

**International Biometric Society  
Multi-Regional Conference**

**Organised by the British, Belgian,  
French and Netherlands Regions**

**Digby Hall & Stamford Hall,  
University of Leicester, UK**

**Tuesday 5 April 2005  
to  
Friday 8 April 2005**



# Programme

## Tuesday 5<sup>th</sup> April

<i>11:00 – 14.00</i>	<i>Registration and Coffee</i>	<i>Stamford Hall, ante-room</i>
<i>12:00 – 13:30</i>	<i>Lunch</i>	<i>Stamford Hall, Dining Room</i>
<b>14:00 – 14:15</b>	<b>Opening of Conference</b>	<b>Digby Hall, Dining Room</b>
	Welcome & Introduction <i>Joe Perry (IBS British Region President)</i>	
<b>14:15 – 15:15</b>	<b>Opening Keynote Session</b>	<b>Digby Hall, Dining Room</b>
	<b>Bioinformatics (Chair, Joe Perry)</b>	
	The Bioinformatics Challenge: are we up for it? <i>Wally Gilks (Medical Research Council, Cambridge, UK)</i>	
<i>15:15 – 15:45</i>	<i>Tea/Coffee</i>	<i>Digby Hall, JCR</i>
<b>15:45 – 17:15</b>	<b>Contributed Papers I</b>	
	<b>I(a) Ecology &amp; Environment</b> (Chair: Tim Sparks)	Digby Hall, Dining Room
	<b>I(b) Medical Statistics</b> (Chair: Hein Putter)	Stamford Hall, Lounge
<i>17:30 – 19:00</i>	<i>British Region Committee Meeting (Digby Hall, Room 11)</i>	
<i>18:00 -</i>	<i>Bar opens</i>	<i>Stamford Hall</i>
<i>19:00 – 20:30</i>	<i>Dinner</i>	<i>Stamford Hall, Dining Room</i>

## Wednesday 6<sup>th</sup> April

<i>07:30 – 09:00</i>	<i>Breakfast</i>	<i>Stamford Hall, Dining Room</i>
<b>09:00 – 10:30</b>	<b>Contributed Papers II</b>	
	<b>II(a) Missing Data</b> (Chair: Tim Cole)	Digby Hall, Dining Room
	<b>II(b) Multivariate Statistics</b> (Chair: Paul Eilers)	Stamford Hall, Lounge

10:30 – 11:00      *Tea/Coffee*      *Digby Hall, JCR*

Note: Authors of posters should put their posters up during this period or lunchtime. All posters should be on display in the area at the rear of the Digby Hall Dining Room by 14:00 today.

**11:00 – 12:30      Invited Session I      Digby Hall, Dining Room**

**Climate Change (Chair: Avner Bar-Hen)**

The influence of climate change on the phenology of plants and animals  
*Tim Sparks (Centre for Ecology and Hydrology, UK)*

Analysis of extremes of synthetic sea surge data  
*Jonathon Tawn (University of Lancaster, UK)*

12:30 – 14:00      *Lunch*      *Stamford Hall, Dining Room*

**14:00 – 15:30      Contributed Papers III**

**III(a) Experimental Design and Sampling**

(Chair: Sue Welham)

Digby Hall, Dining Room

**III(b) Genetics**

(Chair: Robert Curnow)

Stamford Hall, Lounge

15:30 – 16:00      *Tea/Coffee*      *Digby Hall, JCR*

**16:00 – 17:30      Special Contributed      Digby Hall, Dining Room**

***Biometrics Showcase (Chair: John Whitaker)***

**18:00 – 19:00      Poster Session/Reception      Digby Hall, Dining Room**



**We are grateful to Pfizer Global  
Research and Development for their  
sponsorship of the Poster Session**

Note: all posters authors should remain at their poster throughout the poster session. The prize(s) for the best posters will be judged by members of the Scientific Programme Committee during the poster session. Prize winners will be announced and prizes presented at the Conference Dinner.

19:30 – 21:00      *Dinner*      *Stamford Hall, Dining Room*

## Thursday 7<sup>th</sup> April

07:30 – 09:00      *Breakfast*      *Stamford Hall, Dining Room*

**09:00 – 10.30      Invited Session II      Digby Hall, Dining Room**

### **Microarrays (Chair: Peter Colman)**

Statistical models to correct for saturation effects in cDNA microarrays  
*Chris Glasbey (Biomathematics and Statistics Scotland, UK)*

Dense scatterplots  
*Paul Eilers (Leiden University Medical Centre, The Netherlands)*

10:30 – 11:00      *Tea/Coffee*      *Digby Hall, JCR*

**11:00 – 12:30      Contributed Papers IV**

**IV(a) Inference**      Digby Hall, Dining Room  
(Chair: Peter Rigsby)

**IV(b) Statistics in Agriculture and Food**      Stamford Hall, Lounge  
(Chair: Mart de Jong)

12:30 – 14:00      *Lunch*      *Stamford Hall, Dining Room*

**14:00 – 16:20      Invited Session III      Digby Hall, Dining Room**

### **Biometry in the EU (Chair: Fred van Eeuwijk)**

Modelling gene flow at the landscape level  
*Antoine Messéan (INRA, France)*

Quantification of transmission of viruses in livestock with and without  
vaccination  
*Mart de Jong (Wageningen University, the Netherlands)*

15:30 – 15:35      *Short break*

Statistical challenges for the GenomEUtwin project  
*Hein Putter (Leiden University Medical Centre, The Netherlands)*

16:20 – 16:45      *Tea/Coffee*      *Digby Hall, JCR*

**16:45 – 17:50            Contributed Papers V**

**V(a) Spatial/Temporal Statistics**            Digby Hall, Dining Room  
(Chair: Jonathon Tawn)

**V(b) Miscellaneous**                                Stamford Hall, Lounge  
(Chair: Bas Engel)

*18:00 – 19:00            Meeting of Regional Officers (Digby Hall, Room 11)*

*19:30 -                    Conference Dinner            Villiers Hall, Dining Room*

**21:00 (approx).        After Dinner Address**

From guano to Procrustes: a life of addiction to field crop experiments  
*George Dyke*

**Friday 8<sup>th</sup> April**

*07:30 – 09:00            Breakfast                        Stamford Hall, Dining Room*

*Note: Rooms must be vacated before the start of the first session. Luggage may be deposited in Stamford Hall, Luggage Room between 07:30 and 09:15.*

**09:15 – 10:45        Contributed Papers VI**

**VI(a) Microarrays**                                Digby Hall, Dining Room  
(Chair: Chris Glasbey)

**VI(b) Non-linear & Generalized Additive Models**  
(Chair: Robin Thompson)                        Stamford Hall,  
Lounge

*10:45 – 11:15            Tea/Coffee                        Digby Hall, JCR*

**11:15 – 12:15        Closing Keynote Session    Digby Hall, Dining Room**

**Biometry in Society (Chair: Joe Perry)**

Disputed Science and Government Policy  
*Robert Curnow (University of Reading, UK)*

**12:15 – 12:30        Closing Remarks            Digby Hall, Dining Room**

*Joe Perry (IBS, British Region President)*

*12:30 – 14:00            Lunch                                Stamford Hall, Dining Room*

## Sponsors and Exhibitors

### Poster Session Sponsor

We are grateful to Pfizer Global Research and Development for their generous sponsorship of the Poster Session and Reception.



### Poster Prizes

We are grateful to Hodder Arnold for their donation of copies of the 2004 volumes of three journals - Clinical Trials, Statistical Modelling, and Statistical Methods in Medical Research – and to Oxford University Press for their donation of a number of books as prizes for the authors of the best posters presented at this conference.

The prizes will be on display near to the registration desk throughout the conference.

### Manned Exhibitions

#### VSN International

5 The Waterhouse  
Waterhouse Street  
Hemel Hempstead  
HP1 1ES  
U.K.

Tel: +44 (0)1442 450233  
Fax: +44 (0) 8701215653  
Email: [info@vsn-intl.com](mailto:info@vsn-intl.com)  
<http://www.vsn-intl.com/>

#### StatSoft Ltd

21-23 Mill Street  
Bedford  
MK40 3EU

Phone: +44 (0) 1234 341226  
Fax: +44 (0) 1234 341622  
Email: [info@statsoft.co.uk](mailto:info@statsoft.co.uk)  
<http://www.statsoft.co.uk/>

## **Unmanned Exhibitions/Fliers**

### **Oxford University Press**

Great Clarendon Street  
Oxford  
OX2 6DP

<http://www.oup.co.uk/>

### **Hodder Arnold**

338 Euston Road  
London  
NW1 3BH

<http://www.hodderarnoldjournals.com/>

### **John Wiley & Sons, Ltd**

The Atrium  
Chichester  
West Sussex

<http://www.wiley.co.uk/>

### **Chapman & Hall / CRC Press**

Taylor & Francis Group  
23 Blades Court  
Deodar Road  
London  
SW15 2NU

<http://www.crcpress.com/>

# Abstracts

## Keynote and Invited Papers

### Opening Keynote – Bioinformatics

The Bioinformatics Challenge: are we up for it? – **Gilks** 1

### Invited 1 – Climate Change

The influence of climate change on the phenology of plants and animals  
– **Sparks** 1

Analysis of Extremes of Synthetic Sea Surge Data – **Tawn & Butler** 2

### Invited 2 – Microarrays

Statistical models to correct for saturation effects in cDNA microarrays  
– **Glasbey & Khondoker** 3

Dense Scatterplots - **Eilers** 3

### Invited 3 – Biometry in the EU

Modelling gene flow at the landscape level - **Messéan** 4

Quantification of transmission of viruses in livestock with and without  
vaccination – **de Jong** 6

Statistical challenges for the GenomEUtwin project - **Putter** 6

### Closing Keynote – Biometry in Society

Disputed Science and Government Policy - **Curnow** 7

## Contributed Sessions

### Special Contributed Session – *Biometrics Showcase*

Infections with varying contact rates: application to varicella – **Whitaker  
& Farrington** 8

A multiple record systems estimation method that takes observed and  
unobserved heterogeneity into account – **Stanghellini  
& van der Heijden** 8

Prentice's Approach and the Meta-analytic paradigm: A Reflection on the  
Role of Statistics in the Evaluation of Surrogate Endpoints – **Alonso,  
Molenberghs, Burzykowski, Renard, Geys, Shkedy, Tibaldi,  
Abrahantes & Buyse** 9

### I(a) Ecology and Environment

Marked point patterns for herbivore herds – **Stein & Georgiadis** 10

Statistical analysis of oilseed rape dispersion data along a road network  
– **Adamczyk, Pivard, Bouvier, Lecomte, Gouyon & Huet** 10

Mixture models for predation and damage in ecological research – **Morgan** 11

Classification of spatially dependent data: application to ecological data  
– **Bel, Bar-Hen, Allard, Laurent & Cheddadi** 12

### **I(b) Medical Statistics**

Dynamic manganese enhanced MRI signal intensity processing based on non-linear mixed modelling to study changes in neuronal activity – <b>Serroyen, Molenberghs, Verhoye, van Meir &amp; van der Linden</b>	13
Sequential analysis as an efficient test for linkage – <b>Schipper &amp; van der Tweel</b>	13
Age changes in the first four moments of the pubertal height distribution are defined by the shape of the height growth curve – <b>Cole, Pan, Cortina Borja, Sandhu, Ben-Schlomo, Davey Smith &amp; Kelly</b>	14
Investigation of factors predicting function and failure of kidney transplants – <b>McShane, Quiroga, Fuggle &amp; Darby</b>	15

### **II(a) Missing Data**

Multiple imputation in case-control studies: one or two imputation models? – <b>Boshuizen &amp; Doorduyn</b>	17
Longitudinal quality of like studies: the problem of the missing data reconsidered – <b>Post, Buijs, de Vries &amp; le Cessie</b>	18
Multiple imputation and model selection in Cox regression – <b>Schipper, Breteler &amp; Stijnen</b>	19
Model selection in multiply imputed datasets – <b>Wood, White &amp; Royston</b>	20

### **II(b) Multivariate Statistics**

The visualisation of multiplicative interaction – <b>Gower &amp; de Rooij</b>	21
A nonlinear model with latent process using multivariate longitudinal data: application to cognitive aging – <b>Proust, Jacqmin-Gadda, Taylor &amp; Commenges</b>	22
Mokken scale analysis using restricted optimization techniques – <b>van Abswoude</b>	23

### **III(a) Experimental Design and Sampling**

Designs for real experiments using fractions of 2 <sup>n</sup> factorials – <b>Mead</b>	24
Interactive experimental design using modern computer software – <b>Edmondson</b>	25
Are we really that blind? – <b>van der Meulen</b>	26
Relative efficiency of unequal versus equal cluster sizes in multicenter trials – <b>van Breukelen, Kotova, Candel &amp; Berger</b>	26

### **III(b) Genetics**

Joint estimation of Gene-Gene and Gene-Environment interaction effects For numerous loci using (Double) penalised log-likelihood – <b>Tanck, Jukema &amp; Zwinderman</b>	28
Weighted penalised logistic regression to estimate multilocus haplotype effects on dichotomous outcomes – <b>Souverein, Zwinderman &amp; Tanck</b>	29
Collision probabilities for bands in AFLPs – <b>Gort</b>	30
A non-linear mixed model for modelling QTLs related to plant development – <b>Malosetti &amp; van Eeuwijk</b>	31

#### **IV(a) Inference**

Visualizing, summarizing and comparing odds ratio structures – <b>de Rooij &amp; Anderson</b>	32
Reduced rank techniques for multi-state models – <b>Fiocco, Putter &amp; van Houwelingen</b>	32
A goodness-of-fit test for multinomial regression – <b>Goeman &amp; le Cessie</b>	33
Lower and upper bounds on the correlation between treatment and instrumental variables – <b>Martens, Pestman, de Boer, Belitser &amp; Klungel</b>	33

#### **IV(b) Statistics in Agriculture and Food**

Bayesian multiplicative model for assessor performance – <b>Nonyane &amp; Theobald</b>	35
Evaluation of diagnostic tests in the absence of a gold standard – models and applications – <b>Engel, Bouma, Buist, Swildens, van Roermund &amp; de Jong</b>	36
Probabilistic food risk assessment accounting for variability and uncertainty in both chemical exposure and toxicological effects – <b>van der Voet &amp; Slob</b>	37
A comparison of mixed model splines – <b>Welham</b>	38

#### **V(a) Spatial and Temporal Statistics**

Early detection of outbreaks of infectious diseases – <b>Heisterkamp &amp; Heijne</b>	40
Empirical Bayesian time series analysis in air pollution and health studies – <b>Heisterkamp &amp; Dekkers</b>	41
A GLMM approach to study the spatial and temporal evolution of spikes in the small intestines – <b>Faes, Aerts, Bijmens, Geys &amp; Molenberghs</b>	42

#### **V(b) Miscellaneous**

Testing log-linear models with inequality constraints: a comparison of asymptotic, bootstrap and posterior predictive p-values – <b>Galindo Garre</b>	44
A multivariate extension of Halley's method for finding the stationary points of a function – <b>Moncrieff</b>	44
How much information is lost by haplotype uncertainty in SNP case-control analysis – <b>Uh, Houwing-Duistermaat, Putter &amp; van Houwelingen</b>	45

#### **VI(a) Microarrays**

The use of background signal in transformation of cDNA microarray measurements – <b>van Sanden &amp; Burzykowski</b>	46
Graphical exploration of genomic data as biomarkers for drug activity in oncological cell lines using spectral map analysis – <b>Lin, de Bondt, Geerts, Perera &amp; Bijmens</b>	46
On the Benjamini-Hochberg method of multiple testing – <b>Ferreira &amp; Zwinderman</b>	47

## VI(b) Non-linear and Additive Models

Models for growth after bone marrow transplantation during childhood – <b>Geskus</b>	48
Analysis of circular responses using generalised additive models for location, scale and shape – <b>Cortina Borja</b>	49
The use of bivariate generalised additive models to investigate the effectiveness of cycle helmets – <b>Hewson</b>	50
Estimating incidence of HIV infection in women using serial prevalence data from antenatal clinics – <b>Sakarovitch, Alioum, Ekouevi, Msellati &amp; Dabis</b>	51

## Posters

QTL mapping in autotetraploid populations – <b>Hackett &amp; Bradshaw</b>	52
An R routine for fitting time-varying effects in Cox and Reduced Rank models – <b>Perperoglou, le Cessie &amp; van Houwelingen</b>	52
A Bayesian calibration model to predict fungal contamination levels in wheat seed based on a PCR assay – <b>Roberts, Theobald &amp; McNeil</b>	53
Performance of class prediction methods in a microarray setting – <b>van Sanden, Lin &amp; Burzykowski</b>	54
Optimization of sampling schemes for vegetation mapping using fuzzy classification – <b>Stein, Tapia &amp; Bijker</b>	55
Modelling SAGE data with Poisson mixtures – <b>Thygesen &amp; Zwinderman</b>	55
A new test statistics to deal with multiple testing in association between a disease and a multi-allelic marker – <b>el Galta, Stijnen &amp; Houwing-Duistermaat</b>	57
Analysis of a large family study ascertained through hypertensive probands – <b>Avery &amp; Keavney</b>	58
Estimating hidden population sizes in the presence of covariates and prior information – <b>King, Bird, Brooks, Hutchinson &amp; Hay</b>	58
Long-term spatio-temporal variation in abundance of the arden tiger moth ( <i>Arctia caja</i> ) during a population decline – <b>Conrad, Woiwod &amp; Perry</b>	59
MiCoSPA: Microbial Pest Control for Sustainable Per-urban/urban Agriculture in Latin America – <b>Riley</b>	60
Estimation of the false discovery rate in functional genomic studies – <b>Dalmaso &amp; Broët</b>	60
Analysis of microarray data in a dose-response setting: resampling based multiple testing – <b>Bijnens, Shkedy &amp; Lin</b>	62

## **Opening Keynote: Bioinformatics (Tuesday 5<sup>th</sup> April)**

### **The Bioinformatics Challenge: are we up for it?**

**Wally Gilks**

Email: wally.gilks@mrc-bsu.cam.ac.uk

*Medical Research Council, Cambridge, UK*

Bioinformatics is the computational arm of genomics. In its broadest sense, genomics is the study of the genome (the DNA of cells), and of the biology related to it. The rapid expansion of genome sequence databases, including that of the whole human genome, and large databases of results from high-throughput experiments involving DNA, RNA, proteins and other biomolecules, have propelled bioinformatics into a front-line area of research. Now, most biological research involves frequent use of these huge databases, and their associated bioinformatic tools.

