Embedded partially replicated field designs for grain quality testing

Alison Smith
alison.smith@industry.nsw.gov.au

Biometrics
New South Wales Industry and Investment

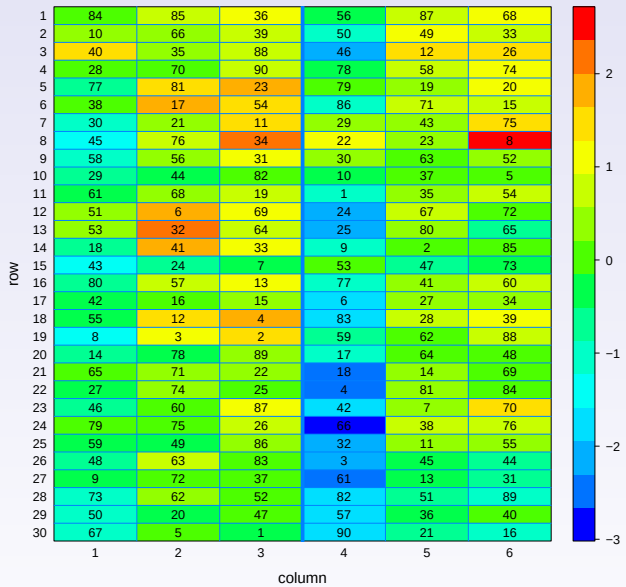
Embedded partially replicated designs

Collaborations and Acknowledgements

- This presentation is joint work with Brian Cullis (NSWI&I) and Robin Thompson (Rothamsted Research, UK).
- Thanks to Neil Coombes for helpful discussions and generation of designs in DiGGeR
- Thanks to Dave Butler for ASReml-R collaboration
- Thanks to Wallace Cowling and Cameron Beeck (CBWA) for data
- Grains Research and Development Corporation for financial support.

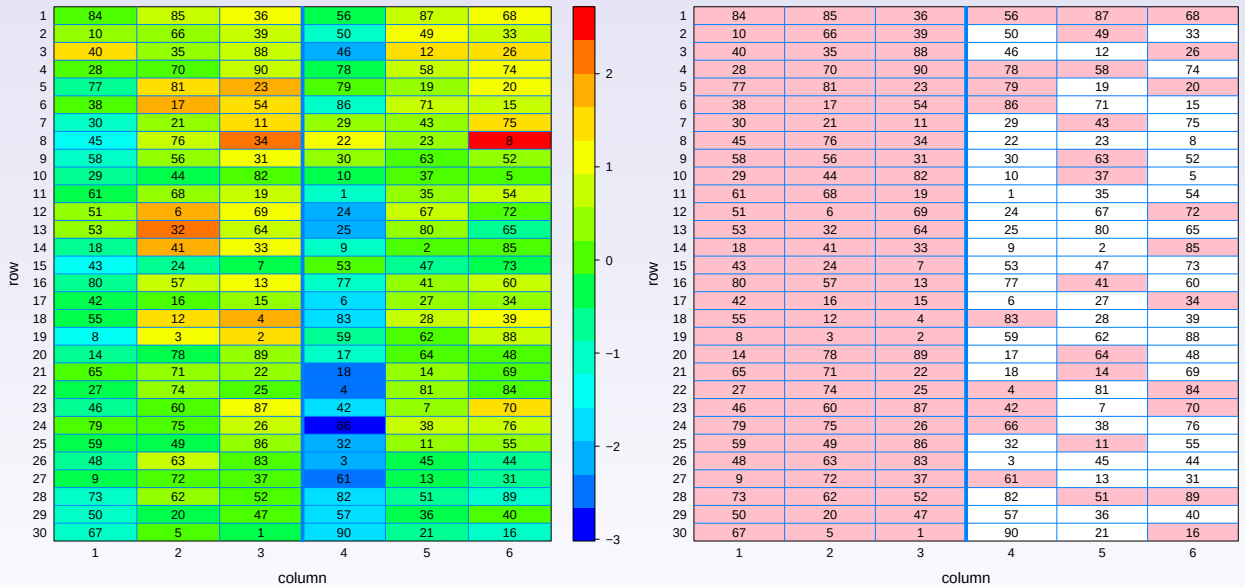
Executive summary

- Plant variety trials: many traits exhibit spatial trend
- Grain yield (relatively) inexpensive to measure so obtain data for all plots
- Many grain quality traits expensive so . . . ?



