Varietal connectivity: Does it affect the accuracy of results from a multi-environment trial analysis?

Chris Lisle

Alison Smith, Carole Birrell and Brian Cullis

Centre for Bioinformatics and Biometrics (CBB)
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong, Australia

clisle@uow.edu.au

Australasian Applied Statistics Conference, Rotorua NZ, 3-7 December 2018

December 5, 2018

Overview of talk

Multi-environment trials

- Varietal testing and evaluation stages
- Varietal connectivity
- MET analysis
- Previous assumptions - Varietal connectivity

Simulation study

- Design of study
- Model-based accuracies

Results

- Result capture
- MSE for variance parameter estimates
- Accuracy of variety effects for base site
- Accuracy of Overall performance
- Simulation study has shown

Multi-environment trials

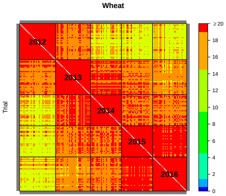
- Often referred to as METs.
- Involves the testing of varieties in designed experiments conducted in a range of environments that are often defined by geographic locations and years.
- Have several objectives. Focus here is in the area of selecting superior varieties.

Varietal testing and evaluation stages

- Varietal progression through successive breeding stages.
  - Early stage (S1/2) comprises a large number of varieties (m) over a small number of experiments (p). Selections of varietal performance are generally made after one season.
  - Late stage (S3/4) comprises smaller m and larger p. Selections are generally made after a couple of seasons.
- In Australia, the final stage of varietal evaluation is conducted through the GRDC National Variety Testing (NVT) program. Here released or near to released varieties (small m) are tested over several seasons and many locations (large p).
- **Selection and evaluation requires the accurate prediction of varietal effects.**
- Variety predictions are obtained from appropriate MET analyses.

Varietal connectivity

- Can be represented as a two-way table of varieties and environments.
    - As discussed earlier, varieties move in and out of the testing stages. Hence, the two-way table of varieties and environments is typically unbalanced (not complete). This degree of unbalance can be measured in terms of varietal connectivity between pairs of environments.
    - Often presented in the way of 'heatmaps' and tables.
        - Example: NVT Main season Wheat (Southern).
          172 experiments, 185 varieties.



	Absolute and relative connectivity				
Year	2012	2013	2014	2015	2016
2012	**51**	0.51	0.53	0.37	0.36
2013	24	**47**	0.69	0.44	0.44
2014	19	25	**36**	0.48	0.48
2015	19	23	25	**52**	0.60
2016	18	22	24	30	**50**

MET analysis

- Many techniques for the analysis of MET datasets.
- The modelling of the variety by environment interaction (V x E) is most important.
- Considered in this talk is the use of a Factor Analytic (FA) mixed model approach (Smith *et al*., 2001).
  - Widely used in Australian breeding programs and the NVT.
  - Provides a parsimonious and informative model for V x E.
  - Mixed model approach allows for the adjustment of spatial field trends.
  - **Can accommodate unbalanced data.**
    At what expense? How low can you go?

Varietal connectivity - Assumptions

- **We believe:**
  - Poor connectivity can cause issues with estimation as well as convergence and model fitting.
- **We do this:**
  - Subset MET datasets to obtain a more balanced dataset.
  - Suggest rules of thumb of having x varieties in common.
- **We don't know this:**
  - **What impact varietal connectivity has on estimation within an FA linear mixed model MET analysis.**

Linear mixed model approach

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

- $\boldsymbol{y}$ is the vector of individual plot data (usually yields) combined across experiments.
- $\boldsymbol{u}_g$ is the vector of random variety effects for individual environments. We model this using a FA multiplicative model (next slide).
- $\boldsymbol{u}_p$ is the vector of random non-genetic effects. e.g. Blocks effects for individual experiments.
- $\boldsymbol{e}$ is the vector of residuals which is modelled using a separable autoregressive spatial structure of order 1 for each experiment.