Genomics and genetics are related fields, with no clear separation. Nevertheless, it can be said that genetics approaches the study of biology through observing and manipulating genome-level differences between individuals within a species, whilst genomics focuses increasingly on finding similarities or patterns within and between the genomes of different species. Statisticians have played a pivotal role in the development of genetics, but are not seen as major players in the development of genomics.

I will review some of the main areas of bioinformatics research, and the contribution that statisticians have so far made. I will argue that the status of the statistician in the front line of biomedical research depends crucially on our greater involvement in bioinformatics, but that we face considerable difficulties in meeting this challenge.

## **Invited Session 1: Climate Change (Wednesday 6<sup>th</sup> April)**

### **The influence of climate change on the phenology of plants and animals**

**Tim Sparks**

Email: ths@ceh.ac.uk

*Centre for Ecology and Hydrology, UK*

The temperature of the Earth has already increased, but by much less than is predicted over the coming century. Already this increase is having an effect on flora and fauna, through changes to population size, to geographic distribution and to the timing of life cycle events. The study of the latter, termed phenology, has a long history and a wealth of observational data exist. Change in phenology has proved to be much easier to detect

than many other aspects of a species' ecology and is relatively easy to record. The speaker has been involved both in the establishment and running of a large scale public participation phenological network ([www.phenology.org.uk](http://www.phenology.org.uk)) and in identifying and analysing long-term data sets. This talk will discuss the various sources of data and their analysis and interpretation.

### **Analysis of Extremes of Synthetic Sea Surge Data**

**Jonathan Tawn** and Adam Butler

Email: [j.tawn@lancaster.ac.uk](mailto:j.tawn@lancaster.ac.uk)

*University of Lancaster, UK*

Hydrodynamical models are routinely used by oceanographers, both for short-term operational forecasting of sea levels and currents and for the assessment of long-term trends in oceanographic characteristics. The extremal behaviour of the model outputs is poorly understood as empirical summaries of the model output do not provide a clear picture due to the sparse data they contain on extreme events. Consequently a statistical analysis of the model output is required for understanding about extremal behaviour.

In this talk we treat the output from a hydrodynamical model as data, and use statistical methods from extreme value theory to investigate changes in extreme sea surge levels in the North Sea during the period 1955-2000. We focus upon meteorological induced sea surge levels, since these are of critical importance when assessing coastal flood risk, but we will also consider the effects of tide-surge interaction which is important in shallow water areas.

Any statistical analysis must take proper account of the high level of spatial and temporal dependence between neighbouring sites and time points. We begin by presenting a univariate site-by-site analysis of temporal trends in declustered extreme surge levels. We identify trends in extreme surges which are complex in their time dependence so that we use nonparametric statistical models for trends instead of standard parametric statistical models. The uncertainty in trends in extreme surges at a single site are large, but can be reduced if we are prepared to assume that the extremal features at nearby sites are likely to be similar. We use an adaptation of the local likelihood methodology to fit a nonparametric model for spatio-temporal extremes, and so to identify regions and periods within which the extremal properties of the surge process are changing. Bootstrap procedures are used to assess the statistical significance of any changes found.

We find that there is a clear spatial structure to temporal trends in extreme levels in the North Sea over the period considered, and that by exploiting the spatial dependence of the process it enables us to make improved estimation of the all features of the extremal process of sea surges in this region.

## **Invited Session 2: Microarrays (Wednesday 6<sup>th</sup> April)**

### **Statistical models to correct for saturation effects in cDNA microarrays**

**Chris Glasbey** and Mizanur Khondoker

Email: [chris@bioss.sari.ac.uk](mailto:chris@bioss.sari.ac.uk)

*Biomathematics and Statistics Scotland, UK*

Digital images obtained by the laser scanning of cDNA microarrays often include saturated pixel values. These arise when scan settings are sufficiently high that some pixels exceed the software limit of 65535. Failure to take this censoring into account leads to biased estimates of gene expression levels. We consider two statistical models that correct for these saturation effects.

Our first model is applicable when we have access to the digital image of a microarray. To impute censored values, we propose a linear model based on the principal components of uncensored spots on the same array. Parameters are estimated by penalised least squares. The method is computationally fast, flexible to adapt to distinctive spot shapes on different arrays, and effective in correcting for the bias.

An alternative model is proposed when we do not have access to individual pixel values, but instead have mean spot values for multiple laser scans at different settings. A functional regression model is used, based on a nonlinear relationship with both additive and multiplicative error terms. A robust M-estimator is used to fit the model, which is able to estimate gene expressions, taking account of random outliers and the systematic bias caused by signal censoring of highly expressed genes. Simulation studies and applications to experimental data both give good results.

### **Dense Scatterplots**

**Paul Eilers**

Email: [P.Eilers@lumc.nl](mailto:P.Eilers@lumc.nl)

*Leiden University Medical Centre, The Netherlands*

Perhaps the most visible products of microarray technology are scatterplots, filled with (tens of) thousands of dots. They lead to large PDF files and take ages to print. In most cases they are completely black in the center and tend to draw too much attention to extreme observations. The visual impression is also very dependent on symbol size and may be quite different on screen or on paper.

A scatterplot represents observations of a two-dimensional density, so it might be a good idea to treat it as such, by computing and smoothing a two-dimensional histogram and presenting this in gray scale or color. This not-so-revolutionary idea works well,

especially if smoothing is fast and flexible. I will present an implementation based on penalized likelihood.

Trends in scatterplots are also important. In the microarray world they are used to eliminate artefacts. LOESS, a local least squares smoother, is popular but quite slow. I will discuss a very fast alternative, again using penalties.

Although this work was motivated by microarrays, smoothed scatterplots turn out to be useful in many other places. They can also be enhanced in several ways. Some examples will show this.

### **Invited Session 3: Biometry in the EU (Thursday 7<sup>th</sup> April)**

#### **Modelling gene flow at the landscape level**

**Antoine Messéan**

Email: [messean@grignon.inra.fr](mailto:messean@grignon.inra.fr)

*INRA, France*

Although gene flow is a common phenomenon for crop species, its implications for Genetically Modified Plants have raised new concerns. Undesirable effects related to gene flow result in **ecological** or **agronomic** considerations (persistence of resistant volunteers; creation of new weeds; multiple resistance) as well as **in commercial** considerations (unintended presence of GMOs in conventional production affecting its competitiveness in the marketplace).

Many results are now available for these concerns. From the available results, it can be stressed that, for example, herbicide tolerant rapeseed cannot be cultivated without applying specific guidelines for crop management. Those crop management guidelines should achieve two main objectives:

- 1) the development or extension of practices aiming at reducing, in time and space, the persistence of undesirable plants (volunteers and hybrids with wild relatives);
- 2) the avoidance of selection pressure on these undesirable plants (a major issue for herbicide tolerance).

Nevertheless, for forecasting the spread and behaviour of transgenes and their impacts in a wide range of agro-ecosystems as well as for designing monitoring tools, modelling is a key element. Models help in:

- structuring knowledge, identifying gaps and reducing the research fragmentation;
- ranking farming systems according to gene flow behaviour;
- forecasting the behaviour of transgenes in cultivated and non-cultivated lands;
- testing *a priori* the efficiency of mitigation measures or regulation schemes;
- implementing monitoring schemes by identifying high risk situations;

- re-assessing the overall balance of the impacts of GM crops when new results are available (from trials as well as from monitoring).

Modelling for forecasting the behaviour of transgenes has been in development for some years. It has been focused mainly on crop-to-crop gene flow and six models have been published so far. However, only two of them, GENESYS<sup>®</sup> for rapeseed (Colbach *et al*, 2001a & b) and MAPOD<sup>®</sup> for corn (Angevin *et al*, 2001), actually take into account the spatial patterns of landscapes and are able to forecast the behaviour of transgenes within the landscape. GENESYS<sup>®</sup> takes into account crop rotations as well seed persistence. An adaptation of GENESYS<sup>®</sup> for sugar beet is under progress and validation over a wide range of available data is being carrying out.

Models have been used to underpin the co-existence studies carried out by INRA in France (Relevance and feasibility of non-GM chain, 2001) and by JRC/IPTS study (Scenarios for co-existence, 2002). Results from these co-existence studies have raised several issues that research should address.

- There exists a wide range of farming systems within Europe that could not be addressed through specific studies. How should we represent or take into account this regional variability when assessing ecological and economical balances or designing regulation rules?
- The landscape fragmentation has a great influence on gene flow and ecological impacts and its effect should be taken into account in modelling.
- Induced costs due to indirect effects of co-existence rules are difficult to estimate and are highly dependent on the local regional variability of landscapes and on agricultural farming systems.
- Available models for gene flow and ecological impacts focus mainly on the field level or on a small region (group of fields). However, mitigation measures and monitoring schemes should involve at least three different decision levels: the field level with crop management practices, the “cropping system” within the farming systems strategy, the landscape or the regional level. Up-scaling of models at different biogeographical levels should thus be made possible and easy to handle.
- Models should be made more generic in order to apply to a wide range of crops, especially those crops forthcoming and they should be more dynamic so that new impacts can be forecasted while keeping the basic gene flow structure.

All these elements lead to a key bottleneck: addressing the various ecological and economical factors and processes linked to GM impacts at the same common landscape level. To cope with these issues, several ongoing studies or programs have been launched over the last years but remain fragmented. A special emphasis has to be given to the integration of current knowledge and the development of generic methods in order to set up a science-based framework, strategies, methods and tools for assessing ecological and economical balances of GM crops and for an effective management of their development within European farming systems.

## **Quantification of transmission of viruses in livestock with and without vaccination**

**Mart de Jong**

Email: Mart.dejong@wur.nl

*Wageningen University, The Netherlands*

European policy making with regard to infectious animal diseases should ideally be based on quantitative knowledge on the effects of prevention and control measures. There are, however, very many aspects of infectious animal disease policy to be studied, for example: the characteristics of diagnostic tests including the sampling schemes used, the effects of vaccines, the effects of regionalisation, etc. Moreover, it is often far from trivial to determine how to obtain quantitative effect measures. Consequently policy making is often based on “expert” appraisal of effects.

However, especially for novel methods and for effects that have complex manifestation at the population level expert estimates of effects are not reliable. Thus more effort should be put in obtaining suitable effect measures. To arrive at the right measures a multidisciplinary approach is needed: insights from biometry, biology, and social sciences for the human behavioural side need to be combined.

As an example one aspect of infectious disease policy is discussed here: the use of vaccines to stop transmission of disease agents. Assessment of vaccines was up to now based on measuring the clinical protection vaccination offers to individual animals exposed to the infectious agent. Expert opinion about vaccine effects based on clinical protection is misleading as these vaccines may not protect against transmission even when they give clinical protection or the vaccine does stop transmission although clinical protection is not complete. Moreover, clinical protection experiments only allow qualitative assessment of effects in a highly artificial exposure situation and thus extrapolation is not possible. We developed methods to assess quantitatively the effects of vaccines on transmission. The stochastic outcome of a transmission process in an experimental group of animals was modelled,. Thus the probability distribution for the number of contact infected animals is obtained and that distribution can be used to do statistical inference on these experiments. Examples will be given from experimental outcomes, of the statistical analysis and of the insights obtained.

## **Statistical challenges for the GenomEUtwin project**

**Hein Putter**

Email: H.Putter@lumc.nl

*Leiden University Medical Centre, The Netherlands)*

The aim of the GenomEUtwin project is the localization of genes responsible for common diseases. Such diseases or traits are called complex traits because there are many genes influencing the trait, possibly interacting with each other and with the environment. The effect of any single gene is likely to be quite modest, and

consequently enormous sample sizes are needed to find them. Twins are very useful for genetic studies because they are intrinsically matched for age and to a large extent for environment, and because the presence of monozygotic and dizygotic twins makes it possible to distinguish between genetic and shared environmental factors. The GenomEUtwin project is a collection of twin registries throughout Europe and Australia. The combined twin registries contain a total of 600,000 twins, not all of which have information on the traits to be studied. I will discuss a number of statistical issues that are specific to this project, including selection of most informative twins for linkage, and meta-analysis for genome scans.

## **Closing Keynote: Biometry in Society (Friday 8<sup>th</sup> April)**

### **Disputed Science and Government Policy**

**Robert Curnow**

Email: r.n.curnow@reading.ac.uk

*University of Reading, UK*

Statisticians are becoming increasingly involved in debates about matters of public concern where the science is disputed. I will suggest that this is due to recent advances in statistical methods and the ability with modern computing facilities to use more realistic and, therefore, respected models. Also, and this has some negative consequences, the increasing specialisation in our profession results in biometricians having greater knowledge of particular areas of application and so being able to contribute more effectively to scientific debates. Drawing on my experience with the Bristol Royal Infirmary Inquiry and Government Committees concerned with tobacco and health and with BSE and vCJD, I will discuss the way in which the statistical aspects were handled with implications for the role of individuals and learned societies. I hope that colleagues from all four Regions will contribute their own experiences.

## **Special Contributed Session**

### ***Biometrics Showcase***

#### **Infections with varying contact rates: application to varicella**

**Heather Whitaker** and Paddy Farrington

Email: H.J.Whitaker@open.ac.uk

*The Open University, Milton Keynes, UK*

Over the last 30 years the incidence of varicella (chickenpox) has increased among pre-school age children in the UK. It has been suggested that this is due to increased contact rates because of wider nursery school provision.

A common assumption in models for infectious disease data is that the infection rates, and contact rates are stationary over time. We extend these methods to allow contact rates to vary slowly over time by assuming that the infection is in approximate equilibrium over a one year period. This requires the numerical solution of an integral equation in several dimensions.

These methods are applied to UK serological survey and case reports data on varicella using likelihood methods. Four models for contact rates were fitted. When contact rates are stationary, the models are usually unidentifiable. However, when contact rates change, more information becomes available. We were thus able to evaluate the fit of each model, and validate our findings using an independent data set. We conclude that the data are consistent with a substantial increase in contact rates in pre-school aged children over the last 30 years.

#### **A multiple record systems estimation method that takes observed and unobserved heterogeneity into account**

Elena Stanghellini<sup>1</sup> and **Peter G.M. van der Heijden**<sup>2</sup>

Email: p.vanderheijden@fss.uu.nl

<sup>1</sup>*Dipartimento di Scienze Statistiche, Universita di Perugia, Italy*

<sup>2</sup>*Department of Methodology and Statistics, Utrecht University, The Netherlands*

We present a model to estimate the size of an unknown population from a number of lists that applies when the assumptions of (a) homogeneity of capture probabilities of individuals and (b) marginal independence of lists are violated. This situation typically occurs in epidemiological studies, where the heterogeneity of individuals is severe and researchers cannot control for independence between sources of ascertainment. We discuss the situation when categorical covariates are available and the interest is not only in the total undercount, but also in the undercount within each stratum resulting

from the cross classification of the covariates. We also present several techniques for determining confidence intervals of the undercount within each stratum using the profile log likelihood, thereby extending the work of Cormack (1992).

**Prentice's Approach and the Meta-analytic paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints**

**Ariel Alonso**<sup>1</sup>, Geert Molenberghs<sup>1</sup>, Tomasz Burzykowski<sup>1</sup>, Didier Renard<sup>1</sup>, Helena Geys<sup>1</sup>, Ziv Shkedy<sup>1</sup>, Fabián Tibaldi<sup>1</sup>, José Cortiñas Abrahantes<sup>1</sup> and Marc Buyse<sup>2</sup>

Email: ariel.alonso@luc.ac.be

<sup>1</sup>*Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium*

<sup>2</sup>*International Drug Development Institute (IDDI), Brussels, Belgium*

The very mention of surrogate endpoints has always been very controversial. This may be due to a number of well-known unfortunate historical events. Thus, while many would like to avoid surrogate endpoints altogether, sometimes surrogates will be the only reasonable alternative, especially when the true endpoint is rare and/or distant in time. It is then best to use *validated* surrogates, but one clearly needs to reflect on the very meaning of validation. We put a perspective on the strengths and limitations of statistical methods for the evaluation of surrogate endpoints. Whereas using several trials overcomes some of the limitations of a single-trial framework (Prentice, 1989), arguably the evaluation of surrogate endpoints can never be done using only statistical evidence but such evidence should be seen as but one component in a decision making process that involves, among others, a number of clinical and biological considerations. We briefly present a hierarchical framework that incorporates ideas from Prentice's work and is uniformly applicable to different types of surrogate and true clinical outcomes.

