

# Embedded partially replicated designs for grain quality testing

Brian Cullis

`brian.cullis@dpi.nsw.gov.au`

Biometrics

New South Wales Department of Primary Industries

Embedded partially replicated designs

## Collaborations and Acknowledgements

- This presentation is joint work with Alison Smith (NSWDPI) and Robin Thompson (Rothamsted Research, UK).
- Thanks to Neil Coombes for generation of designs and helpful discussions
- Grains Research and Development Corporation for financial support.

## Why do the work?

- Motivation for this work originated from our involvement with the National Variety Trials system (NVT), as part of the GRDC funded Statistics for the Australian Grains Industry project (SAGI).
- What is NVT and what is it about?

## National Variety Trials system - NVT

### NVT facts

- NVT generates (independent) information for growers on the performance of newly released crop varieties.
- NVT complements the various plant breeding programs:

## National Variety Trials system - NVT

### NVT facts

- NVT generates (independent) information for growers on the performance of newly released crop varieties.
- NVT complements the various plant breeding programs:
  - Breeders make release decisions prior to nominating lines for testing in NVT
  - NVT tests lines which are either commercial or very close to release.

## National Variety Trials system - NVT

### NVT facts

- NVT generates (independent) information for growers on the performance of newly released crop varieties.
- NVT complements the various plant breeding programs:
  - Breeders make release decisions prior to nominating lines for testing in NVT
  - NVT tests lines which are either commercial or very close to release.
- NVT was established in 2005 by the GRDC and is managed by the Australian Crop Accreditation System Limited (ACAS).

## National Variety Trials system - NVT

### NVT facts

- NVT generates (independent) information for growers on the performance of newly released crop varieties.
- NVT complements the various plant breeding programs:
  - Breeders make release decisions prior to nominating lines for testing in NVT
  - NVT tests lines which are either commercial or very close to release.
- NVT was established in 2005 by the GRDC and is managed by the Australian Crop Accreditation System Limited (ACAS).
- More than 580 trials are sown at over 250 locations each year
- Crops tested are: Wheat; Barley; Triticale; Oat; Canola; Lupin; Lentil; Field Pea; Faba Bean and Chickpea.

## National Variety Trials system - NVT

### SAGI's Involvement

Provision of IT and statistical support for

## National Variety Trials system - NVT

### SAGI's Involvement

Provision of IT and statistical support for

- Design and analysis of (yield) for individual field trials ( $n > 600$ , annually)
- (MET) Analysis of yield across years and locations for  $p = 12$  crops
- Presentation and generation of web-based reports on yield performance
- Development of software and data-base tools (ASReml-R and DiGGeR and KaTmanDoo).

**WEAKNESS?**

## MAJOR WEAKNESS OF NVT information

- Varieties are selected and released on the basis of their improved performance based on a range of traits
- Key economic traits include yield, disease and quality
- Quality traits include both physical (eg. grain density, grain plumpness, grain size) and end-product (eg milling, baking, malting, brewing, digestibility)
- NVT routinely measures and **reports** “information” on physical and some end-product traits for wheat, barley, oats and canola.

**NO ANALYSIS!**

## Current protocols for physical and quality traits NVT and numerous plant breeding programs

- Samples are taken from individual plots and composited
- Raw data is then presented for each trial
- No analysis is attempted due to confounding of plot and  $G \times E$  effects - though some progress here is possible

**NO ANALYSIS!**

## Proposition

### Embedded $p$ -rep designs

- Propose a new approach to design and analysis
- Based on partially ( $p$ -rep) designs of Cullis *et al.* (2006)
- Embed the  $p$ -rep design within a standard replicated trial
- Augment analysis of  $p$ -rep trial data with composited data from remaining plots

**MAXIMIZE GAIN - MINIMIZE COST**

## Model for $j^{\text{th}}$ trial, $j = 1 \dots t$

The model for  $\mathbf{y}_j^{n_j \times 1} = \text{vec}(Y^{r_j \times c_j})$  can be written as

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\tau}_j + \mathbf{Z}_{g_j} \mathbf{u}_{g_j} + \mathbf{Z}_{p_j} \mathbf{u}_{p_j} + \mathbf{e}_j$$

where the vectors  $\boldsymbol{\tau}_j$ ,  $\mathbf{u}_{g_j}$ ,  $\mathbf{u}_{p_j}$  represent fixed effects, random variety effects and random non-genetic (or peripheral, ie design and additional) effects respectively.

