

Supplementary material for “Plant variety selection using interaction classes derived from Factor Analytic Linear Mixed Models: models with information on genetic relatedness.”

Alison B. Smith, Arun S. K. Shunmugam, David G. Butler, Brian R. Cullis

1 MET dataset construction

The construction of the MET dataset for the motivating example in the paper followed the concepts in Smith et al. [2021], the key aim of which is to optimize the amount of data on the varieties of interest, namely the varieties under consideration for selection. This data comprises both direct data from trials in which the varieties of interest were grown and indirect data derived from trials in which genetically related varieties were grown. Accuracy gains for variety selection resulting from the latter require the use of genetic relatedness in the analysis. This was achieved in the paper with the use of pedigree records. Accuracy gains associated with direct data are associated with the inclusion of sufficient trials to trace the selection histories of the varieties of interest. A key selection decision for the motivating example was L3 selections for 2023. It was pointed out in the paper that in order to maximise the direct data for five key varieties in this set it was necessary to include data from L4 trials in 2018, 2019 and 2020. Inclusion of these trials provided an additional two years of data for each of the five varieties and between 22 and 26 additional environments for those varieties. Note that there was a total of 26 L4 trials across the years 2018, 2019 and 2020. Of these trials, 12 were co-located with other trials in an environment, whilst the remaining 14 were singletons so comprised separate environments. The latter will hence-forth be referred to as “L4 only” environments. Thus the inclusion of the L4 trials increased the total number of trials from 134 up to 160, and the total number of environments from 76 up to 90. Note that of the L4 only environments, eight comprised 36 varieties (which was the smallest number per environment in the dataset) and the remaining six comprised 108 varieties.

Smith et al. [2021] use the \mathcal{A} -optimality criterion from model-based design literature to compare datasets in terms of the amount of information they contain for selection decisions. In the current context the focus is on five specific varieties from the L3 cohort of 2023 so instead of computing an \mathcal{A} -value that reflects the average pairwise variance of the entire cohort, the approach is used to compute (design based) reliabilities of prediction for individual varieties. As with the \mathcal{A} -value approach in Smith et al. [2021], this is done with respect to the variety main effects. The resultant values calculated for the full dataset and the dataset from which the L4 trials have been excluded are summarised in Table S1. This shows that

the design based reliability of the variety main effect predictions for the five key varieties increases, on average, from 0.927 to 0.943 when the L4 trials are included. The reliability for the remaining 120 lines in the L3 selection cohort is largely unaffected.

Table S1: Model-based design reliabilities of variety main effect predictions for full lentil MET dataset and reduced dataset that excludes L4 trials. Values presented are means for the five key varieties in the L3 selection set and means for the remaining 120 varieties. The final row in the table gives the mean difference in reliabilities between the full and reduced datasets.

	5 key varieties	120 other varieties
full dataset	0.943	0.883
reduced dataset	0.927	0.882
difference	0.016	0.001

As with all MET datasets that span multiple years and stages of testing, the underlying selection process leads to an incomplete variety by environment table. One method for quantifying this is via “variety connectivity”, that is, the number of varieties in common between pairs of environments. Figure S1 contains a heatmap of connectivities for the MET dataset in the paper. The minimum number of varieties in common between any pair of environments is two; the maximum is 3119 and the mean is 93. Many of the smaller connectivities are associated with the L4 only environments that had 36 varieties.

