Australian Data Science Network Conference 2022

CONFERENCE PROCEEDINGS

21 & 22 NOVEMBER

Hosted by
QUT Centre for Data Science
Brisbane QLD

AUSTRALIAN DATA SCIENCE NETWORK

Welcome

Welcome to the inaugural conference of the Australian Data Science (ADSN).

Data Science focuses on intelligent extraction of information from data. It is an inclusive term that includes at its core the development of mathematical, statistical and computer science tools, data infrastructure (e-research), data management, responsible data science (data governance, sovereignty) and human engagement with data science. The Data Science umbrella also includes the application of these devices to address challenging issues in health, business, engineering, environment, social science, sport, and so on and on.

Data Science is one of the fastest growing fields both nationally and internationally. Government and corporate entities are now aware of the opportunities that data science can provide, the critical demand for skilled data sciences, and the need for both fundamental and applied research in the above areas. This is evidenced by the proliferation of university and corporate data science courses and research centres across Australia.

As a country, we are now at a crossroads: we can either continue as individual entities in data science or link these entities. The aim of the ADSN is to connect the concentrations of expertise in data science across Australia, in order to improve communication, encourage collaboration, expand opportunities, promote our individual and collective capabilities, and grow the profile of Australian Data Science both nationally and internationally.

This inaugural conference of the ADSN brings together members from our partner organisations to build collaborations across the network, share our efforts and achievements, reflect on the opportunities and constraints of Data Science in Australia, and plan a collective strategy to enhance Australian Data Science in 2023 and beyond.

The QUT Centre for Data Science is the host organisation for this year's event. The conference will be held at QUT Gardens Point Campus in Brisbane. It is anticipated that future conferences will be hosted by other nodes of the Network.

On behalf of the members of the QUT Centre for Data Science, and the members of the Australian Data Science Network, I would like to welcome you to this event. May you be inspired, learn and expand your network in this very exciting field of Data Science.

Kind regards
Kerrie

Kerrie Mengersen
Distinguished Professor of Statistics
Director, Centre for Data Science
QUT Australia
18th November, 2022

Table of Contents

Conference Program

Day 1 - Monday 21st November

Time	Duration	Item
8:45am	30 mins	Coffee/sign In
9:15am	15 mins	Uncle Cheg – Welcome to Country
9:30am	45 mins	Keynote - Tomasz Bednarz, NVIDIA
10:15am	45 mins	Welcome by Kerrie and Meet and Greet
11:00am	30 mins	Break
11:30am	30 mins	Invited talk - Yoni Nazarathy, The University of Queensland
12:00pm	30 mins	Invited talk - Astrid Zeman, Melbourne Centre for Data Science
12:30pm	1 hour	Lunch Break
1:30pm	45 mins	Communications Workshop - Tim Macuga, ADSN
2:15pm	30 mins	Invited talk - Emi Tanaka, Monash University Econometrics & Business Statistics
2:45pm	30 mins	Invited talk - Becki Cook, QUT Centre for Data Science
3:15pm	30 mins	Break
3:45pm	1 hour	Industry Panel - "Reimagining Data Science"
4:45pm	15 mins	Closing
5:00pm	1 hour	Posters & Networking until 6pm

Day 2 - Tuesday 22nd November

Time	Duration	Item
8:30am	30 mins	Coffee/sign In
9:00am	15 mins	Kerrie Mengersen Open & Welcome to Country Video
9:15am	45 mins	Keynote - Richard Fox, AFL Data & Analytics
10:00am	30 mins	Meet and Greet Activity
10:30am	30 mins	Invited talk - Joanna Batstone, Monash Data Futures Institute
11:00am	30 mins	Break
11:30am	30 mins	Response to Science Academy Report on Advancing Data-intensive research
12:00pm	30 mins	Invited talk - Ides Wong, CSIRO
12:30pm	1 hour	Lunch Break
1:30pm	75 mins	ADSN Profile - planning
2:45pm	30 mins	Invited talk - Lewis Mitchell, University of Adelaide
3:15pm	30 mins	Break
3:45pm	30 mins	Invited talk - Andrew Zammit Mangion, University of Wollongong
4:15pm	30 mins	Invited talk - Richi Nayak, QUT Centre for Data Science
4:45pm	15 mins	Closing
5:00pm	1 hour	Posters & Networking until 6pm

Speakers

Keynotes

Day 1: Tomasz Bednarz, Director of Strategic Researcher Engagement at NVIDIA
Day 2: Richard Fox, Data & Analytics Manager for the AFL (Australian Football League)

Invited Talks

Julia as a data science research tool

A/Prof Yoni Nazarathy

The University of Queensland

Data Science practice and research thrives on open source computer programming languages and their supporting ecosystems. When taking the statistical viewpoint, the R language is probably the most common choice, and when focusing on machine learning, Python certainly rules. However, most computationally demanding packages for these languages use other machinery under the hood, often written in C or Fortran. This prevailing "multi-language" approach is fine for high level end users, yet it poses a serious entry barrier for research and innovation. It requires those developing and deploying new computationally intensive data science ideas to have expertise not only in data science domains, but also in low level software engineering.

The emerging Julia language and its ecosystem aims to overcome this barrier by leveraging on several 21st century software technologies and ideas. Julia "feels like" a combination of Matlab, Python, and R to the end user, yet runs incredibly fast. Since its birth, about a decade ago, Julia has attracted practitioners and researchers from applied maths, operations research, machine learning, statistics, and general data science domains, and by now it hosts a rich eco-system useful both for applied high level data science analysis, and for quick transfer of new research ideas from pen and paper to software. In this talk we discuss the Julia data science eco-system and highlight the pros and cons of using Julia as data science research tool.

Towards a unified language in experimental designs

Dr Emi Tanaka,

Monash University Econometrics & Business Statistics

Experimental data are hallmarks of scientific evidence to prove or disprove theories or hypotheses. Multiple people with different expertise are typically involved in planning and executing experiments but rarely is the communication easy or seamless, especially across people from different domains, yet we predicate on the assumption that misapprehensions will be somehow sorted out. This assumption leaves the success of an experiment at the mercy of the interpersonal communication skills of people involved. Rather than leaving the success of an experiment to serendipity, I propose a novel framework to robustify the workflow of the construction of experimental designs that encourages users to deliberate on understanding the experimental structure. This framework, called "the grammar of experimental designs", considers an object oriented system to encapsulate the experimental structure in a cognitive programming approach. I demonstrate this approach using the R packages, edibble and deggust.

Analysing images using Deep Convolutional Neural Networks to supplement human decision making

Dr Astrid Zeman

Melbourne Centre for Data Science

Recognising and classifying objects within images is generally a straightforward task for Deep Convolutional Neural Networks (DCNNs), with their performance exceeding humans in benchmark competition datasets since 2015. Their impressive performance on object images allow for relatively simple integration into automated systems. An open question is how well do these networks deal with more challenging datasets, such as those containing medical images, which would require a level of visual expertise in human observers? I describe a case study in collaboration with the UZLeuven hospital in Belgium, where we analysed a dataset of over 30,000 microscopy images of fertilised human embryos. To date, DCNNS have assisted in classifying embryos as early as day 5 after insemination. We investigated whether DCNNs could successfully predict the destiny of each embryo (discard or transfer) at an even earlier stage, namely at day 3. We first assessed whether the destiny of each embryo could be derived from technician scores, to examine whether the ratings that technicians gave to images correlated with the decision made. We then explored whether a DCNN could make accurate predictions using images alone. We found that a simple 8-layer network was able to achieve 75.24% accuracy of destiny prediction, outperforming deeper, state-of-the-art models. Importantly, when analysing cases of transferred embryos, we found that our lean, DCNN predictions were correlated (0.65) with clinical outcomes. I describe some of the known shortcomings of DCNNs compared to human observers, which is especially relevant when integrating this technology within a clinical context for making medical decisions.

Indigenous Data Science: co-designing a framework for best practice enhancement of data literacy

Becki Cook

QUT Centre for Data Science

This session will explore considerations surrounding Indigenous Research, in particular developing Indigenous research projects, undertaking research with Aboriginal and Torres Strait Islander Peoples, Indigenous data sovereignty and data governance. This will be demonstrated though discussing how researchers in the QUT Centre for Data Science are engaging with the Aboriginal and Torres Strait Islander Community Heath Service Brisbane to co-design a framework for best practice enhancement of data literacy.

AI and data science for social good

Prof Joanna Batstone

Monash Data Futures Institute

The Monash Data Futures Institute brings together leading cross-disciplinary expertise, international partnerships and a large affiliate network to address future technologies, social partnerships and advanced applications. The Institute's ifocus areas include using data-driven AI to enhance governance and policy, sustainable development, climate change, health sciences and thriving communities. This talk will include examples of the AI and data science work underway at Monash in the context of Australia's opportunity for AI leadership around social change.

Healthcare, tech-celerated: using data to power an open and equitable health system

Dr Ides Wong

CSIRO

Excellence is never an accident. It represents the wise choice of many alternatives". An excellent healthcare system hinges on using timely and good-quality data to support decision making, at both the clinician-patient level and across systems level.

Our pandemic response highlighted the use of data to support timely and high-quality decisions across all levels of individual-behaviour, clinical operation, system administration and health and intergovernmental policy. Enabled by data and technology, health and medical services, research and industry partners worked together and created a more open and equitable health system.

The pandemic has "tech-celerated" the healthcare sector in developing and implementing innovative breakthroughs there were previously considered fantastical. Iterative improvement and innovation are essential to underpin the safe and effective evolution of our healthcare system to meet a rapidly changing healthcare environment. At the CSIRO Australia e-Heath Research Centre, we undertake research and develop technologies across the full spectrum from genome sequencing to systems-level analytics, that: 1) transform health systems with data and artificial intelligence, 2) transform healthcare delivery with virtual care; 3) improve health system efficiency and readiness with digital health; and 4) speed the transition to precision health.

With our existing conventions molten by the needs of a global pandemic, we have a unique opportunity to create a new conversation around the future of human health and incorporate and build upon these ideas.

Data science and mathematical modelling

Prof Lewis Mitchell

The University of Adelaide

Data science has become one of the "buzzwords" of the past 10 years, in both academic and industry contexts. However, it involves many of the same core skills once associated with mathematical modelling: real-world applications, computation, data analysis, and importantly, assumptions-based modelling. Does data science present an existential threat to mathematical modellers? In this talk I'll attempt to define data science, and discuss its interconnections with mathematical modelling, illustrating with examples from my own research. Far from being a threat, I will argue that data science and mathematics (both pure and applied) have many synergies, and that the two disciplines can work together and interact for mutual benefit.

Deep learning for facilitating parameter estimation in statistical models

A/Prof Andrew Zammit Mangion

University of Wollongong

Parameter estimation is often the computational bottleneck in analyses involving intractable statistical models. In the first part of the talk I will show how deep learning models trained to be Bayes estimators can alleviate this computational burden. The trained "neural Bayes estimators" yield optimal parameter estimates from data at a fraction of the computational cost typically associated with parameter estimation. In the second part of the talk I show how permutation-invariant neural networks are ideal for being trained as Bayes estimators when the data are exchangeable. In experiments involving multiple replicates and spatial models of extremes, I show that these permutation-invariant neural Bayes estimators considerably outperform other neural-network-based estimators that do not account for replication appropriately in their network design, and that they are highly competitive and much faster than traditional likelihood-based estimators. The work is joint work with Matthew Sainsbury-Dale (University of Wollongong) and Raphael Huser (KAUST).

Neural Natural Language Processing Methods with added Context

Prof Richi Nayak

QUT Centre for Data Science

In this talk, I will present novel methods of deep learning models with added contexts to deal with the text data for natural language processing tasks. I will present an Informed Machine Learning model for sentiment mining with prior information. I will also show how topic modelling of text data can be improvised by utilising visual information with a deep learning model.

