Multi-phase design and analysis using a single step multi-experiment approach with factor analytic models to improve accuracy of late maturity α-amylase classification in wheat

Lu Wang[1]

Brian Cullis[1]

Bettina Berger[2] and Daniele Giblot-Ducray[2]

AASC 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]The Plant Accelerator
School of Agriculture, Food and Wine
University of Adelaide
bettina.berger@adelaide.edu.au

Outline

I. Introduction

II. Multi-phase experimental designs in **odw**

III. Single stage MET analysis in **ASReml-R**

Introduction

Introduction

Background

- The enzyme, α-amylase, is responsible for the degradation of starch into sugars in wheat grains.
- Wheat genotypes prone to late maturity α-amylase (LMA) produce high levels of α-amylase if exposed to certain environmental conditions during grain development (Mrva and Mares, 2001).
- Genetic propensity to express LMA (GPE-LMA) is routinely assessed through LMA expression experiments (LMAEEs) that provide LMA classification of Australian wheat genotypes.
- Breeding lines[1] with high levels of LMA are deemed unsuitable for high quality end-products, resulting in significant financial losses for growers.
- The current protocol of phenotyping for LMA uses optical density (OD) readings as a predictor of GPE-LMA of the genotype.

[1] line is synonymous with genotype

Introduction

LMAEEs

- Two sets of experiments are conducted annually, **WIN** and **SUM**, which form a pair of LMAEEs with a high proportion of lines in common.
- The current pair (**WIN21** and **SUM22**) is analysed together with previous seasons[2] in a multi-environment trial (MET) analysis.
- The aim of the LMAEE MET analysis is to classify the current set of test lines against the benchmark, RAC655, which is known to express LMA.
- There have been significant improvements to the protocol since 2019...
  - testing facilities;
  - experimental designs;
  - statistical analyses.

[2] season is synonymous with environment

Introduction

Model-based design approach for multi-phase experiments

- LMAEE is an example of a **multi-phase** experiment (Brien, 2017) that comprises a glasshouse (GH) phase and an ELISA[3] laboratory phase.
- It involves several time periods and has observational units which are completely different from the preceding phase (Butler et al., 2009).
- Non-genetic influences on OD during grain development could lead to exhibition of genotype by environment (GE) interaction (Mrva and Mares, 2001).
- Model-based design approach provides the framework for generating efficient designs for complex multi-phase experiments.
- This approach generates an optimal design under a pre-specified (analysis) model and a design criterion.

[3]ELISA: enzyme-linked immunosorbent assay

Introduction

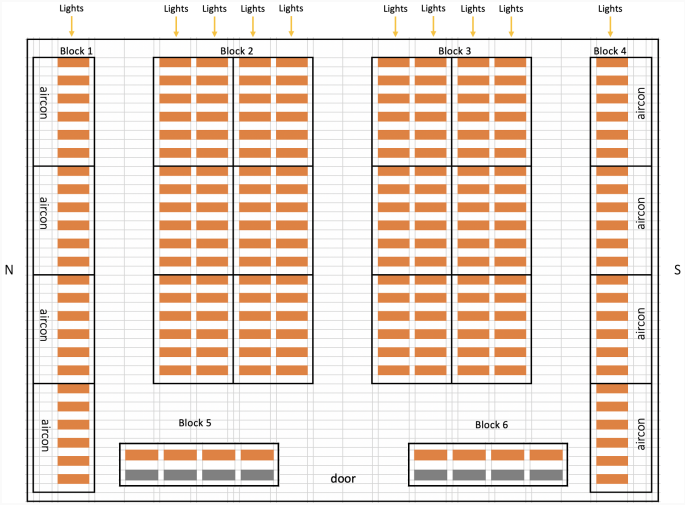
Model-based design approach for multi-phase experiments

- **odw** (Butler, 2022) package is freely available on `mmade.org` & constructs optimal designs under the linear mixed model (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 designs with genetic relatedness (Cullis et al., 2022).
  - ♦ **multi-phase experimental designs** (Smith et al., 2006).
- In the case of LMAEEs, **odw** generates efficient designs that
  - ♦ accommodate sources of non-genetic variation which arise in both phases;
  - ♦ allow design information from the glasshouse phase to be carried on to the ELISA phase.

