Computational Statistics in Data Science Workshop

Event Details

***Date and Time**:* *5 November 2025, 8.30 AM – 5 PM (Register 8.30 – 8.55 AM)*

***Venue**:* *Building 6, Room 201, University of Wollongong, NSW, 2500*

Registration: *FREE – EventBrite Registration Link*

Keynote speaker: SMAS 2025 Data Science and Statistics Lecturer

Howard Bondell

Professor of Statistical Data Science and Head of the School of Mathematics and Statistics, University of Melbourne

Howard Bondell is Professor of Statistical Data Science in the School of Mathematics and Statistics at the University of Melbourne, where he has been since 2018. Since 2021, Professor Bondell has also served as Head of School.

Professor Bondell received his Ph.D. in Statistics from Rutgers University in 2005 and immediately commenced his academic career in the Department of Statistics at North Carolina State University. His current research interests include model selection, robust estimation, regularisation, Bayesian methods, and all aspects of modelling and handling uncertainty in statistical and machine learning approaches.

Professor Bondell was an ARC Future Fellow from 2020-2024 and was elected a Fellow of the American Statistical Association in 2017. He has served on the editorial boards for Journal of the American Statistical Association, Journal of the Royal Statistical Society, Series B, Biometrics, and Biostatistics.

Program

Time	Session (All sessions in Room 201, Building 6)
8.30 – 8.55	REGISTRATION
8.55 – 9.00	Opening: Noel Cressie (University of Wollongong)
9.00 – 9.30	Andrew Zammit Mangion **Title**: Neural conditional simulation for complex spatial processes
9.30 – 9.50	Pratik Nag **Title**: Spatio-temporal modelling with Fourier neural operators
9.50 – 10.10	Josh Jacobson **Title**: Scalable Bayesian inference for global CO_2 flux attribution with WOMBAT v2.S
10.10 – 10.30	Bao Anh Vu **Title**: Calibration of ice-sheet models using a neural ensemble Kalman filter
10.30 – 11.00	MORNING TEA
11.00 – 11.30	Sumeetpal Singh **Title**: Bayesian learning of the optimal action-value function in a Markov decision process
11.30 – 12.30	**SMAS 2025 Data Science and Statistics Lecture** Professor Howard Bondell **Title**: Data and decision-making: Informative missingness, recommender systems, and personalised medicine
12.30 – 14.00	LUNCH RECEPTION
14.00 – 14.20	Anjana Wijayawardhana **Title**: Non-Gaussian simultaneous autoregressive models with missing data
14.20 – 14.40	Daniel Fynn **Title**: Neural inference function for margins for time series copula models
14.40 – 15.00	Vinicius Riffel **Title**: Neural data augmentation for parameter inference from incomplete data
15.00 – 15.30	AFTERNOON TEA
15.30 – 15.50	Bradley Wakefield **Title**: Integrating geophysical ground models with Bayesian spatial mixture modelling for seabed-sediment characterisation
15.50 – 16.10	Laurence Davies **Title**: Amortised variational transdimensional inference
16.10 – 16.40	David Gunawan **Title**: Fast variational boosting for latent-variable models
16.40 – 16.45	Closing: Noel Cressie

Abstracts of Presentations

1. Neural conditional simulation for complex spatial processes

Title: Neural conditional simulation for complex spatial processes

Presenter: Andrew Zammit-Mangion (Joint work with Julia Walchessen (CMU), Raphaël Huser (KAUST), and Mikael Kuusela (CMU))

