Loss of accuracy in the genomic prediction for prevalent two-stage analysis with single site.

Emi Tanaka[1*], Ky Mathews[2], Alison B. Smith[2], Brian R. Cullis[2]

[1]School of Mathematics and Statistics, The University of Sydney, NSW, Australia, 2006
[2]School of Mathematics and Applied Statistics, The University of Wollongong, NSW, Australia, 2522

* emi.tanaka@sydney.edu.au

Abstract

Linear mixed model is the prevailing method in genomic prediction and selection as it fits the structure of the data well. There are many forms of the analysis using linear mixed models however it is widely concurred that a single model that analyses the individual plot data, i.e. one-stage analysis, is superior to a two-stage analysis. Briefly, two-stage analysis involves computation of the adjusted genotype means in the first stage with a weighted or unweighted analysis in the second stage. A prevalent form of two-stage analysis lacks spatial modelling nor considers variance heterogeneity for genotype $\times$ environment effects.

Often in crop breeding trials, analytical approaches do not take into account the non-genetic sources of variation, either by the adoption of more suitable designs (Butler et al, 2014) or an appropriate analysis (Stefanova et al, 2009) or both. In particular, a two-stage analysis is frequently used in the analysis of crop breeding trials. We present one-stage and two-stage models for single trial analysis. Our one-stage model utilises both the pedigree and marker information and is an extension to the approach in Oakey et al (2006). Our simulation results based on 48 early generation wheat selection trials show there is almost always a loss in accuracy for the prevalent weighted and unweighted two-stage analysis over a one-stage analysis. The loss of accuracy was noticeably larger from the lack of use in spatial modelling than the use of a weighted analysis. In addition, the loss was more pronounced for partially replicated designs (Cullis et al, 2006) which are becoming widely adopted in plant improvement programs in Australia.

1 Introduction

With the advent of high-throughput genotyping technologies, molecular markers such as single nucleotide polymorphisms (SNPs) are now readily available for genomic prediction of breeding values (Crossa et al, 2011). Traditionally, breeding values were estimated based on pedigree information alone which can infer the average additive genetic effects (Crossa et al, 2010). Molecular marker information can capture the variation due to Mendelian sampling and has become prolific for use in plant improvement programs worldwide .

Advances on experimental design and analysis that accounts for non-genetic sources of variation of plant breeding programs have shown great improvement in genomic prediction (Cooper et al, 2014). More specifically, the application of linear mixed models that account for effects in relation to the experimental design (such as block effects) and field variation as described in Cullis and Gleeson (1991); Gilmour et al (1997); Stefanova et al (2009) have reduced the undesirable impact of non-genetic sources of variation. For multi-environmental trial (MET) datasets, the modelling of the between trial variance-covariance matrix can be accommodated in the two-stage approach with or without weights (Piepho et al, 2012; Smith et al, 2001a; Welham et al, 2010) or preferably in a one-stage approach with the use of multiplicative mixed models (Smith et al, 2001b). Oakey et al (2006) demonstrated that a one-stage analysis using an appropriate spatial linear mixed model that partitions the total genetic effects to additive and non-additive effects with the use of pedigree information provides better accuracies for both the additive and total genetic effects. In this paper, we extend this model to include three sets of genetic effects, namely, additive effects due to markers, residual additive effects and non-additive effects.

In a simulation study for late-stage variety evaluation trials, Welham et al (2010) demonstrated that genetic gain decreased when using a two-stage analysis, especially without using appropriate weights and when trials with low accuracy were present, compared to the one-stage approach. Their study, however, focussed on the prediction of the total genetic effects and did not incorporate marker nor pedigree data. In their study, Welham et al (2010) used the weights recommended by Smith et al (2001a) in which weights are given as the diagonal of the (asymptotic) variance-covariance matrix of the empirical best linear unbiased estimators (E-BLUEs) of the line means from a single site analysis. Piepho et al (2012) suggest an alternative weighting scheme which use the full variance-covariance matrix of the E-BLUEs. Using an empirical example they suggest that use of these weights results in predictions which are more similar to the prediction of genetic effects from a one stage analysis.

In the selection of best parents for further crosses, the interest lies in the prediction of additive genetic effects as these can be passed onto the progeny in a expected way. The common practice in current genomic prediction methods in plant breeding is to employ a two-stage analysis with a number of studies accounting of no genotype $\times$ environment (G$\times$E) interaction. For studies that consider G$\times$E effects, the model in the first stage generally predicted the main genotype effect with G$\times$E as a nuisance factor resulting in an equivalent formulation to a model that assumes a compound symmetric variance-covariance structure to the G$\times$E effects. The adjusted genotype means per trial, usually obtained from a model with fixed genotype effects, are used as the response in the second stage in either a weighted or unweighted overall mixed model analyses. The main appeal of the two-stage analysis is the computational efficiency, however, the two stage approach may be simply employed due to software restrictions. For MET datasets, a two-stage approach without weights and assuming no genetic correlation between environments in genomic prediction models is commonly employed. The between trial variance and covariance heterogeneity is occasionally accommodated in genomic selection models, albeit in a two-stage approach.

In addition, Bernal-Vasquez et al (2014); Cullis and Gleeson (1991); Gilmour et al (1997); Stefanova et al (2009) have shown clear gains in the prediction of genetic effects by taking into account spatial variation or trends in the data. The lack of incorporating spatial variation or trend is in particular prevalent in two-stage analysis.

The aim of this paper is to use real breeding trials as a basis of simulation to evaluate the loss of accuracy of one-stage analysis over prevalent two-stage approaches in prediction of breeding values using single trial information. This aim is similar to Schulz-Streeck et al (2013) but the major differences lie in that we consider spatial variation or trends and a more realistic comparison by incorporating a model selection procedure for our simulation. Furthermore, Schulz-Streeck et al (2013) did not consider heterogeneity for the genetic effects nor residuals as per Smith et al (2001b) thus we cannot draw conclusions from their model for dataset that exhibit this heterogeneity.

An outline of the paper is as follows. We describe the motivating data set that served as a basis of our simulation. The statistical models for one-stage and two-stage analysis in the context of a single trial are reviewed. The analysis of motivating data set is presented. The simulation setting is described and results are presented. We conclude with a discussion of the results.

2 Motivating dataset

The motivating dataset we consider is 48 early generation selection trials conducted by Australian Grain Technologies (AGT) wheat breeding program in 2010-2013. Each trial was planted as a rectangular array of 12 rows by 16 columns or 12 rows by 24 columns, i.e. a total of 192 or 288 plots. The number of genotypes varied from 138 to 191 with a total of 448 genotypes across the 48 trials. All trials were designed using the DiGGer software (Coombes, 2002) with default trial design parameters. Most trials (28 of the 48 trials) were designed as p-replicate designs (Cullis et al, 2006) in which a proportion p of genotypes were sown with two plots each, and the remaining genotypes sown with a single plot each. Some check varieties had additional plots. The range of p across all p-replicate trials is (0.33, 0.48) with a mean of 0.43. The remaining 20 trials had 2 replicates of all genotypes with some additional plots for check varieties. All trials had 2 resolvable blocks (for the replicated genotypes only in the p-replicate designs), such that the first replicate block comprised of columns 1-6 and the second replicate columns 7-12.

The trait of interest is grain yield and the trial mean yield ranged from 0.88 to 6.70 t/ha (Figure 1) with most trials missing phenotypic data for less than 3 plots ($< 2\%$ missing). We provide a detailed analysis of a single trial in Section 4.1.

The genotypes grown in these trials all have marker and pedigree information. These genotypes were part of the early generation selection trials and have no particular population structure (see Appendix A). The marker genotype information was obtained from a custom Axiom$^{\text{TM}}$ Affymetrix array. For each of the 17,305 SNP markers, the individual was coded as -1, 1, or 0 for homozygous minor allele, homozygous major allele or heterozygous, respectively. Missing marker information was imputed using k-nearest neighbour using the R-package `pedicure` (Butler, 2014). Pedigree information was available on a total of 1831 lines, which comprised all the genotypes grown in the 48 trials together with the ancestral varieties. Note that lines, entries, varieties and genotypes are used synonymously in this paper.

3 Statistical methods

We consider the analysis of a single trial. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of (phenotypic) data, where n is the number of plots in the trial. We assume that m_d genotypes were grown in the trial but that we only have marker data (on r markers) for $m \leq m_d$ genotypes. Pedigree information is available on m_p lines which includes the m genotypes with marker data (so $m_p \geq m$). Note that this is a general model that considers lines with