Industry Panel (Day 1)

- Dr Matt Aburn, WearOptimo
- Emma Black, Black Box Co
- Prof Mark Harvey, QUT VP of Business Development
- Suzy Lynch-Watson, Metso Outotec
- Dr Iain McCowan, Dubber AI

Moderator: Prof Michael Rosemann, Director of QUT's Centre for Future Enterprise

Extended Abstracts

Forecast and optimise with machine learning: applications in wastewater treatment

Matthew Colwell [1], Mahdi Abolghasemi [2]

School of Mathematics and Physics, The University of Queensland

Email: [1] matthew.colwell@uq.edu.au, [2] m.abolghasemi@uq.edu.au

Introduction

Prediction and optimisation are two widely-used techniques that have found many applications in solving real-world problems. While prediction is concerned with estimating the unknown future values of a variable, optimisation is concerned with optimising the decision given all the available data. These methods are used together to solve problems for sequential decision-making where often we need to predict the future values of variables and then use them for determining the optimal decisions. This paradigm is known as "forecast and optimise" and has numerous applications, e.g., forecast demand for a product and then optimise inventory, forecast energy demand and schedule generations, forecast demand for a service and schedule staff, to name a few. In this extended abstract, we review one such model developed and applied in wastewater treatment. While the current study is tailored to the case study problem, the underlying principles can be used to solve similar problems in other domains [1,2].

Purpose

This paper presents a novel methodology applied to a sequential decision-making problem with competing priorities and differing levels of importance.

Case study

Urban Utilities (UU) provides wastewater treatment services to the south-east Queensland region. Over the past year, we have been working with Urban Utilities' (UU) Resource Recovery team whose responsibilities concern the safe and efficient operation of the company's wastewater treatment plant (WWTP) infrastructure. We have implemented a project with this team help operate their wastewater treatment equipment at lower cost, by applying a data-driven approach.

The wastewater treatment process generates solid waste, typically referred to as biosolids [4]. At Urban Utilities, biosolids are stabilised by specialised reactors at their Oxley WWTP. This equipment is colloquially referred to by its manufacturer's name, Cambi. The End of Waste Code (part of the *Waste Reduction and Recycling Act, 2011*) requires that biosolids generated from wastewater treatment, which are a biological hazard, are treated prior to reuse or disposal [8]. Cambi achieves the necessary treatment of biosolids by thermal hydrolysis (destruction by reaction with water) [5]. Thermal hydrolysis sterilises the biosolids and produces a stable product which is suitable for agricultural reuse. However, thermal hydrolysis involves a lot of energy over a brief period (up to an hour), after which the biosolids are typically left in digestion for several weeks [6].

Cambi is currently operated based on heuristics (and the intuition of the plant operators at Oxley) and therein laid the opportunity for this project. The Cambi equipment is complex and subject to several logistical constraints: operation of the system is viewed more as an art than a science, with varying opinions about how to run it. Running Cambi is one of UU's single largest operational costs, so small efficiency gains could result in a significant cost saving [7]. Our primary objective in this project was to help realise this cost saving by using a data-driven forecast and optimise approach to operation.

There are multiple objectives for the Cambi system, however, under the current intuition-based operation, operators are forced to prioritise some objectives and disregard others, due to the system's complexity. In order of priority, the objectives within the Cambi system are, based on communication with UU employees:

- to maintain a low level in the upstream storage such that it can accept new trucked deliveries, via appropriate throughput management;
- to produce an acceptable biosolid quality, within the specification of the End of Waste Code
- (this is a lower priority than flow since this is rarely an issue); and finally,
- to produce at the lowest possible cost.

With rising gas prices [3], this paper highlights the need to further investigate this final objective and provides a framework to do so, using the data available to UU.

Data

The data, which is available to UU includes historical plant data, accessible through UU's various data warehouses [9], and Bureau of Meteorology weather data. Exploration of the data revealed several correlations, which could be exploited for efficiency gains, and regression models demonstrated the non-linearity of the data.

Methodology

To build a decision optimisation system which preserves the current operating objective philosophy of Cambi, we first constructed a mixed-integer programming (MIP) model whose objective function minimises the difference between the level in the upstream storage unit and a target level, over some planning horizon. This ensures that throughput is still the primary objective of the recommendation system. Next, we applied a novel machine learning regression model, which predicts the behaviour of the set of all possible operating scenarios, subject to the constraints described by the MIP outputs. This model reports the case within this set which it predicts is the most efficient, contrary to the conventional predict and optimise framework. This structure allowed us to use the regression model to predict both biosolid quality and energy efficiency, choosing the inputs which resulted in the lowest energy usage, while still subject to the constraint of acceptable biosolid quality. The MIP optimisation model and the machine learning regression model together form the basis for the decision optimisation system, which collates live operation data and provides feedback to the operators.

The MIP model was implemented using Google's Python based OR-Tools library and minimises the function $\sum^{T}_{t=0} z_t + \omega \sum^{T}_{t=^-0^1} \sum r^3_{=1} y_{r,t}$ where z_t is the absolute difference between the level in the upstream storage unit and a target level, ω is a weighting factor determined by trial-and-error, and $y_{r,t}$ is a binary variable representing if reactor r was turned on/off in time step t. The first term of the objective function ensures that the level never deviates too far from the target and the second term ensures that operator intervention is not laborious. Figure 1 shows the performance of the MIP model in comparison to the current manual operation and it can be seen the MIP model maintains a much more stable storage level.

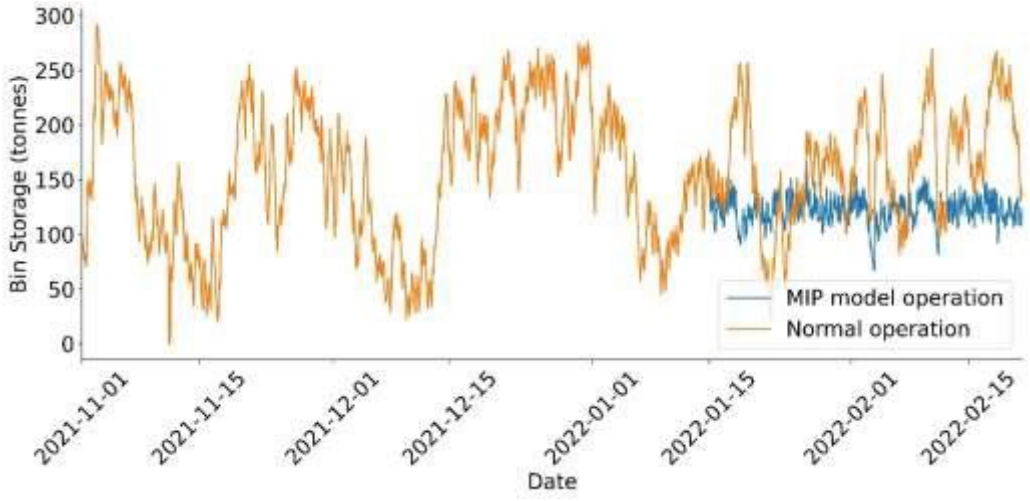

Figure 1: MIP Model Performance Against Actual Operation Level

The machine learning regression model was developed by testing a shortlist of candidate architectures including SVM (with RBF kernel), AdaBoost, Random Forest, LightGBM, kNN and MLP. While not a native multi-output model, LightGBM, with a multi-output wrapper from the Scikit-learn library, was found to predict Cambi's behaviour on a test set better than the other models.

It is estimated that if UU were to adopt the recommendations of this project, they could save approximately $500,000 annually on natural gas consumption, given current gas prices. We implemented the forecasting and optimisation models in Python and developed a dashboard in PowerBI for visualisation.

Limitations

In this current study we developed a forecasting and an optimisation model for the problem and separately applied them to provide a solution. However, these two phases can be integrated into a single model where the forecasting model considers the quality of the final decisions. Thus, the forecasting model generates more accurate forecasts which will be used for improving the quality of decisions in the optimisation model. The whole process can be also integrated into an end-to-end model to directly map the inputs to final optimal decisions.

Conclusion

We conclude this paper by reiterating the cost savings that a data driven approach with machine learning to wastewater treatment plant operation can provide. Furthermore, we highlight where the proposed "forecast and optimise with machine learning" approach could be extended beyond Cambi. Finally, we encourage further research in the domain of forecasting and decision optimisation with machine learning models, especially in automating the process and further improving the quality of forecasts and decisions by integrating them.

Reference

1. Abolghasemi, M., Abbasi, B., Babaei, T., HosseiniFard, Z. (2021). How to effectively use machine learning models to predict the solutions for optimization problems: lessons from loss function. Paper submitted to Computers and Operations Research.
2. Abolghasemi, M., & Esmaeilbeigi, R. (2021). State-of-the-art predictive and prescriptive analytics for IEEE CIS 3rd Technical Challenge. arXiv preprint arXiv:2112.03595.
3. Australian Government. (2022). Gas Market Prices. Australian Energy Regulator.
4. https://www.aer.gov.au/wholesale-markets/wholesale-statistics/gas-market-prices
5. Burton, F.L., Stensel, H.D., Tchobanoglous, G., Tsuchihashi, R. (2013). Wastewater Engineering:
6. Treatment and Resource Recovery 5th ed. New York: McGraw-Hill
7. Cambi. (2021). How does thermal hydrolysis work? Retrieved from https://www.cambi.com/what-we-do/thermal-hydrolysis/how-does-thermal-hydrolysiswork/
8. Dwyer, J., Starrenburg, D., Tait, S., Barr, S., Batstone, D., Lant, P. (2008). Decreasing activated sludge thermal hydrolysis temperature reduces product colour, without decreasing degradability. Water Research, 42, 4699-4709.
9. Malekizadeh, A. (2019, May). Thermal Hydrolysis Process Optimisation in Relation to Anaerobic Digestion. Paper presented at OzWater 2019 in Melbourne, Australia.
10. Waste Reduction and Recycling Act 2011 (Qld). Retrieved from https://www.legislation.qld.gov.au/view/pdf/inforce/current/act-2011-031 [9] Wong, R. (2020). ProcessDataLink Management Procedure. Urban Utilities.

Acknowledgment:

In addition to the above, personal communication with employees of Urban Utilities has been instrumental in developing the problem definition and discussions with Dr. Mahdi Abolghasemi have been vital to developing modelling ideas for this problem.

National Weighted Vulnerability Index Methodology: an Australian Case Study at Fine Temporal and Geographical Resolutions

Aiden Price[1], Flavia Barar[2], Callan Davis[1], Paula Fiévez[3], Rohit Gupta[2], Kerrie Mengersen[1], Michael Rigby[2], and Evan Thomas[2]

[1] Centre for Data Science, Queensland University of Technology, Brisbane, Australia. [2] Data Science Support, Australian Urban Research Infrastructure Network, Melbourne, Australia. [3] Business Development, FrontierSI, Melbourne, Australia.
Email: *a11.price@qut.edu.au*

***Keywords*:** environmental health, vulnerability index, climate, air quality, built environment.

Abstract

Extreme natural hazards are increasing in frequency and intensity and continue to have substantial economic, social, and health impacts globally. The impact of the environment on human health (environmental health) is becoming well understood in international research literature.

However, there are significant barriers to understanding key characteristics of this impact, related to substantial data volumes, data access rights, and the time required to compile and compare data over regions and time. This study aims to reduce these barriers by creating an open data repository of national environmental health data and presenting methodology for the production of weekly health outcome weighted population vulnerability indices related to extreme heat, extreme cold, and air pollution at the statistical area level 2 (SA2) geographical resolution.

The weighted vulnerability index methodology employed in this study offers an advantage over others in the literature by targeting health outcomes in the calculation process. The resulting vulnerability percentile more clearly aligns population characteristics with health risks. The temporal and spatial resolutions available enable national monitoring to the public on a scale never before seen across Australia. Additionally, we show that the weekly temporal resolution can be used to identify spikes in vulnerability due to changes in relative national environmental exposure.