Factor Analytic model for random variety effects

$$\boldsymbol{u}_g = (\boldsymbol{\lambda}_1 \otimes \boldsymbol{I}_m)\boldsymbol{f}_1 + (\boldsymbol{\lambda}_2 \otimes \boldsymbol{I}_m)\boldsymbol{f}_2 + \cdots + (\boldsymbol{\lambda}_k \otimes \boldsymbol{I}_m)\boldsymbol{f}_k + \boldsymbol{\delta}$$

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{\delta} \end{bmatrix} \sim \mathsf{N}\left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{mk} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \otimes \boldsymbol{I}_m \end{bmatrix} \right\}$$

$$\mathsf{var}(\boldsymbol{u}_g) = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m$$

- k represents factors numbers.
- $\boldsymbol{\lambda}_{1:k}$ are the $(p \times 1)$ vectors of environmental loadings.
- $\boldsymbol{f}_{1:k}$ are the $(m \times 1)$ vectors of varietal slopes.
- $\boldsymbol{\delta}$ is the $(mp \times 1)$ vector of lack of fit.
- $\boldsymbol{\Psi}$ is the $(p \times p)$ diagonal matrix of specific trial variances.
- $\boldsymbol{\Lambda}$ is the $(p \times k)$ matrix of environmental loadings.

Variety predictions

- Estimate variance parameters using REML.
- Empirical best linear unbiased predictors (EBLUP) for random effects.
- Variety by environment effects.

$$\widetilde{\boldsymbol{u}}_g = (\hat{\boldsymbol{\Lambda}} \otimes \boldsymbol{I}_m)\widetilde{\boldsymbol{f}} + \widetilde{\boldsymbol{\delta}}$$

- Overall performance (OP) (Smith & Cullis, 2018).

$$\widetilde{\mathsf{OP}} = \bar{\boldsymbol{\lambda}}_1 \widetilde{\boldsymbol{f}}_1$$

Simulation study design

- Variance parameters, trial dimensions, genetic scenarios, and experimental design protocols from four NVT MET datasets were used to generate **realistic** scenarios.
- Two trials
  - Base trial with m varieties of interest.
  - Second trial with varying levels $(1:m)$ of varieties in common.
- Four trial sizes: 12, 24, 48, and 96 varieties.
- 3x3 factorial of low, medium and high genetic scenarios.
  - Genetic variance, based on trial accuracy (a talk by itself!).
  - Genetic correlation between trials.
- 5000 simulations for each scenario (+500) - Pilot study.
- 576 scenarios and **3.2million** FA MET analyses (ASReml-R (Butler *et al.*, 2018)).

Simulation study design - data

$$\text{var}(\boldsymbol{u}_g) = \begin{bmatrix} \lambda_1^2 + \psi & \lambda_1\lambda_2 \\ \lambda_2\lambda_1 & \lambda_2^2 + \psi \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

- Because only two trials we impose constraints of $\psi = \psi_1 = \psi_2$.
- Genetic variance
  - (σ_{11}): 0.064 (L), 0.207 (M), 0.669 (H).
  - (σ_{22}): always 0.207 (M).
- Genetic correlation
  - $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$: 0.2 (L), 0.5 (M), 0.8 (H).
- Non-genetic effects using average NVT variance parameters for blocks and residuals.

True FA parameters

Model*	σ_{11}	σ_{22}	ρ_{12}	λ_1	λ_2	ψ
LML	0.064	0.207	0.2	0.060	0.383	0.060
LMM	0.064	0.207	0.5	0.142	0.404	0.044
LMH	0.064	0.207	0.8	0.212	0.434	0.019
MML	0.207	0.207	0.2	0.203	0.203	0.166
MMM	0.207	0.207	0.5	0.322	0.322	0.103
MMH	0.207	0.207	0.8	0.407	0.407	0.041
HML	0.669	0.207	0.2	0.688	0.108	0.195
HMM	0.669	0.207	0.5	0.726	0.256	0.141
HMH	0.669	0.207	0.8	0.780	0.382	0.061

* characters 1&2: genetic variance level for trials 1 and 2, character 3: levels of between trials genetic correlation.