Typically variance models for the random and residual effects would be:

$$\text{var}(\mathbf{u}_{g_j}) = \sigma_{g_j}^2 \mathbf{G}_g,$$

$$\text{var}(\mathbf{u}_{p_j}) = \bigoplus \sigma_{p_{jk}}^2 \mathbf{I}_{q_{jk}},$$

$$\text{var}(\mathbf{e}_j) = \mathbf{R}_j = \sigma_j^2 \boldsymbol{\Sigma}_{c_j} \otimes \boldsymbol{\Sigma}_{r_j}$$

## MET Model for series of $t$ trials

The MET model for  $\mathbf{y} = (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_t)'$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

where the vectors  $\boldsymbol{\tau}$ ,  $\mathbf{u}_p$ ,  $\mathbf{u}_g$  represent trial specific fixed, random-peripheral and random-genetic effects respectively. Typically variance models for the random and residual effects would be:

$$\begin{aligned} \text{var}(\mathbf{u}_g) &= \mathbf{G}_e \otimes \mathbf{G}_g, \\ \text{var}(\mathbf{u}_p) &= \mathbf{G}_p = \bigoplus_{j=1}^t \bigoplus_{k=1}^{b_j} \sigma_{p_{jk}}^2 \mathbf{I}_{q_{jk}}, \\ \text{var}(\mathbf{e}) &= \mathbf{R} = \text{diag}(\mathbf{R}_j) \end{aligned}$$

and  $\mathbf{G}_e = \boldsymbol{\Lambda}^{t \times k} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}$  - the so-called order  $k$  Factor Analytic model, formulated in terms of loadings and specific variances.

## Canola Oil Content - CBWA

Is there spatial and  $G \times E$  in quality traits?

### Background

Seven early stage canola breeding trials grown in 2007 across Australia.

Total of 260 entries with  $p$ -rep designs (created in DiGGeR ). All laid out in rectangular arrays. We consider oil content (measured using NIR).

### Summary of Trials

Trial	Rows	Columns	Entries	Mean oil	$p$
1	48	6	213	38.2	0.35
2	51	6	232	43.9	0.32
3	52	6	245	40.6	0.27
4	52	6	252	46.0	0.24
5	53	6	254	45.6	0.25
6	49	6	220	38.9	0.34
7	53	6	260	47.5	0.22

## Canola Oil MET

### Spatial and Extraneous Effects

Trial	$\hat{\tau}_0$	$\hat{\tau}_{xrow}$	$\hat{\sigma}_{blk}^2$	$\hat{\sigma}_{col}^2$	$\hat{\sigma}^2$	$\hat{\rho}_c$	$\hat{\rho}_r$
1	38.2	0.012	0.104		0.326	0.13	0.39
2	43.9	-0.044	0.087	0.306	0.377	0.20	0.45
3	40.6	-0.018	0.082		0.594	0.14	0.59
4	45.9		0.000		1.282	0.35	0.78
5	45.7		0.271		2.217	0.26	0.61
6	38.9		0.000	0.123	0.478	0.27	0.55
7	47.6		0.000		0.707	0.21	0.56

## Canola Oil MET

### Genetic variance parameters

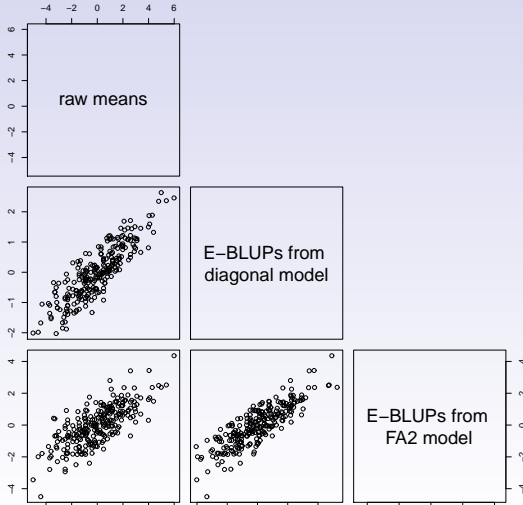
### REML Estimates of $G_e$

#### Model Fit

Model	$n_\gamma$	$n_K$	$\log l + 1103$
1. Diag	7	37	-576.3
2. Unif	2	32	-76.0
3. FA1	14	46	-11.8
4. FA2	20	52	-0.6

Trial	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\Psi}$	diag ( $\hat{G}_e$ )
1	0.762	0.258	0.296	0.944
2	1.163	0.000	0.076	1.429
3	0.986	-0.003	0.201	1.173
4	1.463	-0.040	0.395	2.536
5	1.259	0.278	0.214	1.877
6	0.868	0.410	0.072	0.993
7	1.284	-0.272	0.100	1.822
$\{\hat{\rho}_{e_{jk}}\} \in [0.69, 0.94]$				

## E-BLUPs of variety effects for trial 5



## Embedded $p$ -rep designs

### Design paradigm

Construction of an efficient and contiguous design for an expensive trait, embedded within a replicated and efficient design for a less expensive trait.

Trade-off of efficiency of the latter for the former.

### Design definition & process

An embedded  $p$ -rep design is a partially replicated design contained within a RCB design. The optimisation process is sequential commencing with the  $p$ -rep design, followed by formation of the RCB design conditional on the  $p$ -rep design **embedded** within it. Each design search is undertaken using a supervised learning algorithm which minimises a pre-specified objective function (typically the  $A$ -value) for chosen blocking and spatial correlation models. All with DiGGeR !

