

Embedded partially replicated designs & hybrid mixed models for grain quality testing

Brian Cullis

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

Biometrics

Industry and Investment New South Wales

Embedded designs and hybrid mixed models

Collaborations and Acknowledgements

- This presentation is joint work with Alison Smith (NSWDPI) and Robin Thompson (Rothamsted Research, UK).
- Thanks to Neil Coombes and David Butler 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 individual field trials ($n > 500$, annually) for obtaining yield information
- (MET) Analysis of yield across years and locations for 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.

Based on composited samples and therefore NO analysis has been done

National Oat Breeding Program - NOBP

Late stage evaluation trials and NVT MET

- MET analysis in 2008 requested by Pamela Zwer.
- MET data-set comprised a total of 38 trials, each of three replicates.
- The data involve 8 quality traits:
 - grain weight, hectolitre weight, screenings, measured using technical instruments, and
 - grain protein, groat percent, percent groat oil content, estimated metabolizable energy and minolta L, measured using NIR.

Summary of Trials for NOBP Late Stage MET Protocols and approaches

Year	Trials	Type	Varieties	Data
2005	13	S4	30-80	1 rep
2006	5	S4	48-52	1 rep
2007	16	NVT	15-24	Comp
2007	4	S4	48	1 rep

- Samples are either taken from individual plots and composited, or taken from a single replicate
- Raw data is then *interpreted* or *analysed* for each trial
- We attempt no formal analysis due to confounding of plot and $G \times E$ effects - though some progress here is possible

same protocols apply to many early stage and QTL trials

Proposition

Embedded p -rep designs & hybrid mixed models

- Propose new approaches to design and analysis,
- Mindful of resources and statistical efficiency
- Approaches depend on the type of trial: ie late stage - replication exceeding 2; early stage or QTL - replication 2 or less
- Modify field trial design if necessary - based on partially (p -rep) designs of Cullis *et al.* (2006)
- Undertake hybrid mixed model using all of the plots.

Model for j^{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:

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

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

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

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.

Early Stage Selection trials - Canola Oil Content

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

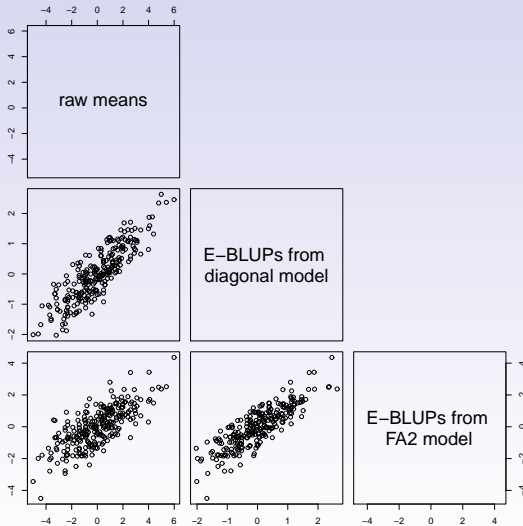
Model Fit

Model	n_γ	n_κ	$\log +1103$
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 30×6 , with block 1 = columns 1-3; block 2 = columns 4-6. Embedded design 20×6 .

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	64
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	62	88
20	14	78	89	17	64	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

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 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

Late Stage Evaluation and NVT trials

NOBP

- In 2008 (after much hassling) NOBP altered protocols for the S4 trials
- Samples were taken from two (out of three) replicates
- Aim was to determine level of spatial variation in oat quality traits

Oat Grain Oil Content NOBP

Spatial and Extraneous Effects

Trial	$\hat{\tau}_0$	$\hat{\sigma}_{blk}^2$	$\hat{\sigma}_{col}^2$	$\hat{\sigma}^2$	$\hat{\rho}_c$	$\hat{\rho}_r$
1	5.16	0.014		0.058		0.16
2	5.74	0.000		0.046	0.29	0.43
3	5.18	0.004		0.104	0.23	0.61
4	5.43	0.004		0.076	0.27	
5	5.25	0.000		0.151		0.20
6	5.97	0.002	0.099	0.090	0.04	0.20
7	5.22	0.002		0.070		
8	5.54	0.010		0.087	0.11	0.52

NOBP hybridized trial

Illustrative Example

Compositing plan

$n_b = 3$, $m = 48$. Full layout
 12×12 , with block 1 = columns
 1-4; block 2 = columns 5-8;
 block 3 columns 9-12. Plots
 shaded pink and blue in block 3
 have been composited with
 those in blocks 1 and 2
 respectively.

Design layout

1	V36	V11	V20	V32	V25	V47	V46	V7	V3	V32	V37	V23
2	V34	V16	V27	V42	V19	V43	V42	V16	V41	V34	V40	V8
3	V21	V41	V43	V35	V31	V3	V39	V48	V5	V44	V9	V18
4	V40	V8	V48	V19	V21	V17	V33	V29	V10	V31	V4	V45
5	V10	V12	V22	V5	V44	V12	V38	V22	V36	V29	V20	V48
6	V18	V7	V6	V14	V41	V9	V6	V2	V6	V38	V47	V21
7	V31	V3	V1	V24	V34	V13	V14	V11	V13	V14	V2	V25
8	V33	V39	V45	V46	V32	V1	V37	V40	V27	V35	V28	V7
9	V17	V38	V37	V44	V30	V10	V36	V28	V24	V19	V39	V30
10	V9	V25	V26	V30	V5	V24	V18	V45	V22	V46	V11	V42
11	V15	V2	V28	V23	V27	V4	V35	V23	V12	V17	V1	V43
12	V47	V13	V29	V4	V26	V15	V8	V20	V16	V33	V26	V15
	1	2	3	4	5	6	7	8	9	10	11	12

What's the gain for NOBP/NVT

Are hybrid linear mixed models worth it?

Simulation - Single Site only

Data generated from real MET data-set (Trial 2). Methods are M1: true model fitted to full data-set, M2: true model fitted to two replicate data-set, M3: true model fitted to two replicate hybrid-composited data-set, Figures are response to selection (top 3 entries), absolute value for M1 then % decrease for other methods.

Results

	True	M1	mean M2	M3
σ_g^2	1.0	1.019	1.000	1.026
σ_{blk}^2	0.1	0.100	0.137	0.109
ρ_c	0.3	0.296	0.279	0.270
ρ_r	0.5	0.489	0.468	0.451
σ^2	1.0	1.006	0.984	0.986
		Response to Selection		
		1.699	4.0	1.0

Conclusions and Further Work

- Hybridized mixed models & embedded designs demonstrate significant gains in response to selection for expensive traits
- Improvements over standard such as one replicate and composite approaches of between 3 and 15% in relative response to selection
- Further research required to investigate efficient design and compositing strategies
- ASReml-R method to be developed for more user friendly interface