Biometrics By The Border

International Biometrics Society Australasian Region ConferenceKingscliff, NSW

26-30th November 2017

Contents

Welcome

This is the very first part of the book.

Programme And Abstracts For Monday 27th Of November

Keynote: Monday 27th 9:40 Mantra

Agricultural And Agri-Environmental Statistics With Support Of Geospatial Information, Methodological Issues

Elisabetta Carfagna University of Bologna

Agri-environmental trade-offs are issues critical for policy makers charged with managing both food supply and the sustainable use of the land. Reliable data are crucial for developing effective policies and for evaluating their impact. However, often the reliability of agricultural and agro-environmental statistics is low.

Due to the technological development, in the last decades, different kinds of geospatial data have become easily accessible at decreasing prices and have started to be an important support to statistics production process.

In this paper, we focus on methodological issues related to the use of geospatial information for sampling frame construction, sample design, stratification, ground data collection and estimation of agricultural and agri-environmental parameters. Particular attention is devoted to the impact of spatial resolution of data, change of support aggregation and disaggregation of spatial data, when remote sensing data, Global Positioning Systems and Geographic Information Systems (GIS) are used for producing agricultural and agro-environmental statistics.

Monday 27th 11:00 Narrabeen

Developing A Regulatory Definition For The Authentication Of Manuka Honey

Claire McDonald[1], Suzanne Keeling[1], Mark Brewer[2], and Steve Hathaway[1] [1]Ministry for Primary Industries [2]BioSS

Manuka honey is a premium export product from New Zealand that has been under scrutiny due to claims of fraud, adulteration and mislabelling. Although there are several industry approaches for defining manuka honey, there is currently no scientifically robust definition suitable for use in a regulatory setting. As such, ensuring the authenticity of manuka honey is challenging.

Here we present the results of a three year science programme which developed scientifically robust definitions for monofloral and multifloral manuka honey produced in New Zealand. The programme involved: selecting

appropriate markers to identify honey sourced from Leptospermum scoparium (manuka), establishing plant and honey reference collections, developing test methods to determine the levels of the markers and analysing the data generated to develop the definitions.

The suitability of 16 markers (chemical and DNA-based) were evaluated for use in a regulatory definition for manuka honey. Plant samples were collected from two flowering seasons representing both manuka and non-manuka species from both New Zealand and Australia. Honey samples, also representing manuka and non-manuka floral types, were sourced from seven New Zealand production seasons. Additionally, honey samples were sourced from another 12 countries to enable comparison. All samples were tested for the markers being evaluated using the developed test methods.

The method of CART (Classification and Regression Trees) was used to develop the monofloral and multifloral manuka honey definitions. The CART outputs were further processed using a simulation approach to determine the sensitivity and the robustness of the definitions. The definitions use a combination of 5 markers (4 chemical and 1 DNA) at set thresholds to classify a sample as manuka honey or otherwise. We discuss the practicalities of using the science-based definitions within a regulatory context.

Monday 27th 11:00 Gunnamatta

On Testing Random Effects In Linear Mixed Models

Alan Welsh[1], Francis Hui[1], and Samuel Mueller[2] [1]ANU [2]University of Sydney

We can approach problems involving random effects in linear mixed models directly through the random effects or through parameters such as that the variance components that describe the distributions of the random effects. Both approaches are useful but lead to different issues. For example, working with the random effects raises questions of how to estimate them and can mean dealing with a large number of random effects, while working with variance components leads to testing hypotheses on the boundary of the parameter space. In both cases, finding good approximations to the null distribution of the test statistic can be challenging so modern approaches often rely on simulation. In this talk, we re-examine the F-test based on linear combinations of the responses, for testing random effects in linear mixed models. We present a general derivation of the test, highlight its computation speed, its generality, and its exactness as a test, and report empirical studies into the finite sample performance of the test. We conclude the presentation by reporting our latest results from ongoing research that investigates connections between testing and model selection by discussing some tests of significance of random effects and exploring their relationship to model selection procedures.

Monday 27th 11:20 Narrabeen

Analysing Digestion Data

Maryann Staincliffe, Debbie Frost, Mustafa Farouk, and Guojie Wu AgResearch

Food technologist are interested in understanding if adding grain and/or vegetables to beef increases the digestibility over meat alone. Digestibility of meat was measured at up to 4 hours using a pepsin and pancreatin in vitro model. This method involves using gels in lanes, where the density of the colour of the gel is an indicator of the presence of a protein or peptide at that level of kilodalton (kDa). Typically kDa above 12kDa are considered to be the proteins and lower than that are the peptides. In the past we have analysed these data using two approaches. The first approach is to select 4 or 5 bands that expected to be important and then use a mixed effects model to compare the mean Trace Quantity of each of the bands, where Meat type, Additive type (vegetable and/or grain) and Time points are fixed factors and the gel is a random effect. The second approach is to fit a curve to the change in the proportion of Trace Quantity above 12kDa from Time zero. The problem with these two approaches is that we struggle to provide a consistent interpretation of the results. Therefore, we will explore alternative methods for analysing this type of data.

Monday 27th 11:20 Gunnamatta

Mixed Models For Complex Survey Data

Xudong Huang and Thomas Lumley University of Auckland

I want to fit a mixed model to a population distribution, but I have data from a complex (multistage) sample. The sampling is informative, that is, the model holding for the population is different from the model holding for the (biased) sample. Ignoring the sampling design and just fitting the mixed model to the sample distribution will lead to biased inference. Although both the model and sampling involve " clusters" , the model clusters and sample clusters need not be the same. I will use a pairwise composite likelihood method to estimate the parameters of the population model under this setting. In particular, consistency and asymptotic normality can be established. Variance estimation in this problem is challenging. I will talk about a variance estimator and how to show it is consistent.

Monday 27th 11:40 Narrabeen

Challenges And Opportunities Working As A Consulting Statistician With A Food Science Research Group

M. Gabriela Borgognone Queensland Department of Agriculture and Fisheries

When an established research group has been functioning for many years without a statistician as an integral part of the team, welcoming one into the group can present challenges as well as opportunities for all involved.

Challenges for the research group include, for example, involving the statistician at the beginning of the study instead of once the experiments have been completed and the data collected; acquiring or increasing knowledge of experimental design principles; understanding the limitations of some statistical analyses, expanding the range of methods they feel familiar with, and learning when/how to apply each one; and improving the presentation of results in this era where poor presentation is perpetuated by the general lack of sound statistical methods in the literature of the research area. Challenges for the statistician include, for example, overcoming his/her lack of general knowledge of the underlying scientific area and its specific vocabulary; determining what experimental designs would work from a practical point of view; developing understanding of their scientific questions, data management practices, and types of data collected; navigating the various software they use and checking their adequacies and limitations; and, above all, communicating with patience and perseverance.

Correspondingly, all challenges present opportunities for improvement and collaboration between scientists and statisticians. Working as a team supports a decision making process that is relevant to industry and that is based on good statistical practices. Additionally, it helps scientists become more statistically aware and empowered. A bit more than a year ago I started working as a consulting statistician with a food science research group. In this presentation I will share some of the challenges and the opportunities to incorporate good statistical practice I have identified, as well as some of the improvements we have made so far working together in this partnership.

Monday 27th 11:40 Gunnamatta

Robust Semiparametric Inference In Random Effects Models

Michael Stewart[1] and Alan Welsh[2] [1]University of Sydney [2]ANU

We report on recent work using semiparametric theory to derive procedures with desirable robustness and efficiency properties in the context of inference concerning scale parameters for random effect models.

Monday 27th 12:00 Narrabeen

Robust Penalized Logistic Regression Through Maximum Trimmed Likelihood Estimator

Hongwei Sun, Yuehua Cui, and Tong Wang Shanxi Medical University

Penalized logistic regression is used to identify genetic markers for many high-dimensional datasets such as in gene expression, GWAS, DNA methylation studies and so on. But outliers sometimes occur due to missed diagnosis or misdiagnosis of subjects, heterogeneity of samples,technical problems in experiments or other problems. They can greatly influence the estimation of penalized logistic regression. Few studies focus on the robustness of penalized methods when the response variable is categorical, which is standard in medical research. This study proposed a robust LASSO-type penalized logistic regression based on maximum trimmed likelihood(MTL-LASSO). The definition of breakdown point (BDP) for penalized logistic regression was given and its property for the proposed method was proved. A modification of FAST-LTS algorithms was used to implement the estimation. The reweighted step was added to improve performance while guaranteeing robustness. The simulation study shows the proposed method can resist against outliers. A real dataset about gene expression profiles of multiple sclerosis patients and healthy controls was analysed. Outliers in the control group identified by reweighted MTL-LASSO behave differently from others. It unveils there may be heterogeneity problem in control group. A much better fit is obtained after removing outliers.

Keynote: Monday 27th 13:30 Mantra

A Multi-Step Classifier Addressing Cohort Heterogeneity Improves Performance Of Prognostic Biomarkers In Complex Disease

Jean Yang University of Sydney

Recent studies in cancer and other complex diseases continue to highlight the extensive genetic diversity between and within cohorts. This intrinsic heterogeneity poses one of the central challenges to predicting patient clinical outcome and the personalization of treatments. Here, we will discuss the concept of *classifiability* observed in multi-omics studies where individual patients' samples may be considered as either *hard* or *easy* to classify by different platforms, reflected in moderate error rates with large ranges. We demonstrate in a cohort of 45 AJCC stage III melanoma patients that clinico-pathologic biomarkers can identify those patients that are most likely to be misclassified by a molecular biomarker. The process of modelling the classifiability of patients was then replicated in independent data from other diseases.

A multi-step procedure incorporating this information not only improved classification accuracy overall but also indicated the specific clinical attributes that had made classification problematic in each cohort. In statistical terms, our strategy models cohort heterogeneity via the identification of interaction effects in a high dimensional setting. At the translational level, these findings show that even when cohorts are of moderate size, including features that explain the patient-specific performance of a prognostic biomarker in a classification framework can significantly improve the modelling and estimation of survival, as well as increase understanding.

Monday 27th 14:20 Narrabeen

How To Analyse Five Data Points With Fun

Pauline O'Shaughnessy[1], Stephen Robson[2], and Louise Rawlings[2] [1]University of Wollongong [2]ANU

While "big data" is one of the biggest buzzwords, we occasionally come cross data with very few data points. This dataset is from a study of the incidence rates of commonly performed medical procedures In Australia, which is available only at the state level. We have five states in Australia thus five data points. So what can we do when we only have five data points? One of the novelty approaches is to fit the data with a regression model. However given the challenging nature of the data with small size, no guarantee can be placed on the satisfactory of the linearity and homoscedasticity assumption of the linear regression, in turns, the inference from the standard linear model theory is no longer valid. Double bootstrap is used to provide solution to the valid statistical inference for the best linear approximation for the relationship between variables in an assumption-lean regression setting.

Monday 27th 14:20 Gunnamatta

An Approach To Poisson Mixed Models For -Omics Expression Data

Irene Suilan Zeng and Thomas Lumley University of Auckland

We are interested in regression models for multivariate data from high-throughput biological assays ('omic' data). These data have correlations between variables, and may also come from structured experiments, so a generalised linear mixed model is appropriate to fit the experimental variables and different types of omics data. However, the number of variables is often larger than the number of observations: a structured covariance model is necessary and sparsity induction is biologically appropriate. In this presentation we describe an approach to Poisson mixed models, suitable for RNAseq gene expression data, based on transcript-specific random effects with a sparse precision matrix. We show by simulations that the optimal sparseness penalty for regression modelling is not the same as in the usual graph estimation problem and compare some estimation strategies in simulations.

Monday 27th 14:40 Narrabeen

Bayesian Spatial Estimation When Areas Are Few

Aswi Aswi, Susanna Cramb, Earl Duncan, and Kerrie Mengersen Queensland University of Technlogy

Spatial modelling when there are few (< 20) small areas can be challenging. Bayesian methods can be beneficial in this situation due to the ease of specifying structure and additional information through priors. However, care is needed as there are often fewer neighbours and more edges, which may influence results. Here we investigate Bayesian spatial model specification when there are few areas, first through a simulation study (number of areas ranging from 4 to 2500) and then apply to a case study on dengue fever in 2015 in Makassar, Indonesia (14 areas). Four different Bayesian spatial models namely, an independent model and 3 models based on a CAR (Conditional Autoregressive) prior: the Besag, York & Mollié, Leroux, and a localised model (augments the CAR prior with a cluster model using piecewise constant intercepts) were applied. Data were generated for the simulation study considering low and high spatial autocorrelation and low and high disease incidence. Model goodness of fit was compared using Deviance Information Criteria. Analysis of variance and Bonferroni's method were also used to determine which models were significantly different. The simulation study showed models differed in their performance mainly in two situations: 1. When there were at least 25 areas and both the disease rate and spatial autocorrelation was low, and 2. For all area sizes when there was low spatial autocorrelation but a high overall disease rate. Likewise, results

from the case study showed that all four models performed similarly. This is probably due to the low number of areas and a low disease incidence.

