Improving the Accuracy of Genetic Predictions for Expensive Multi-Phase Traits

Daniel Tolhurst, Alison Smith & Brian Cullis

NIASRA, University of Wollongong, Australia

Biometrics by the Harbour

The International Biometric Society - Australasian Region

SAGI NIASRA NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA UNIVERSITY OF WOLLONGONG GRDC Grains Research & Development Corporation Your GRDC working with you

Collaborations and Acknowledgements

- This presentation is joint work through the Statistics for the Australian Grains Industry (SAGI) project funded by Grains Research and Development Corporation (GRDC).
- Thanks to Steve Jefferies (AGT) for providing stimulating ideas and collaboration in the current project.

SAGI

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

GRDC Grains Research & Development Corporation
Your GRDC working with you

Motivating Example

Australian wheat quality project

- 11 wheat variety trials located across Australian growing regions in QLD, NSW, SA and WA.
- Aim is to obtain accurate estimates of variety effects and investigate variety by trial interaction in wheat quality traits.
- Data is collected on a specified set of varieties for a range of traits, including grain, flour, dough, noodle and baking characteristics. Key example here is dough strength.

Wheat quality traits

multi-phase experiments

- Data on all traits are obtained from multi-phase experiments with either 2 or 3 phases.
- **First phase** for all traits is a designed field experiment, where plots are harvested to produce bags of grain.
- **Second phase** involves trait dependent processes in a laboratory using bags of grain from phase 1.
- The transition to phase 3 is characterised by the milling of grain into flour using a Buhler mill.
- **Third phase** use samples of flour from phase 2 and trait dependent processing.
- Dough strength is a 3 phase trait; where the third phase requires use of an Extensograph.

Compositing of field plots

Reducing replication from the field phase

- Smith *et. al.* (2006) and Brien *et al.* (2011) showed that non-genetic variation variation can have a substantial impact on many wheat quality traits and so valid experimental design techniques (including replication and randomisation) are required in every phase. However…
- Full replication in all three phases is not possible due to budgetary and time constraints.
- Consequently, field plots are composited according to the methods of Smith *et al.* (2015) in order to reduce replication from the field.
- In particular, a proportion of varieties are composited while the remainder are fully replicated. This ensures that information from all plots is used.

Compositing of field plots

Reducing replication from the field phase

- Consider a typical field trial in our study (say *Trial A*) comprising three replicates (r1, r2 and r3) of 18 varieties (54 plots in total). Smith *et al.* (2015) suggest a scheme in which there is 3 types of compositing:
    - **T3** composite of all 3 field replicate plots, r1 + r2 + r3
    - **T2** composite of 2 field plots, r1+r2 or r1+r3 or r2+r3
    - **T0** individual field replicate plot (no compositing). r1 or r2 or r3
- Then, based on cost and laboratory throughput limits were set at
    - **Phase 1**: 33 bags of grain,
      (54 → 33 bags achieved via compositing)
    - **Phase 2**: 40 samples of grain,
      (33 → 40 samples via duplicating 7 bags from field and process separately)
    - **Phase 3**: 45 samples of flour
      (40 → 45 samples achieved by milling 5 grain samples at double weight, which are then split for separate processing).

Laboratory design for dough strength (Rmax)

Model-based design of Trial A (OD)

- Standard techniques are invalid because designs are non-orthogonal; owing to
  - compositing of field plots in phase 1,
  - limited replication during laboratory phases.

- Despite this, efficient model-based design (MBD) was constructed using optimal design (OD). This involves specifying a LMM, where sources of variation specified match those in the subsequent analysis exactly.

