AASC 2022

Spatial Models for Colocated Trials

Monique Jordan

Supervisors: Prof. Brian Cullis, Dr. Alison Smith, Dr. Daniel Mullan & Dr. Matthew Berryman

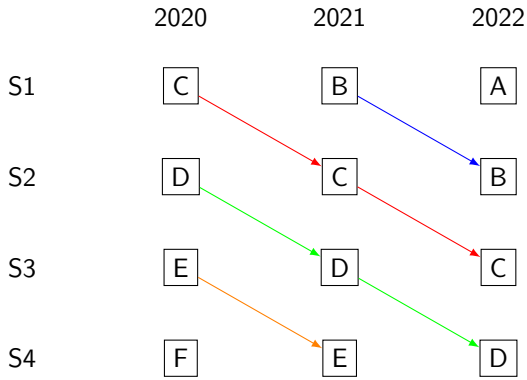
December 1, 2022

A Quick Introduction to Plant Breeding Programs

- The aim of plant breeding programs is to release varieties which are superior for traits of interest such as harvest yield.
- The breeding cycle spans 8-10 years between initial crosses and variety release.
- Lines are evaluated for harvest yield in field trials. Note the terms; lines, varieties and genotypes are used throughtout synonymously.

Introduction to Plant Breeding Programs

- A breeding program involves evaluation in multiple stages such as stage 1 (S1), stage 2 (S2), stage 3 (S3) and stage 4 (S4), with selection decisions made between stages in each year.
- Only the top performing lines are progressed into the next stage. S1: 1000 lines $\rightarrow$ S2: 300 lines $\rightarrow$ S3: 100 lines $\rightarrow$ S4: 60 lines
- There are multiple cycles occuring, with a new cohort of lines entering S1 each year. These cohorts are defined as contemporary groups (CGs).



Introduction

- Trials are conducted across various locations and years in multi-environment trials (METs), where environments are defined as year and location combinations.
- METs are an important tool in plant breeding to measure genotype by environment interaction. As genotypes vary in their response to different environments.
- It is widely accepted that single step factor analytic linear mixed models (FALMMs) are used for the analysis of such MET data.
- The methodology provides numerous advantages over other piecemeal/ad hoc approaches (Smith et al., 2021a).
- Fundamental information for variety selection from an FALMM are the predictions of variety effects for individual environments.

Colocated trials arise as a result of the construction of MET data sets

- Additionally Smith et al. (2021b) highlight the importance of the construction of the appropriate MET data sets for selection decisions.
- They introduced the concept of CGs and data bands.
- MET data sets are formed by combining bands of data to trace the selection histories of lines within CGs, which optimizes the information available for selection decisions.
- Also naturally results in the occurrence of so-called "colocated" trials.

There are two possible scenarios of colocated trials:

- Multiple trials from one stage:
  - Such as early stage trials with large numbers of genotypes which are blocked according to size to ensure that management activities can be completed in a single day or suitable time frame.
- Multiple trials from more than one stage:
  - Such as several S1 trials along side S2 and S3 trials.

An Example InterGrain Hard Wheat Stage 1 & Stage 2 MET 2021

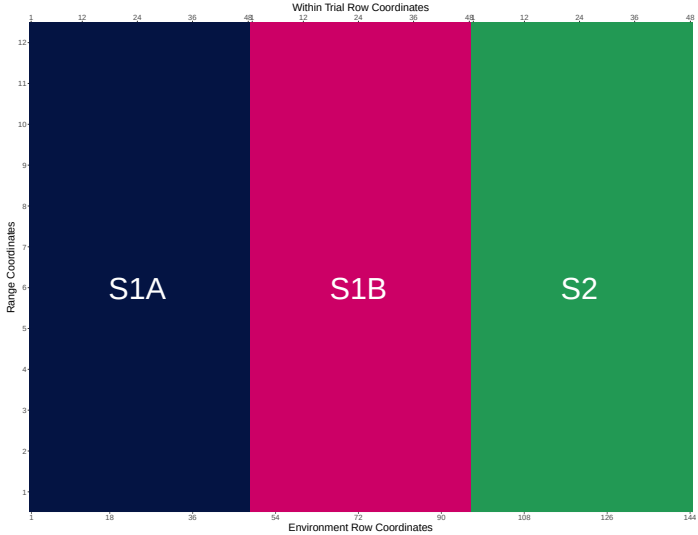
Lets look at an InterGrain MET data set for hard wheat selection decisions for:

