Higher Education
Australian Mathematical Sciences Institute

On the design and analysis of early stage plant breeding selection experiments

Lauren Borg
Supervised by Brian Cullis and Emi Tanaka
University of Wollongong

Australian Government
Department of Education

Contents

1 Introduction

The linear mixed model is regarded as an expansion on the classical linear model and so its fundamental properties may be applied to the analysis of mixed models. This chapter initially presents details on the analysis of linear models and then extends these ideas to the mixed model approach.

1.1 Introduction to Mixed Models: One way classification

1.1.1 Defining the Linear Model

The classical linear model is given by:

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{1.1}$$

where y_{ij} is the observed outcome of the ith treatment ($i = 1, \ldots, t$) on the jth replicate ($j = 1, \ldots, b$), μ is the overall mean for all treatments and α_i is the fixed effect of the ith treatment. The residual term is represented by e_{ij}, which describes the variability within treatments. The model assumes that the residuals are independent and identically distributed such that $e_{ij} \sim N(0, \sigma^2)$.

For the set of all observations, the model in (1.1) may be represented in matrix form as follows,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e} \tag{1.2}$$

Here, $\boldsymbol{y}$ is a vector of length n, where n is the sample size. The matrix, $\boldsymbol{X}$, is the design matrix with dimensions $n \times (t+1)$, where t is the number of treatments. The vector, $\boldsymbol{\tau}$, is the vector containing the fixed effects for each treatment as well as the overall mean and so is of size $(t+1)$ given as,

$$\boldsymbol{\tau} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_t \end{bmatrix}$$

Finally, $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. The assumption on the residual terms is represented as $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I_n})$ and so it follows that $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\tau}, \sigma^2 \boldsymbol{I_n})$.

The aim of the analysis of a linear model is to estimate the unknown parameters $\boldsymbol{\tau}$ and σ^2 using the observed data in order to fit the model (1.2).

1.1.2 Maximum Likelihood Estimation

Typically estimation is achieved using the maximum likelihood approach. The individual observations from the model in (1.1) are distributed such that,

$$y_i \sim N(\boldsymbol{x_i^T}\boldsymbol{\tau}, \sigma^2) \tag{1.3}$$

where $\boldsymbol{x_i}^T$ is the ith row of $\boldsymbol{X}$. Therefore the likelihood is given by,

$$L(\boldsymbol{\tau}, \sigma^2; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(\frac{-1}{2\sigma^2}(y_i - \boldsymbol{x_i^T}\boldsymbol{\tau})^2\Big)$$

The log-likelihood is given by,

$$l(\boldsymbol{\tau}, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\boldsymbol{\tau})^2$$

Now, finding the score vector by differentiating the log-likelihood equation with respect to $\boldsymbol{\tau}$ and σ^2,

$$\frac{\partial l}{\partial \boldsymbol{\tau}} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \boldsymbol{x_i}(y_i - \boldsymbol{x_i^T}\boldsymbol{\tau}) \tag{1.4}$$

and

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\boldsymbol{\tau})^2 \tag{1.5}$$

Equating (1.4) to zero brings about the normal equation for estimating $\boldsymbol{\tau}$ by first noting,

$$\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i^T} = \boldsymbol{X^T X}, \quad \sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{y_i} = \boldsymbol{X^T y}$$

Thus,

$$\boldsymbol{X^T X}\hat{\boldsymbol{\tau}} = \boldsymbol{X^T y} \tag{1.6}$$

Equating (1.5) to zero allows for estimation of σ^2 where,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\hat{\boldsymbol{\tau}})^2 \\
&= \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}) \\
&= \frac{1}{n}\tilde{\boldsymbol{e}}^{\boldsymbol{T}}\tilde{\boldsymbol{e}} \\
&= \frac{1}{n}RSS
\end{aligned} \tag{1.7}$$

The maximum likelihood estimator of variance is biased as it doesn't take into account the estimation of $\boldsymbol{\tau}$. The unbiased estimator of σ^2 is given by

$$s^2 = \frac{1}{n-r} RSS$$

where r is the rank of $\boldsymbol{X}$ and RSS is the residual sum of squares.

The model (1.2), in it's current form, results in a design matrix that is not of full rank; that is, the rank, r, of $\boldsymbol{X}$ is less than the number of columns in $\boldsymbol{X}$. Aliasing occurs in this situation, where there is no information on which to base an estimate for a parameter value. This occurs is linear models that involve an intercept and main effects, where it is said the intercept is intrinsically aliased with the interaction terms.

A standard approach for dealing with intrinsic aliasing is to place constraints on $\boldsymbol{X}$ to obtain a full rank design matrix. This can be achieved by omitting the μ parameter from $\boldsymbol{\tau}$ and the column of ones in $\boldsymbol{X}$ to produce a design matrix that is of full rank. Therefore the matrix $\boldsymbol{X^T X}$ is invertable and the unique maximum likelihood estimate is derived from (1.6) to be

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \tag{1.8}$$

This simple transformation to the model does not always result in a design matrix that is of full rank. There are further manipulations involving generalised inverse matrices, but this concept is not utilised in this report.

1.1.3 REML Estimation

Residual maximum likelihood is a form of maximum likelihood estimation which uses the likelihood function calculated from a transformed data set, removing the effect of nuisance parameters.

The maximum likelihood estimate, $\hat{\sigma}^2$, given by (1.7) doesn't allow for estimation of $\boldsymbol{\tau}$ and so produces a biased estimate. A transformation on $\boldsymbol{y}$ can correct this by first defining $\boldsymbol{L} = \begin{bmatrix} \boldsymbol{L_1} & \boldsymbol{L_2} \end{bmatrix}$ as an $n \times n$ projection matrix where $\boldsymbol{L_1}$ is $n \times r$, $\boldsymbol{L_2}$ is $n \times (n-r)$ and r is the rank of $\boldsymbol{X}$. The matrix $\boldsymbol{L}$ is chosen such that:

$$\boldsymbol{L_2^T X} = \boldsymbol{0}, \quad \boldsymbol{L_2^T L_2} = \boldsymbol{I_{n-r}}, \quad \boldsymbol{L_1^T L_1} = \boldsymbol{I_r}, \quad \boldsymbol{L_1 L_1^T} = \boldsymbol{I_n} \quad \boldsymbol{L_1^T L_2} = \boldsymbol{0}$$

For the transformation, define a new space

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \boldsymbol{L^T y} = \begin{bmatrix} \boldsymbol{L_1^T y} \\ \boldsymbol{L_2^T y} \end{bmatrix} \tag{1.9}$$

Finding the expectation of the new variables $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$,

$$E(\boldsymbol{y_1}) = E(\boldsymbol{L_1^T y}) = E(\boldsymbol{L_1^T X \tau} + \boldsymbol{L_1^T e}) = \boldsymbol{L_1^T X \tau}$$
$$E(\boldsymbol{y_2}) = E(\boldsymbol{L_2^T y}) = E(\boldsymbol{L_2^T X \tau} + \boldsymbol{L_2^T e}) = \boldsymbol{0}$$

Similarly for the variance,

$$\begin{aligned}\text{Var}\left(\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}\right) &= \text{Var}(\boldsymbol{L^T y})\\ &= \boldsymbol{L^T}\text{Var}(\boldsymbol{y})\boldsymbol{L}\\ &= \sigma^2 \boldsymbol{L^T L}\\ &= \sigma^2 \begin{bmatrix}\boldsymbol{L_1^T L_1} & \boldsymbol{L_1^T L_2}\\ \boldsymbol{L_2^T L_1} & \boldsymbol{L_2^T L_2}\end{bmatrix}\\ &= \sigma^2 \begin{bmatrix}\boldsymbol{I_r} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I_{n-r}}\end{bmatrix}\\ &= \sigma^2 \boldsymbol{I_n}\end{aligned}$$

Therefore the distribution of the transformed data set is given by

$$\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{L_1 X^T \tau}\\ \boldsymbol{0}\end{bmatrix}, \sigma^2 \boldsymbol{I_n}\right) \tag{1.10}$$

This provides two independent models for estimation where the marginal distribution for $\boldsymbol{y_1}$ is used to estimate $\boldsymbol{\tau}$ and the marginal distribution of $\boldsymbol{y_2}$ is used to estimate σ^2. The marginal distribution for $\boldsymbol{y_1}$ is

$$\boldsymbol{y_1} \sim N(\boldsymbol{A\tau}, \sigma^2 \boldsymbol{I_r}) \tag{1.11}$$

where $\boldsymbol{A} = \boldsymbol{L_1^T X}$. The normal equations for estimating $\boldsymbol{\tau}$ are then given by

$$\boldsymbol{A^T A \hat{\tau}} = \boldsymbol{A^T y_1}$$

Therefore substituting for $\boldsymbol{A}$ and $\boldsymbol{y_1}$,

$$\boldsymbol{X^T L_1 L_1^T X \hat{\tau}} = \boldsymbol{X^T L_1 L_1^T y}$$

$$\boldsymbol{X^T X \hat{\tau}} = \boldsymbol{X^T y}$$

This estimate for $\boldsymbol{\hat{\tau}}$ using the REML approach is the same as the one produced using the maximum likelihood approach in (1.8).

The REML estimate for σ^2 can be obtained by applying the maximum likelihood method to the marginal distribution for $\boldsymbol{y_2}$ where

$$\boldsymbol{y_2} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_{n-r}}) \tag{1.12}$$

Therefore the log likelihood function for $\boldsymbol{y_2}$ is given by

$$l(\sigma^2; \boldsymbol{y_2}) = -\frac{n-r}{2}\log(2\pi) - \frac{n-r}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{y_2^T y_2}$$

Differentiating by σ^2 and equating to zero produces the estimate

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-r}\boldsymbol{y_2^T y_2}\\ &= \frac{1}{n-r}\boldsymbol{y^T L_2 L_2^T y}\end{aligned} \tag{1.13}$$

As the matrices $\boldsymbol{L_1^T}$ and $\boldsymbol{X}$ are orthogonal (ie.$\boldsymbol{L_2^T X = 0}$) and $\boldsymbol{L_2^T L_2 = I}$ as defined in the original conditions when choosing $\boldsymbol{L}$, then

$$\begin{aligned}\boldsymbol{L_2 L_2^T} &= \boldsymbol{I_n - X(X^T X)^{-1} X^T} \\ &= \boldsymbol{I_n - P_X}\end{aligned}$$

This can be verified in the following proof:

$$\begin{aligned}\boldsymbol{I_n} &= \boldsymbol{[L_2 \quad X][L_2 \quad X]^{-1}} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \\ &= \boldsymbol{[L_2 \quad X]}\Big(\begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \boldsymbol{[L_2 \quad X]}\Big)^{-1} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \\ &= \boldsymbol{[L_2 \quad X]} \begin{bmatrix}\boldsymbol{L_2^T L_2} & \boldsymbol{L_2^T X} \\ \boldsymbol{X^T L_2} & \boldsymbol{X^T X}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \\ &= \boldsymbol{[L_2 \quad X]} \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X^T X}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \\ &= \boldsymbol{[L_2 \quad X]} \begin{bmatrix}\boldsymbol{I^{-1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{(X^T X)^{-1}}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{L_2^T} \\ \boldsymbol{X^T}\end{bmatrix} \\ &= \boldsymbol{L_2^T L_2 + X(X^T X)^{-1} X^T}\end{aligned}$$

Rearranging to find,

$$\boldsymbol{L_2^T L_2 = I_n - X(X^T X)^{-1} X^T = I_n - P_X}$$

Using this information and the fact the matrix $(\boldsymbol{I_n - P_X})$ is idempotent (ie. $\boldsymbol{P^2 = P}$) and symmetric ($\boldsymbol{P^T = P}$), both of which are easily verified, the estimate for σ^2 in (1.13) may be rewritten as

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-r}\boldsymbol{y^T(I_n - P_X)y} \\ &= \frac{1}{n-r}\boldsymbol{y^T(I_n - P_X)^2 y} \\ &= \frac{1}{n-r}\boldsymbol{y^T(I_n - P_X)^T(I_n - P_X)y}\end{aligned} \tag{1.14}$$

Also noting that

$$\begin{aligned}\tilde{e} &= \boldsymbol{y - X\hat{\tau}} \\ &= \boldsymbol{y - X(X^T X)^{-1} X^T y} \\ &= \boldsymbol{y - P_X y} \\ &= \boldsymbol{(I_n - P_X)y}\end{aligned}$$

Therefore

$$\hat{\sigma}^2 = \frac{1}{n-r}\boldsymbol{\tilde{e}^T \tilde{e}} = \frac{RSS}{n-r}$$

This is unbiased and is the standard REML estimate corrected for estimation of $\boldsymbol{\tau}$. This decomposition of the data vector $\boldsymbol{y}$ into components for estimation is residual maximum likelihood (REML) estimation.