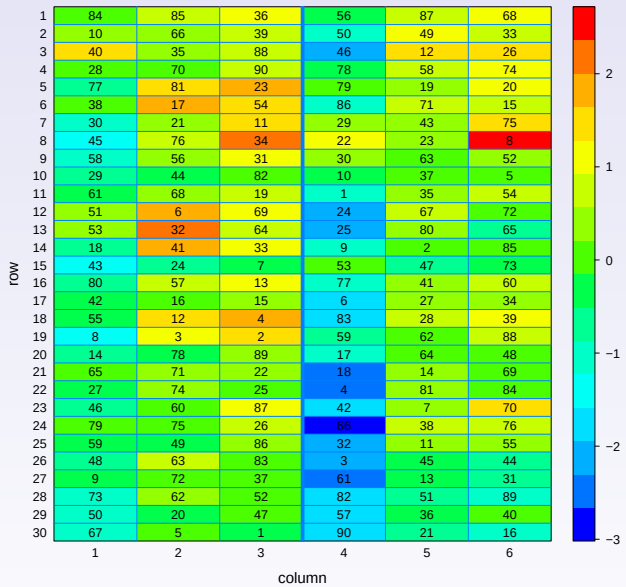
Executive summary

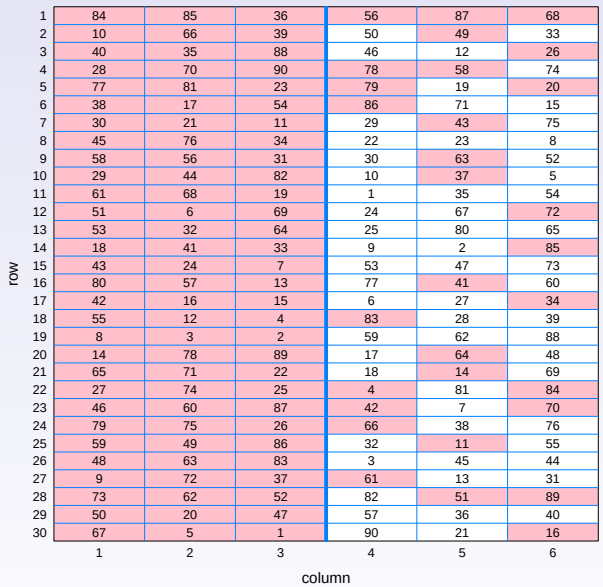
- Plant variety trials: many traits exhibit spatial trend
- Grain yield (relatively) inexpensive to measure so obtain data for all plots
- Many grain quality traits expensive so . . . ?

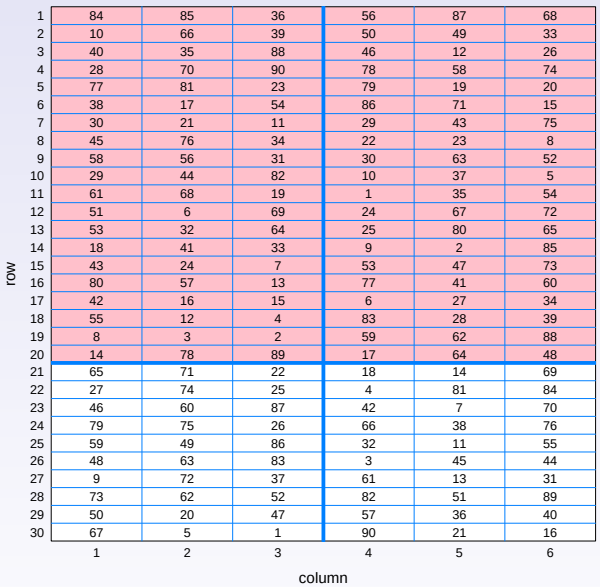


Executive summary

- Plant variety trials: many traits exhibit spatial trend
- Grain yield (relatively) inexpensive to measure so obtain data for all plots
- Many grain quality traits expensive so . . . ?