Monday 27th 14:40 Gunnamatta

Knowledge-Guided Generalized Biclustering Analysis For Integrative –Omics Analysis

Changgee Chang[1], Yize Zhao[2], Mingyao Li[1], and Qi Long[1] [1]University of Pennsylvania [2]Cornell University

Advances in technology have enabled generation of multiple types of -omics data in many biomedical and clinical studies, and it is desirable to pool such data in order to improve the power of identifying important molecular signatures and patterns. However, such integrative analyses present new analytical and computational challenges. To address some of these challenges, we propose a Bayesian sparse generalized biclustering analysis (GBC) which enables integrating multiple omics modalities with incorporation of biological knowledge through the use of adaptive structured shrinkage priors. The proposed methods can accommodate both continuous and discrete data. MCMC and EM algorithms are developed for estimation. Numerical studies are conducted to demonstrate that our methods achieve improved feature selection and prediction in identifying disease subtypes and latent drivers, compared to existing methods.

Monday 27th 15:00 Narrabeen

Understanding The Variation In Harvester Yield Map Data For Estimating Crop Traits

Dean Diepeveen[1], Karyn Reeves[1], Adrian Baddeley[1], and Fiona Evans[2] [1]Curtin University [2]Murdoch University

Our recent exploratory research involves extracting tangible crop traits from images such as yield-maps. Research by Diepeveen et al (2012) demonstrated that genetic information can be extracted from near-infrared (NIR) images using both the implicit knowledge of data, environmental data and using a multivariate approach. Yield map data is geo-referenced data of grain-yield that is generated by a harvester cutting the crop, threshing the straw and extracting the grain. Our results show that the yield-map data has significant issues associated with it. One issue is the delay from time of cutting the crop to entering into the storage-bin for measurement. This is dependent on speed of the harvester and is compounded with maintaining a critical volume going through the harvester to operate efficiency. There are also issues with variation from the density and plant size of the crop within the paddock being harvested. Our preliminary results just highlight the significant challenges in extracting precise crop traits from yield map data.

Monday 27th 15:00 Gunnamatta

Citizen Science To Surveillance: Estimating Reporting Probabilities Of Exotic Insect Pests

Peter Caley[1], Marijke Welvaert[2], and Simon Barry[1] [1]CSIRO [2]University of Canberra

Up until mid-2016, citizen science uploads to the Atlas of Living Australia included c. 400 bug species, and c. 1,000 beetle species. Given the short time period (c. 3 years) over which most of these records have accumulated, this represents a considerable reporting effort. The key applied question from a biosecurity context is how this level of reporting translates to the detection and reporting of exotic insect pests in the event of an incursion.

We use a case-control design to model the probability of existing insect species being reported via citizen science channels feeding into the Atlas of Living Australia. The effect of insect features (size, colour, pattern, morphology) and geographic distribution on reporting rates are explored as explanatory variables. We then apply the model to exotic high priority pest species to predict their reporting rates in the event of their introduction.

Monday 27th 15:50 Narrabeen

Introduction To "Deltagen" - A Comprehensive Decision Support Tool For Plant Breeders Using R And Shiny

Dongwen Luo and Zulfi Jahufer AgResearch

The objective of this presentation is to introduce a unique new plant breeding decision support software tool "DeltaGen", implemented in R and its package Shiny. DeltaGen provides plant breeders with a single integrated solution for experimental design generation, data quality control, statistical and quantitative genetic analyses, breeding strategy evaluation/simulation and cost analysis, pattern analysis, index selection and underlying basic theory on quantitative genetics. This software tool could also be used as a teaching resource in plant breeding courses. DeltaGen is available as Freeware on the link: http://agrubuntu.cloudapp.net/shiny-apps/PlantBreedingTool/

Monday 27th 15:50 Gunnamatta

Estimating Nitrous Oxide Emission Factors

Alasdair Noble and Tony Van Der Weerden AgResearch

Nitrous Oxide (N2O) is an important greenhouse gas with a global warming potential nearly 300 times that of carbon dioxide. Under the Kyoto Protocol New Zealand is required to report a greenhouse gas inventory annually which includes N2O. In New Zealand, 95% of N2O emissions are derived from nitrogen (N) inputs to agricultural soils (e.g. animal excreta and fertiliser). Field experiments are conducted to estimate these N2O emissions, where data is collated and analysed following a standard methodology to determine emission factors, which estimate the amount of N2O lost per unit of N applied to soil. However for individual datasets there are some aspects of the data are incompatible with the proposed model so some ad hoc adjustments are made. A more rigorous Bayesian approach is proposed and some results will be discussed.

Programme And Abstracts For Tuesday 28th Of November

Keynote: Tuesday 28th 9:00 Mantra

Cluster Capture-Recapture: A New Framework For Estimating Population Size

Rachel Fewster University of Auckland

Ask any wildlife manager: their first burning question is "How many are there?", and their second is "Are they trending upwards or downwards?" Capture-recapture is one of the most popular methods for estimating population size and trends. As the name suggests, it relies on being able to identify the same animal upon multiple capture occasions. The pattern of captures and recaptures among identified animals is used to estimate the number of animals never captured.

Physically capturing and tagging animals can be a dangerous and stressful experience for both the animals and their human investigators - or if it transpires that the animals actually enjoy it, biased inference may result. Consequently, researchers increasingly favour non-invasive sampling using natural tags that allow animals to be identified by features such as coat markings, dropped DNA samples, acoustic profiles, or spatial locations. These innovations greatly broaden the scope of capture-recapture estimation and the number of capture samples achievable. However, they are imperfect measures of identity, effectively sacrificing sample quality for quantity and accessibility. As a result, capture-recapture samples no longer generate capture histories in which the matching of repeated samples to a single identity is certain. Instead, they generate data that are informative—but not definitive—about animal identity.

I will describe a new framework for drawing inference from capture-recapture studies when there is uncertainty in animal identity. In the cluster capture-recapture framework, we assume that repeated samples from the same animal will be similar, but not necessarily identical, to each other. Overlap is also possible between clusters of samples generated by different animals. We treat the sample data as a clustered point process, and derive the necessary probabilistic properties of the process to estimate abundance and other parameters using a Palm likelihood approach.

Because it avoids any attempts at explicit sample-matching, the cluster capture-recapture method can be very fast, taking much the same time to analyse millions of sample-comparisons as it does to analyse hundreds. I will describe a preliminary framework for abundance estimation from acoustic monitoring. Cluster capture-recapture can also be used for behavioural studies, and I will show an example using camera-trap data from a partially-marked population of forest ship rats.

Tuesday 28th 10:30 Narrabeen

Propensity Score Approaches In The Presence Of Missing Data: Comparison Of Balance And Treatment Effect Estimates

Jannah Baker[1], Tim Watkins[1,2], and Laurent Billot[1,3] [1]The George Insitute for Global Health [2]University of Sydney [3]University of New South Wales

The use of propensity score methods can potentially improve balance between groups in observational study data, thus minimising confounding. However, a frequent problem with such studies is the presence of missing data. We compare three approaches to generating the propensity score accounting for missing data within the context of a clinical case study examining the effect of telemonitoring on hypertension. Overall, 4,642 patients diagnosed with hypertension receiving online health support in My Health Guardian – a telephone chronic disease support program offered by private health insurer HCF – were offered a telemonitoring intervention. Of these, 2,729 accepted and started treatment between July 2014 to April 2015 (designated "cases"), and 1,913 declined (designated "controls"). Data were available from cases and controls on several baseline variables including demographic, lifestyle and clinical characteristics. Outcomes were the number of hospitalisations, total length of stay and total cost of hospitalisation between 1 January to 31 December 2016. Propensity score methods were used to balance baseline variables between groups. Three approaches were used to generate propensity scores from a logistic regression model accounting for missing data: 1) categorisation of all variables with "Missing" as a category, 2) multiple imputation of treatment effect (MIte) where treatment effect estimates are combined over 20 imputed datasets, and 3) multiple imputation of propensity score (MIps) where propensity scores are first averaged over imputed datasets prior to estimation of treatment effects. The propensity score from each approach was then used in two ways: a) matching cases to controls, and b) inverse probability of treatment weighting (IPTW). The balance achieved by each approach was compared using standardised differences of means and proportions in baseline characteristics between groups. The treatment effect estimates from each approach were also compared. The discussion will canvas our findings and recommendations for handling missing data when using propensity score approaches.

Tuesday 28th 10:30 Gunnamatta

Visualising Model Selection Stability In High-Dimensional Regression Models

Garth Tarr University of Sydney

The mplot R package provides a provides an implementation of model stability and variable inclusion plots for researchers to use to better inform the variable selection process. The initial focus was on exhaustive searches through the model space, however, this quickly becomes infeasible for high dimensional models. An alternative approach for high dimensional models is to combine bootstrap model selection with regularisation procedures. There exist a number of fast and efficient method regularisation methods for variable selection in high dimensional regression settings. We have implemented variable inclusion plots and model stability plots using the glmnet package. We demonstrate the utility of the mplot package in identifying stable regularised model selection choices with respect to two main sources of uncertainty. Firstly, by resampling the data we are able to determine how often various models are chosen when the data changes. Secondly, we are able to evaluate how often competing models are chosen across a range of values for the tuning parameter. Exploring these two sources of uncertainty in model selection generates a large amount of raw data that needs to be processed. The mplot package provides a variety of methods to visualise this raw data to help inform a researcher's model selection choice.

Tuesday 28th 10:50 Narrabeen

The Missing Link: An Equivalence Result For Likelihood Based Methods In Missing Data Problems

Firouzeh Noghrehchi, Jakub Stoklosa, Spiridon Penev, and David I. Warton University of New South Wales

Multiple imputation and maximum likelihood estimation (via the expectation-maximization algorithm) are two well-known methods readily used for analysing data with missing values. These two methods are often considered as being distinct from one another, due to their construction for estimation and their theoretical properties. We show that there is a close relationship between the two methods. Specifically, we show that a type of multiple imputation can be understood as a stochastic expectation-maximisation approximation to maximum likelihood. As a result, we can explore the application of a range of likelihood-based tools in the multiple imputation context in order to improve its performance. In particular, we develop information criteria for selecting an imputation model given a set of competing models, and a flexible likelihood ratio test when models are fitted by multiple imputation. We demonstrate our methods on real and simulated datasets.

Tuesday 28th 10:50 Gunnamatta

Dimensionality Reduction Of LIBS Data For Bayesian Analysis

Anjali Gupta[1], James Curran[1], Sally Coulson[2], and Christopher Triggs[1] [1]University of Auckland [2]ESR

In 2004, Aitken and Lucy published an article detailing a two-level likelihood ratio for multivariate trace evidence. This model has been adopted in a number of forensic disciplines such as the interpretation of glass, drugs (MDMA), and ink. Modern instrumentation is capable of measuring many elements in very low quantities and, not surprisingly, forensic scientists wish to exploit the potential of this extra information to increase the weight of this evidence. The issue, from a statistical point of view, is that the increase in the number of variables (dimension) in the problem leads to increased data demand to understand both the variability within a source, and in between sources. Such information will come in time, but usually we don't have enough. One solution to this problem is to attempt to reduce the dimensionality through methods such as principal component analysis. This practice is quite common in high dimensional machine learning problems. In this talk, I will describe a study where we attempt to quantify the effects of this this approach on the resulting likelihood ratios using data obtained from a Laser Induced Breakdown Spectroscopy (LIBS) instrument.

Tuesday 28th 11:10 Narrabeen

Analysis Of Melanoma Data With A Mixture Of Survival Models Utilising Multi-Class DLDA To Inform Mixture Class

Sarah Romanes, John Ormerod, and Jean Yang University of Sydney

Melanoma is a prevalent skin cancer in Australia, with close to 14000 new cases estimated to be diagnosed in 2017. Survival times are markedly different from one individual to the next. In particular, there appears to be three classes of survival outcome. This talk considers integrating survival time data with microarray gene expression data. We construct a hybrid model that seamlessly integrates a three-class linear discriminant analysis model, mixture of parametric survival models, and model selection components. We fit this model using a variational expectation maximization (VEM) approach. Our model selection component naturally simplifies as a function of likelihood ratio statistics allowing natural comparisons with traditional hypothesis testing methods. We compare our method with several naïve approaches which only addresses the classification aspect or survival model aspect in isolation.

