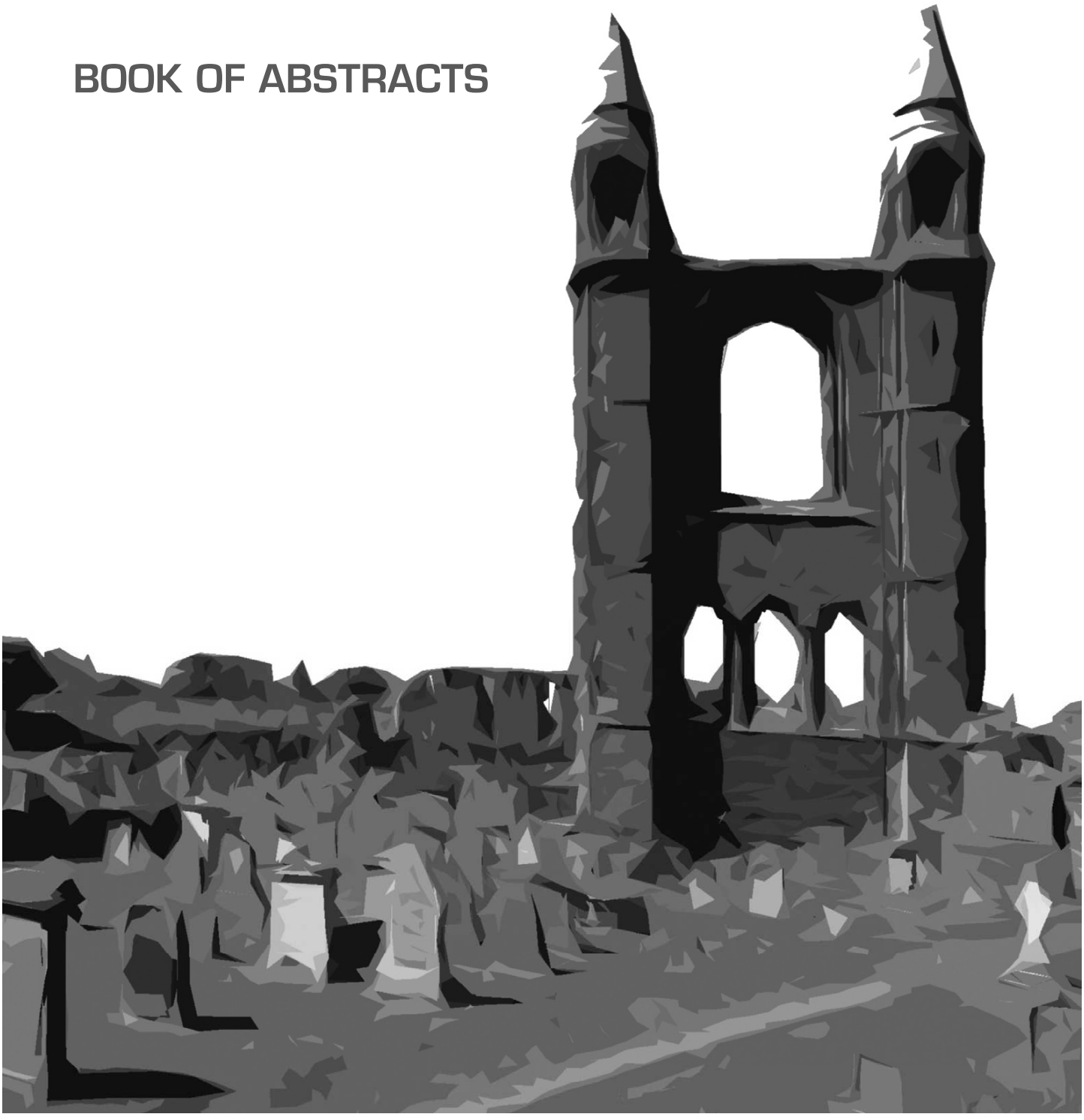
# 4TH CHANNEL NETWORK CONFERENCE

UNIVERSITY OF ST ANDREWS

## BOOK OF ABSTRACTS



# TWO-PART REGRESSION CALIBRATION MODEL TO CORRECT FOR MEASUREMENT ERROR : A SIMULATION STUDY

George O. Agogo[*1], Hendriek Boshuizen[1,2]

[1] *Wageningen University and Research Centre, Wageningen, The Netherlands*
[2] *National Institute for Public Health and the Environment, Bilthoven, The Netherlands*

[*]E-mail: *george.agogo@wur.nl*

Measurement error in predictor variables is a common problem in many research areas such as in epidemiology where the risk-exposure association is of interest. The error is mostly known to attenuate the association leading to estimates biased toward the null.

As a result, most studies, for instance, in nutritional epidemiology, conduct a calibration sub-study, where a more precise but expensive reference instrument, whose measurements are assumed unbiased for the true measurements, is administered in a sample of subjects selected from the study population where the imprecise and inexpensive main instrument is used. Then the measurements from the former are used to correct for error in the imprecise measurements from the latter using regression calibration.

Further complications arise when the reference measurements have excess zeros and taken only once per study subjects, for instance, for episodically consumed foods in nutritional studies. To circumvent this, a two-part regression calibration modeling approach is adapted.

Therefore, this presentation will describe how a two-part regression calibration model is adapted to correct for measurement error in main instrument measurements using zero-inflated reference measurements with single measurement per subject for the predictor of interest.

First, a formulation of the two-part regression calibration is shown for single measures data setting. Secondly, a simulation study design is described to mimic European Investigation into Cancer and Nutrition (EPIC) Study. As a result, a two-part Gamma model is developed. Also included is a naïve one-part linear model. The resulting calibrated values are then used in place of the imprecise ones in the Cox model and the models are assessed using bias, mean squared error and coverage probabilities for the log hazard ratio.

From this study, we conclude that the two-part regression calibration seems to correct for error on the predictor variable better than the one-part model.

Key Words: Attenuation; Cox model; measurement error; two-part regression calibration

# DESIGN AND ANALYSIS OF EXPERIMENTS IN ECOLOGY

Rosemary A. Bailey*[1]

[1] *University of St Andrews, St Andrews, UK*

*E-mail: *r.a.bailey@qmul.ac.uk*

It is now widely believed that biological diversity is good for the environment. One way that ecologists test this is to place random collections of species in mini-environments and then measure some outcome. I have been working with a group of fresh-water ecologists to improve this in two ways. The first is that our subsets of species are carefully chosen, not random. The second is that we fit a nested family of plausible models. Our results suggest that the underlying model is not diversity at all.

# POWERFUL TESTING VIA HIERARCHICAL LINKAGE DISEQUILIBRIUM IN HAPLOTYPE ASSOCIATION STUDIES

Brunilda Balliu*[1] and Stefan Boehringer[1]

[1]*Dept of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands*

*E-mail: *bballiu@lumc.nl*

Haplotypes play key roles in the study of the genetic basis of disease. Studies have shown that haplotype-based methods may provide more power and accuracy in disease gene mapping than those based on single markers. Traditional haplotype association methods test for differences between haplotype distribution of cases and controls using for example a likelihood ratio test.

A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of loci, incurring many degrees of freedom (df) and weakening the power to detect associations. Moreover, the success of these approaches depends on getting the ”right size” haplotypes. If the haplotypes are too long to cover relevant correlations (e.g. more than 10 loci), such approaches are not feasible. These situations can occur when relatively recent mutations have introduced long range correlations in low linkage disequilibrium (LD) regions.

To address these limitations, we propose hierarchical modeling of LD for disease mapping. We develop a new parametrization of the haplotype distribution where every parameter corresponds to the cumulant of each possible subset of a set of loci. That is, the new parametrization consists of the allele frequencies at each locus, the pairwise, and higher-order $(3,\ldots,N)$ LD parameters. This introduces a hierarchy in hour parameters and enables us, for example, to selectively test differences that are described in terms of a certain number of loci, ignoring higher order parameters and sparing df to test for the full haplotype.

We perform a simulation study and show that our approach maintains the type I error at nominal level and has increased power under many realistic scenarios, as compared to traditional haplotype-based association studies. To evaluate the performance of our proposed methodology in real data we analyze data from the WTCCC study, a Genome Wide Association Study on Rheumatoid Arthritis.

# CHAIN EVENT GRAPHS FOR INFORMED VARIABLE CONSTRUCTION IN EPIDEMIOLOGY

Lorna M. Barclay*[1], Jane L. Hutton[1], Jim Q. Smith[1]

[1] *University of Warwick, Coventry, UK*

*E-mail: *L.M.Barclay@warwick.ac.uk*

Chain Event Graphs (CEGs) are a new class of graphical models which are derived from a probability tree by merging the vertices in the tree whose associated conditional probability distributions are the same. It generalises more common graphical models, such as the discrete Bayesian Network, by allowing for asymmetries with the dependence structure of the variables, while still retaining the property that conclusions can easily be read back to the client via its graph.

The advantage of employing CEGs for modelling discrete processes which exhibit strong asymmetric dependence structures can be particularly exploited to represent studies where missingness is influential and data cannot plausibly be hypothesised to be missing at random. Further the CEG retains the paths of the original probability tree within its graphical structure hence giving a detailed explanation of how a combination of variables can affect an outcome, such as survival. Consequently, we show how the CEG can provide a useful framework for defining categories of variables which are informative in competing risk models and survival analysis.

We illustrate this use of the CEG in a study of cause of death for a large cerebral palsy cohort study. We consider survival up to the age of ten with respect to various disabilities. It is already known that each observed severe impairment is associated with poorer survival. We demonstrate how the CEG provides further insight into the way in which the severity and the number of impairments, including missing values, influence survival. This allows us to make informed decisions about the construction of the covariates with reduced degrees of freedom which reflect the number of severe disabilities. These covariates can then be used in multinomial cause of death models, or competing risks models for survival.

# THE SIMEX METHOD FOR CORRECTING THE BIAS IN A SURVIVAL CURE MODEL WITH MISMEASURED COVARIATES

Aurélie Bertrand*[1], Catherine Legrand[1], Ingrid Van Keilegom[1]

[1] *Université catholique de Louvain, Louvain-la-Neuve, Belgium*

*E-mail: *aurelie.bertrand@uclouvain.be*

In traditional survival analysis, all subjects in the population are assumed to be susceptible to the event of interest: every subject has either already experienced the event or will experience it in the future. In many situations, however, it may happen that a fraction of individuals (long-term survivors) will never experience the event: they are considered to be cured. The promotion time cure model is one of the survival models taking this feature into account.

We consider the case where the explanatory variables in the model are supposed to be subject to measurement error. This occurs e.g. when the instrument used to measure the variable (blood pressure, cholesterol level, etc.) has some calibration error. This measurement error should be taken into account in the estimation of the model, to avoid biased estimators of the model. In the literature, several approaches to correct this bias have been proposed. The SIMEX algorithm is one of them: it is a method based on simulations which allows to estimate the effect of measurement error on the bias of the estimators and to reduce this bias. It has already been applied to many different models, but not to the promotion time cure one. For this model, Ma and Yin (2008) have suggested a corrected score approach.

We extend the SIMEX approach to the promotion time cure model. We show via simulations that the suggested method performs well in practice by comparing it with the method proposed by Ma and Yin (2008), which is, as far as we know, the only paper that has studied this problem before in the literature.

## References

Ma, Y. and Yin, G. (2008), Cure rate model with mismeasured covariates under transformation. *Journal of the American Statistical Association*, **103**, 743–756.

# ESTIMATING AND COMPARING DYNAMIC PREDICTIVE ACCURACY OF JOINT MODELS FOR TIME-TO-EVENT AND LONGITUDINAL DATA

Paul Blanche[*1,2], Cécile Proust-Lima[1,2], Lucie Loubère[1,2] and Hélène Jacqmin-Gadda [1,2]

[1]*INSERM, U897 Epidemiology and Biostatistics Research center, Bordeaux, France*
[2]*University of Bordeaux 2, Bordeaux, France*

[*]E-mail: *Paul.Blanche@isped.u-bordeaux2.fr*

Using longitudinal cohort data, it is often of clinical interest to investigate how a biomarker that is repeatedly measured in time is associated with a time-to-event. Such a research question has already given birth to a large litterature about joint modeling of longitudinal and time-to-event data. Due to the growing interest on personalized medicine, joint models have more recently started to be used to predict individual risk of event. Motivated by elderly cohort data, we are interested in our work in predicting the onset of dementia using repeated cognitive test measurements. Individual predictions are dynamic because there are updated when information on the subject's health profile grows with time.

Despite a lot of different statistical approaches for making such dynamic predictions, there is no clear rule to choose among them. We therefore focus in this work on statistical methods for quantifying and comparing dynamic predictive accuracy of different prognostic tools, accounting for right censoring and possibly competing risks. We use dynamic area under the ROC curve (AUROC) and Brier Score to quantify predictive accuracy. Nonparametric inverse probability of censoring weighting technique is used to estimate dynamic curves of AUROC and Brier Score as functions of the time at which predictions are made. Asymptotic results are established and pointwise confidence intervals as well as confidence bands are derived. Comparison tests are also presented to test superiority of dynamic prediction accuracy of one prognostic tool to another.

The finite sample behaviour of the inference procedures are assessed via simulations. We then apply the proposed methodology to compare different prediction tools using repeated measurements of several psychometric tests to predict dementia in the elderly, accounting for the competing risk of death. Data comes from the French PAQUID and 3C cohorts.

# Visualization of classification decisions based on geometric predictors

Brunilda Balliu[1], Stefan Boehringer*[1]

[1]*Leiden University Medical Center, Leiden, The Netherlands*

*E-mail: *s.boehringer@lumc.nl*

Classification problems with geometrically interpretable predictors, e.g. graphs, allow for visualization of classifiers. For regression techniques, coefficients associated with predictor variables have to be associated with their corresponding geometric interpretation and used for highlighting. Often data gets transformed prior to classification, for example by converting graph vertices into pair-wise distances or similar transformations. Such transformations can induce ill-posed problems by creating more predictors than observations which require penalized methods to achieve regularized solutions. We consider visualization of such penalized predictors which are known to have biased coefficient estimates. A first approach assigns weights to each point in a mean graph corresponding to the importance of that point for the classification process and illustrate the choice of weights for various transformations (distances, angles, areas). A second approach tries to create caricatures, i.e. graphs that overemphasize features of one class with respect to the other. We define penalty functions on graph deformations that are chosen such as to produce caricatures when the penalty is minimized. The penalties account for the transformations and the biased nature of the estimates. We discuss algorithms to compute these transformations. We illustrate our visualization techniques on a discrimination problem of 2D images where genetic syndromes are discriminated using LASSO regression.

# DISCRETE- AND CONTINUOUS-TIME MODELS FOR LINE TRANSECT AND CAPTURE-RECAPTURE SURVEYS

David L. Borchers*, Roland Langrock, Greg B. Distiller

*University of St Andrews, St Andrews, UK*

*E-mail: *dlb@st-andrews.ac.uk*

Line transect and capture-recapture methods are two of the most widely used methods for estimating animal abundance. Although both typically operate in continuous time, line transect surveys are usually analysed as if they were instantaneous and capture-recapture surveys are usually analysed using discrete time units (capture occasions). We consider circumstances in which the usual way of dealing with time is not ideal.

In the case of line transect surveys, one has to take account of time when animal availability changes (stochastically) with time. And although the survey process operates in continuous time, it can be convenient to discretise time for analysis. We present and compare discrete-time and continuous-time models for line transect surveys with stochastic animal availability. The discrete-time model integrates a detection process model and a hidden Markov model (or hidden semi-Markov model) for the availability process in discrete time. The continuous-time model does the same in continuous time by using a Markov-modulated Poisson process in place of a hidden Markov model.

We also consider spatially explicit capture-recapture surveys, and present a continuous-time model for these. We show how discrete-time models arise from this model, and that continuous-time models have a number of advantages over discrete-time models when the capture process operates in continuous-time – as it does on most capture-recapture surveys of wildlife populations.

We conclude by pointing out the similarity between line transect and spatially explicit capture-recapture models, and speculating on the future role of continuous-time models in this context.

# ESTIMATING MISCLASSIFICATION RATES OF INCIDENT CASES IN GENERAL PRACTITIONER REGISTRATION NETWORKS

Hendriek C. Boshuizen[*1,2], Nancy Hoeymans[2]

[1]*National Institute of Public Health and the Environment, Bilthoven, the Netherlands*
[2]*Wageningen University, Wageningen, the Netherlands*

[*]E-mail: *hendriek.boshuizen@rivm.nl*

For health policy purposes, population health need to be monitored. Registries in general practice (GP) are key sources for morbidity estimates, as in the Netherlands the general practitioner is the gatekeeper of health care. However, morbidity estimates between different GP registration networks vary considerably, and in previous research could not be explained by differences in characteristics of the patient population. Here we investigate whether these differences can be explained by differences between networks in the accuracy to distinguishing new (incident) cases from existing (prevalent) cases, and if so, whether modeling this enables more reliable estimates. We used data from five Dutch general practice registration networks and data on four chronic diseases (COPD, diabetes, heart failure, and osteoarthritis). Data comprised the number of prevalent cases, incident cases and selective outflow (mortality and institutionalization rates in those with and without the disease) for 2010. We fitted a joint model (DISMOD model), that links these three together, as prevalence results from incidence (inflow) and outflow. The model assumes stable incidence and survival probabilities over time (within the duration of disease), and includes a misclassification factor that allows for misclassification of a percentage of prevalent cases as incident cases. We fitted this model by maximum likelihood, and compared the prevalence and incidence estimates for each network with the observed prevalence and incidence estimates. Including a misclassification term mostly effected incidence estimates, but did not systematically decrease the variation between prevalence and incidence estimates from different networks. Osteoarthritis of the knee showed large misclassifications, especially (but not exclusively) in episode based registrations. For the other diseases, modeling misclassification rates does not systematically decrease the variation between registration networks. However, such a modeling exercise does give qualitative insight in the reliability of estimates.

# THE USE OF INTERACTIVE EBOOKS FOR TEACHING BAYESIAN STATISTICAL MODELLING USING THE STAT-JR PACKAGE

William J. Browne*[1], Richard Parker[1], Christopher Charlton[1], Camille Szmaragd[1] and Danius Michaelides[2]

[1]*University of Bristol, Bristol, UK*
[2]*University of Southampton, Southampton, UK*

*E-mail: *william.browne@bristol.ac.uk*

As part of an ESRC funded grant we have developed a new statistical software package, Stat-JR with many novel features including interoperability with most of the commonly used statistical software packages. The package uses a web browser as a user-friendly interface but also has a novel eBook interface. Our interactive eBooks combine the best features of books and statistical software packages and can embed (interactive) statistical analyses within the text of a web document. Thus as the reader reads the document they interact with the book, for example changing parameters in a model, and the package then performs the modelling for the new inputs and updates the book accordingly. In this talk we will introduce and demonstrate Stat-JR and its eBook interface and show some of its features. These will include displaying MCMC algorithms that are specific to the chosen model and dataset, linking to other packages to use the best of their features or simply compare the estimates across packages for the same model. We will use as examples a standard multilevel statistical model and a capture-recapture model as used in statistical ecology.