Outline of talk

- Motivation for work
- Mixed model analysis of some replicated grain quality data from a multi-environment trial (MET)
  - spatial trend?
  - G$\times$E?
- Description of embedded designs
- A tricky enhancement
- Simulation studies
- Conclusions and future work

Motivation for work

- Statistics for the Australian Grains Industry (SAGI) project (GRDC funded, led by BC).
- Support for plant breeding programs
- Support for National Variety Trials (NVT) system
- NVT generates (independent) information for growers on the performance of newly released crop varieties
- More than 600 trials sown at over 250 locations each year
- Crops tested are: Wheat; Barley; Triticale; Oat; Canola; Lupin; Lentil; Field Pea; Faba Bean and Chickpea
- Key economic traits measured include grain yield (GY), grain quality (GQ): protein, screenings, 1000 grain weight, oil content . . . and disease resistance
- Information made available to growers via NVT Online web-site

Motivation for work

Grain yield information

- Individual replicate data for each trial
- MET analysis: appropriate GxE model + spatial covariance models for errors
- Selection (breeder and farmer) based on best currently available estimates of variety performance . . .
- Large gains for Australian grains industry

Grain quality information

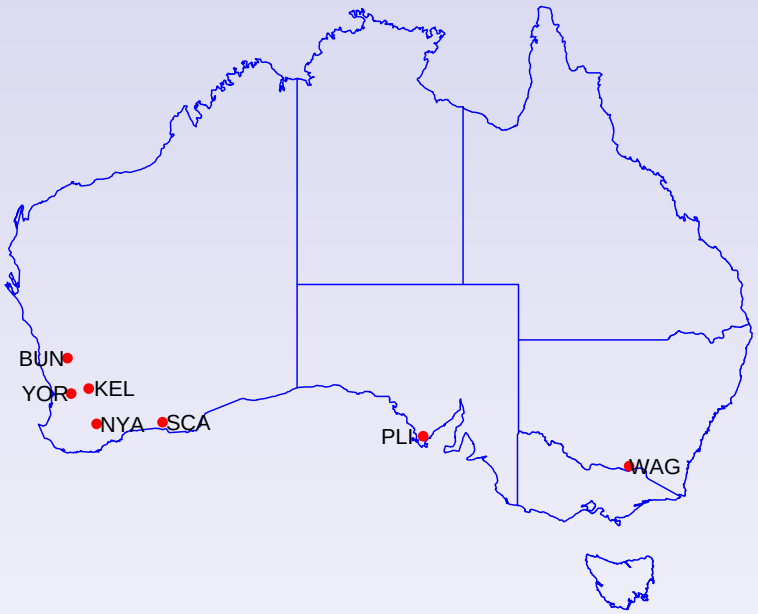
- Typically (including NVT) measured on single grain sample from each trial: composite from all replicates in trial
- Not possible to conduct spatial analysis; no MET analysis conducted
- Selection (breeder and farmer) based on raw data . . .
- Potentially large losses for Australian grains industry

Motivation for work

- Propose much to be gained by applying similar statistical approaches (design, replication and analysis) to GQ as has been done for GY for many years
- Common argument (especially NVT context): GQ can be very costly to measure so individual plot data prohibitively expensive.
- Gains depend on magnitude of spatial variation and G$\times$E for GQ traits
- Largely unknown since need replicated data to examine this. Very little exists - here is one example . . .

Example: replicated GQ data

7 trials from CBWA in 2007



Oil content data

- Total of 260 entries (breeding lines plus commercial varieties) across all 7 trials (between 213 and 260 per trial)
- $p-$ rep designs with either 1 or 2 plots per entry per trial (p varies between 0.22 and 0.35). Laid out as rows $\times$ columns
- 2g sample of grain from each plot $\Rightarrow$ NIR analysis for range of GQ traits including oil% $\Rightarrow$ data for 2148 plots

Single trial analysis

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{e}_j$$

- $\boldsymbol{y}_j$: data for j^{th} trial, ordered as rows within columns
- $\boldsymbol{\tau}_j$: fixed effects
- $\boldsymbol{u}_{\boldsymbol{g}_j}$: random variety effects
- $\boldsymbol{u}_{\boldsymbol{p}_j}$: random non-genetic (or peripheral) effects
- $\boldsymbol{e}_j$: residuals