1.2 Linear Mixed Models

The methods of estimation discussed in the previous section for linear models can then extended for the analysis of the linear mixed model. The distinction of the mixed model, compared to the linear model presented in (1.2), is that the mixed model assumes that some effects are actually realisations of a random variable, termed random effects. In the linear model only the residual terms are regarded as random.

The mixed model is essential in the analysis of plant variety selection data, the context of this report. Growing regions for commercial crops vary greatly in location and season, therefore it is important to consider that not all varieties prosper to the same level in different environments, known as variety by environment (VxE) interaction. Variety effects made be regarded as fixed or random but VxE interactions are always regarded as random. Regarding variety effects as random allows for more reliable predictions of the varieties performance and addresses issues of selection bias.

1.2.1 Defining the Linear Mixed Model

The linear mixed model is defined as,

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e} \tag{1.15}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observed outcomes, $\boldsymbol{\tau}$ is the $t \times 1$ vector of fixed effects and $\boldsymbol{u}$ is the $b \times 1$ vector of random effects. Here, n is the number of data values and t is the number of varieties, if variety effects are regarded as fixed. For mixed models, b is the number of levels of the random factor, typically the number of blocks. $\boldsymbol{X}$ is the $n \times t$ design matrix associated with the fixed effects, assumed to be of full rank. $\boldsymbol{Z}$ is $n \times b$ design matrix associated with the random effects.The joint distribution of the random effects and residuals, $(\boldsymbol{u}, \boldsymbol{e})$, is Gaussian with zero mean and variance

$$\begin{bmatrix} \boldsymbol{G}(\boldsymbol{\gamma}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}(\boldsymbol{\phi}) \end{bmatrix}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are vectors of variance parameters associated with $\boldsymbol{u}$ and $\boldsymbol{e}$ respectively.

Therefore

$$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X\tau}, \boldsymbol{H})$$

where $\boldsymbol{H} = \boldsymbol{ZGZ^T} + \boldsymbol{R}$.

Typically, for data composed of q components, $\boldsymbol{u} = \{\boldsymbol{u_i}\}$ where $\boldsymbol{u_i}$ is the $b_i \times 1$ vector of effects for the ith term with $i = 1 \dots q$. The associated design matrices are $\boldsymbol{Z_i}$ of size $n \times b_i$, so that $\boldsymbol{Z}$ is of the form $[\boldsymbol{Z_1} \dots \boldsymbol{Z_q}]$. The q components are considered independent

so that if $\text{Var}[\boldsymbol{u_i}] = \boldsymbol{G_i}$, then $\boldsymbol{G} = \text{diag}(\boldsymbol{G_i})$. In many situations $\boldsymbol{G_i} = \gamma_i \boldsymbol{I_{b_i}}$ where γ_i is the variance component for the ith random term. This type is used in the standard mixed model for multi-environment trial (MET) data which includes a variance component for the variety effects and the VxE interactions. More complicated structures for $\boldsymbol{G_i}$ can also be adopted depending on the nature of the data to be fitted.

There are also multiple types for the variance matrix, $\boldsymbol{R}$, for the error terms. In IB (incomplete-block) analysis $\boldsymbol{R} = \sigma^2 \boldsymbol{I_n}$ where σ^2 is the within block variance. For spatial analysis of MET data sets, $\boldsymbol{R} = \text{diag}(\boldsymbol{R_j})$ where $\boldsymbol{R_j}$ is the error variance matrix for the jth trial or location.

In mixed model analysis, estimation procedures are required for the fixed effects, variance parameters and random effects.

1.2.2 Estimation of Fixed and Random Effects

There are several derivations for the estimates of the fixed and random effects in the model outlined in (1.15). Here we assume that $\boldsymbol{X}$ is a full rank matrix and so we can use the mixed model equations which are derived from maximising the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$ given by,

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{X\tau} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{H} & \boldsymbol{ZG} \\ \boldsymbol{GZ^T} & \boldsymbol{G} \end{bmatrix} \right)$$

The log-density function for $(\boldsymbol{y}, \boldsymbol{u})$ is written as

$$L(\boldsymbol{y}, \boldsymbol{u}) = L^*(\boldsymbol{y}|\boldsymbol{u}) + L(\boldsymbol{u}) \tag{1.16}$$

where $L(\boldsymbol{u})$ is the marginal log-density function for $\boldsymbol{u}$. The function $L^*(\boldsymbol{y}|\boldsymbol{u})$ has the form of a conditional log-likelihood but is not actually a likelihood as $\boldsymbol{u}$ cannot be observed. The marginal distribution of $\boldsymbol{u}$ is

$$\boldsymbol{u} \sim \boldsymbol{N(0, G)}$$

and the conditional distribution of y is

$$\boldsymbol{y|u} \sim \boldsymbol{N(X\tau + Zu, R)}$$

The function in (1.16) is then given by

$$\begin{aligned} L(\boldsymbol{y}, \boldsymbol{u}) &= -\frac{1}{2}\{\log|\boldsymbol{R}| + (\boldsymbol{y - X\tau - Zu})^{\boldsymbol{T}} \boldsymbol{R^{-1}} (\boldsymbol{y - X\tau - Zu})\} \\ &\quad - \frac{1}{2}\{\log|\boldsymbol{G}| + \boldsymbol{u^T G^{-1} u}\} \\ &= -\frac{1}{2}\{\log|\boldsymbol{R}| + \log|\boldsymbol{G}| + (\boldsymbol{y - X\tau})^{\boldsymbol{T}} \boldsymbol{R^{-1}} (\boldsymbol{y - X\tau})\} \\ &\quad - \frac{1}{2}\{\boldsymbol{u^T}(\boldsymbol{ZR^{-1}Z^T + G^{-1}})\boldsymbol{u - 2}(\boldsymbol{y - X\tau})^{\boldsymbol{T}} \boldsymbol{R^{-1}Zu}\} \end{aligned}$$

The parameters $\boldsymbol{\tau}$ and $\boldsymbol{u}$ are estimated by maximising this function (taking the derivatives with respect to each parameter and equating to zero) to produce the system called the mixed model equations (MME) given as

$$\begin{bmatrix} \boldsymbol{X^T R^{-1} X} & \boldsymbol{X^T R^{-1} Z} \\ \boldsymbol{Z^T R^{-1} X} & \boldsymbol{(Z^T R^{-1} Z + G^{-1})} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T R^{-1} y} \\ \boldsymbol{Z^T R^{-1} y} \end{bmatrix} \tag{1.17}$$

Breaking this system down into two separate equations we have,

$$\boldsymbol{(X^T R^{-1} Z)\tilde{u} + (X^T R^{-1} X)\hat{\tau} = X^T R^{-1} y} \tag{1.18}$$

$$\boldsymbol{(Z^T R^{-1} Z + G^{-1})\tilde{u} + (Z^T R^{-1} X)\hat{\tau} = Z^T R^{-1} y} \tag{1.19}$$

By rearranging (1.18) we obtain,

$$\hat{\boldsymbol{\tau}} = \boldsymbol{(X^T R^{-1} X)^{-1}(X^T R^{-1} y - X^T R^{-1} Z\tilde{u})} \tag{1.20}$$

And then substituting this expression for $\hat{\tau}$ into (1.19) and gathering the terms associated with $\tilde{u}$,

$$\begin{aligned} \boldsymbol{Z^T R^{-1} y} = &\boldsymbol{\big(Z^T[R^{-1} - R^{-1} X(X^T R^{-1} X)^{-1} X^T R^{-1}]Z + G^{-1}\big)\tilde{u}} \\ &\boldsymbol{+ Z^T R^{-1} X(X^T R^{-1} X)^{-1} X^T R^{-1} y} \end{aligned}$$

This is simplified by defining

$$\boldsymbol{S = R^{-1} - R^{-1} X(X^T R^{-1} X)^{-1} X^T R^{-1}} \tag{1.21}$$

and rearranging,

$$\begin{aligned} \boldsymbol{(Z^T S Z + G^{-1})\tilde{u}} &= \boldsymbol{-Z^T R^{-1} X(X^T R^{-1} X)^{-1} X^T R^{-1} y + Z^T R^{-1} y} \\ \boldsymbol{(Z^T S Z + G^{-1})\tilde{u}} &= \boldsymbol{Z^T(R^{-1} - R^{-1} X(X^T R^{-1} X)^{-1} X^T R^{-1})y} \\ \boldsymbol{(Z^T S Z + G^{-1})\tilde{u}} &= \boldsymbol{Z^T S y} \\ \tilde{\boldsymbol{u}} &= \boldsymbol{(Z^T S Z + G^{-1})^{-1} Z^T S y} \end{aligned} \tag{1.22}$$

These solutions to the MME given in (1.20) and (1.22) serve as estimates for the fixed and random effects respectively. Again this assumes that the design matrix, $\boldsymbol{X}$, is of full rank for the solutions to be unique. The representation of the MME solutions above is a convenient form for the data set examined in the upcoming chapter of this report, where the $\boldsymbol{S}$ matrix allow for a more computationally simple solutions.

The solution, $\hat{\boldsymbol{\tau}}$, derived from the MME, is referred to as the best linear unbiased estimate (BLUE). BLUE's are linear functions of the data $\boldsymbol{y}$ and the expectation of the estimates is equal to the expectation of the value estimated, hence they are unbiased. They are termed "best" as they have the minimum mean square error (MSE) among the class of linear unbiased estimators. It should be noted the term, estimator, is used to distinguish that BLUE's are estimates of fixed, not random, effects. The solution, $\tilde{\boldsymbol{u}}$, to the MME is considered the best linear unbiased predictor (BLUP), where here they estimate random effects and so are referred to as predictors.

1.2.3 Estimation of Variance Parameters

When data is balanced and errors follow the assumption of independence and identical distribution, the variance parameters of a mixed model can be estimated using an ANOVA table. In plant breeding, however, it is unlikely data will be balanced and so parameter estimation is more complex. In this case, the residual maximum likelihood (REML) approach is used to estimate the variance parameters. REML is used in preference over standard maximum likelihood as REML accounts for loss of degrees of freedom from fixed effects estimation, producing less bias estimates of the variance parameters.

The residual log-likelihood function is given by,

$$l_R = -\frac{1}{2}\{\log|\boldsymbol{H}| + \log|\boldsymbol{X^T H^{-1} X}| + \boldsymbol{y^T P y}\} \tag{1.23}$$

where $\boldsymbol{P = H^{-1} - H^{-1} X (X^T H^{-1} X)^{-1} X^T H^{-1}}$.

This likelihood is obtained by transforming the mixed model outcome $\boldsymbol{y}$ into two parts; $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$, similarly to the REML approach utilised for the general linear model. The conditional distribution of $\boldsymbol{y_1}$ on $\boldsymbol{y_2}$ can be maximised in order to obtain estimates of the fixed effects and the distribution of $\boldsymbol{y_2}$ is maximised to obtain the variance parameter estimates.