Model based accuracies

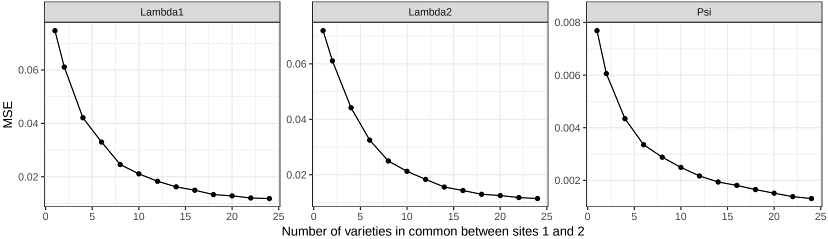
- Model-based values were obtained by fixing the variance parameters to the true values for each scenario, and obtaining the model-based accuracies for the variety predictions of interest.
- These are compared with the simulated accuracies, with the difference highlighting if connectivity was negatively impacting the accuracy of variety predictions.

Result capture

- 36 scenarios in separate R workspaces (100 GB total).
- Capture of model convergence parameters.
  - Number of iterations, updates, singularities.
- Variety effects, predictions and variance parameters estimates
  - Genetic parameters: $\boldsymbol{\Lambda}, \boldsymbol{\psi}, \rho_{12}, \sigma_{11}, \sigma_{22}, \sigma_{12}$.
  - Variety predictions and effects: $\boldsymbol{u}_g, \boldsymbol{f}$, **OP**.
  - Non-genetic: $\sigma^2_{cr}, \sigma^2_{rr}, \sigma^2, \rho_c, \rho_r$.
- (True, Pred): Bias, MSE, Accuracy(correlation).
- In the next few slides is a small set of these results. (24 variety scenario).

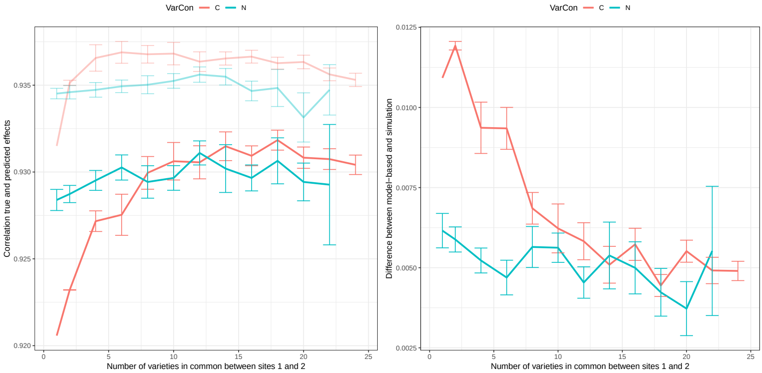
MSE for variance parameter estimates

- We believe varietal connectivity is influencing the accuracy of variance parameter estimates and in turn affecting the accuracy of variety predictions.

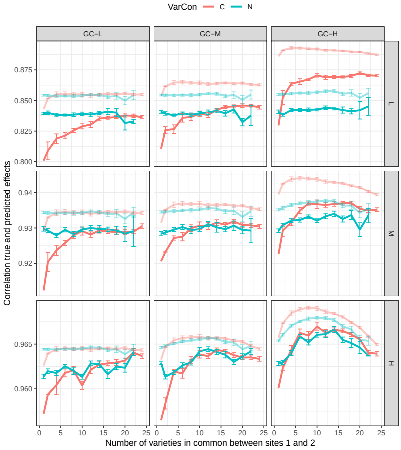


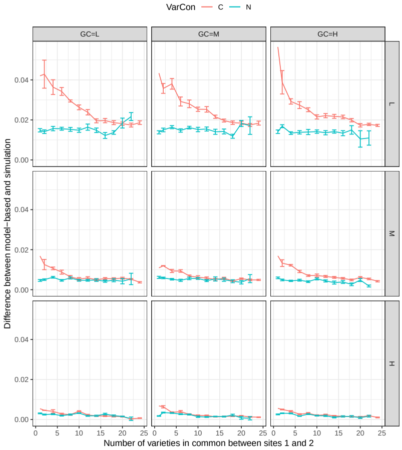
Accuracy of variety effects for base site

