# Use of genetic relationship matrices in the prediction of breeding values and their accuracy assessment

Katia Stefanova[1], Wallace Cowling[1], Arthur Gilmour [2]

[1] The UWA Institute of Agriculture, University of Western Australia, Australia
[2] Cargo Vale, Cargo, Australia

E-mail for correspondence: katia.stefanova@uwa.edu.au

**Abstract:** The paper presents a statistical model to obtain the predicted breeding values with accuracy of selection in the context of plant breeding trials in two cycles of selection by using genetic relationship matrices.

**Keywords:** Linear mixed model; Predicted breeding values; BLUPs; Multi-environment trial; Accuracy of selection.

## 1 Introduction

Analyses of single trials and multi-environment trials (METs) assume IID normal genetic effects, although for METs genetic variances and correlations may differ between trials. The parental covariances arising from common ancestors are ignored. In reality the test lines in single trials or METs are genetically related and using this information can improve the accuracy of prediction of breeding values. Relatedness may be determined from the pedigree or from genetic markers. This allows the individual genetic effects to be partitioned into *additive* and *non-additive* genetic effects (Oakey et al., 2007). The additive effect reflects the ability of a line as a parent and the non-additive effect is associated with dominance and residual effect.
The paper presents a statistical model incorporating genetic relationship matrix for the analysis of two cycles of selection in a plant breeding trial. Based on the model, predicted breeding values (PBV) and the accuracy of selection are obtained. The approach is illustrated with data from a self-pollinating crop (*Pisum Sativum*) for resistance to ascochyta blight (*Didymella pinodes*) complex (Cowling et al., 2015) A new breeding approach, utilizing F1-recurrent selection and the animal breeding model, is

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015. 

used to accelerate response to selection in this self-pollinating crop, including phenotypic and relationship records from self progeny. Using PBV, narrow-sense heritability and the accuracy of selection allows comparison of breeding strategies.

## 2 Motivating example

Traditionally in a plant breeding trial, selfed progenies of F1s are intercrossed, not the F1s themselves. The innovation in this study is that using relationships allows evaluation and hence selection of the F1 plants through their progeny. Each self is identified as a genotype with explicitly stated identical parents inside the pedigree tree.

### 2.1 Phenotypic data

A total of 1139 and 1077 early generation progeny plants were tested in cycles 1 and 2, respectively. Plots were single plants spaced $1m$ apart in rows and columns, with 40 columns by 30 rows in cycle 1 grown in 2010, and 20 columns by 70 rows in cycle 2 grown in 2013. We consider the two cycles as two experiments (trials). Replication and concurrency of the genotypes across experiments were provided by pure lines including founder lines, Australian cultivars and Chinese landraces.

### 2.2 Experimental design

Trials were laid out as rectangular arrays of rows and columns using a partially replicated (*p-rep*) design (Cullis et al., 2006). The design was generated using package DiGGer (Coombes, 2009).

## 3 Statistical model

Denote the vector of plot yields $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_t^{\mathrm{T}})^{\mathrm{T}}$, $t$ is the number of trials. The model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_a + \mathbf{Z}_g\mathbf{u}_{\bar{a}} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e},$$

where $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_{\bar{a}}$ and $\mathbf{u}_g, \mathbf{u}_a, \mathbf{u}_{\bar{a}}$ are respectively the vectors of genetic, additive genetic and non-additive genetic effects, $\boldsymbol{\tau}$ is the vector of fixed effects with design matrix $\mathbf{X}$, $\mathbf{u}_p$ is the vector of random non-genetic effects with design matrix $\mathbf{Z}_p$ and $\mathbf{e}$ is the vector of plot errors combined across trials.
The random effects are assumed to follow Gaussian distribution

$$\begin{pmatrix} \mathbf{u}_a \\ \mathbf{u}_{\bar{a}} \\ \mathbf{u}_p \\ \mathbf{e} \end{pmatrix} \sim N\left[ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{G}_a & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\bar{a}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R} \end{pmatrix} \right]$$

and

$$\text{var}(\mathbf{y}) = \mathbf{Z}_g\mathbf{G}_a\mathbf{Z}_g^{\mathrm{T}} + \mathbf{Z}_g\mathbf{G}_{\bar{a}}\mathbf{Z}_g^{\mathrm{T}} + \mathbf{Z}_p\mathbf{G}_p\mathbf{Z}_p^{\mathrm{T}} + \mathbf{R}.$$