## **I(a) – Ecology and Environment**

### **Marked point patterns for herbivore herds**

**Alfred Stein** and Nick Georgiadis

Email: Alfred.Stein@wur.nl

*Biometris, Wageningen University, PO Box 100, 6700 AC Wageningen, The Netherlands*  
*Mpala Research Centre, PO Box 555, Nanyuki, Kenya*

Quantitative descriptions of animal species' distributions at the ecosystem level are rare. In this study we used marked spatial point pattern analysis to characterize herd spatial distributions of several species comprising a savanna large herbivore community in Laikipia, central Kenya. Points are the herd centers, marks are the herd sizes. Previous research identified possible discrepancies between prey and non-prey species on the basis of the nearest neighbour distance function. In this paper we make a similar distinction and analyze possible consequences. Analysis concentrated on Ripley's K-function on several data subsets. A digitized boundary of the area has been included. The herd patterns of Thomson gazelle was modelled with a Strauss marked point process, showing a single mode, whereas the herd pattern of the plains zebra showed multiple modes. This can be well explained by the ecosystem behavior (habitat specialist versus habitat generalist) of the two species.

### **Statistical analysis of oilseed rape dispersion data along a road network**

**Katarzyna Adamczyk**<sup>1</sup>, Sandrine Pivard<sup>2</sup>, Annie Bouvier<sup>1</sup>, Jane Lecomte<sup>2</sup>, Pierre-Henry Gouyon<sup>2</sup> and Sylvie Huet<sup>1</sup>

Email: Katarzyna.Adamczyk@jouy.inra.fr

<sup>1</sup>*INRA, MIA Department, Jouy-en-Josas*

<sup>2</sup>*Laboratory of Ecology, Systematic and Evolution, University Paris-Sud*

The release of genetically modified oilseed rape may involve some undesirable effects for the environment. Herbicide resistant cultivars may transfer the resistance gene to the conventional crops or some weedy species. The risk of transgene spread is amplified by the presence of abundant feral populations of oilseed rape growing on the road verges. In order to study the origin and the dynamics of these populations a ground survey has been conducted in an agricultural region of winter oilseed rape production in the centre of France each year from 2000 to 2003. The oilseed rape fields and the feral populations were located by a GPS system on about a 100 km long road network and their characteristics were taken.

We present the statistical methods used to analyse the results of this ecological survey. The set of the data contains the information about the factors likely to affect the feral population occurrence and persistence. Because of the large scale of the survey some possibly important variables are not available, like transport intensity or weeding treatment. The purpose of the analysis is to model the probability of feral population occurrence in 2003 conditionally on the explanatory variables taking into account spatio-temporal origin of the feral population and dealing with the unknown sources of the variability. We begin with the exploratory analysis using the Random Forest algorithm for ranking the variables. Then we propose the Mixed Effect Logistic Model to describe the relationship between the probability of feral plant occurrence and the explanatory variables.

*Keywords:* oilseed rape, risk assessment, Random Forest algorithm, Mixed Effect Logistic Model

#### *References*

- [1] Pessel, F. D., Lecomte, J., Emeriau, V., Krouti, M., Messan, A., Gouyon, P-H. (2001) Persistence of oilseed rape in natural habitats : consequence for release of transgenic crops. *Theoretical and Applied Genetics* 102, 841-846.
- [2] Pinheiro, J.C. , Bates, D.M. (2000) Mixed-effects model in S and S-Plus. *Statistics and computing*. Springer-Verlag.
- [3] Breiman, L. (2001) Random Forests. *Machine Learning* 45, 5-32.

### **Mixture models for Predation and Damage in Ecological Research**

**Dr Geoff Morgan**

Email: [geoff.morgan@forestry.gsi.gov.uk](mailto:geoff.morgan@forestry.gsi.gov.uk)

*Forest Research, Alice Holt Lodge, Farnham, Surrey, UK GU10 4LH*

There are many situations occurring in, for example, forestry research in which the data can come from two or more classes with each class having a different distribution that may be related to covariates. An example is the predation of seeds by mice, when some plots have a high risk of predation and others a low risk. The high-risk plots may have a different relationship to the covariates than do the low risk plot, but whether a plot is high or low risk may not be known. Other examples are the susceptibility of plants to attack by a fungus or by chemicals. The distribution of damage in high-susceptible and low-susceptible cases may be too different to allow a common distribution model to be assumed.

This paper will consider some examples and show how models can be fitted using both Bayesian and likelihood based methods. Models may allow for different distributions and covariate models depending on a known or unknown variable.

**Classification of spatially dependant data : application to ecological data**

L. Bel<sup>1</sup>, **A. Bar-Hen**<sup>2</sup>, D. Allard<sup>3</sup>, J.M. Laurent<sup>4</sup>, R. Cheddadi<sup>4</sup>

Email: avner@inapg.fr

<sup>1</sup>*Probabilités, Statistique et Modélisation, Université Paris-Sud, Orsay, France*

<sup>2</sup>*Institut National d'Agronomie, Paris, France*

<sup>3</sup>*Unité de Biométrie, Institut National de la Recherche Agronomique, Avignon, France*

<sup>4</sup>*Institut des Sciences de l'Evolution, CNRS and Université Montpellier II, France*

In environmental and ecological studies, samples have spatial coordinates and are often very irregularly located and sometimes strongly clustered. These data generally exhibit strong spatial dependence due to the physical and/or biological processes driving the phenomenon under study. Statistical studies need to take into account this dependence. This is the scope of spatial statistic and more precisely of geostatistics when continuous variables on a continuous domain are under study.

In this work, we are concerned with a supervised classification problem. Most algorithms are written for independent data. When used for dependent data while ignoring this dependence, the classification rule gives too much importance to redundant data. This might lead to unacceptable error rates when the sampling design is very clustered.

We focus on the CART (Classification and Regression Tree) algorithm and we propose an extension of this algorithm for spatially dependent data: the split criterion (based on a Gini Index to minimize) and the loss function used for pruning the tree are estimated using a kriging technique that explicitly takes into account the estimated dependence of the data.

On simulated data we show how this method improves the classification rates. We then apply it to ecological data exhibiting very strong spatially clustered samples.

## I(b) – Medical Statistics

### Dynamic Manganese Enhanced MRI Signal Intensity Processing Based on Non-linear Mixed Modeling to Study Changes in Neuronal Activity

Jan Serroyen<sup>1</sup>, Geert Molenberghs<sup>1</sup>, Marleen Verhoye<sup>2</sup>,  
Vincent Van Meir<sup>2</sup>, and Annemie Van der Linden<sup>2</sup>

Email: jan.serroyen@luc.ac.be

<sup>1</sup> *Limburgs Universitair Centrum, Center for Statistics, Universitaire Campus, Diepenbeek, Belgium;*

<sup>2</sup> *University of Antwerp, Bio-Imaging Lab, Middelheimcampus, Antwerp, Belgium*

We analyze data on the impact of testosterone on the dynamics of Mn<sup>2+</sup> accumulation measured by magnetic resonance imaging in three songbird brain areas: the nucleus robustus arcopallii (RA), area X, and the high vocal center (HVC). Birds with and without testosterone were included in the experiment, and repeated measurements were available in both a pre and post drug administration period. We formulate a non-linear modeling strategy, allowing for the incorporation of (1) within-bird correlation, (2) the non-linearity of the profiles, and (3) the effect of treatment. For two of the outcomes (RA and area X), biological theory suggests a parametric form, while for HVC this is not the case. Since the HVC outcome bears some resemblance with the two-compartment model known from pharmacokinetics, this model was considered a sensible choice. We use a different model, based on fractional polynomials, as a sensitivity analysis for the latter. All methods used provide good fits to the data, confirm results from previous, simple analyses undertaken in the literature, but were able to detect additional effects of treatment that had so far gone undetected. The fractional polynomial and two-compartment models provide similar substantive conclusions, the two together can be seen as a form of sensitivity analysis.

*Key Words:* Fractional Polynomial; Pharmacokinetics; Random Effect; Testosterone; Two-compartment Model.

### Sequential analysis as an efficient test for linkage

Maria Schipper and Ingeborg van der Tweel

Email: i.vandertweel@bio.uu.nl

*Centre for Biostatistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, the Netherlands*

Spielman *et al* (1993) introduced the Transmission/Disequilibrium Test (TDT) as a method to test for linkage between a genetic marker and a disease when population association has been found. Using data from trios consisting of both parents and one

affected child, the (non)transmission of a marker allele to the child is evaluated. This test is a McNemar test because data within one trio are paired.

Wittkowski and Liu (2002) presented the Stratified McNemar test (SMN) as “a statistically valid alternative to the TDT”. The two tests differ in the way the allele transmission from two heterozygous parents is handled. This difference shows in the variance of the test statistic under the null hypothesis.

In this study, we investigate the statistical properties of the sequential counterparts of both the TDT and the SMN. On average, sequential tests require less observations to come to a decision than their fixed sample size alternatives. Especially for rare recessive diseases this might be beneficial, because the number of families needed for a powerful linkage test can become large.

We will compare the results for the sequential TDT and SMN and discuss differences between them.

#### *References*

- Spielman RS, RE McGinnis, WJ Ewens. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506-516.
- Wittkowski KM, X Liu. (2002) A statistically valid alternative to the TDT. *Human Heredity* 54:157-164.

### **Age changes in the first four moments of the pubertal height distribution are defined by the shape of the height growth curve**

**TJ Cole**<sup>1</sup>, H Pan<sup>1</sup>, M Cortina Borja<sup>1</sup>, J Sandhu<sup>2</sup>, Y Ben-Shlomo<sup>2</sup>, G Davey Smith<sup>2</sup>, FP Kelly<sup>3</sup>

Email: tim.cole@ich.ucl.ac.uk

<sup>1</sup>*Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, UCL*

<sup>2</sup>*Department of Social Medicine, University of Bristol*

<sup>3</sup>*Statistical Laboratory, University of Cambridge*

**Background** Height at puberty has interested statisticians for over 100 years. In individuals it accelerates then decelerates, so height velocity peaks then falls to zero, the timing of the peak varying from person to person. Treated cross-sectionally height is usually assumed to be Normally distributed throughout puberty, though in sufficiently large samples it is transiently skew. Why should this be?

**Aim** To establish how and why skewness and kurtosis in the frequency distribution of height change with age during puberty in boys.

**Methods** 3297 Christ's Hospital School boys born 1927-56 had their heights measured by the School's medical officer 9-10 times per year from 9-20 y (median 42 times). Treating the data (n = 131,710) as independent and cross-sectional, the first four moments of height (mean, standard deviation (SD), skewness and kurtosis) were

calculated in 40 narrow age groups and plotted as functions of age. A simple growth process was developed:

$$H(t) = \alpha + h(t + \varepsilon)$$

with mean growth curve  $h(t)$  offset by subject random effects  $\alpha$  and  $\varepsilon$  reflecting variability in height and timing at the age at peak height velocity (PHV). To validate this growth process the data were reanalysed longitudinally, fitting cubic spline curves to each individual.

**Results** Mean height increased over time, rising most steeply at 14 y. The SD increased until 14 y then decreased again. Skewness increased, decreased steeply then increased. Kurtosis was largely absent except briefly around 14 y when the distribution was appreciably platykurtic. Thus the SD, skewness and kurtosis curves had respectively one, two and three turning points, and surprisingly, corresponded closely in shape to the first, second and third derivatives of the mean height curve. The growth process was shown to predict such a pattern, which meant that each moment curve had a turning point providing an independent estimate of the mean age at PHV. The four estimates, ranging from 14.1 y to 14.4 y, were similar to the mean of 14.4 y calculated from individual height curves.

**Conclusion** Both skewness and kurtosis change cyclically during puberty, consistent mathematically with a growth process of varying time of onset. This generalises the observations of Boas 1 and Merrell 2 that puberty induces puberty and flattens the population mean curve. It also allows longitudinal information such as the individual mean growth curve and mean age at PHV to be estimated from age-related cross-sectional moments.

#### *References*

1. Boas F. The growth of children. *Science* 1892;19:256-7&81-2.
2. Merrell M. The relationship of individual growth to average growth. *Hum Biol* 1931;3:37-70.

### **Investigation of factors predicting function and failure of kidney transplants**

<sup>1</sup>McShane P., <sup>2</sup>Quiroga I., <sup>2</sup>Fuggle S.V. and <sup>2</sup>Darby C.

Email: philip.mcshane@nds.ox.ac.uk

<sup>1</sup>*Nuffield Dept of Surgery, John Radcliffe Hospital, Headington, Oxford, U.K.*

<sup>2</sup>*Oxford Transplant Centre, Churchill Hospital, Headington, Oxford, U.K.*

**Introduction** Kidney transplantation is an important treatment for end-stage renal failure, with about 1700 transplants being performed each year in the U.K. The majority of these are from brain-stem-dead ‘cadaveric’ donors. Despite advances in management, and the introduction of new immunosuppressive drugs, graft failure continues to be a problem. The factors predicting this failure have been the subject of considerable analysis and some disagreement. Here we investigate factors predicting failure and primary non-function of cadaveric renal transplants performed in the Oxford

Transplant Centre. One area of disagreement is whether the effect of various factors is explained better by continuous models, or whether the data support ‘cut- offs’, and we seek to examine this.

**Methods** 518 transplants performed between 1991 and 1999 were included in the analysis (later cases are to be included in subsequent analysis). A considerable number of donor and process variables were recorded. Outcomes included the presence of ‘delayed graft function’ and graft loss. SPSS (v12 for Windows) was used for statistical analysis, including logistic and proportional hazards regression.

**Results** The dominant factor predicting graft failure was ‘delayed graft function’ (dgf) (a failure of the graft to work quickly, leading to a period of dialysis after transplantation). Since this is of some importance itself, requiring a longer period of hospital stay, we then searched for factors predicting this. Of the variables analysed, 4 proved significant in logistic regression: cold ischaemia time (‘cit’: the period between removal of organ from donor and placing in recipient), donor age, sex match (1= F to M. 0 otherwise) , and donor creatinine (a measure of the function of the organ before removal) .

**Logistic regression results**

	B	S.E.	Wald	df	Sig.	Exp(B)
donor age	.028	.007	16.180	1	.000	1.029
creatinine	.005	.002	8.279	1	.004	1.005
cit (hr)	.069	.013	29.998	1	.000	1.071
F to M	.919	.235	15.310	1	.000	2.506
Constant	-4.590	.531	74.725	1	.000	.010

The adequacy of the model was examined in various ways. The ‘Hosmer Lemeshow chisquare statistic produced a value of 6.4 with 8 d.f. Addition of quadratic terms in donor age and cit did not produce a significant improvement in fit, expressed as  $-2\ln L$  or other ways. Since a ‘cut-off’ for cit as high as 36 hours has been suggested, we split the data into those where it is greater than 29 hrs or otherwise and repeated the analysis: the effect of CIT was significant in both groups.

We conclude that there is no reason to reject the logistic model identified.

We have also used the scores estimated by SPSS to construct an ‘ROC curve’ for the prediction of dgf. The area under the curve was 0.726 (95 % CI = 0.675-0.775).

SPSS ‘Classification table’ shows that most cases of dgf occur in the group where it is not predicted. This suggests that there may be other factors contributing to the development of dgf.

## II(a) – Missing Data

### Multiple imputation in case-control studies: one or two imputation models?

Hendriek C. Boshuizen and Yvonne Doorduyn

E-mail: hendriek.boshuizen@rivm.nl

*National Institute of Public Health and the Environment, PO Box 1, 3720 BA Bilthoven,  
the Netherlands*

In case-control studies, recall bias can play an important role. A less frequently discussed aspect is that due to differences in recall, the amount of missing information might also differ between cases and controls. We were confronted with this issue in a large case-control study on gastroenteritis caused by campylobacter and salmonella, looking at food consumption, contacts with animals, and kitchen hygiene. Cases were interviewed on exposures in the week before illness, which by the time of the interview was a few weeks ago, while controls were asked for exposures during the most recent week, yielding less “don’t know” answers on many questions. Apart from bias issues, a multivariate analysis on “complete cases” was not possible in this dataset as hardly any subjects would have remained in the analysis.

Assuming an ignorable missing data mechanism, multiple imputation can be used to get unbiased estimates in datasets like these. In multiple imputation first an imputation model is fitted and missing values are randomly drawn multiple times from the posterior distribution of the imputation model. Next, the analysis model (in our case a logistic model) is fitted multiple times and the parameter estimates are pooled.

When the mechanism generating missing data might be different for cases and controls, it is possible to use separate imputation models for cases and controls. We compared results when using an single imputation model on cases and controls jointly, and when using separate imputation models. For imputation we used MCMC (SAS PROC MI) to generate predicted values for all variables assuming a joint multivariate normal distribution, and imputed missing values from those predicted values by predictive mean matching. We used logistic regression to compare cases of *Campylobacter jejuni* infections (n=1003) with controls (n=3119). Results of separate and joint imputation were generally more similar to each other than to complete case analysis, with the exception of eating undercooked meat (OR 2.03 [1.61-2.56] in complete case analysis, 2.09 [1.61-2.73] with separate imputation of cases and controls, and 1.97 [1.59-2.43] with joint imputation). Differences, however, are too small to be of practical importance and could also be due to the randomness involved in the imputation procedure.

The theory behind multiple imputation assures proper working only when the disease model and imputation model are congenial. We will discuss this issue, as well as the issue of ignorability of the missing data mechanism. We will present some simulation results showing that using a probable mechanism of non-ignorable missingness, separate imputation performs slightly better than joint imputation. However, differences are very small, even in rather extreme cases, especially in view of remaining bias caused by the non-ignorability of the missing data mechanism.