Multi-phase experimental designs in odw

Multi-phase experimental designs in odw

Phase I: Glasshouse experiment



- 6 blocks, 22 benches, 200 trays, 2000 pots.
- Edge is a factor with 2 levels.
- Previous analyses have shown Edge to be a significant source of variation in the GH experiment.

Phase I: Glasshouse experiment

Tray layout and genetic materials

	Block 1 column				
row	1	2	3	4	
1	o	o	o	o	
2	o	×	×	o	tray 1
3	o	o	o	o	
4	o	o	o	o	
5	o	×	×	o	tray 2
6	o	o	o	o	
7	o	o	o	o	
8	o	×	×	o	tray 3
9	o	o	o	o	
10	o	o	o	o	
11	o	×	×	o	tray 4
12	o	o	o	o	
⋮	⋮	⋮	⋮	⋮	⋮
70	o	o	o	o	
71	o	×	×	o	tray 24
72	o	o	o	o	

- Pots within a tray arranged in 3 rows by 4 columns, with middle two positions ('x') on each tray left empty to allow for airflow.
- 2000 plots (pots) to 714 lines (from 8 sources).
- There was no information available on the genetic relatedness of the lines, hence replications for lines were chosen at random.

Phase I: Glasshouse experiment

Design overview

Design construction for the GH design involves two stages:

- Stage One - allocation/randomisation of packet[4] choice (pC) to lines.
- Stage Two - allocation of plots to lines given packet choice status.
- Each step uses a different call to **odw**.

[4] packet refers to a plot/pot in the glasshouse

Phase I: Glasshouse experiment

Stage one - packet choice allocation

	pC2	pC3	pC4	pC50
RAC655	0	0	0	1
Check lines	0	0	17	0
Test lines	206	490	0	0

- Packet choices: 2, 3, 4 and 50.
- In the absence of genetic relatedness, packet choices of 2 and 3 for test lines were determined using simple random sampling, evenly across sources.

Phase I: Glasshouse experiment

Stage two - allocating plots to lines

odw constructs designs under the following LMM (Cullis et al., 2022):

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

- $\boldsymbol{y}$ 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.

Phase I: Glasshouse experiment

odw linear mixed model

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{W}_1\beta_1 + \boldsymbol{W}_2\beta_2 + \boldsymbol{e} \\
&= \underbrace{\textbf{permute}\ \text{set}}_{\textbf{objective}\ +\ \textbf{linked}} + \textbf{static}\ \text{set} + \ \text{errors}
\end{aligned}
$$

- 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.
- Two rows of $\boldsymbol{W}_1$ are interchanged during permutation, the rows of $\boldsymbol{W}_2$ are considered invariant (static).
- **odw** is the ONLY design software that allows for linked factors, which revolutionised multi-phase experimental designs.

Phase I: Glasshouse experiment

Allocating plots to lines

- Call to **odw** to generate the GH design is:

```
gh.des <- odw(fixed=~ 1,
              random=~ Line + Block + Edge + Block:Bench + Tray,
              permute=~ Line ,
              residual=~ units,
              search= "tabu+rw", maxit= 50, data= init.df)
```

  - objective factor: **Line**, from which the $\mathcal{A}$-value is computed.
  - linked factor: *NULL*.
  - static set: Block, Edge, Block:Bench and Tray.

- Spatial arrangement (i.e., rows and columns) of pots within blocks was not considered in the randomisation of GH experiment due to design timeframe.

Multi-phase experimental designs in odw

Phase II: Laboratory (ELISA) phase

- Grains harvested from the 2 plants in a pot milled to form a 1.2g bulk meal, which then soaked in a solution overnight; ELISA sample(s) were taken from each flour soak and stored in the cool room.
- Since the non-genetic variation was mostly from the glasshouse,
  - RAC655 & test lines that had 2 GH pots or less: 2 ELISA samples from each pot;
  - check lines & test lines that had 3 GH pots: 1 ELISA sample from each pot.
- Each ELISA slide has 88 wells (12 columns x 8 rows) available for test material.
- The **WIN21** ELISA experiment required 2462 ELISA wells, hence 28 slides.
- To accommodate management, it was recommended to conduct the ELISA experiment in **3** runs (weeks), **2** days per run.