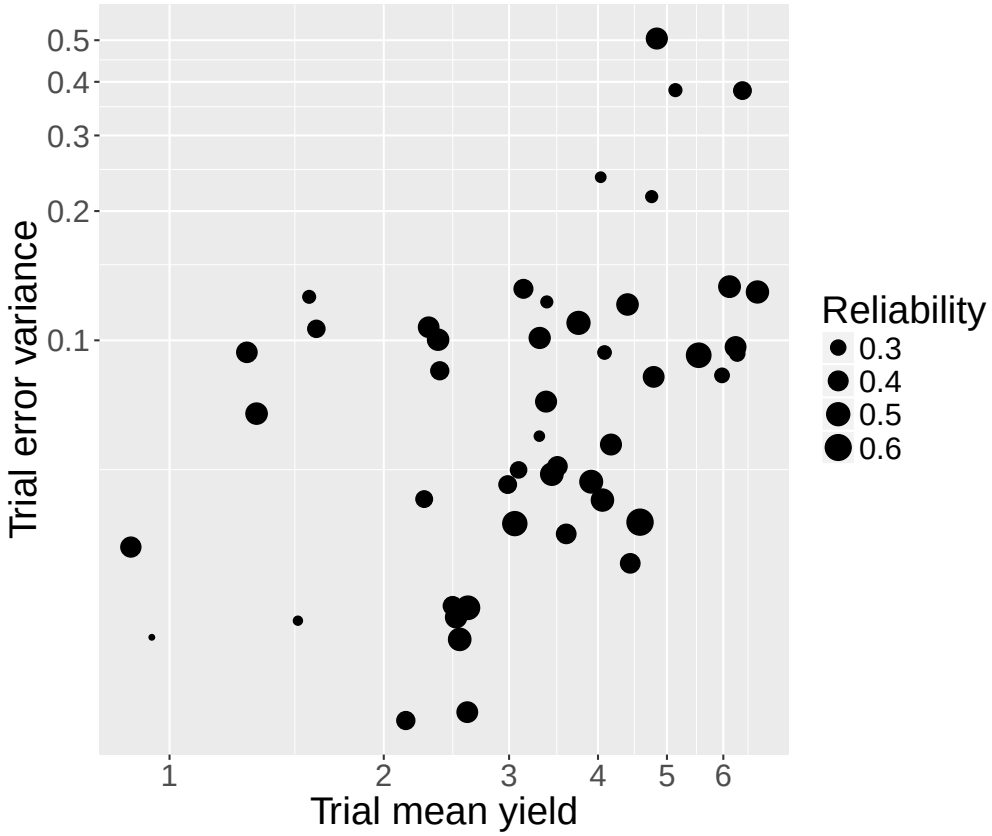

Figure 1. Plot of trial error variance versus trial mean yield (t/ha) in log-scales for the motivating dataset. The size of the points correspond to the average genotype reliability of the trial calculated as in (7).

missing marker or pedigree information and our analysis of the motivating data is a special case of this model since all our genotypes have marker and pedigree information.

3.1 One-stage analysis

Following on from the results in Appendix B and C, we can write the model for the data vector that excludes irrelevant genotypes (such as check varieties) and entries without data as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the $m \times 1$ vector of random genetic effects corresponding to those genotypes with marker data, and has associated $n \times m$ design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_p}$ is a vector of non-genetic or peripheral random effects with associated design matrix $\boldsymbol{Z_p}$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. The fixed effects are partitioned as $\boldsymbol{\tau} = (\boldsymbol{\tau_0}^\top, \boldsymbol{\tau_{g0}}^\top)^\top$ where $\boldsymbol{\tau_{g0}}$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to the genotypes without marker data and we let $\boldsymbol{X_{g0}}$ denote the associated $n \times (m_d - m)$ design matrix. Thus $\boldsymbol{X} = [\boldsymbol{X_0} \;\; \boldsymbol{X_{g0}}]$ where $\boldsymbol{X_0}$ is the design matrix associated with the (non-genetic) fixed effects $\boldsymbol{\tau_0}$. The vector $\boldsymbol{u}_p$ consists of subvectors that may include random terms for extraneous field variation such as random row or column variation and also design and randomisation based blocking factors. We assume that the vectors of random effects and residuals are mutually independent, and distributed as multivariate Normal, with zero means and var $(\boldsymbol{u_p}) = \boldsymbol{G_p}$, var $(\boldsymbol{u_g}) = \boldsymbol{G_g}$ and var $(\boldsymbol{e}) = \boldsymbol{R}$. The matrix $\boldsymbol{G_p}$ may be completely general, but in many applications, it is block diagonal with blocks given by $\sigma^2_{p_k}\boldsymbol{I}_{n_{p_k}}$ with $\sigma^2_{p_k}$ and n_{p_k} corresponding to the variance component and the length of the associated effect, respectively. The structures of $\boldsymbol{G_g}$ and $\boldsymbol{R}$ are discussed in Section 3.1.1 and Section 3.1.2, respectively.

We then consider a simple model for $\boldsymbol{u_g}$ given by

$$\boldsymbol{u_g} = \boldsymbol{u_a} + \boldsymbol{u_e} \tag{2}$$

where the two terms represent the additive and non-additive (or residual) genetic effects. Then we propose that the additive genetic effects be modelled as a linear function of the marker covariates so write

where $\boldsymbol{u_m}$ is the $m \times 1$ vector of additive genetic effects due to r markers; $\boldsymbol{M}$ is the $m \times r$ matrix of marker covariate data; $\boldsymbol{\alpha}$ is the associated $r \times 1$ vector of random marker effects (regression coefficents) and $\boldsymbol{u_\epsilon}$ is the $m \times 1$ vector of lack of fit effects for the marker regressions.

Thus the model in equation (1) can be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g M\alpha} + \boldsymbol{Z_g u_\epsilon} + \boldsymbol{Z_g u_e} + \boldsymbol{Z_p u_p} + \boldsymbol{e}. \tag{4}$$

We assume that the variance matrices of random genetic effects are given by

$$\begin{aligned} \text{var}\,(\boldsymbol{\alpha}) &= \sigma^2_m \boldsymbol{D} \\ \text{var}\,(\boldsymbol{u_\epsilon}) &= \sigma^2_\epsilon \boldsymbol{A} \\ \text{var}\,(\boldsymbol{u_e}) &= \sigma^2_e \boldsymbol{I}_m \end{aligned}$$

where $\boldsymbol{A}$ is the $m \times m$ block of the numerator relationship matrix that relates to the genotypes with marker data; $\boldsymbol{D}$ is an $r \times r$ matrix, often assumed to be the identity matrix $\boldsymbol{I}_r$, and σ^2_m, σ^2_ϵ and σ^2_e are the variances for marker effects, marker lack of fit effects and residual genetic effects, respectively.

The variance matrix for the (total) genetic effects, denoted $\boldsymbol{G_g}$, is therefore given by

$$\begin{aligned} \boldsymbol{G_g} = \text{var}\,(\boldsymbol{u_g}) &= \sigma^2_m \boldsymbol{MDM}^\top + \sigma^2_\epsilon \boldsymbol{A} + \sigma^2_e \boldsymbol{I}_m \\ &= \sigma^2_m \boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{A} + \sigma^2_e \boldsymbol{I}_m \end{aligned} \tag{5}$$

where $\boldsymbol{K}$ is the $m \times m$ genomic relationship matrix. We write $\boldsymbol{G_g} = \boldsymbol{G_g}(\sigma^2_m, \sigma^2_\epsilon, \sigma^2_e)$ to highlight that in the maximal genetic model in which both pedigree and marker information is included, it is a function of three unknown parameters.

3.1.1 Spatial modelling

In general, field trials are arranged in a rectangular array of n_r rows and n_c columns (so $n = n_r n_c$). Every trial generally have considerable sources of non-genetic variation that may not be expected at the experimental design stage, thus, it is important to perform spatial model selection at the analysis stage. Following the spatial modelling approaches of Cullis and Gleeson (1991); Gilmour et al (1997), the residuals are assumed to represent local spatial variation or trend and their variance matrix is chosen from the class of separable processes.

Suppose that $\boldsymbol{y}$ is ordered so that rows are within columns, then we write $\boldsymbol{R} = \sigma^2 \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r$ where the matrices $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_r$ are $n_c \times n_c$ and $n_r \times n_r$ correlation matrices for the column and row dimension, respectively. Often these are assumed to correspond to autoregressive processes of order one, so each is a function of a single autocorrelation parameter, denoted ρ_c and ρ_r. Here-after, this first order separable autoregressive process will be referred to as an AR1×AR1 process.

Simplifications to the correlation model may be appropriate and will be referred to as ID×AR1 if there is independence for the column dimension (i.e. $\boldsymbol{\Sigma}_c = \boldsymbol{I}_{n_c}$), AR1×ID if there is independence for the row dimension (i.e. $\boldsymbol{\Sigma}_r = \boldsymbol{I}_{n_r}$), and ID×ID if there is independence for both dimensions, in which case the standard assumption of independent and identically distributed (iid) residuals applies (i.e. $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$).

Gilmour et al (1997) discuss the need to allow for global trend and extraneous variation in addition to the local trend if any. This may be accommodated in the model by fitting appropriate effects, either fixed or random, that are related to the spatial co-ordinates (rows and columns) of the trial (see Table 2). A white noise component (also called measurement error or a nugget effect) may be added by including a sub-vector of n independent random effects in $\boldsymbol{u_p}$, provided the residual is not assumed iid. We assess the fit of the spatial model using diagnostic tools, such as coverage intervals for the sample variogram, as described in Stefanova et al (2009).

3.1.2 Genetic modelling

The baseline model in (4) assumes the presence of three sets of genetic effects, namely the additive effects associated with the markers, the residual additive effects and the non-additive genetic effects. After the spatial modelling selection, we then consider a selection of genetic effects to form the "best" one-stage model. More specifically, we assess three genetic models given the spatial model using the Akaike Information Criterion (AIC, Akaike, 1973): (1) a model that includes the three set of genetic effects as described above, denoted MPI, (2) a model that includes additive effects associated with markers and residual genetic effects denoted MI, and (3) a model that includes expected additive genetic effects based on pedigree information alone and the residual genetic effects denoted PI. These models are listed in Table 1. The model with the smallest AIC is selected as the best one-stage model.

Note the value predicted is the total additive genetic value or genomic estimated breeding values (GEBV) under the specified model. For a one-stage analysis, the genetic model may differ according to the best fit to the data. Specifically, the predicted value is $\boldsymbol{u_m} + \boldsymbol{u_\epsilon}$, $\boldsymbol{u_m}$, or $\boldsymbol{u_\epsilon}$ under MPI, MI or PI, respectively. For the two-stage analysis, the predicted value is the breeding value by the markers which is analogous to $\boldsymbol{u_m}$ in the one-stage model. We acknowledge that the value compare appear to differ, however we emphasise that both are predictions of additive genetics effects. The difference lies in the model used to predict this. Two stage approach in plant breeding commonly uses the marker information alone and our model is a reflection of what is commonly used.