Historically, variety connectivity was thought to be the key driver of the reliability of genetic variance parameter estimation in a MET analysis and that this in turn affected the reliability of predictions of variety effects. Lisle et al. [2021] developed a diagnostic based on the \mathcal{D} -optimality criteria from the model-based design literature. They showed that this provided a superior diagnostic compared with simple variety connectivity in the sense of better fore-casting the reliability of genetic variance parameter estimates. A key finding was that connectivity is a determining factor of \mathcal{D} -optimality, but that the number of varieties in an environment is even more important. In the remainder of this section, the referencing of figures from Lisle et al. [2021] will be done using their figure number in that paper and the letter “L” (for Lisle). Figure 9L shows the results of a simulation study using additive variety effects and pairs of trials with the same numbers of varieties (either 48 or 96). The numbers of varieties in common between the trials was varied and ranged from only two in common to all varieties in common (either 48 or 96 depending on the trial size). Figure 9L shows that the diagnostic \mathcal{D} -optimality values have very good agreement with the simulation based values (reflecting actual variance parameter reliability from fitting of a model). Additionally, for a given trial size, they have a strong relationship with the log of the number of varieties in common and the trial size itself has a very large impact on the \mathcal{D} -optimality values. The diagnostic \mathcal{D} -values for the trial size of 96 are all less than -7.8 (irrespective of the number of varieties in common), whereas the values for the trial size of 48 ranged from -7.8 up to -7.0 (for very low connectivity of 2 and 4). Lisle et al. [2021] then relate these diagnostic \mathcal{D} -values to the loss in the reliability of EBLUPs. This then represents the loss due to the estimation of genetic variance parameters. Figure 11L shows that for diagnostic \mathcal{D} -values smaller than -7.4, there is negligible loss (of the order of 0.01) but for larger \mathcal{D} -values, the loss increases, with a maximum of the order of 0.04 for a \mathcal{D} -value of -7.0. The conclusion