The REML estimate of $\boldsymbol{\kappa}$, the vector of variance components, is obtained by solving the score equations

$$U_R(\kappa_i) = \frac{\partial l_R}{\partial \kappa_i} = 0$$

for $i = 1 \dots n_k$, where n_k is the number of variance parameters in $\boldsymbol{\kappa}$.

The score for κ_i is given by

$$U_R(\kappa_i) = -\frac{1}{2}\Big(tr(\boldsymbol{P\dot{H}_i}) - \boldsymbol{y^T P \dot{H}_i P y}\Big) \tag{1.24}$$

where $\boldsymbol{\dot{H}_i} = \partial \boldsymbol{H} / \partial \kappa_i$.

2 Multi-Environment Trial Data

This chapter serves as a reference for the analysis carried out in the remainder of the report. Here, we present the data from a multi-environment trial (MET); that is data gathered from plant variety trials carried out at a series of different locations, perhaps over several years. Also, this chapter provides a brief insight into the context of the data analysed; describing how MET data is utilised in selection experiments that aim to breed new varieties of plants with optimal characteristics.

2.1 Selection Experiments

Multi-environment trials are the typical experimental design for selection trials. Selection trials involve comparing multiple test varieties of a crop against existing commercial varieties for beneficial properties, for example, increased yield and disease resistance. This is done in order to make recommendations on which varieties are to be selected for commercial growing as part of the agricultural industry. Ideally, for a crop to be considered appropriate for selection, it must achieve a consistent performance of optimum yield in a diverse range of seasonal and geographical conditions. Thus variety by environment (V×E) interaction is a crucial factor in selection programs and hence why MET's are the common design adopted when making variety recommendations.

There are several stages of testing and selection in the breeding of a new variety for improved yield. Each successive stage involves more trial locations and replicates but fewer varieties, where the highest yielding varieties progress onto the next stage. Selections are based on analysis of yield data, with minimal possible error to ensure accuracy in future grain yield predictions for new varieties.

Traditional MET analysis is comprised of two stages; initially the mean yield for varieties are estimated for the individual locations, followed by a combined overall analysis of varietal performance across all trials. Newer approaches involve spatial modelling of the MET data, discussed fully in Chapter 4. These models include variety (V) and environment (E) main effects aa well as V×E interaction. The effects V and V×E are regarded as random and assumed independent with constant variance. This constant variance assumption implies that all environments have the same genetic variance which is unlikely to be true. Cullis. et al (2006) have recognised the improbability of this

assumption by fitting a separate V×E interaction for each trial but still assume a common genetic covariance for all environment pairs.

A partially replicated (p-rep) design, introduced in Cullis et al (2006), is used on varieties of interest to allow for efficient mixed model analyses in MET's. In the early stages of selection there are large number of varieties grown in a small number of trials making fully replicated designs inefficient and impractical. This design uses replicated plots for a percentage, p, of the test varieties. For this particular data set $p = 22\%$ and the remaining varieties are randomly assigned to single plots. P-rep designs serve as an alternative to grid-plot designs, resulting in higher genetic gains.

2.2 Explanation of the Data

As a motivating example, Chapter 3 presents an analysis of a wheat variety selection experiment carried out by AGT (Australian Grain Technologies) in 2013. The following data was the result of experimentation as part of a plant breeding program aimed at developing varieties of wheat which maximise the yield harvested. This multi-environment experiment took place over five different locations in Australia, namely; Angas Valley, Booleroo, Gnowangerup, Roseworthy and Winulta. A total of 237 different varieties of wheat were used in each of the five locations or trials. Of these varieties, 225 were tested for selection onto the next stage and the remaining 12 were existing commercial varieties to be tested against.

The experiment was a partially replicated (p-rep) design, where 51 of the varieties had two replications at each location, whereas the remaining 186 had no replications. This means that at each location there was a total of 288 plots ($2 \times 51 + 186$) and each location was subdivided into two blocks with 144 plots in each block. For the 51 varieties that were replicated, a replicate was randomly assigned to a plot in each of the two blocks for every location and the other 186 varieties were then randomly assigned to the remaining plots across the two blocks. Thus the data is classified as unbalanced but with replicated varieties balanced across blocks. The trial for each location was organised into 24 rows by 12 columns, the later referred to as ranges. The plots in ranges 1 to 6 belonged to one block and the plots in ranges 7 to 12 belonged to the second block.

Initially, for the simplicity of understanding the models applied to the data, non-genetic sources of variation will be ignored and the genetic effects will be assumed independent. Also, separate models will be fitted at each location and spatial analysis of the trials is to be investigated in Chapter 4.

The data, `newprep.df`, was read into **R** and a summary of the data produced, showing that there are 23 variables in the data frame. Some of these variables show `NA` output meaning no data was collected for these variables. The `location` variable is a factor of five levels corresponding to the five trials and `name` is a factor with 237 levels corresponding to the wheat varieties used including replicates. The `yieldkgha` variable take numerical values indicating the yield in kilograms per hectare for each plot.

```
R> str(newprep.df, strict.width='cut')

'data.frame':        1440 obs. of  26 variables:
 $ plot      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Column    : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1..
 $ Row       : Factor w/ 24 levels "1","2","3","4",..: 1 2 3 4 5 6 7..
 $ entry     : int  1 95 214 94 208 143 124 41 28 76 ...
 $ expt      : Factor w/ 5 levels "ANR2I131","BLR2I132",..: 1 1 1 1 ..
 $ site      : Factor w/ 5 levels "AN13","BL13",..: 1 1 1 1 1 1 1 1 ..
 $ loccode   : Factor w/ 5 levels "AN","BL","GW",..: 1 1 1 1 1 1 1 1..
 $ location  : Factor w/ 5 levels "Angas Valley",..: 1 1 1 1 1 1 1 1..
 $ year1     : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 201..
 $ name      : Factor w/ 237 levels "AGTKATANA","AXE",..: 3 83 228 8..
 $ plotwidth : int  NA NA NA NA NA NA NA NA NA NA ...
 $ yieldkgha : num  2231 2213 2909 2173 2409 ...
 $ hectlitwt : num  NA NA 85.9 NA NA NA 84.8 NA NA NA ...
 $ screens_2 : num  NA NA 2.81 NA NA NA 1.1 NA NA NA ...
 $ height    : logi  NA NA NA NA NA NA ...
 $ lodgescore: int  NA NA NA NA NA NA NA NA NA NA ...
 $ lodgerhizo: int  4 3 2 4 2 2 3 5 2 1 ...
 $ matjuldays: logi  NA NA NA NA NA NA ...
 $ matscore  : logi  NA NA NA NA NA NA ...
 $ matzadoks : int  NA NA NA NA NA NA NA NA NA NA ...
 $ rhizoctsc1: logi  NA NA NA NA NA NA ...
 $ shattering: int  NA NA NA NA NA NA NA NA NA NA ...
 $ vigour    : logi  NA NA NA NA NA NA ...
 $ yield     : num  2.23 2.21 2.91 2.17 2.41 ...
 $ Block     : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Geno      : Factor w/ 237 levels "AGTKATANA","AXE",..: 3 83 228 8..
```

The table, `trials.sum` shows how there 237 varieties grown in 288 plots organised into an array of 24 rows by 12 columns at each location. The column `na` shows the number of plots where no yield value was obtained and the `meanyld` column is the average overall crop yield for each location. The left column contains the codes for the five trial locations. Angas Valley is coded by ANR2I131, Booleroo by BLR2I132, Gnowangerup by GWR2I133, Roseworthy by RSR2I134 and Winulta by WTR2I135.

```
R> trials.sum <-
R> data.frame(nv=tapply(newprep.df$name,newprep.df$expt,function(x)
+    length(unique(x))),
+    ncol=tapply(newprep.df$Column,newprep.df$expt,function(x)
+    length(unique(x))),
+    nrow=tapply(newprep.df$Row,newprep.df$expt,function(x)
+    length(unique(x))),
+    nplot=tapply(newprep.df$plot,newprep.df$expt,function(x)
```

```
+    length(unique(x))),
+    na=tapply(newprep.df$yield,newprep.df$expt,function(x)
+    length((x[is.na(x)]))),
+    meannyld=round(tapply(newprep.df$yield,newprep.df$expt,function(x)
+    mean(x,na.rm=T)),3))
R> trials.sum

          nv ncol nrow nplot na meannyld
ANR2I131 237   12   24   288  0    2.244
BLR2I132 237   12   24   288  0    2.989
GWR2I133 237   12   24   288  0    2.885
RSR2I134 237   12   24   288 10    4.131
WTR2I135 237   12   24   288  0    4.970
```

Another table was produced confirming that there are 51 replicated varieties and 186 unreplicated varieties at each of the five locations.

```
R> trial <- with(newprep.df,table(name,location))
R> apply(trial,2,table)

  location
   Angas Valley Booleroo Gnowangerup Roseworthy Winulta
1          186      186         186        186     186
2           51       51          51         51      51
```

Additionally, a blocking structure was applied to the data that separates the plots in columns 1 to 6 in one block and the plots in columns 7 to 12 in another block for each location.

```
R> newprep.df$Block <- 1
R> newprep.df$Block[newprep.df$Column>6] <- 2
```

3 Simple Analysis of Data

This chapter presents a simple overview of the concepts described in Chapter 1 by application to the data set outlined in Chapter 2. This example involves fitting and comparing several simplified models to a data-set in order to further develop an understanding on the ideas previously presented. This analysis particularly focuses on obtaining BLUE's and BLUP's for each model as well as variance components, using both the **ASReml-R** package and derivation of estimates as solutions to the mixed model equations.

Initially, subsets of the data set `newprep.df` were created for each of the five trial locations in order to perform separate analyses for each trial. In this analysis, there will be a focus on the results obtained in the Angas Valley trial, so the data for this location was extracted and several arbitrary or irrelevant variables were omitted to produce the `angas` data table.

```
R> angas<- subset(newprep.df, location == "Angas Valley",
+    select = c(plot, prange, location, name, yieldkgha))
```

3.1 Simple Linear Model

Fitting a simple linear model is often an inadequate way of describing data but here it will serve as a useful comparison for understanding the impact of the introduction of random effects. This model regards the variety effects as fixed and is defined as,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e} \tag{3.1}$$

For this model, $\boldsymbol{y}$ is a 288×1 vector of yields for each plot, $\boldsymbol{X}$ is a 288×237 design matrix and $\boldsymbol{\tau}$ is a 237×1 vector representing the fixed effect of each variety. Note that $\boldsymbol{\tau}$ does not contain a parameter for the overall mean as this would result in a non-full rank design matrix. Initially, **ASReml** is used to estimate the fitted values of the variety yields,

```
R> angas.rem.fix<- asreml(yieldkgha~ -1+name, data= angas)
R> tau.rem.fix<- coef(angas.rem.fix)$fixed
```

Comparison of the fitted values `tau.rem.fix` against the raw mean, where the raw mean is simply the summation of yields for a variety divided by the number of replicates, revealed the BLUE and raw mean were equal for each variety.

The estimates for the fixed effects can be reproduced using algebraic calculations, where the maximum likelihood estimate was derived in Chapter 1 for the classical linear model to be,

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{3.2}$$

In order to evaluate the expression in (3.2), the necessary matrices are compiled in **R** and then function for $\hat{\boldsymbol{\tau}}$ is evaluated,

```
R> y<- angas$yieldkgha
R> X<- model.matrix(~-1+name,data=angas)
R> tau.fix<- solve(t(X)%*%X)%*%t(X)%*%y
```

The vector `tau.fix` is equivalent to the vector `tau.hat.fix` produced by **ASReml** as expected. This confirms that **ASReml** produces the same effect estimates as the solutions to the normal equations.