# PARAMETER ESTIMATION IN DIFFERENTIAL EQUATIONS WITH ORTHOGONALITY CONDITIONS

Nicolas J-B. Brunel[*1,2], Quentin Clairon[1],

[1]*Fédération de Mathématiques, Université d'Évry Val d'Essonne, Évry, France*
[2]*École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Évry, France*

[*]E-mail: *nicolas.brunel@ensiie.fr*

Differential equations are commonly used to model dynamical deterministic systems in pharmacokinetics, population ecology, or cell biology. These models are often known up to some parameters that need to be estimated from experimental data for mathematical analysis or prediction.
The calibration of Ordinary Differential Equations (ODEs) with Nonlinear Least Squares (or MLE) are often confronted to complex and ill-posed optimization problem. As a consequence, alternative estimators are useful for obtaining reliable estimates. With a "Functional Data Analysis" point of view, we propose a gradient matching approach for the estimation of parametric ODEs observed with noise. Starting from a nonparametric proxy of a true solution of the ODE, we build a parametric estimator based on a variational characterization of the solution. As a Generalized Moment Estimator, our estimator must satisfy a set of orthogonal conditions that are solved in the least squares sense. Despite the use of a nonparametric estimator, we prove the root-$n$ consistency and asymptotic normality of the Orthogonal Conditions estimator. We derive confidence sets thanks to a closed-form expression for the asymptotic variance, and we give a practical way to compute and optimize the variance by adaptive reweighting.
The estimator is compared to classical estimators in several (simulated and real) settings and ODE models in order to show its versatility and relevance with respect to classical Gradient Matching and Nonlinear Least Squares estimators. In particular, we show on a real dataset of influenza infection that the approach gives reliable estimates. Moreover, we show we can deal directly with more elaborated models such as Delay Differential Equation (DDE) by considering a blowfly population model fitted on Nicholson's experiments. We obtain parameters and confidence sets without solving the DDE, and the values are consistent with the ones obtained with the ABC method.

# PREDICTIVE MODELS FOR THE SPATIAL DISTRIBUTION OF SEABIRD FORAGING

Adam Butler[*1], Ellie Owen[2], Mark Bolton[3]

[1]*Biomathematics and Statistics Scotland, Edinburgh, UK*
[2]*Royal Society for the Protection of Birds, Inverness, UK*
[3]*Royal Society for the Protection of Birds, Sandy (Cambridgeshire), UK*

[*]E-mail: *adam@bioss.ac.uk*

Seabirds are a key part of the marine ecosystem, but a substantial number of seabird species have experienced declines over recent years and 97 out of 346 seabird species are currently classed by the IUCN Red List Index as "Globally Threatened". In order to protect seabird species it is important to identify the areas of sea that are most closely associated with foraging, in order to manage human activities within these areas - including fishing and the development of offshore renewables - so that they do not have an adverse impact upon seabird populations.

The widespread availability of GPS tags mean that it has now become possible to collect individual-level data on seabird behaviour and movement at high temporal resolution. In this talk we outline some of the statistical methodologies and challenges that are associated with using GPS tagging data to draw inferences about the spatial distribution of seabird foraging, and illustrate the methods and issues using results from the analysis of data collected for the FAME project (Future of the Atlantic Marine Environment). We will focus primarily upon the problem of inferring the spatial distribution of foraging for birds from unmonitored colonies - those for which we have no GPS data - based on using statistical models to describe the relationship between foraging density and environmental covariates at monitored colonies. We initially present the results associated with using a Binomial GLM to compare the environmental conditions associated with presences and pseudo-absences. These analyzes highlight substantial regional differences in foraging ranges, and identify more subtle relationships with environmental variables. We then compare these results against those obtained using inhomogeneous Poisson point process models (which provide a more theoretically rigorous basis for modelling spatial associations) and hidden Markov models (which incorporate the uncertainty associated with using GPS tags to classfiy behaviour as foraging or non-foraging).

# SAMPLE SIZE CALCULATION FOR TREATMENT EFFECTS IN TWO-ARMED RANDOMIZED TRIALS WITH FIXED CLUSTER SIZES AND HETEROGENEOUS CLUSTERING

Math J.J.M. Candel[*1], Gerard J.P. Van Breukelen[1]

[1]*University of Maastricht, Maastricht, The Netherlands*

[*]E-mail: *math.candel@maastrichtuniversity.nl*

When comparing two different groupwise treatments or two individual treatments where patients within each arm are nested within care providers, clustering of observations may occur in both arms. The arms may differ in terms of (a) the intraclass correlation, (b) the outcome variance, (c) the cluster size and (d) the number of clusters, and there may be some ideal group size or ideal caseload in case of care providers, fixing the cluster size. Employing a flexible cost function, the optimal numbers of clusters for the treatment arms are derived for a linear mixed model analysis of the treatment effect. To account for uncertain prior knowledge on relevant model parameters, also maximin sample sizes are given. Formulas for sample size calculation are derived, based on the standard normal as the asymptotic distribution of the test statistic. For small sample sizes, an extensive numerical evaluation shows that in a two-tailed test employing restricted maximum likelihood estimation, a safe correction for both 80% and 90% power, is to add 3 clusters to the number of clusters in each arm for a 5% type I error rate, and 4 clusters to each arm for a 1% type I error rate.

Keywords: individually randomized group treatment, maximin design, optimal design, sample size calculation, therapist effects

# ANALYSIS OF LONGITUDINAL TRIALS WITH PROTOCOL DEVIATION: A FRAMEWORK FOR RELEVANT, ACCESSIBLE ASSUMPTIONS, AND INFERNECE VIA MULTIPLE IMPUTATION

James R Carpenter[*1,2], James H Roger[1], Michael G Kenward[2]

[1]*London School of Hygiene & Tropical Medicine, London, UK*
[2]*MRC Clinical Trials Unit, London, UK*

[*]E-mail: *james.carpenter@lshtm.ac.uk*

Protocol deviations, for example due to early withdrawal and non-compliance, are unavoidable in clinical trials. Such deviations often result in missing data. Additional assumptions are then needed for the analysis, and these cannot be definitively verified from the data at hand. Thus, as recognized by recent regulatory guidelines and reports, clarity about these assumptions and their implications is vital for both the primary analysis and framing relevant sensitivity analysis.

This poster focuses on clinical trials with longitudinal quantitative outcome data. For the target population, we define two estimands, the de-jure estimand, does the treatment work under the best case scenario and the de-facto estimand what would be the e!ect seen in practice. We then carefully define the concept of a deviation from the protocol relevant to the estimand, or for short a deviation. Each patients post-randomization data can then be divided into pre-deviation data and post-deviation data. We set out an accessible framework for contextually appropriate assumptions relevant to de-facto and de-jure estimands, i.e. assumptions about the joint distribution of pre- and post-deviation data relevant to the clinical question at hand. We then show how, under these assumptions, multiple imputation provides a practical approach to estimation and inference.

We illustrate with data from a longitudinal clinical trial in patients with chronic asthma.

# COMPARING CHANGE-POINT LOCATION OF INDEPENDANT PROFILES

Alice Cleynen*[1], Stéphane Robin[1]

[1]*AgroParisTech, Paris, France*

*E-mail: *alice.cleynen@agroparistech.fr*

We are interested in the comparison of the length of Un-Translated Region of genes (UTR) of an organism grown in different conditions for which we have performed RNA-Seq experiments. This problem can be addressed in a segmentation framework where the issue is the comparison of change-point locations. The data consists in read counts associated to each position along the genome. Using the negative binomial distribution with known over-dispersion parameter $\phi$, we can compute, in a Bayesian framework, the exact posterior distribution of change-point locations.

For example, in order to assess whether the first change-point is located at the same position under all conditions, we compute the posterior probability for such an event to occur given the profiles. Given that $\phi$ is known, this computation can be done exactly and in quadratic time, providing us with a natural decision rule.

We propose to estimate the over-dispersion using a modified version of Johnson, Kotz and Kemp's estimator. We assess the impact of estimating $\phi$ in a simulation study where the data is sampled from the negative binomial distribution. We then illustrate our method on yeast RNA-Seq data, which allows us to suggest that alternative splicing may exist for this organism.

# DETECTING PARAMETER REDUNDANCY IN ECOLOGICAL STATE-SPACE MODELS

Diana J. Cole*[1], Rachel M. McCrea[1]

[1] *University of Kent, Canterbury, UK*

*E-mail: *d.j.cole@kent.ac.uk*

To be able to fit or examine a parametric model successfully all the parameters need to be identifiable. If the parameters are non-identifiable the model can be rewritten in terms of a smaller set of parameters, and is termed parameter redundant. Parameter redundancy is not always obvious, in which case the definitive method for detecting parameter redundancy involves calculating the rank of a derivative matrix, which is expressed symbolically (see for example Cole et al., 2010, Mathematical Biosciences, 228, 16-30).
State-space models can be used in ecology to describe counts or census data. Here we extend this derivative matrix method to diagnose parameter redundancy within state-space models.
State-space models consist of two component processes: an underlying state process which is unobserved and the observation process. Because only part of the system is observed many state-space models alone are parameter redundant. We therefore also discuss how to determine parameter redundancy in models that combine state-space modelling with the modelling of other types of data. One example is provided by integrated population modelling (see for example Besbeas et al, 2002, Biometrics, 58, 540-547).

# THE STOCHASTIC SYSTEM APPROACH FOR CAUSAL MODELING

Daniel Commenges*[1], Mélanie Prague[1], Anne Gégout-Petit[2], Rodolphe Thiébaut[1]

[1] *INSERM, Bordeaux, France*
[2] *Bordeaux University, Bordeaux, France*

*E-mail: *daniel.commenges@isped.u-bordeaux2.fr*

We compare different approaches for estimating causal effects of a treatment in observational studies. As an example, we focus on the analysis of the effect of a HAART on CD4 counts and viral load, where attribution of the treatment may depend on the observed marker values. This problem has been treated using marginal structural models relying on the counterfactual/potential response formalism. We show that the so-called "stochastic system " approach, based on the Doob-Meyer decomposition, can give an equivalent solution with several advantages: avoidance of the complexity of the counterfactuals, more natural mechanistic modeling, easier generalization. Moreover, inference based on maximum likelihood should be more efficient. The formalism of this approach will be developed. Then we show that more realistic models can be developed involving distinguishing model for the system and model for the observations, acknowledging that the system lives in continuous time, and expressing mechanism in the form of a system of differential equations. This leads us to mechanistic models which are much more challenging, particularly from a numerical point of view, but which can yield much richer and reliable results.

# EFFICIENT ESTIMATION OF THE DISTRIBUTION OF TIME TO COMPOSITE ENDPOINT WHEN SOME ENDPOINTS ARE ONLY PARTIALLY OBSERVED

Rhian M. Daniel*[1], Anastasios A. Tsiatis[2]

[1]*London School of Hygiene and Tropical Medicine, London, UK*
[2]*North Carolina State University, Raleigh NC, USA*

*E-mail: *Rhian.Daniel@LSHTM.ac.uk*

Two common features of clinical trials, and other longitudinal studies, are (1) a primary interest in composite endpoints, and (2) the problem of subjects withdrawing prematurely from the study. In some settings, withdrawal may only affect observation of some components of the composite endpoint, for example when another component is death, information on which may be available from a national registry. In this talk, we use the semiparametric theory of augmented inverse probability weighted estimating equations to show how such partial information on the composite endpoint for subjects who withdraw prematurely from the study can be incorporated in a principled way into the estimation of the distribution of time to composite endpoint, typically leading to increased efficiency without relying on additional assumptions above those that would be made by standard approaches. We give some background on semiparametric theory and counting processes, describe possible approaches in this setting, including our proposed augmented inverse probability weighted estimator, discuss its desirable properties, namely semiparametric efficiency and double robustness, and confirm these properties using a simulation study.

# A CONTINUOUS-TIME SEMI-MARKOV MODEL FOR SPERM WHALE BEHAVIOR, WITH AND WITHOUT ACOUSTIC DISTURBANCE

Stacy L. DeRuiter*[1], Patrick J. O. Miller[1], Megan Rose[2], Brandon Southall[3], Alison K. Stimpert[2], Len Thomas[1], Peter L. Tyack[1]

[1]*University of St Andrews, St Andrews, UK*
[2]*Naval Postgraduate School, Monterey, California, USA*
[3]*SEA, Inc., Aptos, California, USA*

*E-mail: *sldr@st-andrews.ac.uk*

In order to quantify changes in sperm whale echolocation-based foraging behaviour in response to acoustic disturbance, we model the whales' behaviour as a semi-Markov process in continuous time. Our dataset consists of recordings from DTAGs placed on sperm whales, including acoustic data as well as the whale's dive depth, posture and movements. In addition to baseline data collected in the absence of acoustic disturbance, we applied the model to data from 8 whales experimentally exposed to the sounds of airgun arrays (a source of intense, impulsive low-frequency sound used for geophysical mapping and prospecting), military mid-frequency sonar (used in anti-submarine warfare), or pseudo-random noise. The tag acoustic records allowed unambiguous determination of the animal's behaviour state: production of echolocation clicks indicated searching for prey, a buzz (a series of very rapid clicks) indicated an attempt to capture prey, and non-foraging periods lacked echolocation clicks. Models were fit using maximum likelihood, and a likelihood-based model selection framework governed inclusion in the model of parameters allowing inter-individual differences in behaviour and the level and type of acoustic disturbance as covariates. A combination of graphical methods and parametric and non-parametric bootstraps allowed further assessment of model goodness-of-fit and construction of confidence bounds on the parameter estimates. Overall, the results allow us to detect and describe subtle changes in sperm whale foraging behaviour in response to acoustic disturbance from several anthropogenic sources, with a general trend toward changes likely to reduce foraging efficiency. These descriptions of short-term behavioural responses to acoustic disturbance can be combined with data on the frequency and durations of such disturbances to understand the longer-term effects of noise on a population, and to inform management actions.

# A SEMIPARAMETRIC FRAMEWORK FOR RANK TESTS

Jan De Neve*[1], Olivier Thas[1,2], Jean–Pierre Ottoy[1]

[1]*Ghent University, Ghent, Belgium*
[2]*University of Wollongong, Wollongong, Australia*

*E-mail: *JanR.DeNeve@UGent.be*

We demonstrate how well known rank tests, such as the Wilcoxon–Mann–Whitney, Kruskal–Wallis, and Friedman tests, and many more, can be embedded in a statistical modelling methodology and how our approach can be used for constructing new rank tests for more complicated designs. In particular, rank tests for unbalanced and multi-factor designs, and rank tests that allow for correcting for continuous covariates are included. In addition to hypotheses testing, the method allows for the estimation of meaningful effect sizes, resulting in a better understanding of the data. Our method results from two particular parametrizations of probabilistic index models (Thas et al., 2012).

Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012) Probabilistic index models (with discussion). *Journal of the Royal Statistical Society - Series B*. 74:623–671.

# MODELLING THE SPATIAL OCCURRENCE OF UK BUTTERFLIES FROM OPPORTUNISTIC PRESENCE-ONLY ATLAS RECORDS

Emily B. Dennis[*1], Byron J.T. Morgan[1], Stephen N. Freeman[2], David B. Roy[2], Tom Brereton[3], Richard Fox[3]

[1]*University of Kent, Canterbury, UK*
[2]*Centre for Ecology & Hydrology, Wallingford, UK*
[3]*Butterfly Conservation, Dorset, UK*

[*]E-mail: *ed234@kent.ac.uk*

Occupancy models (MacKenzie *et al*. 2002) have provided an opportunity for the study of patterns in species occurrence, distribution and range dynamics. However, the formal presence-absence data they require are not always available for the species of interest and in some surveys only a subset of sites with known presence is recorded. This list may substantially underestimate the species' entire range. In this case, there is no clear optimal approach to estimating occupancy, though many have been proposed. Presence records of other "benchmark" species may, for example, be used to deduce absence records. Alternatively, a presence-only model for occupancy has recently been suggested, subject to certain assumptions such as random sampling and constant detection probability (Royle et al. 2012).

Here, we test these methods by applying them to data from the Butterflies for the New Millennium scheme (BNM), consisting of records for UK butterflies collected between 2000 and 2009 via opportunistic, unstructured sampling by the public. Estimation of the spatial distribution of UK butterfly species may allow for more accurate assessment of levels of change, such as contractions in response to degradation of habitats or expansions in response to climate warming.

Presence-only and presence-absence model approaches are compared for this opportunistic dataset, in addition to a presence-absence model applied to two more standardised datasets (the UK Butterfly Monitoring Scheme and Wider Countryside Butterfly Survey), where counts are treated as presence-absence records. Additionally, we investigate the extent to which combining information from multiple sources can improve estimates of occupancy. Our aim is to develop recommendations for occupancy modelling, both specific to this case study and for modelling ad hoc, presence-only records in general, in order to benefit the study of distributions and range dynamics.

References

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology,* **83,** 2248-2255.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution,* **3,** 545-554.

# IDENTIFICATION OF DOG BREED COMPOSITION

Orlando Döhring*[1], Neale Fretwell[1,2], David J. Balding[1]

[1]*University College London, London, UK*
[2]*Mars Veterinary, Leicestershire, UK*

*E-mail: *o.doehring@ucl.ac.uk*

Our goal is to identify the breed composition of mixed-breed dogs. Genetic test data has been provided by Mars Veterinary, using SNP markers.
Firstly, in our algorithm we compute a representation which characterizes the genetic variation within breeds. The genetic variation is given by haplotype frequencies which we compute for each chromosome over all breed sets composed of purebred training dogs. These haplotype frequencies are reconstructed from the genotype data using software PHASE. Clustering of either genotype data or haplotype frequencies using a variety of similarity measures provides a visual measure for ease of breed discrimination. These breed similarity plots show that typically some breeds are very distinct while breeds which have subpopulations, for example based on country (UK vs. USA), may show some overlap.
In our approach we use these haplotype frequencies to predict the breed composition of test dogs under the assumption of lineages formed by up to eight different great-grandparents. However, the search space over all possible lineages is huge. Enumeration and exploration of all possible lineages is prohibitive. To compute breed probabilities given a genotype test dog we approximately sample this search space with a Metropolis-Hastings algorithm. As proposal density we either uniformly sample new breeds for the lineage or we bias the Markov Chain, such that breeds in the lineage are more likely to be replaced by similar breeds. We will report results for a synthetic, mixed breed test dataset in two ways: Firstly, we will present ROC curve results which either show sensitivity versus positive predictive value or F1 score versus breed confidence cut-off. Secondly, we will compare deviation of true breed proportions with estimated breed proportions.

# ESTIMATING SPECIES RICHNESS FROM QUADRAT SAMPLING: A BAYESIAN APPROACH

Jérôme A. Dupuis

*MT, Laboratoire de Statistique et Probabilités, Université Toulouse III, France*

*E-mail: *dupuis@math.univ-toulouse.fr*

The species richness of a community of animals or plants - that is the number of species present within the community - is a basic measure of its biodiversity. Estimating the species richness (denoted by *S*) of a biological community located in some specified region often relies on quadrat sampling. The region is first divided in n quadrats, and inference on *S* is then based on the exploration of a sample. The difficulty is that a given species present in a sampled quadrat may not be detected by the experimenter. To deal with this difficulty different parametric approaches which separate assumptions related to occurences and detections via a hierarchical modeling have been recently developed; they will all be reviewed in detail. However, in spite of a rich modeling of the underlying biological processes, they all have some limitations. Indeed some assume that n is theoretically infinite, while n is not necessarily large in practice; moreover, species richness at a small scale cannot be derived from the one estimated at a large scale. Other require some prior information on S to be efficiently implemented, which prevents one from using them in little explored areas. We will present an approach which applies without limitation on the size of n, and which can be used in the presence, as well as in the absence, of prior information on S. We will also describe the MCMC algorithm perfected for obtaining the Bayesian estimate of S. Our approach will be illustrated by estimating the number of species of a birds community located in a forest. Finally, different directions for future research will be suggested.

## References

Dorazio, R. M. and Royle, J. A. (2005) Estimating size and composition of biological communities by modeling occurrence of species. Journal of the American Statistical Association **100**, 389-398.

Dupuis, J. A. and Goulard, M. (2011) Estimating species richness from quadrat sampling data: a general approach. Biometrics **67**, 1489-1497.

Tjorve, E. (2010). How to solve the SLOSS debate: lesson from species-diversity models. Journal of Theoretical Biology **264**, 604-612.