Introduction

In 2016, the World Health Organisation (WHO) stated that 13.7 million deaths a year, or 24% of all global deaths, are linked to the environment and that a range of environmental health risks such as clean air, stable climate, and health-supportive cities and built environments are all prerequisites for good health [1]. For example, it is well understood that air quality and climate extremes have significant impacts on human health [2, 3, 4, 5]. Even in Australia, where air quality is generally among the cleanest in the world [11], pockets of poor air quality are said to be responsible for the deaths of approximately 3000 Australians per year [12]. Extreme heat is the focus of a large body of research investigating the relationship between climate and human health. However, cold-related mortality is shown to exceed heat related mortality both internationally [6, 7] and in Australia [8], especially when considered alongside increased respiratory conditions [9, 10].

Historically, significant efforts have been made to characterise the human and financial risks related to natural hazards. More recently, these efforts and the corresponding risks have become focused more sharply on the vulnerability of exposed human populations rather than the exposure itself [13]. The term vulnerability, in social science research, generally describes a state of people and populations, and varies significantly both historically and geographically [14]. The changing nature of many hazards, coupled with growing and ageing populations and infrastructure in exposed areas is leading to increased vulnerability across Australia and internationally [15]. This study utilises a range of environment and health data to provide an assessment of environmental health risk through the production of vulnerability indices for heat, cold, and air quality across Australia.

Methodology

The national vulnerability indices created in this study utilise a consistent environmental health vulnerability index framework introduced by the Intergovernmental Panel on Climate Change (IPCC) in 2007 [16]. This framework aims to better explain population vulnerability with respect to climate, and is made up of three sections: exposure, sensitivity, and adaptive capacity. Exposure is a direct measure of household, community, or population exposure to a certain event (extreme heat, cold, air pollution, etc.). Sensitivity is

defined as the susceptibility of a household, community, or population to the exposure. Finally, adaptive capacity captures the capabilities of a household, community, or population to cope with or recover from the impact of the exposure. A number of vulnerability indices exist which focus primarily on population variables. Prominent 21st century examples of these indices include the social vulnerability index (SVI) [17], the social vulnerability index for the United States (SoVI) [18], and Australia's index of relative socio-economic disadvantage (IRSD) [19].

The data used to create the vulnerability indices in this study include climate data from the Bureau of Meteorology (BOM) [20], air pollutant data from the Copernicus Atmosphere Monitoring Service (CAMS) [21], built environment data from the Australian Bureau of Statistics (ABS) [22], Geoscience Australia's Digital Earth Australia (DEA) [23], NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) [24], and the Terrestrial Ecosystem Research Network (TERN) [25], demographics data from the ABS Census [22] and the Public Health Information Development Unit (PHIDU) social health atlas [26, 27], and mortality data from the Australian Institute of Health and Welfare (AIHW) [28].

The majority of vulnerability indices identified in research literature combine data from a wide temporal range to form a single index [29, 30]. When changes over time are considered in the creation of vulnerability indices, this is usually over either a short time period or only a few time points [31]. The vulnerability indices in this study are presented at weekly, monthly, and yearly scales from 2011-2019 by incorporating daily climate and air quality data and considering changes over time to the exposure, sensitivity, and adaptive capacity sections of the created vulnerability indices.

Baseline Equal Weights Approach

The equal weights method [17] is the baseline approach in the creation of vulnerability indices. In this approach, a sub-index is typically created by summing together spatial percentiles of its variables. Let S_{kt} be a sub-index at time t containing N_k variables, where $k = 1,2,3$ corresponds to exposure, sensitivity, and adaptive capacity sub-indices, respectively. With variables $n = 1, \dots, N_k$, a vulnerability sub-index, S_{ik}, for a geographical region, i, at time, t, is calculated as follows.

$$S_{ikt} = \sum_{n=1}^{N_k} f_i(\underline{x}_{nt}), \tag{1}$$

where $\underline{x}_{nt}$ is a vector containing every region's observed value of variable n at time t. The function f_i computes the i^{th} region's spatial percentile of $\underline{x}_{nt}$, ignoring any missing values in $\underline{x}_{nt}$. The final vulnerability index, VI_{it}, is calculated in a similar way.

$$VI_{it} = \sum_{k=1}^{3} f_i(\underline{S}_{kt}), \tag{2}$$

where $\underline{S}_{kt}$ is a vector containing every region's observed value of vulnerability sub-index k. In consideration of missing data when calculating the index, the average of the spatial percentiles is calculated without the related variable for each of the sub-indices and overall vulnerability index.

Health Outcome Weighted Vulnerability Indices

The methodology used in this study uses pairwise correlation values between each variable and age-standardised mortality rates to form a weighted sum. Note that weights can be positive or negative depending on the relationship between a particular variable and the health outcomes considered. The calculation of each weighted sub-index, WS_{ikt}, at time t for region i is as follows.

$$WS_{ikt} = \frac{w_{nk}}{N_{kt}} \sum_{n=1}^{N_{kt}} f_i(\underline{x}_{nt}), \tag{3}$$

where w_{nk} is the Kendall's Tau correlation between variable n and all-cause mortality in sub-index k. Note that equal weights are being used for the exposure sub-index, i.e., $w_{nk} = 1$, for $n = 1, \dots, N_1$ (see Discussion). The calculation of the overall weighted vulnerability index, wVI_{it}, remains the same, replacing the original sub-indices for the weighted sub-indices and taking the average.

$$wVI_{it} = \frac{1}{3} \sum_{k=1}^{3} f_i(\underline{WS}_{kt}). \tag{4}$$

Findings

The drastic improvement in correlation between age-standardised all-cause mortality and the two vulnerability index methodologies can be seen more clearly in Figure 1, where spatially ranked all-cause mortality is plotted directly against both (A) the baseline heat vulnerability index, HVI, and (B) the weighted heat vulnerability indices, $wHVI$. Choropleth maps additionally provide spatial insight into this improvement in correlation in Figure X, where the difference between spatially ranked all-cause mortality and vulnerability index methodologies are shown, again for the heat vulnerability index. The proposed inclusion of weights that are directly tied to targeted health outcomes (e.g., mortality) inherently increase the value of the indices for longer-term risk assessment.

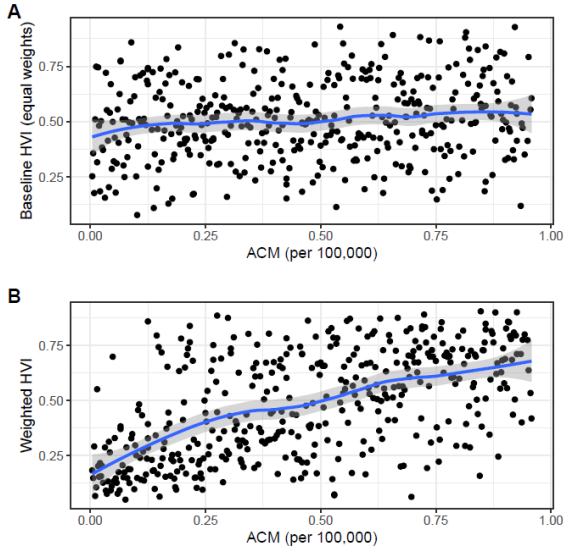

Figure 1: Scatter plot of age-standardised all-cause mortality (ACM) against (A) the baseline heat vulnerability index (HVI) showing low correlation, and (B) the weighted heat vulnerability index ($wHVI$) showing improved correlation.

It is of particular importance to raise the temporal resolution of existing vulnerability indices for the purpose of communicating need. While it has been common historically and recently to combine multiple years of data in the creation of vulnerability indices [29, 32], it is clear that data relating to fine temporal-scale extreme events becomes is largely ignored once data is averaged annually. The range of data collected for the AusEnHealth initiative has enabled the presentation of a finer scale vulnerability index, allowing for variation in region-specific vulnerability to be observed. Given the temporal resolution of the underlying data, these intra-seasonal changes are a direct result of fluctuations in exposure, which when combined with the underlying demographics and built environment characteristics over a larger time period, represent reactive population vulnerability at a resolution never seen before across Australia.

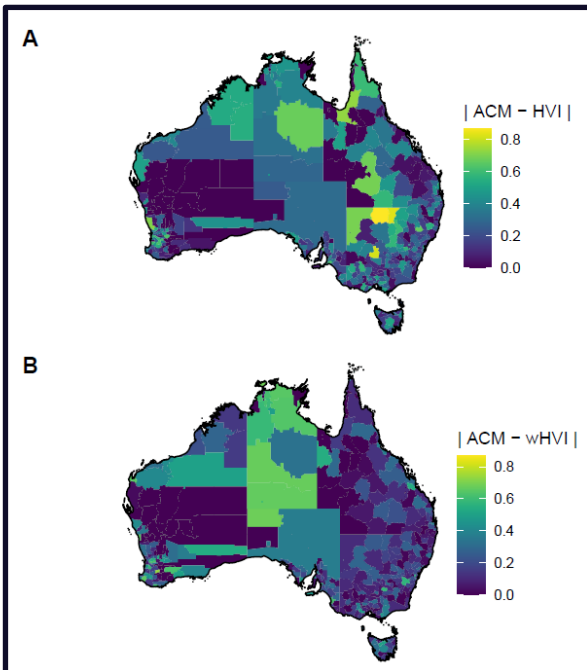

Figure 2: Choropleth map displaying the absolute difference between age standardised all-cause mortality (ACM) and (A) the baseline equally weighted heat vulnerability index (HVI) and (B) the weighted heat vulnerability index ($wHVI$).

Conclusion

Vulnerability indices offer a way to assess population vulnerability across Australia despite variations in population density and in environmental burden of disease. However, common index creation methodology may not accurately reflect timely population health risks due to a heavy focus on population demographics over health outcome data and a lack of data necessary to produce time series data. The pairwise correlation approach to vulnerability index methodology allows for a health outcome of interest to guide population vulnerability calculations, while maintaining the critical exposure, sensitivity, and adaptive capacity framework used in all recent vulnerability index methodological advances. The volume of data collected in this study also allows for high temporal and geographical resolutions, allowing users of the data to observe fine changes in exposure data and hence enabling the ability to determine sharp and sudden changes in population vulnerability due to an emerging natural hazard.

The categories of mortality, as well as the exposure, sensitivity, and adaptive capacity variables used in this study reflect the data available at the time of writing. Data included can be greatly expanded given future data availability, including health service availability data, urban environment variables such as road and building density data, and upcoming 2021 Australian Census data. There are also opportunities to explore the vulnerability index methodology, including a more rigorous assessment of underlying variables and related health outcomes.