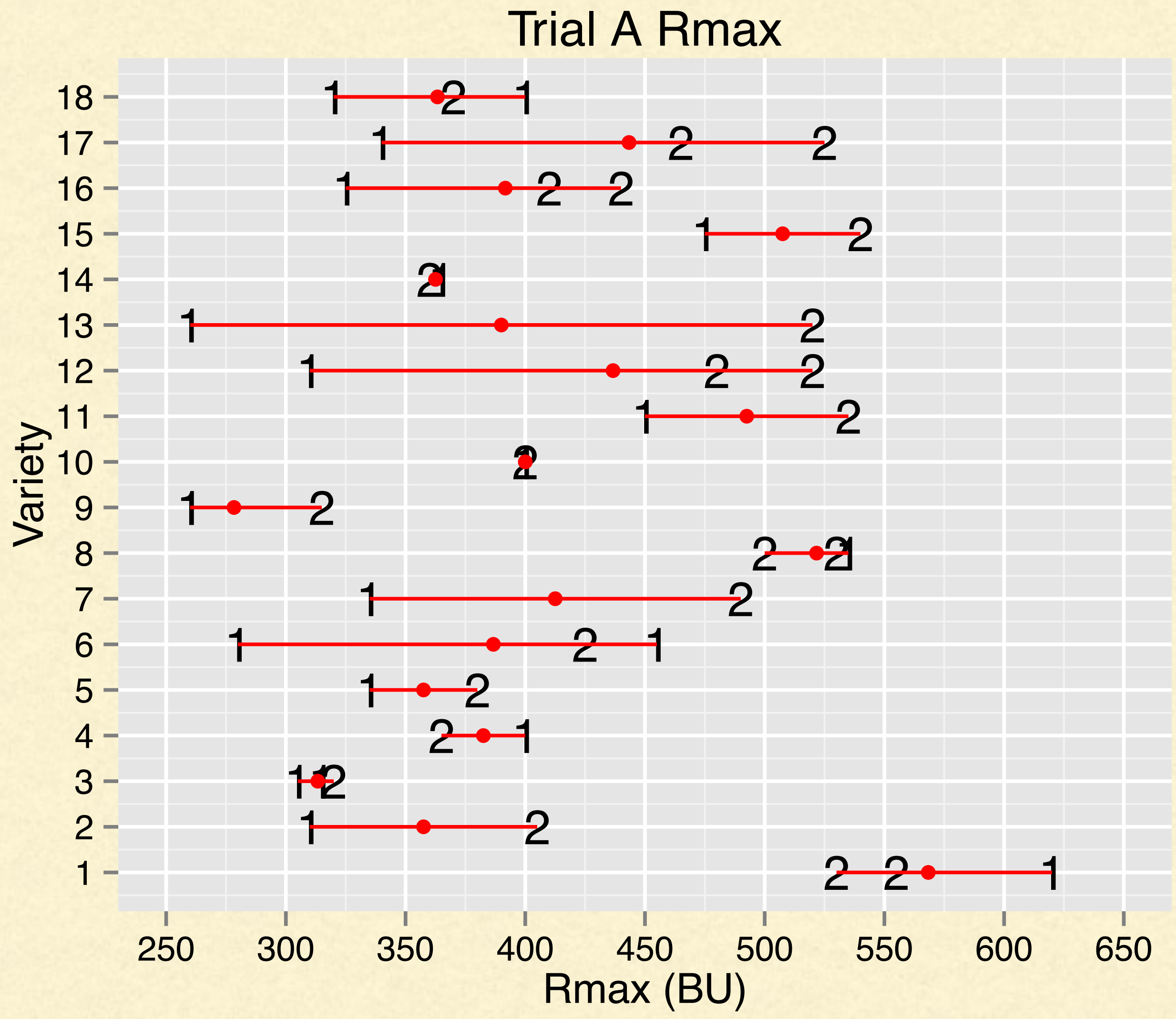
- In particular, LMM in OD contain random effects for
  - varieties,
  - field blocks and plots (phase 1),
  - mills, milling days (MDay) and MDay.MOrdWD (phase 2),
  - extensograph replicate blocks and processing days (phase 3).

MOrdWD- milling order within days

Exploratory analysis of dough strength (Rmax)

Single Site analysis of Trial A

- Genetic variation: between varieties.
- Non-genetic: variation in the field, milling and extensograph processing.
- Raw data suggests substantial non-genetic variation at this trial, but from where?



- Data points labelled according to mill number (1 or 2) used during flour extraction.

LMM Analysis of dough strength (Rmax)

Single Site analysis of Trial A (ASReml-R): specification of field effects

- Specification of field effects with compositing requires non-standard design matrices to enable averaging of effects from composited plots. This requires factors *P1-P6* (54 levels corresponding to plots in Trial A) to be coded in the data-frame according to, for example,

Variety	Type	No. parent plots	P1	P2	P3	P4	P5	P6	Average plot effects
1	T0	1	C1R10	C1R10	C1R10	C1R10	C1R10	C1R10	$1/6(u_{C1R10}+u_{C1R10}+u_{C1R10}+u_{C1R10}+u_{C1R10}+u_{C1R10})$
1	T2	2	C2R5	C2R5	C2R5	C3R18	C3R18	C3R18	$1/6(u_{C2R5}+u_{C2R5}+u_{C2R5}+u_{C3R18}+u_{C3R18}+u_{C3R18})$
2	T3	3	C1R15	C1R15	C2R10	C2R10	C3R1	C3R1	$1/6(u_{C1R15}+u_{C1R15}+u_{C2R10}+u_{C2R1}+u_{C3R1}+u_{C3R1})$

- Syntax in ASReml-R then uses the "and" constructor function. The plot effects are specified as:

str(P1:zero + and(P1, 0.166667) + and(P2, 0.166667) + and(P3, 0.166667) + and(P4, 0.166667) + and(P5, 0.166667) + and(P6, 0.166667), ~idv(P1))

- *P1:zero* is a 45x54 zero matrix that ASReml requires for initial setup of the design matrix.