# BIADDITIVE MODELS, ALTERNATIVE ESTIMATION PROCEDURES AND BETTER BIPLOTS

Paul H.C. Eilers[1,2], Sabine K. Schnabel[1], Fred A. van Eeuwijk[1]

[1]*Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands*
[2] *Erasmus University Medical Center, Rotterdam, The Netherlands*

[*]E-mail: *p.eilers@erasmusmc.nl*

Biadditive models are a useful model class for investigating interactions in two-way tables. An area where biadditive models are popular is plant breeding and genetics, where sets of genotypes are evaluated across a range of environmental conditions, with the results being summarized in two-way tables of genotype by environment (GxE) means. For GxE tables, various biadditive models have been proposed, like the Finlay-Wilkinson model ($Y_{ij} = \mu + G_i + \beta_i E_j + \varepsilon_{ij}$), the additive main effects and multiplicative interactions model ($Y_{ij} = \mu + G_i + E_j + \sum_k \gamma_{ki}\delta_{kj} + \varepsilon_{ij}$), and the GGE or PCA model ($Y_{ij} = \mu + E_j + \sum_k \gamma_{ki}\delta_{kj} + \varepsilon_{ij}$).
For the estimation of parameters in biadditive models, least squares procedures are a common choice. However, inference in a least squares framework offers limited possibilities. We investigate Bayesian and penalized regression methods and discuss their possibilities.
For the interpretation of bilinear model fits, biplots, in which genotypes and environments are assigned coordinates on the basis of their bilinear parameters, are an important tool. Surprisingly, biplots often lack clarity and attractiveness. We propose a number of cosmetic improvements.
Genotypes in the centre of biplots are less interesting, in contrast to those further away. The convex hull has been used to identify the most extreme genotypes. So-called alpha-bags are a generalization; they aim at hull that contains a chosen percentage of the genotypes. They are hard to compute and visually not very attractive. As a quick and pleasing alternative we present expectile hulls, based on asymmetric least squares. The convex hull is useful for the identification of groups of environments, mega-environments, which elicit comparable adaptations in genotypes. We extend this idea to expectile hulls.

# MODELLING HISTORICAL WEATHER EFFECTS ON NATIONAL BIRD INDEX DATA

David A. Elston[*1], Mark J. Brewer[1], Blaise Martay[2], Alison Johnston[2] and James Pearce-Higgins[2]

[1]*Biomathematics and Statistics Scotland, Aberdeen, UK*
[2]*British Trust for Ornithology, Thetford, UK*

[*]E-mail: *d.elston@bioss.ac.uk*

The need to improve the understanding of relationships between species abundance and weather has been given added impetus by recent data on changing weather patterns and model-based projections of future climate change. We have recently been considering how best to model national bird index data as a function of national monthly data on temperature and rainfall. Of particular interest is the challange of identifying empirically how the effects of a weather variable vary between months within the years leading up to the time at which each annual national index value is collected. We will compare and contrast a selection of ways of imposing structure on the sequence of regression coefficients, one for each month, including a smoothing approach based on difference penalties and a parametric approach based on damped oscillations defined by a low-order Fourier series. Finally, we will discuss how to combine information across species to uncover sets of species with similar patterns of sensitivy to weather variables.

# INCORPORATING AGE- AND SEX-DEPENDENT REFERENCE RANGES IN JOINT LONGITUDINAL AND TIME-TO-EVENT MODELLING FOR BONE MARROW TRANSPLANTATION

Markus C Elze[*1], Annekathrin Heinze[3], Stephan Klöß[2], Oana Ciocarlie[4], Melanie Bremm[3], Ulrike Koehl[2], Jane L Hutton[1]

[1]*University of Warwick, Coventry, United Kingdom*
[2]*Hannover Medical School, Hannover, Germany*
[3]*Goethe-University Frankfurt, Frankfurt am Main, Germany*
[4] *Victor Babeş University of Medicine and Pharmacy, Timisoara, Romania*

[*]E-mail: *m.elze@warwick.ac.uk*

The incorporation of time-varying data in survival models is a common objective in areas where longitudinal measurements are collected at arbitrary time points, such as clinical trials or the social sciences. Joint modelling of longitudinal measurements and time-to-event data is a natural solution to this problem, but the amount of available data may limit the use of joint models. Here, we show that transforming the longitudinal data using additional information from external sources may increase the amount of information gained from the data.

'Bone marrow transplantation' is a potentially curative treatment option for different hematologic disorders, such as severe leukaemia. However, it is still associated with high mortality rates due to complications after transplantation. Early identification of high-risk patients is crucial for successful intervention. Thus, predictive models are needed to assist clinical decision making. The development of longitudinal immune measurements is relevant for complication prediction and should be considered in modelling. As studies are often faced with limited patient numbers, assessing the immune recovery may be difficult. Here, we demonstrate how the use of reference data from healthy persons may assist the model fitting.

Age- and sex-dependent reference ranges are created from 100 healthy children for several immune subpopulations using the LMS method (Cole *et al*., 1992). These are then employed to assess the immune recovery for 67 paediatric patients who underwent bone marrow transplantation. The performances of joint models with and without the use of reference ranges are compared. The use of these models in clinical practice is discussed.

# MODELLING SURVIVAL FOLLOWING 'BONE MARROW TRANSPLANTATION' USING LONGITUDINAL IMMUNE MEASUREMENTS AT ARBITRARY TIME POINTS

Markus C Elze[*1], Stephan Klöß[2], Annekathrin Heinze[3], Oana Ciocarlie[4], Melanie Bremm[3], Ulrike Koehl[2], Jane L Hutton[1]

[1]*University of Warwick, Coventry, United Kingdom*
[2]*Hannover Medical School, Hannover, Germany*
[3]*Goethe-University Frankfurt, Frankfurt am Main, Germany*
[4] *Victor Babeş University of Medicine and Pharmacy, Timisoara, Romania*

[*]E-mail: *m.elze@warwick.ac.uk*

Severe leukaemias in paediatric patients may be treated by 'bone marrow transplantation'. However, complications after transplantation may arise and early identification of high-risk patients is crucial for successful intervention. Thus, predictive models are needed for clinical decision making. These models need to incorporate longitudinal immune measurements taken at arbitrary time points. Also, measurement time points and frequency cannot be assumed to be independent of the survival process.

Joint modelling of longitudinal and time-to-event data is a natural approach to data of this nature. Implementations of joint modelling techniques are readily available. However, no consensus has been reached yet on how to best assess goodness of fit and predictive ability of such models.

For use in clinical practice simpler models may be preferred. A simple comparison of measurements at a single time point is sometimes found in biological publications to group patients into 'high-risk' and 'low-risk' categories. However, bias may be associated with such an approach. Bias can be reduced by using appropriate summary measurements and accelerated failure time models can be fitted. This has the advantage that many measures of the prognostic ability of survival models are readily available, e.g. explained variation (Stare et al.). These simpler models may be used to specify a joint model and joint modelling may in turn inform the simpler models.

A possible extension of joint modelling is the inclusion of a proportion cured component. Also, the incorporation of the measurement frequency in the longitudinal model may inform on the risk assessment by the clinicians.

# COVARIANCE MODELLING IN LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION

Daniel Farewell*[1], Chao Huang[1]

[1]*Cochrane Institute of Primary Care and Public Health, School of Medicine, Cardiff University, Cardiff, UK*

*E-mail: *farewelld@cf.ac.uk*

When using generalized estimating equations to model longitudinal data, both inconsistency (due to informative observation) and inefficiency (due to misspecified working covariances) are often of concern. We describe a class of generalized inverses of singular working correlation matrices that allows flexible modelling of covariance within a subject's responses while offering robustness to certain kinds of informative observation. We demonstrate how this class corresponds to dynamic models on the increments in the longitudinal responses, and illustrate its application to a randomized trial of quetiapine in the treatment of delirium.

# MODELLING PRION DYNAMICS IN BUDDING YEAST CELLS

Vasileios Giagos*[1], Martin Ridout[1], Byron Morgan[1]
Wesley Naeimi[2], Tobias von der Haar[2], Mick Tuite[2]

[1]*SMSAS, University of Kent, UK*
[2] *School of Biosciences, University of Kent, UK*

*E-mail: *v.giagos@kent.ac.uk*

Prions are misfolded proteins which have the ability to self-replicate by a nucleated polymerization process and play the role of infectious agents in fatal neuron-degenerative diseases in mammals, e.g. Creutzfeldt-Jakob disease.
Contrary to mammalian prions, yeast prions can be beneficial to the cell. For instance, Saccharomyces cerevisiae cells with the prion form of the Sup35 protein ([PSI+]) show resistance to toxic compounds. Sup35 protein is a translation termination factor and its prion form is inherited through cell division.
Modelling the development of yeast proteins in a growing population of yeast cells is a complex process, due to the need to describe both polymerisation within cells and the partition of prions on cell division. In this poster we shall outline some of the stochastic models which have been developed to describe the [PSI+] biology in budding yeast cells. We shall discuss the associated computational problems and the proposed solutions using moment-closure approximations. Finally, we shall present the statistical challenges which arise when we fit models to the experimental data.

# AUTOMATIC PHENOTYPING: ESTIMATING FRUIT NUMBERS BY CLUSTERING REPEATED, INCOMPLETE OBSERVATIONS

Chris Glasbey*[1], Yu Song[1], Graham Horgan[1], Gerie van der Heijden[2], Gerrit Polder[2]

[1]*Biomathematics & Statistics Scotland, Edinburgh, UK*
[2]*Biometris, Wageningen, The Netherlands*

*E-mail: *chris@bioss.ac.uk*

In our post-genomic world, where we are deluged with genetic information, the bottleneck to scientific progress is often phenotyping, i.e. measuring the observable characteristics of plants and animals. Image analysis is one way to automate phenotyping: in an EU-FP7 project, SPICY (Smart tools for Prediction and Improvement of Crop Yield), we have developed methodology for automatically taking measurements from pepper plants, such as counting fruit numbers. Cameras mounted on a trolley took photographs at regular intervals as they were moved down aisles in a greenhouse, for plants of a range of genotypes in each of 400 experimental plots. Fruits were then located in each photograph by image analysis, except where they were occluded by leaves. We wish to estimate the total number of fruits by combining data from series of photos of each plot, to see whether we can reproduce the true numbers obtained at considerable cost by human counting.

For each plot let $(x_{ij}, y_{ij})$ denote the column & row coordinates of the centre of the $j$th fruit found in the $i$th image. Suppose there are $K$ fruits observed at least once in the plot, and let $(\alpha_k, \beta_k)$ denote the true column & row coordinates of fruit $k$ in the zeroth image, and $\gamma_k$ the true shift in column coordinate between consecutive images (which is inversely proportional to the distance from the fruit to the camera). We propose as our observation model:

$$\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} \sim N\left( \begin{bmatrix} \alpha_{k(ij)} + i\gamma_{k(ij)} \\ \beta_{k(ij)} \end{bmatrix}, \begin{bmatrix} \sigma_x^2, & 0 \\ 0, & \sigma_y^2 \end{bmatrix} \right) \qquad i = 0, \ldots, (I-1), \;\; j = 1, \ldots, J_i$$

where $k(ij)$ denotes the correct fruit label of observation indexed $(i, j)$, and $(\sigma_x^2, \sigma_y^2)$ are the variances of normally distributed observation errors.

MCMC is problematic as an estimation procedure because the dataset has 40,000 observations, so we have instead developed a simpler, faster method. We first estimate $\sigma_y^2$ by fitting a mixture distribution to differences in $y$ between pairs of observations. Similarly we estimate $\sigma_x^2$ by considering triplets of observations. Finally, for each plot we apply a 95% significance threshold rule to identify non-overlapping clusters of observations of a single fruit, starting with the largest possible cluster size ($I$) and progressively reducing until we are left with singletons. The number of clusters is our estimate of $K$.

Simulations show the method to give an unbiased estimate of fruit numbers, and a 74% correlation with true counts for the 400 plots, which is strong enough for use in plant breeding.

# ON THE MODERATED *T*-TEST AND ITS MODERATED *P*-VALUES

Jelle Goeman*[1], Livio Finos[2], Erik van Zwet[1]

[1]*Leiden University Medical Center, The Netherlands*
[2]*Univerity of Padua, Italy*

*E-mail: *j.j.goeman@lumc.nl*

Moderated *t*-tests such as limma are very popular for the analysis of high-dimensional genomics data. Such tests improve the variance estimate of each individual probe by using additional information from the other probes using empirical Bayes arguments. By "borrowing strength" in this way the moderated *t*-test is better able to separate true from false hypotheses than the classical *t*-test, especially in data in which the number of probes is very large and the sample size is very small. But there is a price to pay for using the moderated *t*-test. We argue that *p*-values resulting from a moderated *t*-test lose an important property that classical *p*-values do have. Whereas *p*-values from a classical *t*-test are uniformly distributed if the null hypothesis is true, the *p*-values from a moderated *t*-test are uniform only in some average sense. The lack of this uniformity property is a serious problem if we want to adjust these *p*-values for multiple testing. We give practical examples of problematic situations arising when multiple testing methods are combined with the moderated *t*-test. Most notably, we show that the frequently-used combination of false discovery rate control and a lower bound on the estimated effect size ("fold change") can result in excessive error rates.

# CAR AND P-SPLINE MODELS FOR SHORT-TERM CANCER MORTALITY COUNT PREDICTIONS

Goicoa, Tomás.*[1,2], Etxeberria, Jaione. [1,3], Ugarte, M. Dolores[1], and Militino, Ana.F[1]

[1]*Public University of Navarre, Pamplona, Spain*
[2]*Research Network on Health Services in Chronic Diseases (REDISSEC), Spain*
[3]*Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain*

*E-mail: *tomas.goicoa@unavarra.es*

The cost of cancer (including diagnosis, treatment, research, and lost-person-hours) is huge. The current delay for reporting actual cancer mortality data is about 3 years in most countries and about 4 years for cancer incidence figures. Hence, statistical techniques providing mortality or incidence counts predictions in different health areas are very useful to estimate cancer burden.

Prediction is the next step after smoothing temporal trends of risks in different areas, and it is sensible to extend spatio-temporal models for forecasting purposes. Major Health Agencies predict mortality and incidence for different cancer sites using statistical methods at national level, but in countries like Spain, where health policies depends on the regional Governments, predictions are required for small areas. Consequently, models including a spatial component are required.

In this work, CAR, P-splines and a combination of both are used to provide short-term cancer mortality counts for different health areas or small regions. In particular, we consider models with a CAR structure for space and a random walk structure for time, P-splines for a common time trend and area specific Psplines, and models combining a CAR structure for space and P-splines for time. These models include alternative space-time interactions.

The techniques are illustrated using prostate cancer mortality data in fifty Spanish provinces for the period 1975-2008. Predictions are provided for 2009-2011. These data are also used to assess the predictive performance of the models. The results show that CAR models for space with a random walk of order two for time and the corresponding interactions seem to be close competitors of P-spline models. However, models combining CAR structures for space with P-splines for time including space-time interactions do not seem to improve mortality risks forecasting.

# INCLUSION OF SEGREGATION INFORMATION IN THE GENOTYPING OF TETRAPLOID SPECIES

Gerrit Gort[*1], Roeland E. Voorrips[1], Fred A. van Eeuwijk[1]

[1]*Wageningen University, Wageningen, Netherlands*

[*]E-mail: *gerrit.gort@wur.nl*

The genotype calling of Single Nucleotide Polymorphisms (SNPs) in polyploids, like potato, rose, or sugarcane, is a topic that currently receives quite some attention in the biometric world. In a recent paper we studied the SNP genotype calling of tetraploid potatoes, based on normal mixture models (Voorrips, Gort and Vosman, 2011). In this paper (Golden Gate) SNP data are analyzed, originating from an association panel of 288 diploid and 224 tetraploid potato cultivars. For each SNP a mixture of normal distributions for transformed intensity ratio's is assumed. In the tetraploid case a mixture of five components is needed: from nulliplex (0 times allele A) to tetraplex (4 times allele A). The methods in this paper are made available in the R-package fitTetra, which is available from CRAN.
If SNP's are scored on tetraploid parents and their offspring, the genotyping of the offspring may be facilitated by knowledge of the parental genotype. With known parental genotypes, mixing probabilities following from the tetraploid crossing scheme may be taken as constant in the mixture modelling. If parental genotypes are unknown, the best combination of estimated parental genotypes and corresponding offspring genotypes may be selected. A check on segregation distortion will report SNP's which do not segregate according to (possible) parental genotypes. We further describe how a combination of data from an association panel and from tetraploid crossings can be analyzed simultaneously. These methods are made available in a new version of the fitTetra package.
As an example SNP's from a 20k Illumina Infinium assay on 240 offspring with parents in potato are analyzed.

**Literature**

Voorrips, R.E., G. Gort, B. Vosman (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, 12:172. DOI: 10.1186/1471-2105-12-172

# OUTLIER DETECTION IN POISSON REGRESSION MODELS VIA OVERDISPERSION

Freedom Gumedze, Tinashe Catora

*Department of Statistical Sciences,*
*University of Cape Town, Rondebosch, South Africa 7701*

*E-mail: *freedom.gumedze@uct.ac.za*

This paper introduces a variance shift outlier model (VSOM) for the detection of potential outliers in count data. The model considers outliers as counts with inflated variance and uses random effects to model the overdispersion associated with a single count. A VSOM is formulated as a negative binomial model by assuming a Poisson distribution for all observations with a gamma random effect for a given observation. The status of a given observation as an outlier is indicated by the size of the associated shift in variance for that observation. The model is then extended to longitudinal count data for the detection of individual, and or, groups of observations. In this extension of the model both oversidpersion and correlation between observations, due to clustering, are modelled using conjugate random effects. We illustrate the methodology using two real datasets taken from literature.

# PARAMETER INFERENCE IN A STOCHASTIC REACTION NETWORK WITH TRANSCRIPTIONAL SWITCHES

Kirsty L. Hey*[1], Bärbel Finkenstädt[1], David A. Rand[1]

[1]*University of Warwick, Coventry, UK*

*E-mail: *K.L.Hey@warwick.ac.uk*

In order to estimate the parameters of reaction networks one can distinguish three sources of 'noise', namely the intrinsic variability arising from the biochemical processes, the intercellular extrinsic noise and the error associated with the measurement process. Unfortunately, the likelihood is intractable even for simple reaction network models. We consider two different approximations to the likelihood that enable us to perform inference within a Bayesian framework, namely the linear noise approximation (LNA) and a newly derived approximation that arises from a simplification of the underlying model specific reaction network. We investigate how these approximations deal with the intrinsic stochasticity of the system and also, the often neglected, measurement process and compare their performance for parameter estimation. It will be shown that a particle filter implementation for the model specific approximation will increase the computational efficiency compared to the theoretically simple LNA likelihood.

Coupled with the above methodology, will be a random switch model for gene transcription. The function of transcription will be modelled by a random step function relating to bursts in gene activity, but not restricted to the traditional on-off binary structure. We show how reversible jump Markov chain Monte Carlo can be used to identify this function within the stochastic reaction framework. The resulting methodology will be used to infer the switching regimes of the hPRL (human prolactin) gene within single cells of mammalian pituitary tissue.

# FINITE MIXTURE MODEL CLUSTERING OF SNP DATA

John Hinde*[1], Norma Coffey[2], Augusto Franco Garcias[3]

[1]*National University of Ireland Galway, Ireland*
[2]*University College Dublin, Ireland*
[3]*ESALQ/USP, Piracicaba, Brazil*

*E-mail: *john.hinde@niugalway.ie*

Sugarcane is polypoid and it is important to develop methods that identify the many different alleles and associated genotypes. Single nucleotide polymorphisms (SNPs) can give an indication as to the number of allele haplotypes present for a gene and such information could have implications in sugarcane breeding since high yield potential may be due to the presence of and/or different number of copies of, a specific allele(s) present at a gene locus. Clustering these data provides a means of identifying different genotypes and therefore it is necessary to develop a technique that can determine the number of clusters present, determine the angles between the clusters to identify different genotypes, and provide a probabilistic clustering to identify points that have high probability of belonging to a particular cluster (have a particular genotype) and those that are regarded as an unclear genotype. Standard clustering methods, such as mclust do not perform well because of the radial nature of the data, although the performance can be improved by moving to polar coordinates, as previously proposed in the literature. Here we propose the use of finite mixtures of orthogonal regression lines to cluster the data. We implement this technique in R, show its usefulness in clustering these data and compare the performance with the other methods.