Single trial analysis

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{e}_j$$

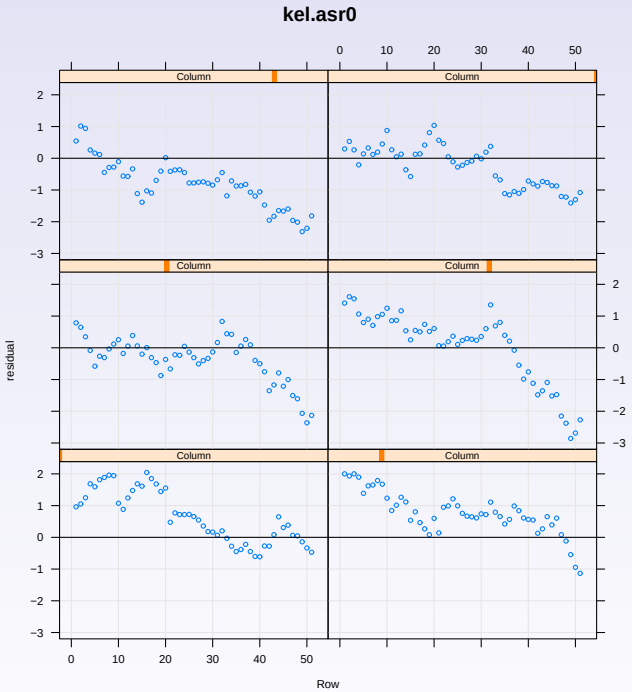
- Genetic model: for simplicity assume $\mathrm{var}\left(\boldsymbol{u}_{\boldsymbol{g}_j}\right) = \sigma^2_{g_j}\boldsymbol{I}_m$ where $\sigma^2_{g_j}$ is the genetic variance for trial j. Other genetic variance models possible (eg. use of relationship matrix).
- Error model: assume separable AR1$\times$AR1 spatial model so $\mathrm{var}\left(\boldsymbol{e}_j\right) = \boldsymbol{R}_j = \sigma^2_j\boldsymbol{\Sigma}_{\boldsymbol{c}j} \otimes \boldsymbol{\Sigma}_{\boldsymbol{r}j}$ where component matrices are correlation matrices for the column and row dimensions

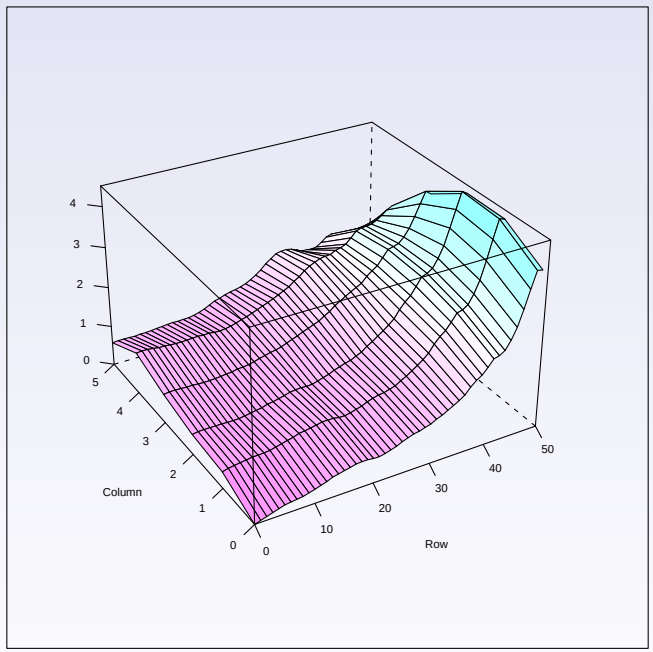
Oil data for KEL: analysis in ASReml-R

Base-line model

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{e}_j$$

kel.asr0 <- asreml(oil ~ 1, random = ~ Entry + Block, rcov = ~ ar1(Column):ar1(Row), data=kel.df)





Oil data for KEL: analysis in ASReml-R

Final model

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{e}_j$$

kel.asr1 <- asreml(oil $\sim$ 1 + linrow, random = $\sim$ Entry + Block + Column, rcov = $\sim$ ar1(Column):ar1(Row), data=kel.df)

Model term	Parameter estimate	Significance
linrow	-0.042	$p < 0.01$
Entry	1.414	–
Block	0.122	NFT
Column	0.257	$p < 0.01$
Residual variance	0.389	–
Row autocorrelation	0.58	$p < 0.01$
Column autocorrelation	0.27	$p > 0.10$