Accuracy (correlation between true and predicted effects) of variety effects for base site for the 'MMM' scenario. Varieties which are present in both trials (red) and those only present in base trial (blue). Dark lines represent simulation results, semi-transparent model-based.

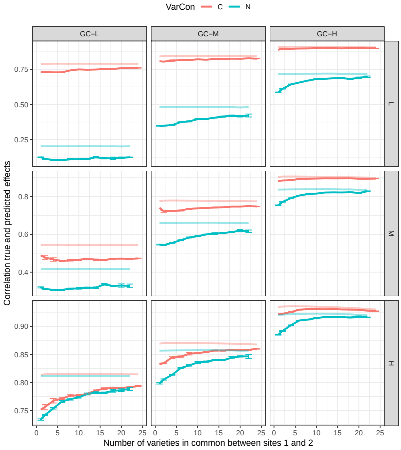


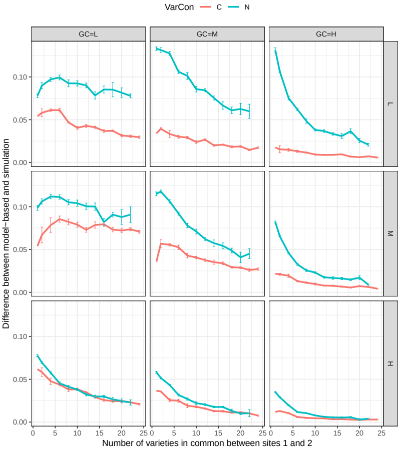
Accuracy of variety effects for base site





Accuracy - Overall performance





Simulation study has shown

- Low varietal connectivity between trials reduces the accuracy of variety predictions.
- The mechanism appeared to be a reduction in the reliability of estimation of genetic variance parameters.

Special thanks

- Thanks to RPBC and GRDC for their generous funding.
- Thanks to my supervisors Alison Smith, Carole Birrell, and Brian Cullis.
- Finally, to my family for their support.

NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

GRDC
GRAINS RESEARCH &
DEVELOPMENT CORPORATION

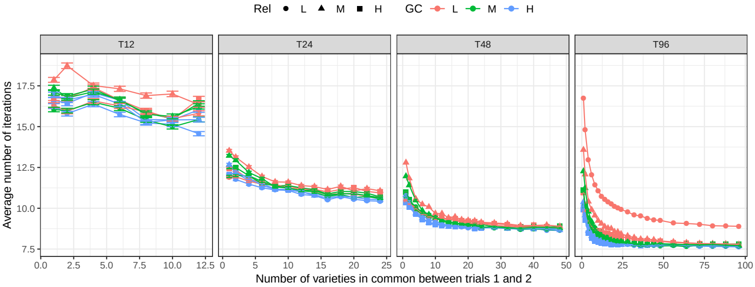
RPBC®
Radiata Pine Breeding Co Ltd
BREEDING QUALITY

Appendix

Appendix

Model convergence

- Each model was given a maximum of 10 sets of 13 iterations (updates) each to reach convergence.
  - 16,371 (0.5%) models did not convergence.
  - 3.1% in METs with 12 varieties, 0.1% with 96 varieties.
  - 67 models in total with singularities in the AI matrix.
- Decrease in the number of iterations required for convergence.


Rel ● L ▲ M ■ H
GC L M H
T12
T24
T48
T96
Average number of iterations
Number of varieties in common between trials 1 and 2