3.2 Linear Mixed Model

A simple linear mixed model was fitted to the data with the variety effects considered random and the overall mean yield considered fixed. The model is represented by

$$\boldsymbol{y} = \boldsymbol{X}\tau + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{3.3}$$

Here, $\boldsymbol{y}$ is a 288×1 vector of yields for each plot, $\boldsymbol{X}$ is a 288×1 vector of ones and τ is the overall mean yield. The random effects design matrix, $\boldsymbol{Z}$, is a 288×237 matrix and $\boldsymbol{u}$ is a 237×1 vector of the random variety effects.

3.2.1 Effects Estimation

This model is initially fitted using **ASReml-R** which produces solutions of the mixed model equations. As the design matrix is full rank these solutions are unique such that `tau.hat` is the BLUE of the mean and `u.hat` is the BLUP of $\boldsymbol{u}$. The variable, `name`, is fitted as a simple random term using a default variance model where a single variance component is associated with the random variable, $\boldsymbol{u}$, given by $\boldsymbol{G} = \sigma_u^2\boldsymbol{I}$. The residual variance is given as $\boldsymbol{R} = \sigma^2\boldsymbol{I}$.

```
R> angas.reml<- asreml(yieldkgha~1, random= ~ name, data= angas)

R> u.hat<- coef(angas.reml)$random
R> tau.hat<- coef(angas.reml)$fixed
R> tau.hat

              effect
(Intercept) 2246.788
```

The overall mean yield, given by `tau.hat`, produced by **ASReml** is not equivalent to the raw mean of the yields as shown below,

```
R> c(tau.hat, mean(angas$yieldkgha))

[1] 2246.788 2243.954
```

This is a result of the nature of the experiment. The BLUE for τ is only the raw mean when the design is orthogonal. The experiment used to obtain the data `newprep.df` had a p-rep design, where there is not an equal number of replicates for each variety and thus the data is non-orthogonal. The raw mean isn't weighted by the the sample size for each variety whereas the BLUE for τ takes into account the number of replicates for each variety.

The solutions to the mixed model equations produced by **ASReml-R** can be replicated algebraically. Recall the solutions to the MME's are given as,

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X^T R^{-1} X})^{-1}(\boldsymbol{X^T R^{-1} y} - \boldsymbol{X^T R^{-1} Z \tilde{u}}) \tag{3.4}$$

$$\tilde{\boldsymbol{u}} = (\boldsymbol{Z^T S Z} + \boldsymbol{G^{-1}})^{-1}\boldsymbol{Z^T S y} \tag{3.5}$$

where

$$\boldsymbol{S} = \boldsymbol{R^{-1}} - \boldsymbol{R^{-1} X}(\boldsymbol{X^T R^{-1} X})^{-1}\boldsymbol{X^T R^{-1}}$$

This form of the mixed model equations is convenient for this model as $\boldsymbol{R} = \sigma^2 \boldsymbol{I_n}$ and $\boldsymbol{X}$ is simply a vector of ones, $\mathbf{1_n}$, and so

$$\begin{aligned}\boldsymbol{S} &= \sigma^{-2}\boldsymbol{I_n} - \sigma^{-2}\mathbf{1_n}(\mathbf{1_n^T 1_n})^{-1}\mathbf{1_n^T} \\ &= \sigma^{-2}(\boldsymbol{I_n} - \frac{1}{n}\mathbf{1_n 1_n^T})\end{aligned}$$

Therefore, looking individually at the terms in the mixed model equation for the random terms in (3.5),

$$\begin{aligned}\boldsymbol{Z^T S Z} + \boldsymbol{G^{-1}} &= \sigma^{-2}(\boldsymbol{Z^T Z} - \frac{1}{n}\boldsymbol{Z^T}\mathbf{1_n 1_n^T}\boldsymbol{Z}) + \boldsymbol{G^{-1}} \\ &= \sigma^{-2}(\text{diag}(\boldsymbol{M}) - \frac{1}{n}\boldsymbol{M M^T}) + \sigma_u^{-1}\boldsymbol{I_n}\end{aligned} \tag{3.6}$$

Here $\boldsymbol{M}$ is an vector of length 237 whose elements correspond to the number of replicates for each variety and diag($\boldsymbol{M}$) is a 237×237 matrix with the vector, $\boldsymbol{M}$, as it's diagonal and the remaining elements are zero. Also,

$$\begin{aligned}\boldsymbol{Z^T S y} &= \sigma^{-2}(\boldsymbol{Z^T y} - \frac{1}{n}\boldsymbol{Z^T}\mathbf{1_n 1_n^T y}) \\ &= \sigma^{-2}(\boldsymbol{Z^T y} - \boldsymbol{M}\bar{y}_{..})\end{aligned}$$

where $\bar{y}_{..}$ is the grand mean and $\boldsymbol{Z^T y}$ is simply a vector of length 237, whose elements are the sum of yields for each variety.

Therefore the estimate for the random effect vector, $\tilde{\boldsymbol{u}}$, can then be calculated by

$$\tilde{\boldsymbol{u}} = \sigma^{-2}\big(\sigma^{-2}(\text{diag}(\boldsymbol{M}) - \frac{1}{n}\boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}}) + \sigma_u^{-1}\boldsymbol{I_n}\big)^{-1}\big(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{M}\bar{y}_{..}\big) \tag{3.7}$$

This is a less computationally expensive means of solving the MME for the random term, using convenient properties of the data set to avoid the inversion of the matrix $\boldsymbol{H}$.

Before solving for $\tilde{\boldsymbol{u}}$ the variance component, σ_u, must be extracted from the **ASReml** output and multiplied by the identity matrix to produce $\boldsymbol{G}$ and similarly for $\boldsymbol{R}$,

```
R> sig.u<- ang.var[1,2]
R> G<- sig.u*diag(1, nrow=237)
R> sig<- ang.var[2,2]
R> R<- sig*diag(1, nrow=288)
```

The remaining necessary matrices are then extracted from the data and the BLUP's for $\boldsymbol{u}$ are derived,

```
R> ones<- matrix(rep(1, 288), ncol=1)
R> S<- sig^{-1}*(R - (1/288)*ones%*%t(ones))
R> Z<- model.matrix(~-1+name,data=angas)
R> X<- matrix(rep(1, 288), ncol=1)
R> M<- t(Z)%*%ones
R> rep<- t(Z)%*%Z #matrix with no. of reps for each variety as diag
R> gy<-as.numeric((1/288)*t(ones)%*%y)
R> sum<-t(Z)%*%y
R> part1<- (1/sig)*(rep - (1/288)*M%*%t(M)) + solve(G)
R> part2<- (1/sig)*(sum- gy*M)
R> BLUP<- solve(part1)%*%part2
```

These `BLUP`'s for the random effects are identical to the estimates of the random variety effects, `u.hat`, produced by **ASReml**. This justifies that, for full rank design matrices, the estimate produced for the fixed and random effects by **ASReml** are in fact the solutions to the mixed model equations.

3.2.2 Variance Parameter Estimation

For unbalanced data such as this, the residual maximum likelihood (REML) approach is adopted for estimation of variance parameters as it produces less bias estimates than maximum likelihood by accommodating for loss of degrees of freedom from the estimation of fixed effects.

The residual log likelihood for the model in (1.15) is given in (1.23) but this formula requires inversion of the variance matrix, $\boldsymbol{H}$, which for typical MET data sets is compu-

tationally expensive. A strategy for overcoming this involves representing the matrix,

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{H}^{-1}$$

in the form,

$$\boldsymbol{P} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{R}^{-1} \tag{3.8}$$

where $\boldsymbol{W} = [\boldsymbol{X}\,\boldsymbol{Z}]$ and $\boldsymbol{C}$ is the coefficient matrix for the MME as depicted in (1.17).

Also residual log-likelihood can be written in the form,

$$l_R = -\frac{1}{2}\{\log|\boldsymbol{G}| + \log|\boldsymbol{R}| + \log|\boldsymbol{C}| + \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{y}\} \tag{3.9}$$

ASReml produces the same value for the REML log likelihood as the one obtained by algebraically solving the likelihood equation as demonstrated below.

```
R> #ASReml log-likelihood
R> angas.reml$loglik

[1] -1758.599

R> #algebraic derivation of log-likelihood
R> W<- cbind(X, Z)
R> #create coefficient matrix
R> C.left<-rbind((288/sig), (1/sig)*as.matrix(M))
R> C.right<-rbind((1/sig)*t(ones)%*%Z, ((1/sig)*rep+solve(G)))
R> C<- cbind(C.left, C.right)
R> P<- solve(R) - solve(R)%*%W%*%solve(C)%*%t(W)%*%solve(R)
R> loglik<- -(1/2)*(determinant(G, logarithm= T)$modulus +
+    determinant(R, logarithm= T)$modulus + determinant(C, logarithm=
+    T)$modulus + t(y)%*%P%*%y)
R> loglik

          [,1]
[1,] -1758.599
attr(,"logarithm")
[1] TRUE
```

ASReml chooses values for the variance components that maximise this residual log-likelihood function. The REML estimate of the variance parameters, $\boldsymbol{\kappa}$, is obtained by solving the system of score equations given by

$$U_R(\kappa_i) = \frac{\partial l_R}{\partial \kappa_i} = 0 \tag{3.10}$$

The score for κ_i is given by

$$U_R(\kappa_i) = -\frac{1}{2}\big(tr(\boldsymbol{P}\dot{\boldsymbol{H}}_{\boldsymbol{i}}) - \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{P}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\boldsymbol{P}\boldsymbol{y}\big) \tag{3.11}$$

where $\dot{\boldsymbol{H}}_{\boldsymbol{i}} = \partial \boldsymbol{H}/\partial \kappa_i$.

For this data set, there are two variance parameters to be estimated where $\boldsymbol{\kappa} = (\sigma_u^2, \sigma^2)$ and so

$$\frac{\partial \boldsymbol{H}}{\partial \sigma_u^2} = \boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{T}} \tag{3.12}$$

$$\frac{\partial \boldsymbol{H}}{\partial \sigma^2} = \boldsymbol{I} \tag{3.13}$$

Using R to evaluate the score function,

```
R> H1<- Z%*%t(Z)
R> H2<- diag(288)
R> -1/2*(sum(diag(P%*%H1)) - t(y)%*%P%*%H1%*%P%*%y)
R> -1/2*(sum(diag(P%*%H2)) - t(y)%*%P%*%H2%*%P%*%y)
```

Therefore the score functions for both variance parameters, produced by **ASReml**, are very small. This confirms that these estimates are both fairly accurate as they are both approximately solutions to the score equations.

3.3 Impact of Random Variety Effects

It can be shown that the best linear unbiased estimator (BLUE), when varieties are regarded as having fixed effects, is given by $(\bar{\boldsymbol{y}} - \mathbf{1}\bar{y}_{..})$, where $\bar{\boldsymbol{y}}$ is the vector of variety means and $\bar{y}_{..}$ is the grand mean. The best linear unbiased predictor (BLUP) of the random effects, $\boldsymbol{u}$, is given by the $E(\boldsymbol{u}|\boldsymbol{y_2})$ after applying the transformation on $\boldsymbol{y}$ discussed in Chapter 1, resulting in

$$\tilde{\boldsymbol{u}} = \frac{t\sigma_u^2}{t\sigma_u^2 + \sigma^2}(\bar{\boldsymbol{y}} - \mathbf{1}\bar{y}_{..}) \tag{3.14}$$

where t is the number of varieties. The coefficient $t\sigma_u^2/(t\sigma_u^2 + \sigma^2)$ is called the mean line heritability and is the genetic variance over the total variance. This form of BLUP in (3.14) demonstrates how BLUP is a shrinkage estimator as the mean variety heritability is a value between zero and one and the remainder of the function is simply the BLUE of the fixed effects.