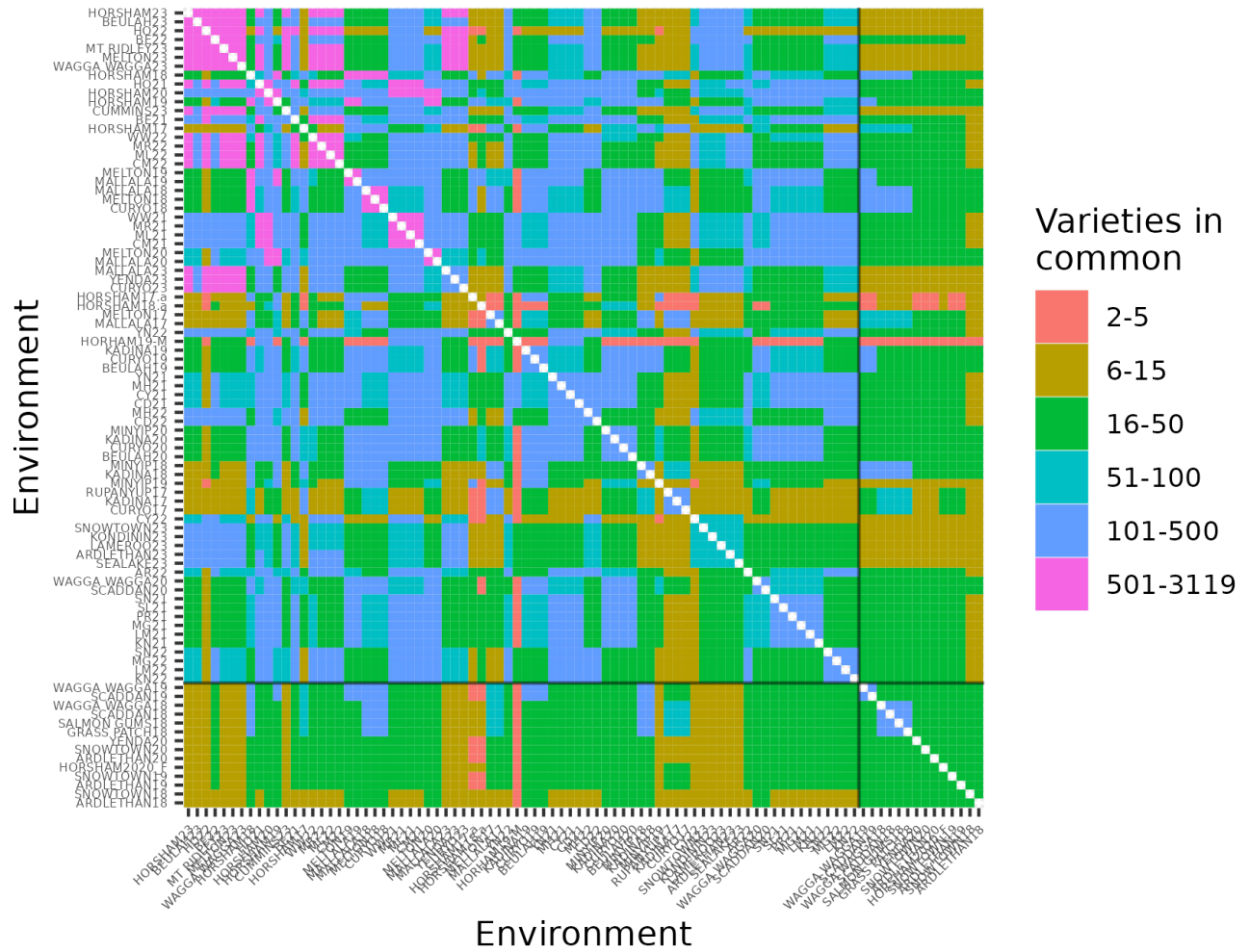


Figure S1: Heatmap of the number of varieties in common between all pairs of environments in the lentil dataset. The horizontal and vertical lines delineate the 14 “L4 only” environments (which comprise a single trial, namely an L4 trial).

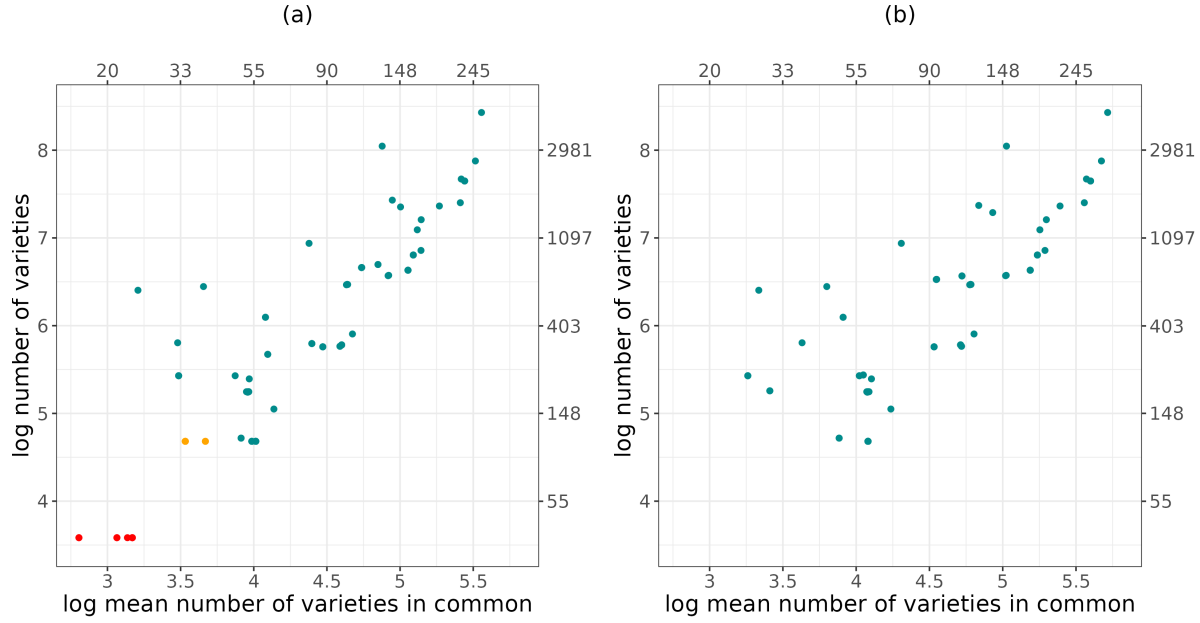


Figure S2: The number of varieties in each environment plotted against the mean number of varieties in common between each environment and all others in the lentil dataset. Panel (a) relates to the full dataset with 90 environments; the 8 “L4 only” environments that contained 36 varieties are coloured red; the 6 “L4 only” environments that contained 108 varieties are coloured orange. Panel (b) relates to the dataset from which all L4 trials have been excluded, resulting in 76 environments. The graphs are on a log scale, but back-transformed values on the original scale are shown on the right and top axes.

here is that the loss is minimal for trials with large numbers of varieties (96 or greater), even with very low levels of connectivity between this and other trials. For smaller trials, there may be a larger loss with lower levels of connectivity.

With this in mind, these two key quantities of trial size and number of varieties in common have been illustrated for the lentil dataset in Figure S2. Given that there may be some concern about the inclusion of L4 trials in the dataset, the plots have been done with respect to the dataset using all trials (panel (a)) and the dataset in which L4 trials have been excluded (panel (b)). Thus from Figures S1 and S2 it is clear that the only potential issues of a loss in reliability due to genetic variance parameter estimation may arise with the full dataset and with the small subset of L4 only environments that have 36 varieties. This is investigated further using the \mathcal{D} -optimality ideas of Lisle et al. [2021].