**Longitudinal quality of life studies: the problem of the missing data reconsidered.**

**Wendy J. Post<sup>1</sup>**, Ciska Buijs<sup>2</sup>, Elisabeth G.E. de Vries<sup>2</sup>, Saskia le Cessie<sup>3</sup>

Email: w.j.post@mta.umcg.nl

<sup>1</sup>University Medical Center Groningen, Office for Medical Technology Assessment, Groningen, the Netherlands

<sup>2</sup>University Medical Center Groningen, Department of Medical Oncology, Groningen, the Netherlands

<sup>3</sup>Leiden University Medical Center, Department of Medical Statistics and Bio-informatics, Leiden, The Netherlands

This paper deals with analyses of longitudinal Quality Of Life (QOL) studies, in which two treatment arms are compared and in which both the (disease free) survival time and the quality of life are relevant outcome measures. The aim is to give a preference for one of the two treatments, taking both differences in (disease free) survival and in QOL into account. However, the comparisons of the QOL are hampered by problems of drop-out due to morbidity and/or death. Several papers are published about the problem of missing data in QOL studies. See for example the special issues 5-7 in *Statistics in Medicine* (1998) and the special issue in *Statistical Methods in Medical Research* 11, 2002

We encountered this problem in a Dutch multi-centre randomised clinical trial for breast cancer, in which the long-term impact of two different kinds of chemotherapy schedules (high dose versus conventional dose) was compared on QOL. The five years disease free survival rates were 65% in the high dose arm, and 59% in the conventional dose ( $p=0.09$ ). Over eight hundred patients were included in the study, which makes it possible to use complex modelling. Analysing the QOL data using a linear mixed model with the assumption of missing at random showed that the treatment arm with the best disease free survival has lower scores on the health related quality of life scores. The effect of relapse on QOL is difficult to see in this approach. Moreover, the assumption of missing at random in longitudinal QOL data is questionable, in particularly problems occurred with respect to the patients who drop-out due to morbidity or relapse (the clinical interesting group) and those who do not survive. In the linear mixed model no distinction is made between these two groups of patients. Moreover, in mixed models for both groups the results are extrapolated to the end of the study, which is not meaningful for those who do not survive. This was also mentioned by Pauler et.al (2003), who analysed this kind of data with pattern mixture models.

We modelled this data in different ways, ranging from very simple approaches like combining marginal means and survival percentages in one plot or imputing low values for QOL for deceased patients, to complex modelling with pattern mixture models. In the latter model, assumptions on the nature of the missing data must be made. Therefore, clinicians involved in this project were asked to identify several groups of patients with their own missing mechanism, and groups with special (expected) quality of life patterns. The main aim of this paper is to give an overview of which models give

which answers in which situations, and which models are problematic to interpret by clinicians.

*Reference*

Pauler, DK, McCoy, S, Moinpour, C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine* 2003, **22**: 795-809.

*Keywords*: missing data, quality of life, longitudinal studies

**Multiple imputation and model selection in Cox regression**

**C.M.A.Schipper, M.M.B. Breteler & Th. Stijnen**

Email: c.schipper@erasmusmc.nl

*Dept. of Epidemiology & Biostatistics, Erasmus MC, Rotterdam, the Netherlands*

In large cohort studies the number of missing observations is often substantial. This complicates the derivation of an adequate prognostic model, when one is interested in an outcome for which the missings appear in the prediction variables.

The naïve approach of using complete cases only, leads to estimates that are overly variable. Moreover, if the missings are not completely at random the estimates can also suffer from substantial bias.

We suggest an approach with multiple imputation via van Buuren's MICE library followed by stepwise model selection. Generalized Wald-statistics that are based on pooled coefficients and variances of Cox proportional hazards analyses provide the diagnostics in each step of the model selection process.

We extend the validation approach of Harrell to quantify how good the final selected prediction model is. The discrimination and calibration properties are obtained using bootstrap techniques that at the same time can produce estimates for shrinkage factors. These shrink the model coefficients to obtain better calibration properties.

We apply the proposed strategy to derive a prognostic model for the risk of developing dementia in the Rotterdam study. This is a population based prospective cohort study in Rotterdam of about 8,000 people of 55 years and above. During 10 years of follow up, over 400 cases of dementia were recorded.

Initially there are 39 covariates of interest. Although only 9% of the observations are missing, deletion of incomplete cases leads to a loss of over 70% of the cohort. Application of the aforementioned imputation and model selection procedure seems an appropriate approach here.

*References:*

- Harrell FE Jr, Lee KL, Mark DB. Tutorial in Biostatistics. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Stat Med* 15:361-387, 1996
- Rubin DB. Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York, 1987
- van Buuren S, Oudshoorn CGM. Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, 1999

**Model selection in multiply imputed datasets.**

**Angela Wood**, Ian White, Patrick Royston

Email: [angela.wood@mrc-bsu.cam.ac.uk](mailto:angela.wood@mrc-bsu.cam.ac.uk)

*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR*

Multiple imputation is a procedure to deal with missing data which can give valid estimates and standard errors under the "missing at random" assumption. Due to recent software development it is becoming an increasingly used procedure. Often one assumes there is an established substantive model to be fitted to the data. However, in some cases the model may not be known and one may wish to perform some model selection. Here we consider covariate selection. There are no guidelines for such a procedure in multiple imputed data sets and ad-hoc procedures are often used. In this work, we compare covariate selection in the following: the complete-cases; a single stochastic imputation; in combined multiple imputations using weighted regression; in the combined imputations using Rubin's rules and also a procedure which combines (using various rules) the covariates selected from each imputed dataset. Simulation results and applications to the UK700 trial and a renal cancer trial show there can be substantial differences in the model selected from these methods.

## II(b) – Multivariate Statistics

### The Visualisation of Multiplicative Interaction

John C. Gower<sup>1</sup> and Mark De Rooij<sup>2</sup>

Email: j.c.gower@open.ac.uk

<sup>1</sup>*Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K.*

<sup>2</sup>*Department of Psychology, Leiden University, Leiden, 2300 RB, The Netherlands.*

Among other things, the biadditive model:

$${}_p\mathbf{Y}_q = m\mathbf{1}\mathbf{1}' + \mathbf{a}\mathbf{1}' + \mathbf{1}\mathbf{b}' + {}_p\mathbf{C}_K\mathbf{D}'_q,$$

is often used to represent genotype/environment interactions (e.g. Kempton, 1984, and Denis and Gower, 1996). When  $K=2$ , the usual visualisation is to plot the rows of  $\mathbf{C}$  (points representing genotypes) and the rows of  $\mathbf{D}$  (points representing environments). Interpretation uses the visually inconvenient inner-product

$$\mathbf{C}_2\mathbf{D}' = (\mathbf{c}_1, \mathbf{c}_2)(\mathbf{d}_1, \mathbf{d}_2)'.$$

We present three more convenient visualisations:

- (i) Using a variation of the PCA biplot where either the genotypes or environments (or both) are represented by moveable scaled axes (see Gower and Hand, 1996)
- (ii) Replacing inner-products by distances (see De Rooij and Heiser, 2005).
- (iii) Replacing inner-products by areas.

The basic model remains unchanged but these visualisations are easier to interpret.

#### References

- Denis, J-B., & Gower, J.C. (1996) Asymptotic confidence regions for biadditive models: Interpreting genotype-environment interactions. *Applied Statistics*, **45**, 479-493.
- De Rooij, M. and Heiser, W.J. (2005). Graphical representations and odds ratios in a distance association model for the analysis of cross-classified data, *Psychometrika*, **70**, \*-\*.
- Gower, J.C. and Hand, D. J. (1996) *Biplots*. London: Chapman and Hall, 277 + xvi pp.
- Kempton, R. A. (1984) The use of biplots in interpreting variety by environment interactions. *J. Agric. Sci. Camb.*, **103**, 123-135.

**A nonlinear model with latent process using multivariate longitudinal data:  
application to cognitive aging**

Cécile Proust<sup>1</sup>, Hélène Jacqmin-Gadda<sup>1</sup>, Jeremy Taylor<sup>2</sup> and Daniel Commenges<sup>1</sup>

Email: cecile.prouts@isped.u-bordeaux2.fr

<sup>1</sup>INSERM E0338, ISPED, Université de Bordeaux2, 146 rue Léo Saignat, 33076  
Bordeaux cedex, France

<sup>2</sup>Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann  
Arbor, MI 48109, USA

As cognition is not directly measurable, it is generally assessed by a battery of psychometric tests.

Roy and Lin (2000) proposed a model for longitudinal data in which various continuous outcomes were linear transformations of a latent variable representing an unmeasurable quantity. The aim of this work is to propose a more general model for longitudinally describing cognition using various psychometric tests.

Global cognitive ability is considered as a latent process defined in continuous time and is modeled by a linear mixed model including covariates and a Brownian motion. Psychometric tests are then defined as parameterized nonlinear transformations (Beta CDF) of the global cognitive ability.

The likelihood of the model is written in the natural scale of the psychometric tests by using the Jacobian of the nonlinear transformation and estimation is performed using a Marquardt algorithm. Graphical methods are also proposed for assessing the adequation of the model.

This model for multivariate longitudinal data is applied for describing the global cognitive ability in the elderly and for assessing the association of covariates both with the global cognitive ability and with the various psychometric tests. This model gives also interesting results concerning the metrologic properties of the psychometric tests thanks to the flexibility of the family of nonlinear transformations used in the analysis.

*References*

Roy, J. and Lin, X. (2000) Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56, 1047-1054

**Mokken scale analysis using restricted optimization techniques**

**A.A.H. van Abswoude**

Email: a.a.vanabswoude@amc.uva.nl

*University of Amsterdam, Academic Medical Center, Departement of Clinical,  
Epidemiology and Biostatistics, P.O. Box 22660, 1100 DD Amsterdam*

Mokken scale analysis (MSA) is a nonparametric item response theory (NIRT) method that can be used to select items sensitive to approximately one dimension (for example an ability, trait or characteristic) from an item pool with items sensitive to multiple dimensions. Creating one or more ordinal scales using MSA can be seen as optimization problem subject to a number of MSA-specific conditions. The aim is to maximize for each scale the overall  $H$  coefficient, which for binary items can be written as normed covariances, subject to an observable consequence of the Monotone Homogeneity Model (a NIRT model) and a lowerbound for the  $H$  of each item. The latter allows the user to choose the minimum discrimination power of items into scales.

Studies have shown that the sequential clustering procedure used for MSA may not yield optimal solutions. Consequently, items joined into a scale may be less unidimensional or reliable than possible. In this study, the aim is to implement a new algorithm in MSA that allows us to keep the general focus of the method intact, but resolves the problems associated with the old algorithm (i.e., suboptimal solutions).

We use stochastic and deterministic versions of non-hierarchical clustering algorithms and three clearly defined objective functions both for unidimensional and multidimensional problems for this purpose. Using simulated data it was shown that the new stochastic methods may be used to obtain globally (or nearly globally) optimal solutions. The globally optimal solutions (in particular for the second objective function) reflected the underlying dimensionality structure of the data better than the solutions of the old method. Finally, suggestions how to incorporate scaling conditions in the optimization problem are given and illustrated.

## **III(a) – Experimental Design and Sampling**

### **Designs for real experiments using fractions of 2<sup>n</sup> factorials**

**Andrew Mead**

Email: [andrew.mead@warwick.ac.uk](mailto:andrew.mead@warwick.ac.uk)

*Warwick HRI, University of Warwick, Wellesbourne, Warwick, CV35 9EF*

Two recent collaborative projects have included experiments requiring novel fractions of 2<sup>n</sup> factorial designs. In both cases, there were constraints that did not appear to allow the use of conventional fractional designs available in textbooks or via computer-generated design packages, such as that available in GenStat.

The first project was concerned with the control of foliar pests of brassica crops, with field experiments considering the effects of different patterns of spray timing on the levels of pest presence and damage at harvest. The growing periods of cauliflower and cabbage were each divided into two-week periods (6 for cauliflower, 8 for cabbage) with a spray application being possible in each period. Each treatment then consisted of combinations of the presence or absence of sprays on each of 6 or 8 occasions, the complete set of treatments forming a 2<sup>6</sup> or 2<sup>8</sup> factorial set. A maximum of 64 plots could be included in each experiment, for practical purposes divided into 4 blocks of 16 plots. The primary aim of each trial was to provide information on the main effects and as many two- and three-factor interactions as possible (preferably including all interactions between two or three consecutive periods), to construct a predictive model to identify the number and most effective combinations of spray timings. It was considered essential to include the “no spray throughout” treatment in every block, and also to allow the observation of all eight treatment combinations in every set of three consecutive periods. A further constraint, however, was that we could only afford to regularly assess 8 plots per block, which therefore needed to include these eight treatment combinations across all sets of three consecutive periods.

The second project was concerned with identifying the sources of variability within the mushroom production system. There are several potential causes of cropping variability, which can be grouped into environmental variables, that cannot easily be controlled, and production variables, that can be manipulated. The production process is conventionally divided into 4 phases, the first two being concerned with compost production, the third with the development of spawn within the compost (spawn-running) and the last being the production of mushrooms (cropping). In the last two phases, trays are conventionally arranged in a three-dimensional array (e.g. in experimental work, 4 trays high, by 4 trays wide by 8 trays along), with differences between layers generally considered to contribute most to tray-to-tray variability. Initial trials just considered the environmental factors, defining 1024 possible combinations of spatial positions, with each trial only able to include 128 combinations. Experimental designs were required allowing the estimation of all main effects and two-factor interactions, plus the three-factor interactions involving three positional factors from either spawn-running or cropping. With two of these factors having 4 levels (those defining layers in each phase), an unusual fraction of the 2<sup>10</sup> factorial was required.

## **Interactive experimental design using modern computer software**

**R.N. Edmondson**

Email: [rodney.edmondson@warwick.ac.uk](mailto:rodney.edmondson@warwick.ac.uk)

*Warwick HRI, University of Warwick, Wellesbourne, Warwick, CV35 9EF*

Over the last 100 years or so, much research has been devoted to the design of efficient experiments in biology and agriculture (for example, see Edmondson 2005 for a review of design of crop experiments). However, the application of good design in agricultural research has always depended on the availability of skilled statisticians with a practical knowledge of the subject and this availability is now becoming limited. Scientists without ready access to statistical advice must look to the design of their own studies or experiments but may lack the experience to choose appropriate designs for their work. In the past, textbooks have provided a major source of design information for scientists but in the future computers, together with sophisticated design software, will provide a major resource for practical design work.

The availability of powerful computers, together with the rapidly increasing availability of powerful open-source software, provides opportunities for design software that is both wide-ranging and simple to use. We are developing web-based design software at <http://biometrics.hri.ac.uk> that is intended to provide options for a wide range of block and treatment designs. To encourage the development and uptake of the software, the core modules are being developed using the freely available open-source R language software available at <http://www.r-project.org/> together with a web-browser interface supported by the Java cross-platform software. We believe that the core modules must be based on open-source software to ensure portability but the peripheral outputs can be displayed via any specialist statistical package and we use GenStat to display a dummy analysis for some of the designs. In addition to the server-based software, we also hope that the software will eventually become available as a simple free-standing application.

We will exemplify the use of the software to construct incomplete block designs. We will also briefly mention some further future developments of the site including fractional factorial design algorithms for response-surface designs (Edmondson 1991) and power studies for block and treatment designs using R algorithm simulations (Horton et. al. 2004).

### *References*

- Edmondson, R. N. (1991) Agricultural response surface experiments based on four-level factorial designs. *Biometrics* 47, 1435-1448
- Edmondson, R. N. (2005) Centenary Review: Past developments and future opportunities in the design and analysis of crop experiments, *Journal of Agricultural Science, Cambridge*, In press
- Horton, N.J, Brown, E.R, Qian, L (2004). Use of R as a Toolbox for Mathematical Statistics Exploration, *American Statistician*, 58, 4, 343-357

## **Are We Really That Blind?**

**Egbert A. Van Der Meulen**

Email: bert.vandermeulen@organon.com

*Clinical Trial Operations /Biometrics, NV Organon, PO Box 20, 5340 BH Oss, The Netherlands*

In double-blind randomized clinical trials it is common practice to randomize patients using randomly permuted blocks. In this paper, it is demonstrated that before unblinding statistical inference of the treatment effects can be conducted yielding consistent and rather precise estimates even in the presence of an additive block-effect. With an even greater precision the within-group standard deviation on which power calculation are usually based can be inferred from blinded data. The use of blocks of random lengths as suggested by ICH-E9 in the (unlikely) case that previous treatment allocation can be guessed by strong pharmacological effects, merely complicates the analysis but blinded inference can still be conducted without much extra loss of information. On the hand, one might argue that this possibility of blinded inference takes away the need of conducting interim-analyses for administrative or business reasons or for sample size re-estimation. On the other hand, however, it most probably will have a disputable, positive or negative effect on the conduct of the remainder of the trial. If regulators and the pharmaceutical world at large would like to avoid this possibility, then other e.g. unrestricted, biased-coin, or more general dynamic allocation randomization procedures may be less controversial alternatives. It at least provides another strong argument in favor of using large blocks as the precision of blinded inference decreases with increasing block lengths.

If blinded inferences are deemed a useful replacement of interim-analyses in non-pivotal trials, then further guidelines will be needed on consequent decision-making aspects.

This paper is to appear in the Journal of Biopharmaceutical Statistics, 2005, Issue 3.