Figure 3.1 shows the fitted values, $\hat{y}$, from the linear model in (3.1) where variety effects are considered fixed, plotted against the predicted values, $\tilde{y}$, from the linear mixed model in (1.15) with random variety effects. This figure graphically demonstrates the concept of shrinkage, where the BLUP estimates are shrunk towards the overall mean ($\mu = 2243.954$) compared to the BLUE's for the fixed effects. There are two lines represented, the top line is for the varieties that were replicated and the bottom line is for unreplicated varieties. The amount of shrinkage is dependent on the variance of the random effects. The extent to which the BLUP for a particular random effect is "shrunken" depends on the amount of information available for predicting that random effect where the bottom line, with

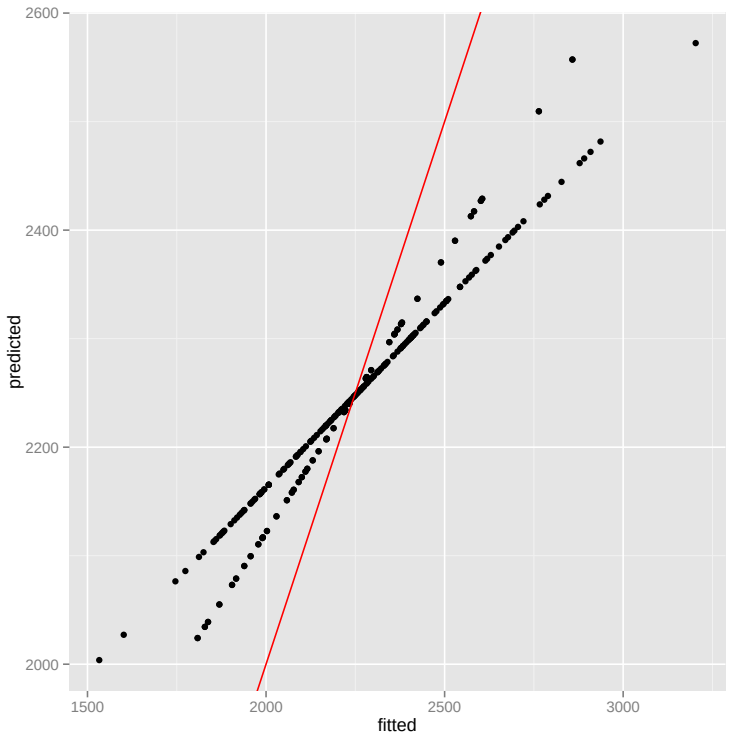

Figure 3.1: Plot of fitted values of the linear model against predicted values of the mixed model

unreplicated varieties, has more shrinkage compared to the BLUPs for the replicated varieties.

By regarding the variety effects as fixed, we obtain BLUE that is the best estimate of the performance of that variety in a particular trial. Regarding the variety effects as random, we obtain the BLUP that is the best estimate of the performance of that variety in future trials. Previously, selection has been based on BLUE's. It is then later observed that performance after release of that variety is not as good as that observed in trials. This is a result of the fact that future performance is predicted by BLUP's not BLUE's, hence why the variety effects are now more commonly treated as random.

Figure 3.1 shows the fitted values against the predicted values rather than a direct comparison of the BLUE's and BLUP's. This is due to the fact that, under the model in (3.1), the BLUE's are not estimatable. The vector $\boldsymbol{\tau}$ does not include a separate term for the intercept as this would require the design matrix, $\boldsymbol{X}$, to contain a column of ones that would result in the design matrix no longer being of full rank.

4 Spatial Modelling of Field Trials

There is a need to identify sources of variation in order to fit an appropriate model to a data set. Estimates of variety effects have improved accuracy using spatial analysis over, say incomplete block design. This is due to spatial analysis taking into account variation that results from plot location in field experiment data. There are three types of spatial variation:

1. smooth local spatial trend; which are the set of stationary trends such as fertility fluctuations.
2. smooth global spatial trend; which are the set of non-stationary trends across a field.
3. extraneous variation; which is variation due to how the field is managed. For example, serpentine harvesting is where crops are harvested systematically up and down rows of the field. This form of trial management leads to extraneous variation and thus reduced efficiency of selection.

Spatial variation modelling is particularly important at the early stages of selection when there are many varieties grown in few locations with few replicates. When modelling spatial variation, the error variation is broken up. Local trends are modeled using a covariance structure while global and extraneous variation are modeled using design factors and functions of the co-ordinates of the plots. It is assumed in field experiments that n plots are arranged in a rectangular array of r rows by c columns such that $n = r \times c$.

The model for spatial analysis is the same as the linear mixed model presented in (1.15),

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e} \tag{4.1}$$

Again, the joint distribution of $(\boldsymbol{u}, \boldsymbol{e})$ is Gaussian with zero mean and variance

$$\begin{bmatrix} \boldsymbol{G}(\gamma) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}(\phi) \end{bmatrix}$$

where γ and ϕ are vectors of variance parameters associated with $\boldsymbol{u}$ and $\boldsymbol{e}$ respectively and $\text{Var}(\boldsymbol{y}) = \boldsymbol{H} = \boldsymbol{ZGZ^T} + \boldsymbol{R}$. For the spatial model, the residual vector, $\boldsymbol{e}$, is broken down such that

$$\boldsymbol{e} = \boldsymbol{\zeta} + \boldsymbol{\eta} \tag{4.2}$$

The $\text{Var}|\boldsymbol{\zeta}| = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\alpha})$ where $\boldsymbol{\Sigma}$ is the spatial correlation matrix and $\boldsymbol{\alpha}$ is the vector of spatial correlation parameters. The vector, $\boldsymbol{\zeta}$, is associated with the smooth local trend. The measurement error, $\boldsymbol{\eta}$, has variance component σ^2_η and so

$$\text{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\alpha}) + \sigma^2_\eta \boldsymbol{I_n} \tag{4.3}$$

The marginal distribution of $\boldsymbol{y}$ is then,

$$\boldsymbol{y} \sim \text{N}(\boldsymbol{X\tau}, \boldsymbol{ZGZ^T} + \boldsymbol{R}) \tag{4.4}$$

The same concepts of estimation for mixed models apply but here the variance parameters to be estimated are $\boldsymbol{\kappa} = (\boldsymbol{\gamma}, \boldsymbol{\phi})$ where $\boldsymbol{\phi} = (\sigma^2, \boldsymbol{\alpha}, \sigma^2_\eta)$.

For accurate estimates of variety effects, an appropriate choice of plot error variance model must be selected.

4.1 Covariance Models for Local Spatial Trend

Anisotropic models are used for modelling variance structure in field trials, where anisotropic refers to plots that are directionally dependent. This means that the local trends show that data from plots closer together are more similar than those from plots further apart when disregarding variety effects. Therefore the elements of $\boldsymbol{\zeta}$ are correlated and the correlations are functions of the spatial distance between plots. Let $\boldsymbol{\Sigma} = \{\rho_{ij}\}$ where $\rho_{ij} = \text{Cor}(\zeta_i, \zeta_j)$ is the spatial correlation between plots i and j. Also, let $\boldsymbol{s_i} = (s_{ir}, s_{ic})$ denote the location of the ith plot in the field, giving the row and column co-ordinate respectively. Therefore, the spatial correlation may be rewritten as a function, V,

$$\rho_{ij} = V(\boldsymbol{s_i}, \boldsymbol{s_j}; \boldsymbol{\alpha}) = V(\boldsymbol{l_{ij}}; \boldsymbol{\alpha})$$

where $\boldsymbol{l_{ij}} = (l_{ijr}, l_{ijc}) = |\boldsymbol{s_i} - \boldsymbol{s_j}|$, as the correlation between two plots depends only by the distance between them. The model assumes the two dimensional process is separable so that V can be given as a product of the correlation for each dimension, that is rows and columns, such that

$$V(\boldsymbol{l_{ij}}; \boldsymbol{\alpha}) = V_r(l_{ijr}; \boldsymbol{\alpha_r}) V_c(l_{ijc}; \boldsymbol{\alpha_c})$$

where V_r and V_c are the row and column correlation functions respectively.

There are many possible forms for the spatial correlation function V. The directional exponential covariance (DEC) model has applications in field experiments and is given by,

$$V(\boldsymbol{l_{ij}}; \boldsymbol{\alpha}) = \exp(-\alpha_r|l_{ijr}| - \alpha_c|l_{ijc}|) \tag{4.5}$$

Field experiments are typically arranged in a regular array such that distances between plots may be measured by the number of rows or columns between them. Let l^*_{ijr} be the number of rows between plots i and j and similarly for l^*_{ijc}. If d_r is the actual distance

between the centre of plots in the row direction then $l_{ijr} = d_r l^*_{ijr}$ and similarly for the columns. Then the function in (4.5) may be rewritten as

$$V(\boldsymbol{l_{ij}; \alpha}) = \rho_r^{|l^*_{ijr}|} \rho_c^{|l^*_{ijc}|} \tag{4.6}$$

where $\rho_r = \exp(-\alpha_r d_r)$ and $\rho_c = \exp(-\alpha_c d_c)$.

The function in (4.6) is the correlation function for a separable first order autoregressive process denoted by AR1×AR1 and $\boldsymbol{\alpha} = (\rho_r, \rho_c)$ are the autoregressive correlation coefficients. This is one of many possible forms of $\boldsymbol{\Sigma}$, however the AR1×AR1 model tends to provide an adequate variance structure for local spatial trend in field trials.

4.2 Diagnostics in the Modelling Process

Residual plots and sample variograms are diagnostic tools used to determine whether the appropriate variance structure has been applied to the data as well as identify any extraneous effects. Residual plots, here, refer to graphs of estimated residuals, $\tilde{\boldsymbol{e}}$, against the plot row or column number.

Initially for the spatial modelling process, $\boldsymbol{\eta}$ is omitted from the model so that $\boldsymbol{e} = \boldsymbol{\zeta}$ and the AR1×AR1 model is the assumed variance structure of the error component. The residuals from this model are used to assess this variance structure for the local trend and identify any global and extraneous variation as well as the need for the measurement error component.

4.2.1 The Variogram

Variograms are tools used to gain a graphical interpretation of the spatial dependence in a data set. The variogram for the general two-dimensional spatially correlated stochastic process, $\boldsymbol{E(.)}$, for two locations $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ is given as

$$\omega(\boldsymbol{s_i}, \boldsymbol{s_j}) = \frac{1}{2}\text{Var}\left[\boldsymbol{E(s_i)} - \boldsymbol{E(s_j)}\right]$$

For a second order stationary spatial trend process, $\boldsymbol{\zeta}$, the theoretical variogram is

$$\omega(\boldsymbol{s_i}, \boldsymbol{s_j}) = \omega(\boldsymbol{I_{ij}}) = \sigma^2\{1 - V(\boldsymbol{I_{ij}; \alpha})\} \tag{4.7}$$

as $\omega(\boldsymbol{s_i}, \boldsymbol{s_j}) = \omega(\boldsymbol{s_i} - \boldsymbol{s_j})$ for stationary processes and where V is as defined in the (4.6). The variogram defined in (4.7) increases monotonically in both the row and column direction as the distance between plots increases and thus the correlation between plots, V, decreases. The spatial dependence increases to a plateau at the variance σ^2, where the greater the autoregressive correlation coefficients, the slower the increase.

The sample variogram is calculated as half of the squared differences of residuals between locations (called semivariances),

$$v_{ij} = \frac{1}{2}[e_i(\boldsymbol{s_i}) - e_j(\boldsymbol{s_j})]^2, \quad i, j = 1, \ldots, n; i \neq j \tag{4.8}$$

In practice, the residuals vector is replaced by the estimate $\tilde{\boldsymbol{e}} = \{\tilde{e}_i(\boldsymbol{s_i})\} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} - \boldsymbol{Z}\tilde{\boldsymbol{u}}$ as the true residual vector is unknown, producing an estimate of the sample variogram, $\tilde{v}_{ij}$. An estimate of the sample variogram is sufficient as it is merely an informal diagnostic tool.