Tuesday 28th 11:10 Gunnamatta

Forecasting Hotspots Of Potentially Preventable Hospitalisations With Spatially Aggregated Longitudinal Health Data: All Subset Model Selection With A Novel Implementation Of Repeated K-Fold Cross-Validation

Matthew Tuson, Berwin Turlach, Kevin Murray, Mei Ruu Kok, Alistair Vickery, and David Whyatt University of Western Australia

It is sometimes difficult to target individuals for health intervention due to limited information on their behaviour and risk factors. In such cases place-based interventions targeting geographical 'hotspots' with higher than average rates of health service utilisation may be effective. Many studies exist examining predictors of hotspots, but often do not consider that place-based interventions are typically costly and take time to develop and implement, and hotspots often regress to the mean in the short-term. Long-term geographical forecasting of hotspots using validated statistical models is essential in effectively prioritising place-based health interventions.

Existing methods forecasting hotspots tend to prioritise positive predicted value (i.e. correct predictions) at the expense of sensitivity. This work introduces methods to develop models optimising both positive predicted value and sensitivity concurrently. These methods utilise spatially aggregated administrative health data, WA census population data, and ABS geographic boundaries, combining all subset model selection with a novel implementation of repeated cross-validation for longitudinal data. Results from models forecasting 3-year hotspots for four potentially preventable hospitalisations are presented, namely: type II diabetes mellitus, heart failure, high risk foot, and chronic obstructive pulmonary disease (COPD).

Tuesday 28th 11:30 Narrabeen

Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process

Mugdha Manda, Thomas Lumley, and Susan Wells University of Auckland

Estimating disease trajectories has increasingly become more essential to clinical practitioners to administer effective treatment to their patients. A part of describing disease trajectories involves taking patients' medical histories and sociodemographic factors into account and grouping them into similar groups, or clusters. Advances in computerised patient databases have paved a way for identifying such trajectories in patients by recording a patient's medical history over a long period of time (longitudinal data): we studied data from the PREDICT-CVD dataset, a national primary-care cohort from which people with diabetes from 2002-2015 were identified through routine clinical practice. We fitted a Bayesian hierarchical linear model with latent clusters to the repeated measurements of HbA1c and eGFR, using the Markov birth-death process proposed by Stephens (2000) to handle the changes in dimensionality as clusters were added or removed.

Tuesday 28th 11:30 Gunnamatta

Challenges Analysing Combined Agricultural Field Trials With Partially Overlapping Treatments

Kerry Bell and Michael Mumford Queensland Department of Agriculture and Fisheries

To make recommendations on which management practices have the potential to increase crop yield there needs to be a consistent pattern demonstrated across trials from many environments. This presentation considers a case study looking at 31 mungbean trials in northern Australian from 2014 to 2016. The trials

did not always have consistent factors (e.g. variety, row spacing or target plant density) or even consistent factor levels. To overcome the issue of inconsistent factors, environments were defined as the combination of site, year and any management factors not common across trials (e.g. time of sowing, irrigation, fertiliser).

There were numerous full factorial combinations within subsets of the data that could be considered for investigation so the first challenge was to determine which factorial combinations to focus on to best address the research questions and reporting requirements. Once this was determined, all the data from the trials that contributed to the factorial were included in a combined analysis using linear mixed models. In this model, the factorial of interest was partitioned in the test of fixed effects while each trials' design parameters and residual variances were estimated using all the data from each trial. An example of the above mentioned factorial combinations is environment by row spacing for one particular variety. The next challenge was that with so many environments there was usually an environment by row spacing interaction which was not useful for making recommendations about row spacing.

Clustering of environments allowed forming groups that did not have a significant interaction between row spacing and environment. These groups were then generalised to types of environments with certain responses to row spacing.

Tuesday 28th 11:50 Narrabeen

A Hidden Markov Model For Sleep Stage Detection Using Raw Tri-Axial Wrist Actigraphy

Michelle Trevenen[1], Kevin Murray[1], Berwin Turlach[1], Leon Straker[2], and Peter Eastwood[1] [1]University of Western Australia [2]Curtin University

Sleep is a complex yet organised process consisting of regular cycles of sleep stages. These stages are rapid-eye movement and non-rapid eye movement (light sleep and slow-wave sleep). Sleep staging is of great importance in the physiological world as sleep disorders occur in around 20% of the population and are associated with a multitude of serious health implications and considerable economic burden. Hidden Markov models have been successfully used in the classification of individual sleep stages measured by polysomnography, which is considered the 'gold standard' in assessing sleep, however, it is intrusive and costly.

Actigraphy is increasingly being considered as a non-intrusive and cost effective alternative method to objectively measure sleep patterns. However, there is limited research on the ability of actigraphy to detect individual sleep stages, furthermore, this research indicates that these current methodologies do not have the ability to do so. Current actigraphic approaches to sleep detection use filtered uni-axial data measured with a low sampling rate, whereas, raw tri-axial data measured at high sampling rates is frequently used in the assessment of day-time activities.

Using simultaneously measured actigraphy and polysomnography data from 100 healthy young adults in the Western Australian Pregnancy Cohort Study we created and validated an algorithm to determine sleep stages utilising raw, tri-axial acceleration data from wrist actigraphy. Ten feature variables were created from each 30-second block of data and 50 subjects were used to train the hidden Markov model with the feature variables used as input parameters. The remaining 50 subjects were used to validate the trained hidden Markov model against polysomnography.

Validation suggested that our model is able to classify sleep stages using raw tri-axial actigraphy data. These results demonstrate that actigraphy-based hidden Markov models can feasibly be used for automatic sleep staging.

Tuesday 28th 11:50 Gunnamatta

Sparse Phenotyping Designs For Early Stage Selection Experiments In Plant Breeding Programs

Nicole Cocks, Alison Smith, David Butler, and Brian Cullis University of Wollongong

The early stages of cereal and pulse breeding programs typically involve in excess of 500 test lines. The test lines are promoted through a series of trials based on their performance (yield) and other desirable traits such as heat/drought tolerance, disease resistance, etc. It is therefore important to ensure the design (and analysis) of these trials are efficient in order to appropriately and accurately guide the breeders through their selection decisions, until only a small number of elite lines remain.

The design of early stage variety trials in Australia provided the motivation for developing a new design strategy. The preliminary stages of these programs have limited seed supply, which limits the number of trials and replicates of test lines that can be sown. Traditionally, completely balanced block designs or grid plot designs were sown at a small number of environments in order to select the highest performing lines for promotion to the later stages of the program. Given our understanding of variety (i.e. line) by environment interaction, this approach is not a sensible or optimal use of the limited resources available.

A new method to allow for a larger number of environments to be sampled for situations where seed supply is limited and number of test lines is large will be discussed. This strategy will be referred to as sparse phenotyping, which is developed within the linear mixed model framework as a model-based design approach to generating optimal trial designs for early stage selection experiments.

Tuesday 28th 12:10 Narrabeen

Comparisons Of Two Large Long-Term Studies In Alzheimer's Disease

Charley Budgeon University of Western Australia

The incidence of Alzheimer's disease (AD), the leading cause of dementia, is predicted to increase at least three fold by 2050. Curing this disease is a global priority. Currently, two major studies are attempting to gain further understanding of this disease; the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging, Biomarker and Lifestyle Study (AIBL). We describe these two cohorts to assess the impact of combining them to provide a larger cohort for analyses.

An initial comparison of the protocols was carried out and recruitment strategies were shown to be marginally different between the studies. Inclusion criteria specified ages between 55 and 90 years in ADNI and > 65 years in AIBL. Marginally different specifications for disease stage classifications of healthy controls (HC), mild cognitively impaired (MCI) and AD individuals were observed, for example, different Mini-Mental State Exam (MMSE) cut-offs. However, both studies had AD diagnosis supported by the NINDS/ARDA criteria. Baseline characteristics were compared between ADNI and AIBL cohorts. Overall, AIBL had more HCs compared to ADNI (69% vs 30%), but fewer MCI individuals (12% vs 50%). The ADNI cohort had a higher level of education and generally, within a disease classification, there were minimal differences in baseline age, sex, MMSE, and Preclinical Alzheimer Cognitive Composite (PACC) scores.

Longitudinal analyses compared the change over time for the two cohorts and disease classifications for PACC and MMSE. There were no significant differences in cohorts within the HC and MCI groups, but within the AD group, subjects in the ADNI cohort had generally higher predicted PACC and MMSE scores over time than those in AIBL.

Our results suggest there is the potential to combine the ADNI and AIBL cohorts for analysis purposes to provide one more powerful data set; however, consideration should be taken for some measures.

Tuesday 28th 12:10 Gunnamatta

A One-Stage Mixed Model Analysis Of Canola Chemistry Trials

Daniel Tolhurst, Ky Mathews, Alison Smith, and Brian Cullis University of Wollongong

The National Variety Trials (NVT) program is used by plant breeding companies to evaluate the yield potential of new crop varieties independently across a large range of Australian growing conditions. By comparison with the remaining NVT crops, grower decisions are further complicated in canola because of its vulnerability to the infestation of weeds. A measure historically used by farmers for the management of weeds is the application of a herbicide (chemistry) treatment. The choice of chemistry is important as it restricts variety selection to those bred with the specific tolerance. The set of varieties currently evaluated in NVT are tolerant to one of three chemistries, namely imidazolinone (I; but marketed as Clearfield), glyphosate (Roundup Ready; R) or triazine (T), or have no specific tolerance (i.e. conventional canola; C). Consequently, canola has a more complex testing regime than the remaining NVT crops as each trial has a nested treatment structure involving both chemistries and varieties.

Canola trials are conducted in locations across the Australian grain belt and reflect best farmer practice for each district. Every site is partitioned into several field blocks and plots are allocated to the treatments according to orthogonal block designs. A spray boom is used to administer each chemistry but is pragmatic in the sense that large areas are treated simultaneously. This precludes the application of different sprays to plots in the same block. Randomisation is therefore restricted so varieties in a single block are tolerant to the same chemistry. However, as the number of chemistries and blocks are exactly equal, there is no information to estimate the experimental error variation and both are statistically confounded. Consequently, growers are limited to evaluating varieties with the same tolerance as comparisons across chemistries are invalid. This also has important implications on the statistical analysis, which is discussed in this talk.

Tuesday 28th 12:30 Narrabeen

A Semi-Parametric Linear Mixed Models For Longitudinally Measured Fasting Blood Sugar Level Of Adult Diabetic Patients

Tafere Tilahun[1], Belay Birlie[1], and Legesse Kassa Debusho[2] [1]Jimma University [2]University of South Africa

This paper focused on longitudinal data analysis of fasting blood sugar (FBS) level of adult diabetic patients at Jimma University Specialized Hospital diabetic clinic using an application of semi-parametric mixed model. The study revealed that the rate of change in FBS level in diabetic patients, due to the clinic interventions, does not continue as a steady pace but changes with time and weight of patients. Furthermore, it clarified associations between FBS level and some characteristics of adult diabetic patients that weight of a diabetes patient has a significant negative effect whereas patient gender, age, type of diabetes and family history of diabetes did not have a significant effect on the change of FBS level. Under various variance structures of subject-specific random effects, the semi-parametric mixed models had better fit than linear mixed model. This was likely due to the localized splines, which captured more variability in FBS level than the linear mixed model.

Tuesday 28th 12:30 Gunnamatta

Individual And Joint Analyses Of Sugarcane Experiments To Select Test Lines

Alessandra Dos Santos[1], Chris Brien[2,4], Clarice G. B. Demétrio[1], Renata Alcarde Sermarini[1,5], Guilherme A. P. Silva[3], and Sandro R. Fuzatto[3] [1]University of São Paulo [2]University of South Australia [3]CTC - Piracicaba [4]Universtiy of Adelaide [5]University of Adelaide

In the early stages of a breeding program many field trials are conducted, considering several soil and weather conditions. In the case of sugarcane, these experiments are installed using a large number of test lines, but limitations in the field and of the amount of genetic material do not allow the replication of many. This work evaluated 21 trials from different regions, from a Brazilian sugarcane breeding program. Each of these experiments occupied a rectangular array of around 20 rows by 25 columns in most instance, the plots were 12 m longer, double-furrows with 0.9m between furrows within the plot and 1.5m spacing between different plots and 1m between columns. All the trials had at least 79% of the area planted with unreplicated test lines, the at most 21% of the plots were occupied by four commercial varieties, check. A special check, interspersed along diagonals, was planted systematically on a diagonal grid, the other three were equally replicated and each replicate was spread out in three neighbouring row plots. Seven of the 21 experiments had no significant direct genetic effects, 11 presented significant competition at the residual level and only one had significant competition at the genetic level. The correlation between the selected test lines for the different experiments in a same region was less than 0.54. However, the genetic correlation was significant in the joint analyses and stronger than that from the individual analyses. Two simulation studies were performed: the first investigated the analysis for a single experiment and the results show that it is difficult to fit a model when there is genetic competition with or without residual competition. The same difficulty was observed in the second study, which compared the results from individual and joint analyses. It showed that, even for the joint analyses, only around 45 to 55% of the true best test lines were selected.