## **Relative efficiency of unequal versus equal cluster sizes in multicenter trials**

**Gerard Van Breukelen**, Larissa Kotova, Math Candel & Martijn Berger

Email: gerard.vbreukelen@stat.unimaas.nl

*Dept. of Methodology and Statistics, Maastricht University, The Netherlands*

Equal cluster sizes are optimal for estimating and testing a treatment effect in a multicenter trial (assuming homogeneous variances), but rarely feasible. This paper addresses the relative efficiency (RE) of unequal versus equal cluster sizes for treatment effect evaluation, assuming a quantitative outcome and 50:50 randomized allocation of clusters or of persons within clusters. It is shown how the RE depends on the intraclass correlation (ICC) and the amount of cluster size variation, attaining a minimum of 0.90 or 0.80 at worst for realistic distributions of cluster sizes. It confirms and generalizes

results by Kerry and Bland (2001) for English and Welsh general practices. And it suggests a simple correction for optimal designs as derived by Moerbeek, Van Breukelen and Berger (2000) under the assumption of equal cluster sizes.

*References*

- Kerry S, Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine* 2001; 20: 377-390.
- Moerbeek M, Van Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. *J of Educational and Behavioral Statistics* 2000; 25: 271-284.

### III(b) – Genetics

#### **Joint Estimation Of Gene-Gene And Gene-Environment Interaction Effects For Numerous Loci Using (Double) Penalised Log-Likelihood**

**M.W.T. Tanck**<sup>1</sup>, J.W. Jukema<sup>2</sup> and A.H. Zwinderman<sup>1</sup>

Email: m.w.tanck@amc.uva.nl

<sup>1</sup> *Dept. of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands.*

<sup>2</sup> *Dept. of Cardiology, Leiden University Medical Center, Leiden, The Netherlands.*

The number of genetic markers available for association studies between these markers and complex traits like e.g. coronary artery disease is growing rapidly. Next to main effects of the markers, gene-gene and gene-environment interactions are becoming increasingly important. However, joint analysis of numerous main and interaction effects is infeasible when the number of effects surpasses the number of observations. One approach is to have a marker selection step preceding the joint analysis (see e.g. Ott and Hoh, 2001). However, since only markers with a significant main effect on the phenotypic trait are included in the joint model, this approach assumes that the interaction effects between selected and non-selected markers are zero.

In our approach, the main and two-way interaction effects of all markers are included in the linear regression model without selection based on single marker effects, under the restriction that the number of total effects to be estimated is smaller than the number of observations. We assume that the majority of the interaction effects are zero and a LASSO (L1) penalty on the interaction effects in the likelihood function is used to select those interactions that are significantly different from zero. Subsequently, only the non-zero interaction effects are included in a reduced unpenalised model resulting in unbiased estimates of the different effects. To further reduce noise, a Ridge (L2) penalty on the main effects can be included either simultaneously with the L1 penalty or in the reduced model only.

Results of simulations showed that with this approach significant interactions were nearly always found when the effect was larger than the associated residual standard deviation. Less often, but usually recognisable, when the effect was smaller. Subsequent re-estimation of the effects using a reduced model with the main and significant interaction effects only (no penalty) resulted in unbiased estimates of the interaction effects.

#### *Reference*

Ott, J. and Hoh, J., 2001. Human Mutation 17: 285-288

**Weighted penalised logistic regression to estimate multilocus haplotype effects on dichotomous outcome**

**Olga W. Souverein**, Aeilko H. Zwinderman, Michael W.T. Tanck

Email: o.w.souverein@amc.uva.nl

*Department of Clinical Epidemiology and Biostatistics, Academic Medical Center,  
Amsterdam, the Netherlands*

Multiple heterozygous individuals present a problem when trying to estimate haplotype effects in association studies of unrelated individuals when phase is not known. Methods have been developed to estimate haplotype frequencies and haplotype effects in such populations. Previously, Tanck et al.<sup>1</sup> have described a weighted penalized log-likelihood method to handle multiple heterozygous individuals in an efficient way. The present study is a generalisation of this method for dichotomous outcome data, namely a (penalised) weighted logistic regression. In short, the method uses estimated haplotype frequencies to assign weights to all possible haplotype combinations of multiple heterozygotes. This is done by calculating posterior weights using Bayes' theorem. Furthermore, these weights are re-estimated using the predicted and observed outcome. We also considered a ridge regression, using the similarity between haplotypes as a penalty function. The penalty function introduces bias for the parameter estimate, but increases precision, which would be especially helpful for estimating effects of infrequent haplotypes.

To investigate properties of our method, we performed a simulation study. A total of six haplotypes consisting of five SNPs were chosen to be present in the population with haplotype frequencies similar to those we previously found in the CETP gene<sup>1</sup>. Haplotypes were randomly assigned to 500 individuals. Disease status was sampled from the binomial distribution with probability depending on the haplotypes using a logistic model. Overall disease prevalence was 10%. We evaluated performance of our method in different scenarios with one or more (comparable or different) haplotypes associated with disease. For each scenario, 500 replicates were carried out.

Preliminary results indicate that the method of weighted logistic regression performs adequately. As expected, including the penalty reduces the standard error of the estimates (most pronounced for haplotype with lowest frequency), but increases the bias. However, the penalty has adverse effects when there is only one functional haplotype with a small effect, in that it reduces power and increases false positive results. The penalty is advantageous when similar haplotypes have similar effects on disease, as is the case when the allele of one SNP has an effect on the disease, irrespective of the surrounding SNPs.

*Reference*

Tanck M.W.T., Klerkx A.H.E.M., Jukema J.W., De Knijff P., Kastelein J.J.P., Zwinderman A.H. (2003) Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. *Annals of Human Genetics* 67,175-184.

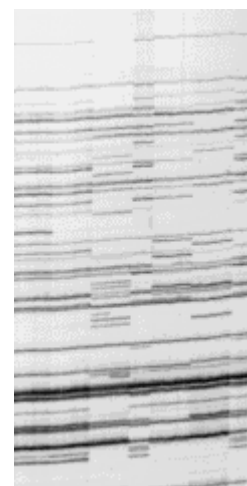
## Collision probabilities for bands in AFLP's

G. Gort

Email: gerrit.gort@wur.nl

*Biometris, Wageningen University and Research Center, Wageningen,  
The Netherlands*

AFLP is a DNA fingerprinting technique, used often in the plant and animal sciences. One of the drawbacks of the technique is the occurrence of multiple DNA fragments of the same length, which we name collisions. The problem is closely connected to the well-known birthday problem. In this paper we quantify the problem, focussing on individual bands.



To calculate collision probabilities a fragment length distribution (fld) is needed. We estimate the fld from the band pattern, using a generalized linear model with binomial distribution and complementary log-log link. The fld is modelled as a monotone smooth function of the fragment length by application of a monotone version of P-splines (Eilers & Marx, 1996). As a by-product of the generalized linear model we get an estimate of the number of fragments, and, hence, of the number of collisions (= #fragments – #bands).

The probability distribution of the number of fragments, given the observed band pattern and the estimated fld, is approximated by application of Bayes rule and a saddle point approximation for multinomial tail probabilities. For an AFLP examples with 56 bands on lettuce (*L. sativa*) the expected number of fragments is 63.9 (s.d. 3.1), so 8 collisions.

Because the fld is highly asymmetrical with a strong preference for shorter fragments, more collisions are expected for the "shorter" bands. Therefore, collision probabilities are calculated for individual bands. Bayes rule, generalized occupancy distributions and saddlepoint approximations for multinomial tail probabilities lead to the wanted probabilities. We find a strong effect of the band position, with much smaller collision probabilities for the longer fragments (up to 20 times smaller).

### References

Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89-102

**A non-linear mixed model for modelling QTLs related to plant development**

Marcos Malosetti & **Fred van Eeuwijk**

Email: Fred.vanEeuwijk@wur.nl

*Laboratory of Plant Breeding, Wageningen University, P.O. Box 386, 6700 AJ  
Wageningen, The Netherlands*

A typical objective of QTL modelling is to predict trait values as expressed at a particular point in time as a function of molecular marker information. However, for many purposes not only the trait value at a particular time is of importance, but also the development of the trait over time. Traditionally, crop growth models are used to study development in plants. Various attempts have been made recently to integrate crop growth models and QTL models for development. A two-stage approach consists in first estimating parameters for developmental curves and next enter those parameters as ‘traits’ in a standard QTL analysis (REYMOND *et al.* 2003; YIN *et al.* 2005). We propose a further integration of crop growth models and QTL models within the framework of non-linear mixed models. Our approach is closely related to work by MA *et al.* (2002), in which physiological processes are modeled using maximum likelihood coupled to the EM algorithm. We illustrate our approach in a QTL model for the senescence process in potato leaves.

*References*

- Ma, C.-X., G. Casella And R. Wu, 2002 Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 161: 1751-1762.
- Reymond, M., B. Muller, A. Leonardi, A. Charcosset And F. Tardieu, 2003 Combining Quantitative Trait Loci Analysis and an Ecophysiological Model to Analyze the Genetic Variability of the Responses of Maize Leaf Growth to Temperature and Water Deficit. *Plant Physiol.* 131: 664-675.
- Yin, X., P. C. Struik, F. A. Van Eeuwijk, P. Stam And J. Tang, 2005 QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. *J. Exp. Bot.*: in press

## IV(a) – Inference

### Visualizing, Summarizing and Comparing odds ratio structures

Mark de Rooij<sup>1</sup> and Carolyn J. Anderson<sup>2</sup>

Email: ROOIJM@fsw.leidenuniv.nl

<sup>1</sup>*Dept. of Psychology, Leiden University, PO Box 9555, 2300 RB Leiden, THE NETHERLANDS*

<sup>2</sup>*University of Illinois, Champaign Urbana, USA*

Odds ratios are the main measure of association in 2 x 2 frequency tables. For larger tables summary measures like the odds ratio have been proposed as well as modelling strategies, where from the parameters of a model the odds ratios can be deduced. We propose a computationally simple scaling methodology, which gives a summary measure of association as well as a visualization of the odds ratio structure. Variants of the methodology will be discussed and compared, both theoretically as well as on an empirical data set. In a second step the methodology will be generalized to multiple tables, where not only the structure per table is important, but also the comparison among tables.

### Reduced Rank Techniques For Multi-State Models

M. Fiocco, H. Putter and J.C. van Houwelingen

Email: m.fiocco@lumc.nl

*Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, PO Box 9604, 2300 RC Leiden, The Netherlands*

Multi-state survival analysis usually involves a series of detailed regression analysis describing transitions between various states. When incorporating covariates influencing the transition intensities, an obvious approach is to use Cox's proportional hazards model for each of the transitions separately. A practical problem then is how to deal with the abundance of regression parameters coming from each transition in the model. We propose a reduced rank multi-state model where a parameterization is used such that the matrix of all regression coefficients is of lower rank  $R$ . For the special case of  $R=1$ , such a reduced rank model is a proportional hazards model where all covariates have the same effect on each permitted transition from state  $i$  to state  $j$  in the multi-state model but with proportionality coefficients. Several advantages derive from the application of such a model: fewer parameters need to be estimated, it allows for clearer interpretation of the parameter estimates and the model can also handle transitions with rare events.

We shall outline an algorithm that estimates the regression coefficients in a reduced rank multi-states model, then show how to compute their standard errors and make predictions of various outcomes from specified patient histories by using multi-state

reduced rank models. The illustration of this approach will be based on leukemia patients from the European Group for Blood and Marrow Transplantation (EBMT).

*Keywords:* multi-state models, survival analysis, reduced rank, prognostic factors

### **A Goodness-of-Fit Test for Multinomial Logistic Regression**

**J. J. Goeman** and S. le Cessie

Email: j.j.goeman@lumc.nl

*Department of Medical Statistics, Leiden University Medical Center*

We present a score test to check the fit of a logistic regression model with two or more outcome categories. The null hypothesis that the model fits well is tested against the alternative that there is an additional random effect with a specified covariance structure, which induces a correlation structure among the residuals. By specifying the covariance matrix of the random effect, the user of the test can choose the alternative against which the test is directed, making it either an omnibus goodness-of-fit test or a test for lack of fit of specific model variables or samples. The test is an extension of a goodness-of-fit test of Le Cessie and Van Houwelingen for two-valued logistic regression. The test is applied to a liver enzyme data set.

### **Lower and upper bounds on the correlation between treatment and instrumental variable**

**Edwin P. Martens**<sup>1,2</sup>, Wiebe R. Pestman<sup>1</sup>, Anthonius de Boer<sup>2</sup>,  
Svetlana V. Belitser<sup>2</sup> and Olaf H. Klungel<sup>2</sup>

Email: e.p.martens@bio.uu.nl

<sup>1</sup>*Centre for Biostatistics, Utrecht University, The Netherlands*

<sup>2</sup>*Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands*

When Ordinary Least Squares (OLS) is used to estimate a treatment effect in observational studies, in general the estimate will be biased. To correct for possible confounding the method of simultaneous structural equations, better known as instrumental variables (IV), is increasingly used in the medical literature. A strong basic assumption is that the IV determines the response variable *only* through its correlation with the treatment variable: there must be no other variable that partly determines both IV and response. When the correlation between treatment and IV is small, the method of IV will still give an asymptotically unbiased treatment effect, but there are some weaknesses. The first is that standard errors become larger when this correlation becomes weaker. Second, the estimator will be inconsistent when the basic assumption of IV estimation is only slightly violated. Third, the estimate will be more and more biased in the same direction as OLS when sample size becomes smaller. In general,

there exist a lower bound on the correlation between treatment and IV for a valid application of this method. Although well known from the literature, this implicit ‘lower bound’ on the correlation between IV and treatment is often ignored in practice. With some examples from studies in which IV estimation has been used we showed the application of the method, the plausibility of the assumptions and the limitations.

As a consequence it is advised to have an instrument that is strongly correlated with treatment when IV estimation is to be used. But when we like to find such a strong instrument not violating the basic assumption, we will face the situation that, given a certain amount of confounding in the data set, also an ‘upper bound’ exists on the correlation between IV and treatment. It turns out that this upper bound is lower when the amount of confounding is larger. In other words: in cases of considerable confounding it is impossible to find a strong instrument, and when one does, it is likely that the assumptions of IV will be violated. When on the other hand confounding is small the transition from OLS to IV estimation is less urgent. In some figures we showed the relationship between the amount of confounding and the strength of the IV.

## **IV(b) – Statistics in Agriculture and Food**

### **Bayesian Multiplicative Model for Assessor Performance**

**Dr Aletta Nonyane<sup>1</sup> and Dr Chris Theobald<sup>2</sup>**

Email: B.A.Nonyane@bham.ac.uk

<sup>1</sup>*Research Fellow, Department of Primary Care and General Practice, University of Birmingham, UK*

<sup>2</sup>*School of Mathematics, University of Edinburgh and Biomathematics & Statistics Scotland, UK*

The food industry is increasingly involved in food tasting experiments for reasons such as establishing consumer preferences of different food and drink varieties on the market. Usually, series of such experiments are conducted with a panel of trained assessors who give their scores or rankings of food products, with respect to certain attributes. The assessors are selected and trained so that their scoring is reliable, consistent and shows high discriminating ability. Analysis of data from apple tasting experiments is presented here, illustrating a multiplicative interaction model for assessor performance. Several experiments were conducted with 14 trained assessors who scored apple varieties on a number of attributes. The analysis presented here is for the sweetness attribute, which was scored on the scale of 0 to 100. The design used in each experiment was a Williams Latin square design which accounts for order and carry-over effects.

A multiplicative model widely used in plant variety trials is used here to model assessor performance in food tasting. Assessors are taken as measuring instruments that quantify the intensity of food attributes for which no standard measure is known. Thus, the differences between food products are realized through the assessors' use of the scale. This results in the multiplicative nature of the interaction between assessor and food product effects.

A Bayesian hierarchical framework enables one to combine data from the several experiments, in order to model assessor performance over time. Furthermore, not all assessors in the apple tasting experiments attended every experiment, and therefore, the Bayesian predictive approach was used to adjust mean product effects for the missing assessors. Another advantage of the Bayesian framework is that information on the assessors' use of the scale may be incorporated into future analysis for more precise estimation.

*Keywords:* Food tasting; Multiplicative interaction model; Bayesian predictive approach

**Acknowledgement:** Apple tasting data were kindly provided by the Hannah Research Institute, Ayr, UK. The work of the speaker was supported by the Cecil Renaud Educational and Charitable Trust, South Africa.

**Evaluation of diagnostic tests in the absence of a gold standard – models and applications**

**B. Engel**<sup>1</sup>, A. Bouma<sup>2</sup>, W.G. Buist<sup>1</sup>, B. Swildens<sup>2</sup>, H. van Roermund<sup>1</sup> and M.C.M. de Jong

Email: bas.engel@wur.nl

<sup>1</sup>*Quantitative Veterinary Epidemiology, Animal Sciences Group of Wageningen University Research, Lelystad, The Netherlands*

<sup>2</sup>*Faculty of Veterinary Medicine, University of Utrecht, The Netherlands.*

The Animal Sciences Group, in cooperation with the University of Utrecht, has been involved in the evaluation of accuracy of diagnostic tests for various diseases in farm animals in The Netherlands, such as classical swine fever (CSF), EF-positive *Streptococcus suis* serotype 2 strains (*S.suis* for short) in sows, para tuberculosis, avian tuberculosis and foot and mouth disease. We will focus on two studies: accuracy of clinical diagnosis of CSF for breeding sows at herd level (Engel et al., 2004a) and accuracy of diagnostic tests for *S.suis* in sows (Engel et al., 2004b).

In the evaluation of clinical diagnosis for CSF at herd level, data from the 1997-1998 CSF outbreak in the Netherlands were analysed. Herds were visited by veterinarians and, on the basis of their reports, each visit was coded as a binary variable (negative or positive clinical diagnosis). A feature of special interest in this study was the dependence of herd sensitivity of clinical diagnosis of CSF on the number of days elapsed since virus introduction in the herd. The true status of the herds (positive or negative) was (eventually) known. However, the moment of introduction of the CSF virus in the herds was unknown. Hence, when a herd was visited by a veterinarian, neither the status of the herd at that moment, nor the number of days since virus introduction was known. There was no gold standard, but a probability distribution for the moment of virus introduction could be derived from serum samples collected at the moment of depopulation of the herds. This distribution was incorporated in a logistic regression model for the binary data derived from the veterinarians' reports.

