CENTRE FOR BIOINFORMATICS AND BIOMETRICS

Pedigree *p*-rep Designs:

A class of designs for early stage variety trials

Nicole Cocks, Brian Cullis, Alison Smith and Dave Butler

AASC 2018, December 5

Acknowledgements

CBB gratefully acknowledges the Grains Research and Development Corporation (GRDC) for their funding, without which we would not be able to contribute to these projects and many others with the statistical support we have provided. We would also like to thank Jeff Paull for his data that provided the motivating example for this seminar.

Overview

Breeders Equation

$$RGG = \frac{i\sigma_g r}{t}$$

- RGG: Rate of Genetic Gain
- i : selection intensity
- σ_g^2 : genetic variance of trait (yield)
- r : accuracy of selection - the correlation between the true and predicted genetic effects. Note that r^2 is the reliability (Mrode 1995)
- t : generation interval

Breeders Equation

$$RGG = \frac{i\sigma_g r}{t}$$

- For fixed i, σ_g^2 and t, an increase in the accuracy of EBLUPs increases the RGG
- This requires implementing procedures that are (and remain) best practice
- We can contribute statistically by
  - Optimal experimental design
  - Appropriate construction of a MET dataset (contemporary groups, co-located trials, etc.)
  - Sophisticated analysis of the dataset (Smith et. al. 2014)
  - Summary and dissemination of results (selection tools (Smith & Cullis 2018), etc.)

p-rep Designs

- Since Cullis et. al. (2006), adoption of *p*-rep designs in plant breeding and crop research has replaced traditional grid-plot designs
- The 25% *p*-rep threshold advocated by Cullis et. al. (2006) was based on the ratio of test plots to grid plots in traditional grid-plot designs
- Cullis et. al. (2006) stated "At present pedigrees are not generally used in routine analyses of EGVTs"
- From 2017 we incorporated pedigree data in MET analysis of PBA breeding programs

Pedigree p-rep designs

- Historically, the design of PBA early generation variety trials (EGVT) were generated in-house using DiGGer (Coombes 2002)
- From 2017 we transitioned to od (Butler & Cullis 2018) and exploited known genetic relationships for the allocation of varieties to both plots and sites
- This gives rise to two areas of research
  - Is the 25% level of p-rep necessary to achieve a given level of accuracy (r)
  - Is there an advantage to using pedigree information in the design of EGVT

Outline of the simulation study

- A simulation study was designed to primarily investigate
  - Various levels of p-rep
  - The advantage (if any) of including pedigree information

  in the design of the PBA Southern Faba Bean S1 trial that consisted of 256 breeding lines and 4 check varieties
- The random component of the linear mixed model for the analysis of a single PBA trial is:
  $\texttt{random} \sim \texttt{vm(Line, A)} + \texttt{ide(Line)} + \texttt{Block} + \texttt{Column} + \texttt{Row}$
  $\texttt{residual} \sim \texttt{ar1(Column)} : \texttt{ar1(Row)}$
- Where $\mathbf{A}$ is the (pedigree derived) relationship matrix
- Variance parameters: $\sigma_g^2 = \bar{a}\sigma_a^2 + \sigma_i^2$, σ_b^2, σ_c^2, σ_r^2, ρ_c and ρ_r

Outline of the simulation study

- Simulation treatment factors
  - 7 levels of p-rep: $p = 0, 5, 10, 15, 25, 50, 100\%$
  - 3 levels of proportion of additive genetic variance to the total: $k = 0.5, 0.7, 0.9$
  - 3 levels of null (baseline) reliability: $r_0^2 = 0.33, 0.5, 0.66$
  - 3 design types $d = od_{od}, od_{gg}, od_{\alpha}$
    - od_{od} *od* design with pedigree information included Butler and Cullis (2018)
    - od_{gg} DiGGer style row-column design Coombes (2002)
    - od_{α} augmented α design for single site Williams et. al. (2011)
- $7 \times 3 \times 3 = 63$ treatments $\{T\} \times 3$ designs

CENTRE FOR BIOINFORMATICS AND BIOMETRICS

Outline of the simulation study

- For each $\{T_i\}$ generate n=4000 datasets
- Values of non-genetic variance components chosen from previous analyses of data
- σ_a^2 and σ_i^2 chosen to realise pre-specified values of r_0^2 for a design with no replication and sub-optimal allocation of varieties to plots
- $T_{(id)_j} \rightarrow$ the j^{th} simulation allocated to plots by design strategy d

Outline of the experiment designs

- With reference to the LMM used for analyses, the following indicates a) if a term was fitted and b) whether it was fitted as fixed or random for each of the three design types d

Term	od_{od}	od_{gg}	od_{α}
`Line`		F	F
`Additive`	R		
`Non-Additive`	R		
`Block`	R	R	F
`Column`	R	R	
`Row`	R	R	
`Column:Row (Plot)`	R	R	