Programme And Abstracts For Wednesday 29th Of November

Keynote: Wednesday 29th 9:00 Mantra

Statistics On Street Corners

Dianne Cook Monash University

Perceptual research is often conducted on the street, with convenience sampling of pedestrians who happen to be passing by. It is through experiments conducted using passer-bys that we have learned about the effect of change-blindness (https://www.youtube.com/watch?v=FWSxSQsspiQ) is in play outside the laboratory.

In data science, plots of data become important tools for observing patterns, making decisions, and communicating findings. But plots of data can be viewed differently by different observers, and often provoke skepticism about whether what you see "is really there?" With the availability of technology that harnesses statistical randomisation techniques and input from crowds we can provide objective evaluation of structure read from plots of data.

This talk describes an inferential framework for data visualisation, and the protocols that can be used to provide estimates of p-values, and power. I will discuss the experiments that we have conducted that (1) show that the crowd-sourcing does provide results similar to statistical hypothesis testing, (2) how this can be used to improve plot design, (3) p-values in situations where no classical tests exist. Examples from ecology and agriculture will be shown.

Joint work with Heike Hofmann, Andreas Buja, Deborah Swayne, Hadley Wickham, Eun-kyung Lee, Mahbubul Majumder, Niladri Roy Chowdhury, Lendie Follett, Susan Vanderplas, Adam Loy, Yifan Zhao, Nathaniel Tomasetti

Wednesday 29th 10:30 Narrabeen

Estimating Overdispersion In Sparse Multinomial Data

Farzana Afroz University of Otago

The phenomenon of overdispersion arises when the data are more variable than we expect from the fitted model. This issue often arises when fitting a Poisson or a binomial model. When overdispersion is present, ignoring it may lead to misleading conclusions, with standard errors being underestimated and overly-complex models being selected. In our research we considered overdispersed multinomial data, which can arise in many research areas. Two approaches can be used to analyse overdispersed multinomial data; the use of the quasilikelihood method or explicit modelling of the overdispersion using, for example, a Dirichlet-multinomial or finite-mixture distribution. Use of quasilikelihood has the advantage of only requiring specification of the

first two moments of the response variable. For sparse data, such as in a contingency table with many low expected counts, use of quasilikelihood to estimate the amount of overdispersion will be particularly useful, as it may be difficult to obtain reliable estimates of the parameters in a Dirichlet-multinomial or finite-mixture model. I consider four estimators of the amount of overdispersion in sparse multinomial data, discuss their theoretical properties and provide simulation results showing their performance in terms of bias, variance and mean squared error.

Wednesday 29th 10:30 Gunnamatta

Assessing Mud Crab Meat Fullness Using Non-Invasive Technologies

Carole Wright, Steve Grauf, Brett Wedding, Paul Exley, John Mayze, and Sue Poole Queensland Department of Agriculture and Fisheries

The decision of whether a mud crab should be retained at harvest has traditionally been based on shell hardness. This is most commonly assessed by using thumb pressure applied to the carapace of the mud crab. The carapace of a recently moulted mud crab will flex considerably and is therefore returned to the water. This assessment has also been used to divide mud crabs into three meat fullness grades (A, B and C). The higher meat fullness grade A mud crabs fetch a greater price at market compared to the lower B and C grades. The subjective nature of this assessment will always result in disputes at the boundaries of the grades. By developing a more objective science-based method downgrades at the market will be reduced while consumer satisfaction and the overall industry profitability will increase.

A scoping study was conducted that evaluated innovative non-invasive technologies to assess mud crab meat fullness based on percentage yield recovery of cooked meat from the dominant individual mud crab claws. The non-invasive technologies assessed included near infrared spectroscopy (NIRS), candling using visible light, and acoustic velocity. NIRS showed the most potential and was reassessed in a second study with slight improvements to spectra capture methods and NIR light sources.

94 live mud crabs from the Moreton Bay area were used in the second study. Partial least squares regression (PLS-R) was performed to build a calibration model to predict the percentage yield recovery of cooked meat based on the spectral data. The PLS-R had $R^2 = 0.77$ and $RMSECV = 4.8$.

A principal components linear discriminant analysis (PC-LDA) was also conducted to discriminate between the standard three grades of mud crab meat fullness. This was compared to the industry standard shell hardness method. The NIRS PC-LDA achieved a minimum of 76% correct classification for each of the three grades, compared to 24% for the shell hardness method.

The non-invasive technologies trialled along with the results will be discussed in this talk.

Wednesday 29th 10:30 Bundeena

Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition

Geoff Jones Massey University

Metabolic syndrome (MetS) is a major multifactorial condition that predisposes adults to type 2 diabetes and CVD. It is defined as having at least three of five cardiometabolic risk components: 1) high fasting triglyceride level, 2) low high-density lipoprotein cholesterol, 3) elevated fasting plasma glucose, 4) large waist circumference (abdominal obesity), and 5) hypertension. In the US Study of Women's Health Across the Nation (SWAN), a 15-year multi-centre prospective cohort study of women from five racial/ethnic groups, the incidence of MetS increased as midlife women underwent the menopausal transition (MT). A model is sought

to examine the interdependent progression of the five MetS components and the influence of demographic covariates.

Wednesday 29th 10:50 Narrabeen

Statistical Analysis Of Coastal And Oceanographic Influences On The Queensland Scallop Fishery

Wen-Hsi Yang[1], Anthony J. Courtney[2], Michael F. O'neill[2], Matthew J. Campbell[2], George M. Leigh[2], and Jerzy A. Filar[1] [1]Universtiy of Queensland [2]Queensland Department of Agriculture and Fisheries

The saucer scallop (Ylistrum balloti) otter-trawl fishery used to be the most valuable commercially-fished species in Queensland ocean waters. Over the last few years, there has been growing concern among fishers, fishery managers and scientists over the decline in catch rates and annual harvest. A quantitative assessment conducted in 2016 showed that scallop abundance was at an historic low level. The assessment used data sourced from the fishery and independent surveys. Further information on coastal and oceanographic influences are available and may reveal new factors that influence population abundance of scallops and improve management of the fishery. In this study, scallop catch rate abundance data and coastal and physical oceanographic variables (e.g. sea surface temperature anomalies, coastal freshwater flow and Chlorophyll-a) were modelled to identify spatial and temporal environmental processes important for consideration in fishery management procedures.

Wednesday 29th 10:50 Gunnamatta

Saved By The Experimental Design: Testing Bycatch Reduction And Turtle Exclusion Devices In The Png Prawn Trawl Fishery

Emma Lawrence and Bill Venables CSIRO

In trawling for prawns, the prawn catch is often only a small part of the results of any one trawl, with the remainder called "bycatch". Reducing the bycatch component, while maintaining the prawn catch, is an important industry goal, primarily for environmental accreditation purposes, but also for economic reasons.

We designed an at-sea trial for the Gulf of Papua Prawn Fishery, involving four vessels each towing "quad gear" (that is, 4 separate, but linked trawl nets) in each trawl shot, over 18 days. The experiment was designed to assess the effectiveness of 27 combinations of Turtle Excluder Devices (TEDs) and Bycatch Reduction Devices, (BRDs), with a control net, without any attached device as one of the nets in each quad. At Biometrics 2015 we discussed how we used simulated annealing to generate a highly efficient design, in several stages, to meet the large number of highly specific logistical constraints.

The focus of this talk will be the analysis, which also proved somewhat challenging. We will present the results of our analysis and demonstrate why putting the time into thinking about and generating a non-standard experimental design allowed us to accommodate the various glitches and misfortunes that always seem to happen at sea.

Wednesday 29th 10:50 Bundeena

Rethinking Biosecurity Inspections: A Case Study Of The Asian Gypsy Moth (AGM) In Australia

Petra Kuhnert[1], Dean Paini[1], Paul Mwebaze[1], and John Nielsen[2] [1]CSIRO [2]Department of Agriculture and Water Resources

The Asian gypsy moth (AGM) (Lymantria dispar asiatica) is a serious biosecurity risk to Australia's forestry and horticultural industries. While similar in appearance to the European gypsy moth (Lymantria dispar dispar), the Asian gypsy moth is capable of flying up to 40 kilometres and therefore has the potential to establish and spread in other areas like Australia. In addition, females are attracted to light and will oviposit (lay eggs) indiscriminately. As a result, females are attracted to shipping ports at night and will oviposit on ships. These ships therefore have the potential to spread this moth around the world.

The life-cycle of the moth has been well documented and is heavily dependent on temperature, with eggs undergoing three phases of diapause before hatching. Current inspections of vessels arriving into Australian ports from what is deemed an "at risk" port is a lengthy and costly process.

To assist the Department of Agriculture with their prioritisation of ships, we developed an AGM Tool in the form of an R Shiny App that (1) shows the shortest maritime path from an at risk port to an Australian port for a vessel of interest and (2) predicts the probability of a potential hatch and it's reliability using a classification tree model that was developed to emulate the lifecycle of the moth from simulated data. In this talk we will discuss the methodology that (1) simulates the AGM biology and potential hatches of eggs, along with how we extracted relevant temperature data that was the primary driver of the lifecycle for AGM, and (2) emulates this simulated data using a statistical model, namely a classification tree to predict the probability of a potential hatch. We will also discuss a bootstrap approach to explore the reliability of the potential hatch predicted.

Wednesday 29th 11:10 Narrabeen

Subtractive Stability Measures For Improved Variable Selection

Connor Smith[1], Samuel Müller[1], and Boris Guennewig[1,2] [1]University of Sydney [2]University of New South Wales

This talk builds upon the Invisible Fence (Jiang et al., 2011) a promising model selection method. Utilizing a combination of coeffcient, scale and deviance estimates we are able to improve this resampling based model selection method for regression models, both linear and general linear models. The introduction of a variable inclusion plot allows for a visual representation for the stability of the model selection method as well as the variables bootstrapped rank. The suggested methods will be applied to both simulated and real examples with comparisons about both computational time and effectiveness made to selections through alternative selection procedures. We will report on our latest results from ongoing work in scaling up subtractive stability measures when the numbers of features is large.

References: Jiang, J., Nguyen, T., & Rao, J. S. (2011). Invisible fence methods and the identification of differentially expressed gene sets. Statistics and Its Interface, 4(3), 403-415.

Wednesday 29th 11:10 Bundeena

A Comparison Of Multiple Imputation Methods For Missing Data In Longitudinal Studies

Md Hamidul Huque[1], Katherine Lee[1], Julie Simpson[2], and John Carlin[1] [1]Murdoch Childrens Research Institute [2]University of Melbourne

Multiple imputation (MI) for imputing missing data are increasingly used in longitudinal studies where data are missing due to non-response and lost to follow-up. Standard multivariate normal imputation (MVNI) and fully conditional specifications (FCS) are the principle imputation framework available for imputing cross-sectional missing data. A number of methods has been suggested in the literature to impute longitudinal data including (i) use of standard FCS and MVNI with repeated measurements as separate distinct variables (ii) use of imputation methods based on generalized linear mixed models. No clear evaluation of the relative performance of available MI methods in the context of longitudinal data. We present a comprehensive comparison of the all the available methods for imputation longitudinal data in the context of estimating coefficient for both linear regression model and linear mixed effect model. We also compared the performance of the methods to impute both binary and continuous data. A total of 10 different methods (MVNI, JM-pan, JM-jomo, standard FCS, FCS-twofold, FCS-MTW, FCS-2lnorm, FCS-2lglm, FCS-2ljomo and FCS-Blimp) are compared in terms of bias, standard error and coverage probability of the estimated regression coefficients. These methods are compared using a simulation study based on a previously conducted analysis exploring the association between the burden of overweight and quality of life (QoL) using data from the Longitudinal Study of Australian Children (LSAC). We found that both standard FCS and MVNI provide reliable estimates and coverage of the regression parameters. Among other methods linear mixed models based methods, JM-jomo and FCS-Blimp approaches hold great promise.

