Design of a CIMMYT Australian ICARDA Germplasm Evaluation (CAIGE) experiment using an Incomplete multi-environment trial (IMET) design approach incorporating genetic relatedness

Lu Wang[1]

Brian Cullis,[1] Beverley Gogel,[1]
Christopher Lisle[1] and Julie Nicol[2]

Eucarpia 2022

[1]Centre for Biometrics and Data Science for Sustainable Primary Industries (CBADS-SPI)
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
luw@uow.edu.au

[2]Plant Breeding Institute
Faculty of Science
University of Sydney
julie.nicol@sydney.edu.au

Outline

Introduction

Introduction

The analysis model: FALMM

- A key objective of Australian plant breeding programs is to increase genetic gain by selecting superior individuals in the analysis of multi-environment trial (MET) data.
- The factor analytic (FA) linear mixed model (LMM) has been widely used as a superior method for modelling the genotype by environment (GE) effects in MET data-sets (Smith and Cullis, 2018).
- Most implementations of FALMMs in plant breeding programs incorporate genetic relatedness either through a Numerator Relationship Matrix (NRM) or a Genomic Relationship Matrix (GRM).
- Factor analytic selection tools (FAST) and more recently iClasses (Smith et al., 2021) use the outputs from the fit of an FALMM to produce meaningful summaries for selection.

Introduction

Towards model-based design

- Maximal gains from the use of FALMMs require appropriate design and construction of the MET data-set.
- There has been very little attention given to the design of MET data-sets, even less utilising genetic relatedness.
- Classical approaches to design are **incapable** of constructing optimal designs which
  - include genetic relatedness;
  - take into account seed supply issues and resource allocation constraints which commonly present in early stage trials;
  - provide optimal allocation of genotypes across environments.

Introduction

Motivating example 1 - S1 Desi Chickpea MET design

The design of 2022 S1 Desi Chickpea experiment comprised 7 sites (trials), with two home sites both at Narrabri, on different soil types, one each for northern and southern adapted genotypes, respectively.

- The aim is selection - promotion of lines from stage 1 to stage 2.
- All regional adapted genotypes must be present at least once at their corresponding home-site.
- There were two northern and three southern satellite sites.
- Satellite sites typically include subsets of genotypes due to seed supply and land availability.
- There was a total of 4240 genotypes, including test and check varieties.

Motivating example 2 - CAIGE MET design

In 2022, the design of CIMMYT Australian ICARDA Germplasm Evaluation (CAIGE) for bread wheat yield comprised 12 environments (trials), with a single environment being the home site.

- 337 unique lines (genotypes) including 14 Australian check lines.
- All genotypes must be present at least once at the home site.
- Checks should each be allocated to at least 2 plots in each environment.
- Different trial sowing rates must be accommodated.
- Seed limitation for a number of test lines.
- Budgetary constraints - trial size.

Introduction

Model-based approach to design

- Model-based designs can provide the framework for generating designs with the required properties.
- This approach generates an optimal design under a pre-specified (analysis) model and a design criterion.
- **odw** (Butler, 2022) package is freely available in R (R Core Team, 2020) & constructs optimal designs under the LMM framework & can adapt to a wide range of scenarios:
  - classical designs such as latinised row-column designs;
  - single site ***p*-rep designs** (Cullis et al., 2006) with or without genetic relatedness;
  - **incomplete MET (IMET) designs** with genetic relatedness.

Model-based designs using the odw design software

Model-based designs in odw

Design model in odw

Cullis et al. (2022) rewrite the LMM in terms of sets of effects that they refer to as the **permute** set and the **static** set.

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \\
&= \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e} \\
&= \boldsymbol{W}_1\boldsymbol{\beta}_1 + \boldsymbol{W}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \textbf{permute}\ \text{set} + \textbf{static}\ \text{set} + \ \text{errors}
\end{aligned}
$$

- $\boldsymbol{y}$ is the $n \times 1$ vector of observations.
- $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$ (assumed to have full column rank).
- $\boldsymbol{u}$ is a vector of random effects with associated design matrix $\boldsymbol{Z}$.
- $\boldsymbol{e}$ is the vector of residuals.