The *S.suis* study involved data from a field study with 3 diagnostic tests that are (possibly) conditionally dependent. No gold standard was available, i.e. the true disease status of the animals was unknown. Because one of the tests, involving bacterial examination, was quite expensive, it was only performed for a subset of the samples. The model is an instance of a latent class model. Because the data set was incomplete, a particular parameterization was chosen involving product conditional binomial distributions rather than multinomial distributions.

For both studies, Bayesian posterior inference was performed with the Gibbs sampler, as implemented in the WinBUGS package. Aspects of modeling, problems with choice of priors, goodness of fit of the model and numerical problems with the Gibbs sampler will be discussed.

*References*

- Engel, Bouma, Stegeman, Buist, Elbers, Kogut, Döpfer, De Jong (2004a). When can a veterinarian be expected to detect classical swine fever virus among breeding sows in a herd during an outbreak? *Prev. Vet. Med.* (in press).
- Engel, Swildens, Stegeman, Buist, De Jong (2004b). Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. Submitted for publication.

**Probabilistic food risk assessment accounting for variability and uncertainty in both chemical exposure and toxicological effects**

**Hilko van der Voet<sup>1</sup> & Wout Slob<sup>2</sup>**

Email: hilko.vandervoet@wur.nl

<sup>1</sup>*Wageningen University and Research centre, Biometris, P.O. Box 100, 6700 AC Wageningen, Netherlands*

<sup>2</sup>*RIVM, P.O. Box 1, 3720 BA Bilthoven, Netherlands*

Both dietary intake (exposure) assessment and toxicological critical effect (benchmark) dose modelling are developing as probabilistic techniques. It is natural to integrate both approaches for a complete risk assessment for a substance under consideration.

In exposure assessment it is usual to characterise the distribution of individual intakes by a high tail percentile. Toxicological dose-effect studies typically result in a dose level which is considered borderline safe (e.g. NOAEL, or CED). Probabilistic modelers of exposure tend to refer to a fixed toxicological limit value, whereas probabilistic modelers of toxicology tend to refer to a fixed exposure level. It is the purpose of this paper to present an integrated model where both exposure and toxicology are included in a probabilistic way.

The purpose of food or drug safety risk managers is to provide protection for the general population or for specific subpopulations, such as children or workers. The proposed approach is *population-based*, and the protection will be in the form of specifying the probability that a random individual from a defined (sub)population will have an exposure high enough to produce a predefined critical health effect (e.g. 10 % rise of cholesterol blood level). Such an exposure will be called a *critical exposure*, and the lowest critical exposure for each individual will be called the *individual critical effect dose* (ICED).

Individuals in a population typically show variation, both in food or drug intake and in toxicological sensitivity. In our approach we quantify both the variation in exposure and the variation in sensitivity in the form of probability distributions. Then, assuming independence between exposure and sensitivity, these distributions are easily combined into a distribution of *individual margin of exposure* (IMoE). This is a probabilistic standardisation, and IMoE values lower than 1 represent critical exposures. The proportion of the IMoE distribution below 1 is the *probability of critical exposure* (PoCE) in the defined (sub)population.

To quantify uncertainty we concentrate on the bootstrap as a practical device for arbitrarily complex situations. In the bootstrap procedure data sets and/or uncertainty distributions are repeatedly resampled, and in each iteration the statistic of interest is calculated. Accumulated over a large number of iterations this gives the bootstrap uncertainty distribution of the statistic of interest. In our proposed method for integrated risk assessment we choose as primary statistic of interest the probability of critical exposure (PoCE). Additional statistics of interest are for example specific percentiles (e.g. p5, p1) of the IMoE distribution.

## **A Comparison of Mixed Model Splines**

**Sue Welham**

Email: [sue.welham@bbsrc.ac.uk](mailto:sue.welham@bbsrc.ac.uk)

*Rothamsted Research, Harpenden*

Over the past 10 years, various forms of spline have been put into the mixed model framework, with the smoothing parameter estimated by the method of residual maximum likelihood (REML). This formulation has the advantage that semi-parametric terms can be added to complex mixed models accounting for all sources of variation in the data, including correlated errors.

Cubic smoothing splines were used within the mixed model framework by Verbyla *et al.* (1999), and introduced by several other authors around the same time. Eilers and Marx (1996) introduced P-splines, which used a B-spline basis with reduced knot points and a discrete penalty based on differencing, and were motivated as a computationally efficient approximation to polynomial smoothing splines. In addition, the authors suggested that the order of differencing used in the penalty could be varied. Eilers, in discussion of Verbyla *et al.* (1999), showed that P-splines could also be fitted within the mixed model framework, although the computational advantage of using a B-spline basis is then lost. More recently, Ruppert, Wand & Carroll (2003) introduced penalised splines, which use a truncated power function basis with an ‘ad-hoc’ penalty.

These spline methods are now becoming increasingly popular, but there has been little comparison of the different models. There are in fact close connections between the models, so that the recommended ‘default’ splines (cubic P-spline with second-order differencing, linear penalised spline) can both be considered as low-rank approximations to the cubic smoothing spline. However in general, the user has a choice of the order of the spline basis, the number and position of knots and the penalty (order of differencing). The minimum order of spline basis is usually determined by the required properties of the fitted spline, in terms of differentiability. The issue of knot selection has been widely covered in the regression spline literature, and is summarised by Wand (2000). The choice of penalty can have a large impact on the fitted spline, but there are currently no satisfactory inferential tools for selecting between these splines when fitted by REML.

This talk will examine the impact of different penalties on the fitted spline in an agricultural example (response of crop to soil properties) and a small simulation study. Diagnostics that explain the differences between the fitted splines will be considered.

*References*

- Eilers PHC & Marx BD (1996) Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**, 89-112
- Ruppert D, Wand MP & Carroll RJ (2003) *Semiparametric regression*. Cambridge University Press.
- Verbyla, A P, Cullis B R, Kenward M G & Welham S J. (1999) The analysis of designed experiments and longitudinal data using smoothing splines (with discussion). *Applied Statistics*, **48**, 269-311.
- Wand MP (2000) A comparison of Regression Spline Smoothing Procedures. *Computational Statistics*, **15**, 443-462.

## **V(a) – Spatial and Temporal Statistics**

### **Early detection of outbreaks of infectious diseases**

Siem Heisterkamp and **Janneke Heijne**,

Email: [janneke.heijne@rivm.nl](mailto:janneke.heijne@rivm.nl)

*National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands*

Since 1994 the National Institute for Public Health and the Environment (RIVM) receives electronically all tests results of more than 350 pathogens from a growing number of laboratories on a daily basis. At present, data is received from eleven laboratories, covering approximately 3 million of the 16 million inhabitants of the Netherlands. The laboratory reports are anonymous, but include unique identifiers such as date of report and date of sampling. The Ministry of Public Health is interested in early detection of outbreaks of infectious diseases and our task is to use the laboratory results to advise on this matter. Therefore a time series record is kept and alerts are generated when the number of cases of a certain pathogen is above a given threshold for that pathogen. For some (highly infectious) pathogens the threshold is simply 0 (zero), but for most other pathogens one wants to know whether the number of cases is higher than expected, or, for a seasonal disease, if a season starts earlier than expected. Thus a threshold could depend on season, or the foregoing day etcetera.

This raises the following question: is it possible to provide one universal algorithm for all pathogens that produces a threshold such that we get an alert as early as possible, but not too early (false alert) and not too late (miss an outbreak). The alerts are discussed on a weekly basis with infectious diseases professionals and, if appropriate, the alerts are spread to Municipal Health Services and to the Dutch Inspectorate of Health.

In the recent past several attempts have been made to counter the problem of setting threshold, an overview of the most popular algorithms is given in [1]. The considered algorithms are largely based on heuristics and not so much on rigorous statistical modelling. RIVM had implemented the linear variant of the algorithm of Farrington [2] which gave problems with low number of cases- due to using a generalised linear model of Poisson counts and the identity link function.

We proposed another track- the use of a hierarchical time series analysis model [3]. We assume an unobserved auto-correlated process with regard to the means of the number of cases and a discrete distribution- in particular a Poisson distribution- for the observed counts. To avoid lengthy daily computations, we trained for each pathogen a time series model on the last two years and used its parameters- only two or three for the on-line daily analysis.

Each day the observed counts are compared with the predicted counts of that particular day- an alert is generated if appropriate- and a new predicted value for the following day is computed using the new observation. The daily computations involve the solving of a non-linear equation and the computation of the threshold involves a numerical

integration. We will give examples for different thresholds- depending on its use- for different pathogens (with high and low number of cases and with highly variable number of cases).

*References*

- [1] RolfHamre, AVP, Outbreak detection of Communicable Diseases, Smyttskyinstitutets rapportserie nr 3: 2003. Sweden
- [2] Farrington CP, Andrews NJ, Beale AD and Catchpole MA: A Statistical Algorithm for the Early detection of outbreaks of Infectious Disease. JRSS, A, 159, (1996) 547-563
- [3] Dekkers ALM, Heisterkamp SH, NPbats Software for Bayesian timeseries analysis, Technical report 550002006/2004, 2004 (in Dutch)

**Empirical Bayesian time series analysis in Airpollution and Health studies**

**S.H. Heisterkamp** and A.M.L. Dekkers

Email: sh.heisterkamp@rivm.nl

*National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands*

In this paper we propose the use of an empirical Bayesian time series analysis instead of the more frequently used GAM and GLM models as advocated Schwartz [1] in the analysis of air pollution and health studies. The proposed model resembles the Dynamic Generalised Model (DGLM) as proposed in [2], but is more flexible and is in fact a generalization of the statistical method used in [3]. The comparison of different statistical techniques in the study of air pollution and public health has been triggered by the recent discussion on the topic of convergence of GAM models when using default criteria and the under estimation of the standard errors of the regression coefficients of those [4,5]. We show that the empirical Bayesian analysis can be done with a standard package, e.g. S-Plus, using a modified version of the existing glm() function of Splus, indeed yielding larger standard errors for the effects of interest. The proposed model can easily be generalised for a full blown Bayesian analysis using standard software like BUGS.

*References*

- [1] Schwartz J, (1994) Nonparametric smoothing in the analysis of air pollution and respiratory illness, The Canadian Journal of Statistics, vol. 22, no. 4, 471-487
- [2] Chiogna M, Gaetan C, Dynamic generalized linear models with application to environmental epidemiology, Appl. Statist., 2002, 51, Part 4, pp 453-468
- [3] S.H. Heisterkamp, J.C. van Houwelingen, A.M. Downs, (1999) Empirical Bayesian Estimators for a Poisson Process Propagated in Time, Biometrical Journal 41 4, 385-400
- [4] Dominici F., McDermott A., Zeger S.L., Samet J.M., (2002) On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health, AJE, , 156,3, 193-203

- [5] Katsouyanni K., Touloumi G., Samoli E., Gryparis A., Monopolis Y. (2002) Different Convergence Parameters Applied to S-Plus GAM function, letter to the Editor, *Epidemiology*, 2002, vol 13, no. 6, 742

**A GLMM approach to study the spatial and temporal evolution of spikes in the small intestines.**

**Christel Faes<sup>1</sup>**, Marc Aerts<sup>1</sup>, Luc Bijmens<sup>2</sup>, Helena Geys<sup>2</sup>, Geert Molenberghs<sup>1</sup>

Email: christel.faes@luc.ac.be

<sup>1</sup>*Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium*

<sup>2</sup>*Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica N.V. Beerse, Belgium*

Mixed models can be applied in a wide range of settings. Probably, they are most commonly used to handle grouping in the data. In addition, mixed models can be used for smoothing purposes as well. Speed (1991) explicitly made the connection between nonparametric regression and mixed models. When dealing with non-normal data, the use of smoothing methods within the generalized linear mixed models (GLMM) framework is less familiar (Ruppert, Wand and Carroll, 2003). We explore the use of GLMM for smoothing purposes in both a spatial and longitudinal dimension. The methodology is illustrated by analysis of spike potentials in the small intestines of different cats (Lammers *et al.* 1996, 2000).

In the small intestines, coordinated contractions of smooth muscle participate in several ways to facilitate digestion and absorption. This is controlled predominantly by signals from the enteric nervous system. Two basic patterns of electrical activity are important: the slow waves and spike potentials. Until now, there is no information as to the behaviour of spike potentials in successive slow waves. Both the spatial and temporal evolution of spikes is of interest. We propose a generalized linear mixed model for the analysis of the spatial and temporal effects of spike potentials. Spatio-temporal models that use two-dimensional smoothing splines across the spatial dimension and random effects to account for the correlations during successive slow-waves are developed. A major advantage of the mixed model approach is that it can handle smoothing together with grouping (or other types of correlations) in a unified model. In this way, areas with high spike incidence compared with other areas can be detected. Also, the temporal and spatial characteristics of spikes during successive slow waves can be identified.

**Keywords:** gastroenterology, generalized linear mixed models, smoothing splines, spatiotemporal model.

*References.*

- Lammers, W. J. E. P., Stephen, B., Arafat, K., and Manefield, G.W. (1996). High resolution mapping in the gastrointestinal system: initial results. *Neurogastroenterology and Motility*, **8**, 207-216.

- Lammers, W. J. E. P., Faes, C., Stephen, B., Bijmens, L., Ver Donck, L., and Schuurkes, J.A.J. (2004). Spatial Determination of Successive Spikes in the Isolated Cat Duodenum. *Neurogastroenterology and Motility*, **16**, 775-783.
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Speed, T. (1991). Comment on paper by Robinson. *Statistical Science*, **6**, 42-44.

## **V(b) – Miscellaneous**

### **Testing log-linear models with inequality constraints: a comparison of asymptotic, bootstrap and posterior predictive p values**

**Francisca Galindo Garre**

Email: f.galindogarre@amc.uva.nl

*Academic Medical Centre, University of Amsterdam*

In the analysis of contingency tables the assessment of the goodness-of-fit of an estimated statistical model usually involves determining the discrepancy between observed and estimated frequencies using the likelihood-ratio statistic. In models with inequality constraints as are used for ordinal data, the asymptotic distribution of this statistic depends on unknown model parameters whose values are often on the boundary of the parameter space. As a result, there no longer exists a unique p value. Bootstrap p values obtained by replacing the unknown parameters by their maximum likelihood estimates may also be inaccurate, especially if many of the imposed inequality constraints are violated in the available sample. In this talk, the various problems associated with the use of asymptotic and bootstrap p values will be described and the use of Bayesian posterior predictive checks will be proposed as a better alternative for assessing the fit of log-linear models with inequality constraints.

#### *Reference*

Galindo-Garre, F., and Vermunt, J.K. (in press) Testing Log-linear Models with Inequality Constraints: A Comparison of Asymptotic, Bootstrap, and Posterior Predictive P Values. *Statistica Neerlandica*.

### **A Multivariate extension of Halley's method for finding the stationary points of a function.**

**Clive Moncrieff**

Email: cbm@nhm.ac.uk

*Natural History Museum, London*

Halley proposed a variant of the Newton-Raphson method for finding the solutions of  $f(x)=0$ . The Newton-Raphson method is widely used for finding the stationary points, for example the maxima of a likelihood function, of a function  $F(X)$  of several parameters. The Halley method has a cubic rather than quadratic rate of convergence. A simple generalisation of his method to the particular problem of finding the stationary points of a function has little extra cost and can be expected to have significantly enhanced convergence properties. The circumstances which led to its being devised arose when trying to find the parameters for models describing grouped data assumed normal with substantial asymmetric grouping and where the Newton Raphson method led to circling around the optimal root..

**How much information is lost by haplotype uncertainty in SNP case-control analysis?**

**Hae-Won Uh**, Jeanine J Houwing-Duistermaat, Hein Putter, Hans van Houwelingen

Email: h.uh@lumc.nl

*Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*

Due to current high-throughput genotyping technologies, there is considerable interest in using Single Nucleotide Polymorphism (SNP) markers to conduct association studies for complex diseases. Such studies often involve comparing of haplotypic variants between cases and controls. A haplotype-based analysis requires information about which alleles at each genotyped locus are transmitted from which parent. When this phase information cannot be determined with certainty, it can be inferred using statistical algorithm such as EM.

As Hodge et al. (1999) showed, the probability of the individual ambiguity increases with the number of the loci. This ambiguity increases the variance of the estimated haplotype frequencies. Consequently accepting the “best” configuration of haplotypes from EM-algorithm as the “real” haplotype might lead to inaccurate and misleading impression in terms of haplotype effects on disease status.

Therefore we endeavor to develop methods for a better understanding of the missing structure. Using all possible configurations of haplotype reconstruction we first quantify the information loss per individual and per haplotype due to phase ambiguity in the sense of Louis (1982). Then we determine the order of the most informative individuals for additional parental genotyping who would actually contribute most for information gain. This might hopefully lead to more accurate results and would reduce the genotyping costs and efforts.