Wednesday 29th 11:30 Narrabeen

Species Distribution Modelling For Combined Data Sources

Ian Renner[1] and Olivier Gimenez[2] [1]University of Newcastle [2]Centre d'Ecologie Fonctionnelle et Evolutive

Increasingly, multiple sources of species occurrence data are available for a particular species, collected through different protocols. For single-source models, a variety of methods have been developed: point process models for presence-only data, logistic regression for presence-absence data obtained through single-visit systematic surveys, and occupancy modelling for detection/non-detection data obtained through repeat-visit surveys. In situations for which multiple sources of data are available to model a species, these sources may be combined via a joint likelihood expression. Nonetheless, there are questions about how to interpret the output from such a combined model and how to diagnose potential violations of model assumptions such as the assumption of spatial independence among points.

In this presentation, I will explore questions of interpretation of the output from these combined approaches, as well as propose extensions to current practice through the introduction of a LASSO penalty, source weights to account for differing quality of data, and models which account for spatial dependence among points. This approach will be demonstrated by modelling the distribution of the Eurasian lynx in eastern France.

Wednesday 29th 11:30 Gunnamatta

A Factor Analytic Mixed Model Approach For The Analysis Of Genotype By Treatment By Environment Data

Lauren Borg, Brian Cullis, and Alison Smith University of Wollongong

The accurate evaluation of genotype performance for a range of traits, including disease resistance, is of great importance to the productivity and sustainability of major Australian commercial crops. Typically, the data generated from crop evaluation programmes arise from a series of field trials known as multi-environment trials (METs), which investigate genotype performance over a range of environments.

In evaluation trials for disease resistance, it is not uncommon for some genotypes to be chemically treated against the afflicting disease. An important example in Australia is the assessment of genotypes for resistance

to blackleg disease in canola crops where it is common practice to treat canola seeds with a fungicide. Genotypes are either grown in trials as treated, untreated or as both.

There are a number of methods for the analysis of MET data. These methods, however, do not specifically address the analysis of data with an underlying three-way structure of genotype by treatment by environment (GxTxE). Here, we propose an extension of the factor analytic mixed model approach for MET data, using the canola blackleg data as the motivating example.

Historically in the analysis of blackleg data, the factorial genotype by treatment structure of the data was not accounted for. Entries, which are the combinations of genotypes and fungicide treatments present in trials, were regarded as 'genotypes' and a two-way analysis of 'genotypes' by environments was conducted.

The analysis of our example showed that the accuracy of genotype predictions, and thence information for growers, was substantially improved with the use of the three-way GxTxE approach compared with the historical approach.

Wednesday 29th 11:30 Bundeena

The Impact Of Cohort Substance Use Upon Likelihood Of Transitioning Through Stages Of Alcohol And Cannabis Use And Use Disorder: Findings From The Australian National Survey On Mental Health And Well-Being

Louisa Degenhardt[1], Meyer Glantz[2], Chrianna Bharat[1], Amy Peacock[1], Luise Lago[1], Nancy Sampson[3], and Ronald Kessler[3] [1]National Drug and Alcohol Research Centre [2]National Institute on Drug Abuse [3]Harvard University

The aims of the present study were to use population-level Australian data to estimate prevalence and speed of transitions across stages of alcohol and cannabis use, abuse and dependence, and remission from disorder, and consider the potential impacts that an individual's age and sex cohort's level of substance use predicted transitions into and out of substance use. Data on lifetime history of use, DSM-IV use disorders, and remission from these disorders were collected from participants (n=8,463) in the 2007 Australian National Survey of Mental Health and Wellbeing using the Composite International Diagnostic Interview.

Lifetime prevalence of alcohol use, regular use, abuse, dependence, and remission from abuse and dependence were 94.1%, 64.5%, 22.1%, 4.0%, 16.1% and 2.1%, respectively. Unconditional lifetime prevalence of cannabis use, abuse, dependence, and remission from abuse and dependence were 19.8%, 6.1%, 1.9%, 4.0% and 1.5%. Increases in the estimated proportion of people in the respondent's sex and age cohort who used alcohol/cannabis as of a given age were significantly associated with most transitions from use through to remission beginning at the same age. Clear associations were documented between cohort-level prevalence of substance use and personal risk of subsequent transitions of individuals in the cohort from use to greater substance involvement. This relationship remained significant over and above associations involving the individual's age of initiation. These findings have important implications for our understanding of the causal pathways into and out of problematic substance use.

Wednesday 29th 11:50 Narrabeen

The LASSO On Latent Indices For Ordinal Predictors In Regression

Francis Hui[1], Samuel Mueller[2], and Alan Welsh[1] [1]ANU [2]University of Sydney

Many applications of regression models involve ordinal categorical predictors. A motivating example we consider is ordinal ratings from individuals responding to questionnaires regarding their workplace in the Household Income and Labour Dynamics in Australia (HILDA) survey, with the aim being to study how workplace conditions (main and possible interaction effects) affect their overall mental wellbeing. A common approach to handling ordinal predictors is to treat each predictor as a factor variable. This can lead to a very high-dimensional problem, and has spurred much research into penalized likelihood methods for handling categorical predictors while respecting the marginality principle. On the other hand, given the ordinal ratings are often regarded as manifestations of some latent indices concerning different aspects of job quality, then a more sensible approach would be to first perform some sort of dimension reduction before entering the predicted indices into a regression model. In applied research this is often performed as a two-stage procedure, and in doing so fails to utilize the response in order to better predict the latent indices themselves.

In this talk, we propose the LASSO on Latent Indices (LoLI) for handling ordinal categorical predictors in regression. The LoLI model simultaneously constructs a continuous latent index for each or groups of ordinal predictors and models the response as a function of these (and other predictors if appropriate) including potential interactions, with a composite LASSO type penalty added to perform selection on main and interaction effects between the latent indices. As a single-stage approach, the LoLI model is able to borrow strength from the response to improve construction of the continuous latent indices, which in turn produces better estimation of the corresponding regression coefficients. Furthermore, because of the construction of latent indices, the dimensionality of the problem is substantially reduced before any variable selection is performed. For estimation, we propose first estimating the cutoffs relating the observed ordinal predictors to the latent indices. Then conditional on these cutoffs, we apply a penalized Expectation Maximization algorithm via importance sampling to estimate the regression coefficients. A simulation study demonstrates the improved power of the LoLI model at detecting truly important ordinal predictors compared to both two-stage approaches and using factor variables, and better predictive and estimation performance compared to the commonly used two-stage approach.

Wednesday 29th 11:50 Gunnamatta

Whole-Genome QTL Analysis For Nested Association Mapping Populations

Maria Valeria Paccapelo[1], Alison Kelly[1], Jack Christopher[2], and Arunas Verbyla[3,4] [1]Queensland Department of Agriculture and Fisheries [2]Queensland Alliance for Agriculture and Food [3]Data61 [4]CSIRO

Genetic dissection of quantitative traits in plants has become an important tool in breeding of improved varieties. The most commonly used methods to map QTL are linkage analysis in bi-parental populations and association mapping in diversity panels. However, bi-parental populations are restricted in terms of allelic diversity and recombination events. Despite the fact that association mapping overcomes these limitations, it has low power to detect rare alleles associated with a trait of interest. Multi-parent populations such as multi-parent advanced generation inter-cross (MAGIC) and nested association mapping (NAM) populations have been developed to combine strengths of both mapping approaches, capturing more recombination events and allelic diversity than bi-parental populations and in a greater frequency than a diversity panel. Nested association mapping uses multiple RIL families connected by a single common parent. Such a population structure presents some additional challenges compared to traditional mapping, in particular the population design and the large number of molecular markers that need to be integrated simultaneously into the analysis. We present a method for QTL mapping for NAM populations adapted from multi-parent whole genome average interval mapping (MPWGAIM) where the NAM design is incorporated through the probability of inheriting founder alleles for every marker across the genome. This method is based on a mixed linear model in a one-stage analysis of raw phenotypes together with markers. It simultaneously scans the whole-genome through an iterative process leading to a multi-locus model. The approach was applied to a wheat NAM population in order to perform QTL mapping for plant height. The method was developed in R, with main dependencies being the R packages MPWGAIM and asreml. This approach establishes the basis for further

studies and extensions such as the combination of multiple NAM populations.

Wednesday 29th 11:50 Bundeena

An Asymmetric Measure Of Population Differentiation Based On The Saddlepoint Approximation Method

Louise McMillan and Rachel Fewster University of Auckland

In the field of population genetics there are many measures of genetic diversity and population differentiation. The best known is Wright's Fst, later expanded by Cockerham and Weir, which is very widely used as a measure of separation between populations. More recently a multitude of other measures have been developed, from Gst to D, all with different features and disadvantages. One thing these measures all have in common is that they are symmetric, which is to say that the Fst between population A and population B is the same as that between population B and population A. Following my work on GenePlot, a visualization tool for genetic assignment, I am now working on the development of an asymmetric measure, where the fit of A into B may not be the same as the fit of B into A. This measure will enable the detection of scenarios such as "subsetting", the relationship between a large, diverse population A and a smaller population B that has experienced genetic drift since being separated from A. The measure has several features that distinguish it from existing measures, and is constructed using the same saddlepoint approximation method underlying GenePlot, and which is used to approximate the multi-locus genetic distributions of populations.

Wednesday 29th 12:10 Narrabeen

Fast And Approximate Exhaustive Variable Selection For GLMs With APES

Kevin Wang, Samuel Mueller, Garth Tarr, and Jean Yang University of Sydney

Obtaining maximum likelihood estimates for generalised linear models (GLMs) is computationally intensive and remains as the major obstacle for performing all subsets variable selection. Exhaustive exploration of the model space, even for a moderately large number of covariates, remains a formidable challenge for modern computing capabilities. On the other hand, efficient algorithms for exhaustive searches do exist for linear models, most notably the leaps and bound algorithm and, more recently, the mixed integer optimisation algorithm. In this talk, we present APES (APproximated Exhaustive Search) a new method that approximates all subset selection for a given GLM by reformulating the problem as a linear model. The method works by learning from observational weights in a correct/saturated generalised linear regression model. APES can be used in partnership with any other state-of-the-art linear model selection algorithm, thus enabling (approximate) exhaustive model exploration in dimensions much higher than previously feasible. We will demonstrate that APES model selection is competitive against genuine exhaustive search via simulation studies and applications to health data. Extensions to a robust setting is also possible.

Wednesday 29th 12:10 Gunnamatta

Order Selection Of Factor Analytic Models For Genotype X Environment Interaction

Emi Tanaka[1], Francis Hui[2], and David Warton[3] [1]University of Sydney [2]ANU [3]University of New South Wales

Factor analytic (FA) models are widely used across a range of disciplines owing to computational advantages from dimension reduction and possible ability to interpret the factors. In plant breeding, FA model provides a natural framework to model the genotype x environment interaction. An FA model is dictated by the number of factors (order of the model). A higher order lends to more parameters in the model and this necessitates the order selection to achieve parsimony. We introduce an order selection method via the ordered factor lasso (OFAL). We illustrate its performance based on a simulation on a real wheat yield multi-environmental trial.

Wednesday 29th 12:10 Bundeena

Multiple Sample Hypothesis Testing Of The Human Microbiome Through Evolutionary Trees

Martina Mincheva[1], Hongzhe Li[2], and Jun Chen[3] [1]Temple University [2]University of Pennsylvania [3]Mayo Clinic

Next generation sequencing technologies make it possible to survey microbial communities by sequencing nucleic acid material extracted from multiple samples. The metagenomic read counts are summarized as empirical distributions on a reference phylogenetic tree. The distance between them is evaluated by the Kantorovich-Rubinstein (kr) metric, equivalent to the commonly used weighted UniFrac distance on a tree. This paper proposes a method to test the hypothesis that two sets of samples have the same microbial composition. The asymptotic distributions of the kr distance between the two Frechet means and the Frechet variances are derived and are shown to be independent. The test statistic is defined as the ratio of those distances and it is shown to follow an asymptotic F-distribution. Its generality stems from the fact that the test is nonparametric and requires no assumptions on the probability distributions of the count data. It is also applicable for varying set sizes and sample sizes. It is an extension of Evans and Matsen (2012) who suggest a test to only compare two single samples. The computational efficiency of the proposed test comes from the exact asymptotic distribution of the proposed test statistic. Extensive data analysis shows that the test is significantly faster than the permutation-based multivariate analysis of variance using distance matrices (permanova) (McArdle and Anderson, 2001). At the same time, it has correct type 1 errors and comparable power, which makes it preferable in the analysis of large scale microbiome data.