# A MIXTURE MODEL FOR THE ANALYSIS OF DATA DERIVED FROM RECORD LINKAGE

M.H.P Hof[1] and A.H. Zwinderman[1]

[1]*Academic Medical Center - University of Amsterdam, Amsterdam, the Netherlands*

*E-mail: *m.h.hof@amc.uva.nl*

In record linkage problems, unique identifiers per record are often not available and a linkage strategy depending on partially identifying variables $G$ is necessary. A risk of relying on these variables is that records from both datasets are wrongly linked to each other. Because current regression methods for linked data are based on strong assumptions, we propose a less restricted model.

Because we do not know which record $i$ from data source **A** belongs to record $j$ from data source **B**, we treat this indicator $d_{ij}$ as missing. Our goal is to estimate the relation between outcome $y$ (located in **B**) and covariates $X$ (located in **A**) using the records that belong together (i.e. $d_{ij} = 1$). To obtain this estimate, we have to optimize the likelihood of the complete data which is

$$\prod_{i=1}^{n}\prod_{j=1}^{m} L(y_j|X_i,d_{ij}=1)L(X_i|d_{ij}=1)L(G_{ij}|d_{ij}=1)L(d_{ij}=1) + L(y_j|d_{ij}=0)L(X_i|d_{ij}=0)L(G|d_{ij}=0)L(d_{ij}=0)$$

where $L()$ is a (parametric) conditional density function, and $n$ and $m$ are the number of records in respectively **A** and **B**. The mixture model is optimized with an EM-type algorithm. We simulated different scenarios to illustrate the performance of this estimator. In addition, we compared its performance with our previous proposed, more restricted, WLS estimator[1]. Most importantly, the WLS estimator assumes that $L(X_i|d_{ij}=1) = L(X_i|d_{ij}=0)$ which is only satisfied in particular situations.

The simulations showed that the WLS estimator and our newly proposed mixture model have similar performance when the WLS assumptions hold. However, when the WLS assumptions did not hold, the WLS estimator was highly biased to the mean. Our new proposed method retained the same accuracy which was reflected by the bias, MSE, and coverage of the 95% confidence interval.

In almost all situations, our proposed mixture model can be used for wrongly matched records in the analysis of a dataset derived from a linking process with partially identifying variables.

[1]Hof and Zwinderman, Statistics in Medicine, 31:4231–4242, 2012

# KEEPING WEIGHT ON TRACK –A NEUTRAL MODEL OF HUMAN BODY MASS REGULATION

Graham W Horgan[*1]
[1]*Biomathematics and Statistics Scotland, Aberdeen, UK*

[*]E-mail: *g.horgan@bioss.ac.uk*

While there are many mechanisms that may be involved in the regulation of body mass in humans and other animals, it is not so clear how much regulation is needed beyond the negative feedback effect of body mass itself. This occurs because increased body mass requires more energy, which promotes negative energy balance and so weight loss. The opposite occurs with reduced body mass. Here we model weight change as a stochastic process, and show that it behaves approximately as an autoregressive process. This leads to a model

$$(W_{i+1} - W_0) = (1 - C\mu_P\beta)(W_i - W_0) + C(\varepsilon_I - \varepsilon_E)$$

where $W_i$ denotes weight at time point $i$, C is the energy cost of weight gain and loss, $\mu_P$ is the mean physical activity level, $\beta$ is the coefficient linking weight and resting metabolic rate and $\varepsilon_I$ and $\varepsilon_E$ are fluctuations in energy intake and expenditure.

Using published estimates of the energy cost of weight gain, the effect of weight on resting metabolic rate and the daily variation in intake and activity, we show that fluctuations in weight will be small, about 2% of mean weight. The effect of excess intake is also examined, and the assumptions and limitations of the model are discussed.

# MODIFICATIONS OF THE DUNNETT MULTIPLE COMPARISON PROCEDURE: WITHOUT PRE-TEST, INCLUDED INTO ORDER-RESTICTED WILLIAMS TEST

Ludwig A. Hothorn[*1]
[1]*Leibniz University, Hannover, Germany*

[*]E-mail: *hothorn@biostat.uni-hannover.de*

Dunnett's procedure (Dunnett, 1955) belongs to the most cited (bio)statistical approaches. Several treatment groups are compared to a control or placebo in a randomized one-way layout, where multiplicity-adjusted p-values and/or simultaneous confidence intervals are available for both one- or two-sided comparisons.

A common (mis)-use is pre-testing by an ANOVA F-test, i.e. performing the Dunnett-test only after a significant global test. The power of the quadratic form F-test is analytically compared to the linear forms multiple contrast tests, particularly all-pairs, comparison-to-grand mean and many-to-one alternatives under both least favourable configurations and particular alternatives. These power comparisons indicate to avoid pre-testing at all.

The US-National Toxicology Program recommends the analysis of either Dunnett- or Williams (Williams, 1971) test for designs including a zero-dose control and several dose groups. However, the guidance does not recommend under which condition a test without order restriction (Dunnett) and with order restriction (Williams) should be performed. And, such a recommendation seems to be difficult either a-priori defined or data-dependent. Therefore, the simultaneous use of both tests will be recommended. The principle of multiple contrast tests is used. The tolerable loss of power, but the increased interpretability of many interesting individual alternatives is presented. The basic principle behind is " correlation matters".

Real data examples are presented and the data are analysed using the R packages "multcomp" and "mratios".

# STATISTICAL METHODS FOR THE ANALYSIS OF HUMAN MICRO BIOME DATA

Jeanine J. Houwing-Duistermaat

*Dept of Medical Statistics and Bioinformatics, LUMC, Leiden, The Netherlands.*

*E-mail: *j.j.houwing@lumc.nl*

Statistical analysis of new data types in complex models is typically challenging. We will consider the statistical modeling of Human Micro Biome data for 62 subjects from Indonesia. A part of the subjects were infected with helminth. The data are from a randomized clinical trial were half of the subjects received anti-helminth treatment and the other half placebo. We have data on baseline as well as data after treatment. For 51 subjects we have complete data. The Human Micro Biome data consist of 3000 peaks, where each peak corresponds to a bacterium. The aim of the study is to identify differences in bacteria profiles between treated and non-treated and between infected and healthy subjects. The standard analysis of this type of data is to compute and compare diversity parameters and to perform dimension reduction by PCA analysis and test the first components for association with covariates. However such an analysis is not correct for missing and correlated data. Other challenges of these data are the skewness of the data due to inflation of zeros (left side of the distribution) and bias due to overflow for large peaks. In this presentation we will introduce the type of data. For each peak, we will consider standard linear mixed models as well as a mixture model for repeated measurements (1).

(1) Mahmud et al. BMC Medical Research Methodology. 2010. **10**:55

# MODELLING THE RELATIONSHIP BETWEEN CARDIOVASCULAR EVENTS AND VARIABILITY/INSTABILITY OF BLOOD PRESSURES VIA DOUBLE INTEGRATED SEMIPARAMETRIC GENERALIZED LINEAR MODEL

Chao Huang[*1], Jianxin Pan[2]

[1]*Caridff University, Cardiff, UK*
[2]*University of Manchester, Manchester, UK*
[*]E-mail: *huangc12@cf.ac.uk*

Systolic and diastolic blood pressure are the most essential cardiovascular risk factors. And high blood pressure (hypertension) has been regarded as the common-sense cause of cardiovascular events including strokes and myocardial infarction (MI). Rothwell (2010) proposed that variability and instability of blood pressure are also associated to cardiovascular events. This new finding could bring innovative cardiovascular disease treatment strategy and potential cost-benefit improvement. However, little rigorous statistical evidence has been collected to consolidate it. To this end, we propose a double integrated semiparametric generalized linear model to assess the impact of variability/instability of blood pressure. The variability/instability, in addition to the mean blood pressure, is linked into the modelling of cardiovascular events, in which all the historic blood pressure information are combined in the form of weighted integration. The orthogonal B-spline technique is also adopted for the nonparametric modelling. The simulation results show that our proposed method could identify the potential relationships between risk events and historical and current blood pressures. This proposed method could also be applied to any other area involving the variability/instability and historic information in modelling.

# PARAMETER REDUNDANCY: IDENTIFIABILITY ISSUES IN OCCUPANCY MODELS

Ben A. Hubbard*, Diana J. Cole, Byron J.T. Morgan

*The University of Kent, Canterbury, UK*

*E-mail: *bah21@kent.ac.uk*

To be able to fit a parametric model successfully using maximum likelihood, we expect that all the parameters can be estimated. If it is not possible to estimate all the parameters, the model is known as parameter redundant or non-identifiable and then the model can then be rewritten in terms of a smaller set of parameters. Parameter redundancy can be detected be forming a derivative matrix and calculating its rank; the model is parameter redundant if the rank is not equal to the number of parameters in the model. When a model is parameter redundant, not all the parameters in the model can be explicitly estimated, which can be irrespective of the data collected.

Occupancy models are used to estimate site occupancy for a particular species, over a certain length of time, using a number of different detection sites. MacKenzie et al. (2003) suggests models which are capable of estimating occupancy, detection and colonization/extinction probabilities. These models are also flexible enough to allow the age-, time- and site-dependency of certain parameters to be relaxed when required. We first consider intrinsic parameter redundancy, which looks at identifiability due to the inherent structure of the model. We develop simple to use methods involving a reparameterisation of the model that uses symbolic algebra to generate results. These intrinsic results provide the best case scenario in terms of parameter redundancy results, but parameter redundancy can also be caused by a particular set of data. This is known as extrinsic parameter redundancy. We show what happens in practice for sparse data sets and how this affects parameter redundancy for a range of models when parameter dependencies are relaxed. We illustrate these extrinsic parameter redundancy methods using data on amphibian breeding in Yellowstone and Grand Teton National Parks, and on House Finches in North America.

# ESTIMATING CONDITIONAL SURVIVAL AFTER SPINAL CORD INJURY: AN APPROACH TO META-ANALYSIS OF OBSERVATIONAL STUDIES

Jane L. Hutton*[1]

[1]*University of Warwick, Coventry, UK*

*E-mail: *J.L.Hutton@warwick.ac.uk*

The impact of spinal cord injury on life expectancy is of interest to those with such injuries and their families, and to insurers who have to meet the costs of care. Patients and insurers want an estimate of life expectancy, conditional on age, sex, country, severity of injury and time since injury.
Ten studies published after 1980 gave results as either standardised mortality rates (SMRs) or some more direct statistic of a survival function. Mortality is worse for those with injuries higher up the spinal cord, and greater age at injury. However, estimates of the effects of time since injury, calendar year of injury, sex and country vary substantially. The general population mortality rates for the USA, and UK and Australia, three of the countries for which data are reported, also differ.
The challenges of meta-analysis of comparisons of time-to-event outcomes from randomised trials have been studied, and approaches such as assuming underlying Weibull distributions have been useful in providing estimates based on specific assumptions. Estimation of mortality rates from observational studies in which the focus is not on a comparison of treatments allocated at random, but on severity of injury measured on various different scales offers additional adventures in advocating and assessing assumptions.
The natural assumption to use when dealing with SMRs is that the numbers of deaths are distributed according to Poisson processes, with rates dependent on covariates including age, sex, country and severity of injury. Extra-Poisson variation is conveniently captured through a Gamma random effect. As the distribution of a random variable which is exponentially distributed conditional on a Gamma random effect is a Lomax distribution, this distribution, rather than the Weibull distribution, is used in modeling outcomes such as mean survival times or hazard ratios.
.

# FITTING COMPLEX MODELS IN INLA – DEVELOPMENTS AND EXTENSIONS

Janine B Illian

*University of St Andrews, St Andrews, UK*

*E-mail: *janine@mcs.st-and.ac.uk*

Integrated nested Laplace approximation INLA) may be used to fit a large class of (complex) statistical models. While MCMC methods use stochastic simulations for estimation, integrated nested Laplace approximation (INLA) is based on deterministic approximations where there are no convergence issues. INLA is a very accurate and computationally superior alternative to MCMC and may be used to fit a large class of models, latent Gaussian models.

Since INLA is fast, complex modelling has become greatly facilitated and has also become more accessible to non-specialists. In addition, due to the fact that the fitting approach is embedded in a large and general class of statistical models, very general types of models may be considered. This allows us a lot more flexibility in the choice of model than previously – and hence the models to capture interesting aspects of the data and consequently the system they are relevant for. In the context of spatial statistics, for example, we can now fit models to spatial point patterns of high dimensionality, replicated point patterns, hierarchically marked point patterns etc. In many cases, analysing these data sets with MCMC approaches would be very cumbersome and computationally prohibitive.

The INLA-methodology has been implemented in `C`, and the associated numerical calculations and algorithms rely on an efficient implementation of numerical procedures for Gaussian Markov random fields (GMRF), in particular the algorithms in the `C`-library `GMRFLib`. However, most users do not need to worry about this, as the INLA-methodology has been made accessible through a user-friendly `R`-library, `R-INLA`, described and available for download at `www.r-inla.org`. Specifying and fitting models using `R-INLA` is just as easy as applying standard routines in `R`, for example fitting generalised linear models, and it also provides great flexibility with regard to the models that may be fitted.

In order to illustrate INLA's versatility I will discuss a range of spatial and non-spatial examples and present a number of recent developments. This concerns generalisations of the methodology as well as new functionality within the `R-INLA` library.

# SELECTING THE STATES THAT DEFINE MARKOV HEALTH ECONOMIC MODELS

Howard Thom, Christopher Jackson, Linda Sharples

*MRC Biostatistics Unit, Cambridge, United Kingdom*

*E-mail: *chris.jackson@mrc-bsu.cam.ac.uk*

Health economic evaluations are typically based on Markov multi-state models. Patients in these models move between discrete states of health and treatment, accumulating costs and quality-adjusted years of life. This leads to estimates of the long-term benefits and costs of different interventions. When designing these models, the states should include any event or health condition which may be affected by the intervention. However there are not always sufficient data to inform the transition probabilities between all potential states, or the cost and quality of life associated with each state. Similar states could then be merged, for example, adjacent states of disease severity. It may then be uncertain whether the larger or smaller model gives better estimates of the outcomes of interest, such as expected survival. Since we are comparing estimators using aggregated data with estimators using disaggregated data, this is not a standard statistical model comparison. The corresponding likelihoods will be on different scales, since the disaggregated dataset contains more information. Instead, we adapt a recently-developed modification of Akaike's criterion, and a cross-validatory criterion, to compare the predictive ability of both models on the information which they share. We apply the methods to a model for comparing the cost-effectiveness of different diagnostic tests for coronary artery disease.

# JOINT MODELING OF LONGITUDINAL AND TIME-TO-EVENT DATA WITH APPLICATION TO THE PREDICTION OF PROSTATE CANCER RECURRENCE

Mbéry Séne[*1,2], Hélène Jacqmin-Gadda[1,2], Cécile Proust-Lima [1,2]

[1] *INSERM, ISPED, F-33000 Bordeaux, France*
[2]*Univ. Bordeaux, ISPED, F-33000 Bordeaux, France*

[*]E-mail: Mbery.Sene@isped.u-bordeaux2.fr

In the last decade, the joint modelling has rapidly developed in the field of biostatistics and medical research to study simultaneously a longitudinal marker and a correlated time-to-event. Among joint models, the shared random-effects models, that define a mixed model for the longitudinal marker and a survival model for the time-to-event in which characteristics of the mixed model are included as covariates, received the main interest. Indeed, they extend naturally the survival model with time-dependent covariates and offer a flexible framework to explore the link between a longitudinal biomarker and a risk of event.

The objective of this work is to present the shared random-effects models methodology, and illustrate its implementation and evaluation in practice through a real example from the study of prostate cancer progression after a radiation therapy. In particular, different specifications of the dependency between the longitudinal biomarker, the prostate specific antigen (PSA), and the risk of clinical recurrence are investigated to better understand the link between these two processes.

The different joint models are compared in terms of goodness-of-fit and assessment of the joint model assumptions but also in terms of predictive accuracy using the expected prognostic cross-entropy. Indeed, in addition to better understand the link between the PSA dynamics and the risk of clinical recurrence, the perspective in prostate cancer studies is to provide dynamic prognostic tools of clinical recurrence based on the biomarker history.

# NEW ANALYTICAL METHODS FOR ESTIMATING KEY ECOLOGICAL PARAMETERS FROM CAMERA TRAP DATA.

Natoya O.A.S. Jourdain*[1], Diana J. Cole[1], Martin S. Ridout[1], Marcus Rowcliffe [2]

[1] *University of Kent, Canterbury, UK*
[2] *Institute of Zoology, Living Conservation, London, UK*

*E-mail: *nj64@kent.ac.uk*

This project expands on the Random Encounter Model (REM) developed by *Rowcliffe et al. (2008)*. The REM eliminates the requirement of individual recognition of animals by modelling the underlying process between animals and cameras. The focus is to develop a more unified statistical modelling framework to estimate robust behavioural parameters of animals from camera trap images. Maximum likelihood estimation will be adopted as it gives a unified approach to estimation, allowing robust model-based inference of animal abundance by including covariates such as climate and habitat type. The derivation of the components; trap rate, speed, activity, detection distance and detection angle, in abundance estimation currently requires five independent steps under the REM, hence, the need for a unified single analytical framework, which will be developed to incorporate all five components. As multiple sources of variance will be included in the analysis, the overall precision will be investigated to give guidance on the minimum and optimal survey effort required to achieve a given power in detecting density differences in time or space. This will be done by exploring how variance partitions between the different sources empirically, and model how sample sizes in the different sources affect overall precision. An application to an extensive data set from Panama will illustrate the methods.

Rowcliffe, M.J. *et al.* (2008), *Estimating animal density using camera trap methods without the need for individual recognition*, Journal of Applied Ecology **45**, 1228-1236.

# JOINT SPATIO-TEMPORAL MODELING OF MYCOBACTERIUM BOVIS INFECTIONS IN BADGERS AND CATTLE

Gabrielle E. Kelly*

*University College Dublin, Dublin 4, Ireland*

*E-mail: *gabrielle.kelly@ucd.ie*

Introduction

We outline how linear geostatistical models (LGMs) may be used to assess spatio-temporal associations in bovine TB infection across two species and present and interpret the results of such an analysis on data from the Irish Four Area Project.

Methods

In Ireland and the UK bovine bTB infects cattle and wildlife badgers (*Meles meles Linnaeus*) and badgers contribute to the spread of the disease in cattle. Cattle herd and badger sett bTB incidence data are drawn from a large scale field trial, the Four Area Project (FAP), a formal badger removal project undertaken in four counties in Ireland from September 1997 to August 2002, to assess the effect of badger culling on the incidence of bTB. Sequences of LGMs with different spatial correlation structures are fitted to these data and statistical methodology is outlined whereby hypotheses related to spatial correlation structure is tested.

Results

Using combined data, association was found between the spatial distribution of the disease in cattle and badgers in two of three areas. This is evidence for cross-infection between the species. Separately, spatial association of bTB in badger setts varies over time, between areas and with direction within an area. Similar results were found for cattle herds.

Discussion

The results on cross-infection are unique and important in informing control measures. The spatial methodology used to establish them is widely applicable. The limited association may be due to irregularity of sett territories, fragmentation of farms, TB-test insensitivity, temporal lags associated with transmission or non-spatial transmission or because the relative contribution of infection of cattle herds by badgers is small as was seen in the FAP and Randomized Badger Culling Trial in GB. Conclusions from models fitted separately to badgers and to cattle herds are in agreement with previous analyses in the literature.

