel Approach to Classify Projects Applying Value Chain Interventions: A Study using Agriculture-related Development Project Reports

Prabhath K. A. K.[1*], Tillekeratne L. N. T.[2], Hettiarachchi D. I.[3], Silva T. de[4] and Dassanayake S. M.[5]

[1,2,3,4,5]Department of Decision Science, University of Moratuwa, Sri Lanka

[1]kaveesha.18@business.mrt.ac.lk, [2]nisali.18@business.mrt.ac.lk, [3]duvindu.18@business.mrt.ac.lk
[4]tiloka.desilva@gmail.com, [5]sandund@uom.lk

[1]0009-0000-9590-4333, [2]0009-0003-4239-8926, [3]0009-0008-7439-0001, [4]0000-0002-0543-2182, [5]0000-0001-6809-5016

Modern natural language processing (NLP) techniques allow users to analyze textual data better. Past project reports that summarize the outcomes, interventions and failures of development projects create a significant, yet untapped, dataset that can be effectively used for future project planning. Development practitioners evaluate these reports by manually extracting excerpts, classifying them into interventions and assessing their effects on project outcomes. This process is time-consuming and labor-intensive. This research tested and presented a novel approach based on NLP and supervised learning to identify projects applying Value Chain Interventions (VCI). The model proposed in this study classifies each given excerpt as either a 'Value Chain Intervention' or a 'Non-Value Chain Intervention'. Specifically, the study utilizes a dataset consisting of 1,438 text excerpts obtained from the executive summaries of development project reports to automate this classification. Of these excerpts, 719 are Value Chain Intervention excerpts, while the remaining 719 are Non-Value Chain Intervention excerpts that were collected using the random sampling method. In the first phase, exploratory data analysis using word clouds, scatterplot and a word dictionary highlighted a clear distinction in token distribution among the two categories. The researchers then selected the Term Frequency-Inverse Document Frequency (TF-IDF) score-based model as the preferred method for keyword extraction after comparing it with the Bidirectional Encoder Representations from Transformers (BERT) and Rapid Automatic Keyword Extraction (RAKE), which are algorithm-based keyword extraction techniques, on the basis that the extracted key phrases from the latter two models often possessed little semantic significance. The keyword extraction exercise was then used to vectorize the data set that was fed into the classification model. The study used and compared linear Support Vector Machine (SVM), logistic regression and Naive Bayes as alternative models for classification purposes. given that these models can handle high dimensional data and small sample complexity. Furthermore, the research tested an ensemble model to further improve the classification accuracy. Through the analysis of the data and the accuracies obtained for each model, it was determined that the model constructed utilizing the Naive Bayes algorithm exhibited the highest accuracy of 77% in identifying the value chain interventions within the provided excerpts. The research provides a framework for the development of a more broad-based model that would enable development practitioners to automatically extract necessary information in future evaluations, allowing them to easily identify projects applying Value Chain Intervention. This would help to enhance the design and implementation of future development projects with VCI.

Keywords: Classification models, Development projects, Keyword extraction, NLP, Text analytics

Author Index

Sponsors

Bronze Sponsors

almas EQUITIES

ALGONOMY

CIC Nurturing Life

aura insights

Financial Sponsors

WITCO

LANKEM ලංකෙම් லங்கெம்

POLYPAK

Industry Collaborator

MAS

OCTAVE

John Keells Group

Publicity Supporters

Statistical Society of Australia

Souvenir Sponsor

Prym INTIMATES GROUP

Lipton

ISC 2024 Case Study Competition Sponsors

International Water Management Institute (IWMI)

RETINA Project - University of Moratuwa, Sri Lanka funded by OWSD-UNESCO and the International Development Research Centre, Ottawa, Canada

Printing of this publication is sponsored by

NATIONAL
SCIENCE
FOUNDATION