- Selection Decision Stage 1:
  - Stage 1 2021 lines progressed to Stage 2 2022.
- Selection Decision Stage 2:
  - Stage 2 2021 lines progressed to Stage 3 2022.

Lets take a deeper dive into the environment Mingenew 2021.

- Three colocated trials: `S1A`, `S1B` and `S2`.
- Two trials `S1A` and `S1B` are Stage 1 trials with a p-rep design.
- One trial `S2` is a Stage 2 trial with nearly all varieties having at least two replicates.
- All have been designed with 12 ranges and 48 rows.
- The two Stage 1 trials `S1A` and `S1B` here are "management blocks".

Mingenew 2021 Stage 1 and Stage 2 Trials



Definition of Colocated Trials

- Conducted in the same environment.
- Very similar agronomic management practices (such as fertiliser and herbicide application).
- Within a certain sowing date window.
- Within a certain harvesting date window.

Models for Colocated Trials

Recall our example $\boldsymbol{y}$ ordered `Row` (coded within trials) within `Range` (coded within trials) within `MBlock`,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_a + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e},$$

where $\boldsymbol{u}_a$ and $\boldsymbol{u}_e$ are additive and non-additive genetic effects so that we have partitioned the (total) genetic effects i.e. $\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e$. Also

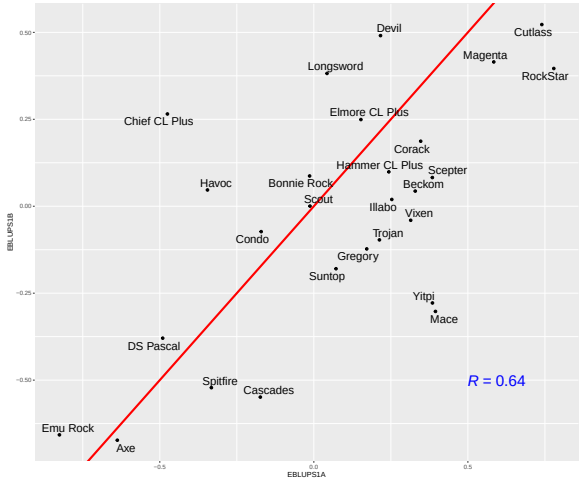
$$\text{var}\left(\begin{bmatrix} \boldsymbol{u}_a \\ \boldsymbol{u}_e \\ \boldsymbol{u}_p \\ \boldsymbol{e} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{G}_a(\gamma_a) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_e(\gamma_e) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}_p(\gamma_p) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}(\phi) \end{bmatrix}.$$

Two branches of considerations in model specification for colocated trials:

- Estimation of non-genetic variance parameters.
- Estimation of genetic variance parameters.

- People often analyse colocated trials as separate trials. This practice cannot produce meaningful results:
  - Results with specifying Variety by Trial interactions within an Environment.
  - Also results with specifying different spatial processes within an Environment and leads to poor precision of the genetic and non-genetic variance parameter estimation.
  - Lets show what the results look like from analysing `S1A` and `S1B` as two trials.
  - These trials have common check varieties so we can compare the genetic predictions.

Predicted Genetic Effects of S1A vs S1B for check varieties in common when analysing each trial separately.