Keynote: Wednesday 29th 13:40 Mantra

Statistical Strategies For The Analysis Of Large And Complex Data

Louise Ryan[1,2], Stephen Wright[1,3], and Hon Hwang[1] [1]University of Technology Sydney [2]Harvard T. H. Chan School of Public Health [3]Australian Red Cross

This talk will focus on challenges that arise when faced with the analysis of datasets that are too large for standard statistical methods to work properly. While one can always go for the expensive solution of getting access to a more powerful computer or cluster, it turns out that there are some simple statistical strategies that can be used. In particular, we'll discuss the use of so called "Divide and Recombine" strategies that relegate some of the work to be done in a distributed fashion, for example via Hadoop. Combining these strategies with clever subsampling and data coarsening ideas can result in datasets that are small enough to manage on a standard desktop machine, with only minimal efficiency loss. The ideas are illustrated with data from the Australian Red Cross.

Wednesday 29th 14:30 Narrabeen

Optimal Experimental Design For Functional Response Experiments

Jeff Zhang and Christopher Drovandi Queensland University of Technology

Functional response models are important in understanding predator-prey interactions. The development of functional response methodology has progressed from mechanistic models to more statistically motivated models that can account for variance and the over-dispersion commonly seen in the datasets collected from functional response experiments. However, little information seems to be available to those wishing to prepare optimal parameter estimation designs for functional response experiments. We develop a so-called exchange design optimisation algorithm suitable for integer-valued design spaces, which for the motivating functional response experiment involves selecting the number of prey used for each observation. Further, we develop and compare new utility functions for performing robust optimal design in the presence of parameter uncertainty, which are generally applicable. The methods are illustrated using a published beta-binomial functional response model for an experiment involving the freshwater predator Notonecta glauca (an aquatic insect) preying on Asellus aquaticus (a small crustacean) as a case study.

Wednesday 29th 14:30 Gunnamatta

To PCA Or Not To PCA

Catherine M. McKenzie[1], Wei Zhang[1], Stuart D. Card[1], Cory Matthew[2], Wade J. Mace[1], and Siva Ganesh[1]
[1]AgResearch [2]Massey University

When there are groupings of observations present in the data, many researchers resort to utilising a Principal Components Analysis (PCA) a priori for identifying patterns in the data, and then look to map the patterns obtained from PCA to differences among the groupings, or attribute biological signal to them. Is this appropriate, given that PCA's are not designed to discriminate between the groupings? Is a group-oriented multivariate methodology such as multivariate analyses of variance (MANOVA) or Canonical Discriminant Analysis (CDA) preferable? Which method has more relevance when investigating factor effects of biochemical pathways? We explore this question via a biological example.

Two biological materials (B1 & B2) were analysed for the same 19 primary metabolites, with three factors of Methods (M1 & M2), Treatment (T1, T2, T3 and T4), and Age (A1 & A2), with three replicate values giving a total of 48 observations for each biological material. Univariate and multivariate analyses of variance (ANOVA and MANOVA, respectively) were carried out, for which there were many statistically significant interaction effects. In addition, other multivariate techniques such as PCA and CDA were used to explore relationships between the variables. The question remains as to the appropriateness of carrying out PCA to explore biochemical pathways, the comparison being between tailoring the pattern extraction a priori to match the known groupings within the data versus starting with an unrestrained pattern analysis and seeking to explain the patterns detected post-analysis?

Wednesday 29th 14:50 Narrabeen

An Evaluation Of Error Variance Bias In Spatial Designs

Emlyn Williams[1] and Hans-Peter Piepho[2] [1]ANU [2]University of Hohenheim

Spatial design and analysis are widely used, particularly in field experimentation. However, it is often the case that spatial analysis does not enhance more traditional approaches such as row-column analysis. It is then of interest to gauge the degree of error variance bias that accrues when a spatially-designed experiment is analysed as a row-column design. This talk builds on the work of Tedin (1931) who, with R.A. Fisher as advisor, studied error variance bias in knight's move Latin squares.

Wednesday 29th 14:50 Gunnamatta

New Model-Based Ordination Data Exploration Tools For Microbiome Studies

Olivier Thas[1], Stijn Hawinkel[1], and Luc Bijnens[2] [1]Ghent University [2]Janssen Pharmaceutics

High-throughput sequencing technologies allow easy characterization of the human microbiome, but the statistical methods for analysing microbiome data are still in their infancy. Data exploration often relies on classical dimension reduction methods such as Principal Coordinate Analysis (PCoA), which is basically a Multidimensional Scaling (MDS) method starting from ecologically relevant distance measures between the vectors of relative abundances of the microorganisms (e.g Bray-Curtis distance).

We will demonstrate that these classical visualisation methods fail to deal with microbiome-specific issues such as variability due to library-size differences and overdispersion. Next we propose a new technique that is based on a negative binomial regression model with log-link, and which relies on the connection between correspondence analysis and the log-linear RC(M) models of Goodman (Annals of Statistics, vol. 13, 1985); see also Zhu et al. (Ecological Modelling, vol. 187, 2005). Instead of assuming a Poisson distribution for the counts, a negative binomial distribution is assumed. To better account for library size effects, we adopt a different weighting scheme, which naturally arises from the parameterisation of the model. An iterative parameter estimation method is proposed and implemented into R. The new method is illustrated on several example datasets, and it is empirically evaluated in a simulation study. It is concluded that our method succeeds better in discovering structure in microbiome datasets than with other conventional methods.

In the second part of the presentation we extent the model-based method to a constrained ordination method by using sample-specific covariate data. The method looks for a two-dimensional visualisation that optimally discriminates between species with respect to their sensitivity to environmental conditions. Again we build upon results of Zhu et al. (2005) and Zhang and Thas (Statistical Modelling, vol. 12, 2012). The method is illustrated on real data.

All methods are available as an R package.

Wednesday 29th 15:10 Narrabeen

Always Randomize?

Chris Brien[1,2] [1]University of South Australia [2]Universtiy of Adelaide

Fisher gave us three fundamental principles for designed experiments: replication, randomization and local control. Consonant with this, Brien et al. (2011) [Brien, C. J., Harch, B. D., Correll, R. L., & Bailey, R. A. (2011) Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. Journal of Agricultural, Biological, and Environmental Statistics, 16, 422-450.] exhort the use of randomization in multiphase experiments via their Principle 7 (Allocate and randomize in the laboratory). This principle is qualified with 'wherever possible', which leads to the question 'when is randomization not possible?'.

Situations where randomization is not applicable will be described for both single-phase and multiphase experiments. The reasons for not randomizing include practical limitations and, for multiphase experiments, difficulty in estimating variance parameters when randomization is employed. For the latter case, simulation studies canvassing a number of potential difficulties will be described. A Nonrandomization Principle, and an accompanying analysis strategy, for multiphase experiments will be proposed.

Wednesday 29th 15:10 Gunnamatta

Comparing Classical Criteria For Selecting Intra-Class Correlated Features For Three-Mode Three-Way Data

Lynette Hunt[1] and Kaye Basford[2] [1]University of Waikato [2]University of Queensland

Many unsupervised learning tasks involve data sets with both continuous and categorical attributes. One possible approach to clustering such data is to assume that the data to be clustered come from a finite mixture of populations. There has been extensive use of mixtures where the component distributions are multivariate normal and where the data would be described as two mode two way data. The finite mixture model can also be used to cluster three way data. The mixture model approach requires the specification of the number of components to be fitted to the model and the form of the density functions of the underlying components.

This talk illustrates the performance of several commonly used model selection criteria in selecting both the number of components and the form of the correlation structure amongst the attributes when fitting a mixture model to the finite mixture model to cluster three way data containing mixed categorical and continuous attributes

Wednesday 29th 15:50 Narrabeen

Exploring The Social Relationships Of Dairy Goats

Vanessa Cave[1], Benjamin Fernoit[2], Jim Webster[1], and Gosia Zobel[1] [1]AgResearch [2]Agrosup Dijon

Goats are sentient beings capable of an emotional response to their lives. Yet despite this, the social relationships between animals are largely overlooked in commercial systems. Goats have been shown to recognise, and make decisions based on, the presence of other specific goats. The degree to which they choose to associate with individuals has not been established, but social bonds in other animals have been shown to buffer against stressful situations, and conversely can be a significant source of stress when such bonds are disrupted.

To investigate whether social relationships exist among dairy goats, 4 non-consecutive days of video focusing on a group of 12 goats was analysed. At one minute scan intervals, the proximity of every goat relative to every other goat was recorded as an ordinal variable with four levels (in contact, within a head length, within a body length, or alone). Each goat's location in the pen was also noted (feeding, bedding, climbing platform).

A variety of statistical techniques, including heatmaps and network analyses, were used to study the social relationships among goats based on proximity. Social relationships were characterised by specific pairs of goats reliably spending a lot of time in close proximity.

Results indicate that whilst some goats were "sociable" (e.g., spending more than 60% of their time with other goats), others tended to be "loners" (e.g., spending more than 60% of their time alone). Interestingly, there was evidence of both preferred and avoided companionships.

This small-scale study provides the first evidence to suggest that common management practices resulting in the regrouping of dairy goats could have an impact on their welfare.

Wednesday 29th 15:50 Gunnamatta

Bayesian Semi-Parametric Spectral Density Estimation With Applications To The Southern Oscillation Index

Claudia Kirch[1], Matt Edwards[2], Alexander Meier[1], and Renate Meyer[2] [1]University of Magdeburg [2]University of Auckland

Standard time series modelling is dominated by parametric models like ARMA and GARCH models. Even though nonparametric Bayesian inference has been a rapidly growing area over the last decade, only very few nonparametric Bayesian approaches to time series analysis have been developed. Most notably, Carter and Kohn (1997), Gangopadhyay (1998), Choudhuri et al. (2004), and Rosen et al (2012) used Whittle's likelihood for Bayesian modeling of the spectral density as the main nonparametric characteristic of stationary time series. On the other hand, frequentist time series analyses are often based on nonparametric techniques encompassing a multitude of bootstrap methods (Kreiss and Lahiri, 2011, Kirch and Politis, 2011).

As shown in Contreras-Cristan et al. (2006), the loss of efficiency of the nonparametric approach using Whittle's likelihood approximation can be substantial. On the other hand, parametric methods are more powerful than nonparametric methods if the observed time series is close to the considered model class but fail if the model is misspecified. Therefore, we suggest a nonparametric correction of a parametric likelihood that takes advantage of the efficiency of parametric models while mitigating sensitivities through a nonparametric amendment. We use a nonparametric Bernstein polynomial prior on the spectral density with weights induced by a Dirichlet process. Contiguity and posterior consistency for Gaussian stationary time series have been shown in a preprint by Kirch et al (2017). Bayesian posterior computations are implemented via a MH-within-Gibbs sampler and the performance of the nonparametrically corrected likelihood is illustrated in a simulation. We use this approach to analyse the monthly time series of the Southern Oscillation Index, one of the key atmospheric indices for gauging the strength of El Nino events and their potential impacts on the Australian region.

Wednesday 29th 16:10 Narrabeen

Efficient Multivariate Sensitivity Analysis Of Agricultural Simulators

Daniel Gladish CSIRO

Complex mechanistic computer models often produce multivariate output. Sensitivity analysis can be used to help understand sources of uncertainty in the system. Much of the literature around sensitivity analysis has focused on univariate output, with some recent advances using multivariate correlated output. One promising method for multivariate sensitivity analysis involves decomposition through basis function expansion. However, these methods often require several model runs and may still be computationally intensive for practical purposes. Emulators have been a proven method for reducing computational time for univariate sensitivity analysis, with some recent development for multivariate computer models. We propose the use of generalized additive models and random forests combined with a principal component analysis for emulation for a multivariate sensitivity analysis. We demonstrate our method using a complex agricultural simulator.

Wednesday 29th 16:10 Gunnamatta

Bayesian Hypothesis Tests With Diffuse Priors: Can We Have Our Cake And Eat It Too?

John Ormerod, Michael Stewart, Weichang Yu, and Sarah Romanes University of Sydney

We introduce a new class of priors for Bayesian hypothesis testing, which we name "cake priors". These priors circumvent Bartlett's paradox (also called the Jeffreys-Lindley paradox); the problem associated with the use of diffuse priors leading to nonsensical statistical inferences. Cake priors allow the use of diffuse priors (having ones cake) while achieving theoretically justified inferences (eating it too). We demonstrate this methodology for Bayesian hypotheses tests for scenarios under which the one and two sample t-tests, and linear models are typically derived. The resulting test statistics take the form of a penalized likelihood ratio test statistic. By considering the sampling distribution under the null and alternative hypotheses we show for independent identically distributed regular parametric models that Bayesian hypothesis tests using cake priors are strongly Chernoff-consistent, i.e., achieve zero type I and II errors asymptotically. Lindley's paradox is also discussed.