Due to the assumption of equal plot sizes, an average of values of $\tilde{v}_{ij}$ with the same plot displacement $\boldsymbol{l^*_{ij}}$, can be taken, denoted by $\bar{v_{ij}}$. Thus the sample variogram is defined by the axes $(l^*_{ijr}, l^*_{ijc}, \bar{v_{ij}})$ in a 3-dimensional plot.

4.2.2 Identification of Global Trends

A global trend in the data can be identified using a residual plot, where an indication of global trend in the row direction will appear as a smooth linear or non-linear trend in the residuals over the row number for every column.

If the sample variogram doesn't plateau at the variance σ^2 in either the row or column direction, this is another indication of a global trend. This can be demonstrated by examining the rows of a single column, assuming the the errors are linearly related to the row co-ordinates such that,

$$e_i = \zeta_i + \beta s_{ir}$$

The variogram for $\boldsymbol{e}$ is given by

$$\begin{aligned}
\omega(l_{ijr}) &= \frac{1}{2}\mathrm{E}\big[(e_i - e_j)^2\big] \\
&= \frac{1}{2}\mathrm{E}\Big[\big((\zeta_i - \zeta_j) + \beta(s_{ir} - s_{jr})\big)^2\Big] \\
&= \frac{1}{2}\mathrm{E}\big[(\zeta_i - \zeta_j)^2\big] + \frac{1}{2}\beta^2(s_{ir} - s_{jr})^2 \\
&= \sigma^2\{1 - \rho_r^{|l^*_{ijr}|}\} + \frac{1}{2}\{\beta d_r l^*_{ijr}\}^2 \qquad (4.9)
\end{aligned}$$

As the row displacement l^*_{ijr} increases, the first term approaches σ^2 but the second term keeps increasing and so the variogram doesn't plateau.

Non-stationarity induced by global trends can be modeled by fitting a polynomial or cubic spline to the row or column co-ordinates of the plots.

4.2.3 Identification of Extraneous Variation

Extraneous data is a result of how a field is managed, typically the procedures of how plots are sown and harvested which can result in row and column effects in the data. This form of variation can be visualised again, in a variogram. A saw-tooth appearance indicates the presence of cyclic row or column effects that occur from the movement of the harvester up and down the rows/columns of a field. As this pattern is systematic it can be fitted in the model as fixed effect term.

For non-systematic variation associated with rows or columns, random effects can be fitted in the model to accommodate for this. Again, the variogram can be used to diagnose this random variation. To demonstrate, consider the two dimensional process as a sum of three random terms,

$$e_i = \alpha_r + \beta_c + \eta_i$$

where $e_i = e_i(s_{ir}, s_{ic})$ is the ith plot error, α_i is the effect of row r, β_c is the effect of column c and $\eta_i = \eta_i(s_{ir}, s_{ic})$ is the residual. The variance components for each term are denoted by σ_r^2, σ_c^2 and σ^2 so that the variogram for $\boldsymbol{e}$ is

$$\omega(l_{ijr}, l_{irc}) = \begin{cases} 0 & \text{if } l^*_{ijr} = l^*_{ijc} = 0 \\ \sigma_r^2 + \sigma^2 & \text{if } l^*_{ijr} \neq 0, l^*_{ijc} = 0 \\ \sigma_c^2 + \sigma^2 & \text{if } l^*_{ijr} = 0, l^*_{ijc} \neq 0 \\ \sigma_r^2 + \sigma_c^2 + \sigma^2 & \text{otherwise} \end{cases}$$

Therefore if there are random row effects, the value of the variogram will be lower at zero row displacement compared to other row displacement and similarly for columns.

4.2.4 Measurement Error

Until now the measurement error, $\boldsymbol{\eta}$, has been omitted from the residuals but recall the original spatial model defined in (4.1) had $\boldsymbol{e} = \boldsymbol{\zeta} + \boldsymbol{\eta}$. Inclusion of the measurement error results in a jump discontinuity in the variogram at zero displacement. Referring back to the variogram outlined in (4.7) but now including the measurement error in the residuals vector, the variogram becomes

$$\omega(\boldsymbol{l_{ij}}) = \begin{cases} \sigma_\eta^2 + \sigma^2\left(1 - \rho_r^{|l^*_{ijr}|}\rho_c^{|l^*_{ijc}|}\right) & \text{if } \boldsymbol{l^*_{ij}} \neq 0 \\ 0 & \text{if } \boldsymbol{l^*_{ij}} = 0 \end{cases}$$

If σ_η^2 is near zero the jump will be negligible, the variogram will appear to have a zero intercept and thus it is unlikely that the measurement error is needed in the model. This can be difficult to view on a 3D surface of a variogram and so slices corresponding to zero displacement in the column/row direction are examined. For the zero column displacement slice, if superimposing the fitted values without $\boldsymbol{\eta}$ is a poor fit, this indicates the need for the measurement error. The fitted values for this case are $\hat{\sigma^2}\{\hat{\rho_r}^{|l_{irj^*}|}\}$. If it is unreasonable to constrain a zero intercept and thus omit the measurement error, the fitted varigoram will rise too quickly to the plateau as the autoregressive correlation, ρ_r, has been underestimated. The same concept can be applied to the variogram slice with zero row displacement.

5 Example of Spatial Analysis

The following analysis is conducted on MET data of a wheat variety selection experiment carried out by the AGT in 2013, presented in Chapter 2. In order to develop a fuller understanding of the concepts discussed in Chapter 4, the spatial model outlined in (4.1) was fitted to the data for each location individually. A residuals plot against column number and a variogram are produced for each location and examined in order to identify the types of variation in the data. This is done in order to achieve more accurate estimates of variety effects. After examination of these diagnostic tools, if determined necessary, the choice of plot error variance model is updated in order to produce a truer model. For this report, the analyses of the Angas Valley and Gnowangerup trials are presented here.

5.1 Angas Valley Trial

The model presented in (4.1) was fitted to the Angas Valley data, where Table 5.1 gives an overview of the processes used to update the model. For this location, there are 288 plots arranged in a rectangular array of 12 columns by 24 rows. The first 6 columns make up the first block and columns 7 to 12 make up the second block.

Table 5.1: Summary of error models for Angas Valley trial: residual log-likelihood, l_R, and likelihood ratio test, REMLRT.

	Sources of Variation				
Model	global/ extraneous	local	variance parameters	l_R	REMLRT
1.		AR1xAR1 (0.6506, 0.2847)	4	265.9405	
2.	ran(col)	AR1xAR1 (0.16897, 0.24554)	5	273.4744	15.06781 ($p < 0.01$)
3.	ran(col)	AR1xAR1 + me (0.9619, 0.5834, 0.78774)	6	276.005	5.16292 ($p = 0.02$)

For Model (1), the covariance model for the local trend is initially assumed to be a first order two-dimensional autoregressive (AR1×AR1) structure whereby

$$\sigma^2\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{\Sigma_c} \otimes \boldsymbol{\Sigma_r}) \tag{5.1}$$

where $\boldsymbol{\Sigma_c}$ and $\boldsymbol{\Sigma_r}$ are (12×12) and (24×24) matrix functions of the column (ρ_c) and row (ρ_r) autoregressive parameters. Gilmour et al. (1997) suggest this is a useful model to commence the spatial modelling process. The measurement error, $\boldsymbol{\eta}$, is initially omitted from this model. Variety effects and the block effects are regarded as random and so included in $\boldsymbol{u}$ and $\boldsymbol{\tau}$ is fixed as the grand mean.

For the ease of this example, the `yieldkgha` variable in the `newprep.df` data set was converted from kilograms per hectare to tonnes per hectare and the `name` variable was renamed `Geno`,

```
R> newprep.df$yield <- newprep.df$yieldkgha/1000
R> newprep.df$Geno <- factor(as.character(newprep.df$name))
```

Firstly the data for the Angas Valley trial is extracted from the data-set into a table `ANR.df`, including the relevant variables for spatial analysis.

```
R> ANR.df<- subset(newprep.df, location == "Angas Valley", select =
+    c(plot, Column, Row, expt, Geno, yield, Block))
```

ASReml is then used to fit the initial model which is updated until convergence of the log-likelihood function,

```
R> ANR.asr <- asreml(yield~1,random=~Geno+ Block ,
+    rcov=~ar1(Column):ar1(Row), data=ANR.df, na.method.X='include')
R> ANR.asr <- update(ANR.asr) #likelihood converges
```

A summary of the variance components was then produced and the iterative sequence converged to row and column correlation parameters of (0.6505629, 0.2841678) respectively.

```
R> gg <- ANR.asr$gammas
R> gg
```

```
Block!Block.var   Geno!Geno.var      R!variance    R!Column.cor
      0.4417453       1.0959736       1.0000000       0.2841678
      R!Row.cor
      0.6505629
```

By examining the residuals plots in Figure 5.1, there appears to be a non-linear trend in several of the columns, particularly in columns 3 and 4. Potential outliers might be identified in column 12 (row 22) and column 11 (row 24). However as there is still much

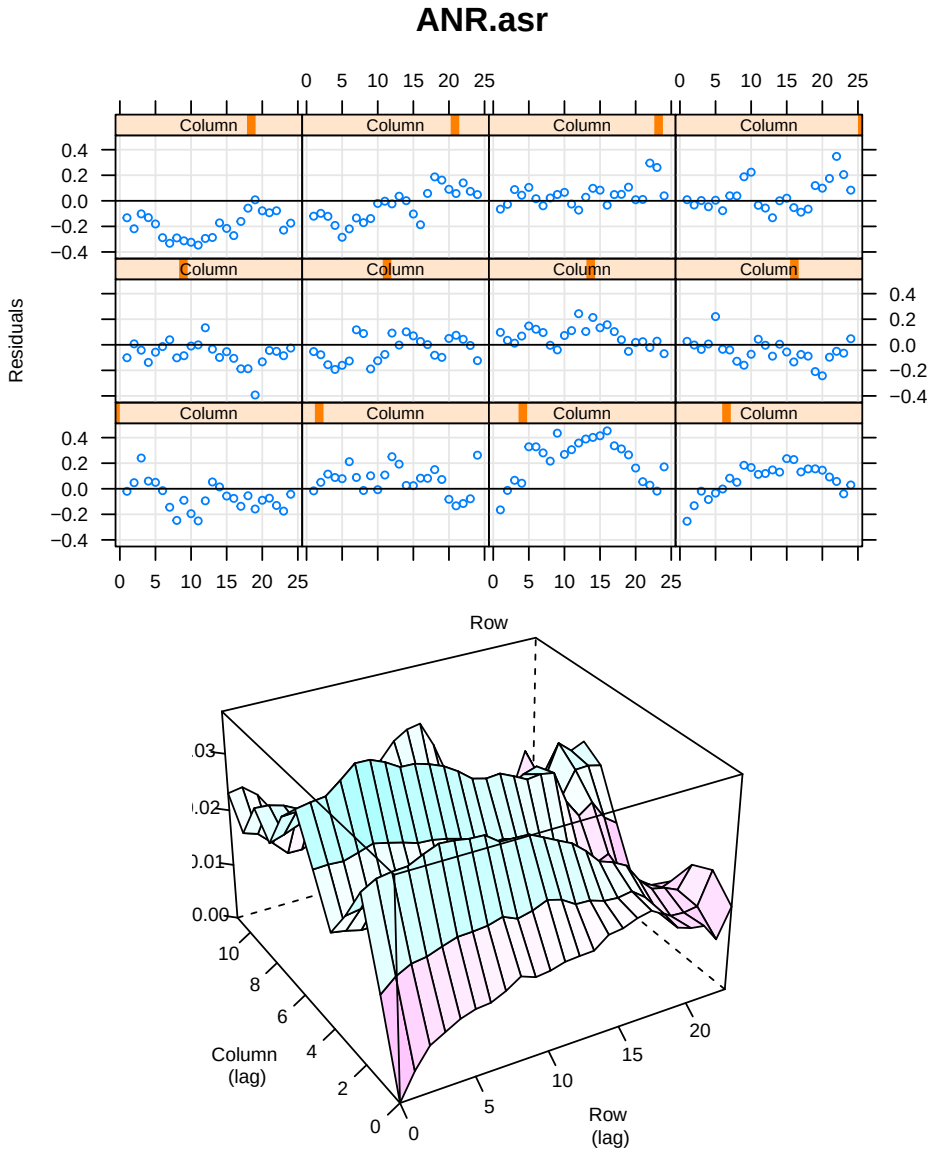

Figure 5.1: Angas Valley Model (1). Top: Residuals against row number for each column. Bottom: 3-dimensional plot of sample variogram. x- and y- ordinates are displacements in the row and column directions respectively, measured as the difference in the number of rows/columns.

investigation into the formal testing of outliers. Consultation with the experimenters would be required in order to determine whether these data points are errors and to be excluded from analysis.