Phase II: ELISA phase

Generating the ELISA design in odw

- Due to the constraint that ELISA duplicates must be processed within a run, a design that allocates GH pots across ELISA Run was generated prior to expanding the data frame to incorporate the ELISA duplicates.
- Call to **odw** to generate the final ELISA design is:

```
elisa.des <- odw(fixed=~ 1,
                 random=~ Line + Block + Edge + Block:Bench + Tray + PotBarcode + Run +
                          Run:Day + Run:Day:Slide + Run:Day:Slide:SRow + Run:Day:Slide:SCol,
                 permute=~ Line | Block + Edge + Block:Bench + Tray + PotBarcode ,
                 residual=~ units,
                 swap=~ Run, search= 'tabu+rw', maxit= 50, data= w21eli.df)
```

  - objective factor: **Line**, from which the $\mathcal{A}$-value is computed.
  - linked factors: Block, Edge, Block:Bench, Tray and PotBarcode from the GH experiment that need to be permuted with Line in parallel, however, they do not contribute to the $\mathcal{A}$-value.
  - static set: all factors associated with the ELISA experiment.

Single stage MET analysis in ASReml-R

Analysis overview - FA modelling of the GE effects

- 6 environments, 2025 genotypes and 15948 data entries; genotype connectivity: 19-700.
- The current recommended method of analysis follows Smith and Cullis (2018) and involves a linear mixed model with factor analytic (FA) variance structure for the genotype by environment random effects ($\boldsymbol{u}_g$); and
- appropriate modelling of the non-genetic effects and residuals in a combined single stage MET analysis.
- The FA model of order k (FAk) for $\boldsymbol{u}_g$ within an LMAEE can be written as

$$\begin{aligned} \boldsymbol{u}_g &= (\lambda_1 \otimes \boldsymbol{I}_m)\boldsymbol{f}_1 + (\lambda_2 \otimes \boldsymbol{I}_m)\boldsymbol{f}_2 + \cdots + (\lambda_k \otimes \boldsymbol{I}_m)\boldsymbol{f}_k + \boldsymbol{\delta} \\ &= (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta} \end{aligned}$$

where

- $\boldsymbol{\Lambda}$ is the $p \times k$ matrix of environment loadings for individual factors.
- $\boldsymbol{f}$ is the mk-vector of genotype scores (ordered as genotypes within factors).
- $\boldsymbol{\delta}$ is the mp-vector of GE lack of fit effects.

FA modelling of the GE effects - continued

$$\boldsymbol{u}_g = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}$$

It is assumed that $\boldsymbol{f}$ and $\boldsymbol{\delta}$ are independent and distributed as multivariate Gaussian with zero means and variance matrices given by

$$\text{var}(\boldsymbol{f}) = \boldsymbol{D} \otimes \boldsymbol{I}_m \text{ and } \text{var}(\boldsymbol{\delta}) = \boldsymbol{\Psi} \otimes \boldsymbol{I}_m$$

where

- $\boldsymbol{D}$ is a $k \times k$ symmetric positive (semi)-definite matrix that is referred to as the factor score variance matrix.
- $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with elements referred to as specific variances.

These assumptions lead to a variance matrix for the GE effects of the form

$$\text{var}(\boldsymbol{u}_g) = (\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m$$

The between environment genetic variance matrix is then given by $(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{\Psi})$.