*Keywords:* case-control studies, haplotype frequency estimates, phase ambiguity, information loss

*References*

- Fedorov VV (1972) Theory of Optimal Experiments. New York: Academic Press;  
Hodge SE, Boenke M, Spence MA (1999) Loss of information due to ambiguous haplotyping. *Nat Genet*, 21:360—361  
Lehmann EL (1983): Theory of Point Estimation. New York: John Wiley & Sons;  
Louis T (1982) Finding the observed information matrix when using the EM algorithm. *J R Stat Soc B*, 44:226-233  
Stram D *et al.* (2003) Modelling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, 55:179-190

## VI(a) – Microarrays

### The Use of Background Signal in Transformation of cDNA-Microarray Measurements

Suzy Van Sanden, Tomasz Burzykowski

Email: suzy.vansanden@luc.ac.be

*Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B 3590 Diepenbeek, Belgium*

As the application field of cDNA-microarrays grows, so does the need for appropriate statistical tools to analyze the signal intensity measurements. Certain procedures are necessary to make the signals from different channels and arrays comparable. According to Cui et al. (2003) this process of removing systematic effects can be separated into three stages: background correction, data transformation, and data normalization. The focus of this paper is on the transformations aiming at the removal of the curvature in the RI-plots (Kerr et al., 2002). In particular, using data from two cDNA microarray experiments we show that some of the previously proposed transformations (like shift, arsinh or linlogshift; see Cui et al., 2003) do not always suffice. Some of them are based on the assumption of a shift between the measurements of the two channels. We explore the possibility of using of background intensity measurements to estimate that shift and to correct for it. The method shows good results when applied to microarray data obtained from the two case studies.

#### References

- Cui, X., Kerr, M.K., and Churchill, G.A. 2003. Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* 2, Article 4.
- Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J., and Churchill, G.A. 2002. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 12: 203–218.

### Graphical Exploration Of Genomic Data As Biomarkers For Drug Activity In Oncological Cell Lines Using Spectral Map Analysis

Dan Lin<sup>1</sup>, An De Bondt<sup>2</sup>, Tamara Geerts<sup>2</sup>, Tim Perera<sup>2</sup> and Luc Bijmens<sup>2</sup>

Email: LBIJNENS@PRDBE.jnj.com

<sup>1</sup>*Center for Statistics Limburgs Universitair Centrum, Diepenbeek, Belgium*

<sup>2</sup>*Johnson and Johnson Pharmaceutical Research and Development a division of Janssen Pharmaceutica, Beerse Belgium*

We illustrate the use of multivariate projection methods to compare them for their ability to identify clusters of biological samples and genes using data on gene expression levels with Affymetrix chips. The dataset contains intensities of thousands of

genes coming from 48 samples in 16 drug treatment combinations applied on a cervix cancer cell line. From those genes a selection is made to focus on two functional pathways including respectively 101 and 99 genes. We compare the spectral map analysis (SMA) method developed by Wouters et al (2003) an adaptation of the method developed by Lewi, 1976 with other two methods: principal component analysis (PCA) and correspondence factor analysis (CFA). PCA has the disadvantage that the resulting principal factors are not very informative to find differential gene expression or clusters of treatment groups, while CFA is sensitive to single large values and has difficulties regarding interpretation of the distances between genes and samples. The importance of weighting for the level of gene expression is highlighted by SMA. Proper weighting allows less reliable data to be down-weighted and more reliable information to be emphasized. It is shown that weighted SMA is better than PCA and CFA in finding clusters of treatment groups and identifying gene expressions related to treatment differences. SMA addresses the gene intensity data in a more appropriate manner than CFA with respect to the scale of measurement. SMA allows the use of a flexible weighting to the gene expressions and treatments.

#### *References*

- Lewi, P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneimittel forschung (drug research)* 26, 1295-1300.
- Wouters L., Göhlmann H., Bijmens L., Kass S., Molenberghs G. and Lewi, P. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* 59: 1133-1141 dec.

### **On the Benjamini-Hochberg method of multiple testing**

**J.A. Ferreira** and A.H. Zwinderman

Email: J.A.Ferreira@amc.uva.nl

*Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre,  
University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands*

The Benjamini-Hochberg procedure has become a popular method for testing several hypotheses simultaneously, especially in contexts where the number of hypotheses is very large such as those of "microarray analysis" and "gene discovery". Under certain conditions, the method is known to control the false discovery rate - the expected proportion of incorrect rejections among rejected hypotheses - and to be less conservative than Bonferroni-type procedures. Several variants of the method (Storey's, for instance) have recently been proposed in the literature with the aim of increasing its power (of rejecting false hypotheses) while maintaining the control of the false discovery rate. In this talk we shall look at some of the properties of the Benjamini-Hochberg method and of another variant of it, namely those related to the convergence of the false discovery rate and of the proportion of incorrect rejections (as the sample size goes to infinity).

## **VI(b) – Non-linear and Generalized Additive Models**

### **Models for growth after bone marrow transplantation during childhood**

**R.B. Geskus**

Email: [statistics@inter.nl.net](mailto:statistics@inter.nl.net)

*Municipal Health Service Amsterdam  
and  
Leiden University Medical Center, Leiden, The Netherlands*

Studies of growth usually involve repeated measurements per individual. Hence a mixed effects model looks like the most suitable approach for analysing this type of data. Each child has his own growth curve, but the characteristics of these curves can be seen as random deviations around a population average pattern. The difference between the measured values and the fitted values are a combination of measurement error and serial correlation.

The basic growth curve pattern is highly nonlinear. When children enter puberty, growth velocity increases. Furthermore, since length can be measured accurately, the amount of measurement error is very small.

The aim of the study was to analyse patterns of growth after total-body irradiation and haematopoietic stem cell transplantation in children with leukemia. Length is measured as lengthSD: the difference between a child's length and the average length of a child of the same age and gender, measured in units of standard deviation. The standard deviation describes the overall variation in length in children of similar age and gender.

Models are presented and compared that analyse growth and include the effect of puberty as a covariate. Since lengthSD is used to quantify growth, puberty influences the outcome variable in two ways: not only can the individual himself enter puberty, but if children of similar age and gender enter puberty, that will also influence the child's lengthSD.

Since measurement error is very small, outcomes are highly sensitive to the model used. Some examples of models will be given. Finally it was decided not to use the model with the best fit, but to use the model that to our opinion better described the growth processes.

**Analysis of circular responses using generalised additive models for location, scale and shape**

**M Cortina Borja**

Email: M.Cortina@ich.ucl.ac.uk

*Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, UCL*

**Background** Circular responses arise in many epidemiological studies where the date of an event is the main outcome. The von Mises distribution is a member of the exponential family and may be used to analyse unimodal and symmetric seasonal patterns, which in many instances correspond to annual cycles. It has one location parameter corresponding to the mean direction, and one scale parameter which defines its concentration with respect to the mean direction. Rigby and Stasinopoulos (JRSS-C in press) have defined a class of generalised additive models for location, scale and shape (GAMLSS) and provided a flexible library in R to fit and analyse such models. Given suitable link functions GAMLSS can incorporate a wide range of distributions.

**Aim** To analyse population-based data on the timing of sudden infant death syndrome (SIDS) and suicide taking advantage of the GAMLSS class and its associated estimation, diagnostics, and model selection methods to fit regression models for circular responses.

**Methods** Using census data, the month of death of all cases of SIDS in the UK between 1983 and 1998, and the dates of death of all suicides by gender in Scotland between 1979 and 2001 were modelled using smooth functions of year of death and extreme daily temperature.

**Results** Regression models assuming a von Mises response fitted the data adequately, suggesting the presence of a single, symmetric underlying annual process in both datasets. Non-symmetric and multimodal models for circular data were also explored. SIDS had a constant mean direction corresponding to deaths in January and a smooth, non-linear, association between the scale parameter and year of death suggesting the effect of public health policies. For suicides there was a constant mean direction around June and significant effects of year of death, gender and daily temperature on the scale parameter.

**Conclusion** GAMLSS provided a convenient framework for building and assessing regression models for circular responses. Strong seasonal patterns were found in presentation of both SIDS deaths and suicides. These analyses may lead to a better understanding of the aetiology of the two diseases and offer new strategies for prevention and treatment.

**The use of bivariate generalised additive models to investigate the effectiveness of cycle helmets.**

**Paul Hewson**

Email: paul.hewson@plymouth.ac.uk

*School of Mathematics and Statistics, University of Plymouth, Drake Circus, Plymouth  
PL4 8AA*

Case control studies indicate that cycle helmets can be effective in reducing morbidity due to head injury. Population studies rarely find evidence of improved morbidity in a pattern that can be associated with known patterns in helmet wearing, such as the large increases in helmet wearing which followed compulsion in Australia or New Zealand. The legislation itself, as well as the very advocacy of cycle helmets can be a controversial issue where there is a need to balance the health promotion potential of exercise with the obvious deterrent effect. Whilst the ecological fallacy is an obvious concern in population level studies, bias and generalisability concerns also arise from the published case control studies. Case control studies have featured large numbers of children injured in lower severity incidents often involving no other vehicle and therefore we will consider a population level analysis of child cyclist casualty data.

Survey data available from England enumerates cycle helmet wearing rates since 1994, monthly time series can provide information on road traffic injuries reported to the police throughout this period. More importantly, hospital episode statistics can provide information on patients injured either as pedestrians or whilst cycling who required hospital treatment. In this case, it is also possible to obtain data on the occurrence of head injury. It should be noted that comparisons with police reported data reveal a large level of cycling morbidity that has not been reported through the police system either because of "under-reporting" or because the injury took place away from the road.

We contrast results on child pedestrians and child cyclists using bivariate generalised additive models (Yee and Wild, 1996) to examine evidence for differential trends in time in terms of head injury that can be associated with cycle helmet wearing rates. There is clear evidence of a decline in morbidity in terms relative numbers of patients suffering head injury over time. Although the proportion of head injuries has fallen, there was evidence of a relative increase in the number of hospitalised young male cyclists, a feature that is also apparent when considering the police collected data. There was rather less evidence that head injuries had increased significantly among young male cyclists; it should be noted helmet wearing rates among young male cyclists have fallen both in absolute terms and have fallen considerably relative to young female cyclists. However, a different picture is seen in terms of young pedestrians. Again there is an overall decrease in the proportion of children suffering head injury. Whilst the gender balance appears constant since 1996 the head injury rates appear to be falling faster in young male pedestrians than young female pedestrians.

The observation that the proportion of children hospitalised as pedestrians or cyclists with head injuries has fallen is clearly welcome. However, there is no clear evidence to associate the published rates of cycle helmet wearing with this fall. Two notable

features emerge from these data; that the proportion of hospitalised young female pedestrians with head injury has stopped declining as quickly as with young males and secondly that relative numbers of male cyclists injured have been increasing. One possible explanation for these results is that helmet wearing is acting as marker for underlying beliefs and behaviours, falling wearing rates among young males may indicate either more risk taking (such as performing stunts), or more acceptance of risk (such as commuting in traffic). The role of primary injury prevention strategies can be discussed in this context.

#### *References*

Yee, T.W. and C.J. Wild (1996) Vector generalised additive models. *Journal of the Royal Statistical Society Series B, Methodological* 58:481-493

### **Estimating incidence of HIV infection in women using serial prevalence data from antenatal clinics**

**Charlotte Sakarovitch**<sup>1,2</sup>, Ahmadou Alioum<sup>1</sup>, Didier Ekouevi<sup>3</sup>, Philippe Msellati<sup>4</sup> and François Dabis<sup>2</sup>

Email: Charlotte.Sakarovitch@isped.u-bordeaux2.fr;

<sup>1</sup> *Unité EMI INSERM 03 38, Université Victor Segalen, Bordeaux – France*

<sup>2</sup> *Unité INSERM 593, ISPED, Université Victor Segalen, Bordeaux – France*

<sup>3</sup> *Projet ANRS 1201, Programme PACCI, Abidjan - Côte d'Ivoire*

<sup>4</sup> *UR 036, Institut de Recherche pour le Développement, Montpellier - France*

Incidence of HIV infection is a key indicator for the planning and evaluation of national AIDS control programs, but direct measures are difficult to obtain in Africa. By contrast, seroprevalence data on specific sentinel groups such as pregnant women are widely available.

Ades and Medley (1994) developed a method to estimate age and time-specific incidence of HIV infection, using a maximum likelihood estimation method and serial prevalence data, accounting for the effect of differential inclusion rate according to HIV status. Our paper presents a modification of Ades and Medley method in three respects. Firstly, we introduced a penalized component in the likelihood to smooth the incidence curves. Secondly, we improved the model of the relative inclusion rate adapted to pregnant women, studying its sensitivity on the hypotheses. Thirdly, we computed simulations in order to assess the variability of the estimates due to the sample size and the identifiability of the parameters estimated. We applied this method to prenatal HIV testing data recorded for several years in Abidjan, Côte d'Ivoire, to estimate the HIV annual incidence rate among women aged between 12 to 40 years old, from the beginning of the epidemic to nowadays. We highlight the relevance of such a method in monitoring the HIV epidemic in Africa.

#### *References*

Ades AE, Medley GF. Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York City. *Stat Med* 1994;13(18):1881-94.

## Posters

### QTL mapping in autotetraploid populations

C. A. Hackett<sup>1</sup> and J. E. Bradshaw<sup>2</sup>

Email: christine@bioss.ac.uk

<sup>1</sup>*Biomathematics and Statistics Scotland, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA*

<sup>2</sup>*Genome Dynamics Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA*

Statistical methods for QTL mapping in diploid species have advanced rapidly over recent years, but methods for QTL mapping in autotetraploid species have received less attention because of the complexities of inheritance in such species. Here we propose a maximum likelihood approach for QTL mapping in a full sib autotetraploid population.

The first step is to use marker phenotype information from the parents and offspring to infer the offspring genotypes for the chromosome. We use a branch and bound approach to identify genotypes with the minimum number of crossovers consistent with the marker phenotype data. From here, we can infer the possible QTL genotypes and their probabilities conditional on the marker data at any position between the markers.

We then use the EM algorithm to assess the likelihood of a QTL at each location along the chromosome to produce a likelihood profile. In an iterative process, trait values are regressed on the QTL genotypes, weighted by the conditional probabilities, and the conditional probabilities are then updated. This continues until the likelihood converges.

This method is applied to map QTLs for foliage blight and maturity in a population of tetraploid potato, and to investigate the relationship between these traits.

### An R routine for fitting time-varying effects in Cox and Reduced Rank models

Aris Perperoglou, Saskia le Cessie, Hans C. van Houwelingen

Email: a.perperoglou@lumc.nl

*Department of Medical Statistics, Leiden University Medical Center, The Netherlands*

Cox's proportional hazard model has become the standard method to analyze time to event data. In cases where the proportionality assumption does not hold, the model can be extended to include time dependent covariates that change through time or time varying effects of fixed covariates. Although standard statistical software includes routines to fit proportional hazards, there are still serious limitations when it comes to non-proportional models.

S-plus and R statistical packages have implemented a counting process setup to estimate Cox models with time varying effects. The data set has to be rearranged in a repeated measurement setting: the time is divided to small time intervals where a single event occurs and the history of each subject (and its covariates values) is repeated for all the sets that the subjects are still under observation. This is the known (Tstart,Tstop] algorithm implemented in Therneau's Survival library (S-plus). A similar routine is used in Stata.

However, the expansion of a data set can lead to a larger set which is hard to handle. The size of such a data set (assuming no ties or censoring) is given by  $1/2(n(n+1)pq)$ , where  $n$  is the number of subjects,  $p$  the number of covariates and  $q$  the number of time functions. That results to data sets of enormous size which cannot be treated even with fast modern computers.

We propose the use of a fast and efficient algorithm, written in R, which works on the original data without the use of an expansion. The computations are done on the original data set, with significant less memory resources used.

This improves the computational time by orders of magnitude and at the same time the model fitting can be done without any restriction on the size of the original data.

We will illustrate this method on a large data set of 2700 breast cancer patients, and compare the computational times on simulated data of up to 10.000 cases. We will also present an efficient algorithm of dimension reduction of Cox models with time varying effects via reduced rank hazard regression.

*References:*

Perperoglou A., le Cessie S., van Houwelingen H.C. Reduced-rank hazard regression for modelling non-proportional hazards, submitted to Statistics in Medicine.1

**A Bayesian calibration model to predict fungal contamination levels in wheat seed based on a PCR assay**

**Adrian M.I. Roberts<sup>1</sup>**, Chris M. Theobald<sup>1</sup> and Marian McNeil<sup>2</sup>

Email: adrian@bioss.ac.uk.

<sup>1</sup>*Biomathematics and Statistics Scotland, James Clerk Maxwell Building, King's Buildings, Edinburgh, EH9 3JZ, UK*

<sup>2</sup>*Scottish Agricultural Science Agency, Edinburgh, UK*

In the UK, samples of seed are sent by cereal seed producers to official seed testing stations for disease testing. Common bunt, *Tilletia caries*, is an important fungal disease of wheat, transmitted on the surface of the seed. Control of the disease relies on treatment of seed with fungicide. The decision to treat is based on the level of disease found in the seed, usually assessed by a microscopic assay. Modern molecular techniques offer a high-throughput alternative. Quantitative polymerase chain reaction (PCR) technology has been developed to quantify the *T. caries* DNA in a sample of seed.

Understanding of the biology indicates that the amount of *T. caries* DNA in a sample should be proportional to the number of spores present. Typically, the quantity of DNA found is related to the level of seed contamination using linear regression, after log-transforming both variates.

We present a Bayesian calibration model relating the amount of DNA with the contamination level. This accounts for the sampling variation inherent in both types of assay as well as measurement error attributable to the PCR assay both within and between run. The proposed model provides superior prediction to the straightforward linear regression approach and can accommodate censored DNA values and zeros in the microscopic assay.

