

# **CURRENT RESEARCH PROJECTS**

## **– BIOMETRICS –**

**(Last updated 27 August 2009)**

## CONTENTS

<b>MFX – Mixed effects models .....</b>	<b>3</b>
<i>Damian Collins.....</i>	<i>3</i>
<i>[Note: This is a collaborative project with the Queensland Department of Employment, Economic Development &amp; Innovation (David Butler) and VSN-International, Hemel-Hempstead, UK (United Kingdom (Simon Harding)).....</i>	<i>3</i>
<b>DiGger experimental design generator.....</b>	<b>4</b>
<i>Dr Neil Coombes.....</i>	<i>4</i>
<b>ASReml .....</b>	<b>5</b>
<i>Brian Cullis.....</i>	<i>5</i>
<b>Statistics for the Australian Grains Industry .....</b>	<b>6</b>
<i>Brian Cullis.....</i>	<i>6</i>
<b>Linear mixed models with mixture distributions and an efficient hybrid algorithm for factor analytic mixed models.....</b>	<b>7</b>
<i>Simon Diffey, Brian Cullis, Alison Smith.....</i>	<i>7</i>
<b>Enhancing NIRS calibrations of grains for livestock – Statistical analysis and advice.....</b>	<b>8</b>
<i>Sharon Nielsen .....</i>	<i>8</i>
<b>Development and implementation of <i>Katmandoo</i>, bioscience data management system.....</b>	<b>9</b>
<i>Mr Avishesh Shrestha.....</i>	<i>9</i>
<i>[Note: This is a collaborative project with the Queensland Department of Employment, Economic Development &amp; Innovation (Mr David Butler and Mr David Rodgers) and Diversity Arrays Technology Pty Ltd (Mr Grzegorz Uszynski)] .....</i>	<i>9</i>

## Research Project Description

### Project Title

MFX – Mixed effects models  
(Activity #773)

### Principal Investigators

Damian Collins

*[Note: This is a collaborative project with the Queensland Department of Employment, Economic Development & Innovation (David Butler) and VSN-International, Hemel-Hempstead, UK (United Kingdom (Simon Harding))]*

### Funding sources

Industry & Investment NSW and industry funds (income from ASReml).

### Objective

The objective is to produce a well documented library of core routines for mixed model analysis, which is freely available to those producing mixed model applications.

### Summary

The linear mixed models package ASReml has been the flagship research project for the I&I NSW Biometrics section. It uses a highly computationally efficient approach (the average information (AI) algorithm) for fitting linear mixed models. Linear mixed models are statistical models which allow for multiple (and possibly complex) sources of variation. These models have wide application in the analysis of biological and agricultural data. In particular, the ASReml software has been particularly useful within I&I NSW for datasets arising from large scale crop variety evaluation programmes and animal breeding. VSN International have been marketing and selling ASReml since February 2003. The code base is maintained primarily by Dr Arthur Gilmour, who is now retired from I&I NSW.

With the need to extend the computational approach to a greater range of mixed models and associated statistical methodologies, such as non-linear or generalized linear mixed models, a new code base or platform is required. This project has been called MFX (an acronym for “mixed effects” models). The employment of a new code base also provides an opportunity for the code base to be modular and cater for future extensions.

Damian Collins is a core developer of the MFX project group. The other two core developers, David Butler (Qld DEEDI) and Simon Harding (VSN-International), have extensive experience in the commercial development of software. In addition, the core developers are supported by Brian Cullis (supervisor, Biometrics) and Robin Thompson (Rothamsted Research), who are both internationally recognized in the development of linear mixed model applications.

The primary benefit of this project to I&I NSW and to the broader scientific community is to provide a code base for incorporation into other research software and applications.

## Research Project Description

### Project Title

DiGGer experimental design generator  
(Activity #774)

### Principal Investigator

Dr Neil Coombes

### Funding Sources

Industry & Investment NSW and the Grains Research & Development Corporation.

### Objective

The objective of this project is to provide a flexible experimental design generation tool which will incorporate complex blocking, correlation between experimental units and unequal treatment replication in finding efficient designs.

### Summary

DiGGer was developed to provide experimental designs which are suited to the patterns of correlation typically found in cereal breeding field experiments. The program uses a modified Reactive Tabu search algorithm to find efficient designs for user specified correlation and blocking. DiGGer is a Fortran program which can be used as a standalone or as a package in the R programming environment.

The DiGGer package for R allows complex searches to be set up in functions and script files which provide simplicity of use for inexperienced users. The R framework also allows users to develop their own functions to interface with the DiGGer search program to solve specific design problems. Functions for incomplete block, split-plot, strip-plot and partially replicated designs are available and will continue to be improved in future releases of the DiGGer package.

DiGGer software is available at <http://www.austatgen.org/files/software/downloads/>.

## Research Project Description

### Project Title

ASReml  
(Activity #775)

### Principal Investigator

Brian Cullis

### Funding Sources

Industry & Investment NSW and the Grains Research & Development Corporation.

### Objective

To continuously upgrade and improve the ASReml software platform.