Examination of the variogram of the residuals should ideally reveal little structure other than that resulting from the AR1×AR1 local covariance structure in order for the model to be considered appropriate. The theoretical variogram for the AR1×AR1 model smoothly increases in both the x and y direction to an asymptote at the process variance. Any

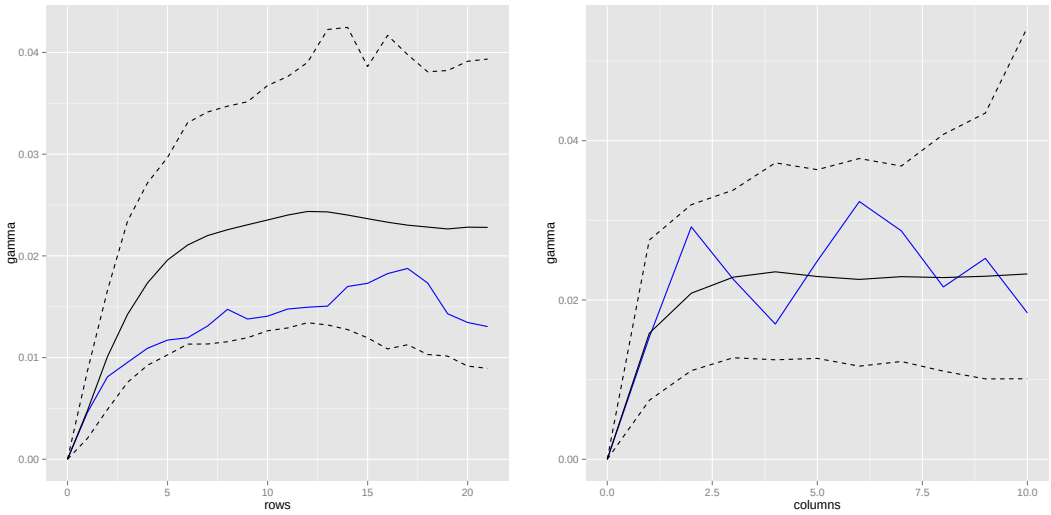

Figure 5.2: Angas Valley: Plot of row face (left) and column face (right) of sample variogram (blue line) augmented with mean (solid black line) and approximate 95% confidence intervals (dashed line) for Model (1).

additional structure is indicative of the presence of a global trend or some form of extraneous variation. Initial observations of the variogram in Figure 5.1 suggest that the data exhibits a departure from this form.

Achieving a comprehensive understanding of the appropriateness of the covariance structure fitted is difficult when examining the complete variogram and so plots of the two faces are produced. These column face and row face correspond to the zero row displacement and zero column displacement splices of the variogram, respectively. The 95% coverage intervals for the theoretical AR1×AR1 variogram are added to the plot to aid in the analysis. These are obtained by producing simulations of the the theoretical model. Firstly the spatial model in (4.1) is fitted to the Angas Valley data and estimates of the variance parameters obtained. Given these estimates, a large number (N) of data values are generated by simulating the random effects from the distribution $\mathbf{N}(\mathbf{0}, \sigma_g^2 \boldsymbol{G}(\boldsymbol{\gamma_g}))$ and residuals from $\mathbf{N}(\mathbf{0}, \sigma_g^2 \boldsymbol{\Sigma}(\boldsymbol{\alpha_g}))$, where the subscript g indicates that these are variance parameter estimates. The spatial model is then fitted to the generated data and the sample variogram ordinates computed for the row and column faces. For each face, the mean ordinate for the each displacement is calculated as well as the 2.5% and 97.5% percentiles, and these are augmented onto the sample variogram faces for the observed data.

The column and row faces of the variogram for the Angas Valley data are shown in Figure 5.2. Examining the row face, it can be seen that the semivariances of the observed data are consistently lower than the mean variogram ordinates generated under Model (1), suggesting that the presence of random column effects in the data. This is supported by the sample variogram in Figure 5.1 where there is an increase in the variogram ordinates from zero to a non-zero column displacement for any row displacement.

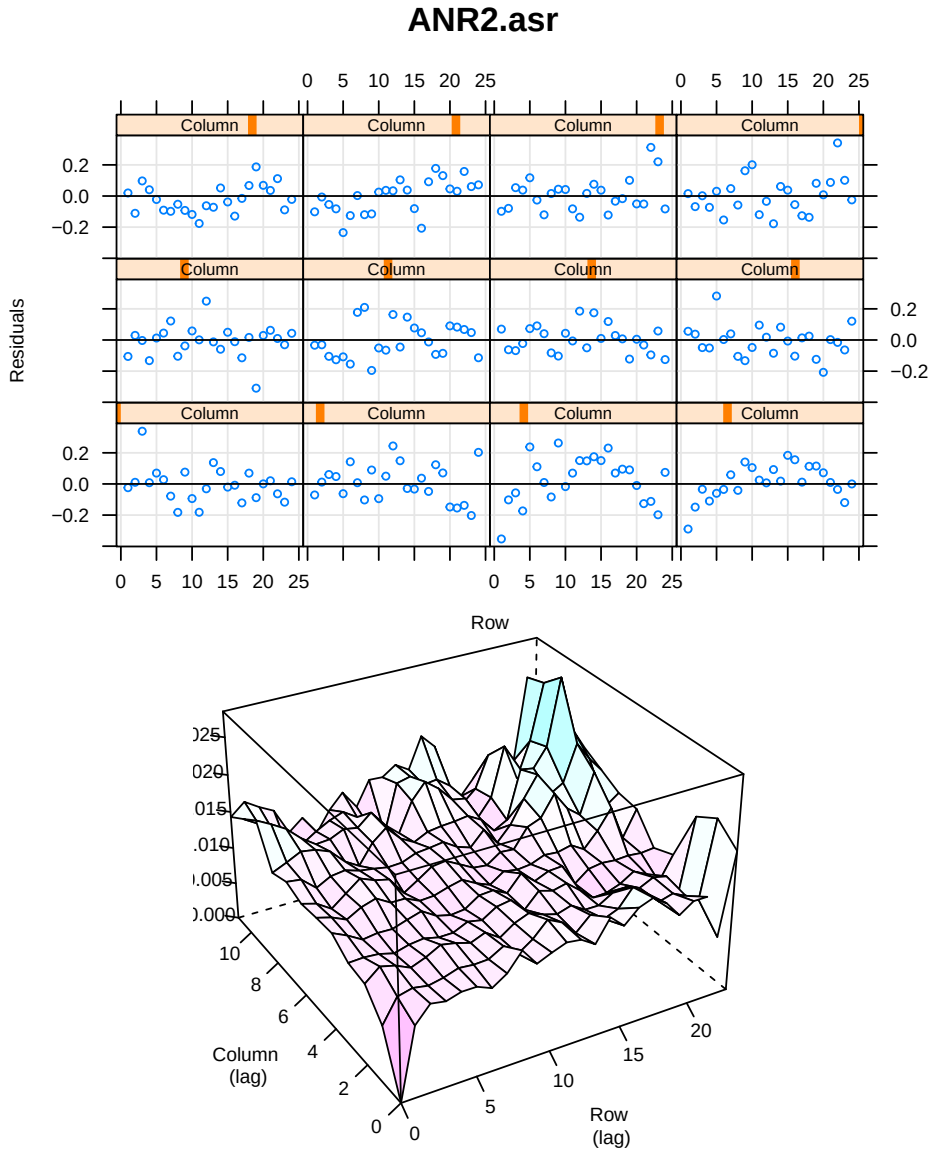

Figure 5.3: Angas Valley Model (2). Top: Residuals against row number for each column. Bottom: 3-dimensional plot of sample variogram for Model (2). x- and y- ordinates are displacements in the row and column directions respectively, measured as the difference in the number of rows/columns.

Model (2) is an updated spatial model fitted, now including the addition of random column effects. The variogram for Model (2) in Figure 5.3 more closely resembles the structure of the theoretical variogram for the given variance structure and thus the addition of random column effects appears to have resulted in an improved model. This is further supported by the plots of the row and column faces in Figure 5.4 where the sample variogram for the observed data much more closely follows the mean variogram ordinates based on the simulated data. Model (2) was compared formally to Model (1) using the residual maximum likelihood ratio test (REMLRT) and the results shown in Table 5.1. All of

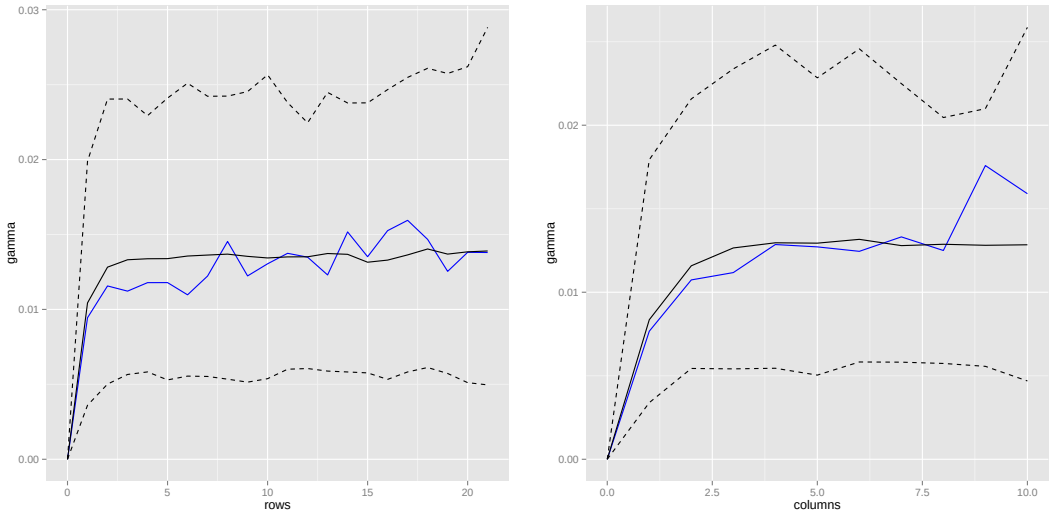

Figure 5.4: Plot of row face (left) and column face (right) of sample variogram (blue line) augmented with mean (solid black line) and approximate 95% confidence intervals (dashed line) for Model (2).

these models for the Angas Valley trial have nested variance structures but the same fixed effects model, allowing for comparisons between them. The test revealed Model (2) has a significantly higher residual log-likelihood and thus is an improvement on the original model. There is an apparent anomaly in the column face in Figure 5.2, with a peak in the pairwise residual difference for plots greater than 8 columns apart. This the result of there being fewer plots making up the mean variogram ordinate for higher column lags. A suggestion to improve this might be to trim the variogram when the number of plot pairs for higher displacements is too small.

Finally the measurement error term, $\boldsymbol{\eta}$, is included in Model (3). It has been discussed that there is significant statistical motivation to include this component of the error term in the model (Wilkinson et al. 1983). Referring to Table 5.1, the inclusion of the measurement error resulted in a significantly improved residual likelihood ($p = 0.2$).