Model-based designs in odw

The design function

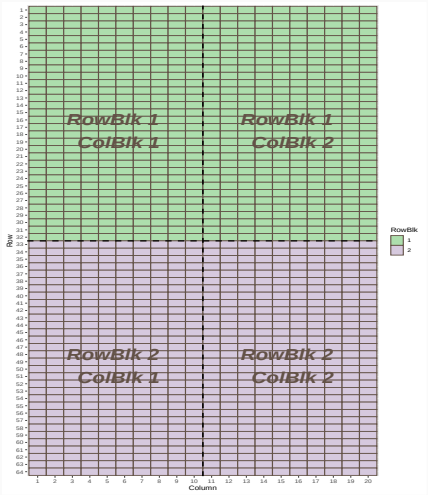
$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e} \\
&= \boldsymbol{W}_1\boldsymbol{\beta}_1 + \boldsymbol{W}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \textbf{permute}\ \text{set} + \textbf{static}\ \text{set} + \ \text{errors}
\end{aligned}
$$

- The **permute** set consist of effects associated with the design search.
- The **static** set consist of effects associated with the plot structure of the experiment, including covariates if any.
- The **odw** package adopts $\mathcal{A}$-optimality which seeks to minimise the average pair-wise error variance of all elementary treatment contrasts.
- The permutation algorithm operates only on the rows of $\boldsymbol{W}_1$.
- Two rows of $\boldsymbol{W}_1$ are interchanged, subject to the interchange being permissible.
- The rows of $\boldsymbol{W}_2$ are considered invariant (static).

Using genetic relatedness in odw

Single site *p*-rep design in odw

S1 Desi Chickpea experiment - south home site



- 1139 genotypes and 1280 plots, only sensible approach is to use a partially replicated (*p*-rep) design.
- To assist trial management, the south home site was split into two management blocks, here we refer to as row blocks (RowBlk).
- Row blocks are 32 rows by 20 columns each.
- Column blocks (ColBlk, 2) are large gradient/extraneous variation blocks of size 64 rows by 10 columns each.
- The mean genetic relatedness across the 1139 genotypes ranged from a minimum of 0.078 to a maximum of 0.263 and a mean of 0.204.

Single site *p*-rep design in odw

S1 Desi Chickpea experiment - south home site

Design construction for a single site *p*-rep design generally involves two stages:

- Stage One - allocation of packet [1] choice (pC) to genotypes.
- Stage Two - allocation of plots to genotypes given packet choice status.
  - Step 2.1 - allocation of plots to genotypes to ensure that the design will be resolvable with respect to replicated test lines (and checks) having only one plot in each RowBlk and each ColBlk.
  - Step 2.2 - finds a design that is optimal with respect to rows (Row) and columns (Column) of the experiment while maintaining the two-way blocking achieved in Step 2.1.
- Each step uses a different call to **odw**.

[1] packet refers to plots in the experiment

Single site p-rep design in odw

Each call to odw requires:

- An R data frame with initial configuration.
- A linear mixed model that sets up the design model.
- A **permute** [2] factor - generally is the Treatment factor.
- A set of **static** factors - generally are block factors.
- A design quality measure, the $\mathcal{A}$-criterion.
- A swap factor that determines legal interchanges during the design search, can be *NULL*.

[2] a *permute* factor can be a set of *objective* and *linked* factors in some cases. When the *linked* factor is *NULL*, the *permute* factor is the *objective* factor.

Single site p-rep design in odw

The search process:

1. Initialise the iteration number $N = 1$, and calculate $\mathcal{A}$ for the initial design - set as the current design.
2. Undertake a legal interchange of the **permute** factors between any two plots, subject to the interchange being a legal swap.
3. Calculate the $\mathcal{A}$ for the new design obtained from this interchange.
4. Accept the new design as the current design if the $\mathcal{A}$ of the new design is less than the $\mathcal{A}$ of the current design
5. $N = N + 1$.
6. If $N < N_{max}$ return to 2, else terminate the search. N_{max} is set by the user in the call to **odw**.

Stage one: packet choice to genotypes

The linear mixed model

In the case of pedigree information, the total GE effects ($\boldsymbol{u_g}$) are partitioned into additive ($\boldsymbol{u_a}$) and non-additive (residual GE, $\boldsymbol{u_e}$) effects (Smith and Cullis, 2018).

At Stage one the LMM is for $\bar{\boldsymbol{y}}$, the "pseudo" data vector of 1139 genotype "means", is:

$$\begin{aligned} \bar{\boldsymbol{y}} &= \boldsymbol{\mu} + \boldsymbol{u_g} + \boldsymbol{e} \\ &= \boldsymbol{\mu} + \boldsymbol{u_a} + \boldsymbol{u_e} + \boldsymbol{e} \end{aligned}$$

- The total genetic effects are $\boldsymbol{u_a} + \boldsymbol{u_e}$.
- The total genetic variance is then

$$\text{var}(\boldsymbol{u_a}) + \text{var}(\boldsymbol{u_e}) = \sigma_a^2 \boldsymbol{A} + \sigma_e^2 \boldsymbol{I}_{1139}$$

where $\boldsymbol{A}$ is the NRM, σ_a^2 and σ_e^2 are the additive and non-additive variance parameters, respectively.