### Summary

The project is a collaboration between Rothamsted Research and I&I NSW. Implementation of the core of our software package ASReml into GENSTAT and S/R statistical languages has been undertaken by Sue Welham and David Butler respectively. VSN-International has an agreement with I&I NSW and Rothamsted Research to distribute ASReml. The major stakeholders of the project (other than the developers and distributors) are the users. Users of ASReml provide an important role in the project. They often provide inspiration for new enhancements and new initiatives as the role of linear mixed models is further expanded into new and developing technologies such as microarrays. Users are guaranteed support as part of the license for ASReml. This support is provided by VSN-International, I&I NSW and Rothamsted Research. Major partnerships are also supported with individual consultancies (e.g. Pioneer Hi-Bred and BSES Ltd).

The ASReml project encompasses the design, development and implementation of a software platform undertaking high level statistical analyses using the linear mixed model. The software integrates a range of unique statistical algorithms developed by the project team with robust and efficient sparse matrix methods coded in a contemporary framework. It is an original and complex system representing 45 person years of international collaboration.

## Research Project Description

### Project Title:

Statistics for the Australian Grains Industry  
(Activity #762)

### Principal Investigator

Brian Cullis

### Funding Sources

Industry & Investment NSW and the Grains Research & Development Corporation (Project No. DAN 00124).

### Objectives

1. This project will deliver statistical support and innovative statistical technologies and statistical software to the grains industry of Australia. Its primary remit is to provide high level support to national, public pre-breeding and breeding programs and the National Variety Testing (NVT) system, AWBMMP-GA and other GRDC funded research projects.
2. The development of statistical methods and algorithms for implementation into software of relevance to the activities and collaborations in components 1 and 3.
3. Course and workshop preparation and delivery. This output also includes a software component, as the adoption and training in statistical methods can only occur with suitable user-friendly, yet powerful software.

### Summary

This project will deliver statistical science to the grains industry of Australia. Its primary remit is to provide high level support to plant improvement programs and, in collaboration with the companion I&I NSW service agreement, provide support to the National Variety Trials (NVT) system. The project will also deliver high quality statistical support to the pilot breeding project at Barley Australia and the defect elimination project. Key marker projects which include the AWBMMP-Genetic Analysis module, CMMP and the (future) PMMP will also be supported. Breeding programs for sorghum, durum wheat, triticale, narrow leaf lupins, Pulse Breeding Australia (PBA) and Barley Breeding Australia (BBA) will benefit from this statistical approach.

The project will continue to develop innovative methodologies and implement these in user-friendly software. It will also promote scientific collaboration and sound statistical thinking through a nationally coordinated training scheme in statistics for plant improvement.

## Research Project Description

### Project Title:

Linear mixed models with mixture distributions and an efficient hybrid algorithm for factor analytic mixed models  
(Activity #776)

### Principal Investigator

Simon Diffey, Brian Cullis, Alison Smith

### Funding Sources

Industry & Investment NSW and the Grains Research & Development Corporation.

### Objectives

1. Implementation of a stochastic EM algorithm for fitting linear mixed models with mixture distributions to QTL data.
2. Implementation of an efficient hybrid EM / AI algorithm for the fitting of factor analytic models to plant breeding data.

### Summary:

This research project will make a contribution in both the early (development) and later stages (evaluation) of the plant breeding process.

#### *Development of experimental breeding lines*

Molecular marker technology plays an important role in the development of experimental breeding lines as appropriate markers are used to screen breeding lines for such things as disease resistance, aluminium tolerance, and protein content. Early screening of experimental breeding lines leads to genetic enrichment as lines with undesirable characteristics are discarded early in the breeding process. The identification of appropriate molecular markers often involves a quantitative trait loci (QTL) analysis and an aim of this project is the development of a more efficient method of identifying the number, location, and genetic effect of these loci. An area that shows promise in meeting this aim is the use of mixture models within an existing linear mixed model framework. Traditional methods for fitting linear mixed models cannot be applied, but a stochastic version of the EM algorithm shows promise in overcoming this problem.

#### *Evaluation of experimental breeding lines*

In the later stages of the plant breeding process a large number of experimental breeding lines are evaluated by being grown at a number of trial locations. These trials are known as multi-environment trials (MET). To identify elite experimental lines it is common to fit factor analytic mixed models to the data generated by these trials. The fitting of these models is computationally intensive. Currently a fast and efficient algorithm known as the Average Information (AI) algorithm is used to fit these models. However, the AI algorithm can be unstable when fitting factor analytic mixed models. A slower but more stable algorithm is the EM algorithm. The EM (Expectation-Maximisation) algorithm has monotonic convergence properties but can be very slow to converge for some problems. An aim of this project is the implementation of a hybrid algorithm that combines the stability of the EM algorithm with the speed and efficiency of the AI algorithm.

## Research Project Description

### Project Title

Enhancing NIRS calibrations of grains for livestock – Statistical analysis and advice  
(Activity #769)

### Principal Investigator

Sharon Nielsen

### Funding Sources

Industry & Investment NSW, the Grains Research & Development Corporation and the Pork Cooperative Research Centre (Project No. 1B-105 0607).

