APBC

Spatial Models For Colocated Trials

Monique Jordan

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

June 23, 2022

Outline

Introduction

- Field experiments are conducted to evaluate the performance of varieties in plant breeding programs.
- Major objective is to produce varieties which are superior for traits of interest such as harvest yield.
- We promote the use of best linear unbiased predictions (BLUPs) of the additive or total genetic effects for selection. Ideally obtained from the analysis of a multi-environment trial (MET).
- Correlation between the BLUPS and the true genetic effects is the accuracy.
- Squared correlation between the BLUPS and the true genetic effects is reliability, also may be defined as the heritability (Oakey et al., 2006).
- Maximising accuracy is crucial to make the best selection decisions to increase genetic gains.

Motivation

- Early stage selection experiments.
  - Large trials with many genotypes.
  - Partitioned into smaller management blocks call these `MBlock`.
- Assume for simplicity that all management blocks are from the same stage of selection (e.g. Stage 1).

Definition of Colocated Trials

- Conducted in the same environment, where environment is the combination of year and location.
- Very similar management practices.
- Within a certain sowing date window.
- Within a certain harvesting date window.

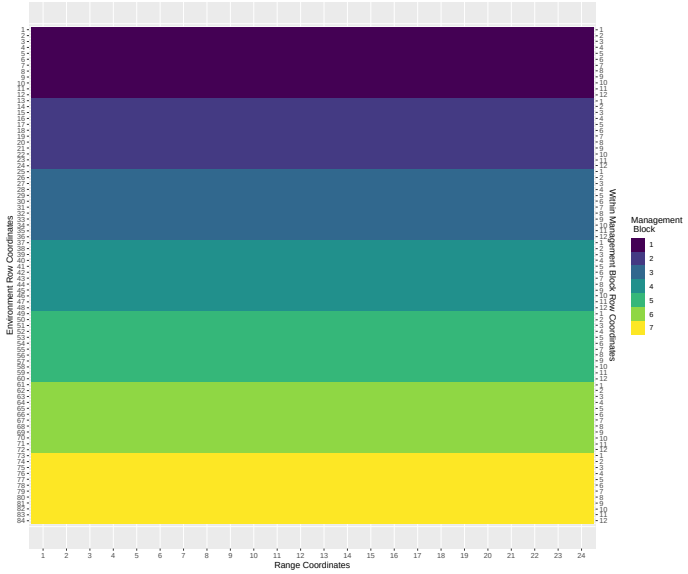
A Typical Stage 1 Experiment

Typically breeding programs evaluate lines in several stages, usually there are 3-4 stages. Our focus here is experiments conducted in stage 0 or 1, where there are large numbers of genotypes and minimal replication.

- 1440 genotypes with yield data.
- Pedigree information available for 4514 genotypes.
- Marker information available for 1438 genotypes.

- 24 ranges/columns.
- 84 rows.
- 2016 plots.
- 7 management blocks.
- 12 rows within each management block.
- Minimal replication
  - Test lines are only allocated to one management block (no connectivity) with one or two replicates (p-rep design (Cullis et al., 2006)).
  - Only check lines are replicated in each management block.

An Example with Seven Colocated Management Blocks



There is a gap in the literature which details appropriate statistical models for colocated trials. Our aim is to investigate several models and recommend which ones are most appropriate.

So what models should we use to analyse these trials?

Some considerations:

- Genotype is usually fit at the environment level.
- Each management block is often designed as a single trial.
- We should expect variance homogeneity within environments.
- Generalised heritabilities are calculated for environments.

What has been done before?

- Management blocks treated as single trials, assuming variance heterogeneity within environments. Not sensible as the shrinkage factors of BLUPs will be incorrect. Also there is a loss of information of the distance between plots in different management blocks.

The Data Frame

Often the columns Range and Row are only given in data sets and after obtaining knowledge of the configuration of management blocks RangeRecoded and RowRecoded can be coded over the entire environment.

yield	RangeRecoded	Range	RowRecoded	Row	MBlock	...
.	1	1	1	1	1	...
.	1	1	2	2	1	...
.	1	1	3	3	1	...
⋮	⋮	⋮	⋮	⋮	⋮	...
.	24	24	11	11	1	...
.	24	24	12	12	1	...
⋮	⋮	⋮	⋮	⋮	⋮	...
.	1	1	73	1	7	...
.	1	1	74	2	7	...
⋮	⋮	⋮	⋮	⋮	⋮	...
.	24	24	82	10	7	...
.	24	24	83	11	7	...
.	24	24	84	12	7	...