- Analysing each colocated trial separately is like fitting Variety by Replicate in a single trial.
- In a MET the only sensible option for estimating variety by environment interaction is to fit Variety by Environment (V x E) effects, not Variety by Trial within Environment interaction.
- When fitting V x E then the non-genetic model terms and variance parameters MUST be at the environment level to get sensible predictions of the genetic effects at the V x E level.
- Thus for `S1A` and `S1B` we specify $\boldsymbol{G_a} = \gamma_a \boldsymbol{A}$ where γ_a is the additive genetic variance and $\boldsymbol{A}$ is the numerator relationship matrix. Also specify $\boldsymbol{G_e} = \gamma_e \boldsymbol{I}$ where γ_e is the non-additive genetic variance.

Estimation of Non-genetic Variance Parameters

Lets look at Models for S1A and S1B:

Three model specifications were compared with genetic effects specified at the environment level.

1. `equal`: rows and ranges are coded across the single contiguous array (i.e. 12 ranges and 96 rows).

- In the model random row, column effects and the spatial variance matrix are conmmensurate with this layout.
- The residual variance matrix is specified so that given the appropriate ordering

$$\boldsymbol{R}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Sigma}(\phi_c) \otimes \boldsymbol{\Sigma}(\phi_r),$$

  where $\boldsymbol{\phi} = (\sigma^2, \phi_c, \phi_r)^\top$ and $\boldsymbol{\Sigma}(\phi_c)$ and $\boldsymbol{\Sigma}(\phi_r)$ are autoregressive lag one (AR1) correlation matrices of order (12×12) and (96×96).
- Range effects are specified so that $\boldsymbol{u_c}$ is a 12-dimensional vector with variance matrix $\gamma_c \boldsymbol{I}_{12}$.
- Row effects are specified so that $\boldsymbol{u_r}$ is a 96-dimensional vector with variance matrix $\gamma_r \boldsymbol{I}_{96}$.

Estimation of Non-genetic Variance Parameters

Models

1. `equal constrained`: rows and ranges are coded within trials (i.e. 12 ranges and 48 rows).

- In the model a separate set of random effects for both rows and ranges and a separate spatial variance model is specified for each trial.
- BUT the variance parameters are constrained to be equal across trials.
- The residual variance matrix is specified so that

$$\boldsymbol{R}(\boldsymbol{\phi}) = \sigma^2 \oplus_{i=1}^{2} \boldsymbol{\Sigma}_i(\phi_c) \otimes \boldsymbol{\Sigma}_i(\phi_r),$$

  where $\boldsymbol{\phi} = (\sigma^2, \phi_c, \phi_r)^\top$ and $\boldsymbol{\Sigma}_i(\phi_c)$ and $\boldsymbol{\Sigma}_i(\phi_r)$ are AR1 correlation matrices of order (12×12) and (48×48) for $i = 1, 2$.
- Range effects are specified so that $\boldsymbol{u_c}$ is a 24-dimensional vector with variance matrix $\gamma_c \boldsymbol{I}_{24}$.
- Row effects are specified so that $\boldsymbol{u_r}$ is a 96-dimensional vector with variance matrix $\gamma_r \boldsymbol{I}_{96}$.

Estimation of Non-genetic Variance Parameters

Models

3. `unequal`: rows and ranges are also coded within trials (i.e. 12 ranges and 48 rows).

- In the model a separate set of random effects for both rows and ranges and a separate spatial variance model is specified for each trial.
- BUT the variance parameters are allowed to vary across trials.
- The residual variance matrix is specified so that

$$\boldsymbol{R}(\boldsymbol{\phi}) = \oplus_{i=1}^{2} \sigma_i^2 \boldsymbol{\Sigma}_i(\phi_{c_i}) \otimes \boldsymbol{\Sigma}_i(\phi_{r_i}),$$