References

1. Prüss-Üstün, A., Wolf, J., Corvalán, C., Bos, R., Neira, M.: Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks (2016)
2. Seltenrich, N.: Between extremes: health effects of heat and cold. Environmental Health Perspectives 123(11), 276–280
3. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E.: Environmental health impacts of air pollution: a review. Frontiers in Public Health 8, 14

4. IPCC: Climate change 2021: The physical science basis: Summary for policymakers
5. Kinney, P.: Climate change, air quality, and human health. American Journal of Preventive Medicine 35(5), 459–467
6. Berko, J., Ingram, D., Saha, S., Parker, J.: Deaths Attributed to Heat, Cold, and Other Weather Events in the United States, 2006-2010, National Health Statistics Report. US Department of Health and Human Services
7. Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklov, J., Forsberg, B., Leone, M.: Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet 386(9991), 369–375
8. Vardoulakis, S., Dear, K., Hajat, S., Heaviside, C., Eggen, B., McMichael, A.: Comparative assessment of the effects of climate change on heat-and cold-related mortality in the United Kingdom and Australia. Environmental Health Perspectives 122(12), 1285–1292
9. Wondmagegn, B., Xiang, J., Dear, K., Williams, S., Hansen, A., Pisaniello, D., Nitschke, M., Nairn, J., Scalley, B., Xiao, A., Jian, L., Tong, M., Bambrick, H., Karnon, J., Bi, P.: Increasing impacts of temperature on hospital admissions, length of stay, and related healthcare costs in the context of climate change in Adelaide, South Australia. Science of the total Environment 773, 145656
10. Kinney, P., Schwartz, J., P, M., Elisaveta, P., Terte, A.L., Medina, S., Vautard, R.: Winter season mortality: will climate warming bring benefits? Environmental Research Letters 10(6), 064016
11. IQAir: Air quality in Australia. In: IQAir. [Online]. Available: https://www.iqair.com/au/australia
12. Australian Institute of Health and Welfare: Australian burden of disease study: impact of illness and death in Australia. In: AIHW, Canberra
13. Burton, G., Rufat, S., Tate, E.: Social vulnerability: conceptual foundations and geospatial modelling. In: Vulnerability and Resilience to Natural Hazards, pp. 53–81. Cambridge University Press
14. Wisner, B., Blaikie, P., Cannon, T., Davis, I.: At Risk: Natural Hazards, People's Vulnerability, and Disasters. Taylor and Francis Group London, Routledge
15. National Resilience Taskforce: Profiling Australia's vulnerability: the interconnected causes and cascading effects of systemic disaster risk. Department of Home Affairs, Canberra
16. Parry, M., Parry, M., Canziano, O., Palutikof, J., Linden, P., Hanson, C.: Climate Change 2007: Impacts, Adaptation and Vulnerability. Cambridge University Press, Cambridge
17. Flanagan, B., Gregory, E., Hallisey, E., Heitgerd, J., Lewis, B.: A social vulnerability index for disaster management. Journal of Homeland Security and Emergency Management 8(1)
18. Cutter, S., Boruff, B., Shirley, W.: Social vulnerability to environmental hazards. Social Science Quarterly 84(2), 242–261
19. Australian Bureau of Statistics: Technical Paper: Census of Population and Housing: Socio-economic Indexes for Areas (SEIFA). Australian Bureau of Statistics, Canberra
20. Bureau of Meteorology: Long-range weather and climate. [Online]. Available: http://www.bom.gov.au/climate/
21. Copernicus: CAMS global reanalysis (EAC4). [Online]. Available: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview
22. Australian Bureau of Statistics: What the Census is. [Online]. Available: https://www.abs.gov.au/census/learn/about
23. Geoscience Australia: DEA Waterbodies. [Online]. Available: https://www.ga.gov.au/dea/products/dea-waterbodies
24. NASA: MODIS Vegetation Index Products (NDVI and EVI). [Online]. Available: https://modis.gsfc.nasa.gov/data/dataprod/mod13.php
25. TERN: Seasonal ground cover - landsat, JRSRP algorithm, Australia
26. coverage. [Online]. Available: https://portal.tern.org.au/ seasonal-ground-cover-australia-coverage/22022
27. Australian Institute of Health and Welfare: Australian Bushfires 2019-20: Exploring the Short-term Health Impacts. AIHW, Canberra
28. Public Health Information Development Unit: Notes on the data: chronic diseases and conditions. [Online]. Available: https://phidu.torrens.edu.au/notes-on-the-data/
29. health-status-disability-deaths/est-diabetes
30. Australian Institute of Health and Welfare: Mortality over regions and time (MORT) books. [Online]. Available: https://www.aihw.gov.au/reports/life-expectancy-death/mort-books/contents/mort-books
31. Beaty, M., Varghese, B.: Reducing illness and lives lost from heatwaves. In: Bureau of Meteorology, Canberra
32. Yu, J., Castellani, K., Forysinski, K., Gustafson, P., Lu, J., Peterson, E., Tran, M., Yao, A., Zhao, J., Brauer, M.: Geospatial indicators of exposure, sensitivity, and adaptive capacity to assess neighbourhood variation in vulnerability to climate change-related health hazards. Environmental Health 20(1), 1–20
33. University of New South Wales: Heat vulnerability index for Sydney. City Futures Research Centre. [Online]. Available:
34. Conlon, K.C., Mallen, E., Gronlund, C.J., Berrocal, V.J., Larsen, L., O'neill, M.S.: Mapping human vulnerability to extreme heat: a critical assessment of heat vulnerability indices created using principal components analysis. Environmental health perspectives 128(9), 097001 (2020)

Low-cost Paretonian DBSCAN Parameter Estimation for Sklearn

T. N. Stenborg[1,2], K. Silversides[1,2]

[1] *The University of Sydney, Sydney, New South Wales 2006, Australia;*

[2] *ARC Training Centre in Data Analytics for Resources and Environments (DARE);*

travis.stenborg@sydney.edu.au

Abstract

DBSCAN is a well-known algorithm for accurate clustering, a key problem in data science. A popular Python library implementing DBSCAN is sklearn. Poor parameter specification with sklearn's DBSCAN can significantly inflate runtimes or crash sessions after memory exhaustion. A low computational cost parameter estimation scheme, incorporating the Pareto principle, was devised to avoid such issues. The scheme was implemented as simple, publicly available, Python code. Testing with sample magnetic and gravity survey data from north-western Queensland showed promise for bounding sensible clustering parameters.

Introduction

Clustering is a common unsupervised learning activity for data science (Xu and Tian 2015, Chhabra et al. 2021). A well-known density-based clustering algorithm is Density-based Spatial Clustering of Applications with Noise (DBSCAN; Ester et al. 1996). For the case of many data points n, DBSCAN can be computationally intensive, even "intolerably intensive" (Gan and Tao 2015). Naïve parameter specification can exacerbate this, so far as to cause session crashes by exhausting system memory (see §3).

DBSCAN builds clusters by checking, for data points, if they have some minimum number of neighbouring points minPts, within a test radius ε. Devising approaches to estimating those two key parameters is not new (e.g. Sander et al. 1998, Esmaelnejad et al. 2010, Starczewski et al. 2020). Notable heuristics include dimensionality-based approaches for minPts and "elbow" identification in k-nearest-neighbour distance plots (Schubert et al. 2017).

Techniques for estimating minPts and ε can be daunting to the non-specialist. Additionally, there's no guarantee that the default values supplied in DBSCAN implementations will work well for a given dataset. We thus devised an objectively simple, low-computational-cost, gross parameter estimation method, designed for novice clustering practitioners. Implemented in Python, it was tested with DBSCAN in the popular (Hao and Ho 2019) Scikit-learn (hereafter, sklearn) machine learning library.

Methods

Parameter Estimation

Estimating minPts was done by taking the maximum of sklearn's default (five) and twice the dataset dimensionality. This was a minor variation of an approach discussed in Schubert et al. 2017.

To estimate ε, the nearest-neighbour distance was calculated for each data point. Then, a Pareto-like assumption was made; 80% of the clustering structure arises from a DBSCAN test radius greater than the smallest 20% of the nearest-neighbour distances. Implementation has low computational cost, even for large n. Note, this approach should not be confused with application of Pareto optimality (as discussed with DBSCAN in, e.g. Azhir et al. 2021).

The ε estimation scheme, and its underlying 80/20 clustering assumption, isn't presented with a rigorous mathematical underpinning. Indeed, its validity is deemed likely to be data dependent. Nonetheless, it was explored here empirically as a potential DBSCAN aid.

Test Platform and Data

Sample magnetic and gravity survey data from a site near Cloncurry (east of Mount Isa, Queensland) were used as test data. Clustering of n = 177,720 in the 2D space of reduced-to-pole magnetic anomaly and the first vertical derivative of Bouguer gravity anomaly was attempted. Data were scaled according to interquartile range, via sklearn's RobustScaler. Sklearn's DBSCAN was then tested with a)

test parameters chosen by a novice user ($\varepsilon = 1$, *minPts* = 20), b) sklearn's default parameters ($\varepsilon = 0.5$, *minPts* = 5) and c) the new low-cost estimation scheme ($\varepsilon \approx 9\times10^{-4}$, *minPts* = 5, dataset specific).

The test platform was Windows 11, version 22H2, on an Intel "Comet Lake" Core i7-10750H, 16GB RAM system. Python 3.9.13 and sklearn 1.0.2 were invoked from Visual Studio Code 1.73.1.

Results

Parameters	Clusters	Outliers	Notes
$\varepsilon = 1$, *minPts* = 20	—	—	Crash, exhausts memory.
$\varepsilon = 0.5$, *minPts* = 5	2	0	Runtime ~ minutes. Highly asymmetric clusters.
$\varepsilon \approx 9\times10^{-4}$, *minPts* = 5	232	176,195	Runtime ~ seconds.

Table 1. Clustering parameter selection scheme testing; novice user selection (top), sklearn default (mid), Paretonian scheme (bottom).

Discussion

DBSCAN's propensity to exhaust system resources under naïve parameter specification was demonstrated. Use of sklearn's default parameters by contrast, clustered all points, albeit highly asymmetrically {25, 177695}. Though the default parameter clustering results may reflect real structure in the underlying data, experimenting with a smaller ε would be prudent in this case.

Using parameters based on the new Pareto-like scheme produced many small clusters, whilst designating the majority (>98%) of points as outliers. Testing larger values of ε is prima facie merited. Clustering runtime with this scheme was however seconds, compared to minutes with sklearn's default parameters. The empirical evaluation thus suggests the Paretonian scheme may be of value in quickly identifying sensible lower bounds for ε.

Using a combination of sklearn's default parameters, and the Paretonian scheme presented here, can guide clustering for novice DBSCAN users not ready to adopt more advanced parameter estimation techniques. Performance of the scheme for other datasets remains as future work.

The Python code and test data associated with this work is available at GitHub (https://github.com/tstenborg/Paretonian-DBSCAN-Parameters).

Acknowledgements

This research was conducted by the Australian Research Council Training Centre in Data Analytics for Resources and Environments (project number ICI9010031). Data products of Geoscience Australia were used, ostensibly available from the Geoscience Australia Portal, https://portal.ga.gov.au/persona.

References

Azhir, E, Jafari Navimipour, N, Hosseinzadeh, M, Sharifi, A, & Darwesh, A, 2021, "An efficient automated incremental density-based algorithm for clustering and classification", Future Generation Computer Systems, vol. 114, pp. 665–678.

Chhabra, A, Masalkovaite, K, & Mohapatra, P, 2021, "An Overview of Fairness in Clustering", IEEE Access, vol. 9, pp. 130698–130720.

Esmaelnejad, J, Habibi, J, & Yeganeh, SH, 2010, "A Novel Method to Find Appropriate ε for DBSCAN", in Intelligent Information and Database Systems, vol. 5990, pp. 93–102.

Ester, M, Kriegel, H-P, Sander, J, & Xu, X, 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in Proc. 2nd ACM International Conference on Knowledge Discovery and Data Mining, vol. 96, no. 34, pp. 226–231, AAAI Press.

Gan, J & Tao, Y, 2015, "DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation", in Proc. 2015 ACM SIGMOD International Conference on management of data, pp. 519–530, ACM.

Hao, J & Ho, TK, 2019, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language", Journal of Educational and Behavioral Statistics, vol. 44, no. 3, pp. 348–361.

Sander, J, Ester, M, Kriegel, H-P, & Xu, X, 1998, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications", Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 169–194.

Schubert, E, Sander, J, Ester, M, Kriegel, H, & Xu, X, 2017, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN", in ACM Transactions on Database Systems, vol. 42, no. 3, pp. 1–21.

Starczewski, A, Goetzen, P, & Er, MJ, 2020, "A New Method for Automatic Determining of the DBSCAN Parameters", Journal of Artificial Intelligence and Soft Computing Research, vol. 10, no. 3, pp. 209–221.

Xu, D & Tian, Y, 2015, "A Comprehensive Survey of Clustering Algorithms", Annals of Data Science, vol. 2, no. 2, pp. 165–193.