The large size of the lentil dataset precluded the use of the software provided by the authors of Lisle et al. [2021] to calculate the \mathcal{D} -optimality criteria. Research is currently being conducted to provide more computationally efficient methods to calculate the criteria. As a guide to assess the status of the lentil dataset, the values in the Figures presented in

Lisle et al. [2021] have been used for extrapolation. Specifically, the diagnostic \mathcal{D} -values in Figure 5L were regressed against the two key explanatory variables of trial size and number of varieties in common. This revealed a near perfect fit for the model that included the log of each of the explanatory variables, together with their interaction. This model was then applied to the diagnostic \mathcal{D} -values in Figure 9L, with the aim of predicting the diagnostic for a trial size of 36. Note that the initial modelling was done with reference to the case of independent variety effects in Figure 5L as this provided more trial sizes to examine model fit. The final values required, however, relate to the case where a numerator relationship matrix has been included (Figure 9L). The observed and predicted values from the regression are shown in Figure S3.

The predicted \mathcal{D} -values for a trial size of 36 provide a guide to the “worst case” scenario for this trial size as they relate to the connectivity between two trials, each with 36 varieties, whereas in the actual MET dataset the connectivities are mainly between a trial with 36 and other environments with far more varieties. Never-the-less, the worst mean connectivity across all environments for an L4 only trial was 13 (see Figure S2). The \mathcal{D} -value for this scenario can be read from Figure S3 as approximately -7.3. Finally, using Figure 11L, this value would suggest a loss in the reliability of EBLUPs for the trial of the order of 0.02.

1.1 Conclusion

The results on the reliability of variety predictions from the \mathcal{A} -optimality and \mathcal{D} -optimality work presented above enables a statistical assessment of the impact of including the L4 trials in the lentil dataset. The \mathcal{A} -optimality work demonstrated an average improvement in the design based reliability of variety main effect predictions for the five L3 varieties of interest from 0.927 to 0.943 when the L4 trials were included. This represents an important gain. Of course, these gains in reliability are based on the assumption that the variance parameters in the underlying linear mixed model are either known or have been estimated with minimal uncertainty. The \mathcal{D} -optimality work assessed the loss associated with variance parameter estimation when the L4 trials were included. This revealed that the maximum loss across all 90 environments in the dataset was associated with the 14 “L4 only” environments, each of which comprised 36 varieties. This loss was of the order of 0.02 but it is critical to recognise that this loss is for variety predictions for individual environments whereas the \mathcal{A} -optimality work was with respect to variety main effects. In the actual analysis of the MET data, the key variety predictions relate to the factor scores (and thence CVEs). It is not possible to assess these directly in the pre-analysis assessment of the “design” of the MET dataset because the order of the FA model is unknown a priori, as are the factor loadings. However, the factor scores are more akin to variety main effects compared with predictions for individual environments in the sense that they utilise replication across the entire MET. This, combined with the finding that the majority of environments in the dataset showed negligible loss due to variance parameter estimation, leads to the expectation that the loss in reliability for variety factor score predictions due to variance parameter estimation would be orders of magnitude less than 0.02 and certainly far less than the gains shown in the design based reliabilities. It can therefore be concluded that the inclusion of the L4 trials in the lentil dataset is justified on a statistical basis.