# ESTIMATING HIDDEN POPULATION SIZES USING MULTI-LIST DATA IN THE PRESENCE OF OBSERVED NON-TARGET INDIVIDUALS

Antony M. Overstall[1], Ruth King[*1], Sheila M. Bird[2,4], Sharon J. Hutchinson[3,4] and Gordon Hay[5]

[1] *University of St Andrews, St Andrews, UK*
[2] *Medical Research Council, Biostatistics Unit, Cambridge, UK*
[3] *Health Protection Scotland, Glasgow, UK*
[4] *University of Strathclyde, Glasgow, UK*
[5] *Liverpool John Moores University, Liverpool, UK*

[*]E-mail: *ruth@mcs.st-and.ac.uk*

Estimating the size of hidden or difficult to reach populations is often of interest for economic, sociological or public health reasons. In order to estimate such populations, administrative data lists are often collated to form multi-list cross-counts and displayed in the form of an incomplete contingency table. Log-linear models are typically fitted to such data to obtain an estimate of the total population size by estimating the number of individuals not observed by any of the data-sources. These models assume that all individuals observed by each of the sources belong to the hidden (or target) population of interest. However, this assumption is not valid in all cases. For example, we consider the hidden population corresponding to the current number of people who inject drugs (PWID) in Scotland. For such data the Hepatitis C virus (HCV) diagnosis database is used as one of the data-sources to identify PWID. However, the HCV diagnosis data-source does not distinguish between current and former PWID, which, if ignored, will lead to over-estimation of the total population size of current PWID. We extend the standard set of log-linear models to allow for a data-source to contain a mixture of target and non-target individuals (i.e. in this case current and former PWID in the HCV database) and fit the model to data using a Bayesian (model-averaging) approach using the recently developed `R` package `conting`.

# LONGITUDINAL ANALYSIS OF SELF-REPORTED HUNGER IN APPETITE STUDIES

Olena Kravchuk[*1]

[1]*The Biometry Hub, University of Adelaide, Adelaide, Australia*

[*]E-mail: *olena.kravchuk@adelaide.edu.au*

The author has been involved for some time into analysing small-scale satiety trials designed to supplement research projects in food properties and food nutrition. In trials, participants are asked to subjectively evaluate their hunger/fullness on a structured visual analogue scale every 15 – 30 minutes for three hours after consumption of the test meal. In longitudinal analyses of fullness responses conducted by the author, an interesting trend has been observed in the covariance structure of repeated measures. This trend appears to correspond to a widespread 'Satiety Cascade' – a conceptual model that includes cognitive, post-ingestive and post-absorptive stages. Intensive research in human nutrition over the last two decades has linked these stages to specific physiological processes and hormones that supports Satiety Cascade as a feasible model of human satiety. Despite the fact that different processes are clearly identified by nutrition research, in the discipline literature, the fullness responses are still commonly analysed cumulatively as the area under the curve. Unfortunately, little attention, if any, has been given to the correlation structure of repeated measures of fullness, which would contribute to a better understanding of satiety phenomenon. In this presentation, several candidates for covariance structures in the process-based fullness response are investigated. Implications for design and analysis of human satiety and appetite studies are discussed.

# ACQUAINTING ENTRY-LEVEL RESEARCHERS IN AGRICULTURAL SCIENCES WITH PUBLISHED STATISTICS RESEARCH

Olena Kravchuk[*1], David Rutley[1]

[1]*The Biometry Hub, University of Adelaide, Adelaide, Australia*

[*]E-mail: *olena.kravchuk@adelaide.edu.au*

The Grains Research and Development Corporation (GRDC) of Australia recognise the importance of biometrics for the industry. In 2010 – 2014, the GRDC has provided the funding of the 'Capacity Building for Statistics' multi-university project led by Prof. Brian Cullis, NIASRA, University of Wollongong. This project will develop and provide university and industry-based training in biometrics with the aim of improving the pool of statistical expertise in agriculture. Our team addresses this goal for entry-level researchers (Honours year or the first year of Masters in Australia).

It is well understood that a substantial improvement in the collaboration and dialogue between statisticians and agricultural scientists is needed. This is one of many reasons that current agriculture science graduates are not aware of modern, or even not so modern, statistical principles and techniques. Hence, part of our approach is to introduce published statistical research to entry-level agricultural researchers to expand their horizons in biometrics. In this poster we present our method to create an annotated bibliography of key statistics papers for a typical entry-level research project. We annotate statistics research papers in the language of the agricultural student's discipline with supporting visual aids. Informal comments from students and supervisors are briefly summarised. We suggest that "translating" statistics research in a discipline-specific language allows entry-level researchers to appreciate advances in statistics and encourage them to explore statistics research literature.

# MARKOV-MODULATED NONHOMOGENEOUS POISSON PROCESSES FOR DEALING WITH AVAILABILITY BIAS IN SURVEYS OF MARINE MAMMAL ABUNDANCE

Roland Langrock*[1], David L. Borchers[1], Hans S. Skaug[2]

[1] *University of St Andrews, St Andrews, UK*
[2] *University of Bergen, Bergen, Norway*

*E-mail: *roland@mcs.st-andrews.ac.uk*

In shipboard or aerial surveys of marine mammals, detection of an animal is possible only when it surfaces, and with some species a substantial proportion of animals is missed because they are diving and thus not available for detection. This needs to be adequately accounted for in order to avoid biased abundance estimates. In this talk, we discuss an approach that addresses this issue by simultaneously modelling both the animal's availability process and the observer's probability of detecting an animal, given that it is available. The tendency of surfacing events of marine mammals to occur in clusters motivates consideration of the flexible class of Markov-modulated Poisson processes in this context. We embed these models in distance sampling models, introducing nonhomogeneity in the process to account for the fact that the observer's probability of detecting an animal decreases with increasing distance to the animal. We derive approximate expressions for the likelihood of Markov-modulated nonhomogeneous Poisson processes that enable us to estimate the model parameters through numerical maximum likelihood. An application to minke whale data will demonstrate the relevance of the method in abundance estimation.

# ANALYSIS OF CROSS-SECTIONAL TIME SERIES GENE EXPRESSION DATA USING PARAMETRIC REGRESSION MODELS

Philip J. Law[*1], Vicky Buchanan-Wollaston[1,2], Andrew Mead[2]

[1]*Systems Biology DTC, University of Warwick, Coventry, UK*
[2]*School of Life Sciences, University of Warwick, Coventry, UK*

[*]E-mail: *P.J.Law@warwick.ac.uk*

Time series expression experiments (such as microarrays) are often used to determine the dynamic changes in gene expression levels in an organism. A major challenge is the examination of these data to identify associations between the responses of different genes.

Typically when analysing such data, expression profiles are grouped together based on the similarity of responses to each other. This similarity is generally determined through the use of some form of distance metric. In contrast to this type of analysis, clustering of expression profiles in this project is based initially on the identification of expression profiles that can be characterised by linear and non-linear (sigmoidal, exponential, Gaussian and hyperbolic) functions. These functions were selected as being both biologically feasible and having biologically interpretable parameters, and best fitting models were selected based on goodness-of-fit statistics. Genes are then grouped based on the similarity of one or more of the fitted, biologically interpretable parameters. The biological interpretation about the co-expression and co-regulation of genes can thus be associated with particular responses.

This approach is illustrated using a yeast dataset, as well as two large-scale datasets from a project investigating the response of Arabidopsis to a variety of environmental stresses, namely senescence and *Botrytis cinerea* infection. The overall aim of this project is to develop statistical analyses to identify the relationships between genes during multiple stress responses, as well as identifying gene networks that respond to a particular external stress.

# FACTOR MODELLING OF MULTIPLE TIME SERIES TO DETECT TEMPORAL VARIATIONS IN COMMON DYNAMICS

Anne Lehébel[*1,2,3], David Causeur[3]
[1]*INRA, F-44307 Nantes, France*
[2]*LUNAM Université, Oniris, Nantes-Atlantic College of veterinary medicine, Food sciences and Engineering, F-44307 Nantes, France*
[3]*AgrocampusOuest, CNRS, Rennes, France*

[*]E-mail: *anne.lehebel@oniris-nantes.fr*

In veterinary epidemiology, the increased risk of disease emergence has led to the development of surveillance systems to early detect these diseases and reduce their economic impacts. Because the disease to emerge and its consequences are usually unknown, surveillance systems have to rely on the spatiotemporal monitoring of several non specific indicators. Generally, the methods used to monitor these indicators are based on the modelling of the expectation of each time series separately and on the detection of important differences between expected and observed data over time. Preliminary analysis on data collected during Bluetongue virus epidemic (2007) in France showed that during the epidemic, not only expectation but also dependencies between indicators are modified. The objective of this work was to propose a method of combined analysis for multiple response variables to detect changes in time dependencies between variables. We propose the use of a factor model to investigate the specific and common temporal dynamics of multiple variables. We have adopted a two step procedure. After filtering out seasonal and auto-regressive components of each time series, a factor model is fitted on moving time windows. To analyze the variations of dependencies between variables over time, we have used a statistics based on the specific variances of response variables obtained from the variance-covariance matrix decomposition of the factor analysis model. Simulation studies based on the principal characteristics of observed data were used to investigate the ability of this method to detect changes in correlation matrix of variables. Results on simulated data have showed that the monitoring statistics allows detecting changes in common dynamic over time. The impact of the level of dependencies between variables and the impact of several emergence scenarios on the ability to detect changes of dynamics have also been studied thanks to simulation studies and will be presented.

# MICROBIOME, METAGENOMICS AND HIGH-DIMENSIONAL COMPOSITIONAL DATA ANALYSIS

Hongzhe Li

*University of Pennsylvania Perelman School of Medicine*

*E-mail: *hongzhe@upenn.edu*

With the development of next generation sequencing technologies, researchers have now been able to study the microbiome composition using direct sequencing, whose outputs are short sequence reads for each of the microbiome samples. We first introduce a model-based method to quantify the bacterial compositions. We then discuss the issues related to associating the microbiome compositions with environmental covariates or clinical outcomes, including (1) identification of the biological/environmental factors that are associated with bacterial compositions; (2) identification of the bacterial taxa that are associated with clinical outcomes. Statistical models to address these problems need to account for the high-dimensional and sparse and compositional nature of the data. In addition, the prior phylogenetic tree among the bacterial species provides useful information on bacterial phylogeny. We introduce kernel regression and constrained sparse regression models for addressing these issues. We demonstrate the methods using a data set that links human gut microbiome to diet intake in order to identify the micro-nutrients that are associated with the human gut microbiome and the bacteria that are associated with body mass index.

# COMBINING EXPRESSION AND CHROMATIN IMMUNOPRECIPITATION DATA TO UNDERSTAND TRANSCRIPTION FACTOR BEHAVIOUR

Andy G Lynch*[1], Jonathan M Cairns[1], Charlie E Massie[1], Shamith A Samarajiwa[1] and colleagues at the Cancer Research UK Cambridge Institute

[1] *University of Cambridge, Cambridge, UK*

*E-mail: *andy.lynch@cruk.cam.ac.uk*

Transcription Factors (TFs) alter the transcription rates of certain target genes. We wish to identify these TF targets, but the products of these genes can affect the transcription of other genes, meaning that gene expression data alone are not enough for our purpose. Chromatin ImmunoPrecipitation sequencing (ChIP-seq) data can tell us where a TF binds, but a TF doesn't only bind in locations where it will be active, and determining the locations of binding (a step often referred to as 'peak calling') is itself a challenge.

In this poster we present two approaches to combining ChIP and expression data to identify the direct targets of TFs. We also highlight two software packages that we have developed to address these problems. These packages are freely available from www.bioconductor.org.

The first motivating TF is the androgen receptor (AR) in the context of prostate cancer. AR is a key regulator of prostate growth and the principal drug target for the treatment of prostate cancer. AR binds to a wide range of genomic locations, and these must first be identified. We present the BayesPeak pacakge for this purpose. Once identified, genomic targets can be annotated and associated with the expression data.

Our second example TF is the tumour-suppressor p53, whose DNA-damage-response and apoptosis-inducing functions are well documented. Our interest is in the influence of p53 on phenotypes such as senescence. We identify that p53 binds near to the transcription start sites of genes, and use this information to bypass the 'peak-calling' step and take a fully Bayesian approach to identifying its direct targets. We present the Rcade package for performing this analysis.

# QUANTIFYING OVERALL CORRELATION IN HIGH-DIMENSIONAL OMICS DATA

Claus D. Mayer[*1]
[1]*Biomathematics & Statistics Scotland, Aberdeen, UK*

[*]E-mail: *claus.mayer@bioss.ac.uk*

Extremely high-dimensional data sets from gene expression (microarrays, RNAseq experiments) or metabolomic studies are commonly generated in biological and medical experiments. Variables (genes, metabolites) measured in these experiments typically interact with each other in gene regulatory networks or metabolic pathways, leading to correlation between variables. From a purely statistical point of view this can be a nuisance. In a highly multiple testing setting methods to control the family wise error rate (FWER) or the false discovery rate (FDR) usually assume independence or only weak correlation of variables. From a biological point of view, however, strong correlations are often of particular interest because they indicate the activation of important processes.

For either case it is useful to have a method that measures and quantifies the overall correlation either in the whole data set or in relevant subsets (e.g. pathways). Here, we will focus on methods that estimate the "effective number of tests" in the area of genome wide association studies (GWAS). These are based on the linkage disequilibrium (LD) matrix and can be transferred to our situation by applying them to the correlation matrix instead.

We will show that some popular "effective number of tests" method are not useful for controlling the FWER as they do not take the actual error level (typically 5%) into account. We will also discuss whether these methods might still be suitable for quantifying and comparing correlation strength across data sets or subsets of data.

# INTEGRATED ANALYSIS OF CAPTURE-RECAPTURE-RESIGHTING DATA AND COUNTS OF UNMARKED BIRDS AT STOP-OVER SITES

Eleni Matechou*[1], Byron J. T. Morgan[2], Shirley Pledger[3], Jaime A. Collazo[4], Jim E. Lyons[5]

[1]*University of Oxford, Oxford, UK*
[2] *University of Kent, Canterbury, UK*
[3]*Victoria University, Wellington, NZ*
[4]*North Carolina State University, Raleigh*
[5] *Patuxent Wildlife Research Center, Laurel*

*E-mail: *matechou@stats.ox.ac.uk*

The models presented in this paper are motivated by a stop-over study of semipalmated sandpipers, *Calidris pusilla*. Two sets of data were collected at the stop-over site: a capture-recapture-resighting data set and a vector of counts of unmarked birds. The two data sets are analysed simultaneously by combining a new model for the capture-recapture-resighting data set with a binomial likelihood for the counts. The models use finite mixtures to group the birds according to their unknown time of arrival and link the probability of remaining at the site to the unknown time spent at the site. The aim of the analysis is to estimate the total number of birds that used the site and the average duration of stop-over. The combined analysis is shown to be highly efficient, even when just 1% of birds are recaptured, and is recommended for similar investigations.

# MULTIPLE TESTING METHOD FOR THE DIRECTED ACYCLIC GRAPH, USING SHAFFER COMBINATIONS

Rosa J. Meijer*[1], Jelle J. Goeman[2]

[1,2] *Leiden University Medical Center, Leiden, the Netherlands*

*E-mail: *r.j.meijer@lumc.nl*

We present a novel multiple testing method for testing null-hypotheses that correspond to nodes in a directed acyclic graph (DAG). Such DAG-structured multiple testing problems can be encountered in various settings. One well-known example is the problem of performing a gene-set analysis, in which multiple gene-sets and individual genes are tested on their association with a clinical outcome. Each hypothesis about a gene or gene-set can be considered a node in a DAG in which the edges correspond to subset-relationships between the nodes. The gene ontology (GO) graph is a specific example of such a DAG.

Although several multiple testing methods have been developed for gene-set analysis, the novelty of our method is that it uses the specific DAG structure to make statements on the possibility of certain configurations of true and false null hypotheses. By constructing/extending the DAG in such a way that every node is the intersection of its child-nodes, it will often happen that the logical relationships between the hypotheses will create *restricted combinations*, which means that not all remaining hypotheses can simultaneously be true. Using this information can reduce the multiple testing burden. Our method can be seen as an extension of Meinshausen's familywise error rate controlling procedure for tree-structured hypotheses.

Implementing our method requires repeated solutions of instances of the *minimum hitting set problem*, which is known to be an NP-hard problem. Depending on the size of the DAG, we either calculate these solutions exactly by using an ILP (integer linear programming) solver or we use approximations based on a greedy algorithm.

The method will be illustrated by testing Gene Ontology terms for evidence of differential expression in a survival setting.

# COMPETING RISKS REGRESSION WITH MISSING CAUSES OF DEATH: ASSESSING THE SENSITIVITY OF INFERENCES TO MISSING DATA ASSUMPTIONS

Margarita Moreno-Betancur[*1,2], Grégoire Rey[3], Aurélien Latouche[4,2]

[1] *Inserm Centre for research in Epidemiology and Population Health, Villejuif, France*
[2]*Univ Paris-Sud, Villejuif, France*
[3]*Inserm Centre for research on the Epidemiology of Causes of Death, Le Kremlin-Bicêtre, France*
[4]*Conservatoire national des arts et métiers, Paris, France*
[*]E-mail: *margarita.moreno@inserm.fr*

In many studies of cause-specific mortality, especially in general population studies, it is difficult to obtain complete cause-of-death information for all individuals. Thus, regression strategies for modeling the cause-specific hazards when some causes of death are missing have been proposed by several authors. Nevertheless, most approaches found in the literature rely on the assumption that the causes of death are missing at random (MAR), that is, that the probability of missingness is independent of the cause of death when conditioning on the observed data. Unfortunately, this assumption can never be verified from the data available. This issue, common to all missing data problems, has led to an increasing awareness of the need to perform sensitivity analyses to assess the robustness of inferences to departures from the MAR assumption. However, no standard method exists for setting up such an analysis, each specific scientific context requires different considerations and this is still an active area of research. In the missing cause of death setting, we propose a flexible procedure to perform sensitivity analyses following an initial MAR analysis of the cause-specific hazards. The methodology relies on the pattern-mixture model factorization of the full data likelihood and allows the analyst to formulate assumptions about the missing data distribution in an explicit manner. The approach was prompted by a study of the socio-economic differentials in suicide mortality in France. The French national cause-of-death register was suspected to contain many suicides mistakenly coded as deaths with unknown cause because of an area-specific reporting issue. The MAR assumption was therefore implausible. This study illustrated the practical value of our approach and underlined the need for sensitivity analyses when analyzing competing risks data with missing causes of death.

# THE SINH-ARCSINH DISTRIBUTION FOR BINARY DOSE-RESPONSE DATA

Martin S. Ridout and Byron J. T. Morgan*

*University of Kent, Canterbury, UK*

*E-mail: *B.J.T.Morgan@kent.ac.uk*

There is substantial, continuing interest in the use of extended models for binary dose-response data. See for example Bazn et al (2006) and Eyheramendy et al (2007). We investigate the use of a four-parameter model based on the sinh-arcsinh distribution, described in Jones and Pewsey(2009), which flexibly allows for skewness and heavy tails relative to symmetric reference distributions such as the logistic and normal. It is shown that this new model has appreciable advantages, as it easy to check how well the model improves upon simpler nested alternatives, using goodness-of- link tests, it provides a simple explicit expression for quantiles, which are frequently used to summarise binary dose-response data, and it readily produces profile confidence intervals for such quantiles. Illustrations of the model performance are provided using real and simulated data.

## References

Bazán, J.L., Bolfarine, H. and Branco, M. D. (2006) A skew item response model. *Bayesian Analysis*, **1**, 861-892.

Eyheramendy, S. and Madigan, D. (2007) A flexible Bayesian generalized linear model for dichotomous response data with an application to text categorization. Pp 76-9, in Liu, R., Strawderman, W. and Zhang, C.-H. (Eds.) *Complex data sets and inverse problems: tomography, networks and beyond*, Beechwood Ohio, US, Institute of Mathematical Statistics.

Jones, M.C. and Pewsey, A. (2009) Sinh-arcsinh distributions. *Biometrika*, **96**, 761-780.