Table 1. The list of genetic models fitted for the selection of genetic effects in one-stage analysis. For MPI, $\boldsymbol{G_g}$ is given as in (5); for MI, the term $\boldsymbol{u_\epsilon}$ in (4) is dropped; and for PI, the term $\boldsymbol{\alpha}$ in (4) is dropped. The corresponding $\boldsymbol{G_g}$ is then $\sigma^2_m \boldsymbol{K} + \sigma^2_e \boldsymbol{I}_m$ and $\sigma^2_\epsilon \boldsymbol{A} + \sigma^2_e \boldsymbol{I}_m$, respectively.

Model	Marker	Pedigree	Residual
MPI	✓	✓	✓
MI	✓	✗	✓
PI	✗	✓	✓

3.2 Accuracy of prediction of additive effects

We compute the reliability of the trait for each genotype as the squared accuracy of each genotype as defined in **?**. Specifically, the accuracy of the i-th genotype is the correlation of the corresponding additive effect to its predicted value and given as

$$r_i = \sqrt{1 - \frac{C^{aa}_{ii}}{G_{a_{ii}}}} \tag{6}$$

where C^{aa}_{ii} is the i-th diagonal element of the prediction error variance matrix for $\boldsymbol{u_m} + \boldsymbol{u_\epsilon}$ under the best one-stage model ($\boldsymbol{u_m} = \boldsymbol{0}$ or $\boldsymbol{u_\epsilon} = \boldsymbol{0}$ if not included in the best model) and $G_{a_{ii}}$ is the i-th diagonal element of the variance matrix of the total additive genetic effects given as $\boldsymbol{G_a} = \sigma^2_\epsilon \boldsymbol{A} + \sigma^2_m \boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^\top$ ($\sigma^2_\epsilon = 0$ or $\sigma^2_m = 0$ if not included in the best model). We call this model-based accuracy to make a distinction to the simulation-based accuracy later. Then the (mean genotype) reliability is given as

$$\bar{r}^2 = \frac{1}{m} \sum_{i=1}^{m} r_i^2. \tag{7}$$

3.3 Two-stage analysis

The two-stage analysis comprises of the analysis of plot data for prediction of genotype means (Stage 1) followed by either an unweighted or weighted analysis of the genotype means (Stage 2).

Stage 1

The plot data for the first stage in the two-stage analysis can be modelled as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{X_g}\boldsymbol{\tau_g} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{8}$$

where $\boldsymbol{X_g}$ is the same as $\boldsymbol{Z}_g$ in (1) except the associated effects $\boldsymbol{\tau_g}$ are treated as fixed effects; $\boldsymbol{\tau}$, $\boldsymbol{X}$, $\boldsymbol{u_p}$ and $\boldsymbol{Z_p}$ are as defined before in (1); and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals that includes the residual genetic effects. We assume that $\boldsymbol{u_p}$ and $\boldsymbol{e}$ are mutually independent and $\boldsymbol{u_p} \sim N(\boldsymbol{0}, \boldsymbol{G_p})$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$. Note the terms in $\boldsymbol{u_p}$ and the variance matrix $\boldsymbol{G_p}$ and $\boldsymbol{R}$ may differ from (1) due to differences in spatial modelling.

The prediction of genotype means, the adjusted genotype means, obtained as a set of E-BLUEs are taken as the vector of observations, which we denote $\boldsymbol{y}^*$ in the next stage.

Stage 2

In the second stage we fit the following model

$$\boldsymbol{y}^* = \boldsymbol{X_0}^*\boldsymbol{\tau_0}^* + \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{e}^* \tag{9}$$

where $\boldsymbol{y}^*$ is the $m \times 1$ vector of predicted genotype means from (8), $\boldsymbol{\tau_0}^*$ is the non-genetic fixed effects with the associated design matrix $\boldsymbol{X_0}^*$; $\boldsymbol{\alpha}$ is the $r \times 1$ random marker effects; $\boldsymbol{M}$ is the $m \times r$ matrix of marker covariates where rows are ordered to correspond to the genotypes in $\boldsymbol{y}^*$; and $\boldsymbol{e}^*$ is the $m \times 1$ vector of residuals.

Stage 2: unweighted analysis

For an unweighted analysis, we assume $\boldsymbol{\alpha}$ and $\boldsymbol{e}^*$ are mutually independent and $\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \sigma^2_m\boldsymbol{D})$ and $\boldsymbol{e}^* \sim N(\boldsymbol{0}, \boldsymbol{R}^*)$. Typically, $\boldsymbol{R}^*$ and $\boldsymbol{D}$ are scaled identity matrices.

Stage 2: weighted analysis

For a weighted analysis we use two methods: one from Smith et al (2001a) and other from Piepho et al (2012) which we will refer to as `diag` and `full` weights, respectively. Specifically, in (9), we fit $\boldsymbol{e}^* = \boldsymbol{\xi} + \boldsymbol{\eta}$, the sum of two components where $\boldsymbol{\xi}$ represents the residual genetic effect and $\boldsymbol{\eta}$ reflects within-trial plot error variation and assume that

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma^2_m\boldsymbol{D} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right) \tag{10}$$

where $\boldsymbol{\Sigma}$ is a known matrix, though in practice is unknown and replaced with the approximation, $\tilde{\boldsymbol{\Sigma}}$, from the data in the first stage. Note that $\text{var}\left(\boldsymbol{y}^* | \boldsymbol{\alpha}, \boldsymbol{\xi}\right) = \boldsymbol{\Sigma}$. The approximation $\tilde{\boldsymbol{\Sigma}}$ is generally based on the covariance of the predicted means which we denote $\tilde{\boldsymbol{C}}$. The standard approach is to use inverse-variance weighting which equates to assuming a diagonal form for $\tilde{\boldsymbol{\Sigma}}$ where the diagonal elements are given as the variance

of the predicted genotype means. An alternative suggestion by Smith et al (2001a) (`diag`) was to assume a diagonal form for $\tilde{\boldsymbol{\Sigma}}$ with diagonal elements given as the reciprocal of the diagonal elements of $\tilde{\boldsymbol{C}}^{-1}$. This approach was motivated as $\tilde{\boldsymbol{\Sigma}}^{-1}$ plays a direct role in the estimation and prediction of fixed and random effects in (9). Piepho et al (2012) uses the approximation $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{C}}$ (`full`) and assumes $\boldsymbol{\xi} = \mathbf{0}$ (we do not adopt the latter assumption). Piepho et al (2012) further uses a transformation for (9) referred to as rotated means which is equivalent to the model in (9) with the aforementioned assumption. These approximations take into account the uncertainty in estimated data $\boldsymbol{y}^*$ by taking into consideration the differing number of replication of the genotype within the trial.

Note that if we follow on from our one stage model then $\boldsymbol{\xi} = \boldsymbol{u}_{\boldsymbol{\epsilon}} + \boldsymbol{u}_{\boldsymbol{e}}$. Subsequently, we can assume $\boldsymbol{\Omega} = \sigma^2_{\epsilon}\boldsymbol{A} + \sigma^2_e\boldsymbol{I}_m$ as before however we will not take this approach, rather we approximate $\boldsymbol{\Omega}$ by $\sigma^2_e\boldsymbol{I}_m$ as commonly done.

4 Analysis of motivating data

All 48 trials were analysed separately using the stage model (4) where we assume $\boldsymbol{D} = \boldsymbol{I}_r$. The baseline model given in (4) included three sets of genetic effects, namely the additive effects associated with the markers, the residual additive effects associated with the pedigree information and the non-additive genetic effects; random peripheral effects comprised the replicate block effects; the fixed effects related to the overall mean; and we chose to fit an AR1×AR1 process for the residuals.

Subsequently extraneous variables, such as random factors based on the column or row indices or fixed covariates to model a linear trend across rows or columns (or its interaction), were fitted. Each fit was assessed by a combination of diagnostic tools to identify the most appropriate model as described in Gilmour et al (1997) and Stefanova et al (2009).

Given the terms for extraneous variation and global trend, the adequacy of the correlation model for local trend is then examined by considering other correlation structures to the residual; namely ID×AR1, AR1×ID, and ID×ID. This was examined using a combination of AIC and graphical diagnostic tools as such as the sample variogram and the coverage intervals.

After the spatial modelling selection, we then select the genetic effects to form the final one-stage model corresponding to the models listed in Table 1. We assumed that all trials have some additive genetic components. The average genotype reliability and the estimate of the error variance for each trial is graphically depicted in Figure 1. There is an approximate linear trend, with moderate variability, observed for the trial mean yield and trial error variance. The non-genetic terms added to the model and

the correlation structures fitted is shown in Table 2 along with the REML estimate of the total genetic variance with its percentage contribution by marker, pedigree-based and non-additive effects.

4.1 An example of trial analysis

We select one trial as an example of our analysis. The diagnostic plots from the fit of the baseline model (Model 1 in Table 3) are shown in Figure 2. The plot of the E-BLUPs of residuals against the row number for each column number in Figure 2(a) shows a similar variation in E-BLUPs of the residuals for any column. The 3D plot of the sample variogram in Figure 2(b) also shows a clear departures from the sample variogram assumed AR1×AR1 variance model, which is clearly visible also in Figure 2(d) - this is indicative of random column effects. The sill of the face in Figure 2(c) is well below the mean from the simulations for majority of the row lags and sits close to the boundary of the coverage interval. This also suggests the presence of random column effects. The full sequence of models fitted to these data is listed in the first column of Table 3.