Accelerating MCMC-driven Gaussian Plumes with Numba

T. N. Stenborg[1,2], S. C. Davis[1,2]

[1] *The University of Sydney, Sydney, New South Wales 2006, Australia;*

[2] *ARC Training Centre in Data Analytics for Resources and Environments (DARE);*

travis.stenborg@sydney.edu.au

Abstract.

Data science can be limited by the computational intensity of numerical analysis. Technologies or methods that ameliorate this thus expand the range of tractable problems. Numba, a just-in-time compiler for Python and NumPy, is one such technology. This study quantifies the performance boost Numba provides to Python code. MCMC-driven Gaussian plume modelling of seawater droplets was used as a test case. Benchmarking showed an 88.8% speedup.

Introduction

Data science problems can be computationally intensive (e.g., Tapiador et al. 2017, Polson and Sokolov 2020, Kenett et al. 2022). This problem can be naively addressed by making more computational resources available (e.g. faster processors, more RAM, multicore computing). Alternatively, it can be addressed by reducing the amount of computation needed to solve a given problem.

Python is common data science language (Ordonez 2021). It can suffer performance issues however, for certain computational problems due to being an interpreted, rather than a compiled language (Strout et al. 2019). Numba addresses this issue by producing compiled Python and NumPy code, at least for a subset of commands (Lam et al. 2015). Moreover, it's typically implementable via a modest level of refactoring and decorator addition (De Pra and Fontana 2020). For problems hampered by Python's computational limits, incorporating Numba can thus provide a performance boost for a smaller time investment than competing amelioration strategies, such as rewriting a program in C, C++ or Cython.

This study quantifies the performance enhancement Numba provides to a test program. The program tested was prototype MCMC-driven Gaussian plume modelling, essentially a fluid dynamics simulation of droplets moving through a carrier fluid. For an introduction to Gaussian Plumes, see e.g., De Visscher 2014.

Methods

MCMC-driven Gaussian Plumes

The prototype MCMC-driven Gaussian plume code developed here was used to model dispersion of seawater droplets in air. This modelling incorporated third-party 3D droplet concentration data (a 6997 row subset), collected via drone over the Great Barrier Reef near Townsville, Queensland. The test platform was Windows 11, version 22H2, on an Intel "Comet Lake" Core i7-10750H (with 16GB RAM) system. Python 3.9.13 and Numba 0.56.3 were invoked from Visual Studio Code 1.73.1.

Two versions of the Gaussian plume code were compared: one standard Python version, and another differing only by the addition of Numba decorators. Execution was configured to run for an arbitrary 10,000 samples. Each version of the code was tested 200 times, using the same random number generator seed each time. Numba was confirmed as running exclusively in *nopython* mode, the fastest of its two available operating modes. All foreground applications were closed, and networking connectivity terminated, to reduce benchmarking noise.

Results

Performance testing results are given in Figure 1, below. The first iteration of the Numba version was noticeably slow, as the overhead of code compilation was incurred. Subsequent calls to that code invoke the compiled version. Numba provided an average speedup of 88.8%.

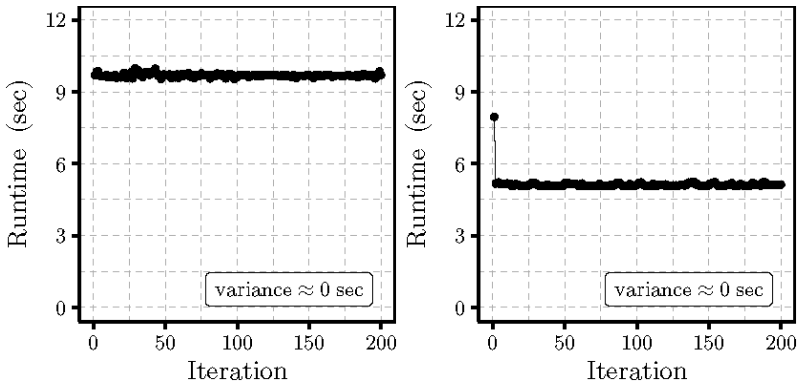

Figure 1. Runtime testing of unmodified (left) vs Numba-enhanced (right) Gaussian plume models. Each iteration performed 10,000 MCMC samples.

Discussion

Application of Numba to the example application provided an unambiguous performance boost (average 88.8% over 200 iterations). Refactoring measures adopted in incorporating Numba included e.g., switching from Pandas to Numpy data structures, avoiding Python lists, and some changes to use of function types (viz. keyword, positional).

Acknowledgements

This research was conducted by the Australian Research Council Training Centre in Data Analytics for Resources and Environments (project number ICI9010031).

References

De Pra, Y & Fontana, F, 2020, "Programming real-time sound in Python", Applied Sciences, vol. 10, no. 12, p. 4214.

De Visscher, A, 2014, "Air Dispersion Modeling: Foundations and Applications", Wiley, Somerset, New Jersey.

Lam, SK, Pitrou, A & Seibert, S, 2015, "Numba: A LLVM-based Python JIT Compiler", in Proc. Second Workshop on the LLVM Compiler Infrastructure in HPC, pp. 1-6, ACM.

Kenett, RS, Gotwalt, C, Freeman, L, & Deng, X, 2022, "Self-supervised cross validation using data generation structure", Applied Stochastic Models in Business and Industry, vol. 38, no. 5, pp. 750–765.

Ordonez, C, 2021, "A Comparison of Data Science Systems", in Big Data Analytics, vol. 12581, pp. 3–11, Springer International Publishing, Cham.

Polson, N & Sokolov, V, 2020, "Deep learning: Computational aspects", Wiley Interdisciplinary Reviews Computational Statistics, vol. 12, no. 5.

Strout, MM, Debray, S, Isaacs, K, Kreaseck, B, Cárdenas-Rodríguez, J, Hurwitz, B, Volk, K, Badger, S, Bartels, J, Bertolacci, I, Devkota, S, Encinas, A, Gaska, B, Neth, B, Sackos, T, Stephens, J, Willer, S, & Yadegari, B, 2019, "Language-Agnostic Optimization and Parallelization for Interpreted Languages", in Languages and Compilers for Parallel Computing, pp. 36–46, Springer International Publishing, Cham.

Tapiador, D, Berihuete, A, Sarro, LM, Julbe, F, & Huedo, E, 2017, "Enabling data science in the Gaia mission archive: The present-day mass function and age distribution", Astronomy and Computing, vol. 19, pp. 1–15.

Counterintuitive Outcomes from PowerShell OS Noise Mitigation

T. N. Stenborg[1,2]

[1]*The University of Sydney, Sydney, New South Wales 2006, Australia;*

[2]*ARC Training Centre in Data Analytics for Resources and Environments (DARE);*

travis.stenborg@sydney.edu.au

Abstract

Data science tasks typically involve selecting some subset of available computational tools. Examples include choosing between competing platforms, database systems, programming languages, algorithms, data structures, etc. Where operational speed drives choice, benchmarking can distinguish options. Background activity by an operating system can however add noise to such benchmarking, compromising accuracy. Running the data science language R on Windows is examined here, where background processes, services and the Task Scheduler combine to frustrate benchmarking. PowerShell calls can be embedded in R that reduce such activity, expected to improve benchmarking accuracy. An example problem in multicore Bayesian inference was used to evaluate the noise mitigation strategy. The counterintuitive outcome was an increase in runtime variance.

Introduction

Performance benchmarking can identify efficient implementation choices (e.g. programming language, algorithms, data structures, etc.) for data science problems. A key aspect of proper benchmarking is repeatability (Huppler 2009). Operating system (OS) noise can compromise repeatability, a well-known issue in the high performance computing community (Petrini et al. 2003, Morari et al. 2011, Gerofi et al. 2019). It's not however, routinely considered in benchmarking workstation-level R (e.g. Eddelbuettel & Balamuta 2018, DeRaad 2022, Herrando-Pérez et al. 2021).

The efficacy of dampening OS noise for the case of R on Windows 11 was examined. Specifically, terminating selected unneeded processes and services via inline PowerShell in R was tested. Additionally, dynamic disabling of the Task Scheduler, capable of spawning new activity, was performed via programmatic manipulation of the Windows registry. A test case incorporating Stan, able to trivially parallelise Bayesian inference over all of a processor's cores from R, was used to contrast noisy vs dampened benchmarking quantitatively.

Methods

Windows Noise Dampening

Windows noise dampening was performed via programmatic dynamic shutdown of processes and services (PowerShell from R). No network connection was active, and the system was configured to never enter a sleep state (background tasks can be triggered by network availability or power state change (Wright & Plesniarski 2016)). By automating Registry (Carvey 2016) modification, the Task Scheduler was amongst the services dynamically shut down, further disabling event or schedule based task triggers. Only background applications deemed essential were left active.

Stan Test Case

Existing public domain Stan code for Bayesian inference of cosmological parameters for type Ia supernovae was used as a test case (listing 10.26, Hilbe et al. 2017). An accompanying dataset was also in the public domain (joint light-curve analysis *JLA* sample, Betoule et al. 2014). The test platform was Windows 11, version 22H2, running on an Intel "Comet Lake" Core i7-10750H, 16GB RAM system. Implementation used RStan 2.26.13, R 4.2.2, RStudio 2022.07.1 and Windows PowerShell 5.1.22621.608. Execution was parallelised over all physical cores to load stress the processor.

The inference program was run 200 times under standard OS conditions, then another 200 times with noise dampening implemented via PowerShell (termination of the Task Scheduler plus 12 processes and 25 services). The number of Stan samples was configured to reach probable convergence, using the metrics R-hat < 1.01 and $\{x \mid x \geq 0.1n \text{ and } x \geq 100m,\ x \in \{\text{bulk-}n_{\text{eff}}, \text{tail-}n_{\text{eff}}\}\}$, where R-hat is a convergence diagnostic, n number of samples, m number of Markov chains, and bulk-n_{eff} and tail-n_{eff} are the effective sample sizes of

the bulk and tails (5% and 95% quantiles) of the posterior distribution, as defined by Vehtari et al. (2021). The same random number generator seed was used for each iteration.

Results

Results of performance benchmarking are given in Figure 1. A counterintuitive higher average runtime (≈32.9 vs ≈31.1 sec) and runtime variance (≈0.6 vs ≈0.3 sec) occurred for the dampened case.

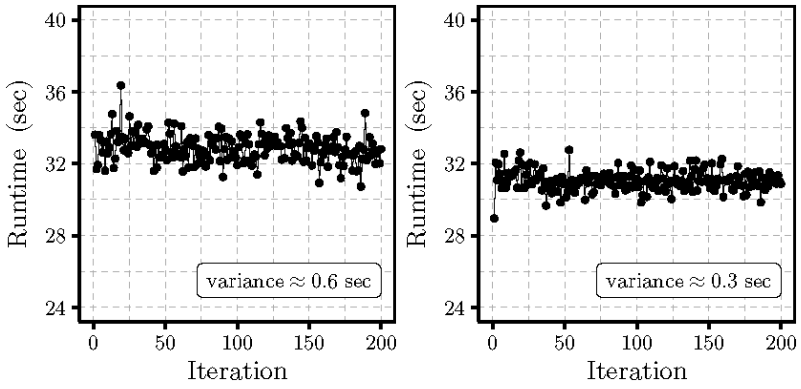

Figure 1. Runtimes over 200 iterations, with OS noise dampening (left) vs without dampening (right). A counterintuitive increase in runtime variance and average runtime occurred for the dampened case.

Discussion

Terminating background processes, services and the Task Scheduler was expected to dampen OS noise and decrease benchmarking variability. Instead, benchmarking variability increased.

The intent of terminating the Task Scheduler was to reduce new OS activity starting mid-benchmark. Terminating it however, may have had some unexpected impact on the way Windows manages internal resources.

The counterintuitive result obtained here warrants further investigation.

Acknowledgements

This research was conducted by the Australian Research Council Training Centre in Data Analytics for Resources and Environments (project number ICI9010031) and funded by the Australian Government (including contributions by the Australian National Health and Medical Research Council Ideas Grant GNT1186572).