Abstract: A key objective in spatial statistics is to simulate from the distribution of a spatial process at a selection of unobserved locations conditional on observations (i.e., a predictive distribution) to enable spatial prediction and uncertainty quantification. However, exact conditional simulation from this predictive distribution is intractable or inefficient for many spatial process models. In this paper, we propose neural conditional simulation (NCS), a general method for spatial conditional simulation that is based on neural diffusion models. Specifically, using spatial masks, we implement a conditional score-based diffusion model that evolves Gaussian noise into samples from a predictive distribution when given a partially observed spatial field and spatial process parameters as inputs. The diffusion model relies on a neural network that only requires unconditional samples from the spatial process for training. Once trained, the diffusion model is amortized with respect to the observations in the partially observed field, the number and locations of those observations, and the spatial process parameters, and can therefore be used to conditionally simulate from a broad class of predictive distributions without retraining the neural network. We assess the NCS-generated simulations against simulations from the true conditional distribution of a Gaussian process model, and against Markov chain Monte Carlo (MCMC) simulations from a Brown--Resnick process model for spatial extremes. In the latter case, we show that it is more efficient and accurate to conditionally simulate using NCS than classical MCMC techniques implemented in standard software. We conclude that NCS enables efficient and accurate conditional simulation from spatial predictive distributions that are challenging to sample from using traditional methods.

2. Spatio-temporal modelling with Fourier neural operators

Title: Spatio-temporal modelling with Fourier neural operators

Presenter: Pratik Nag (Joint work with Andrew Zammit-Mangion, Sumeetpal Singh, and Noel Cressie)

Abstract: Spatio-temporal process models are often used for modelling dynamic physical and biological phenomena that evolve across space and time. These phenomena often exhibit environmental heterogeneity and complex interactions that are difficult to capture using traditional statistical process models such as Gaussian processes. This work proposes the use of Fourier neural operators (FNOs) for constructing spatio-temporal models for forecasting. An FNO is a flexible mapping of functions that approximates the solution operator of non-linear partial differential equations (PDEs) in a computationally efficient manner. It does so using samples of inputs and their respective outputs, without requiring explicit knowledge of the underlying PDE.

Here, we introduce FNO-based models within a hierarchical statistical framework, and through extensive simulations we compare their forecasts to those from state-of-the-art spatio-temporal-forecasting methods. Using sea surface temperature data over the Atlantic Ocean and precipitation data across Europe, we demonstrate the ability of FNO-based hierarchical statistical modelling to capture complex real-world spatiotemporal dependencies, and to provide accurate and valid probabilistic forecasts.

3. Scalable Bayesian inference for global CO2 flux attribution with WOMBAT v2.S

Title: Scalable Bayesian inference for global CO_2 flux attribution with WOMBAT v2.S

Presenter: Josh Jacobson (Joint work with Michael Bertolacci (UWA), Andrew Zammit-Mangion (UOW), Andrew Schuh (CSU), and Noel Cressie (UOW))

Abstract: Contributions from photosynthesis and other natural components of the carbon cycle present the largest uncertainties in our understanding of carbon dioxide (CO_2) sources and sinks. While the spatio-temporal distribution of Earth's net surface flux (the sum of all contributions) can be inferred from atmospheric CO_2 concentrations through flux inversion, attributing the net flux to its individual components remains challenging. The advent of solar-induced fluorescence (SIF) satellite observations provides an opportunity to separate natural components by isolating gross primary productivity (GPP), the photosynthetic component of the net flux. Here, we introduce a novel statistical flux-inversion framework that simultaneously assimilates millions of observations of SIF and CO_2 concentration, extending WOMBAT v2.0 (WOllongong Methodology for Bayesian Assimilation of Trace-gases, version 2.0) with a hierarchical model of spatio-temporal dependence between GPP and SIF processes. We call the new framework WOMBAT v2.S, and we apply it to SIF and CO_2 data from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite and other instruments to estimate natural fluxes over the globe during a recent six-year period. MCMC inference on this high-dimensional inverse problem relies on high-performance computing and several computational strategies: we employ a basis-function decomposition for dimension reduction and implement an exact Hamiltonian Monte Carlo algorithm for multivariate constrained Gaussian distributions, which ensures that physical constraints on the component fluxes are properly enforced. In a simulation experiment, the inclusion of SIF improves accuracy and uncertainty quantification of component flux estimates. Comparing estimates from WOMBAT v2.S and v2.0, we observe that linking GPP to SIF leads to spatial redistribution and amplified seasonal structure in natural flux components.