The variance matrix for non-genetic random effects $\mathbf{G}_p$ is usually a diagonal matrix of scaled identity matrices. The variance matrix for the plot error effects is assumed block diagonal with $\mathbf{R} = \text{diag}(\mathbf{R}_j)$, where $\mathbf{R}_j$ is the plot error matrix for the $j$th trial. The data are spatially modelled and respectively ordered as rows within columns, reflecting the field layout $\mathbf{R}_j = \sigma_j^2\boldsymbol{\Sigma}_{c_j} \otimes \boldsymbol{\Sigma}_{r_j}$, where $\sigma_j^2$ is a scale parameter and $\Sigma_{c_j}$ and $\Sigma_{r_j}$ are the correlation matrices for column and row dimensions of the trial $j$, corresponding to autoregressive processes of order one. The variance matrix for the additive genetic effects $\mathbf{G_a}$ is assumed to have the separable form $\mathbf{G}_a = \mathbf{G}_{e_a} \otimes \mathbf{G}_{v_a}$, where matrix $\mathbf{G}_{e_a}$ is the matrix of additive genetic variances/covariances between environments, matrix $\mathbf{G}_{v_a}$ is the matrix of additive genetic variances/covariances between genotypes. We set $\mathbf{G}_{v_a} = \mathbf{A}$ where the known numerator relationship matrix $\mathbf{A} = \{a_{ij}\}$ is given by $a_{ii} = 1 + F_i$ and $a_{ij} = 2f_{ij}$. $F_i$ is the inbreeding coefficient of genotype $i$ and $f_{ij}$ is the coefficient of parentage between genotypes $i$ and $j$. In a similar manner we assume that the non-additive effects may be presented as a two-way structure of genotype by environments effects, respectively with variance $\mathbf{G}_{\bar{a}} = \mathbf{G}_{e_{\bar{a}}} \otimes \mathbf{G}_{v_{\bar{a}}}$. We assume independence between the non-additive genetic components and set $\mathbf{G}_{v_{\bar{a}}} = \mathbf{I}_m$. The matrix $\mathbf{A}^{-1}$ and the models are computed using ASReml-R (Butler et al., 2009).

## 4 PBV, accuracy and heritability

PBV is the obtained best linear unbiased predictor (BLUP) for each genotype. More precisely, it is the empirical BLUP (E-BLUP), since the unknown variance parameters have been replaced by their REML estimates in the mixed model equation.

The accuracy $r$ of PBVs is the correlation between the true and predicted breeding values, and is sometimes reported as *reliability*, the squared correlation ($r^2$) (Mrode, 2005). The accuracy $r$ of the predicted breeding value for the $i$th genotype at an individual environment was calculated following Gilmour et al. (2009) for the animal model

$$r_i = \sqrt{1 - \frac{s_i^2}{(1 + f_i)\sigma_a^2}},$$

where $s_i^2$ is the prediction error variance for the $i$th genotype, $\sigma_a^2$ is the additive genetic variance and $1 + f_i$ is the diagonal element of the relationship matrix $\mathbf{A}$. Narrow-sense heritability $h^2$ was calculated as $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_{\bar{a}}^2+\sigma_e^2}$.

## 5 Results

The fitted model identified significant linear row and column effects for both trials, the local variation ($AR1 \times AR1$) was not that strong. The variance/covariance structure fitted for the genotype by environment interaction effect was general correlation with heterogeneous variances, equivalent to the unstructured in the case of two environments. The latter produced separate estimates of the trials variances and the genetic correlation between the trials which appeared quite high, 0.82. The inclusion of the pedigree in the model showed significantly ($p < 0.001$) better fit. The average accuracy of predicted breeding values in the combined analysis without pedigrees was 0.562 (genotypes in cycle 1) and 0.696 (genotypes in cycle 2), and increased to 0.894 (cycle 1) and 0.807 (cycle 2) when pedigree information was included.

**Acknowledgments:** The financial support of the Grains Research and Development Corporation of Australia is gratefully acknowledged.

### References

Butler, D.G., Cullis, B.R., Gilmour, A.R., and Gogel, B.J. (2009). ASReml-R reference manual. Technical Report, *Queensland Department of Primary Industries* URL http://www.vsni.co.uk/software/asreml/.

Coombes, N. (2009). *DiGGer design search tool in R.* http://www.austatgen.org/software.

Cowling, W.A., Stefanova, K.T., Beeck, C.P., Nelson, M.N., Hargreaves, B.L.W., Sass, O., Gilmour, A.R., and Siddique, K.H.M. (2015). Using the animal model to accelerate response to selection in a selfpollinating crop. *G3: Genes, Genomes, Genetics* (accepted).

Cullis, B.R., Smith, A.B., and Coombes, N.E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 381 – 393.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. (2009). *ASReml User Guide Release 3.0.* VSN International Ltd, UK. http://vsni.de/downloads/asreml/release3/UserGuide.pdf

Mrode, R.A. (2005). *Linear Models for the Prediction of Animal Breeding Values*, (2nd ed.). CABI Publishing, UK.

Oakey, H., Verbyla, A.P., Cullis, B.R., Wei, X., and Pitchford, W.S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics*, **114**, 1319 – 1332.