The addition only of random column effects was deemed sufficient in terms of extraneous variation and global trend. Simplified structures for the variance of the residuals were considered as outlined in Section 3.1.1 - these correspond to the Models 3-5 in Table 3. Based on AIC we have chosen Model 3 (MPI + ID×AR1 + rc) as our best model fit from our spatial modelling.

We then examined the three genetic models, fixing the spatial terms from previous steps and varying the genetic effects as outlined in Table 1. The best selection was based on AIC, which in this case was provided by the MPI model (Model 3).

The diagnostic plots from the fit of Model 3 are shown in Figure 3 and suggest that this model provides an adequate fit to the data. The REML estimates of the variance parameters for Model 3 were $\hat{\sigma}^2_m = 0.1353$, $\hat{\sigma}^2_\epsilon = 0.0244$, $\hat{\sigma}^2_e = 0.0064$, 0.0419 (Block variance), 0.0277 (Column variance), $\hat{\sigma}^2 = 0.0374$ and $\hat{\rho}_r = 0.0101$.

5 Simulation study

For each trial, the best one-stage model is used in a parametric bootstrap to obtain 200 samples of simulated data. There are eight models we fit to this simulated data, namely, the one-stage data-generated model, the one-stage model selected via AIC and the six two-stage models from a factorial combination of 3 weighting schemes (none, `diag`, `full`) and spatial modelling (no, yes). The list of models and their names are provided in Table 4.

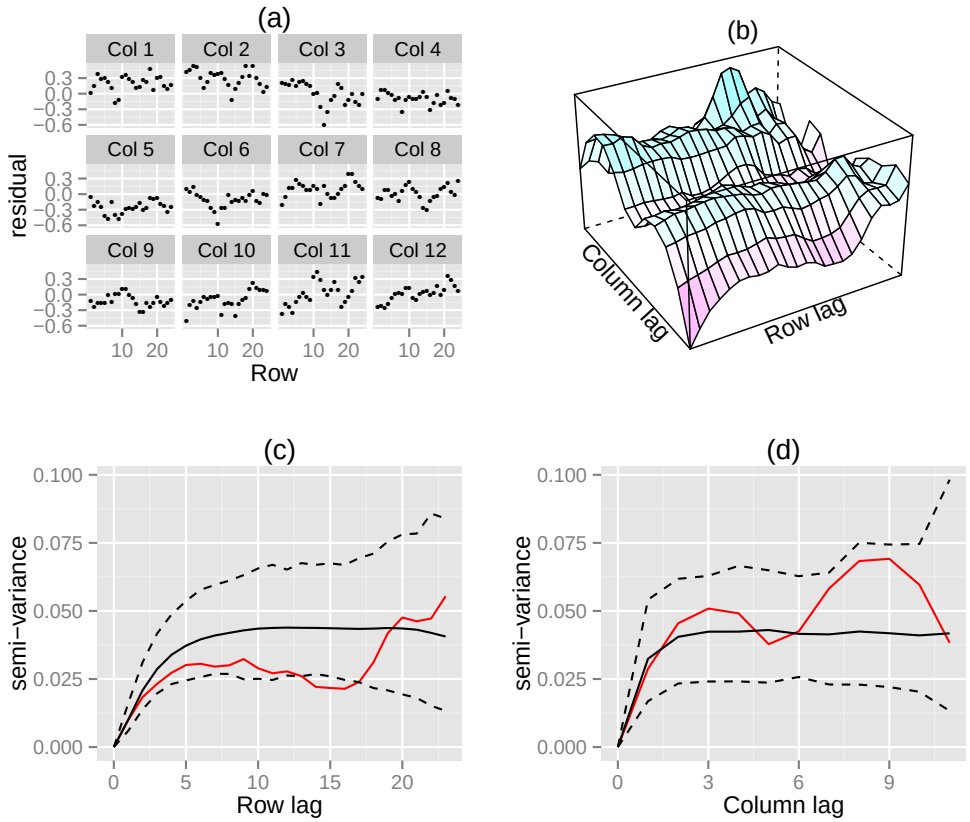

Figure 2. Diagnostic plots from fit of baseline spatial model to the trial data: (a) E-BLUPs of residuals against the row number for each column, (b) 3D plot of the sample variogram, (c) row face of the sample variogram (red line) and mean (black solid line) and approximate 95% coverage intervals (dashed lines), (d) as for (c) but column face.

The measure of the accuracy of each model was computed as follows. For the i-th variety in trial t, we predict the (total) additive effects under each model - specific computation under each model is described in more detail under each model headings below. This was then used to obtain the sample correlation between the true and predicted additive effects of the i-th variety in trial t across the simulations which we denote as $\hat{\rho}_{i,t}$ and term as *simulation-based* accuracy. Note this measure is different to r_i in (6) which we term as the *model-based* accuracy. We then further averaged the simulation-based accuracy across varieties for each trial to obtain the average simulation-based accuracy at trial t:

$$\hat{\bar{\rho}}_t = \frac{1}{m}\sum_{i=1}^{m}\hat{\rho}_{i,t}.$$

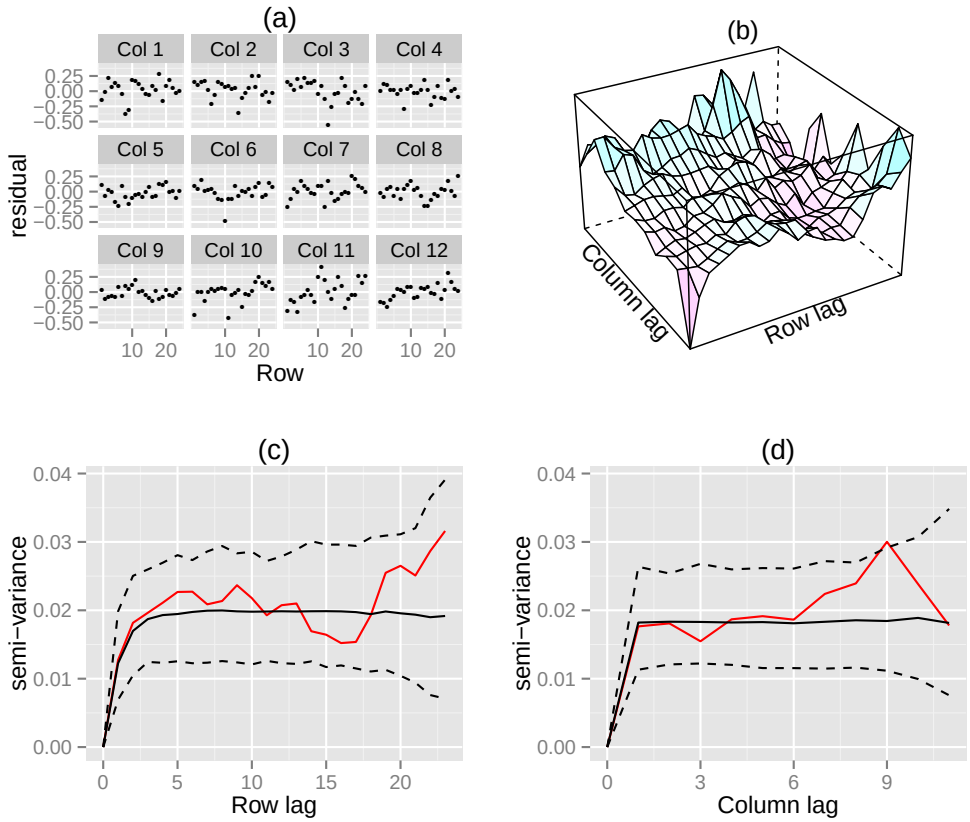

Figure 3. Diagnostic plots from fit of final model to the trial data: (a) E-BLUPs of residuals against the row number for each column, (b) 3D plot of the sample variogram, (c) row face of the sample variogram (red line) and mean (black solid line) and approximate 95% coverage intervals (dashed lines), (d) as for (c) but column face.

We also compute the (average genotype) reliability per trial (**?**), as

$$\hat{r}_t^2 = \frac{1}{m}\sum_{i=1}^{m} \hat{\rho}_{i,t}^2.$$

5.1 One-stage analysis with the best model (1SB)

For each of the 200 simulated data of the 48 trials, we fitted the model that generated the data. We then obtain the GEBV from the prediction of the total additive effects, $\boldsymbol{u_m} + \boldsymbol{u_e}$ (where $\boldsymbol{u_m} = \mathbf{0}$ or $\boldsymbol{u_e} = \mathbf{0}$ if the best model did not include these effects).

5.2 One-stage analysis with model selection (1SM)

Clearly, the prediction in 1SB does not reflect what occurs in reality as we do not know what the data generated model is. To reflect reality, we perform another one-stage analysis which incorporated a model selection procedure to identify the best spatial and genetic model. Specifically, for each generated datasets, we initially fit the baseline model that included random replicated block effects; random genetic effects partitioned into three sets related to additive effects due to markers, residual additive effects and residual genetic effects; fixed overall mean effect; and fixed covariates that model the linear trend across rows, columns and its interaction. These fixed effects were fitted regardless of the generated model as to avoid a model selection step for fixed effects. This is because model comparison for different sets of fixed and random effects remain unclear (Müller et al, 2013) and our inclusion in the first step penalises the one-stage model selection approach by the loss of degrees of freedom in the prediction of random effects (and the estimation of variance components).

A spatial model selection step is incorporated as follows: 16 models consisting of factorial combination of random column effects; random row effects; AR1 correlation structure to row and column (where here non-AR1 structure means assume ID structure) as outlined in Table 5 are fitted and the model with the lowest AIC is automatically selected. Given the spatial model, a model selection for the additive genetic effects is carried out. Specifically, we assumed the presence of some additive effect and three models are fitted, namely MPI, MI and PI models with the same spatial trends selected in the previous step. Note that all three of these models contain the non-additive genetic effects. We again select the model with the lowest AIC and this selected model is used to obtain the GEBV, as in 1SB, by the prediction of the total additive effects.