Fitting FALMM in ASReml-R

```
rr.asr <- asreml(y~ Env,
          random=~ rr(Env):Line + diag(Env):Line +
                   at(Env):Edge + at(Env):Block + at(Env):Block:Bench +
                   at(Env):Tray + at(Env):Block:ghColumn +
                   at(Env):Block:ghRow + at(Env):PotBarcode +
                   at(Env):Run + at(Env):Run:Day + at(Env):Run:Day:Slide +
                   at(Env):Run:Day:Slide:SCol + at(Env):Run:Day:Slide:SRow,
         residual=~ dsum(~id(units)|Env),
         data= lma.df, na.action= na.method(x='include',y='include'))
```

- **ASReml-R** (Butler et al., 2019) provides residual maximum likelihood estimates of the variance parameters and empirical best linear unbiased predictions of random effects.
- Final model was an FA1, the first factor explained 93.5% of the genetic variance.
- The correlation between pairs of environments ranges from 0.86 to 0.98.
- Classification was performed for 714 lines submitted in **WIN21** and **SUM22**, saved in this.seas.
- The predictions for the *common* line by environment effects can be obtained by

```
predict(rr.asr,classify='Line',levels=list(Line=this.seas), only=rrterm , vcov=T)
```

Multiple testing with FDR at q^*

H_{0_i}: The true GPE-LMA of a breeding line i (u_{g_i}) is greater than the true mean GPE-LMA of the benchmark, RAC655 (u_{g_c}).

- 714 - 2 = 712 tests, denoted by m^*.
- Rejection of H_0 was determined using the Benjamini and Hochberg (1995) approach[5] of controlling the false discovery rate at a significance level $q^* = 0.05$ or 0.01.
- Results are returned as `TRUE` or `FALSE`.
- A value of `TRUE` implies that H_0 is rejected.

[5]FDR is designed to control the (expected) proportion of false positives among the set of rejected hypotheses

Multiple testing with FDR at q^*

- To control FDR at level q^*:
  1. Order the p-values: $p_1 \leq p_2 \leq \cdots \leq p_{m^*}$.
  2. Find the test with the highest rank, i, for which the p-value, p_i, is less than or equal to $\frac{i}{m^*} q^*$.
  3. Reject H_0 for tests that had

$$p(i) \leq \frac{i}{m^*} q^*$$

- Some results:

Rank (u_{g_i} - u_{g_c})	p-values	$(i/m^*)q^*$	Rejection at $q^* = 0.01$	Result
Wyalkatchem	≈ 0	1.69e-04	TRUE	PASS
Emu Rock	1.12e-13	4.21e-04	TRUE	PASS
Chara	5.50e-04	7.15e-03	TRUE	PASS
Cranbrook	2.46e-01	9.49e-03	FALSE	NOT PASS
Kennedy	4.75e-01	9.68e-03	FALSE	NOT PASS

Conclusion and future work

- Demonstrated how we used **odw** (Butler, 2022) to generate efficient designs for the complex scenario of multi-phase experiments.

  D. Butler and B. Cullis. On Model Based Design of Comparative Experiments in R. Manuscript in Prep., 2022.

- Demonstrated how we implemented a single stage factor analytic linear mixed model approach in the MET analysis in **ASReml-R** (Butler et al., 2019), which leads to improvement in accuracy of LMA classification in wheat.

- Future work:
  - Investigate data reliability (glasshouse vs. field).
  - Investigate multiple testing methods suitable for correlated tests.
  - Investigate the potential improvement in accuracy with the inclusion of genomic data.

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 LMA classification project.

Thank you!

References i

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300, 1995.

C. J. Brien. Multiphase experiments in practice: A look back. *Australian & New Zealand Journal of Statistics*, 59(4):327–352, 2017. doi: https://doi.org/10.1111/anzs.12221.

D. Butler, A. Smith, B. Cullis, B Gogel, A.R Gilmour, and R. Thompson. *ASReml-R Reference Manual Version 4*, 2019.

D. G. Butler, M. K. Tan, and B. R. Cullis. Improving the accuracy of selection for late maturity α-amylase in wheat using multi-phase designs. *Crop and Pasture Science*, 60(12):1202–1208, 2009. ISSN 18360947. doi: 10.1071/CP09124.

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

References ii

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.

K. Mrva and D. Mares. Induction of late maturity alpha-amylase in wheat by cool temperature. *Journal of Agricultural Research*, 4(52):477–484, 2001.

A. Smith, P. Lim, and B. Cullis. The design and analysis of multi-phase plant breeding experiments. *The Journal of Agricultural Science*, 144:393, 2006. ISSN 0021-8596. doi: 10.1017/S0021859606006319.

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.