Models for Colocated Trials

Consider three models

We fit genetic effects at the environment level where

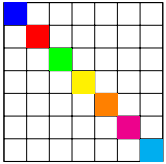
$$\text{yield} = \text{mean} + \text{genetic effect} + \text{non-genetic effects} + \text{errors}.$$

We consider the models:

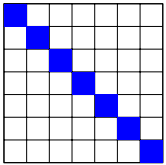
		Variance Across `MBlock`	
Model	Genetic	Non-Genetic	Errors
unequal (`ue`)	equal	unequal	unequal
equal constrained (`ec`)	equal	equal	equal constrained
equal (`e`)	equal	equal	equal

Variance Models for the errors!

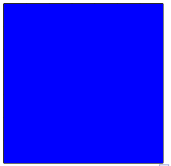
1. Unequal - ue



2. Equal Constrained - ec

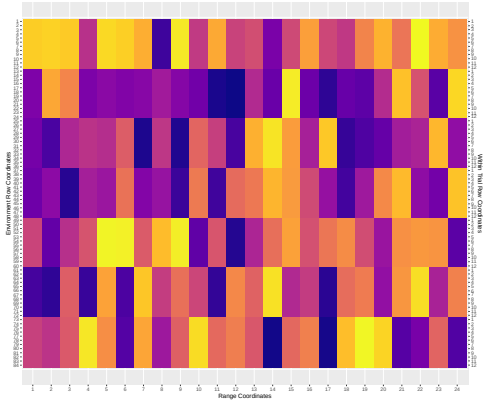


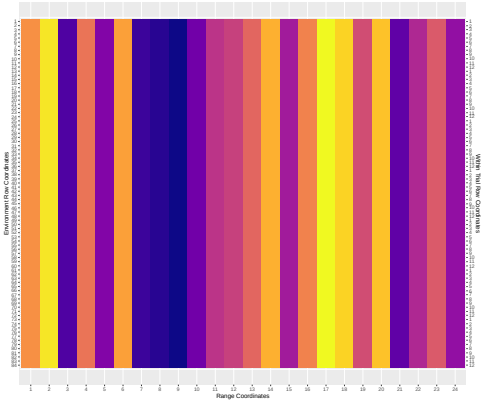
3. Equal - e



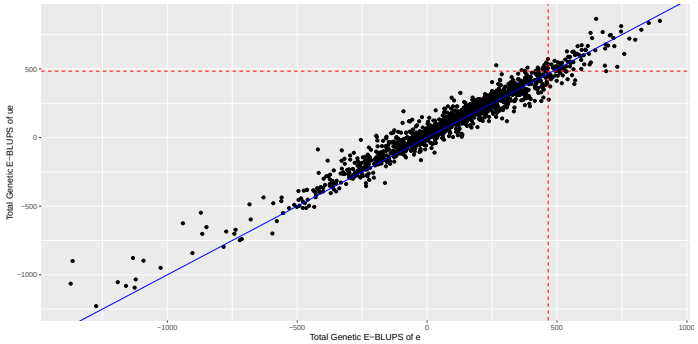
Range Effects

- Consider Range within environment.
- Consider Range within `MBlock`.





Comparing the Total Genetic E-BLUPs of e and ue from the example.



Selection Error if we selected top 100 lines?

- From the BLUPS of either models we would select 79% of the same lines.
- Discrepancy of 21% of the lines selected using the BLUPs of either models.

Comparing the Reliability of Genetic E-BLUPs of e and ue from the example by MBlock.

		Reliability	
Model	MBlock	Two Replicate Test Lines	Single Replicate Test Lines
ue	1	0.604	0.575
ue	2	0.603	0.563
ue	3	0.616	0.584
ue	4	0.663	0.640
ue	5	0.646	0.612
ue	6	0.638	0.598
ue	7	0.641	0.608
		Mean (Range)	Mean (Range)
e	all	0.637 (0.0114)	0.601(0.00410)

An In Silico Experiment

Aim

To compare the several spatial models in realistic scenarios using both pedigree and marker information.

- 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 variance models 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 genomic based generalised heritability and pedigree based generalised heritability for single replicate individuals are both approximately 0.89. Also so that the percentage of additive genetic variance to total genetic variance is approximately 81% for both the pedigree and marker models.