5.3 Two-stage analysis with no weights (2SN)

For each simulated data, we proceed as described in Section 3.3. Specifically, $\boldsymbol{X}_0 = \mathbf{1}_n$ in Stage 1 model, and $\boldsymbol{u}_p$ include only the block effects that is assumed iid. We either assumed that $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ (2SN-N) or $\boldsymbol{R} = \sigma^2 \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r$ (2SN-S). For the latter, we then incorporate a model selection procedure in the same fashion as 1SM. In the Stage 2 model, we fitted a model to the adjusted genotype means with a fixed overall mean; random marker effects assuming iid and $\boldsymbol{R}^* = \sigma^{*2} \boldsymbol{I}_m$. The sum of the predicted marker effects of a genotype is given as the GEBV.

5.4 Two-stage analysis with weights 1 (2SW & 2SV)

For the weighted analysis, Stage 1 model is the same as 2SN (hyphenated with N or S to indicate no spatial modelling or incorporating spatial model selection, respectively). In Stage 2, we fit a model that partitions the residual into two components as described in Section 3.3 with assumption that $\boldsymbol{\Omega} = \sigma^{*2}\boldsymbol{I}_m$ and $\boldsymbol{\Sigma}$ estimated based on Smith et al (2001a) (`diag`) or Piepho et al (2012) (`full`) denoted 2SW and 2SV, respectively. As in 2SN, the sum of the predicted marker effects of a genotype is given as the GEBV under the aforementioned assumption.

6 Results

Table 6 shows the five number summary of the average relative percentage difference in the simulation-based accuracy per variety across trials for selected comparison of the two methods specified in the first column. For example, we observe that there is a maximum average of 46.6% decrease in a trial in the accuracy of GEBV using 2SN-N compared to 1SB, computed from

$$\max_{t=1,\ldots,48} \frac{100}{m} \sum_{i=1}^{m} \frac{\hat{\rho}_{i,t}^{\text{1SB}} - \hat{\rho}_{i,t}^{\text{2SN}}}{\hat{\rho}_{i,t}^{\text{1SB}}}.$$

The average relative loss in accuracy using a two-stage approach compared to a one-stage approach ranged from negligible to 46.6% with an overall loss of 7.0%. The more realistic one-stage approach (1SM) against a slightly more advantageous two-stage approach (2SV-S) shows that on average the loss is negligible however the loss may be particularly high (e.g. 19.7%) for some sites as reflected in Figure 4. This is also reflected in Figure 4 where almost all sites showed a higher average simulation-based accuracy of GEBV for 1SM compared to 2SV-S.

The one-stage analysis was clearly superior in the accuracy of the GEBV compared to the two-stage analysis with almost all yielding a higher accuracy for all the 9600 simulations (Figure 4). The loss of accuracy is noticeably larger with the lack of spatial modelling in the two-stage analysis, as commonly is the case, than between the different weighting schemes (Figure 4).

The difference between the different approaches within one-stage analysis (1SB vs 1SM) appeared negligible (Table 6). As a formal statistical comparison between methods, we carry out a Wilcoxon signed rank test for the mean difference of the accuracy of GEBV between 1SB and 1SM on the 48 observed mean difference of simulation-based accuracy. This test was carried out using the `wilcox.test` function in the statistical computing software `R` (R Development Core Team, 2008) with parameter `exact=T`, yielded a p-value of 7.60×10^{-12}. For the weighted two-stage

analysis, 2SV-S performed better than 2SW-S as was demonstrated by Piepho et al (2012). In the rest of the results, we concentrate on the comparison between the more realistic one-stage method (1SM) and the better weighted two-stage method (2SV-S).

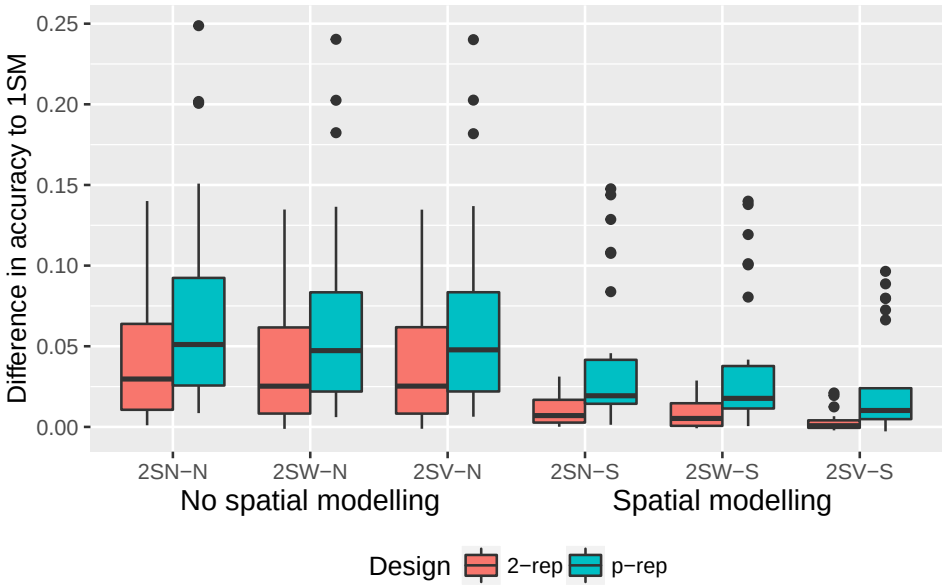

Figure 4. The boxplot of the difference in average accuracy of GEBV (across genotypes) between 1SM and the 2SN, 2SV and 2SW that incorporated (-S) or not incorporated (-N) spatial modelling by experimental design. Positive difference indicates that 1SM is more accurate.

Figure 5 show the plot of simulation-based accuracy of GEBV for each varieties per four selected sites, namely, CS11 showed the most difference in average accuracy over genotype, $\bar{\hat{\rho}}_t$, between 1SM and 2SV-S; LC11 showed a high difference where the value of $\bar{\hat{\rho}}_t$ is high for both 1SM and 2SV-S (note $\bar{\hat{\rho}}_t$ of 1SM and 2SV-S correspond to the average of the values along x- and y-axes of Figure 5.b, respectively); KM10 and HR10 showed very little difference with low and moderate $\bar{\hat{\rho}}_t$ respectively. The latter two sites employ a design with two replicates for each test lines (2-rep) whereas the other two have a partial replicate design (p-rep). Interestingly, as pointed out by the Editor, former two sites have a better fit with the pedigree-based one-stage model while the latter two are fitted with the marker-based one-stage model. This motivates the following comparisons (we omit a factorial comparison as we there is very few sites with only pedigree-based models).

There are 6 and 39 sites with only pedigree-based and only

marker-based models and the mean difference between 1SM and 2SV-S are 0.0182 and 0.013, respectively. While the average accuracies, as one might expect, are higher for trials that employed a 2-rep design than the p-rep design, the differences in the average accuracy between one-stage and two-stage analysis appear to be more pronounced depending on the design. For example, the mean difference between 1SM and 2SV-S for p-rep trials is 0.017 compared to 0.006 for 2-rep trials.

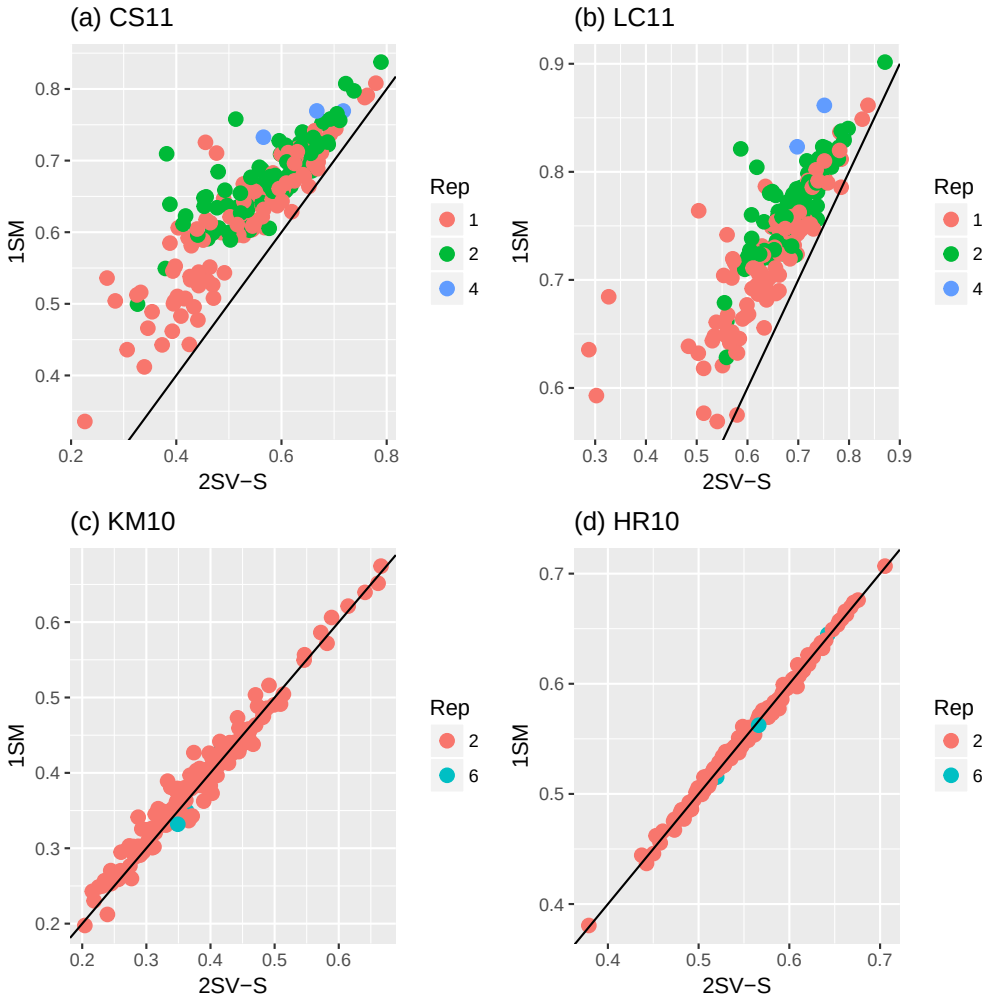

Figure 5. The plots of the simulation-based accuracy of GEBV ($\hat{\rho}_{i,t}$) at four sites: (a) CS11, (b) LC11, (c) KM10, and (d) HR10, for 1SM vs 2SV-S. The colours indicate the number of plots for the genotype in the trial. The solid straight line corresponds to the line $y = x$, where the two methods are equivalent with each other. The points above the diagonal line indicate that the 1SM approach is faring better than the 2SV-S approach.