  where $\boldsymbol{\phi} = (\sigma_1^2, \phi_{c_1}, \phi_{r_1}, \sigma_2^2, \phi_{c_2}, \phi_{r_2})^\top$ also $\boldsymbol{\Sigma}_i(\phi_{c_i})$ and $\boldsymbol{\Sigma}_i(\phi_{r_i})$ are AR1 correlation matrices of order (12×12) and (48×48) for $i = 1, 2$.
- Range effects are specified so that $\boldsymbol{u_c}$ is a 24-dimensional vector with variance matrix $\oplus_{i=1}^{2} \gamma_{c_i} \boldsymbol{I}_{12}$.
- Row effects are specified so that $\boldsymbol{u_r}$ is a 96-dimensional vector with variance matrix $\oplus_{i=1}^{2} \gamma_{r_i} \boldsymbol{I}_{48}$.

Model Specification in ASReml-R

1. equal:

```
as <- asreml(yield~GDrop, random=~vm(GKeep, Aiming) + ide(GKeep) +
Experiment + RecodedRow  + Range,
residual=~ar1(Range):ar1(RecodedRow), data=ming.df,
na.action = na.method(x='include'))
```

2. equal constrained:

```
Vcms1 <- c(1, 1)%x%diag(3)  ## constraint matrix
ascons1 <- asreml(yield~GDrop, random=~vm(GKeep, Aiming) + ide(GKeep) +
Experiment +  Experiment:Row + Experiment:Range,
residual=~dsum(~ar1(Range):ar1(Row)|Experiment), data=ming.df,
na.action = na.method(x='include'), vcm=Vcms1)
```

3. unequal:

```
asuneq <- asreml(yield~GDrop, random=~vm(GKeep, Aiming) + ide(GKeep) +
 Experiment + at(Experiment):Row + at(Experiment):Range,
residual=~dsum(~ar1(Range):ar1(Row)|Experiment), data=ming.df,
na.action = na.method(x='include'))
```

An In Silico Experiment

Aim

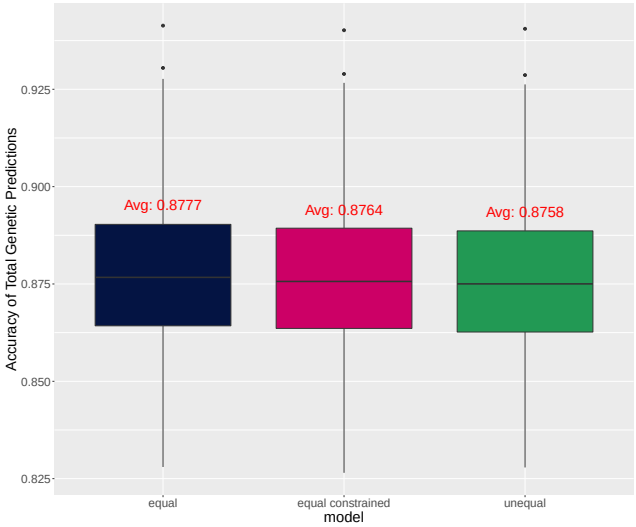
To compare the several spatial models in realistic scenarios.

- Often extraneous variation is seen in the form of range and row effects i.e. from serpentine harvesting and management practices (Gilmour et al., 1997). We assess the methods for three different row and column variances respectively.
- Simulate data from the model which biologically and statistically is most sensible.
- Use correlation parameters for range and row using knowledge from our experience of analysing early stage trials.
- Use additive and non-additive genetic variance parameters so that the generalised heritability for single replicate individuals is approximately 0.838. Also so that the percentage of additive genetic variance to total genetic variance is approximately 73.1%.