An In Silico Experiment

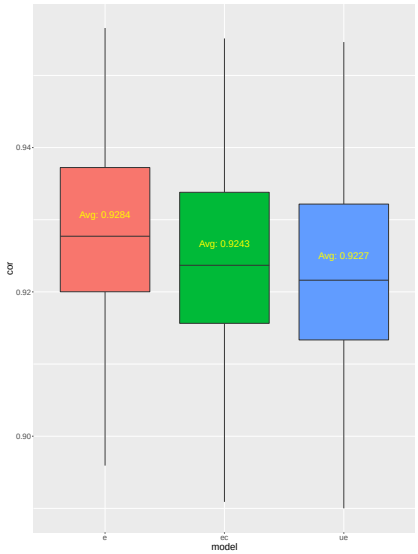
- Simulate data from the single model with different scenarios and subsequently analyse said data with the three analysis methods.
- Compare the correlations of total genetic BLUPs. The spatial process is a 'nuisance parameter', we are most interested in the genetic effects.
- Compare the log D-optimality 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. (See Chris Lisle's talk)

Variance Parameters for Data Simulation

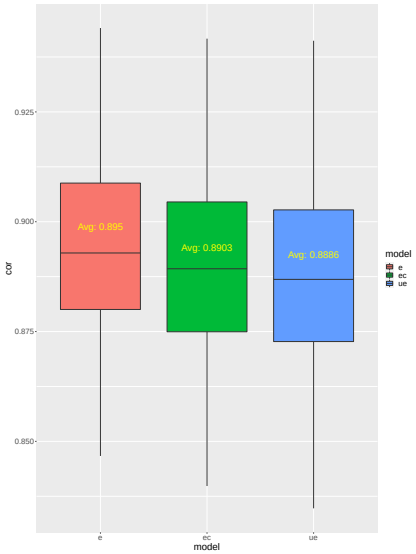
Genetic Effects	Model Terms								
	Overall Scale	Random			Spatial		Genetic		
							Additive		Non-additive
	σ^2	Trial γ_t	Column γ_c	Row γ_r	Column AR ρ_c	Row AR ρ_r	γ_γ	γ_π	γ_e
K + I	1	0.0779	0	0	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	0	0.1	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	0.1	0	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	0.1	0.1	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	0.1	1	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	1	0.1	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	1	0	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	0	1	0.171	0.646	2.197	-	0.531
K + I	1	0.0779	1	1	0.171	0.646	2.197	-	0.531
A + I	1	0.0779	0	0	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	0	0.1	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	0.1	0	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	0.1	0.1	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	0.1	1	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	1	0.1	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	1	0	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	0	1	0.171	0.646	-	1.388	0.659
A + I	1	0.0779	1	1	0.171	0.646	-	1.388	0.659

Results: Correlation of genetic E-BLUPS, $\gamma_c = 1$ and $\gamma_r = 0.1$

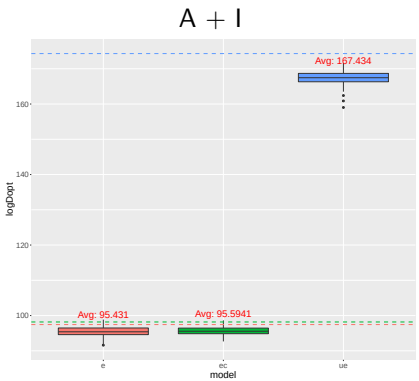
A + I

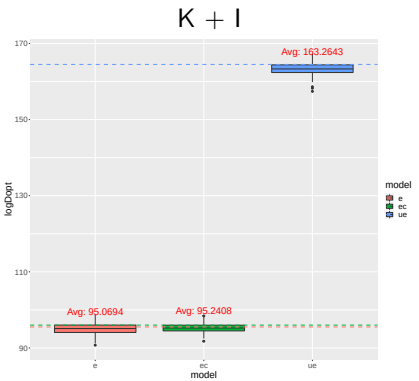


K + I



Results: Log D-optimality Criterion, $\gamma_c = 1$ and $\gamma_r = 0.1$





Conclusions

- For prediction accuracy of genetic effects over all possible scenarios the e model performed best followed by the ec model, last was the ue model.
- For the log D-optimality criterion overall possible scenarios the rankings are the same with the e model performing best. Thus e has the most stability in terms of variance parameter estimation.
- Although the ec model would be the most appropriate if information was unknown about the spatial separation of plots in different trials. There is not a great loss in prediction accuracy or quality of variance parameters estimates like the ue model.
- The timings of e in comparison to ue and ec performed a bit slower for the K+I models but little difference was found for the A+I models.
- Timings are interesting to our group as the ec models are often used by CBADS-SPI using ASReml-R.

References I

- Oakey, H., Verbyla, A. P., Cullis, B. R., Pitchford, W. S., & Kuchel, H. (2006). Joint modelling of additive and non-additive genetic line effects in single field trials. Theoretical and Applied Genetics, 113, 809–819.
- Cullis, B. R., Smith, A. B., & Coombes, N. (2006). On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics, 11(4), 381–393. https://doi.org/10.1198/108571106X154443
- 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).
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). ASReml-R Reference Manual Version 4.

Questions?