#### *References*

- Edwards, K.J., Logan, J.M.J. and Saunders, N.A. (2004) Real-time PCR: an Essential Guide. Wymondham: Horizon Bioscience.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) Markov chain Monte Carlo in Practice. London: Chapman and Hall.
- McNeil, M., Roberts, A.M.I., Cockerell, V. and Mulholland, V. (2004) Real-time PCR assay for quantification of *Tilletia caries* contamination of UK wheat seed. *Plant Pathology* 53, 741-750.
- Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2004) WinBUGS User Manual, Version 1.4.1. Cambridge: Medical Research Council Biostatistics Unit.

### **Performance of Class Prediction Methods in a Microarray Setting**

Suzy Van Sanden, **Dan Lin** and Tomasz Burzykowski

Email: dan.lin@luc.ac.be

*Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B 3590 Diepenbeek, Belgium*

Dudoit et al. (2002) investigated performance of several class prediction methods when applied to microarray data. In their investigation, Dudoit et al. used real-life microarray datasets. This allowed them to evaluate performance of the methods subject to the complexity of microarray measurements. However, due to the use of a few real-life datasets, only a limited number of settings could be evaluated. Second, the true classification, as well as the set of truly differentially expressed genes, was unknown. In order to overcome these limitations, we conduct a simulation study, in which a linear mixed effects model is used to simulate microarray data under several different scenarios. The parameters of the model are chosen based on a real-life dataset. We simulate gene expression level measurements coming from cDNA microarrays with a common reference design. Using the BW criteria (Dudoit et. al, 2002) to pre-select genes, we compare classification methods like classification trees, k-nearest neighbours and discriminant analysis with respect to their ability to discriminate between two classes of biological samples in different experimental settings. Preliminary results show that methods like random forest and diagonal discriminant analysis yield lower misclassification error. The results also indicate that the error very much depends on the

choice of the genes which are to be used in the class prediction. Thus, appropriate gene-selection procedures are needed to obtain a satisfactory performance of class prediction methods.

*References*

Dudoit S., Fridlyand J., and Speed T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 98, 77-87.

**Optimization of sampling schemes for vegetation mapping using fuzzy classification**

**A. Stein**, R. Tapia and W. Bijker

Email: Alfred.Stein@wur.nl

*Biometris, Wageningen University, PO Box 100, 6700 AC Wageningen, The Netherlands*

*International Institute for Aerospace Survey and Earth Sciences (ITC), PO Box 6, 7500 AA Enschede, The Netherlands*

Image segmentation segments an image, for example obtained with remote sensing, into different segments that may have a physical interpretation. Modern segmentation procedures allow us to not only delineate the classes where we can be certain about class memberships, but also to define areas where we are less certain. This presentation considers the design of an optimal sampling scheme for a multivariate fuzzy-*k*-means classifier. Fuzzy classification is applied to delineate vegetation patterns from remote sensing data. The confusion index distinguishes sub-areas with high uncertainty due to class overlapping from those with low uncertainty. These sub-areas govern allocation of sample points. A simulated annealing approach minimizes the mean of shortest distances between samples. Optimization was done by prioritizing the survey to areas with high uncertainty. The methodology is tested on a site located in the Amazonian region of Peru. It resulted into an almost equilateral triangular scheme at those parts of the area where uncertainty was highest. The study shows that optimal sampling can be successfully combined with fuzzy classification, using an appropriate weight function.

**Modeling SAGE data with Poisson mixtures**

**Helene H. Thygesen** and Aeilko H. Zwinderman

Email: h.h.thygesen@amc.uva.nl

*University of Amsterdam*

SAGE (Serial Analysis of Gene Expressions) produces gene expression measurements on a discrete scale, due to the finite number of molecules in the sample. This means that part of the variance in SAGE data should be understood as the sampling error in a binomial or Poisson distribution, whereas other variance sources, in particular

biological variance, should be modeled using a continuous distribution function, i.e. a prior on the intensity of the Poisson distribution. In two recent studies, the Poisson model was used for gene clustering[1] and for identification of tumor specific genes[2]. Both studies showed that the Poisson model performed better than alternative models. However, both authors modeled the data for each gene separately, and as noted by[3], there is a need for a model that accounts for the distribution of the expression levels across genes. One challenge is that such a model predicts a large number of genes with zero expression, which cannot be observed. This calls for a censored Poisson model.

We present a censored Poisson model with a conjugate Gamma prior and a non-parametric component accounting for a small population of very strongly expressed genes. We show how the parameters can be estimated with two different heuristics. One heuristic provides robust estimates of the shape and rate in the Gamma distribution in the uncensored case. The other heuristic works in the censored case as well, but only the rate parameter can be reliably estimated. We argue that this is an inherent problem with this kind of data. Fortunately, the shape parameter has little influence on the posterior distribution of the gene expression, so for most practical purposes it doesn't matter.

We analyzed 73 SAGE libraries [4] from human tissue samples, in which non-expressed genes were included because genes that were expressed in at least one library were included in all libraries. This means that both the censored and the non-censored model could be used. We compare the results from the two models and show how the posterior distribution of the gene expression can be estimated and how it can be used to identify tumor-specific or tissue-specific genes.

*References:*

- [1] Cai L et al (2004) Clustering analysis of SAGE data using a Poisson approach. *Genome Biology* 5:RS1
- [2] Vencio RZN et al (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expressions (SAGE). *BMC Bioinformatics*, 5:119
- [3] Baggerly KA. et al (2003) Differential expression in SAGE: Accounting for normal between-library variation. *Bioinformatics* 19(12): 1477-1483
- [4] Caron H. et al (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291(5507):1289-92

A new test statistic to deal with multiple testing in association between a disease and a multi-allelic marker.

**R. el Galta**<sup>1</sup>, T. Stijnen<sup>2</sup> and J.J. Houwing-Duistermaat<sup>1</sup>

Email: r.elgalta@lumc.nl

<sup>1</sup>*Department of Medical Statistics and Bioinformatics, LUMC, Leiden, The Netherlands*

<sup>2</sup>*Department of Epidemiology and Biostatistics, Erasmus MC, Rotterdam, The Netherlands*

In genetic association studies, the observed case-control data are often summarized in  $m$  by 2 tables, with  $m$  ( $>2$ ) the number of alleles or haplotypes. Under the alternative, the assumption can be made that one or a few alleles are positively associated with the disease. For this situation the classical Pearson's chi-square has small power, especially when the number of alleles is large.

One approach to obtain a more powerful test-statistic is to take the maximum of the chi-square statistics of one allele versus the rest ( $Z_{\max}$ ). When more than one allele is associated, one can also take the maximum of chi-square statistics corresponding to all possible 2-by-2 tables of combinations of alleles against the rest ( $Z_{\text{clump}}$ ) [Sham & Curtis. *Ann Hum Genet.* 59:97-105, 1995.] P-values can be computed by using Monte-Carlo simulations.

Another approach is to sum over the possible associated alleles. Terwilliger proposed to use a weighted sum of conditional likelihood's given that an allele is associated. As weights he used the allele frequencies [*Am. J. Hum. Genet.* 56: 777-787, 1995.] We derived the corresponding score statistic. It appeared that this score statistic does not make any assumption about the number of associated alleles. The significance level of the score statistic can be easily derived by means of Monte-Carlo methods.

We compared heuristically and empirically the power of the new score test, Pearson's chi-square,  $Z_{\max}$  and  $Z_{\text{clump}}$ . The performance of the statistics was studied for two models, namely (1) one positively associated allele and (2) two positively associated alleles. As illustration we applied the tests to a published case control study on association between COL2A1 gene and radiographic osteoarthritis [*Ann. Hum. Genet.* 63:393-400, 1999.] The observed marker had five alleles with frequencies  $\geq 0.05$ . The remaining alleles were combined into one group. One of the alleles was positively associated.  $Z_{\max}$ , the score statistic and the classical Pearson's chi-square statistic gave an empirical p-value of about 0.01.  $Z_{\text{clump}}$  gave a less significant p-value of 0.04. We conclude that the new score test provides a good power regardless the number of positively associated alleles and regardless the number of marker alleles.

**Analysis of a large family study ascertained through hypertensive probands**

**Peter Avery** and Bernard Keavney

Email: P.J.Avery@newcastle.ac.uk

*University of Newcastle*

Data on extended families ascertained through individuals with hypertension have been analysed. The aim of the study was to look at the genetic and environmental effects on various characters, mainly those related to blood pressure and heart function. The problem of which environmental covariates to correct for will be discussed; this being a relatively complex problem due to the large number of missing values for some variables and the fact that correlation between variables can have a genetic or an environmental cause. The results of a comprehensive genome-wide scan will be examined. The problem of allowing for multiple testing and the relatively large effect of small deviations from Normality assumptions will be highlighted. The alternative approach of looking for candidate genes will also be discussed.

**Estimating hidden population sizes in the presence of covariates and prior information**

**Ruth King**<sup>1</sup>, Sheila Bird<sup>2</sup>, Steve Brooks<sup>3</sup>, Sharon Hutchinson<sup>4</sup> and Gordon Hay<sup>5</sup>.

Email: ruth@mcs.st-andrews.ac.uk

<sup>1</sup> *Centre for Research into Ecological and Environmental Modelling, University of St. Andrews*

<sup>2</sup> *Medical Research Council Biostatistics Unit, Cambridge*

<sup>3</sup> *Statistical Laboratory, University of Cambridge*

<sup>4</sup> *Public Health and Health Policy Section, University of Glasgow*

<sup>5</sup> *Centre for Drug Misuse Research, University of Glasgow*

We consider a stratified approach to estimating the number of injector drug users (IDUs) in Scotland between 2000-2002. Data are collected from four different sources, and the overlaps between the four different sources (or captures) are the basis for estimating the number of hidden (or uncaptured) IDUs. We have additional covariate information for all individuals observed, corresponding to their age, sex and locality. We use a log-linear model to describe the relationship between the probability of being observed by each combination of sources and the different sources and covariates, including possible interactions. The estimates obtained for each combination of covariates are combined with the number observed to provide an estimate of the total population. In addition, from the estimates obtained, annual drug-related death rates can be estimated. We consider a Bayesian approach, which allows the direct incorporation of prior information that is available from independent studies. We use Bayesian model discrimination techniques to determine, which, if any, interactions between the data sources and/or covariates are supported by the data. These allow us to obtain model-averaged estimates of the statistics of interest, taking into account both parameter and model uncertainty.

**Long-term spatio-temporal variation in abundance of the garden tiger moth (*Arctia caja*) during a population decline.**

Kelvin Conrad, Ian Woiwod and **Joe N. Perry**

Email: joe.perry@bbsrc.ac.uk

*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK*

The garden tiger moth (*Arctia caja*) is a well-known and attractive moth that was once regarded as common in the UK. It is polyphagous and overwinters as small larvae. Over the past 35 years numbers of garden tiger moth have declined severely. Conrad *et al.* (2002) examined this decline using almost 2700 trap-years of data collected at 407 Rothamsted Insect Survey light-traps from all over Great Britain, spanning 1968-98. The annual collated index, commonly used to assess relative changes in lepidopteran abundance, confirms this long-term trend. However, our examination of the geometric mean abundance across occupied sites has revealed a somewhat different pattern. The annual geometric mean fluctuated around approximately 4.2 individuals/year until 1983, and then fell suddenly to approximately 3.0 individuals/year and continued oscillating near that new, lower level thereafter. In contrast, the proportion of sampled sites occupied (incidence) remained high at approximately 0.60 until 1987-88, when it fell to 0.46 and continued to decline. Thus, garden tiger moth density fell across Great Britain initially in 1983, but the moths did not begin disappearing from individual sites until several years later. Populations may have hovered near some threshold level with local extinctions lagging behind local declines in abundance. Since 1989, the garden tiger moth has remained at low densities and low incidence. The general trend over time has been for the species to become more restricted to the north and west and almost completely absent from the Southeast. Multiple regression analysis using monthly mean values from the Central England data set suggest garden tiger moth abundance is adversely affected by warm wet winters and warm springs. However, the sudden collapse in abundance between 1983 and 1984 is more likely associated with an extreme meteorological event. The sudden drop in abundance and the four to five year lag before the accompanying decrease in incidence underscore the value of long-term monitoring in determining changes in abundance and distribution, even of species considered to be widespread and common. The poster can be viewed online at: <http://www.squirreldance.f9.co.uk/kfconrad/poster/DECPoster2001KFC.htm>

*Reference*

Conrad, K.F., Woiwod, I.P. & Perry, J.N. (2002). Long-term decline in abundance and distribution of the garden tiger moth (*Arctia caja*) in Great Britain. *Biological Conservation*, **106**, 329-337.

**MiCoSPA: Microbial Pest Control for Sustainable Peri-urban/urban Agriculture  
in Latin America** (A European Union Funded Project (ICA4-2001-10185))

**Janet Riley**

Email: janet.riley@bbsrc.ac.uk

*Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, AL5  
2JQ, UK*

Through multidisciplinary research in Europe, Cuba, Mexico and Australia this project aims to provide socially and economically sustainable strategies for pest control in peri-urban and urban vegetable production in Latin America.

A significant limitation to the sustainable production of crops world-wide is the management of foliar and soil invertebrate pests. Safe and effective control strategies as alternatives to expensive and often environmentally damaging pesticides are urgently required. The development of fungi as microbial control agents offers potential in this regard.

We will address the current limitations to uptake of fungal control agents which include an incomplete knowledge of the parameters affecting transmission, lack of field and farm-scale evaluation and quality control problems in production, formulation and application.

I am addressing the first work package: to evaluate farmer and community agro-ecosystem practices and biotechnology potential in Latin America using socioeconomic surveys. These surveys are directed at two entirely different systems in Mexico (Guanajuato: large scale farms and Mixquic: small scale farms) and in Cuba (CREES, governmental farmers, private farmers and workers). The outcome will be a set of surveys for the first year and I will redo the surveys in the third year to assess the difference between the first and third years.

**Estimation of the false discovery rate in functional genomic studies**

**Cyril Dalmasso & Philippe Broët**

Email: dalmasso@vjf.inserm.fr

*INSERM U472 – Faculté de Médecine Paris-Sud, 16, avenue Paul Vaillant –Couturier,  
94807 Villejuif*

New transcriptome-oriented biotechnologies make nowadays possible the comparative analysis of thousands of genes expression in parallel for selecting relevant genes the transcriptional changes of which are related to a clinical or biological outcome. In such a case, a major multiple testing problem arises due to the fact that a large number of statistical tests are simultaneously performed.

Up to now, statistical procedures devoted to this multiple testing problem mostly focused on controlling or estimating false positive error criteria. For cDNA microarray

experiments, the most used criterion is nowadays the false discovery rate (FDR) [1] which is defined as the expected proportion of false discoveries among all discoveries.

Here, we consider the framework of estimating procedures for the FDR based on the marginal distribution of the p-values without any assumption on the conditional distribution related to the modified genes. Thus, estimators of the FDR obtained from the formula introduced by Storey [2] are necessarily conservatively biased. Indeed, only an upper bound estimate can be obtained for the key quantity  $\pi_0$ , which is the probability for a gene to be unmodified. In this setting, the main existing procedures are QVALUE [2], BUM [3] and SPLOSH [4].

The method Location Based Estimator (LBE) [5] that we have proposed is a class of estimators for an upper bound of  $\pi_0$  based on the expectation of the transformed p-values and from which we can obtain results on its asymptotic distribution. As for the procedures QVALUE, BUM and SPLOSH, our procedure does not make any assumption on the distribution related to modified genes. From our proposed estimators, we can easily obtain estimators of the FDR.

After discussing the three methods QVALUE, BUM and SPLOSH, we will present the method LBE. Then, results from a simulation study that compare the four procedures in finite samples will be shown. These results show the good performances of the proposed estimator which has the best mean square error in most of cases. The different methods are applied to real microarray datasets. Finally, extensions of this work will be presented.

#### References:

- [1] Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, 57, 289-300.
- [2] Storey JD, Tibshirani R. (2003b) Statistical significance for genome-wide studies. *Proc Natl Acad Sci*, 100, 9440-9445.
- [3] Pounds S, Morris SW. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*; 19(10), 1236-42.
- [4] Pounds S, Cheng C. (2004) Improving false discovery rate estimation. *Bioinformatics*; 20(11), 1737-45.
- [5] Dalmaso, C; Broet, P.; Moreau, T. (2004) A simple procedure for estimating the false discovery rate. *Bioinformatics*. Advance Access published on 12 Oct.

**Analysis of Microarray Data in a Dose-Response Setting: Resampling Based Multiple Testing**

Luc Bijmens<sup>1</sup>, Ziv Shkedy<sup>2</sup> and Dan Lin<sup>2</sup>

Email: ziv.shkedy@luc.ac.be

<sup>1</sup>*Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica N.V. Beerse, Belgium*

<sup>2</sup>*Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B-3590 Diepenbeek, Belgium*

The biotechnology of DNA microarrays allow the monitoring expression levels of thousands of genes simultaneously, and identifying those genes that are differentially expressed. As a result type I error (the probability for false identification) increase sharply when the number of tested genes gets large. In this talk we focus on a dose-response setting in which DNA microarrays are available for four dose levels (3 microarrays at each dose level). A gene is differentially expressed if there is a trend (with respect to dose) of the gene intensity. We discuss several approaches to test the null hypothesis of no dose effect versus an order alternative. Resampling based False Discovery Rate (Benjamini and Hochberg, 1995, Ge et al. 2003, SAM et al. 2003) and resampling Family-Wise Error Rate (Westfall and Young, 1993) are used for controlling type I error.

*References*

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R.Statist. Soc. B* 57, 289-300.
- Ge, Y., Dudoit, S. and Speed, P.T. (2003). Resampling based multiple testing for microarray data analysis. *University of Berkeley, technical report #633*.
- Westfall, P.H. and Young, S.S (1993). *Resampling based multiple testing*. Wiley.