# A BAYESIAN JOINT MODEL FOR REPEATED EVENTS OF DIFFERENT TYPES AND MULTIPLE BIOMARKERS

J.Z. Musoro*[1], R.B. Geskus[1] and A.H. Zwinderman[1]

*Academic Medical Center - University of Amsterdam, The Netherlands*

*E-mail: *z.j.musoro@amc.uva.nl*

We fitted a joint model for the development of multiple biomarkers over time and repeated infections of different types. Our study was motivated by post kidney transplantation records of 467 patients who could experience up to ten infection types, all multiple times. Patients immune states were monitored longitudinally using five biomarkers. With primary interest on the event process, we propose a multivariate joint model comprising of 1? a multivariate spline based mixed effects submodel to explicitly capture the biomarker trajectories, accounting for the dependency between the repeated measurements of a biomarker over time as well as the relationship between different biomarkers, and 2? an infection-specific Cox submodel with random effects to account for the association within and between repeated infections of different types. The baseline risk functions were infection type specific and unspecified. The association between the two submodels of the joint model was modeled via shared latent terms. We implemented the parameterization used in joint models which includes the fitted intermediate longitudinal measurements as time-dependent covariates in a relative risk model. Our proposed model was implemented in OpenBUGS using the MCMC approach. Findings suggested that low biomarker values were related to high infection risks, with patients previously infected having a higher risk for future infections of any type. We also showed via a simulation study that the proposed algorithm works well.

# PREDICTING METHANE EMISSIONS FROM CATTLE – WHERE META-ANALYSIS AND RANDOM COEFFICIENT MODELLING MEET

Ian M. Nevison[*1], Patricia Ricci[2], Anthony Waterhouse[2], John Rooke[2]

[1]*Biomathematics and Statistics Scotland, Edinburgh, UK*
[2]*SRUC, Edinburgh, UK*

[*]E-mail: *ian.nevison@bioss.ac.uk*

Methane ($CH_4$) emissions from ruminants both represent an efficiency loss for agricultural production systems and are estimated to account for the 2.5% of total UK greenhouse gas emissions. Devising reduction strategies requires the ability to predict emissions for a range of scenarios.

58 suitable methane studies were identified from a literature review of beef and dairy cattle. These covered various management and physiological states. They also collectively encompassed a wide range of values for intake covariates and bodyweight. All studies published methane emission means and associated standard errors for at least two distinct combinations of the covariate values, enabling compilation of a database. This was split for calibration and validation purposes.

Random coefficient models are commonly fitted to data from individual studies to allow different levels of a categorical variable to exhibit heterogeneous responses to a covariate of interest. We have applied this general methodology in a meta-analysis context with a categorical variable representing the methane studies and covariates such as metabolic body weight and gross energy intake. Weightings proportional to the published variances of the recorded covariate combination means for methane in each study were used to account for their differing precisions.

An initial screening process of potential predictors was performed. Methane emissions were regressed on each potential predictor in turn and its associated P-value computed. Any with P-values greater than 0.25 were discarded. A random coefficients prediction model was then built up using a process analogous to stepwise regression. This considered all predictors passing the initial screening. Where a continuous covariate was added to the stepwise model, study-specific slopes for that covariate were also fitted unless their variance component was estimated as negative.

The approach has shown its value in exploring relationships based on multiple studies and obtaining predictions with associated standard errors for a range of scenarios.

# BAYESIAN NETWORK BASED DIFFUSION ANALYSIS OF STARLING DATA

Glenna Nightingale[*1],

[1]*University of St Andrews, St Andrews, UK*

[*]E-mail: *glenna.evans@gmail.com*

Here we examine the robustness of a recently developed method for studying social transmission of behaviour in groups of animals: network based diffusion analysis (NBDA). We fit NBDA models to diffusion data derived from observations of foraging bouts in starlings (*Sturnus vulgaris*) given knowledge of their patterns of association, foraging times, and covariate morphometric data.

We employ a reversible jump Markov chain Monte Carlo (RJMCMCO algorithm to discriminate between the 14 different models derived from various combinations of parameters involved. Our analysis extends the current use of NBDA models to incorporate random effects and facilitate model discrimination. This methodology is likely to be particularly useful to deal with datasets that include many covariates and that can be fitted with a correspondingly large number of competing models.

# A WHOLE GENOME PREDICTION TYPE APPROACH FOR THE GENOME WIDE ASSOCIATION ANALYSIS OF MULTI-ENVIRONMENT TRIAL DATA

Helena Oakey*, Brian P Cullis, Robin Thompson, Jordi Comadran, Nicola Uzrek, Claire Halpin, Robbie Waugh

*James Hutton Institute, University of Dundee, Dundee, UK*

*E-mail: *h.oakey@dundee.ac.uk*

Genome-wide Association Studies (GWAS) aim to determine the location of QTL which control phenotypic traits through the association that occurs between QTL and molecular markers of known position. The relationship between the individuals in GWAS may impose structure within the population which can induce spurious associations. Mixed model approaches that adjust for this structure have been developed, however the methods do not necessarily capture all the structure.

In crops, phenotypic information on varieties is often collected across multiple trials referred to a multi-environment trial (MET). In a GWAS, a MET offers the opportunity to assess environment impact on QTL and thus determine QTL by environment interactions. However, most methods of multi-environment QTL analysis are multi-stage and involve a genome scan of markers making them inefficient. Approaches used in Whole Genomic Prediction (WGP) which use similar populations to GWAS could offer useful insights for GWAS. A WGP type approach which fits all markers simultaneously is attractive as it accounts for population structure and eliminates the need for multiple models and stages.

In this paper a WGP type approach for the GWAS analysis of MET data is proposed. The approach fits a mixed model to the data which incorporates environments, environment by marker interactions, residual genetic by environment interactions as well environment specific field and randomisation based terms. Accounting for linkage disequilibrium between markers is also possible within this model. The approach simultaneously allows genetic and non-genetic variation to be accounted for it does not involve QTL scans or determination of population structure making it a very efficient approach to multi-environment association analysis.

# BUILDING HIERARCHICAL MODELS WITH AN INTEGRATED LIKELIHOOD FOR DISTANCE SAMPLING DATA

Cornelia S. Oedekoven[*1], Stephen T. Buckland[1], Monique L. Mackenzie[1], Ruth King[1], Kristine O. Evans[2] and Loren W. Burger, Jr.[2]

[1]*University of St Andrews, St Andrews, UK*
[2]*Mississippi State University, Mississippi State, MS, USA*

[*]E-mail: *cornelia@mcs.st-and.ac.uk*

Buckland et al. 2009 developed a two-stage approach for analysing count data from large-scale experimental studies. In a first step, a detection function is fitted to the distance data and the effective area estimated. In a second step, the effective area is included in a log-linear count model to adjust for imperfect detectability. Precision estimates are obtained using a nonparametric bootstrap. We propose an integrated likelihood approach which includes a random effect in the count model to accommodate correlated counts due to repeat visits to the same sites. Here parameter and precision estimates for both steps are obtained in one step by maximising one combined likelihood function pertaining to both the detection function and count model.

We use this integrated likelihood function but rather than obtaining parameter estimates by maximising this function we use a Bayesian approach to obtain summary statistics for parameters. A reversible Jump MCMC algorithm is used to explore model and parameter space simultaneously. Each iteration includes a between-model and a within-model move. For the between-model move, a new model is proposed and accepted based on some probability. For the within-model move, a Metropolis Hastings algorithm is used to update the parameters in the current model. It is assumed that the chain improves during each iteration and that – after a burn-in phase – the chain reaches convergence. We illustrate the method using a large-scale point-transect study of northern bobwhite coveys where the interest lies in the effect of establishing conservation buffers along field margins.

# USING THE DIRICHLET PROCESS CLUSTERING MODEL TO ASSIST THE INVESTIGATION OF HIGH ORDER INTERACTIONS BETWEEN ENVIRONMENTAL/GENETIC COVARIATES

Michail Papathomas[*1], Sylvia Richardson[2]

[1]*University of St Andrews, St Andrews, UK*
[2]*MRC Biostatistics Unit, Cambridge, UK*

[*]E-mail: *michail@mcs.st-and.ac.uk*

Detecting high order interactions between covariates in a linear model framework is not straightforward, due to the difficulty in investigating an unwieldy large space of competing models. One approach for reducing the dimensionality of the problem is to create homogenous clusters created using the Dirichlet process, a Bayesian flexible clustering algorithm. In profile regression, the covariate profiles of the subjects are clustered into groups, and the groups are associated via a regression model to a relevant outcome. A variable selection approach is implemented in tandem with profile regression, for the detection of important environmental or genetic risk factors. We present one example where profile regression is applied to a large cohort study in order to examine the effect of environmental carcinogens and explore possible gene-environment interactions. In another illustration, we analyse data from a GWA study on lung cancer, in order to explore gene-gene interactions. We briefly discuss the recently released R package PReMiuM for profile regression mixture models using the Dirichlet Process. The package allows for Bernoulli, Binomial, Poisson and Normal outcomes, with Normal and discrete covariates. Finally, we briefly discuss the relation between interactions within the profile regression framework, and interaction terms in a standard linear model.

# ANALYSING ANTI-MALARIAL TRIALS: A MULTIPLE ENDPOINT APPROACH

Alice Parry[*1], Thomas Jaki[1], Ian Hastings[2] and Katherine Kay[2]

[1]*Lancaster University, Lancaster, UK*
[2]*Liverpool School of Tropical Medicine, Liverpool, UK*

[*]*Email: a.parry1@lancaster.ac.uk*

The aims of a successful malaria treatment are to primarily cure the original infection and then secondly to prevent new infections. Since malaria could be described as a chronic illness with new infections being inevitable at some point, prevention of future new infections would be difficult, if not impossible. Nevertheless it would be beneficial if the length of time between infections could be increased. It would therefore be advantageous to include the time that a patient is free from the parasites into the assessment of the efficacy of a treatment.

We have devised a multiple endpoint approach which incorporates proportion cured along with the time to a new infection in the same analysis (i.e. a binary endpoint jointly with a survival endpoint), using score statistics, to give an overall assessment of the efficacy of the treatment. The talk will discuss the initial set up of the conditional hypotheses and then explore two approaches for finding the critical values and sample size using a 'false-claim' error rate as opposed to the traditional family-wise error rate. The motivation for using a false-claim error rate will be discussed and the score statistics and corresponding covariances between the score statistics will be described.

The features which arise from the simulations of these methods and the resulting critical values and sample sizes required for the different data types will be explored and issues discussed. We found there was little change in the required critical values and, as expected, the sample size reduced as we moved from binary to survival data, as more information was included.

# THE EFFECT OF ANIMAL MOVEMENT ON POINT TRANSECT ESTIMATES OF ABUNDANCE

Rocío Prieto González*[1], Len Thomas[1], Tiago A. Marques[1]

[1] *University of St Andrews, St Andrews, UK*

*E-mail: *rpg2@st-andrews.ac.uk*

Distance sampling is one of the most widely used methods for estimating animal population abundance. The main methods are line and point transects. In both, the observer performs a survey along a randomly located series of lines or points and measures distances to detected animals. These are used to estimate the average probability of detecting an animal, allowing missed animals to be accounted for. They rely on three assumptions: (i) animals are detected with certainty on the line or point, (ii) animals dont move while within detection range and (iii) measurements are exact.

We focus on the second assumption, assessing the effects of animal movement in point transect sampling. In particular, for estimating cetacean abundance, the advantages of using their vocalizations collected from fixed long-term passive acoustic sensors are many. However, animal movement can be a major issue and therefore, the estimates of abundance are potentially biased.

First of all, we suppose that all animals in the covered area are detected (circular plot). Each animal is assumed to be moving at the same constant speed and random direction. An analytic expression for the bias from random animal movement is derived, as well as a corrected abundance estimate.

Secondly, some animals in the covered area may be undetected (point transect sampling). We assess the effects of their movement by simulation. Within the study region, we simulate animals moving in constant but random directions and constant speed or according to a correlated random walk. The detection process is modelled as a two-dimensional hazard-rate process. Our goal is to describe how movement leads to bias and how that bias can be corrected for, leading to more robust density estimates.

# DRAWING THE LINE : CHANGING THE FUNNEL PLOT FOR SCREENING HOSPITAL PERFORMANCE

Bart Van Rompaye[*1], Marie Eriksson[2], Els Goetghebeur[1]

[1] *Ghent University, Ghent, Belgium*
[2] *USBE, Umeå University, Umeå, Sweden*

[*]E-mail: *bart.vanrompaye@ugent.be*

The funnel plot shows how hospital-specific effect estimates vary in function of their precision, influenced by sample size. Limits of prediction intervals around the overall average effect can then typically guide the labeling of centers with high or low effects worthy of further examination. Corresponding decision lines are driven by statistical significance but could be misleading, for instance when unduly small effect sizes are found significant in larger centers.

We propose new decision lines based on a balanced testing procedure that weights evidence against standard performance versus evidence against some pre-specified clinically relevant deviation. We show how this procedure supports the different importance attributed to type I and type II errors in different settings. We find the new boundaries leading to meaningful discussions with clinicians, and a useful set of alternative decision boundaries on the funnel plot. Its cost-efficiency in care-intervention settings with limited resources is illustrated.

This is applied to the evaluation of acute stroke-care in Sweden, using data from the Riks-Stroke registry (http://www.riks-stroke.org). The procedure is used here to evaluate stroke units through an indirectly standardized excess cause-specific cumulative incidence (ECSCI) in a competing risks setting. This directly relevant outcome forms a basis for discussions on benchmarks of interest. The targeted powerful screening tool is readily adapted to other evaluation contexts.

Finally, we extend the discussion to the corresponding evaluation of the evolution of quality of care over time.

# MODELS FOR CLUSTERED INTERVAL-CENSORED OUTCOME WITH A DEPENDENT TERMINAL EVENT: INVESTIGATING THE INTRA-COUPLE CORRELATION FOR DEMENTIA

Virginie Rondeau[*1,2], Alexandre Laurent[1,2], Pierre Joly[1,2], Catherine Helmer[1,2]

[1]*INSERM U897, ISPED, Bordeaux, France*
[2]*University Bordeaux Segalen, Bordeaux, France*

[*]E-mail: *virginie.rondeau@isped.u-bordeaux2.fr*

Event history analysis may be particularly complex because of interval-censored and clustered event times but also because of a dependent terminal event. While methods have been developed and are easily practically implemented, for the analysis of correlated survival outcomes in settings where observations are either left or right censored, analysis methods for settings where observations are correlated and interval censored, with also a dependent terminal event, are not as well developed. Most standard survival analyses are still based on the assumption of independence between time to event endpoints and the terminal event. When the independence assumption is questionable, the inference based on standard methodologies may be biased and possibly misleading. Often the occurrence of a serious event, may be associated with an elevated risk of death. This dependence should be accounted for in the joint modelling of time to the event of interest and death.

The approach we develop in this paper is motivated by a population-based study ("3 cities" study) of 1308 couples of subjects over 65 years with questionnaires and performed clinical examinations at baseline and 2, 4, 7 and 10 years after. Its main objective is to estimate the risk of dementia attributable to vascular factors. In this application scheme, couples are natural clusters and an intra-couple clustering might be present. Furthermore, the time to dementia is not known exactly; it is only known that the event occurred between the last visit without dementia and the first visit with dementia; therefore, the time to dementia is interval censored. Furthermore, censoring by death might be associated to worsening of symptoms, i.e. the patient is at higher risk of dementia. Individual prediction of dementia or death can be obtained with this approach.

So we propose, a semi parametric penalized likelihood method for estimating hazard functions in a general joint frailty model for time to dementia and death, with clustered, interval-censored data. These methodological developments are associated with a user-friendly package for R: "Frailtypack"( *http://cran.r-project.org/web/packages/frailtypack/index.html*).

# PROBABILISTIC MODEL FOR PREDICTING THE PATHOGENICITY OF SEQUENCE VARIANTS

Dace Ruklisa*[1], James Ware[2], Roddy Walsh[2], Stuart Cook[2], David Balding[1]

[1]*UCL Genetics institute, London, UK*
[2]*Molecular Cardiology, Medical Research Council Clinical Sciences Centre, Imperial College London, London, UK*

*E-mail: *d.ruklisa@ucl.ac.uk*

One of the most challenging tasks for genetic analysis of genetically heterogeneous syndromes is distinguishing between pathogenic and harmless variation within associated genes. High-throughput sequencing technology is crucial in pointing to pathogenic rare variants, which will help to identify individuals at risk. This requires statistical methods to help interpret the role of novel sequence variants.

We propose a method for probabilistic prediction of pathogenic variants that takes into account the characteristics of syndrome and is gene- and domain-specific. Our approach is based on a hierarchical logistic regression model that incorporates properties of an individual variant alongside gene and domain level predictors. As the pathogenicity does not depend linearly on some of the features, we model these by splines with a small number of knots. The knots are selected manually, from inspecting the distributions of predictor values in our database of pathogenic and benign variants. The pathogenicity prediction model is an outcome of a model selection process, which guided the choice of domain level predictors and of various characteristics of the sequence variants.

We assess model performance via cross-validation and illustrate it by ROC curves. We show that syndrome-specific models incorporating prior knowledge about the genetic architecture of the disease can achieve high sensitivity and specificity and improve upon more general genome-wide approaches to pathogenicity prediction. Our aim is to implement the proposed pathogenicity model in a tool that can be used by clinicians, when a novel rare variant is encountered for a patient.

We applied the method to the following inherited cardiac conditions: Brugada syndrome, long QT syndrome and hypertrophic cardiomyopathy (HCM). Our model was trained on a carefully-curated set of annotated variants that are classified as pathogenic or benign according to a rigorous definition, separately for each syndrome.

# USING MULTIPLE DIAGNOSTIC TESTS TO RECOVER INCIDENCE TRENDS

Gustaf Rydevik[123], Mike R Hutchings[2], Piran C L White[3], Ross S Davidson[12], Glenn Marion[1], Giles Innocent[1]
[1]*Biomathematics and Statistics Scotland, Edinburgh, UK*
[2]*SRUC, Edinburgh, UK*
[3]*University of York, York, UK*

[*]E-mail: *gustaf.rydevik@bioss.ac.uk*

Disease surveillance data can often have less than complete coverage in time and space, missing vital information that limits the ability to best respond to pathogen outbreaks. Here we describe a statistical approach which partially addresses this problem by combining data from multiple diagnostic tests to recover information on the past spread of an infectious disease.

Consider a cross-sectional sample of individuals infected with a pathogen and diagnostic tests whose real-valued responses vary with time since infection. We define a likelihood function for the time since infection given two or more such diagnostic test results, which incorporates simple models for their temporal response. We conduct Bayesian inference obtaining estimates of the posterior distribution of time since infection for a population of samples via Markov chain Monte Carlo. In cases where the test characteristics differ as a function of time since infection, this approach may provide valuable insights into the development of the spread of the pathogen prior to the times of sampling.

We investigate the properties and limitations of this approach by applying it to data from simulations that mimic diagnostic test results from a Bluetongue virus outbreak. The resulting estimated incidence curves reflect the recent rate of change of incidence in the population as a whole. We also investigate how the ability to recover the curve is affected by different levels of individual test variation and what combinations of temporal test response curves provide useful information. Finally, we discuss the potential benefits that this approach might offer for the surveillance and management of infectious diseases epidemics.

# STATISTICAL CHALLENGES OF HIGH DIMENSIONAL METHYLATION DATA

Maral Saadati*[1], Axel Benner[1]

[1]*German Cancer Research Center (DKFZ), Heidelberg, Germany*

*E-mail: *m.saadati@dkfz-heidelberg.de*

With the fast growing field of epigenetics comes the need to better understand the intricacies of methylation data analysis. Challenges arise from the fact that methylation values (so-called beta values) are proportions between 0 and 1, often from a bimodal distribution with peaks close to 0 and 1. Therefore, the majority of standard statistical approaches do not apply. The logit transformation into so-called m-values is a common approach to circumvent this problem and allow the use of common statistical methods. However, it can be observed that the transformation from beta to m-values does not necessarily result in an approximately normal distribution. Often bimodality, asymmetry and heteroscedasticity are conserved even after transformation.