7 Discussion

In this paper, we have considered simulation based on 48 wheat breeding trials to quantify the loss in GEBV accuracy in using a prevalent two-stage approach to the one-stage approach that takes into account spatial variation or trend. Our trait of interest was yield and the average genotype reliability of yield in these trials ranged from 0.21 to 0.61.

The loss in GEBV accuracy by adopting a two-stage approach which does not adequately accommodate non-genetic sources of variation is substantial (Figure 4). Routine use of such approaches appears common-place in many studies in genomic prediction for plant improvement. Our results indicate, that for a single site, the loss is modest if the two-stage approach incorporates an efficient analysis in the stage 1. This conclusion supports the recommendation by Bernal-Vasquez et al (2014).

We have purposefully restricted our focus to the simple problem of a consideration of a single trial and predictions of GEBV for genotypes within trial. We omitted the consideration of G×E effects as a viable or sensible approach of the two-stage approach in a MET analysis poses another layer of complexity and is beyond the focus of this paper.

Schulz-Streeck et al (2013) raised the question of the implications of the experimental design for genomic selection purposes. From our simulation study, the application of the different experiment designs (2-rep and p-rep) show there may be implications in the difference of accuracy between the one-stage and two-stage methods with differences notably smaller in the 2-rep designs. This is expected as the single site one-stage analysis of an experiment with a randomised complete block design (where all genotypes appear exactly once in each block) would be the same as the two-stage analysis (even without weights), provided that the variance components are known (similar argument is also echoed by Schulz-Streeck et al (2013)).

As Bernal-Vasquez et al (2014) mentions "if feasible, a single-stage approach is preferable to a stage-wise analysis", there is general consensus that one-stage analysis is superior, however a two-stage analysis remains in wide usage due to its computational efficiency. It becomes increasingly important to be aware that the usage of two-stage approaches (or multi-stage analysis for that matter) almost surely results in a loss of accuracy and this loss is particularly pronounced if no spatial modelling is taken into account. We strongly recommend that one-stage approach be used in favour of any multi-stage approach where computational possible. Future research would no doubt benefit greatly from continuing to examine ways to improve computational efficiency such that the use of one-stage analysis with G×E effects pose no computational hurdle.

A Population Structure

The 448 genotypes from the motivating dataset are part of the early generation selection trials. An eigenanalysis of the genomic relationship matrix as described in Patterson et al (2006) revealed that there was no evidence of a population structure. More specifically, we first centre and scale the marker matrix $\boldsymbol{M}$, which we denote as $\boldsymbol{Z}$ with the (i,j)-th entry given as

$$Z_{ij} = \frac{M_{ij} - \mu_j}{\sqrt{p_j(1-p_j)}}$$

where μ_j is the j-th column mean of $\boldsymbol{M}$, $p_j = \frac{1}{2}(\mu_j + 1)$ (note the entries of $\boldsymbol{M}$ are coded as -1, 0 and 1). We then apply a singular value decomposition to the $m \times m$ matrix $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{Z}^\top$. Suppose that λ_i $(1 \leq i \leq m)$ are the eigenvalues of $\boldsymbol{K}$ in descending order. Note that the first principle component accounted 8.7% of the total variation. We form a test statistic

$$T^* = \frac{L - \mu}{\sigma}$$

where

$$\begin{aligned}
L &= \frac{(m-1)\lambda_1}{\sum_i^m \lambda_i}, \\
\mu &= \frac{(\sqrt{r'-1} + \sqrt{m})^2}{r'}, \\
\sigma^2 &= \frac{\sum_i^m \lambda_i}{(m-1)r'}, \text{ and} \\
r' &= \frac{(m+1)\left(\sum_{i=1}^m \lambda_i\right)^2}{(m-1)\sum_{i=1}^m \lambda_i^2 - \left(\sum_{i=1}^m \lambda_i\right)^2}.
\end{aligned}$$

The above scaling (σ) and r', the effective sample size, takes into account of the dependence between columns (i.e. linked markers). The above test statistic, T^* has an approximate Tracy-Widom distribution under standard population genetics assumptions. The p-value was calculated using the R-package `RMTstat` (Johnstone et al, 2014) and yielded 0.066 indicating there is no evidence of population structure.

B Exclusion of genotypes with no phenotypic data

We consider the case where there are m_d genotypes with phenotypic data, but there is pedigree information available on $m_p > m_d$ lines. Without loss of generality, we consider the analysis for a single site and we exclude the

random peripheral (non-genetic) effects so write the linear mixed model for the $n \times 1$ data vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_{\boldsymbol{g}}\boldsymbol{u}_{\boldsymbol{g}} + \boldsymbol{e} \tag{11}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_{\boldsymbol{g}}$ is the $m_p \times 1$ vector of genetic effects with associated $n \times m_p$ design matrix $\boldsymbol{Z}_{\boldsymbol{g}}$ and $\boldsymbol{e}$ is the vector of residuals.

We write the genetic effects as $\boldsymbol{u}_{\boldsymbol{g}} = (\boldsymbol{u}_{\boldsymbol{g}1}^{\top}, \boldsymbol{u}_{\boldsymbol{g}2}^{\top})^{\top}$ where $\boldsymbol{u}_{\boldsymbol{g}1}$ and $\boldsymbol{u}_{\boldsymbol{g}2}$ represent the genetic effects for entries without and with phenotypic data, respectively. The design matrix is therefore given by $\boldsymbol{Z}_{\boldsymbol{g}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{Z}_{\boldsymbol{g}2} \end{bmatrix}$ where $\boldsymbol{0}$ is an $n \times (m_p - m_d)$ matrix of zeros. The genetic variance matrix and its inverse are partitioned conformably as

$$\text{var}\,(\boldsymbol{u}_{\boldsymbol{g}}) = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \text{with} \quad \boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \tag{12}$$

The mixed model equation (MME) for the model in equation (11) are given by

$$\begin{bmatrix} \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{0} & \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Z}_{\boldsymbol{g}2} \\ \boldsymbol{0} & \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{G}^{21} & \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Z}_{\boldsymbol{g}2} + \boldsymbol{G}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_1} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \tag{13}$$

From the second equation in (13) we have that

$$\tilde{\boldsymbol{u}}_{\boldsymbol{g}_1} = -(\boldsymbol{G}^{11})^{-1}\boldsymbol{G}^{12}\tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \tag{14}$$

and substituting this into the third equation in (13) yields the reduced set of MME given by

$$\begin{bmatrix} \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Z}_{\boldsymbol{g}2} \\ \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Z}_{\boldsymbol{g}2} + \boldsymbol{G}_{22}{}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}_{\boldsymbol{g}2}^{\top}\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \tag{15}$$

Therefore, instead of working with the linear mixed model of equation (11), in which the vector of genetic effects, $\boldsymbol{u}_{\boldsymbol{g}}$, is of length m_p and corresponds to all lines in the pedigree, we could use the model commensurate with the MME in equation (15), namely

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_{\boldsymbol{g}2}\boldsymbol{u}_{\boldsymbol{g}2} + \boldsymbol{e} \tag{16}$$

In this model the vector of genetic effects, $\boldsymbol{u}_{\boldsymbol{g}2}$, is of length m_d and corresponds only to those genotypes grown in the trial, that is, those genotypes with phenotypic data.

Then we would obtain the E-BLUPs of the genetic effects for entries with data via a solution of the MME in equation (15) and the genetic effects for entries without data using equation (14). Note that we propose

the form of the model in equation (16) for ease of illustration of the concepts presented in this paper. When the variance matrix $\boldsymbol{G}$ involves the numerator relationship matrix, and when, as is typically the case, the majority of entries with data are non-parental entries, then it is computationally more efficient to use the model as in equation (11) with MME in (13). This is due to the fact that the block of the inverse of the numerator relationship matrix that relates to non-parental entries is diagonal (see Cullis et al, 2014).

C Exclusion of irrelevant genotypes

We consider the case where there are m_d genotypes with phenotypic data, but we are only interested in $m < m_d$ of these genotypes. For example, parental or check varieties may have been grown in the field trial but may not be of interest, or, we may not have marker data for all of the entries grown in the trial. In order to preserve the spatial structure of the trial, we choose not to remove any phenotypic data but instead exclude effects from the genetic model. Without loss of generality, we consider the analysis for a single site and we exclude the random peripheral (non-genetic) effects so write the linear mixed model for the $n \times 1$ data vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{e} \tag{17}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the $m_d \times 1$ vector of genetic effects with associated $n \times m_d$ design matrix $\boldsymbol{Z_g}$ and $\boldsymbol{e}$ is the vector of residuals.