4. Calibration of ice-sheet models using a neural ensemble Kalman filter

Title: Calibration of ice-sheet models using a neural ensemble Kalman filter

Presenter: Bao Anh Vu

Abstract: The Antarctic ice sheet (AIS) is the largest freshwater reservoir on Earth, and a major contributor to sea level rise. Ice loss from the AIS will have significant impact on the biodiversity of Antarctica and its surrounding oceans as well as other coastal communities, including Australia. Therefore, estimates of the AIS' contribution to the sea level budget are of paramount importance.

Ice sheet models are used routinely to quantify and forecast the AIS' contribution to sea level rise. For an ice sheet model to generate valid forecasts, its parameters must first be calibrated using observational data; this is challenging due to the model's high degree of nonlinearity, its complex parameterisation, and limited data availability. This study leverages the emerging field of neural posterior estimation for efficiently calibrating ice sheet parameters, namely the bed elevation and basal friction, based on velocity and surface elevation data. Samples from the approximate posterior distribution are then used with an ensemble Kalman filter to infer ice thickness and velocity in a 1D Shallow-Shelf Approximation (SSA) model. We apply our approach to infer the bed elevation and basal friction along a transect in Thwaites Glacier, Antarctica. This work is joint research with Andrew Zammit-Mangion (UOW), David Gunawan (UOW), Felicity McCormack (Monash University), and Noel Cressie (UOW).

5. Bayesian learning of the optimal action-value function in a Markov decision process

Title: Bayesian learning of the optimal action-value function in a Markov decision process

Presenter: Sumeetpal Singh (Joint work with Jiaqi Guo (University of Cambridge) and Chon Wai Ho (University of Cambridge))

Abstract: The Markov Decision Process (MDP) is a popular framework for sequential decision-making problems, and uncertainty quantification is an essential component of it to learn optimal decision-making strategies. In particular, a Bayesian framework is used to maintain beliefs about the optimal decisions and the unknown ingredients of the model, which are also to be learned from the data, such as the rewards and state dynamics. We focus on infinite-horizon and undiscounted MDPs, with finite state and action spaces, and a terminal state. We provide a full Bayesian framework, from modelling to inference to decision-making. For modelling, we introduce a likelihood function with minimal assumptions for learning the optimal action-value function based on Bellman's optimality equations, analyse its properties, and clarify connections to existing works. For deterministic rewards, the likelihood is degenerate and we introduce artificial observation noise to relax it, in a controlled manner, to facilitate more efficient Monte Carlo-based inference. For inference, we propose an adaptive sequential Monte Carlo algorithm to both sample from and adjust the sequence of relaxed posterior distributions. For decision-

making, we choose actions using samples from the posterior distribution over the optimal strategies. While commonly done, we provide new insight that clearly shows that it is a generalisation of Thompson sampling from multi-arm bandit problems. Finally, we evaluate our framework on the Deep-Sea benchmark problem and demonstrate the exploration benefits of posterior sampling in MDPs.

6. SMAS 2025 Data Science and Statistics Lecture - Data and decision-making: Informative missingness, recommender systems, and personalised medicine

Title: Data and decision-making: Informative missingness, recommender systems, and personalised medicine

Presenter: Howard Bondell (University of Melbourne)

Abstract: In this talk, we will discuss two topics associated with the use of data for decision-making.

The first part of the talk investigates informative missingness in the framework of recommender systems. In this setting, we envision a potential rating for every object-user pair. The goal of a recommender system is to predict the unobserved ratings and then recommend an object that the user is likely to rate highly. A typically overlooked piece is that the data is not missing at random. For example, in movie ratings, a relationship between the user ratings and their viewing history is expected, as human nature dictates the user would seek out movies that they anticipate enjoying. We model this informative missingness, and place the recommender system in a shared-variable regression framework which can aid in prediction quality.