Simulations

- The simulated data from each $\{T_{(id)_j}\}$ were analysed in ASReml-R, Butler et. al. (2018).
- For T_i we have

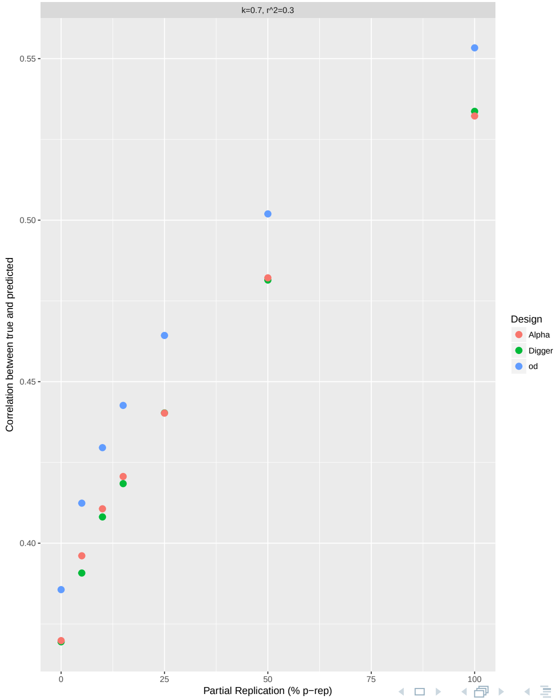
$$\mathbf{u}_g, \mathbf{u}_a, \mathbf{u}_i$$

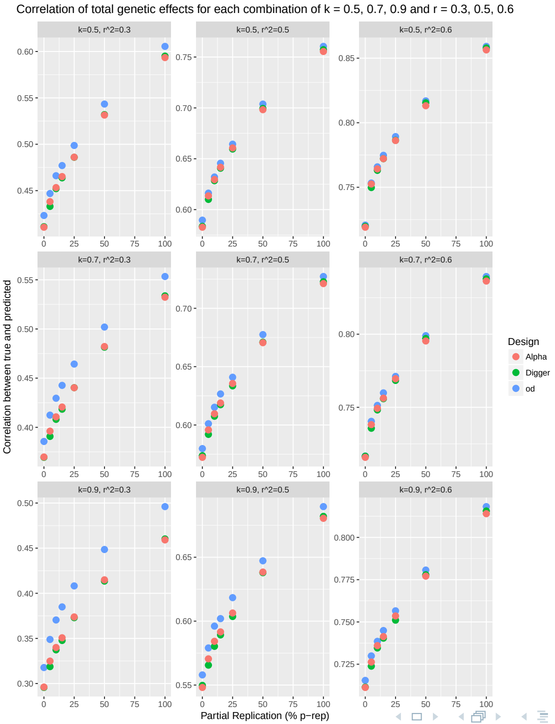
  and for $T_{(id)}$

$$\tilde{\mathbf{u}}_{gd}, \tilde{\mathbf{u}}_{ad}, \tilde{\mathbf{u}}_{id}$$

  For d in $\{od_{od}, od_{gg}, od_{\alpha}\}$, where each $\mathbf{u}$ is of length 260.
- Summarise
  - Bias associated with variance parameter estimates for each design method d for the 63 combinations of p, k and r_0^2
  - Correlation between the "true" (simulated) value and corresponding predicted value for for each design method d and 63 combinations of p, k and r_0^2

Correlation of total genetic effects $k = 0.7$ and $r^2 = 0.3$




Correlation of total genetic effects for each combination of k = 0.5, 0.7, 0.9 and r = 0.3, 0.5, 0.6
k=0.5, r^2=0.3
k=0.5, r^2=0.5
k=0.5, r^2=0.6
k=0.7, r^2=0.3
k=0.7, r^2=0.5
k=0.7, r^2=0.6
k=0.9, r^2=0.3
k=0.9, r^2=0.5
k=0.9, r^2=0.6
Correlation between true and predicted
Partial Replication (% p−rep)
Design
Alpha
Digger
od

Correlation between true and predicted: Total

- There is little to no distinction between the performance of od_{gg} and od_{α} allocations for any level of p, k, and r_0^2 for total correlations
- This suggests that spatial models (i.e. fitting AR1$\times$AR1 to the residual) are not as critical in the design of trials such as these
- As k increases for fixed p and r_0^2 the advantage of od_{od} to od_{gg} and od_{α} designs increases
- For fixed r_0^2 and p, as k increases, the total correlation decreases as the bias associated with the non-additive genetic variance increases