- Similar process applied for the block effects.

LMM Analysis of dough strength (Rmax)

Single Site analysis of Trial A (ASReml-R)

- Syntax for analysis in ASReml-R with Compositing (recall, sources of variation match those included during construction of MBD):

```
Rmax.asr <- asreml(Rmax  ~ 1, random = ~Variety +
        str(B1:zero + and(B1, 0.166667) + and(B2, 0.166667) + and(B3, 0.166667) +
          and(B4, 0.166667) + and(B5, 0.166667) + and(B6, 0.166667), ~idv(B1)) +
        str(P1:zero + and(P1, 0.166667) + and(P2, 0.166667) + and(P3, 0.166667) +
          and(P4, 0.166667) + and(P5, 0.166667) + and(P6, 0.166667), ~idv(P1)) +
        Mill + MDay + MDay:MOrdWD + EBlock + F9Day,
        rcov =  ~ F9Day:F9OrdWD, data = TrialA.df)
```

- Genetic effects: *Variety*.
- Non-genetic effects:
    - Phase 1 sources: field replicate blocks and plots,
    - Phase 2 sources: *Mills*, milling days (*MDays*) and *MDay:MOrdWD*,
    - Phase 3 sources: extensograph replicate blocks (*EBlock*), processing days (*F9Day*) and order within days (*F9OrdWD*).
- Residual effects from phases 1 and 2 must be included.

LMM Analysis of dough strength (Rmax)

Single Site analysis of Trial A (ASReml-R): sources of variation

Source	Variance Component
Mean	
Phase 3	
F9Day	86
EBlock	177
Phase 2	
Mill	2172
MDay	1929
Phase 1	
Block	16
Variety	4737
Phase 1 residual (Plots)	1823
Phase 2 residual (MDay.MOrdWD)	165
Residual	202

REML estimates of variance components.

Simulation study of dough strength (Rmax)

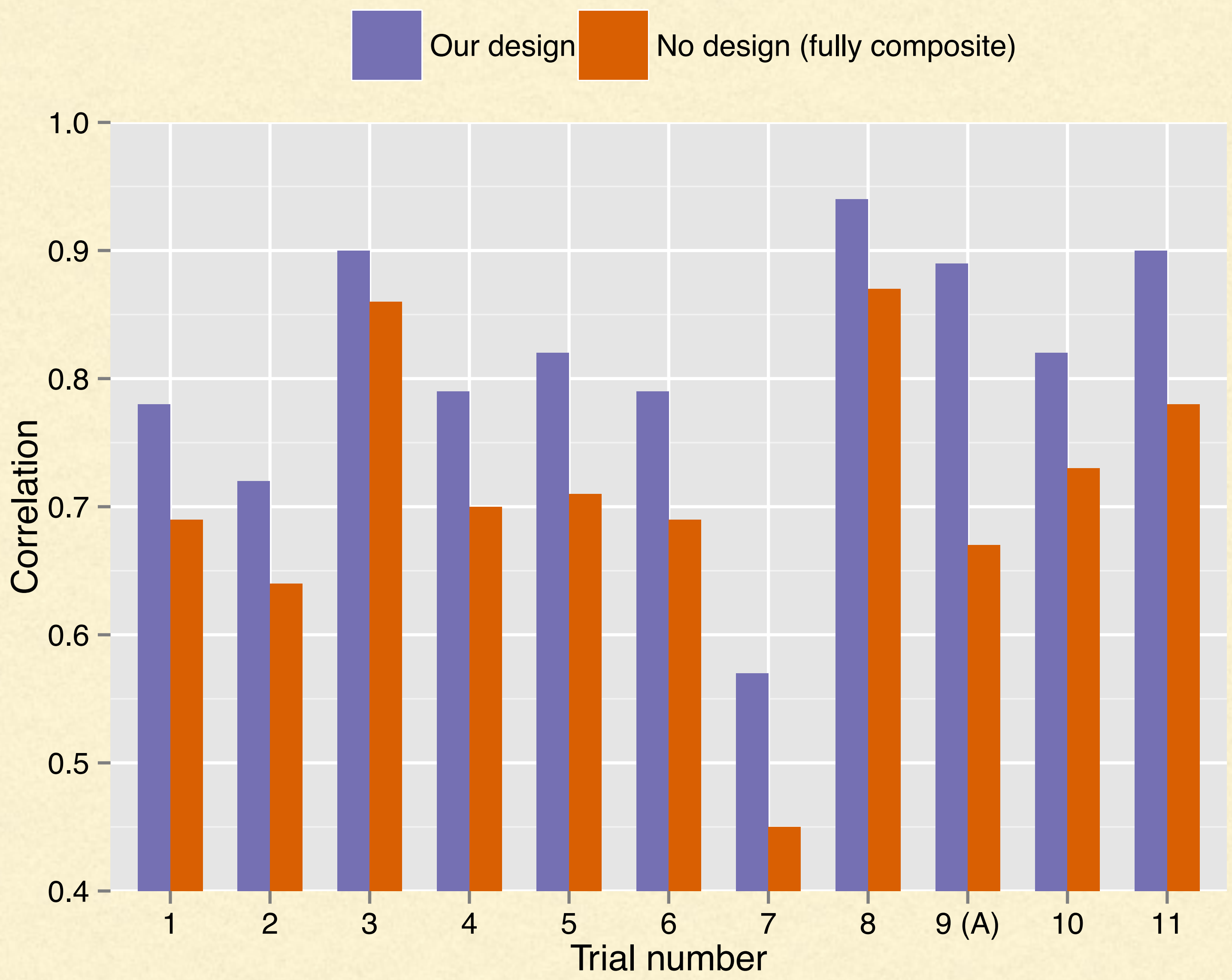
Accuracy of variety predictions for each trial