References

Betoule, M., Kessler, R., Guy, J., Mosher, J., Hardin, D., Biswas, R., Astier, P., El-Hage, P., Konig, M., Kuhlmann, S., Marriner, J., Pain, R., Regnault, N., Balland, C., Bassett, B. A., Brown, PJ, Campbell, H., Carlberg, RG., Cellier-Holzem, F., Cinabro, D., Conley, A., D'Andrea, CB., DePoy, DL., Doi, M., Ellis, RS., Fabbro, S., Filippenko, AV., Foley, RJ., Frieman, J. A., Fouchez, D., Galbany, L., Goobar, A., Gupta, RR., Hill, GJ., Hlozek, R., Hogan, CJ., Hook, IM., Howell, DA., Jha, S. W., Le Guillou, L., Leloudas, G., Lidman, C., Marshall, JL., Möller, A., Mourão, AM., Neveu, J., Nichol, R., Olmstead, MD., Palanque-Delabrouille, N., Perlmutter, S., Prieto, JL., Pritchet, C. J., Richmond, M., Riess, AG., Ruhlmann-Kleider, V., Sako, M., Schahmaneche, K., Schneider, DP., Smith, M., Sollerman, J., Sullivan, M., Walton, NA. & Wheeler, CJ., 2014, "Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples", Astronomy and Astrophysics, vol. 568, pp. 1–32.

Carvey, HA, 2016, "Windows registry forensics: advanced digital forensic analysis of the Windows registry", 2nd edition, Syngress, Cambridge, MA.

DeRaad, DA, 2022, "snpfiltr: An R package for interactive and reproducible SNP filtering", Molecular Ecology Resources, vol. 22, no. 6, pp. 2443–2453.

Eddelbuettel, D & Balamuta, JJ, 2018, "Extending R with C++: A Brief Introduction to Rcpp", The American Statistician, vol. 72, no. 1, pp. 28–36.

Gerofi, B, Ishikawa, Y, Riesen, R, & Wisniewski, RW, 2019, "Operating Systems for Supercomputers and High Performance Computing", eds. B. Gerofi, Y. Ishikawa, R. Riesen, & RW. Wisniewski, 2019, Springer Singapore.

Herrando-Pérez, S, Tobler, R, & Huber, CD, 2021, "SMARTSNP, an R package for fast multivariate analyses of big genomic data", Methods in Ecology and Evolution, vol. 12, no. 11, pp. 2084–2093.

Hilbe, JM, De Souza, RS, & Ishida, EEO, 2017, "Bayesian models for astrophysical data : using R, JAGS, Python, and Stan", Cambridge University Press, Cambridge.

Huppler, K, 2009, "The Art of Building a Good Benchmark", in Lecture Notes in Computer Science, vol. 5895, pp. 18–30, Springer, Berlin, Heidelberg.

Morari, A, Gioiosa, R, Wisniewski, RW, Cazorla, FJ, & Valero, M, 2011, "A Quantitative Analysis of OS Noise", in 2011 IEEE International Parallel & Distributed Processing Symposium, pp. 852–863, IEEE.

Petrini, F, Kerbyson, D, & Pakin, S, 2003, "The Case of the Missing Supercomputer Performance: Achieving Optimal Performance on the 8,192 Processors of ASCI Q", in ACM/IEEE SC 2003 Conference, p. 55, ACM.

Wright, B, & Plesniarski, L, 2016, Microsoft Specialist Guide to Microsoft Windows 10 (Exam 70-697, Configuring Windows Devices) (Cengage Learning).

Vehtari, A, Gelman, A, Simpson, D, Carpenter, B, & Bürkner, P-C, 2021, "Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC", Bayesian Analysis, vol. 16, no. 2, p. 667–718.

A comparison of maximum likelihood estimation and median rank regression method in quantile estimation for Weibull data

D.N.S. Attanayake1, N. Armstrong, and A. P. Robinson

nayomi.attanayake@murdoch.edu.au[1]

School of Mathematics and Statistics, Murdoch University, Australia[1]

November 2022

Keywords: Weibull distribution, MLE, MRR, MSE, Absolute bias, Right Censored Data

Abstract

Accurate statistical analysis of failure data is vital in vehicle maintenance. Commonly, the Weibull distribution is used as the failure model and reliability measures are estimated. Frequently, in reliability analysis the data are heavily censored because we observe only a few failures. Hence, the robustness of the statistical methods in estimating the quantities is vital in probabilistic decision making. We consider two parameter estimation methods namely; Maximum Likelihood Estimation (MLE) and Median Rank Regression (MRR) method to fit the Weibull distribution to observed data and estimate the quantiles of the distribution.

We perform a set of simulation studies to compare the performance of MLE and MRR in quantile estimation. The data are right-censored with Type-II censoring. The simulations were carried out based on different shape parameters, sample sizes and failure rates (degree of censoring). Absolute bias and mean squared error (MSE) were used to compare the performance of the estimated quantiles. The results show that the MLE outperforms MRR in quantile estimation when the data are heavily censored in terms of MSE and absolute bias.

Background

The Weibull distribution is one of the widely used distribution to model failure data. the probability density function (pdf) of the Weibull distribution can be expressed as,

$$f(x|\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right), \quad x > 0. \quad (1)$$

Here, α is the shape parameter and β is the scale parameter of the Weibull distribution.

The shape parameter of the Weibull distribution reflects the failure rate of a system. If the shape parameter $\alpha < 1$, the system has a decreasing failure rate over the time. If $\alpha = 1$, the failure rate of the system decreases but at a slower rate. Figure 1 shows the behaviour of the Weibull distribution for different α values.

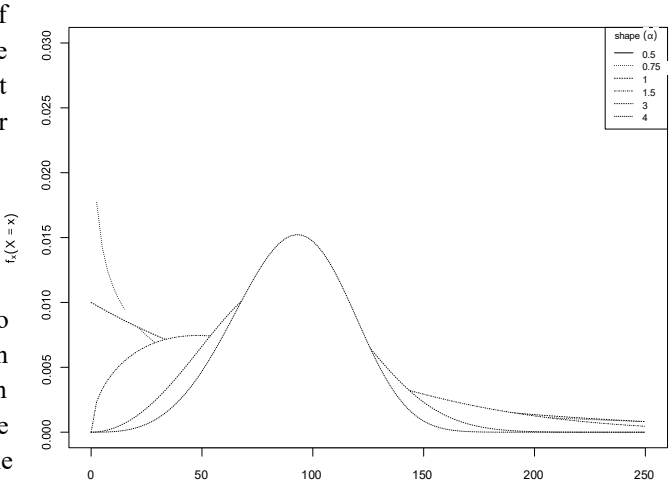

Figure 1: Distribution of the Weibull model for different α values.

The MLE and MRR methods are frequently used parameter estimation methods in reliability analysis. However, the optimality of the parameter estimation method can vary according to the type and degree of censoring, sample size, and the evaluation criteria used. In order to utilise the MLE, the first two derivatives of the log-likelihood function (see section 2) must be defined. In addition to that, the Fisher information matrix must not be zero and must be a continuous function of the parameters. Some of the MLE problems can encounter with more than one maximum solution where some problems may have no ML parameter estimates. For an example, when the number of failures is equal to zero, MLE does not exist [5]. A detailed explanation of the common uses of MLE is given by [4] [6, 8, 2, 3, see also]. The reliability of the decisions made based on the MLE can be questionable when the data are contaminated such as the presence of outliers.

MRR method is comparatively easy to implement in terms of computations, especially with the censored data. In MRR method, the information related to the position of the censored data after the final failure is not considered. Whereas, in MLE that information is taken into account. Additionally, the MRR method cannot be applied to some competing failure models such as gamma distribution and inverse-Gaussian distribution because their quantile functions cannot be written in closed forms in terms of the cumulative distribution functions (cdf).

Literature provides many comparisons of MLE and MRR method in estimating parameters and functions of parameters. [4] compared MLE and MRR method for the Weibull distribution under different censoring schemes. The performance of the two parameter estimation methods under heavy censoring is discussed by [7].

Methodology

If the failure time of a subject is not observed due to another event, then the resulting data are called censored. If the subjects have not failed by the end of the considered time or at the inspection time, then the data are "right-censored". Additionally, Type II censoring occurs when the study is terminated after observing a predetermined number of failures.

MLE is a procedure of determining an estimator $\hat{\theta}$, of θ which maximizes the likelihood of a function. The likelihood $L(\theta)$ of the observed data is defined as a multiple of the joint distributions of the observed data [9]. The likelihood function of n independent and identically distributed random variables $T_1, T_2, \ldots, T_n$ with pdf $f(t;\theta)$ can be given by,

$$L(\theta, t) = c \prod_{i=1}^{n} f(t_i; \theta). \qquad (2)$$

When the data are right-censored, the likelihood function of the censored sample is given by,

$$L(\theta, t) = c \prod_{i=1}^{n} \left[f(t_i, \theta)^{\delta_i} \right] \left[1 - F(t_i, \theta)^{1-\delta_i} \right]. \qquad (3)$$

Here, $F(t,\theta)$ is the cdf and

$$\delta_i = \begin{cases} 1 & \text{if the component fails at time t} \\ 0 & \text{if the component is censored.} \end{cases}$$

Then, the ML estimator $\hat{\theta}$ is the value which maximizes L over the values of θ in the parameter space and thus, requires a solution to the equation $L'(\theta) = 0$. In MRR, it is assumed that the data are linearly related to the quantiles of the distribution. Firstly, all the observations; both censored and failures, are ranked from the lowest to the highest. The ranks are reversed for the observed failures by, $r_k = n - i_k + 1$, where i_k is the rank of the k^{th} failure among all the ranked observations. Then, the ranks are adjusted for the censored data by computing the rank for the k^{th} failure from the recursive formula,

$$r_k^A = r_{k-1}^A + \frac{n + 1 - r_{k-1}^A}{n + 1 - (k_i - 1)}. \qquad (4)$$

Here, $r_0^A = 0$, and n is the total number of observations. The adjusted ranks are then used to calculate the median rank plotting positions according to the Bernard's formula [1].

$$MR[k] = \frac{r_k^A - 0.3}{n + 0.4}. \qquad (5)$$

Next, a least-squares model is fitted to the observed failures, and then the MRR parameters are estimated.

Results & discussion

A simulation study was designed to compare the two parameter estimation methods in estimating the quantiles of the Weibull distribution when the data are censored. We estimated the 10$^{\text{th}}$ and the 70$^{\text{th}}$ quantiles for complete data, and for the censored data with 90% censoring. The quantiles were selected as a lower quantile and an upper quantile to examine the differences according to the degree of censoring

and sample size. After observing approximately similar patterns for different scale values (the scale parameter only stretches the values across the x-axis), we set the scale parameter to 1000 in the simulation.

The simulation study compromised of the following factors for the censored data.

- one scale value set to 1000.
- four different shape values (1.5, 3, 4, 6).
- five different sample sizes (20, 50, 100, 500, 1000).
- two different quantiles (10^{th}, 70^{th})
- 10% failures (or 90% censoring).
- 1000 replicates.