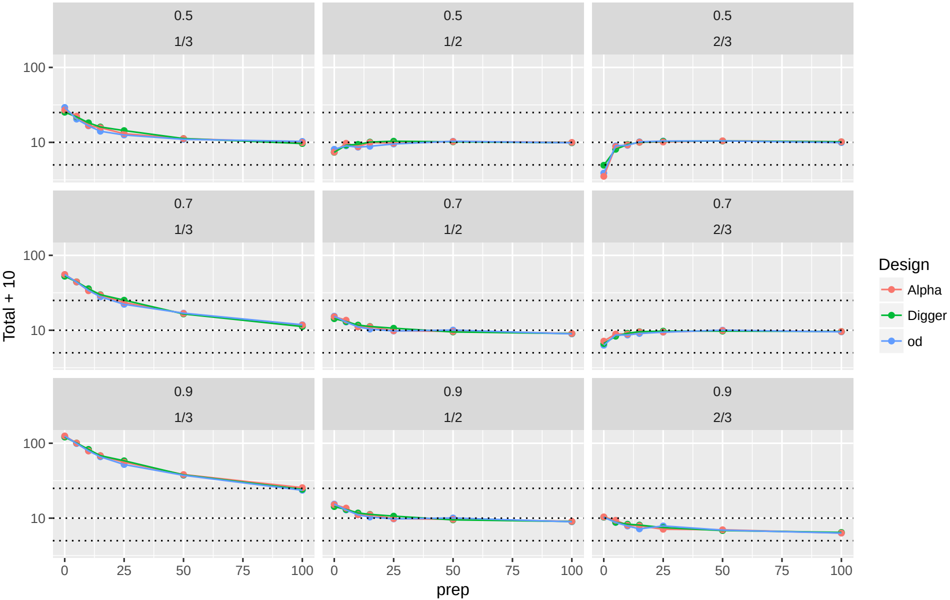
Correlation between true and predicted

- For $k = 0.7$ and $r_0^2 = 0.3$
- Using pedigree information, we can reduce the level of p and still maintain the same level of accuracy as other allocations in this situation.

Design	0	5	10	15	25	50	100
od_α	0.370	0.396	0.411	0.421	0.440	0.482	0.532
od_{gg}	0.369	0.391	0.408	0.418	0.440	0.481	0.534
od_{od}	0.386	0.412	0.430	0.443	0.464	0.502	0.553

Conclusions

- There is still recovery of information for p=0?!
- BUT the bias of variance parameter estimation for low p is large
- Bias associated with σ_i^2 particularly large
- Bias associated with σ_g^2 significantly affected by the bias with σ_i^2 (if it is large) $\forall k$


0.5
1/3
0.5
1/2
0.5
2/3
0.7
1/3
0.7
1/2
0.7
2/3
0.9
1/3
0.9
1/2
0.9
2/3
100
10
Total + 10
0
25
50
75
100
prep
Design
Alpha
Digger
od

Conclusions

- For each level of p and across all levels of k and r_0^2
  - od_{od} achieved the highest accuracy of prediction out of all three design strategies
  - od_{od} had the lowest design criterion (**A**-value) i.e. smallest average pairwise variance of variety contrasts
- For all combinations of p and r_0^2
  - od_{od} achieves a given level of accuracy with smaller p than od_{gg} and od_{α}
- od_{od} is the best design strategy for the true model
- It can be shown the **A** value is proportion to the expected level of accuracy of the predictions in the subsequent analysis

Conclusions

- Is the 25% level of p-rep the gold-standard for EGVTs?
- No... however, this is up to the discretion of the breeder taking into account
  - cost(\$)/plot i.e. how large can the trial be
  - desired level of accuracy
  - population diversity (σ_g^2) and proportion (k) of additive genetic variance
- Is there an advantage to using pedigree data in the design of EGVTs?
- Yes... results indicate for every level of p and all combinations of k and r_0^2
  - the accuracy of the predicted values and
  - relative response to selection (not presented today) are higher for allocations using od_{od} compared to od_{gg} and od_{α} designs (Cullis et. al. in prep.)

CBB
CENTRE FOR BIOINFORMATICS AND BIOMETRICS

References

BUTLER D. & CULLIS B. (2018). od: Generate optimal experimental designs, R package version 2.0.0

BUTLER D., CULLIS B., & THOMPSON R. (2018). Asreml: An R package to fit the linear mixed model. Journal of Statistical Software, in prep.

COOMBES, N. (2002). The Reactive Tabu Search for Efficient Correlated Experimental Designs. PhD thesis, Liverpool John Moores University.

References

CULLIS, B., SMITH, A., & COOMBES, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 381-393.

CULLIS, B., SMITH, A., COCKS, N. & BUTLER D. (2018). Efficient designs for early generation variety trials using genetic relationships. In prep.

MRODE, R.A., (1995) Linear models for the prediction of animal breeding values. CABI Publishing

CENTRE FOR BIOINFORMATICS AND BIOMETRICS

References

SMITH, A., GANESALINGAM, A., KUCHEL, H., & CULLIS, B. (2014). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics* **128**, 55-72.

SMITH A. & CULLIS B. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. Accepted for publication *Euphytica*

WILLIAMS, E., PIEPHO, H., & WHITAKER, D. (2011). Augmented p-rep designs. *Biometrical Journal* **53**(1), 1927.