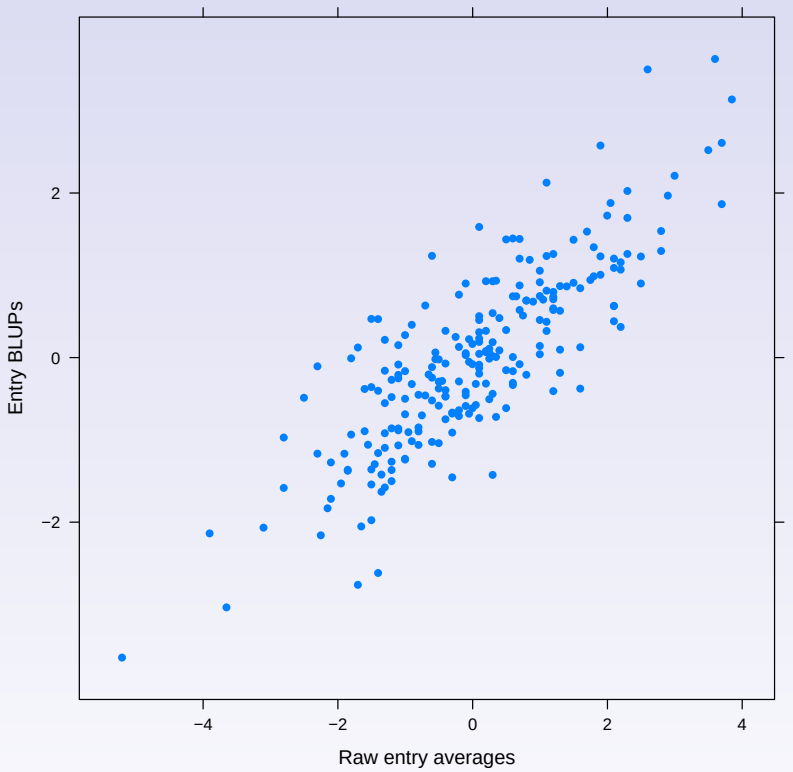
Oil data for KEL: analysis in ASReml-R

Entry predictions

Entry BLUPs vs raw averages

Assessing entry performance

- Mixed model approach: use Best Linear Unbiased Predictions (BLUPs) of $\boldsymbol{u}_{\boldsymbol{g}_j}$
- Standard composite approach: use raw value for composite sample. Analogy here is average of raw replicate data



Multi-environment trial analysis

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g u_g} + \boldsymbol{Z_p u_p} + \boldsymbol{e}$$

- $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots \boldsymbol{y}_t')'$
- $\boldsymbol{\tau} = (\boldsymbol{\tau}_1', \boldsymbol{\tau}_2', \ldots \boldsymbol{\tau}_t')'$
- $\boldsymbol{u_g} = (\boldsymbol{u_{g_1}}', \boldsymbol{u_{g_2}}', \ldots \boldsymbol{u_{g_t}}')'$
- $\boldsymbol{u_p} = (\boldsymbol{u_{p_1}}', \boldsymbol{u_{p_2}}', \ldots \boldsymbol{u_{p_t}}')'$
- $\boldsymbol{e} = (\boldsymbol{e}_1', \boldsymbol{e}_2', \ldots \boldsymbol{e}_t')'$

Multi-environment trial analysis

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g u_g} + \boldsymbol{Z_p u_p} + \boldsymbol{e}$$

- Genetic model:
  - $\mathrm{var}\left(\boldsymbol{u_g}\right) = \boldsymbol{G_e} \otimes \boldsymbol{I}_m$ where $\boldsymbol{G_e}$ is $t \times t$ genetic variance matrix
  - Use Factor Analytic (FA) model so $\boldsymbol{G_e} = \boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi}$ where $\boldsymbol{\Lambda}$ is $t \times k$ matrix of trial loadings (for k factors) and $\boldsymbol{\Psi}$ is $t \times t$ diagonal matrix of trial specific variances
- Error model:
  - $\mathrm{var}\left(\boldsymbol{e}\right) = \boldsymbol{R} = \mathrm{diag}\left(\boldsymbol{R}_j\right)$
  - Use separable AR1$\times$AR1 spatial model for each trial