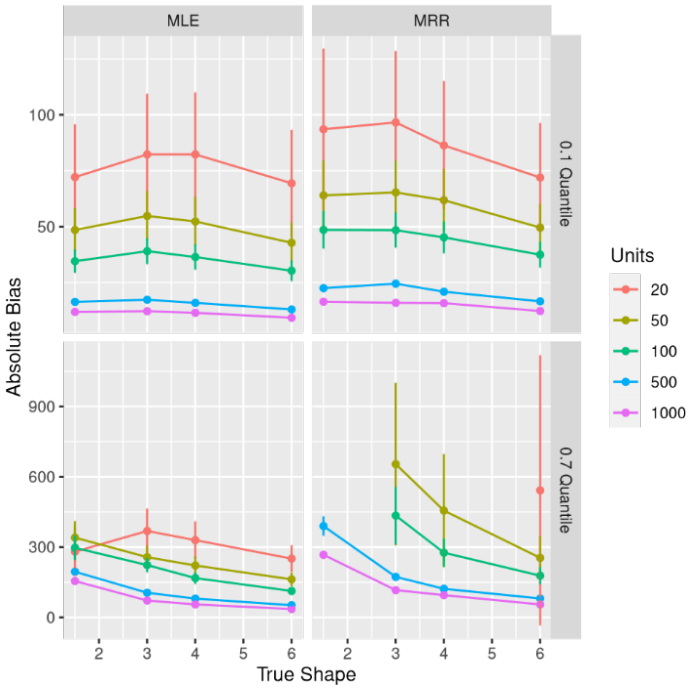

Figure 2: Absolute bias of the estimated 10^{th} and 70^{th} quantiles using MLE and MRR method. The data are 90% censored with only 10% failures. Each point is the mean from 1000 simulations. Absolute biases greater than 1000 are omitted.

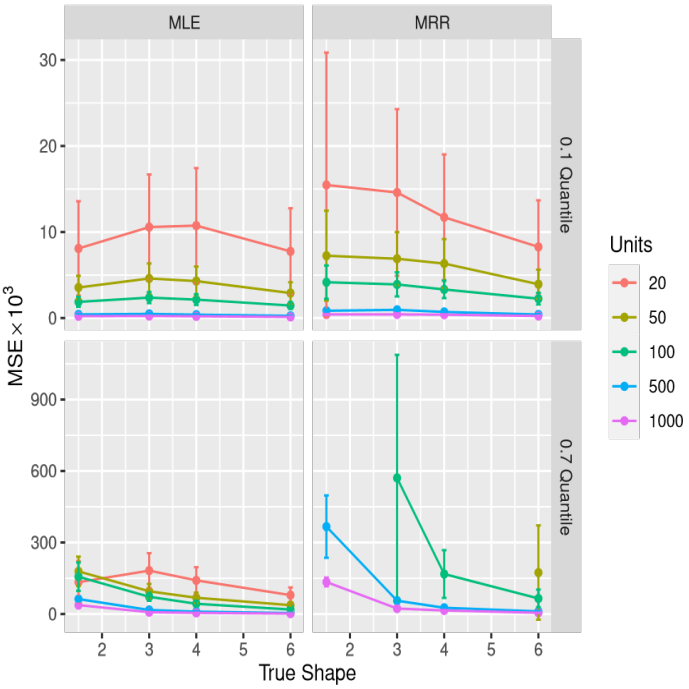

Figure 3: Mean squared error of the estimated 10^{th} and 70^{th} quantiles using MLE and MRR method with 10% fails. Each point is the mean of 1000 simulations. MSE greater than 1000% are omitted

Figure 2 presents the absolute bias of the quantile estimates from both MLE and MRR methods. It appeared that heavy censoring has dramatically affected the absolute bias of the estimates for both quantiles compared to the complete data. The absolute biases of the 70^{th} quantile show noticeably larger values compared to the 10^{th} quantile. MRR method produces higher biases compared to the MLE. Absolute biases decrease towards zero with increasing sample sizes and increasing shape parameter.

Figure 3 illustrates the MSE values computed for both 10^{th} and 70^{th} quantiles from MLE and MRR. MSE values have significantly increased as a result of heavy censoring. Overall, MSE values decrease with increasing sample sizes. Similar to the absolute biases, MRR method has produced higher values of MSE.

It is evident the introduction of heavy censoring has increased the bias and the MSE of the quantile estimates. It appears the reduction of effective sample size by introducing heavy censoring to the data has caused these variations.

Conclusions

When the data are heavily censored with only 10% failures, both absolute bias and the MSE show comparatively larger values. In quantile estimation, MLE outperformed the MRR method, when absolute bias and MSE were considered as measures of precision. Not surprisingly, both MSE and absolute bias values decrease towards zero with increasing sample size. The behaviour of MSE and absolute bias were approximately similar for different values of the scale parameter of the Weibull distribution.

In summary, MLE method produces more reliable quantile estimates in terms of absolute bias and MSE when the data are heavily censored. Additionaly, introduction of heavy censoring to the data affects the robustness of the quantile estimates.

References

1. A. Benard and E. Bos-Levenbach. *The Plotting of Observations on Probability-paper*. Stichting Mathematisch Centrum. Statistische Afdeling, 1955.
2. C. A. Cohen and B. Whitten. Modified maximum likelihood and modified moment estimators for the three-parameter weibull distribution. *Communications in Statistics-Theory and Methods*, 11(23):2631–2656, 1982.
3. L. Dumbgen, K. Rufibach, and D. Schuhmacher.¨ Maximum-likelihood estimation of a log-concave density based on censored data. *Electronic Journal of Statistics*, 8(1):1405–1437, 2014.
4. U. Genschel and W. Q. Meeker. A comparison of maximum likelihood and median-rank regression for weibull estimation. *Quality Engineering*, 22(4):236–255, 2010.
5. S.-L. Jeng and W. Q. Meeker. Comparisons of approximate confidence interval procedures for type i censored data. *Technometrics*, 42(2):135–148, 2000.
6. W. Q. Meeker and L. A. Escobar. *Statistical methods for reliability data*. Wiley series in probability and statistics. Applied probability and statistics section. New York : Wiley, c1998., 1998.
7. D. Olteanu and L. Freeman. The evaluation of median-rank regression and maximum likelihood estimation techniques for a two-parameter weibull distribution. *Quality Engineering*, 22(4):256–272, 2010.
8. B. Reiser and S. B. Lev. Likelihood inference for life test data. *Reliability, IEEE Transactions on*, R28(1):38–43, 1979.
9. P. J. Smith. *Analysis of failure and survival data*. Chapman & Hall/CRC texts in statistical science series. Boca Raton : Chapman & Hall/CRC, c2002., 2002.

Poster presentations

Extracting features from Ecological Audio using Frequency Preserving Autoencoders

Benjamin Rowe
benjamin.rowe@hdr.qut.edu.au
Queensland University of Technology

Continuous audio recordings are playing an ever more important role in conservation and biodiversity monitoring, however, listening to these recordings is often infeasible, as they can be thousands of hours long. Automating analysis using machine learning is in high demand. However, these algorithms require a feature representation.

Several methods for generating feature representations for these data have been developed, using techniques such as domain-specific features and deep learning. However, domain-specific features are unlikely to be an ideal representation of the data and deep learning methods often require extensively labeled data.

We propose a method for generating a frequency-preserving autoencoder-based feature representation for unlabeled ecological audio. We evaluate multiple frequency-preserving autoencoder-based feature representations using a hierarchical clustering sample task. We compare this to a basic autoencoder feature representation, MFCC, and spectral acoustic indices. Experimental results show that some of these non-square autoencoder architectures compare well to these existing feature representations.

Area level estimates of social cohesion in Australia using a Bayesian spatial meta - analysis approach

Dr Farzana Jahan
farzana.jahan@murdoch.edu.au
Lecturer in Statistics, Murdoch University

Social cohesion is the sense of connectedness and unity within communities and is seen as a key influence on individual well-being and on social, economic, and political stability. Surveys that monitor social cohesion have become common tools for researchers and policymakers, and there is increasing need for measures at finer levels of detail than what surveys alone can provide. In this work responses from a national survey of Australian adults were combined with census and other auxiliary data to generate estimates of attitudes to social issues and immigration at the local government area (LGA) level. The paper describes the application of small area statistical methods to deriving measures of social cohesion for local government areas. In particular, two methods were used - the first was a nested error linear regression model utilising the individual level responses and the second was a Bayesian spatial meta-analysis built on top of estimates from the first model. The use of spatial component in the Bayesian hierarchical meta-analysis in estimating area level estimates of social cohesion is a novel application in this field which included the spatial random effects to consider the spatial autocorrelation between score estimates of neighbouring LGAs. This also involved significant data integration from multiple sources of LGA-level data. The study compared the estimates of scores with and without spatial smoothing in terms of validity, meaningfulness, and precision, with a view to recommending methods for future waves of the survey.

On the effectiveness of auxiliary virtual epidemics in epidemic estimation

Aminath Shausan
a.shausan@uq.edu.au
The University of Queensland

The Safe Blues experiment is designed to evaluate how physical interactions over time and between individuals affect the spread of epidemics. In the experiment, the Safe Blues app spreads multiple virtual safe virus strands between mobile phones via Bluetooth. The evolution of the virtual epidemics is recorded as they spread through the mobile phone population.

In this poster, I describe the Safe Blues experiment, the dataset, and some preliminary predictive analysis based on the Safe Blues data. The analysis highlights that using multiple strands (i.e. Safe Blues) provide better predictive performance than the traditional single strand based perditions

Deep learning-based multi-modal data fusion strategies

Duoyi Zhang
zhangduoyi222@gmail.com
QUT

Our world is multi-modal. A scenario or object can be represented by data with multiple modalities. For example, while we post on Twitter, we use text, images, and videos to express our feelings or record a moment. Here, text, image, and video can be identified as different modalities. A critical characteristic of multi-modal data is that different modalities can provide complementary information that can better represent a scenario. However, since the underlying data distributions for each modality are different, combining multi-modal data is a challenging task. In this project, the aim is to explore different fusion strategies and propose novel fusion methods for multi-modal tasks. Specifically, it aims to address the challenges, such as dynamically informative modalities, in multi-modal fusion.

ARDC for ADSN Researchers: How Could we help?

Dr Gnana Bharathy
gnana.bharathy@ardc.edu.au
ARDC

The Australian Research Data Commons (ARDC) is well-positioned to address data challenges across data science lifecycle, as it has a unique depth of expertise in facilitating cross-organizational collaboration to establish national data and infrastructure capabilities.

We believe in the value of research data assets to:

- maximise research quality and impact
- accelerate Australian research and innovation

How: By treating data, software, models, tools etc as a first class asset of the research process. We want to ensure best practice is undertaken in the creation, analysis and retention and reuse of those items.

#IStandWithPutin versus #IStandWithUkraine: The interaction of bots and humans in discussion of the Russia/Ukraine war

Joshua Watt
joshua.watt@adelaide.edu.au
The University of Adelaide

The 2022 Russian invasion of Ukraine emphasises the role social media plays in modern-day warfare, with conflict occurring in both the physical and information environments. There is a large body of work on identifying malicious cyber-activity, but less focusing on the effect this activity has on the overall conversation, especially with regards to the Russia/Ukraine Conflict. Here, we employ a variety of techniques including information theoretic measures, sentiment and linguistic analysis, and time series techniques to understand how bot activity influences wider online discourse. By aggregating account groups we find significant information flows from bot-like accounts to non-bot accounts with behaviour differing between sides. Pro-Russian non-bot accounts are most influential overall, with information flows to a variety of other account groups. No significant outward flows exist from pro-Ukrainian non-bot accounts, with significant flows from pro-Ukrainian bot accounts into pro-Ukrainian non-bot accounts. We find that bot activity drives an increase in conversations surrounding angst (with $p = 2.450 \times 1e\text{-}4$) as well as those surrounding work/governance (with $p = 3.803 \times 1e\text{-}18$). Bot activity also shows a significant relationship with non-bot sentiment (with $p = 3.76 \times 1e\text{-}4$), where we find the relationship holds in both directions. This work extends and combines existing techniques to quantify how bots are influencing people in the online conversation around the Russia/Ukraine invasion. It opens up avenues for researchers to understand quantitatively how these malicious campaigns operate, and what makes them impactful.