### Objectives

1. Analyse data from the PGLP and/or Pork CRC on request from grain and animal industry representatives (data mining) as directed by Pork CRC sub-program manager.
2. Analyse data returned to GRDC from research licences granted to organisations evaluating the PGLP NIR calibrations under commercial conditions.
3. Design all experiments undertaken in the Pork CRC sub-program 1B, which includes editing the design protocols, developing a design and upload experiment information to the Pork CRC web-site.
4. Statistically analyse data to produce predicted treatment means, determine significance of treatment differences and determine statistically significant relationships. Write a statistical report for every experiment analysed.
5. Undertake general statistical consulting and education as required to improve the efficiency of experiments conducted in sub-program 1B of the Pork CRC or other areas of the Pork CRC as requested by the sub-program manager.

### Summary

Experience during the Premium Grains for Livestock Project showed that it is essential for every experiment to be properly designed before it is undertaken so that the effects of all controllable or known variables can be removed during the statistical analysis. Hence, for every experiment conducted in the Pork CRC sub-program 1B, a full experimental protocol is developed that includes a description of all grains, methods used, samples collected and result calculations. Each final protocol and design is then 'signed off' by the sub-program manager and the protocol, including animal ethics number and design are posted on the CRC website in a secure location for access by all 1B personnel. The results from each experiment are then placed into the design spreadsheet by the experimenters and input into the database ready for analysis.

The statistical analysis involves deriving predicted treatment means where the influence of all known factors likely to contribute to variation in data, such as time, period, pen, animal, sex etc. are accounted for to produce the most accurate value for the effect of 'grain' on each measurement. The predicted treatment means are calculated and the significance of differences between grains and treatments are determined as well as determining the grain chemical and/or physical characteristics that may have contributed to these differences. A statistical report is written for each experiment analysed. Additional reports are written for comparisons across experiments where similar grains have been used.

Ongoing statistical consultation is provided to researchers in the Pork CRC sub-program 1B, mainly through phone conversations and email correspondence.

## Research Project Description

### Project title

Development and implementation of *Katmandoo*, bioscience data management system  
(Activity #777)

### Principal investigators

Mr Avishesh Shrestha

[Note: This is a collaborative project with the Queensland Department of Employment, Economic Development & Innovation (Mr David Butler and Mr David Rodgers) and Diversity Arrays Technology Pty Ltd (Mr Grzegorz Uszynski)]

### Funding sources

Industry & Investment NSW, Queensland Department of Primary Industries & Fisheries and Diversity Arrays Technology Pty Ltd.

### Objective

The primary objective is to develop *Katmandoo* and help clients to implement the database system.

### Summary

*Katmandoo* is a joint project between Industry & Investment NSW, Queensland Department of Primary Industries & Fisheries and Diversity Arrays Technology Pty Ltd to develop a bioscience data management system which is essential to the operational requirements of any collaborative breeding effort and will provide a robust platform for the implementation of advanced statistical, breeding and visualisation methods that integrate phenotypic data with known genetic information such as pedigree and molecular data. High throughput genotyping is likely to become commonplace; *Katmandoo* applications will be essential tools in storing, analysing, displaying and interpreting this information.

#### *System architecture*

*Katmandoo* has consensus generic data schema for managing phenotypic and genomic data, where key entities for phenotypic observations respect the statistical notions of experimental units and sampling units. Primary keys identifying experimental units are not predefined so observations from differing sources, such as field trials or laboratory experiments, are stored in a single consistent manner. Treatment structures (such as *Genotypes*) are considered a property of the experimental unit and factorial regimes are admitted. The actual data is stored at a *sampling unit* level and the *trait* entity is extended to the notion of a *variate* so repeated measures, multivariate and infinite sub-sampling cases are allowed.

*Katmandoo* is being developed as a suite of modules:

- phenotypic data management (completed)
- genealogy management (in alpha testing phase)
- tools to capture trait score using Windows-Mobile OS (in alpha testing phase)
- crossing tool (v2.5)
- molecular marker subsystem (v2.5)
- tools to integrate phenotypic and genomic data (v3.0)
- seed inventory subsystem (v3.5)

where not all need to be present depending on the level of functionality required.

#### *Implementation*

The data model is implemented in SQLServer 2005 and MySQL 5.1. The application software is being developed in the Microsoft .NET Framework 2.0. A middleware layer manages data access requests and acts as an applications programming interface for higher-level applications. The .NET environment allows seamless mixed-language development, so that scientific applications may be retained in C++ while Windows or Web applications may be developed in alternative languages.

#### *Development*

The phenotypic data management component was released in November 2008. Genealogy management component is under alpha testing in QDPI&F and I&I NSW. Its beta version will be released August 2009.

Kollect-Trait has been developed to allow the user to enter trait score using PDA operating Windows-Mobile. *Katmandoo* generates a data-file and imports data captured by the application. It is being tested at CSIRO PI, QDPI&F, I&I NSW and its beta version will be released in August 2009.

More information is available from <http://www.katmandoo.org/Help/Index.htm>

Users can download Katmandoo and its update from <http://www.katmandoo.org/forum/>