Stage one: packet choice to genotypes

The linear mixed model

- The error variance for an observation (that is a genotype mean) depends on the packet choice (No. plots) and is given by:

$$\text{var}(\boldsymbol{e}) = \begin{cases} \sigma^2 & : \text{ pC1 - one packet} \\ \sigma^2/2 & : \text{ pC2 - two packets} \end{cases}$$

- To save time for **odw**, $\boldsymbol{u_e}$ and $\boldsymbol{e}$ are combined into one term with the variance parameters given by:

$$\text{var}(\boldsymbol{u_e}) + \text{var}(\boldsymbol{e}) = \begin{cases} \sigma_e^2 + \sigma^2 : \text{pC1 - one packet} \\ \sigma_e^2 + \sigma^2/2 : \text{pC2 - two packets} \end{cases}$$

Stage one: packet choice to genotypes

Initial data frame

The initial data frame has 1139 rows and contains the following key fields:

- Genotype is a factor with 1139 levels - to be used as the **permute** factor.
  - The check variety, CBA CAPTAIN, must have two packets.
  - 677 test lines only have enough seed for one packet.
  - 461 test lines have enough seed for two packets.
- swp is a factor with 3 levels to be used as the swap factor in **odw**.
- pC is a factor with 2 levels to set up the variance model, i.e. the factor for packet choice. The two-way contingency table between pC and swp is:

swp	pC1	pC2
capt	0	1
one	677	0
two	321	140

Stage two: plots to genotypes given pC status

Step 2.1: allocation of plots to genotypes - RowBlk & ColBlk

- The LMM:

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e} \\
&= \boldsymbol{W}_1\boldsymbol{\beta}_1 + \boldsymbol{W}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= [1 \;\; \boldsymbol{Z}_g][\mu \;\; \boldsymbol{u}_g]^{\mathsf{T}} + [\boldsymbol{Z}_{rb} \;\; \boldsymbol{Z}_{cb}][\boldsymbol{u}_{rb} \;\; \boldsymbol{u}_{cb}]^{\mathsf{T}} + \boldsymbol{e}
\end{aligned}
$$

- The data frame now has 1280 rows.
- The **permute** factor is Genotype.
- The **static** factors are RowBlk and ColBlk.
- The total genetic effects are used for calculation of the $\mathcal{A}$- criterion.

Stage two: plots to genotypes given pC status

Step 2.2: allocation of plots to genotypes

- The LMM:

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{W}\beta + \boldsymbol{e} \\
&= \boldsymbol{W}_1\beta_1 + \boldsymbol{W}_2\beta_2 + \boldsymbol{e} \\
&= [1 \;\; \boldsymbol{Z}_g][\mu \;\; \boldsymbol{u}_g]^{\mathsf{T}} + [\boldsymbol{Z}_{rb} \;\; \boldsymbol{Z}_{cb} \;\; \boldsymbol{Z}_r \;\; \boldsymbol{Z}_c][\boldsymbol{u}_{rb} \;\; \boldsymbol{u}_{cb} \;\; \boldsymbol{u}_r \;\; \boldsymbol{u}_c]^{\mathsf{T}} + \boldsymbol{e}
\end{aligned}
$$

- The data frame is the resulting design from Step 2.1.
- The **permute** factor is Genotype.
- The **static** factors are RowBlk, ColBlk, Column and Row.
- swap factor in **odw** is set to RowBlk:ColBlk to keep the resolvable design obtained from Step 2.1.
- The total genetic effects are used for calculation of the $\mathcal{A}$- criterion.

Impact of using genetic relatedness on genetic gain

A small study

- A small study to assess the impact of using genetic relatedness in the design.
- The S1 Desi Chickpea experiment was used.
- Four designs were generated using **odw**.
- These designs were the factorial combinations of using (+) or not using (-) genetic relatedness in stages one and two of the design construction.

SG+/+	Uses genetic relatedness in both stages.
SG+/-	Uses genetic relatedness in stage one only & random allocation of plots to genotypes in stage two.
SG-/+	Uses random allocation of those test lines with enough seed to packet choice in stage one & genetic relatedness in stage two.
SG-/-	Does not use genetic relatedness in stages one and two - default *p*-rep designs from Cullis et al. (2006).