A simple approach to cold start learning for image classification using space-filling design, self-supervised and semi-supervised techniques

Nathaniel Bloomfield
nathaniel.bloomfield@unimelb.edu.au
Melbourne Centre for Data Science,
Centre of Excellence for Biosecurity Risk Analysis,
University of Melbourne

In many machine learning applications, curating a sufficiently large labelled dataset can be an expensive and time consuming undertaking. While it has been shown that high accuracy can be achieved with self-supervised and semi-supervised techniques with very few labels, it can be challenging to identify the most informative examples to label in the first place. This choice can have a large impact on the accuracy of the trained image recognition models. In this poster we propose a novel approach to solving this cold start learning problem, using self-supervised learning and space-filling design methods. On CIFAR-10, we show that (1) we obtain significant improvements in performance within the small label regime when using the proposed approach and (2) that these improvements in performance persist when using semi-supervised learning approaches. We also demonstrate how these methods can lead to improved outcomes on a practical application by training models to identify the presence of biofouling on vessel hulls, resulting in improved data labeling efficiency.

.

Transport Reversible Jump Proposals

Laurence Davies
laurence.davies@hdr.qut.edu.au
QUT

Reversible jump Markov chain Monte Carlo (RJMCMC) proposals that achieve reasonable acceptance rates and mixing are notoriously difficult to design in most applications. Inspired by recent advances in deep neural network-based normalizing flows and density estimation, we demonstrate an approach to enhance the efficiency of RJMCMC sampling by performing transdimensional jumps involving reference distributions. In contrast to other RJMCMC proposals, the proposed method is the first to apply a non-linear transport-based approach to construct efficient proposals between models with complicated dependency structures. It is shown that, in the setting where exact transports are used, our RJMCMC proposals have the desirable property that the acceptance probability depends only on the model probabilities. Numerical experiments demonstrate the efficacy of the approach.

Statistical computing with vectorised operations on distributions

Mitchell O'Hara-Wild
mitch.ohara-wild@monash.edu
Monash University

The distributional nature of a model's predictions is often understated, with default output of prediction methods usually only producing point predictions. Some R packages (such as forecast) further emphasise uncertainty by producing point forecasts and intervals by default, however the user's ability to interact with them is limited. To improve this, I have developed the distributional package which allows you to represent multiple parameterised distributions in a vector. This allows data scientists, package developers and educators to directly interact with distributions alongside their data, which is particularly useful for plotting and evaluating predictions and testing hypothesis.

Predictive capabilities in the Livestock Supply Chain

Kalpani Ishara Duwalage
k.duwalage@qut.edu.au
Queensland University of Technology/QUT Center for Data Science

The livestock supply chain, particularly, the cattle industry, significantly contributes to the Gross Domestic Product in Australia. Over recent years, the use of historical data for decision-making has become more popular in the agricultural sector. Currently, in Australia, large volumes of data are collected on cattle farms, but the value of these data is often wasted due to a lack of appropriate analysis tools and insights. Strategies to enhance the profitability and sustainability of production systems are therefore vital. In this research project, we use statistical and machine learning techniques to address some key issues in the industry related to production and re-production.

National Weighted Vulnerability Index Methodology: an Australian Case Study at Fine Temporal and Geographical Resolutions

Aiden Price
a11.price@qut.edu.au
Queensland University of Technology

Extreme natural hazards are increasing in frequency and intensity and continue to have substantial economic, social, and health impacts globally. The impact of the environment on human health (environmental health) is becoming well understood in international research literature.

However, there are significant barriers to understanding key characteristics of this impact, related to substantial data volumes, data access rights, and the time required to compile and compare data over regions and time. This study aims to reduce these barriers by creating an open data repository of national environmental health data and presenting methodology for the production of weekly health outcome weighted population vulnerability indices related to extreme heat, extreme cold, and air pollution at the statistical area level 2 (SA2) geographical resolution.

The weighted vulnerability index methodology employed in this study offers an advantage over others in the literature by targeting health outcomes in the calculation process. The resulting vulnerability percentile more clearly aligns population characteristics with health risks. The temporal and spatial resolutions available enable national monitoring to the public on a scale never before seen across Australia. Additionally, we show that the weekly temporal resolution can be used to identify spikes in vulnerability due to changes in relative national environmental exposure.

A comparison of maximum likelihood estimation and median rank regression method in quantile estimation for Weibull data

Nayomi Attanayake
nayo.attanayake@gmail.com
Murdoch University

Accurate statistical analysis of failure data is vital in vehicle maintenance. Commonly, the Weibull distribution is used as the failure model and reliability measures are estimated. Frequently, in reliability analysis the data are heavily censored because we observe only a few failures. Hence, the robustness of the statistical methods in estimating the quantities is vital in probabilistic decision making. We consider two parameter estimation methods namely; Maximum Likelihood Estimation (MLE) and Median Rank Regression (MRR) method to fit the Weibull distribution to observed data and estimate the quantiles of the distribution.

We perform a set of simulation studies to compare the performance of MLE and MRR in quantile estimation. The data are right-censored with Type-II censoring. The simulations were carried out based on different shape parameters, sample sizes and failure rates (degree of censoring). Absolute bias and mean squared error (MSE) were used to compare the performance of the estimated quantiles. The results show that the MLE outperforms MRR in quantile estimation when the data are heavily censored in terms of MSE and absolute bias.

The Quality Guardian Improving Activity Label Quality in Event Logs through Gamification

Sareh Sadeghianasl
s.sadeghianasl@qut.edu.au
Queensland University of Technology

Data cleaning, the most tedious task of data analysis, can turn into a fun experience when performed through a game. This thesis shows that the use of gamification and crowdsourcing techniques can mitigate the problem of poor quality of process data. The Quality Guardian, a family of gamified systems, is proposed, which exploits the motivational drives of domain experts to engage with the detection and repair of imperfect activity labels in process data. Evaluation of the developed games using real-life data sets and domain experts shows quality improvement as well as a positive user experience.

Invasive species management: to monitor or control?

Thomas Waring
tom.waring@unimelb.edu.au
The University of Melbourne

Management of invasive species is complicated by uncertainty about the size of the population. To reduce uncertainty, we can expend effort to monitor the species, but monitoring typically doesn't reduce the size of the population. Instead, monitoring gives us information. Hence, environmental managers are often faced with the question of how to divide effort between monitoring and control. Building on techniques from Partially Observed Markov Decision Processes (POMDP), we define a general framework for determining whether it is best to monitor or control . Given an estimate (with uncertainty) of the disease abundance, we compute the optimal action, accounting for multiple control interventions, monitoring with varying uncertainty, and interventions which combine the two. The key advance in our work is the generality and broad applicability of the problem framing.

Exploring topic models to discern cyber threats on Twitter: A case study on Log4Shell

Yue Wang
y355.wang@hdr.qut.edu.au
QUT Centre for Data Science

Recent research shows that Twitter has demonstrated advantages in providing timely Cyber Threat Intelligence(CTI) about zero-day vulnerabilities and exploits. However, the overwhelming volume of unstrctured tweets make it difficult for cybersecurity professionals to investigate critical cyber threat incidents. To assist cybersecueity professionals recognize cyber threats and improve the incident response time, we propose a noval Threat Discovery and Monitoring framework(TDM) that can collect and organize unstructured tweets into meaningful topics with dynamic representations and high explainability. A case study on the Log4Shell vulnerability incident was conducted to demonstrate the feasibility of the proposed framework. Results show that the TDM framework can uncover emerging topics related to the realwork cyberthreat incidents. The emerging Log4Shell related topics are found before the well-established public disclosure date by the National Vulnerabilitily Database.

Probabilistic models of functional trajectories for young people with emerging mood and psychotic disorders

Rafael Oliveira
rafael.oliveira@sydney.edu.au
Brain and Mind Centre, University of Sydney

Context: Youth mental health is characterised by a heterogenous pattern of illness which tends to oscillate between health and disorder as a function of vulnerability, protective and treatment factors. In clinical practice this makes predicting an individual's trajectory challenging, particularly at entry into care, and limits the accuracy and appropriateness of allocating the right type and intensity of treatments which could mitigate poor clinical and functional outcomes. This study aimed to develop a probabilistic prediction model which combines an individual's demographic and clinical information at entry into services to determine their most likely social and occupational functional impairment trajectory over the next 3 months.

Methods: The sample includes 718 young people aged between 12 and 25 who had at least one follow-up visit within 3 months of entering youth mental health care. A Bayesian random effects model was designed, where the coefficients of a linear regression model have a probability distribution which is conditioned on a categorical cluster variable describing the change trajectory (i.e., improvement, deterioration or constant). We applied Bayesian inference via Markov chain Monte Carlo (MCMC) using the no-U-turn sampler (NUTS) for continuous random variables combined with a Gibbs sampling scheme to sample the discrete random variables. We considered different choices of covariates for the random effects model, which were evaluated via the Watanabe-Akaike information criterion and evaluated the model's predictive performance on held-out test data via 5- fold cross-validation.

Results: Of the 30 variables available, a subset of 8 was identified as the optimal set for prediction, and included: not in employment, education or training status, self-harm, psychotic-like experiences, physical health comorbidity, childhood development, illness type, clinical stage, and circadian disturbances. The probabilistic clinical prediction tool was internally validated and predicted functional impairment trajectory at 3 months with an area under the curve of 0.70 (SD = 0.1). Self-harm and physical health comorbidity were found to be strong predictors for a deteriorating trajectory.

Conclusions: We have developed and validated a brief probabilistic clinical prediction tool for use in youth mental health settings which can be used to determine functional impairment trajectories over the first three months of care.

Implications: The use of this model in clinical practice has the potential to guide clinical decision-making by assigning probabilities to the range of potential functional impairment trajectories. This type of individual-level prediction (and associated uncertainty regarding predicted outcomes) could enhance the decision about type and intensity of assessment and treatment needed for early intervention.

Counterintuitive Outcomes from PowerShell OS Noise Mitigation

Travis Stenborg

travis.stenborg@sydney.edu.au

ARC Centre in Data Analytics for Resources and Environments (DARE)

Data science tasks typically involve selecting some subset of available computational tools. Examples include choosing between competing platforms, database systems, programming languages, algorithms, data structures, etc. Where operational speed drives choice, benchmarking can distinguish options. Background activity by an operating system can however add noise to such benchmarking, compromising accuracy. Running the data science language R on Windows is examined here, where background processes, services and the Task Scheduler combine to frustrate benchmarking. PowerShell calls can be embedded in R that reduce such activity, expected to improve benchmarking accuracy. An example problem in multicore Bayesian inference was used to evaluate the noise mitigation strategy. The counterintuitive outcome was an increase in runtime variance.

Joint Deep Non-Negative Matrix Factorization for Learning Consensus and Complementary Information for Multi-View Data Clustering

Sohan Gunawardena

sohan.gunawardena@hdr.qut.edu.au

QUT

Exploring Multi-view data has gained a lot of attention nowadays due to the vast availability of digitally stored information. Many real-world data sets illustrate data using multiple representations or views, where each view represents different heterogeneous angles for the given information. Researchers have proposed many algorithms to take maximum advantage of consensus (shared information among multiple views) and complementary (information that is specific to a given view/s) information that is embedded in the multi-view data. Researchers have mainly focused on dimensionality reduction techniques to deal with the problem of identifying intrinsic features of the data. This is mainly due to their ability to identify hidden latent information with higher interpretability and computational efficiency. Among various dimensionality reduction methods, algorithms that are based on non-negative matrix factorization (NMF) have become popular due to their interpretability and part-based representations. Even though there are numerous methods that already use MF and its variants against multi-view data, most of these methods only focus on the shallow linear structure which limits them from capturing complex hierarchical information. Even though there are some Deep MF methods that exist for multi-view data, many of them focus either only on consensus or complementary information at a given time.

To address these limitations, in this work we presented a novel deep MF algorithm, which hierarchically factorizes multi-view data by leveraging the benefit of both consensus and complementary information. Besides, we also impose manifold regularizations on each view to preserve both local and global geometric structures of the original data upon reducing the dimensionality. The proposed method was evaluated on five real-world data sets and demonstrated to significantly outperform the state-of-the-art multi-view learning approaches.