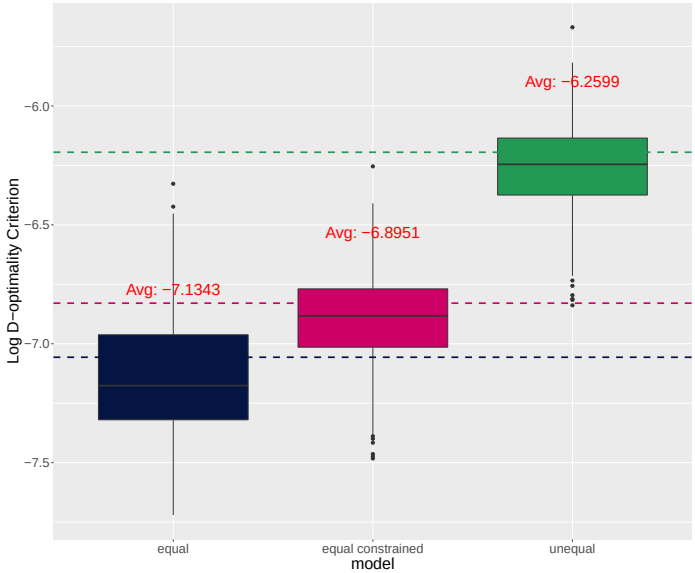
An In Silico Experiment

- Simulate data with different scenarios and subsequently analyse said data with the three models `equal`, `equal constrained` and `unequal`.
- Compare the accuracies of the total genetic BLUPs.
- Compare the log D-optimality criterion of the three models, which is the determinant of the variance matrix (inverse expected information matrix) of the estimated variance parameters. It is a measure of the generalised variance (i.e. the quality) of the estimated variance parameters and thus should be minimised.

Accuracy of Predictions



Quality of Variance Parameter Estimation - log D-optimality criterion



Results of the Simulation Study for Non-Genetic Effects

- `equal` performs best with the highest prediction accuracy of the total genetic effects and with the lowest log D-optimality criterion for the variance parameters.
- `equal constrained` performs second in front of `unequal` for both prediction accuracy and reliability of variance parameter estimation.

Estimation of Genetic Variance Parameters

The Effects of Selection on Genetic Variance Parameter Estimation:

Thompson (1979 & 2008) describe the effect of selection on genetic variance parameter estimation in the context of animal breeding. We describe this in the context of plant breeding.
We consider two trials conducted across two years, the first which evaluates S1 lines in year 1 (labelled S1Y1). The second in year 2 evaluates the subsequent top performing lines from S1Y1 in S2 (labelled S2Y2). Two methods of analyses we consider are:

1. A univariate analyses of S2Y2 and S1Y1 data alone, where we do not account for the selection histories of lines in variance parameter estimation.
2. A bivariate analyses of both S2Y2 and S1Y1 data, where we account for the selection histories of lines in variance parameter estimation.

The univariate analyses will lead to biased inferences since the data of S2Y2 can be regarded as conditional on the yield data of S1Y1. So we need to use a multivariate approach.

Model Specification for Colocated trials in Summary.

For the quality of variance parameter estimation:

- A multivariate approach should be used with the inclusion of data pertaining to the selection histories of lines in the form of data bands.
  - Analysing colocated trials corresponding to different stages strengthens genetic predictions by the addition of indirect information.
  - It is reasonable to assume a common genetic variance between stages since lines are usually derived from a common breeding population.
- Genetic effects should be specified at the environment level with a single genetic variance. In the context of METs this is the specification of V x E effects to estimate variety by environment interaction.
- Non-genetic effects should also be specified at the environment level, for appropriate shrinkage of the genetic BLUPs.

References I

- Smith, A., Norman, A., Kuchel, H., & Cullis, B. (2021a). Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. Frontiers in plant science, 12.
- Smith, A., Ganesalingam, A., Lisle, C., Kadkol, G., Hobson, K., & Cullis, B. (2021b). Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs. Frontiers in Plant Science, 11, 623586.
- Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. In Journal of Agricultural, Biological, and Environmental Statistics (Vol. 2).
- Thompson, R. (1979). Sire evaluation. Biometrics, 339-353.
- Thompson, R. (2008). Estimation of quantitative genetic parameters. Proceedings of the Royal Society B: Biological Sciences, 275(1635), 679-686.

Questions?