## Embedded $p$ -rep designs

### Illustrative Example

### Design specifications

$n_b = 2$ ,  $m = 90$ , for  $p = 1/3$ .

Full layout  $30 \times 6$ , with block 1 = columns 1-3; block 2 = columns 4-6. Embedded design  $20 \times 6$ .

Plots shaded grey are those assigned to replicated entries in  $p$ -rep portion.

### Design layout

1	84	85	36	56	87	68
2	10	66	39	50	49	33
3	40	35	88	46	12	26
4	28	70	90	78	58	74
5	77	81	23	79	19	20
6	38	17	54	86	71	15
7	30	21	11	29	43	75
8	45	76	34	22	23	8
9	58	56	31	30	63	52
10	29	44	82	10	37	5
11	61	68	19	1	35	54
12	51	6	69	24	67	72
13	53	32	64	25	80	65
14	18	41	33	9	2	85
15	43	24	7	53	47	73
16	80	57	13	77	41	60
17	42	16	15	6	27	34
18	55	12	4	83	28	39
19	8	3	2	59	52	88
20	14	78	89	17	54	48
21	65	71	22	18	14	69
22	27	74	25	4	81	84
23	46	60	87	42	7	70
24	79	75	26	66	38	76
25	59	49	86	32	11	55
26	48	63	83	3	45	44
27	9	72	37	61	13	31
28	73	62	52	82	51	89
29	50	20	47	57	36	40
30	67	5	1	90	21	16
	1	2	3	4	5	6

column

## The next twist

### Hybrid linear mixed models

Motivation for a hybrid analysis

## The next twist

### Hybrid linear mixed models

Motivation for a hybrid analysis

- Embedded  $p$ -rep design involves  $s < n$  plots
- But what of the remaining  $n - s$  plots?

## The next twist

### Hybrid linear mixed models

#### Motivation for a hybrid analysis

- Embedded  $p$ -rep design involves  $s < n$  plots
- But what of the remaining  $n - s$  plots?
- Why not composite those  $(1 - p)s$  plots in the embedded design with the plots from the full design?
- Can we then formulate a hybrid linear mixed model for the mixture set of observations?
- Answer - **YES**

## Hybrid linear mixed model

### Individual Trial

- We consider a transformation of  $\mathbf{y}_j$  commensurate with a compositing process, ie. averaging of individual replicate data for the subset of genotypes which are not replicated in the embedded design.

## Hybrid linear mixed model

### Individual Trial

- We consider a transformation of  $\mathbf{y}_j$  commensurate with a compositing process, ie. averaging of individual replicate data for the subset of genotypes which are not replicated in the embedded design.
- Denote  $\mathbf{z}_j = \mathbf{D}_j \mathbf{y}_j$  to be the vector of original and composited data, for some  $\mathbf{D}_j^{s_j \times n_j}$ , then linear mixed model for  $\mathbf{z}_j$  is

$$\mathbf{z}_j = \mathbf{D}_j \mathbf{X}_j \boldsymbol{\tau}_j + \mathbf{D}_j \mathbf{Z}_{g_j} \mathbf{u}_{g_j} + \mathbf{D}_j \mathbf{Z}_{p_j} \mathbf{u}_{p_j} + \mathbf{D}_j \mathbf{e}_j$$

where the all of the fixed, random and residual vectors have the same meaning as before.

## Hybrid linear mixed model

### Estimation and Extensions

This model involves some non-standard design matrices and estimation requires specialist software (eg. ASReml-R )

#### Syntax

```
pcomp.asr <- asreml(z ~ 1+lr2, random=~Entry + grp('blk') +  
grp('range') + str( grp('plot'), ~ar1v(6):ar1(30)), data=site2.df,  
family=asreml.gaussian(dispersion=.00001),  
control=asreml.control(group=  
list('blk'=184:185,'range'=186:191,'plot'=4:183)))
```

Extensions to MET data are trivial (says me!)

## What's the gain?

Are  $ep$ -rep designs and  $p$ -comp linear mixed models worth it?

### Simulation Details

Data generated from real MET data-set ( $N = 192$ ). Methods are M1: true model fitted to full data-set, M2: true model fitted to  $ep$ -data-set, M3: true model fitted to  $p$ -composited data-set, M4 'best possible' model fitted to  $full$ -comp data-set and M5 raw  $full$ -comp. Figures are response to selection (top 5 entries), absolute value for M1 then % decrease for other methods.

### Results

Trial	M1	M2	M3	M4	M5
1	1.82	4.8	1.9	10.9	3.2
2	2.31	1.2	0.6	2.1	6.3
3	2.06	2.2	0.6	5.3	6.1
4	3.06	1.3	1.2	5.8	7.2
5	2.61	1.6	1.2	3.7	15.1
6	1.91	3.0	2.0	4.9	6.0
7	2.58	1.2	0.6	2.9	4.9
Mean		2.2	1.2	5.1	7.0

## Further Thoughts

- Plan to implement the new approach in the NVT system for the coming season
- Encourage use in other applications including QTL and association mapping studies
- Further research on optimality of design search