5.2 Gnowangerup Trial

Similarly to the Angas Valley trial, there are 288 plots for the Gnowangerup experiment arranged in a rectangular array of 12 columns by 24 rows. The model outlined in (4.1) is fitted to the data, with random variety effects, $\boldsymbol{u}$, and the grand mean regarded as fixed. For Model (1), the two dimensional first order autoregressive (AR1×AR1) covariance structure is assumed, omitting the measurement error term.

The data for the Gnowangerup trial is extracted from the `newprep.df` table as well as the variables used in the spatial analysis and **ASReml** is used to fit Model (1). Examining Figure 5.6, there is a clear non-linear trend in the residuals across the rows for each column and the sample variogram for Model (1) exhibits deviation from the theoretical

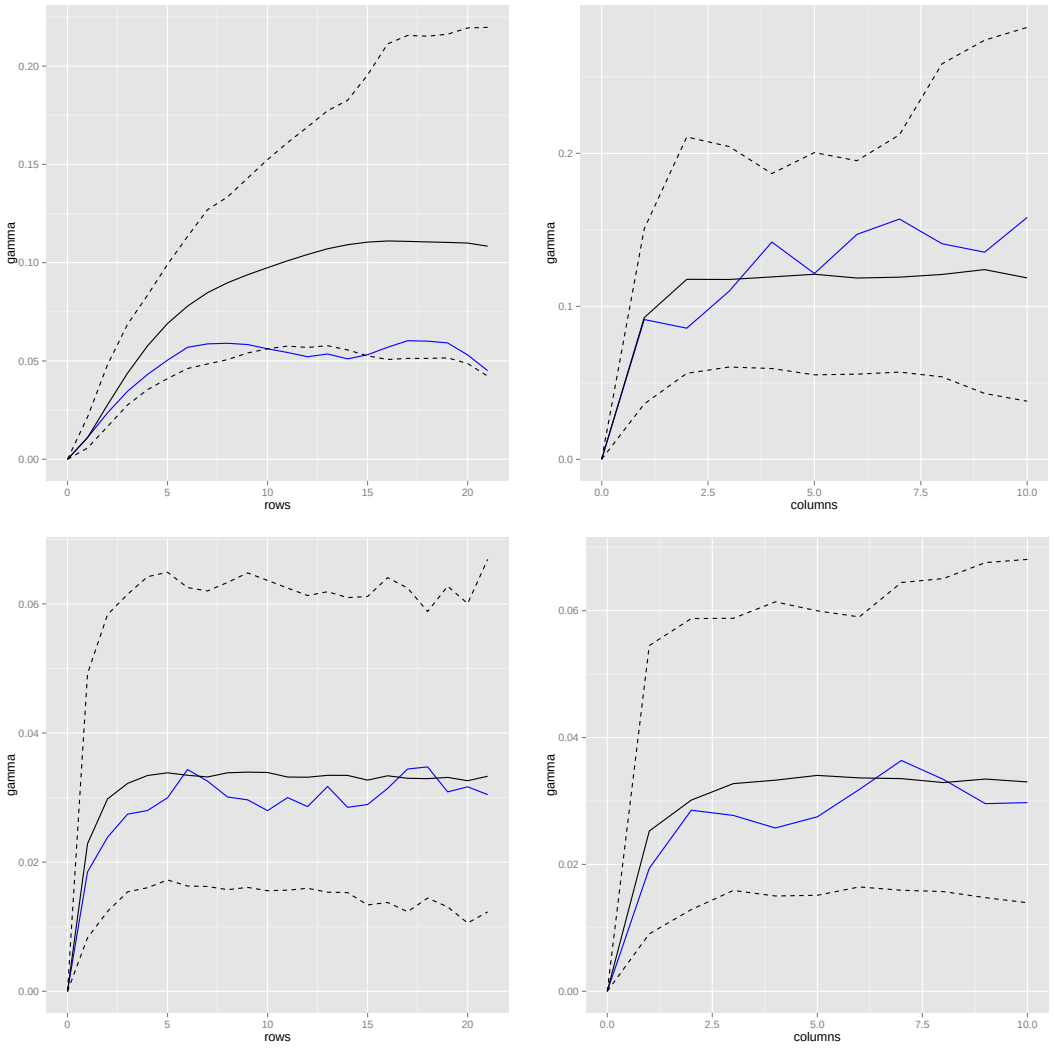

Figure 5.5: Top: Plot of row face and column face for Model (3). Bottom: Plot of row face and column face for Model (4).Line for sample variogram (blue line) augmented with mean (solid black line) and approximate 95% confidence intervals (dashed line)

variogram, where the semivariances fail to plateau in the row direction. Instead, there is a steady increase in semivariance with increasing row displacement, implying the presence of a linear drift in the residuals of the observed data. The effects of this non-stationarity can be accommodated for by including a linear regression of yield on row number. This global trend is fitted in Model (3) as the sum of a linear regression, which introduces a regression coefficient to the fixed effects, and a random curve component with the addition of random spline effects.

The column and row faces for the variogram for Model (3) are depicted at the top of Figure 5.5. It is clear that there is a departure of the sample variogram from the theoretical model. This is seen most apparently for the row face, where the semivariances of the observed data are lower than the mean of the simulated data for all row displacements and the sample variogram even drops out of the confidence interval at several displacements. This

is corrected for in Model (4) by the inclusion of random column effects in the assumed model. The row and column faces at the bottom of Figure 5.5 for Model (4) demonstrate an improvement to the model, where the sample variogram for the observed data now more closely follows the mean variogram ordinates of the simulation.

Table 5.2 provides a summary of the models fitted to the Gnowangerup trial including the correlation parameter estimates for each model and the residual likelihood. The final update to the process was Model (5) that allowed for the inclusion of the measurement error. This model was determined to be the most appropriate for the data set, with a significant result in the REML ratio test suggesting an improved fit. Comparisons are made between models not separated by a line in Table 5.2, as these share the same fixed effects.

Table 5.2: Summary of error models for Gnowangerup trial: residual log-likelihood, l_R, and likelihood ratio test, REMLRT.

	Sources of Variation				
Model	global/ extraneous	local	variance parameters	l_R	REMLRT
1.		AR1xAR1 (0.8516, 0.5523)	4	16.04754	
2.	lin(row)	AR1xAR1 (0.8251, 0.4535)	4	14.92534	
3.	lin(row)+spl(row)	AR1xAR1 (0.7826, 0.2133)	5	18.97028	8.089873 ($p = 0.004$)
4.	lin(row)+spl(row) +ran(column)	AR1xAR1 (0.2458, 0.1803)	6	27.18168	16.42279 ($p < 0.01$)
5.	lin(row)+spl(row) +ran(column)	AR1xAR1+me (0.8497, 0.6547, 1.106)	7	29.1339	3.904456 ($p = 0.04$)

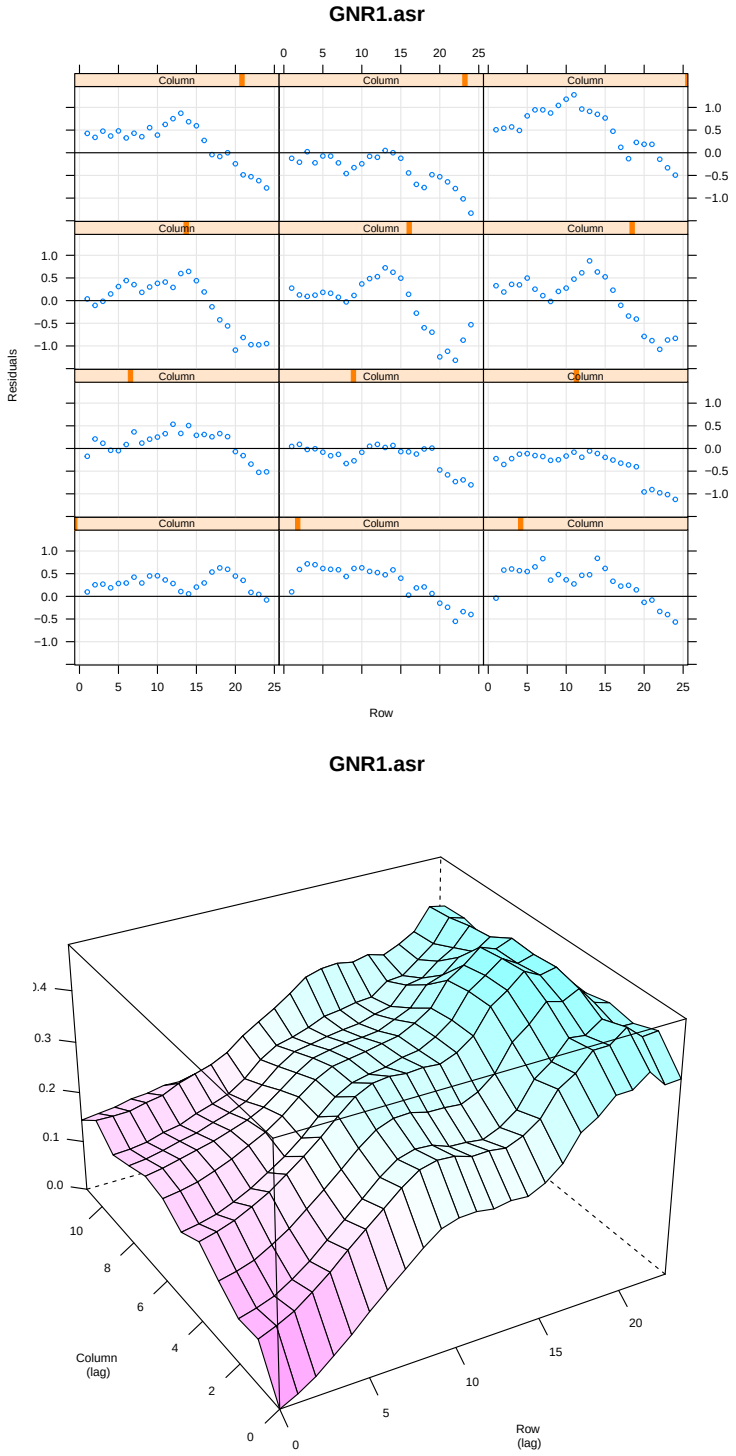

Figure 5.6: Gnowangerup Model (1). Top: Residuals against row number for each column. Bottom: 3-dimensional plot of sample variogram. x- and y- ordinates are displacements in the row and column directions respectively, measured as the difference in the number of rows/columns.

Bibliography

[1] Cullis B.R, Smith A.B, Gogel B.J, Butler D.G, Verbyla A.P 2007, *ASReml-R: Theory and Practise*, The State of Queensland, Department of Primary Industries and Fisheries, Queensland.

[2] Cullis B. R, Gogel B. J, Verbyla A. P, Thompson R. 1998, 'Spatial analysis of multi-environment early generation trials', *Biometrics*, vol.54, pp.1-18.

[3] Cullis B.R, Smith A.B, Coombes N.E 2006, 'On the Design of Early Generation Variety Trials With Correlated Data', *Journal of Agriculture, Biological and Environmental Statistics*, vol.11, no.4, pp.381-393.

[4] Gilmour A.R, Cullis B.R, Verbyla A.P 1997, 'Accounting for natural and extraneous varitation in the analysis of yield experiments, *Journal of Agricultural, Biological, and Environmental Statistics 2*, pp. 269-273.

[5] R Core Team 2012, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[6] Robinson, G.K 1991, 'That BLUP is a Good Thing: The Estimation of Random Effects', *Statistical Science*, vol.6, no.1, pp. 15-51.

[7] Smith, A.B, Cullis B, Thompson R 1999, 'Multiplicative mixed models for the analysis of multi-environment trial data', *Biometrics*, vol. 57, no.4, pp. 1138-1147.

[8] West B.T, Welch K.B, Galecki A.T, 2007, *Linear Mixed Models: A Practical Guide Using Statistical Software*, Chapman and Hall, London.