Programme And Abstracts For Thursday 30th Of November

Keynote: Thursday 30th 9:00 Mantra

A General Framework For Functional Regression Modelling

Sonja Greven and Fabian Scheipl LMU Munich

Recent technological advances generate an increasing amount of functional data, data where each observation represents a curve or an image (Ramsay and Silverman, 2005). Examples of technologies that generate functional data include imaging techniques, accelerometers, spectroscopy and spectrometry. Any kind of measurement collected over time - data usually referred to as longitudinal - can also be viewed as potentially sparsely observed functional data.

Researchers are increasingly interested in regression models for functional data to relate functional observations to other variables of interest. We will discuss a comprehensive framework for additive (mixed) models for functional responses and/or functional covariates. The guiding principle is to reframe functional regression in terms of corresponding models for scalar data, allowing the adaptation of a large body of existing methods for these novel tasks. The framework encompasses many existing as well as new models. It includes regression for 'generalized' functional data, mean regression, quantile regression as well as generalized additive models for location, shape and scale (GAMLSS) for functional data. It admits many flexible linear, smooth or interaction terms of scalar and functional covariates as well as (functional) random effects and allows flexible choices of bases - in particular splines and functional principal components - and corresponding penalties for each term. It covers functional data observed on common (dense) or curve-specific (sparse) grids. Penalized likelihood based and gradient-boosting based inference for these models are implemented in R packages refund and FDboost, respectively. We also discuss identifiability and computational complexity for the functional regression models covered. A running example on a longitudinal multiple sclerosis imaging study serves to illustrate the flexibility and utility of the proposed model class. Reproducible code for this case study is also available online with the recent discussion paper Greven and Scheipl (2017) this talk is based on.

Thursday 30th 10:30 Narrabeen

Hockey Sticks And Broken Sticks – A Design For A Single-Arm, Placebo-Controlled, Double-Blind, Randomized Clinical Trial Suitable For Chronic Diseases

Hans Hockey[1] and Kristian Brock[2] [1]Biometric Matters Ltd. [2]Cancer Research UK Clinical Trials Unit

This work is motivated and exemplified by a genetic disorder causing early onset diabetes, blindness and deafness, which is extremely rare, inevitably fatal and has no current direct treatment. While the standard placebo-controlled RCT is the gold standard required by the regulatory agency for a new proposed drug study, it is conjectured that potential study participants will prefer a design which guarantees that they are always assigned to the drug under study. A single-arm design is proposed which meets this patient need and hence probably increases recruitment and compliance. At the same time, it meets the requirement for full randomization. Analyses which follow naturally from this design are also described and were used in trial simulations for sample sizing and for examination of the effect of underlying assumptions.

Thursday 30th 10:30 Gunnamatta

Bounding IV Estimates Using Mediation Analysis Thinking

Theis Lange[1,2] [1]University of Copenhagen [2]Peking University

In this paper we initially demonstrate that the well-known assumptions for conducting IV-analyses can equivalently be expressed using natural effects from mediation analysis. Viewing the assumptions, an indeed the whole IV analysis, from a mediation analysis perspective opens up a novel possibility for bounding the true causal effect of the exposure when the distributional assumptions of the IV analysis fail; e.g. if there is an interactions between the unmeasured confounders and exposure. The procedure works across all effect scales (risk difference, hazard ratio etc.) and types of violations of the distributional assumptions. The proposed method can also be viewed as a sensitivity analysis for the IV-analysis where there is only a single tuning parameter, which can be straightforwardly interpreted. For the purely binary case the proposed bounds converges to the Balke and Pearl bounds when the tuning parameter tends to infinity (i.e. when even the most extreme misspecification is considered). As the proposed method is computationally demanding some time is spent on implementation considerations.

Thursday 30th 10:50 Narrabeen

Correlated Bivariate Normal Competing Risks—Structuring Estimation In An Ill-Posed Problem

Malcolm Hudson Macquarie University

Estimation of correlation between competing risks has long been known to be an ill-posed problem, due to lack of identifiablity, termed the identifiability crisis [Crowder, Scand J Stat 1991].Recently, many semi-parametric models and a fully parametric model have been used to permit estimation and assess sensitivity of results to degree of correlation [Jeong and Fine Biostatistics 2007; see also Tai et al, SIM 2008]. Parametric models are non-standard and calculations complex; Jeong and Fine's parametric model employs a Gompertz distribution. For log-Normal data, complexity of calculations in a bivariate Normal (BVN) model is forbidding and direct optimization unstable.

We describe a parametric solution available in the BVN case. Our approach uses an EM algorithm for competing risks (generalising the Aitkin's EM for univariate survival). Complex calculations are evaluated in closed form using a lemma of Stein [Liu, Statist Prob Letters 1994]. This probabilistic and statistical platform for exploring the ill-posedness is implemented within the **bnc** R-package (in preparation).

We introduce these various components in the talk.

Thursday 30th 10:50 Gunnamatta

Bayesian Regression With Functional Inequality Constraints

Joshua Bon[1], Berwin Turlach[1], Kevin Murray[1], and Christopher Drovandi[2] [1]University of Western Australia [2]Queensland University of Technology

We investigate how to conduct Bayesian inference with functional inequality constraints over the parameter space. This problem is analogous to semi-infinite programming in optimisation. A novel method using Sequential Monte Carlo (SMC) is given. The SMC algorithm approximates the distribution of minima in order to successively iterate towards the constrained parameter space. We demonstrate on forensic morphometric data of human skulls where monotonicity is required in more than one dimension. Methods are compared to those currently available with one dimensional constraints and the results discussed.

Thursday 30th 11:10 Narrabeen

Genetic Analysis Of Renal Function In An Isolated Australian Indigenous Community

Russell Thomson[1], Brendan McMorran[2], Wendy Hoy[3], Matthew Jose[4], Tim Thornton[5], Gaétan Burgio[2], and Simon Foote[2] [1]Western Sydney University [2]ANU [3]University of Queensland [4]University of Tasmania [5]University of Washington

In close consultation with the local land council, and with ethical approval from many ethics committees, we have performed a genome-wide association study (GWAS) on a sample cohort from the 1990s. We also have a follow up data set of 120 study participants from 2014, with DNA sequence data.

I will discuss the statistical analyses of these data sets, with respect to issues of degraded DNA, high error rates and correlated samples.

Thursday 30th 11:10 Gunnamatta

The Performance Of Model Averaged Tail Area Confidence Intervals

Paul Kabaila La Trobe University

Commonly in applied statistics, there is some uncertainty as to which explanatory variables should be included in the model. Frequentist model averaging has been proposed as a method for properly incorporating this "model uncertainty" into confidence interval construction. Such proposals have been of particular interest in environmental and ecological statistics.

The earliest approach to the construction of frequentist model averaged confidence intervals was to first construct a model averaged estimator of the parameter of interest consisting of a data-based weighted average of the estimators of this parameter under the various models considered. The model averaged confidence interval is centered on this estimator and has width proportional to an estimate of the standard deviation of this estimator. However, the distributional assumption on which this confidence interval is based has been shown to be completely incorrect in large samples.

An important conceptual advance was made by Fletcher & Turek (2011) and Turek & Fletcher (2012) who put forward the idea of using data-based weighted averages across the models considered of procedures for constructing confidence intervals. In this way the model averaged confidence interval is constructed in a single step, rather than first constructing a model averaged estimator.

We review the work of Kabaila et al (2016, 2017) which evaluates the performance of the model averaged tail area confidence interval of Turek & Fletcher (2012) in the "test scenario" of two nested normal linear

regression models. Our assessment of this confidence interval is that it performs quite well in this scenario, provided that the data-based weight function is carefully chosen.

References:

1. Kabaila, P., Welsh, A.H., & Abeysekera, W. (2016) Model-averaged confidence intervals. Scandinavian Journal of Statistics.
2. Kabaila, P., Welsh, A.H. and Mainzer, R. (2017) The performance of model averaged tail area confidence intervals. Communications in Statistics - Theory and Methods.
Thursday 30th 11:30 Narrabeen

Deconstructing The Innate Immune Component Of A Molecular Network Of The Aging Frontal Cortex

Ellis Patrick[1], Mariko Taga[2], Marta Olah[2], Hans-Ulrich Klein[2], Charles White[3], Julie Schneider[4], Lori Chibnik[5], David Bennett[4], Sara Mostafavi[6], Elizabeth Bradshaw[2], and Philip De Jager[2] [1]University of Sydney [2]Columbia University [3]The Broad Institute [4]Rush University [5]Harvard University [6]Universtiy of British Columbia

Alzheimer's disease is pathologically characterized by the accumulation of neuritic -amyloid plaques and neurofibrillary tangles in the brain and clinically associated with a loss of cognitive function. The dysfunction of microglia cells has been proposed as one of the many cellular mechanisms that can lead to an increase in Alzheimer's disease pathology. Investigating the molecular underpinnings of microglia function could help isolate the causes of dysfunction while also providing context for broader gene expression changes already observed in mRNA profiles of the human cortex.

We have used mRNA sequencing to construct gene expression profiles of microglia purified from the cortex of 11 subjects from a longitudinal cohort of aging, Rush Memory Aging Project (MAP). By studying these microglia gene expression profiles in the context of tissue-level profiles of the cortex of 542 subjects from the MAP and Religious Orders Study (ROS) we address dual problems. By using information from the large ROSMAP cohort, we are able to isolate the genes which are strongly associated with immune response. Conversely, we illustrate that the microglia signature can be used to highlight predefined sets of coexpressed genes in ROSMAP that are highly enriched for microglia genes. Addressing these two questions allows us to identify sets of microglia specific genes which are associated with various Alzheimer's disease traits further emphasizing the molecular consequences of microglia dysfunction in this disease. Specifically, we are able to identify a set of microglia related genes associated with tau and amyloid pathology as well as an activated microglial morphology.

Thursday 30th 11:30 Gunnamatta

Bias Correction In Estimating Proportions By Pooled Testing

Graham Hepworth[1] and Brad Biggerstaff[2] [1]University of Melbourne [2]Centers for Disease Control and Prevention

Pooled testing (or group testing) arises when units are pooled together and tested as a group for the presence of an attribute, such as a disease. We have encountered pooled testing problems in plant disease assessment and prevalence estimation of mosquito-borne viruses.

In the estimation of proportions by pooled testing, the MLE is biased, and several methods of correcting the bias have been presented in previous studies. We propose a new estimator based on the bias correction method introduced by Firth (1993), which uses a modification of the score function. Our proposed estimator is almost unbiased across a range of problems, and superior to existing methods. We show that for equal pool sizes the new estimator is equivalent to the estimator proposed by Burrows (1987), which has been used by many practitioners.

Thursday 30th 11:50 Narrabeen

The Parametric Cure Fraction Model Of Ovarian Cancer

Serifat Folorunso, Angela Chukwu, and Akintunde Odukogbe University of Ibadan

We propose incorporating the Gamma link function to the generalized Gamma using a mixture cure fraction model. The mathematical properties of the proposed model were explored and the inferences for the models were obtained. The proposed model called Gamma- generalized gamma mixture cure model (GGGMCM) will be validated using a real life data set on ovarian cancer.

Thursday 30th 11:50 Gunnamatta

The Skillings-Mack Statistic For Ranks Data In Blocks

John Best University of Newcastle

Skillings and Mack gave a statistic for testing treatment differences for ranks data in blocks - both complete and incomplete blocks or blocks with missing values. This general statistic is thus quite useful. There is an R package for its calculation. However we illustrate a problem with tied ranks. Sensory evaluation data is used.

Poster Abstracts

Multi-Environment Trial Analysis Of Agronomy Trials Using Established Plant Population Density

Michael Mumford and Kerry Bell Queensland Department of Agriculture and Fisheries

There is a growing interest within the grains industry to determine the yield response to plant population density. This can vary depending on the trial location, the hybrid that is planted and the management practices that are employed.

In the past, the analysis of plant population density has been performed using the targeted plant population density as determined by the number of seeds initially planted. However, it has been found that the targeted plant population density may differ significantly from the established plant population density, i.e. the plant population density observed in the field. In such circumstances, it becomes important to use the established rather than targeted plant population densities in order to provide the most informative measure of yield response to plant density. This is especially true when performing a multi-environment trial (MET) analysis due to variation between the targeted and established plant population density at each trial.