MET analysis of oil data

Final model: non-genetic effects

	Fixed effects			Variances		Autocorrelations	
	Trial	linrow	block	column	residual	column	row
BUN	38.2	0.012	0.104		0.326	0.13	0.39
KEL	43.9	-0.044	0.087	0.306	0.377	0.20	0.45
NYA	40.6	-0.018	0.082		0.594	0.14	0.59
PLI	45.9		0.000		1.282	0.35	0.78
SCA	45.7		0.271		2.217	0.26	0.61
WAG	38.9		0.000	0.123	0.478	0.27	0.55
YOR	47.6		0.000		0.707	0.21	0.56

MET analysis of oil data

Final model: genetic effects

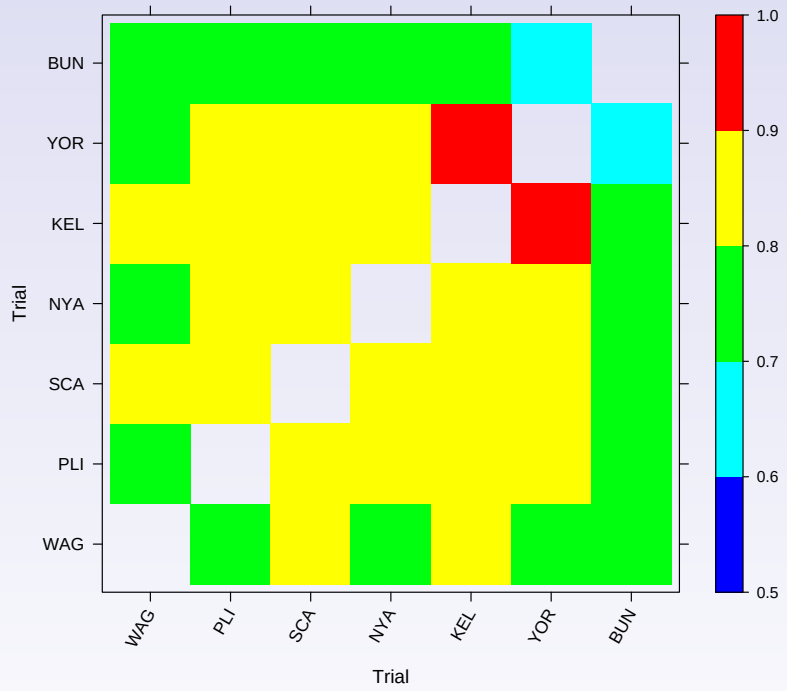
Genetic variance matrix: FA2 parameters

$$G_e = \Lambda\Lambda' + \Psi$$

	λ_1	λ_2	Ψ
BUN	0.762	0.258	0.296
KEL	1.163	0.000	0.076
NYA	0.986	-0.003	0.201
PLI	1.463	-0.040	0.395
SCA	1.259	0.278	0.214
WAG	0.868	0.410	0.072
YOR	1.284	-0.272	0.100

Genetic correlation matrix

$$G_e = D^{1/2} C_e D^{1/2}$$



Motivation for work

- There is spatial variation and G$\times$E in oil% in this data-set
- Other replicated GQ data-sets suggest varying degrees of spatial and G$\times$E
- There is evidence to support need for replicated GQ data followed by proper statistical analysis
- Recall: may be too expensive (particularly NVT) to GQ test all plots
- What follows is a solution for mid-late stage testing where fully replicated designs are used.

Embedded p-rep designs

- Consider trial with 2 replicates of 90 entries laid out as 6 columns by 30 rows. Block (rep) 1 = columns 1-3; block 2 = columns 4-6.
- For GY obtain data on 2 reps for all entries
- For GQ use Cullis *et al* (2006) p-rep design idea so test a proportion (eg. $p = 1/3$) of entries as 2 plots and remainder as 1
- *Do not* generate a replicated and efficient design for GY then later sample spatially disjoint set of plots for GQ
- *Do* generate (a priori) an efficient and contiguous design for GQ embedded within a replicated and efficient design for GY
- Our layout: choose top 20 rows by 6 columns for embedded part