We write the fixed effects as $\boldsymbol{\tau} = (\boldsymbol{\tau_0}^\top, \boldsymbol{\tau_g}^\top)^\top$ where $\boldsymbol{\tau_g}$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to the entries to be excluded and we let $\boldsymbol{X_g}$ denote the associated $n \times (m_d - m)$ design matrix. Thus $\boldsymbol{X} = [\boldsymbol{X_0} \;\; \boldsymbol{X_g}]$ where $\boldsymbol{X_0}$ is the design matrix associated with the (non-genetic) fixed effects $\boldsymbol{\tau_0}$.

In an analogous manner we write the genetic effects as $\boldsymbol{u_g} = (\boldsymbol{u_{g_1}}^\top, \boldsymbol{u_{g_2}}^\top)^\top$ where $\boldsymbol{u_{g_1}}$ is the $(m_d - m) \times 1$ vector of genetic effects corresponding to the entries to be excluded and $\boldsymbol{u_{g_2}}$ is the $m \times 1$ vector of genetic effects of interest. The design matrix is therefore given by $\boldsymbol{Z_g} = \left[\boldsymbol{X_g} \;\; \boldsymbol{Z_{g_2}}\right]$. The genetic variance matrix and its inverse are partitioned conformably as

$$\text{var}\,(\boldsymbol{u_g}) = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \;\text{ with }\; \boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \tag{18}$$

The MME for the model in equation (17) are given by

$$\begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{R}^{-1}\boldsymbol{X_0} & \boldsymbol{X_0}^\top \boldsymbol{R}^{-1}\boldsymbol{Z_{g_2}} & \boldsymbol{X_0}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} & \boldsymbol{X_0}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{R}^{-1}\boldsymbol{X_0} & \boldsymbol{Z_{g_2}}^\top \boldsymbol{R}^{-1}\boldsymbol{Z_{g_2}} + \boldsymbol{G}^{22} & \boldsymbol{Z_{g_2}}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} + \boldsymbol{G}^{21} & \boldsymbol{Z_{g_2}}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} \\ \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_0} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{Z_{g_2}} + \boldsymbol{G}^{12} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} + \boldsymbol{G}^{11} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} \\ \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_0} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{Z_{g_2}} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} & \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\boldsymbol{\tau}}_{\boldsymbol{0}} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_1} \\ \hat{\boldsymbol{\tau}}_{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \quad (19)$$

Absorbing the equation for $\hat{\boldsymbol{\tau}}_{\boldsymbol{g}}$ gives

$$\begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{X_0} & \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{Z_{g_2}} & \boldsymbol{0} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{X_0} & \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{Z_{g_2}} + \boldsymbol{G}^{22} & \boldsymbol{G}^{21} \\ \boldsymbol{0} & \boldsymbol{G}^{12} & \boldsymbol{G}^{11} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}}_{\boldsymbol{0}} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \quad (20)$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X_g}\left(\boldsymbol{X_g}^\top \boldsymbol{R}^{-1}\boldsymbol{X_g}\right)^{-1}\boldsymbol{X_g}^\top \boldsymbol{R}^{-1}$. Thus, in a similar manner to Appendix B, the third equation in (20) gives

$$\tilde{\boldsymbol{u}}_{\boldsymbol{g}_1} = -(\boldsymbol{G}^{11})^{-1}\boldsymbol{G}^{12}\tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \quad (21)$$

and substituting this into the second equation in (19) yields the reduced set of MME, after absorbing $\hat{\boldsymbol{\tau}}_{\boldsymbol{g}}$, given by

$$\begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{X_0} & \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{Z_{g_2}} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{X_0} & \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{Z_{g_2}} + \boldsymbol{G_{22}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}}_{\boldsymbol{0}} \\ \tilde{\boldsymbol{u}}_{\boldsymbol{g}_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_0}^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{Z_{g_2}}^\top \boldsymbol{S}\boldsymbol{y} \end{bmatrix} \quad (22)$$

Therefore, instead of working with the linear mixed model of equation (17), in which the vector of random genetic effects, $\boldsymbol{u_g}$, is of length m_d and corresponds to all genotypes grown in the trial, that is, all genotypes with phenotypic data, we could use the model commensurate with the MME in equation (22), namely

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\tau}_0 + \boldsymbol{Z_{g_2}}\boldsymbol{u_{g_2}} + \boldsymbol{e} \quad (23)$$

In this model the vector of random genetic effects, $\boldsymbol{u_{g_2}}$, is of length m and corresponds only to those entries of interest, for example, those with marker data. Additionally, the model (17) includes fixed effects, $\boldsymbol{\tau_g}$, corresponding to the genotypes to be excluded.

D Software

All models in this paper were fitted using the ASReml-R package (Butler et al, 2009) within the R statistical environment (R Development Core Team, 2008) which uses the average information algorithm (Gilmour et al,

1995) for residual maximum likelihood (REML) estimation for variance parameters. Once the REML estimates of the variance parameters are obtained, a solution of the mixed model equations (MME) is used to provide the empirical best linear unbiased estimates (E-BLUEs) of the fixed effects and empirical best linear unbiased predictions (E-BLUPs) of the random effects (Gilmour et al, 2004).

In general the number of markers is much larger than the number of genotypes with marker data ($r \gg m$). Thus the dimension of the MME coefficient is large. To reduce the computational burden these effects are fitted in the linear mixed model using the high-dimensional approach of Strandén and Garrick (2009).

Acknowledgement

Emi Tanaka, Brian Cullis and Alison Smith gratefully acknowledge the financial support of the Australian Research Council linkage grant. We are grateful to the staff at Australian Grain Technologies, in particular Haydn Kuchel and Adam Norman, for providing access to the grain yield, pedigree and marker data used in this study. Emi thanks Chong You for reviewing the manuscript.

Author's contributions

KM and BRC helped to draft the manuscript. ET wrote the manuscript and ran the simulations. ABS, ET and BRC developed the simulation scheme. ABS wrote the one-stage model and the appendix A and B.

References

Akaike H (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov BN, Csaki F (eds) 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, pp 267–281

Bernal-Vasquez AM, Möhring J, Schmidt M, Schönleben M, Schön CC, Piepho HP (2014) The importance of phenotypic data analysis for genomic prediction-a case study comparing different spatial models in rye. BMC Genomics 15(1):1–17, DOI 10.1186/1471-2164-15-646

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype environment interaction using pedigree and dense molecular markers. Crop Science 52(2):707–719, DOI 10.2135/cropsci2011.06.0299

Butler DG (2014) pedicure: pedigree tools. URL `www.mmontap.org`

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) Mixed models for S language environments ASReml-R reference manual

Butler DG, Smith AB, Cullis BR (2014) On the Design of Field Experiments with Correlated Treatment Effects. Journal of Agricultural, Biological, and Environmental Statistics 19(4):547–557, DOI 10.1007/s13253-014-0191-0

Coombes NE (2002) The Reactive TABU Search for Efficient Correlated Experimental Designs. PhD thesis, Liverpool John Moores University

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G (2014) Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. Crop and Pasture Science 65(4):311–336, DOI 10.1071/CP14007

Crossa J, De Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724, DOI 10.1534/genetics.110.118521

Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic Selection and Prediction in Plant Breeding. Journal of Crop Improvement 25(3):239–261, DOI 10.1080/15427528.2011.558767

Cullis BR, Gleeson AC (1991) Spatial Analysis of Field Experiments-An Extension to Two Dimensions. Biometrics 47(4):1449–1460, DOI 10.2307/2532398

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics 11(4):381–393, DOI 10.1198/108571106X154443

Cullis BR, Jefferson P, Thompson R, Smith AB (2014) Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a Pinus radiata breeding programme. Theoretical and Applied Genetics 127(10):2193–2210, DOI 10.1007/s00122-014-2373-0

de Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1):375–385, DOI 10.1534/genetics.109.101501

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome 4(3):250–255, DOI 10.3835/plantgenome2011.08.0024

Gilmour AR, Thompson R, Cullis BR (1995) Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics 51(4):1440–1450, DOI 10.2307/2533274

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. Journal of Agricultural, Biological, and Environmental Statistics 2(3):269–293, DOI 10.2307/1400446

Gilmour AR, Cullis BR, Welham SJ, Gogel BJ, Thompson R (2004) An efficient computing strategy for prediction in mixed linear models. Computational Statistics and Data Analysis 44(4):571–586, DOI 10.1016/S0167-9473(02)00258-X

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and Applied Genetics 127(2):463–480, DOI 10.1007/s00122-013-2231-5, URL `http://www.ncbi.nlm.nih.gov/pubmed/24264761`

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics 127(3):595–607, DOI 10.1007/s00122-013-2243-1

Johnstone IM, Ma Z, Perry PO, Shahram M (2014) RMTstat: Distributions, Statistics and Tests derived from Random Matrix Theory

Müller S, Scealy JL, Welsh A (2013) Model Selection in Linear Mixed Models. Statistical Science 28(2):135–167, DOI 10.1214/12-STS410, `arXiv:1306.2427v1`

Oakey H, Verbyla AP, Pitchford W, Cullis BR, Kuchel H (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. Theoretical and Applied Genetics 113(5):809–19, DOI 10.1007/s00122-006-0333-z

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS genetics 2(12):e190, DOI 10.1371/journal.pgen.0020190

Piepho HP, Mohring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal 54(6):844–860, DOI 10.1002/bimj.201100219

R Development Core Team (2008) R: A Language and Environment for Statistical Computing

Schulz-Streeck T, Ogutu JO, Piepho HP (2013) Comparisons of single-stage and two-stage approaches to genomic selection. Theoretical and Applied Genetics 126(1):69–82, DOI 10.1007/s00122-012-1960-1