We give an overview and discussion of methods suggested in the recent years that attempt to address the characteristics of methylation data in certain research questions. For example, beta regression models with fixed and random effects for screening of "differential" methylation between groups while adjusting for confounders, model based clustering to derive methylation phenotypes using mixtures of beta distributions, random forests for classification and prediction of patient survival in high dimensional settings. Our goal is to sensitise researchers to the challenges and issues that arise from this type of data as well as to present possible solutions.

# ESTIMATION IN MIXED MODELS DEFINED BY STOCHASTIC DIFFERENTIAL EQUATIONS

Adeline Samson

*UMR CNRS 8145, Laboratoire MAP5, Université Paris Descartes, France*

*E-mail: *adeline.samson@parisdescartes.fr*

Biological processes are generally measured repeatedly along time for several subjects. The classical statistical approach to analyze these longitudinal data is mixed models (Pinheiro et Bates 2000). In biology, the regression function of these mixed models is often described by deterministic models based on ordinary differential equations. However, these functions are not satisfactory when biological processes involved a random behavior which can not be neglected. Thus we consider physiological models based on stochastic differential equations (SDE). Our aim is to propose estimation methods for mixed models defined by SDE. Estimation of SDE has been widely studied (Ait Sahalia 2002). However, their extension to mixed models is not simple because the likelihood of these models is not ex- plicit. Several approaches have been proposed for mixed models with a deterministic regression function (Davidian and Giltinan 1995, Pinheiro and Bates 2000, Kuhn and Lavielle 2005). Especially, Kuhn and Lavielle (2005) propose to couple the SAEM algorithm with a Markov Chain Monte Carlo (MCMC) method. In the context of mixed models defined by SDE, Tornoe et al (2005) propose a method based on the extended Kalman filter but the convergence of their method is not proved. We will present several estimation methods, either exact when based on the continuous time observations of the trajectories (Delattre, Genon-Catalot and Samson 2013) or based on the SAEM algorithm coupled with a MCMC algorithm (Donnet and Samson 2008) or a particle filter (Donnet and Samson 2013). Nonparametric approach could also be adapted to this case (Comte, Genon-Catalot and Samson 2013).

**Bibliography**

[1] Ait-Sahalia, Y. (2002), Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach, Econometrica, 70, 223-262; [2] Andrieu, C., Doucet, A. et Holenstein, R. (2010), Particle Markov chain Monte Carlo methods, J. R. Statist. Soc. B, 72, 1-33; [3] Comte F, Genon-Catalot V, Samson A. (2013) Nonparametric estimation for stochastic differential equations with random effects. submitted; [4] Davidian, M. et Giltinan, D.M. (1995), Nonlinear models to repeated measurement data, Chapman and Hall; [5] Delyon, B., Lavielle, M. et Moulines, E. (1999), Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist., 27, 94128; [6] Delattre M, Genon-Catalot V, Samson A (2013). Maximum likelihood estimation for stochastic differential equations with random effects, Scandinavian Journal of Statistics, to appear; [7] Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, Jr. R. Stat. Soc. B, 39, 138; [8] Donnet, S. et Samson, A. (2008), Parametric inference for mixed models defined by stochastic differential equations, ESAIM P&S, 12, 196218; [9] Donnet, S. et Samson, A. (2013), EM algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models, submitted; [10] Kuhn, E. et Lavielle, M. (2005), Maximum likelihood estimation in nonlinear mixed effects models, Comput. Statist. Data Anal., 49, 10201038; [11] Pinheiro, J. et Bates, D. (2000) Mixed-effect models in S and Splus, Springer-Verlag. [12] Torne, C.W., Overgaard, R.V., Agers, H., Nielsen, H.A., Madsen, H. et Jonsson, E.N. (2005), Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations, Pharm. Res., 22, 124758.

# PREDICTION OF SUCCES OF INSEMINATION WITH PAST DAILY MILK YIELD WITH A FUNCTIONAL GENERALIZED LINEAR MODEL

Cécile Sauder*[1,2], Catherine Disenhaus[1,2], Yannick Le Cozler[1,2], Hervé Cardot[3]

[1]*INRA, Saint-Gilles, France*
[2]*Agrocampus Ouest, Rennes, France*
[3] *Université de Bourgogne, Dijon, France*

*E-mail: *cecile.sauder@rennes.inra.fr*

In modern dairy cows farming, the recent availability of numerous automatically collected data to phenotypic functional traits opens new opportunities. Body weight, milk yield, progesterone level in milk, food intake are now often available on a daily or more frequently basis. This information which completes the traditionally manually collected one is mostly under exploited. These daily records result in new functional data from different kind of variables, which are of interest to improve animals performance and longevity.
Being able to identify and analyse such interesting and original curves is then a new and exciting challenge in dairy management. Parametric approaches are traditionally used and regarding milk yield curves, Wood model (1967) is being used for decades. These models are very efficient on a population level, but are out of interest on an individual level.
B-spline modeling appear to be more accurate to perform such studies and were used to study daily milk yield in dairy lactating cattle. 362 milk yield curves of the 42 first days of lactation are available to predict success of first insemination, done after these 42 days if a heat is observed on a cow. Milk yield curves are smoothed on a basis of 21 splines and a little smoothing parameter to keep variations. A functional generalized linear model with a logit link and the binomial family is performed to predict success at first insemination. Spline smoothing functions and their first and second derivatives are considered as functional predictors. Breed (Normande or Holstein) and parity (primiparous or multiparous) are added in the model as non functional predictors. Comparing to a classical model with the milk yield sum instead of functional variables, functional model decrease the resubstitution error rate of 2%. Adding others functional variables in the model such as body weight or body condition score should improve quality prediction.

# MODELLING PLANT HEIGHT DATA WITH SCALED AND SHIFTED PROTOTYPE CURVES

Sabine K. Schnabel[1], Paul H.C. Eilers[1,2], Fred A. van Eeuwijk[*1]

[1]*Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands*
*Erasmus University Medical Center, Rotterdam, The Netherlands*

[*]E-mail: *fred.vaneeuwijk@wur.nl*

In plant research phenotypic data are mainly collected through field and greenhouse experiments. Measurements are often taken as time series. The data is collected either by hand or completely automatically. We will use plant height in potato as an example for such typical data. A lot of growth traits show a characteristic monotonically increasing shape. To combine such measurements with genetic data meaningful summaries need to be developed. These values are used in subsequent QTL mapping.
In the literature several methods have been proposed. They include classic parametric approaches with the logistic model and semi-parametric flexible splines. Here, we study models assuming one mother curve that has been shifted and stretched on both the time and the vertical axes. The transformation parameters as well as the underlying curve itself are estimated from the data. In our example the response has been log-transformed. Therefore a vertical shift translates into a stretch on the original axis. The approach is loosely based on a three-parameter mixed model for the height of children. In contrast to this model, we use P-splines for a flexible functional form and work with a fixed model.
Our model has been successfully applied to height measurements in a potato field experiment for a population with more than 100 genotypes. Genotype-specific characteristics can be extracted such as the horizontal shifts and vertical stretches to be used in QTL analyses.
While the example data are typical for plant breeding trials, they are a simplification of the real situation. To complicate matters data are often measured for several replications of the same plant. Additionally often measurements come in the form of a mix of longitudinal and cross-sectional data due to intermediate harvests. In order to provide more powerful tools for the plant research community these topics will be investigated in future work.

# SPECIES INTERACTIONS AS A MULTIVARIATE FEEDBACK SYSTEM

Hideyasu Shimadzu[*1,3], Maria Dornelas[1], Peter A. Henderson[2], Anne E. Magurran[1]

[1]*University of St Andrews, St Andrews, UK*
[2]*Pisces Conservation, Lymington, UK*
[3]*Keio University, Yokohama, Japan*

[*]E-mail: *hs50@st-andrews.ac.uk*

Understanding the interactions between species within an ecological community is a key challenge in biodiversity research. Here, we focus on monthly time series records of an exceptionally well-documented estuarine fish assemblage in the Bristol Channel. Given the multi-species time series data, we have developed a model for a multivariate feedback system in which the outputs can be the inputs and *vice versa*. The model assumes linear interactions between species as a tractable approximation. To examine the extent to which the abundance of a given species is driven by the other species, and by environmental variables, we have analysed the model in the spectrum domain and calculated the contribution ratio of each species at each frequency.

The result suggests that our modelling approach dealing with an ecological community as a multivariate feedback system provides new insights into species interactions. We demonstrate how it enables further analysis into ecologically relevant groups of species that underpin the functioning of the system as a whole.

# ESTIMATING SURVIVAL AND FECUNDITY IN GREY SEALS *HALICHOERUS GRYPUS*

Sophie Smout*[1], Ruth King[2], Paddy Pomeroy[1]

[1]*School of Biology, University of St Andrews, St Andrews, UK*
[1]*School of Mathematics and Statistics, University of St Andrews, St Andrews, UK*

*E-mail: *scs10@st-andrews.ac.uk*

Numbers and population dynamics of UK grey seals (*Halichoerus grypus*) are of considerable applied interest: the UK is a stronghold for these animals internationally and their conservation is required by legislation, but they also prey on stocks of commercial fish. In general, the population of UK grey seals is growing. However trends in pup production and recruitment vary between breeding colonies. Adult female grey seals were observed from 1985-2006 at the Isle of May breeding colony, and from 1985- 2006 at the North Rona colony. Associated measurements of individual covariates (mass and breeding status) were also recorded. Using such data, it is straightforward to estimate the fecundity of females when present at the colony and this rate is very high. However, we cannot easily estimate the overall fecundity of these females because there are years in which they are not observed at the colony. In these cases it is possible that they are present at the colony but unobserved, pupping elsewhere during these years, or skip reproduction during these years. A flexible Bayesian hidden process framework is used to perform an analysis of the mark re-sighting data and covariate data in order to obtain inference on overall fecundity, making use of the relationship between mass and breeding status. The process model includes survival, changes in mass that are related to breeding and lactation, and pupping-probability as a function of female mass at the end of the previous breeding season. Nuisance parameters such re- sighting probability and tag loss rates are estimated along with the biologically interesting parameters e.g. for the relationship between fecundity and maternal mass, the cost of lactation, and differences in mass gain between breeding and non-breeding animals. The model is fitted using a data augmentation approach.

# HERITABILITY ANALYSIS: BEYOND VARIANCE EXPLAINED

Doug Speed* and David Balding

*UCL Genetics Institute, University College London UK*

*E-mail: *doug.speed@ucl.ac.uk*

In 2010, Yang, Visscher, et al., showed how mixed model analysis could be applied to genome-wide SNP data to estimate how much of a complex trait's variation can be attributed to common variants. Despite the method's popularity, it has also been subject to criticism, because on its own, an estimate of variance explained reveals little about a trait's genetic architecture. However, SNP-based heritability analysis has other uses aside from estimating variance explained. It allows us to examine the relationship between a causal variant's frequency and its effect size. I will illustrate how this can be used to assess the impact of selection. Heritability analysis can be used to assess the importance of gene promoter regions, and enables a test of polygeneity within and concordance between traits. I use the latter approach to demonstrate the shared aetiology of Rheumatoid Arthritis and Type 1 Diabetes. Finally, heritability analysis provides an elegant method for gene-based association testing, which for Crohn's Disease proves superior to methods relying on single-SNP-based testing. All tools can be readily performed using the softwares GCTA and LDAK.

**FISHER MEMORIAL LECTURE**

# PUTTING LIFE INTO NUMBERS: THE HIGHS AND LOWS OF COMMUNICATING STATISTICS TO THE PUBLIC

David Spiegelhalter[*1]

[1]*University of Cambridge, Cambridge, UK*

[*]E-mail: *D.Spiegelhalter@statslab.cam.ac.uk*

Statisticians do a great and valuable service in designing and analysing biomedical studies. But they have a tendency to feel it is someone else's job to take the final step in communicating the findings in a readily-accessible form. I shall argue that this should be an integral part of a professional statistician's role, and try and get cheap laughs by showing the disastrous things that have been done when numbers get in the wrong hands.

# SPARROWHAWKS AND A DECLINE IN SPARROWS: IS THERE A LINK?

Ben Swallow*[1]

[1]*University of St Andrews, St Andrews, UK*

*E-mail: *bts3@st-andrews.ac.uk*

The Eurasian Sparrowhawk *Accipeter nisus* has had a long and varied history as a predator in the UK. The mixed nature of population trends and distribution, in particular declines linked to the use of organochlorine pesticides during the 1950s and 60s, has coincided with similar varied trends in population dynamics of many of the species of birds that the Sparrowhawks are adept predators of. This (coincidental?) occurrence has led to diverse approaches and opinions amongst a wide audience of people, ranging from ecologists to the media to members of the general public, as to the effects of Sparrowhawks and their changes in population and distribution on the species of birds they prey on. We look to add to the growing literature on declines in songbirds by analysing long-term trends in Sparrowhawk abundance, to see if there is any support for a Sparrowhawk effect on a nationally declining prey species, the House Sparrow *Passer domesticus*. We analysed data from the British Trust for Ornithologys Garden Bird Feeding Survey, controlling for environmental covariates such as winter temperature and latitude/longitude, using a Bayesian framework. A Tweedie distribution, a type of exponential dispersion model, was used to link House Sparrow abundance and the covariates, allowing for the fact that the data relate to a mean of weekly counts, which were expected to consist of a discrete mass at zero and be continuous elsewhere. This presentation will discuss the modelling methodology used in the context of our data and will hope to offer conclusions on what effect, if any, Sparrowhawks are having on the declining populations of House Sparrows in the UK.

# NON-LINEAR MONOTONE REGRESSION FOR HIGH-DIMENSIONAL DATA

Linn C. Bergersen, Kukatharmini Tharmaratnam*, Ingrid K. Glad

*Department of Mathematics, University of Oslo*

*E-mail: *kukathat@math.uio.no*

In recent years, several methods are proposed to model nonlinear relationships in high- dimensional data by using spline basis functions and group penalties. We focus on the special case of nonlinearity as nonlinear monotone effects on the response, as is often a natural assumption in medicine and biology. We construct the monotone splines lasso (MS-lasso) method to estimate and select variables using monotone spline basis functions (I-splines). The additive components in the model are represented by the I-spline basis function expansions and the component selection becomes that of selecting the groups of coefficients in the I-spline basis function expansion. We use a recent procedure called cooperative lasso to select sign-coherent groups, that is selecting the groups with either non-negative or non-positive coefficients. This leads to the selection of the important covariates that have nonlinear monotone increasing or monotone decreasing effect on the response in high- dimensional regression problems. Simulated data and real data examples from genomics illustrate the effectiveness of the proposed method. Results indicate that the (adaptive) MS-lasso has excellent properties compared to the other methods both by means of estimation and selection, and can be recommended for high-dimensional monotone regression.

*Keywords*: cooperative lasso, high-dimensional data, I-splines, lasso, monotone regression, nonparametric additive models

# CHALLENGE OF ANALYSING OMICS DATA

Hae-Won Uh*

*Dept. Of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands*

*E-mail: *h.uh@lumc.nl*

The omics technologies such as genomics, proteomics and metabolomics are promising for obtaining new biological insights. The challenges to be addressed here are heterogeneous omics data obtained by different platforms and efficient statistical methods for combining genetic and omics data.
Two types of data from the Leiden Longevity Study (LLS) are considered: 2.5 million imputed genotypes and glycomics data. Five N-Glycosylation of IgG were measured using MALDI-TOF-MS, and 25 plasma N-glycan profiles using HPLC. We discuss the problems encountered in analysing and interpreting genome-wide association study (GWAS) results caused by different platforms and inadequate data pre-processing. Within long-running studies genotypes of controls are reused and cases are genotyped on more novel platforms yielding a casecontrol study unmatched for genotyping platforms. This can lead to false positive findings by extrapolating differences between arrays in the process of imputation, and stringent quality controls (QCs) are required. We developed an imputation quality measure that also reflects impact on the test statistic [1].
The second example is GWAS on IgG. Due to the heterogeneous platforms and data pre- processing, a meta-analysis could not be performed [2]. Testing for association under H0 provided an extra QC, when GWASs were performed on 25 plasma profiles. The false positive hits were caused by the combination of outliers and a plate of poor quality.
Despite a large number of SNPs the effects detected are typically small, and the challenge is how to efficiently combine multiple omics variables with genetic data. To jointly model multivariate traits and genotype data we combine two strategies. A flexible two stage approach is proposed: multiple variants in a gene are summarized and included as a covariate in the final model [3], and multivariate traits are summarized by a principal components approach based on heritability. This approach is applied to the data from LLS.

[1] Uh et al. (2011) How to deal with the early GWAS data when imputing and combining different arrays is necessary. Eur J Hum Genet 20:572-6
[2] Lauc et al. (2013) Loci associated with N-glycosylation of human immunoglobulin g show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet 9:e1003225
[3] Tsonaka et al. (2012) A two-stage mixed-effects model approach for gene-set analyses in candidate gene studies. Stat Med 31:1190-202.

# OBJECTIVE BAYESIAN SURVIVAL ANALYSIS USING SHAPE MIXTURES OF LOG-NORMAL DISTRIBUTIONS

Catalina A. Vallejos*[1], Mark F.J. Steel[1]

[1]*University of Warwick, Coventry, UK*

*E-mail: *C.A.Vallejos-Meneses@warwick.ac.uk*

Survival models such as the Weibull or log-normal lead to inference that is not robust to the presence of outliers. They also assume that all heterogeneity between individuals can be modelled through covariates. We consider the use of infinite mixtures of lifetime distributions as a solution for these two issues. This can be interpreted as the introduction of a random effect in the survival distribution. We introduce the family of Shape Mixtures of Log-Normal distributions, which covers a wide range of shapes. Bayesian inference under non-subjective priors based on the Jeffreys rule is examined and conditions for posterior propriety are established. The existence of the posterior distribution on the basis of a sample of point observations is not always guaranteed and a solution through set observations is implemented. This also accounts for censored observations. In addition, a method for outlier detection based on the mixture structure is proposed. Finally, the analysis is illustrated using a real dataset.

# ShrinkBayes: BAYESIAN DIFFERENTIAL EXPRESSION ANALYSIS OF RNA SEQUENCING DATA

Mark van de Wiel

*VU University Medical Center*

*E-mail: *mark.vdwiel@vumc.nl*

Next generation sequencing is quickly replacing microarrays as a technique to probe different molecular levels of the cell, such as DNA or mRNA. The technology has the advantage to provide higher resolution, while reducing biases, in particular at the lower end of the spectrum. mRNA sequencing (RNAseq) data consist in counts of pieces of RNA called tags. This type of data imposes new challenges for statistical analysis.

We present a Bayesian framework for differential expression analysis that al- lows for a) various count models b) flexible designs c) random effects and d) multi-parameter shrinkage. The latter is implemented using Empirical Bayes principles by several procedures that estimate hyper-parameters of (mixture) priors or non-parametric priors. The framework provides Bayesian multiplicity correction. In data-based simulations, we show that our method outperforms other popular methods (edgeR, DESeq, baySeq, NOISeq). We illustrate our approach on two data sets. The first contains 25 samples representing five regions of the human brain from seven individuals. This data motivates use of the zero-inflated negative binomial as a powerful alternative to the negative binomial, because it diminishes bias of the overdispersion parameter and improves detec- tion power for the low-count tags. The second is a large, standard two-sample RNAseq data set that we repeatedly split into a small data set and its large complement. Compared to other methods, our results from the small sample data sets validate much better on their large sample complements, illustrating the importance of multi-parameter shrinkage.

The framework is not restricted to RNAseq data nor to differential expression analysis. We currently study multivariate, graphical applications using Bayesian ridge regression, which will be discussed as well. The R software package, termed ShrinkBayes, is build upon INLA, which provides the machinery for computing marginal posteriors in a variety of models.