Embedded p-rep designs

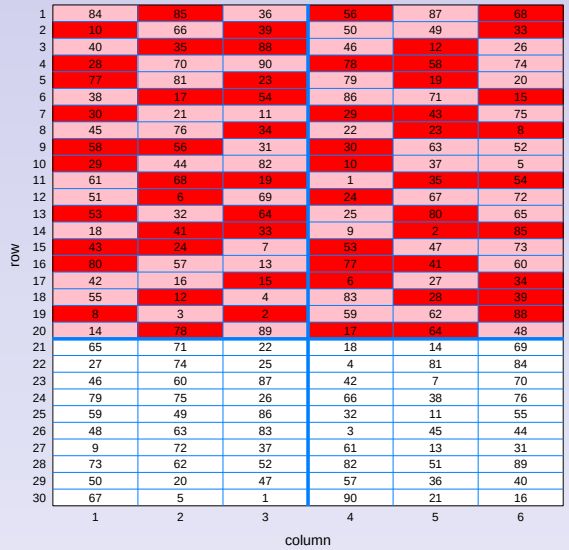
A partially replicated design contained within an RCB design

row	1	2	3	4	5	6
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	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

column

Plots shaded red correspond to replicated entries in embedded p-rep portion.

Embedded p-rep designs



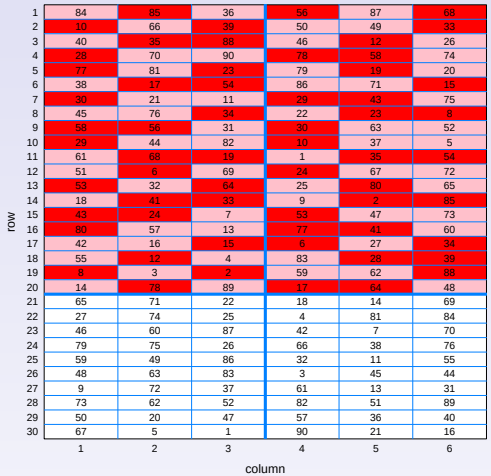
- Optimisation process is sequential commencing with the p-rep design, followed by formation of the RCB design conditional on the p-rep design embedded within it
- Each design search is undertaken using algorithm that minimises pre-specified objective function (typically the A-value) for chosen blocking and spatial correlation models. All with DiGGeR (thanks Neil)
- May be some loss of efficiency for “outer” (RCB) design

Embedded p-rep designs

A further enhancement

Standard approach

- Test 120 samples (from 120 plots) for GQ (and 180 for GY)
  - 30 entries with 2 samples (plots)
  - 60 entries with single sample (plot)
- Allows modelling of
  - spatial trend in p-rep section
  - G$\times$E



Embedded p-rep designs

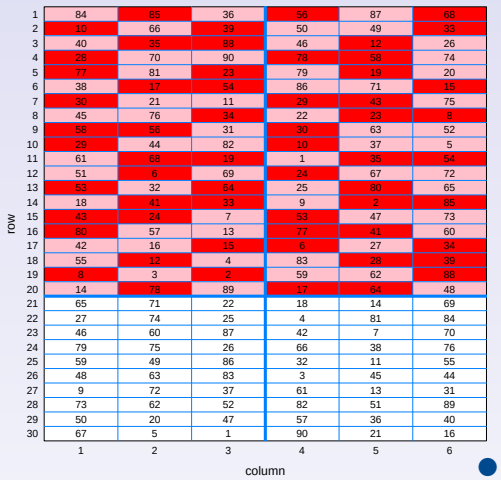
A further enhancement

Standard approach

- Test 120 samples (from 120 plots) for GQ (and 180 for GY)
  - 30 entries with 2 samples (plots)
  - 60 entries with single sample (plot)
- Allows modelling of
  - spatial trend in p-rep section
  - G$\times$E

Partial compositing

- Test 120 samples (from 180 plots) for GQ (and 180 for GY)
  - 30 entries with 2 samples (plots)
  - 60 entries with single sample (composite of 2 plots)
- Allows modelling of
  - spatial trend across whole trial (tricky!)
  - G$\times$E



Single trial analysis

Partially composite data

$$\boldsymbol{D}_j\boldsymbol{y}_j = \boldsymbol{D}_j\boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{D}_j\boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{D}_j\boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{D}_j\boldsymbol{e}_j$$