Smith AB, Cullis BR, Gilmour AR (2001a) The analysis of crop variety evaluation data in Australia. Australian \& New Zealand Journal of Statistics 43(2):129–145, DOI 10.1111/1467-842X.00163

Smith AB, Cullis BR, Thompson R (2001b) Analyzing Variety by Environment Mixed Models and Adjustments Data Using Multiplicative for Spatial Field Trend. Biometrics 57(4):1138–1147, DOI 10.1111/j.0006-341X.2001.01138.x

Stefanova KT, Smith AB, Cullis BR (2009) Enhanced diagnostics for the spatial analysis of field trials. Journal of Agricultural, Biological, and Environmental Statistics 14(4):392–410, DOI 10.1198/jabes.2009.07098

Strandén I, Garrick DJ (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of dairy science 92(6):2971–2975, DOI 10.3168/jds.2008-1929

Welham SJ, Gogel BJ, Smith AB, Thompson R, Cullis BR (2010) A comparison of analysis methods for late-stage variety evaluation trials. Australian \& New Zealand Journal of Statistics 52(2):125–149, DOI 10.1111/j.1467-842X.2010.00570.x

Table 2. Columns 2-9 shows the non-genetic parameters included in the final model for individual trial analysis. The ρ_r and ρ_c correspond to the AR1 structure in the residual for row and column direction, respectively; rr and rc are the random row and column effects; lrow, lcol and lrlc correspond to the fixed effects for linear trend across row, column, and their interaction respectively. The tenth column shows REML estimates of the total genetic variance per trial where variances are scaled to reflect their contribution to the total variance (i.e. $\hat{\sigma}^2_m$ and $\hat{\sigma}^2_\epsilon$ are multiplied with the mean of the diagonal elements of $\boldsymbol{K}$ and $\boldsymbol{A}$, respectively). The percentage contribution to the total genetic variance from the total marker additive effects (M), residual or pedigree-based additive effects (P) and non-additive effects (I) are shown in columns 3-5. The last column is the reliability as given in (7). '-' means that the effect was not in the chosen best model and $*$ signifies where the REML estimate of the variance component hit the boundary and was fixed at 0.

Trial	Parameters in the model								Total Genetic	Contribution (%)			$\bar{r}^2$
	ρ_c	ρ_r	rc	rr	block	lrow	lcol	lrlc	Variance	M	P	I	
AN11	-	✓	-	✓	✓	✓	✓	✓	0.125	77.6	-	22.4	0.376
AN13	-	✓	-	-	✓	✓	✓	✓	0.036	80.2	-	19.8	0.327
BL10	✓	✓	✓	-	✓	-	-	-	0.064	100.0	-	0.0*	0.444
BL11	✓	✓	-	-	✓	-	✓	-	0.074	94.7	-	5.3	0.445
BL13	✓	✓	-	-	✓	-	-	-	0.101	100.0	-	0.0*	0.435
BN10	-	-	✓	✓	✓	-	-	-	0.029	77.2	-	22.8	0.358
BN11	✓	✓	✓	-	✓	-	-	-	0.271	87.5	-	12.5	0.472
CD13	-	✓	-	-	✓	-	-	-	0.054	83.1	-	16.9	0.409
CG11	-	✓	✓	-	✓	-	-	-	0.059	83.7	-	16.3	0.423
CH13	✓	✓	-	-	✓	-	-	-	0.075	79.7	-	20.3	0.348
CM10	✓	✓	-	-	✓	-	-	-	0.038	100.0	-	0.0*	0.418
CM11	-	✓	-	-	✓	-	-	-	0.052	96.0	-	4.0	0.441
CM13	-	-	✓	-	✓	-	-	-	0.192	81.8	-	18.2	0.394
CS10	✓	✓	✓	✓	✓	-	-	-	0.138	80.3	-	19.7	0.263
CS11	✓	✓	-	-	✓	-	-	-	0.067	-	100.0	0.0*	0.432
EL11	-	✓	-	✓	✓	-	✓	-	0.021	-	84.9	15.1	0.358
GW13	-	✓	✓	-	✓	-	-	-	0.169	60.2	-	39.8	0.321
HR10	-	-	-	-	✓	-	-	-	0.408	84.7	-	15.3	0.348
HR11	✓	✓	-		✓	-	-	-	0.055	-	34.1	65.9	0.250
HR13	-	✓	-	✓	✓	-	-	-	0.183	52.3	-	47.7	0.271
KM10	✓	✓	-	-	✓	-	✓	-	0.010	36.0	-	64.0	0.213
LC10	✓	✓	-	-	✓	-	-	-	0.138	91.0	-	9.0	0.429
LC11	-	-	✓	-	✓	-	-	-	0.061	-	99.5	0.5	0.608
LC13	✓	✓			✓	-	-	-	0.103	58.5	-	41.5	0.235
LG13	-	✓	✓	✓	✓	-	-	-	0.021	46.7	-	53.3	0.224
MN10	-	✓	✓	-	✓	-	-	-	0.135	89.7	-	10.3	0.429
MN11	-	✓	✓	-	✓	-	-	-	0.189	71.8	24.8	3.4	0.535
MNG13	-	✓	-	-	✓	-	✓	-	0.022	-	47.0	53.0	0.261
MT10	✓	✓		-	✓	-	-	-	0.247	92.5	-	7.5	0.452
MT11	✓	✓		-	✓	-	-	-	0.124	52.3	-	47.7	0.300
NH11	-	✓	✓	-	✓	-	-	-	0.206	97.4	-	2.6	0.488
NH13	-	✓	✓	-	✓	✓	-	-	0.204	77.0	-	23.0	0.393
PN11	-	✓	-	-	✓	-	-	-	0.016	-	70.3	29.7	0.341
PN13	✓	✓	✓	-	✓	-	-	-	0.023	59.0	34.4	6.5	0.477
RS10	✓	-	✓	-	✓	-	-	-	0.166	94.1	-	5.9	0.424
RS11	✓	✓	-	-	✓	-	-	-	0.146	84.8	-	15.2	0.423
RS13	-	✓	-	-	✓	-	-	-	0.358	93.8	-	6.2	0.445
SH11	✓	✓	-	-	✓	✓	-	-	0.054	48.5	-	51.5	0.250
TC10	-	✓	✓	-	✓	-	-	-	0.355	100.0	-	0.0*	0.496
TC13	✓	✓	-	-	✓	-	-	-	0.413	91.7	-	8.3	0.479
WG11	✓	✓	-	✓	✓	-	-	-	0.042	41.0	-	59.0	0.234
WN10	-	-	✓	✓	✓	-	-	-	0.054	80.8	-	19.2	0.373
WN11	✓	✓	✓	-	✓	-	-	-	0.125	97.8	-	2.2	0.420
WT10	-	✓	-	✓	✓	✓	✓	✓	0.271	100.0	-	0.0*	0.469
WT11	✓	✓	✓	-	✓	-	-	-	0.171	46.7	-	53.3	0.287
WT13	✓	✓	✓	-	✓	-	-	-	0.962	55.8	33.6	10.6	0.546
BH13	✓	-	✓	-	✓	-	-	-	0.076	100.0	-	0.0*	0.392
MN13	✓	✓	-	-	✓	-	-	-	0.127	100.0	-	0.0*	0.515

Table 3. Residual log-likelihood and AIC of models fitted to the example trial. Initial model was base-line with the three set of genetic effect (MPI) and AR1×AR1 spatial correlation model (1). Subsequent models involved addition of random column effects (rc) (2) and simplification to the spatial correlation model (3-5). The last 3 models (6-8) are fit with different genetic models as defined in Table 1. The horizontal lines separate between the selection of global trends (of which there was none that can be seen in this example) and extraneous variation, different correlation structure for the local trend, and selection of genetic effects.

Model	Residual log-likelihood	AIC
1. MPI + AR1×AR1	180.35	-346.70
2. MPI + AR1×AR1 + rc	187.39	-358.79
3. MPI + ID×AR1 + rc	187.16	-360.32
4. MPI + AR1×ID + rc	185.16	-356.32
5. MPI + ID×ID + rc	185.13	-358.26
6. PI + ID×AR1 + rc	179.20	-346.39
7. MI + ID×AR1 + rc	185.59	-359.18
8. I + ID×AR1 + rc	153.21	-296.42

Table 4. The list of models compared in the simulation study with the number of stages in the model; whether spatial modelling was included and the weighting scheme used if relevant. *for 1SB, the data generation models matched the data fitted models.

Model name	No. of stages	Spatial modelling	Weights
1SB	1	✓*	N/A
1SM	1	✓	N/A
2SN-N	2	✗	none
2SN-S	2	✓	none
2SW-N	2	✗	`diag`
2SW-S	2	✓	`diag`
2SV-N	2	✗	`full`
2SV-S	2	✓	`full`

Table 5. The sixteen models fitted in the spatial model selection in simulation for 1SM. Note rc and rr represent random factors based on the column and row indices, respectively.

Local trend	Extraneous variation			
	rc only	rr only	rc + rr	none
ID×ID	1	5	9	13
AR1×ID	2	6	10	14
ID×AR1	3	7	11	15
AR1×AR1	4	8	12	16

Table 6. Five number summary of the average relative percentage difference of the simulation-based accuracy per variety across trials for selected comparisons where relative percentage difference reference to the first method written in the first column.

Method comparison	Min	Q1	Median	Q3	Max
1SB vs. 2SN-N	0.3	4.1	7.0	15.8	46.6
1SM vs. 2SV-S	-0.4	0.7	0.8	2.7	19.7
1SB vs. 1SM	-0.2	0.2	0.7	1.3	7.4
2SV-S vs. 2SW-S	0.0	0.3	0.9	2.6	15.2