# CAUSAL INFERENCE WITH PSEUDO-OBSERVATIONS

Erik van Zwet

*Leiden University Medical Center, The Netherlands*

*E-mail: *vanzwet@lumc.nl*

Consider an observational study of a large cohort of patients with baseline covariates $C$, treatment $T$ and outcome $Y$. Following the arrow of time, we factorize the joint distribution of $C$, $T$ and $Y$ as $P(C,T,Y) = P(C)P(T \mid C)P(Y \mid C,T)$. In this expression, $P(T \mid C)$ represents how the treatment is decided based on the covariates. Under certain assumptions, the modified distribution

$$P^*(C,T,Y) = P(C)P^*(T \mid C)P(Y \mid C,T)$$

represents what would happen if we decided the treatment according to $P^*(T \mid C)$ instead of $P(T \mid C)$. For instance, taking $P^*(T = 1 \mid C = c) = 1$ for all $c$ would mimic the situation where treatment 1 is given to everyone. We can estimate $P(C)$ and $P(Y \mid C,T)$ from the observed data and then assemble an estimate of $P^*(C,T,Y)$.

If we want to know how the covariates affect the outcome in the observed cohort, we can choose an appropriate model and do a regression of $Y$ on $C$. In the modified situation, this is not as straightforward. Of course, we can derive estimates of $P^*(Y \mid C = c)$ from the estimate of $P^*(C,T,Y)$, but then the effects of the covariates on the outcome will not be described by simple relations with interpretable parameters. In particular, it will not be clear how to test for dependence of the outcome on a specific covariate.

Alternatively, we can use pseudo-observations as proposed by Andersen et al. (2003) in the context of multi-state models. This works as follows. We use the entire sample to estimate some parameter of interest, such as $P^*(Y = 1)$ for a binary outcome or $E^*(Y)$ for a continuous one. By repeatedly leaving out one subject and recomputing the estimate, we can gauge the effect of the covariates on the estimator. More precisely, we compute the so-called jackknife pseudo-values and use these as outcomes in a GEE to obtain estimates of the covariate effects. We also get standard errors and $p$ values.

We demonstrate our approach on data from a trial with considerable non-compliance.

**Reference:**

Andersen, P.K., Klein, J.P., and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, **90**, 15–27.

# A COMPARISON OF STATISTICAL METHODS FOR BENCHMARKING CLINICAL CENTERS IN TERMS OF QUALITY OF CARE

Machteld Varewyck*[1], Els Goetghebeur[1], Stijn Vansteelandt[1]

[1]*Ghent University, Ghent, Belgium*

*E-mail: *machteld.varewyck@ugent.be*

Inspired by a study on quality insurance of rectal cancer (Goetghebeur et al., 2011), we will discuss various statistical methods for benchmarking centers on a dichotomous quality indicator in a causal inference framework. Adjustment for differential case-mix between centers will be based on direct standardization either under a fixed or random center effects model which incorporates patient characteristics or using doubly robust estimation procedures incorporating the propensity to belong to the observed center (on the basis of patient characteristics). We discuss the relative advantages and disadvantages of the different approaches and used simulations to evaluate their performance when classifying centers of different quality. Thereby, we focus on realistic settings where some centers may contribute low numbers of patients. When aiming to provide confidential feedback to the centers, the power to detect centers with outlying low/high mortality risk is of prime concern. Yet, following shrinkage, the random effects modeling lost substantial power compared to the suggested alternatives, especially for the smallest centers. Adding the Firth correction to the fixed effects model and the outcome model for the doubly robust method recovered power, while maintaining convergence in the presence of very small centers. We apply our findings and compare the detection of low/high mortality risk outliers for the centers in the Swedish Stroke register, Riks-Stroke (http://www.riks-stroke.org).

# MIXED MODELS: MISCONCEPTIONS, PITFALLS, AND MANY OPPORTUNITIES

Geert Verbeke[*1,2]

[1]*KU Leuven, University of Leuven, Leuven, Belgium*
[2]*Hasselt University, Hasselt, Belgium*

[*]E-mail: *geert.verbeke@med.kuleuven.be*

In many contexts hierarchical, multilevel, or clustered data are collected. Examples are longitudinal studies in which subjects are measured repeatedly at various time points (measurements within subject), surveys in which all members of a sample of families are questioned (members within families), educational data in which students from various schools are tested (students within schools), etc. From a statistical perspective, the challenge is to account for the fact that the measurements within clusters are not necessarily independent, implying that standard models such as linear regression or generalized linear regression are not applicable.

Mixed models are currently amongst the most flexible models for the analysis of hierarchical data. They can be interpreted as standard linear, generalized linear, or non-linear models, with cluster-specific random effects shared by all measurements within the cluster, hereby implicitly accounting for within-cluster associations. Many commercial software packages have developed procedures for fitting mixed models, adding to their popularity. However, their similarity to standard statistical models comes with many misconceptions concerning parameter interpretation, inference, and model diagnostics. On the other hand, their flexibility for modelling associations offers many opportunities for modelling complex hierarchical and high-dimensional multivariate data structures.

In this presentation, some of the most frequently encountered misconceptions and pitfalls will be illustrated, and examples of solutions will be discussed. Furthermore, it will be shown that recent advances in this area have allowed answering complex research questions which could not be addressed previously. Emphasis will be on intuitive rationale rather than mathematical detail, and all topics discussed will be extensively illustrated using real data from various contexts.

# MODELING THE ASSOCIATION BETWEEN TWO SEASONAL TIME SERIES : A SIMULATION STUDY

Marie-Anne Vibet[*1,2,3], Didier Guillemot[1,2,4], Laurence Watier[1,2,4]

[1]*Inserm, Paris, France*
[2] *Institut Pasteur, Paris, France*
[3] *Université Paris-Sud, Le Kremlin Bicêtre, France*
[4] *Université Versailles Saint-Quentin, Versailles, France*

[*]E-mail: *marie-anne.vibet@pasteur.fr*

Upon considering the association between two time series, a fundamental problem arises when both series have in common some seasonal pattern. Indeed, this common pattern will act as a confounder and will most likely create a significant association between series. However, such an association may only reflect the seasonal pattern of a common predisposing factor. Thus, to refine the analysis, it becomes necessary to make some seasonal adjustment. Several methods, from the simplest ones to the most complex ones, have been proposed in the literature: parametric approaches, such as trigonometric functions, and semi-parametric methods, as regression spline functions. Although complex methods show a better fit, the risk of over-adjustment could reduce or mask the true association. Conversely, for simplest approaches, the risk of under-adjustment could increase the strength of the link. So far, there has not been any consensus of an optimal approach. Hence, in order to identify the optimal methodology, we compared performances of available smoothing approaches and different regression models to estimate a temporal link between two seasonal time series using extensive simulations.

We developed a new methodology to simulate Gaussian seasonal time series as met in epidemiological surveillance data. We reviewed the main seasonal patterns: sinusoidal, triangular and epidemic, and considered them all in the simulation schedule.

As expected, when no adjustment for seasonality is made, the association between time series is over-estimated; a false association may even be created. Adjusting for seasonal pattern only the outcome series severely decreases the link. On the contrary, controlling for seasonality only the input series gives suitable results. Indeed, the bias of parameter and the coverage probability become reasonable mainly when working with trigonometric functions.

As a conclusion, we would highly recommend the use of trigonometric functions to control for seasonality when estimating the association between two seasonal time series.

# INFERRING CAUSAL RELATIONSHIPS BETWEEN ASSOCIATED PHENOTYPES USING BOTH PHENOTYPIC AND QTL INFORMATION

Huange Wang*, Fred van Eeuwijk

*Biometris, Wageningen University, 6708 PB Wageningen, the Netherlands*

*E-mail: *huange.wang@wur.nl*

**Motivation**: In addition to the analysis of genotype-phenotype relationships, mapping interactions between phenotypes also provides structural insight into the functional mechanism of biological systems. Various methods have been proposed to reconstruct directed phenotype networks. A recent interesting proposal, the QTL-directed dependency graph (QDG) approach, uses QTL information on phenotypes to infer causal directions for edges in an undirected phenotype network. A prerequisite for this approach is that at least one QTL has been identified for each trait studied. In practice, however, this prerequisite is often not met due to factors such as limited sample size, weak QTL effects and measurement noise.

**Results**: We developed a general method to infer causal directions for edges in a large-scale undirected phenotype network, using both the relevant phenotypic interactions and the detected QTLs. Our method does not require QTLs for each and every trait. We evaluated and compared the performance of our method with the benchmark QDG algorithm via simulations. Results show that our method is applicable to general cases and leads to more accurate overall orientations. Finally, we illustrated our method with a real example involving metabolic and QTL data in ripe tomato fruits.

# MODELLING A REPEATED ORDERED CATEGORICAL RESPONSE WITH PENALISED SPLINES USING MCEM

Sue Welham*[1], Tu Ho[2], James Carpenter[2]

[1]*VSN International, Hemel Hempstead, UK,*
[2]*London School of Hygiene and Tropical Medicine, UK*

*E-mail: *sue.welham@vsni.co.uk*

Longitudinal ordinal data occurs frequently in practice in contexts as varied as clinical and agricultural trials, where scores are used to simultaneously summarise several aspects of patient/plant disease. The generic aim of analysis is to model and evaluate any treatment.time interaction. We use the proportional odds model, extended so that progress over time can be described using penalised splines to account for non-linear trend, and random coefficient regression accounts for correlation between the observations within each subject. In this representation, the smoothing parameter can be shown to be computationally equivalent to a variance component in the model. To avoid the use of approximations known to give bias in the variance components, we use a Monte-Carlo EM (MCEM) algorithm to obtain provisional parameter estimates. We then refine these provisional estimates using simulated maximum likelihood (SML) to obtain an estimate of the log-likelihood at the maximum.

We illustrate the method using disease data from a replicated plant variety trial, with disease assessed on all plots at several dates during the season. Disease scores may either increase or decrease during the season due to a complex infection cycle related to weather conditions. Disease progress curves therefore rarely take a form that can be described by a standard response curve, but are often modelled well by smoothing or penalised splines. The fitted curves can be used to infer differences in variety resistance at important growth stages, and to give an overall measure of resistance.

# PREDICTIVE GENOMIC SIGNATURES
# BIOMARKER DISCOVERY IN HIGH-DIMENSIONAL DATA

Wiebke Werft*, Martina Fischer, Axel Benner

*German Cancer Research Center, Heidelberg, Germany*

*E-mail: *w.werft@dkfz.de*

No treatment works the same for every patient. Few therapies will benefit all patients, and some may even cause harm. Hence, biological markers ("biomarkers") are required that can guide patient tailored therapy. Using omics technologies the challenge is to derive a predictive genomic signature from a large number of candidates.

Commonly the identification of potentially predictive biomarkers is addressed by inference of regression models including interaction terms between the (continuous) biomarkers and the treatment assignment. To derive a prediction model based on a list of potentially predictive biomarkers we propose to combine componentwise screening with a final modelling step comprising a forward stepwise selection of interactions.

To screen for predictive biomarkers we investigated several extensions to standard approaches including multivariable fractional polynomials, concordance regression, and the application of the permutation of regressor residuals test. In the modelling step grouped penalization was applied.

We used simulation studies to assess the utility of the proposed procedures. Applications to two prospectively planned, randomized clinical trials will illustrate our findings.

# NONPARAMETRIC ESTIMATION OF THE SURVIVAL FUNCTION IN SCREENING AND SURVEILLANCE CORRECTED FOR MISCLASSIFICATION

Birgit I. Witte*[1], Marianne A. Jonker[2], Johannes Berkhof[1]

[1]*VU University Medical Center, Amsterdam, The Netherlands*
[2]*VU University Amsterdam, Amsterdam, The Netherlands*

*E-mail: *B.Witte@vumc.nl*

In many longitudinal studies, including screening and surveillance studies, the primary outcome measure is the time until a certain event occurs. Typical examples of events are death and clinical manifestation of disease of which the onset times are unambiguous. However, many studies have as end-point a sub-clinical disease stage that can only be detected by a screening test. This screening test has limited sensitivity and/or specificity leading to possible misclassification of the events. As a consequence, the naive nonparametric maximum likelihood estimator (NPMLE) will be a biased estimator of the survival function.
We describe an Expectation Maximization (EM) algorithm to compute the NPMLE in case of interval censored event times with correction for misclassification. Our algorithm has a closed form solution for the combined E- and M-step and is computationally undemanding. In a simulation study, we show that the mean squared error of our estimator is in general lower than the mean squared error of the estimator that ignores misclassification. The gain is largest if the time to event is short or the sensitivity of the test is low. We illustrate the estimator with follow-up data from women treated for high-grade cervical intraepithelial neoplasia.

# SMOOTHING, RANDOM EFFECTS AND CORRELATION

Simon N. Wood

*University of Bath, BA2 7AY, UK*

*E-mail: *s.wood@bath.ac.uk*

The link between smoothers and random effects allows seamless integration of smooth terms into generalized mixed regression models, and of random effects into smooth regression models. This talk reviews how this integration can most efficiently and flexibly be achieved using reduced rank spline like smoothers. I will then consider how sparse smoothing methods can be integrated with such representations of generalized additive mixed models, in order to provide effective modelling of nuisance auto-correlation effects in large datasets.

# INTEGRATED STOPOVER MODELS - A MARK-RECAPTURE STUDY ON GREY SEALS ON THE ISLE OF MAY

Hannah Worthington[*1], Ruth King[1], Patrick Pomeroy[1], Sophie C. Smout[1], Rachel S. McCrea[2]

[1]*University of St Andrews, St Andrews, UK*
[2]*University of Kent, Canterbury, UK*

[*]E-mail: *hw233@st-andrews.ac.uk*

Many standard capture-recapture models, such as the Cormack-Jolly-Seber model, condition on the first capture of an individual. Stopover models extend this approach by removing this condition, explicitly modelling the 'arrival' of individuals into the study population and specifying an 'age'-dependence on the departure probability from the study. In other words, the probability that an individual leaves the study at any given time is a function of its residence time in the population. Unknown arrival (and departure) times of individuals into the study can be accounted for within the likelihood expression by summing over all possible arrival (and departure) times.

In this study the data is partitioned into cohorts of seal pups born each year and the arrival times correspond to the recruitment of these seals into the breeding population. It is of particular interest to determine whether the distribution of arrival times differs across cohorts. It is known from previous studies that the apparent survival of the adult seals is constant for the Isle of May colony. This permits a level of parsimony by specifying common parameters across the different cohorts. Additional complexity arises due to the nature of identifying individuals within this population. Seals are identified using two main methods: flipper tags and brands. It is known that animals identified by brands are resighted more effectively than those with tags and that tags may be lost. The model is extended to a multi-state model to explicitly model the different tagging methods and to model tag-loss. The model developed is applied to grey seal pup data from the Isle of May across a number of different cohorts.

# DEALING WITH MISSING COVARIATE DATA USING MULTIPLE IMPUTATION - A MARK-RECAPTURE-RECOVERY STUDY ON SOAY SHEEP

Hannah Worthington[*1], Ruth King[1], Stephen T. Buckland[1]

[1]*University of St Andrews, St Andrews, UK*

[*]E-mail: *hw233@st-andrews.ac.uk*

We consider a mark-recapture-recovery study where the survival probabilities are expressed as a function of individual time-varying continuous covariates. For such data, an issue arises with the collection of the covariate data for (at least) the occasions when an individual is not observed. However, in the presence of missing covariate data, the corresponding likelihood is analytically intractable. We propose a two-step multiple imputation approach to obtain estimates of the mark-recapture-recovery model parameters. In the first step, a model is fitted to only the observed covariate values. In the second step, complete data sets are imputed (i.e. all missing covariate values are imputed), conditional on the fitted covariate model. For each complete data set, a complete-data-likelihood can now be maximised. The results from the multiple imputations are then combined to obtain an overall estimate of the model parameters. Confidence intervals are obtained using a non-parametric bootstrap in order to account for the additional uncertainty of the underlying covariate model. We apply the proposed approach to a real data set relating to Soay sheep.

# PARAMETER REDUNDANCY OF MIXTURE MODELS IN CAPTURE RECAPTURE

Chen YU[*1], Byron J.T. Morgan[1], Diana Cole[1]

[1]*University of Kent, Kent, UK*

[*]E-mail: *cy52@kent.ac.uk*

The use of mixture models in statistical ecology is now well established, involving both finite and infinite mixtures and their combination. Particularly influential have been the papers of Pledger (2000) and Pledger et al (2010). The models provide a structure for model-selection, a convenient description of individual heterogeneity, and also permit a study of the robustness of models that ignore such heterogeneity when it is present.

Of interest to us is finite mixture models, which can introduce a large number of parameters that may require constraints to be applied in order for all the parameters to be estimated. This work is motivated by Pledger et al (2003, 2010), which specifies the numbers of constraints required for a range of alternative mixture models for survival in open population analysis. We provide for the first time a formal analysis of these models, using the methods of computational symbolic algebra introduced by Catchpole and Morgan (1997). This provides an illustration of the different steps that need to be carried out, using the recent developments of Cole et al (2010). This includes investigating how missing sparse data sets can affect parameter redundancy. The work is illustrated using data on possums, Trichosurus vulpecula.

References:

Catchpole, E.A. and Morgan, B.J.T. (1997) Detecting parameter redundancy *Biometrika*, **84**, 187-196.

Cole, D. J., Morgan, B.J.T. and Titterington, D. M. (2010) Determining the Parametric Structure of Non–Linear Models. *Mathematical Biosciences*, **228**, 16-30.

Pledger, S. (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434-442.

Pledger, S., Pollock, K. H. and Norris, J. L. (2003), Open Capture–Recapture Models with Heterogeneity: I. Cormack–Jolly–Seber Model. *Biometrics,* **59**, 786-794

Pledger, S., Pollock, K. H. and Norris, J.L. (2010), Open Capture–Recapture Models with Heterogeneity: II. Jolly–Seber Model. *Biometrics,* **66**, 883-890.

# QUANTIFYING TEMPORAL TURNOVER IN BIODIVERSITY, AND HOW IT VARIES SPATIALLY

Joyce Yuan*, Stephen T. Buckland, and Phil J. Harrison

*University of St Andrews, St Andrews, UK*

*E-mail: *joyce@mcs.st-and.ac.uk*

Quantifying species compositional change over time plays an important role in measuring biodiversity trends. In the literature, spatial compositional heterogeneity is often referred to as beta diversity. Temporal compositional heterogeneity is usually measured by species turnover over time using presence-absence data. However, if available, it is more informative to use species abundance estimates to measure the compositional change over time, and how it varies spatially. We use spatial-temporal data analysis to predict the density surface for each species over time, and use the predictions to quantify the rate at which individuals of one species are replaced by individuals of other species. We refer to this turnover of individuals between species as individual- based species turnover, to distinguish it from traditional species turnover measures. We introduce several indices that quantify the dissimilarity of species composition over time. We use a dataset of North Sea fish from the International Bottom Trawl Survey to illustrate the methods. We also show how temporal trends in the individual-based species turnover vary spatially, which helps to identify how climate change is affecting the North Sea fish community.

# CONSTRAINED ORDINATION ANALYSIS IN THE PRESENCE OF ZERO INFLATION

Yingjie Zhang*[1], Olivier Thas[1,2]

[1]*Ghent University, Gent, Belgium*
[2]*University of Wollongong, Wollongong, Australia*

*E-mail: *liv.zhangcn@gmail.com*

Constrained ordination analysis, with canonical correspondence analysis (CCA) as its best known method, is a class of popular techniques for analyzing species abundance studies in ecology. These methods rely on distributional assumptions on the conditional abundance distributions. For abundance observations, the Poisson and the negative binomial distributions are the most frequently considered distributions. However, many large abundance studies result in many zero abundances. This may happen because of several reasons. To name one, in microbial community ecology the abundances of a very large number of species are nowadays often obtained by means of sequencing the pooled DNA sample. Due to the small sensitivity for rare species, too many observed zeroes are to be expected. Moreover, more zeroes are expected with increasing number of species. We propose a constrained ordination method based on zero-altered count distributions (e.g., zero-inflated Poisson, hurdle models). We show how the parameters and the environmental gradients can be estimated. In simulation studies we examine the behaviour of the estimators, and we apply the method to a real data set. We conclude that in the presence of zero inflation our methods give better results than the Poisson-based approaches.