- Data has been "averaged" commensurate with compositing process, ie. started with n_j plots and have reduced to s_j samples (a mixture of composites and individual replicates)
- $\boldsymbol{D}_j$ is $s_j \times n_j$ averaging matrix
- Our example: $n_j = 180,\ s_j = 120$ and we have 60 samples that are composites of 2 replicates and 60 that are individual replicates
- Model involves non-standard design matrices: use "grp" facility in ASReml-R

Single trial analysis

Partially composite data

- Model fitted to real KEL data (individual reps only):

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{e}_j$$

kel.asr1 <- asreml(oil ∼ 1 + linrow,
random = ∼ Entry + Block +
Column, rcov = ∼ ar1(Column):ar1(Row), data=kel.df)

- Now assume partially composite data and fit same model

$$\boldsymbol{D}_j\boldsymbol{y}_j = \boldsymbol{D}_j\boldsymbol{X}_j\boldsymbol{\tau}_j + \boldsymbol{D}_j\boldsymbol{Z}_{\boldsymbol{g}_j}\boldsymbol{u}_{\boldsymbol{g}_j} + \boldsymbol{D}_j\boldsymbol{Z}_{\boldsymbol{p}_j}\boldsymbol{u}_{\boldsymbol{p}_j} + \boldsymbol{D}_j\boldsymbol{e}_j$$

kelpc.asr <- asreml(oil ∼ 1 + linrow,
random = ∼ Entry + grp('Block') + grp('Column') +
str(∼('Plot'), ∼ar1v(6):ar1(30)),
family = asreml.gaussian(dispersion=0.0001), data=kelpc.df,
control=asreml.control(group=list(Block=184:185,Column=186:191,Plot=4:183))

Single trial analysis

Partially composite data

- How reliable is spatial analysis of partially composite data?
- Simulation study based on model (for KEL) and layout as described. Generate data for all 180 plots then fit true model to:
  - Full: full data-set of 180 plots
  - Embed: sub-set of 120 plots corresponding to embedded design (ie. top 20 rows)
  - Pcomp: partially composite data (120 samples = 60 individual plot samples and 60 composite)

Single trial analysis of partially composite data

Simulation study results

	Parameter value			
Model term	True	Full*	Embed*	Pcomp*
linrow	-0.044	-0.045	-0.046	-0.046
Entry	1.429	1.412	1.414	1.412
Block	0.087	0.105	0.114	0.112
Column	0.306	0.284	0.270	0.270
Residual variance	0.377	0.377	0.396	0.395
Row autocorrelation	0.450	0.436	0.432	0.387
Column autocorrelation	0.200	0.199	0.187	0.181

* Means over 400 simulations

MET analysis of partially composite data

- So we *can* conduct spatial analysis of partially composite data
- Can extend to MET analysis (trivial ASReml-R code!??)
- How does embedded concept perform compared with fully replicated design in terms of response to selection (genetic gain)?
- Simulation study based on parameters from MET model for real oil example but 2 replicate $\times$ 90 entries (30 row $\times$ 6 column) layout as described.

MET analysis of partially composite data

Simulation study results

Methods compared

- M1: true model fitted to full data-set
- M2: true model fitted to embedded data-set
- M3: true model fitted to partially composited data-set
- M4: “best possible” model fitted to fully composited data-set and
- M5: raw fully composited data

MET analysis of partially composite data

Simulation study results

Methods compared

- M1: true model fitted to full data-set
- M2: true model fitted to embedded data-set
- M3: true model fitted to partially composited data-set
- M4: “best possible” model fitted to fully composited data-set and
- M5: raw fully composited data

Results

- Figures are response to selection (top 5 entries), absolute value for M1 then % decrease for other methods.

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

Conclusions and Further Work

- Using individual replicate data and proper statistical analysis for GQ traits likely to lead to superior estimates of variety performance and information on G$\times$E
- GQ traits often measured on samples from fully replicated trials
- If cost not limiting test all plots, otherwise . . .
- The embedded p-rep approach (particularly with partial compositing) provides statistically and economically efficient solution

Conclusions and Further Work

- Further research required to investigate efficient design and compositing strategies (especially for NVT with 3 replicate trials)
- ASReml-R method to be developed for more user friendly interface
- Paper submitted to Applied Statistics (JRSS Series C)