The second part of the talk deals with personalised medicine, which relies on the ability to prescribe patient-specific treatments. In this context, it is crucial to identify the variables that impact the optimal treatment decision. Typical variable selection techniques target on selecting variables that are important for prediction, which are not necessarily those that are important for treatment assignment. We propose a framework to directly target the variables related to the optimal decision rule to thus identify the important variables in treatment decision making.

7. Non-Gaussian simultaneous autoregressive models with missing data

Title: Non-Gaussian simultaneous autoregressive models with missing data

Presenter: Anjana Wijayawardhana (joint work with David Gunawan and Thomas Suesse)

Abstract: Standard simultaneous autoregressive (SAR) models typically assume normally distributed errors, an assumption often violated in real-world datasets that frequently exhibit non-normal, skewed, or heavy-tailed characteristics. New SAR models are proposed to capture these non-Gaussian features. The spatial error model (SEM), a widely used SAR-type model, is considered. Three novel SEMs are introduced that extend the standard Gaussian SEM by incorporating Student's t-distributed errors after a one-to-one transformation is applied to the response variable. Variational Bayes (VB) estimation methods are developed for these models,

and the framework is further extended to handle missing response data. Standard variational Bayes (VB) methods perform well with complete datasets; however, handling missing data requires a Hybrid VB (HVB) approach, which integrates a Markov chain Monte Carlo (MCMC) sampler to generate missing values. The proposed VB methods are evaluated using both simulated and real-world datasets, demonstrating their robustness and effectiveness in dealing with non-normal data and missing data in spatial models. Although the method is demonstrated using SAR models, the proposed model specifications and estimation approaches are widely applicable to various types of models for handling non-Gaussian data with missing values.

8. Neural inference function for margins for time series copula models

Title: Neural inference function for margins for time series copula models

Presenter: Daniel Fynn (Joint work with David Gunawan and Andrew Zammit-Mangion)

Abstract: Copula models are widely used for multivariate time series data because they allow the modeling of marginal distributions separately from the dependence structure, which is captured by the copula function. Implicit copulas are the most commonly used approach for modeling dependence in high-dimensional settings. This broad class includes various types of copulas, such as elliptical copulas and skew t-copulas. Estimation of these copula models in high dimensions is challenging. Motivated by the classical inference function for margins approach, which first estimates the parameters of the marginal models and then estimate the copula parameters, we propose a new approach based on neural amortised inference, which we call neural inference function on margins. In the first step, the method uses neural networks to obtain an approximate posterior distribution of the marginal models. Given the estimated marginal parameters, we use another neural network model to obtain the approximate posterior distribution of the copula parameters in the second step. This approach provides rapid parameter estimation given new data, fast sequential prediction and efficient model comparisons using time-series cross validation. The proposed approach is evaluated using simulated and real datasets and compared to the popular Hamiltonian Monte Carlo method.

9. Neural data augmentation for parameter inference from incomplete data

Title: Neural data augmentation for parameter inference from incomplete data

Presenter: Vinicius Riffel

Abstract: Estimating parameters in high-dimensional hierarchical models from missing data is often computationally prohibitive. Classical methods like the Tanner and Wong (1987) data augmentation algorithm rely on iterative Markov chain Monte Carlo (MCMC) sampling, which can be too slow for practical use in time-sensitive, real-world settings. Here, we propose an approach that accelerates this classical algorithm by replacing its most intensive steps with neural networks. Specifically, our hybrid strategy replaces the computationally expensive MCMC parameter update step with a rapid forward pass through a neural network trained to

approximate the posterior distribution. The proposed method is showcased on a complex, high-dimensional spatial statistical model used for inferring the three-dimensional geotechnical properties of subsea sediments. Our results show that our neural data augmentation algorithm achieves high accuracy while being substantially faster than traditional MCMC approaches.