Impact of using genetic relatedness on genetic gain

A small study

- The quality of each design was assessed by calculating the $\mathcal{A}$- criterion of the design against the "correct" linear mixed model, using all static terms and the appropriate variance model for the total genetic effects.
- The design with the smallest $\mathcal{A}$- criterion will result in a higher probability of selecting the best subset of genotypes for progression (Bueno Filho and Gilmour, 2007).
- The $\mathcal{A}$- criteria for each design, expressed as the difference from design SG+/+, and multiplied by $1e4$ were

$$\text{SG } +/+ = 0 \qquad \text{SG } -/+ = 27$$
$$\text{SG } +/- = 15 \qquad \text{SG } -/- = 43$$

CAIGE MET design

IMET design in odw

- Principles are similar to those presented in this talk.
- There were severe seed supply issues for a number of lines - must use genetic relatedness.
- Different site sowing rates introduced further complexity in determining packet choices.
- Genetic relatedness is used to:
  - Allocate packets to genotypes subject to constraints in terms of seed supply, home sites, numbers of sites and plots within sites and so on.
  - Allocate sites to genotypes subject to home sites and other regional sites.
  - Allocate plots to genotypes - achieved in two steps, allowing only interchanges within sites.

CAIGE MET design

Summary of design output

Table 1: Summary of the IMET design for each trial, including the number of plots, seed quantity/plot, number of unique lines, number of unique lines that had one plot/environment (p1), number of unique entries that had two plots/environment (p2), partial replication percentage (p-rep %) and incompleteness percentage (%). [3]

Env	nPlot	plot.wt	#lines	p1	p2	p-rep (%)	incompleteness (%)
W22BALA5	288	30	263	238	25	9.5	22.0
W22BELL2	288	30	265	242	23	8.7	21.4
W22GOOM6	288	30	261	234	27	10.3	22.6
W22JUNE2	288	30	257	226	31	12.1	23.7
W22BREE2	288	50	232	176	56	24.1	31.2
W22COND4	288	50	221	154	67	30.3	34.4
W22LONG3	288	50	231	174	57	24.7	31.5
W22MING6	288	50	195	102	93	47.7	42.1
W22NARR2	432	50	337	242	95	28.2	0
W22NORT2	288	50	238	188	50	21.0	29.4
W22ROSE5	288	50	214	140	74	34.6	36.5
W22YORK6	288	50	218	148	70	32.1	35.3

[3] Due to time constraint, a random allocation of packet choice to lines was used in Stage one. Genetic relatedness was used in Stage two.

Conclusion

Summary

- Demonstrated the potential gains in accuracy of selection by using genetic relatedness for simpler designs.
- The potential increases in genetic gain from the use of IMET designs would most likely exceed those obtained from the design of single experiments.
- **odw** can also be used for:
  - Selective phenotyping (Huang et al., 2013).
  - IMET designs using reduced animal models.
  - Multi-phase experimental designs.
- **odw** is freely available from `mmade.org`; written and maintained by David Butler.
- Two publications are in preparation - Stay tuned!
  - *B. Cullis, A. Smith and D. Butler. The construction of incomplete multi-environment trial designs using a model-based approach. Manuscript in Prep., 2022.*
  - *D. Butler and B. Cullis. On Model Based Design of Comparative Experiments in R. Manuscript in Prep., 2022.*

Acknowledgement

- To Professor Brian Cullis for his supervision and guidance on this work.
- To all the co-authors for their inputs towards this presentation.
- To Grains Research & Development Corporation for funding the CAIGE project.
- To all collaborating breeders for many helpful discussions and the use of data.
- To CBADS-SPI team.

Thank you!

References i

J. Bueno Filho and S. Gilmour. Block designs for random treatment effects. *Journal of Statistical Planning and Inference*, 137(4):1446–1451, 2007. doi: 10.1016/j.jspi.2006.02.002.

David Butler. *Optimal experimental design under the linear mixed model*, 2022. odw package manual, mmade.org.

B. Cullis, A. Smith, and N. Coombes. On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4): 381–393, 2006. ISSN 1085-7117. doi: 10.1198/108571106X154443.

Brian Cullis, Alison Smith, and David Butler. The construction of incomplete multi-environment trial designs using a model-based approach. Unpublished Manuscript, 2022.

Emma Huang, David Clifford, and Colin Cavanagh. Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. *Theoretical and Applied Genetics*, 126(2):379–88, 2013. doi: 10.1007/s00122-012-1986-4.

References ii

R Core Team. R: A Language and Environment for Statistical Computing, 2020.

A.B Smith and B.R Cullis. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, 214(143), 2018. doi: 10.1007/s10681-018-2220-5.

Alison Smith, Adam Norman, Haydn Kuchel, and Brian Cullis. Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. *Frontiers in Plant Science*, 12, 2021.