We propose a procedure for performing a MET analysis using established plant population density as an explanatory variable. Using a linear mixed model framework, separate regression lines/curves are fit to each combination of trial and management practice, which we define as 'environment'. Environments are then grouped according to the results of a cluster analysis performed using the estimates of the regression coefficients, ensuring that environments within a cluster share a similar response to plant density. The lines/curves fitted to environments within clusters are then parallel, allowing us to then perform Fisher's least significant difference testing for yield differences between environments within each cluster.

This presentation will focus on the application of the proposed method to a large set of maize and sorghum agronomy trials conducted across New South Wales and Queensland over the last three years. For simplicity, a separate analysis was performed for each hybrid.

Adventures In Digital Agriculture In New Zealand

Esther Meenken and Vanessa Cave AgResearch

New Zealand is a nation for which primary industries are the economic backbone. New Zealand's highly variable landscape means there are a wide range in climatic and soil conditions, with multiple land uses existing in relatively small spatial areas. This means there are many possibilities to be explored in Digital Agriculture as advances in sensory technologies are made. The technical challenges known to be related to the use of sensors include issues around big data, missing data, biased and variable sensors, and various types of structured and non-normal data. We will describe a suite of studies that have been recently undertaken that can be loosely grouped under the Digital Agriculture umbrella. Case studies explore a range of scientific and statistical challenges in the primary industries arena, and include the use of classical and Bayesian Principled

Experimental Design, Hierarchical Bayesian Models, neural net classifiers, visualisation, animal tracking via GPS, and data management.

Modelling Canopy Greenness Over Time Using Splines And Non-Linear Regression

Bethany Macdonald[1], Jack Christopher[2], and Alison Kelly[1] [1]Queensland Department of Agriculture and Fisheries [2]Queensland Alliance for Agriculture and Food

The ability of a plant to retain leaf greenness for an extended time after anthesis, known as stay-green, has been linked to greater yield in wheat and can vary amongst genotypes. Many aspects of a plant's senescence contribute to stay-green; therefore, accurately modelling senescence and understanding the genetic variation in senescence dynamics is integral to selecting for the stay-green phenotype. Normalised difference vegetative index (NDVI) provides a measure of canopy greenness and when collected over time, can give an indication of when senescence begins and the rate at which it continues. Christopher et al. (2014) captured genetic variation in senescence dynamics using multiple stay-green traits derived from a logistic regression of longitudinal NDVI measurements. These traits were estimated independently for each plot within the experiment, ignoring any sources of covariance, and then subsequently analysed using a linear mixed model. However, no estimation errors were carried from the first to the second stage of the analysis.

We discuss two alternative one-stage methods for modelling senescence dynamics based on longitudinal NDVI measurements. Both of these methods enable the variability in senescence patterns to be partitioned into genetic and non-genetic components, while also incorporating the experimental structure. The first method involves the use of splines in a linear mixed model framework. This method allows an appropriate covariance structure for repeated measures to be utilised and, unlike other approaches, enables flexibility in the senescence patterns. The second method involves fitting logistic equations in a non-linear mixed model framework. The estimated parameters have biological interpretations, aiding in the genetic comparison of senescence dynamics. These two methods offer a more statistically robust approach to modelling crop senescence and the underlying genetic variability.

A Permutation Test For Comparing Predictive Values In Clinical Trials

Kouji Yamamoto and Kanae Takahashi Osaka City University

Screening tests or diagnostic tests are important for early detection and treatment of disease. There are four well-known measurements, sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) in diagnostic studies. For comparing SEs/SPs, McNemar test is widely used, but there are only few methods for the comparison of PPVs/NPVs. Moreover, all of these methods are based on large-sample theory.

So, in this talk, firstly, we investigate the performance of those methods when the sample size is small. In addition, we propose a permutation test for comparing two PPVs/NPVs we can apply even if the sample size is small. Finally, we show the performance of the proposed method with some existing methods via simulation studies.

On Testing Marginal Homogeneity For Square Contingency Tables With Ordinal Categories

Kouji Tahata Tokyo University of Science

For the analysis of square contingency tables with ordered categories, the marginal homogeneity model, which indicates the row marginal distribution is equal to the column marginal distribution, were considered. Stuart (1955) proposed the test statistics (denoted by Q) for the marginal homogeneity model. However it does not use the information of the category ordering. Agresti (1983) compared between Q and the Mann-Whitney test about the power by simulations. In this paper, we consider some measures to represent the degree of departure from the marginal homogeneity. For example, Tomizawa, Miyamoto and Ashihara (2003) and Tahata, Iwashita and Tomizawa (2008). In the situation given by Agresti (1983), we compare between Q and test based on these measures about the power.

A Factor Analytic Approach To Modelling Disease Progression Across Leaf Layers And Time

Clayton Forknall, Greg Platz, Lisle Snyman, and Alison Kelly Queensland Department of Agriculture and Fisheries

Controlling and limiting the impact of foliar diseases is a challenge faced by the Australian grains industry. Foliar diseases compromise and destroy photosynthetic area, thus limiting plant resources and adversely affecting crop productivity. Such diseases often infect the lower leaf layers of plants early in the season and, dependent on their compatibility with the host and environment, progress towards the topmost leaf layers over time.

Synthesizing the complicated dynamics of disease progression, over different leaf layers and across a growing season, into a measure to estimate the impact of loss of leaf area on productivity requires the assessment of the proportion of leaf area compromised by disease (LAD) at multiple times throughout the season. A simple measure that captures both LAD and disease duration on a given leaf layer is the Area Under the Disease Progress Curve (AUDPC), formed by applying the trapezoid rule to LAD assessments over time. The AUDPC is then often correlated to a measure of productivity for estimating the impact of disease at either a variety or experimental unit level.

We propose an alternative approach to the AUDPC to model disease progression over time and leaf layers more efficiently. Using a factor analytic approach, the covariance between leaf layers, both within and across assessment times, is captured on a variety basis in a linear mixed model framework. This approach will not only provide more reliable estimates of the disease pressure exerted on varieties, but also inform of relationships between leaf layers and how consistent these relationships are across time. Better quality information regarding the relationship between loss of leaf area and productivity can be made available to industry through the improved estimates and understanding of disease progression obtained from this modelling approach.

Associating Straw Strength With Likelihood Of Head Loss In Barley For Western Australia

Dean Diepeveen[1], Kefei Chen[1], Chengdao Li[2], and David Farleigh[3] [1]Curtin University [2]Murdoch University [3]Department of Primary Industries and Development

Barley headloss is characterised by straw breakage just below the head at or near plant maturity. Our recent research into barley headloss has enabled us to evaluate the various components of plant maturity and have found a clear genetic component to this industry issue. The more challenging components of this research is to understand why our more tolerant cultivars also show substantial headloss in certain unseasonal conditions during harvest. Our preliminary analyses uses environmental data with plant measurement data and looks at the incidences of extreme weather events on barley headloss. What confounds the analysis is the nutrient/growth conditions of the plant and the predisposition of these barley crops to headloss during these weather events. This paper will provide some lessons learnt and future directions.

A Deep Learning Neural Network Model To Identify The Important Genes In Metastatic Breast Cancer From Censored Microarray Data

Quoc-Anh Trinh Vietnam National University

Microarray data have been shown to correlated with survival in breast cancer. The most difficult in analysis of this data consist in a simultaneous measure of huge gene expression levels over a few patients available. The conventional survival models such as the proportional hazard model of Cox are no more appropriated including the hazard proportional hypothesis and the linear covariates effect hypothesis. Prognostic studies involve the identification of genes that correlate with survival in order to provide new information on pathogenesis. This result may aid in the research of drug design for new targets. We introduce in this paper a deep learning recurrent neural network approach to predict survival times for the individual patient based on microarray measurement. This neural network defines the hazard function of survival not to mention the proportionality and linearity hypothesis. We present a deep learning approach to the identification of genes critical for the metastases of breast cancer. We identified a set of highly interactive genes by analysing the connectivity matrices.

Using R And Shiny To Develop Web Based Sampling Applications For The Agricultural And Education Sectors

Peter Kasprzak, Olena Kravchuk, and Andy Timmins Univeristy of Adelaide

Sampling is an important aspect of agricultural statistics and an important feature of industrial reality. Enabling an efficient dialogue with researchers and agronomists about sound methods for field data collection will increase the efficiency and reliability of agricultural projects. The Shiny package in R allows interactive web applications to be created based upon existing R packages. This presentation will outline the steps involved in creating an app that implements sampling techniques such as SRS, Stratified Sampling, and Ratio Sampling, within a visually appealing graphical environment for use in the education and agricultural sectors.

Graphical Network Analyses Informs Biomarker Discovery Via Quantification Of Key Disease Related Connections

James Doecke CSIRO

Determination of the optimal set of biomarkers often relies on large scale feature selection, that selects a minimal set of independent biomarkers, without consideration of any information regarding the biological network. This research is aimed at defining biological network structures related to disease phenotype, through a three-level statistical process. Briefly, we first implement an unsupervised correlation structure across the set of biomarkers, used to define clusters of biomarkers via strict thresholding. Next, using the graphical network analysis method Differential Network Analysis in Genomics (DINGO), we define the differential network structure for healthy and disease states. Lastly from the derived differential connectivity score we cluster the individual groups of biomarkers. From this process, we are able to produce a network of similarly grouped sets of biomarkers that can separate healthy from disease subjects. Lastly we perform a random effects model for each cluster to estimate the association with disease status. Using this design, we have been able to incorporate information from multiple correlated biomarkers to increase the proportion of variance explained in disease status. We test this design on two data sets, 1) a set of ~2000 peptide biomarkers, with the express aim of defining sets of biomarkers associated with Alzheimer's Disease pathology, and 2) a set of ~2000 mRNA probe sets combined with a set of ~2000 DNA methylation probes with the express

aim of defining sets of biomarkers associated with breast cancer. We present network plots and connectivity structures for each data set, and highlight the increased performance of clustered biomarkers as compared to individual biomarkers to separate disease status. Lastly, we show that by utilizing a network approach, we can incorporate more biological information from the data to explain outcome.

Spatio-Temporal Cortical Brain Atrophy Patterns Of Alzheimer's Disease

Marcela Cespedes[1], James McGree[1], Christopher Drovandi[1], Kerrie Mengersen[2], James Doecke[3], and Jurgen Fripp[3] [1]Queensland University of Technology [2]Queensland University of Technlogy [3]CSIRO

The degeneration of the cerebral cortex is a complex process which often spans decades. This degeneration can be evaluated on regions of interest (ROI) in the brain through probabilistic network analysis. However, current approaches for finding such networks have the following limitations: 1) analysis at discrete age groups cannot appropriately account for connectivity dynamics over time; and 2) morphological tissue changes are seldom unified with networks, despite known dependencies. To overcome these limitations, a probabilistic dynamic wombled model is proposed to simultaneously estimate ROI cortical thickness and network continuously over age, and was compared to an age aggregated model. The inclusion of age in the network model was motivated by the interest in investigating the point in time when connections alter as well as the length of time required for changes to occur. Our method was validated via a simulation study, and applied to healthy controls (HC) and clinically diagnosed Alzheimer's disease (AD) groups. The probability of a link between the middle temporal (a key AD region) and the posterior cingulate gyrus decreased from age 55 (posterior probability > 0.9), and was absent by age 70 in the AD network (posterior probability < 0.12). The same connection in the HC network remained present throughout ages 55 to 95 (posterior probability 0.75). The analyses presented in this work will help practitioners choose suitable statistical methods to identify key points in time when brain covariance connections change, in addition to morphological tissue estimates, which could potentially allow for more targeted therapeutic interventions.

Empirical Modelling Of Fruit Firmness Change During Colour Conditioning

Lindy Guo and Ringo Feng Plant and Food

Gold3 (Actinidia. chinensis *Zesy002*, marketed as Zespri® SunGold Kiwifruit) is a new yellow-fleshed kiwifruit cultivar being used to replace *Hort16A* that has been almost all destroyed by the outbreak of Pseudomonas syringae pv. actinidiae (Psa) in 2010. The Smart Monitoring programme started in 2011, records fruit development, maturation, harvesting and storage performance of Gold3 kiwifruit from 12 orchards in a fairly consistent manner year after year.

The aim of this project is to summarize fruit development and storage potentials over the 5 years, and hence provide a better understanding of the roles of seasonal weather conditions, orchard management, and fruit attributes at harvest affecting storage potential.

For this project, storage life was defined as storage time for a batch of fruit (fruit harvested from an orchard on a particular date) to soften to 1 kfg. Several ways to calculated storage life based on firmness monitoring data are compared and the storage lives were correlated with weather conditions, orchard management, and fruit attributes at harvest using different algorithms. The results will be presented to highlight experiences learnt in this data analysis. Recommendations will be made on dealing with multiyear data with collinear variables.