- For each of *N* simulations:
  - Generate effects for each source of variation (based on estimates of variance components from LMM analysis of "real" data for that trial).
  - Form two types of simulated data:
    - Form data corresponding to actual design, so 45 samples based on our compositing and replication scheme. Then LMM analysis to obtain a set of variety predictions,
    - Form data corresponding to "no design" so a fully composite sample for each variety (with no replication). Then variety predictions are "raw" data (corrected for mean).
- Accuracy for each variety within a trial measured as correlation between the *N* true (generated) and predicted variety effects.
- Accuracy reported for each trial is then the average of correlations across varieties.

Simulation study of dough strength (Rmax)

Accuracy of variety predictions for each trial

- Substantial gains in accuracy using the compositing scheme of Smith *et al.* (2015) together with model-based design and subsequent LMM analysis.



MET analysis of wheat quality traits

- Efficient one-stage multi-environment trial analysis conducted on full set of 11 trials.
- First time this type of analysis has been conducted on wheat quality data involving valid model based design (with replication and randomisation at every phase).
- Separate MET fitted for each trait (11 in total) and produced
    - Accurate estimation of variety effects including an informative model for the variety by trial interaction.

Concluding remarks 1

- The wheat quality project called for pragmatic designs, in the sense that strict sample limits were imposed according to laboratory throughput and cost.
- Historically, both field and laboratory replication would be sacrificed in order to satisfy such restrictions.
- Recent work has produced methods for reducing replication in the field and laboratory while staying true to experimental design (see 'p-rep' of Cullis *et al.* 2006, 'p-q-r' of Smith *et al.* 2006 and compositing scheme of Smith *et al.* 2015).
- The compositing approach of Smith *et al.* (2015) has achieved substantial improvements in the accuracy of genetic predictions compared to traditional testing methods (i.e. single composite samples).

Concluding remarks 2

- For full account of the statistical methodology see:
  - Smith, A. B. , Butler, D. G., Cavanagh, C. , & Cullis, B. R. (2015). The design and analysis of multi-phase variety trials using both composite and individual replicate samples. *Journal of Agricultural Science*, 153, pp 1017–1029.
- The wheat quality project is on-going with numerous interesting statistical problems still to explore.

References

- Brien, C., Harch, B., Correll, R., & Bailey, R. (2011). Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *Journal of Agricultural, Biological and Environmental Statistics* 16, pp422–450.
- Smith, A. B., Lim, P., & Cullis, B. R. (2006). The design and analysis of multi-phase plant breeding experiments. *Journal of Agricultural Science, Cambridge* 144, pp393–409.
- Smith, A. B. , Butler, D. G., Cavanagh, C. , & Cullis, B. R. (2015). The design and analysis of multi-phase variety trials using both composite and individual replicate samples. *Journal of Agricultural Science* 153, pp1017–1029.