Reference: Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

10. Integrating geophysical ground models with Bayesian spatial mixture modelling for seabed-sediment characterisation

Title: Integrating geophysical ground models with Bayesian spatial mixture modelling for seabed-sediment characterisation

Presenter: Bradley Wakefield (Joint work with Andrew Zammit Mangion (UOW), David Gunawan (UOW), Michael Bertolacci (UWA), Fraser Bransby (UWA), and Phil Watson (UWA))

Abstract: This talk presents GeoMix, a Bayesian spatial mixture model developed to integrate geotechnical cone penetrometer test (CPT) data with interpreted geophysical ground models for offshore sediment characterisation. Building on the warped Gaussian process framework of GeoWarp (Bertolacci et al., 2025)[1], GeoMix introduces latent sediment classes with class-specific depth–strength profiles and a Markov random field prior to encourage spatial coherence. The model is fitted via MCMC, enabling joint posterior inference on both class labels and geotechnical responses. Applied to the Ijmuiden Ver wind farm dataset (238 CPTs, 437 seismic lines, 422 km²), GeoMix improves predictive accuracy for cone resistance and better resolves near-surface stratification, while providing uncertainty quantification around geological boundaries. This work illustrates how spatial mixture modelling and hierarchical computation can enhance offshore site investigations by integrating heterogeneous data sources and applying principled computational methods to support scalable and valid statistical inference.

[1]Bertolacci, Michael, et al. "GeoWarp: Warped spatial processes for inferring subsea sediment properties." *Journal of the American Statistical Association* (2025): 1-22.

11. Amortised variational transdimensional inference

Title: Amortised variational transdimensional inference

Presenter: Laurence Davies (Joint work with Dan Mackinlay (CSIRO/data61), Rafael Oliveira (CSIRO/data61), and Scott Sisson (UNSW))

Abstract: The expressiveness of flow-based models combined with stochastic variational inference (SVI) has, in recent years, expanded the application of optimization-based Bayesian inference to include problems with complex data relationships. However, until now, SVI using flow-based models has been limited to problems of fixed dimension. We introduce CoSMIC

normalizing flows (COntextually-Specified Masking for Identity-mapped Components), an extension to neural autoregressive conditional normalizing flow architectures that enables using a single amortized variational density for inference over a transdimensional target distribution. We propose a combined stochastic variational transdimensional inference (VTI) approach to training CoSMIC flows using techniques from Bayesian optimization and Monte Carlo gradient estimation. Numerical experiments demonstrate the performance of VTI on challenging problems that scale to high-cardinality model spaces.

12. Fast variational boosting for latent-variable models

Title: Fast variational boosting for latent-variable models

Presenter: David Gunawan (Joint work with David Nott (NUS) and Robert Kohn (UNSW))

Abstract: We consider the problem of estimating complex statistical latent variable models using variational Bayes methods. These methods are used when exact posterior inference is either infeasible or computationally expensive, and they approximate the posterior density with a family of tractable distributions. The parameters of the approximating distribution are estimated using optimisation methods. This article develops a flexible Gaussian mixture variational approximation, where we impose sparsity in the precision matrix of each Gaussian component to reflect the appropriate conditional independence structure in the model. By introducing sparsity in the precision matrix and parameterising it using the Cholesky factor, each Gaussian mixture component becomes parsimonious (with a reduced number of non-zero parameters), while still capturing the dependence in the posterior distribution. Fast estimation methods based on global and local variational boosting moves combined with natural gradients and variance reduction methods are developed. The local boosting moves adjust an existing mixture component, and optimisation is only carried out on a subset of the variational parameters of a new component. The subset is chosen to target improvement of the current approximation in aspects where it is poor. The local boosting moves are fast because only a small number of variational parameters need to be optimised. The efficacy of the approach is illustrated by using simulated and real datasets to estimate generalised linear mixed models and state space models.