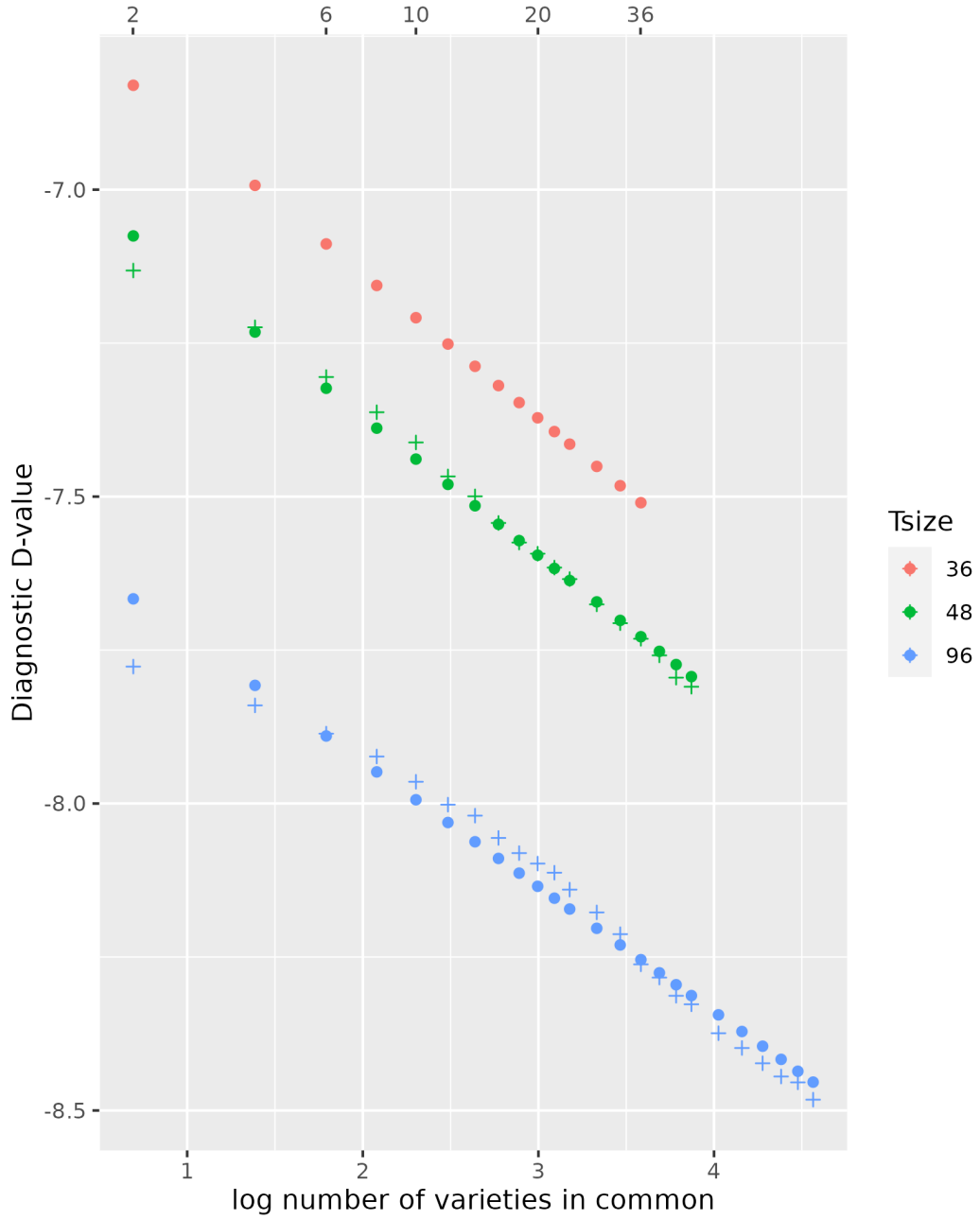


Figure S3: Diagnostic \mathcal{D} -values from Lisle et al. [2021] simulation study with additive variety effects plotted against the log of the number of varieties in common between two trials, each of which has the same number of varieties (48 and 96). The observed data from Lisle et al. [2021] for trial sizes of 48 and 96 are shown using a plus (+) symbol and the predicted values are shown using solid circles. Predicted values are also shown for a trial size of 36. The back-transformed values for the number of varieties in common are shown on the top axis.

2 Variance models for errors

The variance matrix for the errors in equation (1) of the main document is given by $\text{var}(\mathbf{e}) = \mathbf{R}$ and is assumed to be block diagonal, so that $\mathbf{R} = \oplus_{j=1}^p \mathbf{R}_j$ where $\mathbf{R}_j = \text{var}(\mathbf{e}_j)$ is the variance matrix for the errors for the j^{th} environment. In the LMM of Smith et al. [2001], spatial models are used for the errors so that the matrices \mathbf{R}_j correspond to separable autoregressive processes [Cullis and Gleeson, 1991]. These spatial models are applied with respect to a two-dimensional block of plots indexed by rows and columns. The terminology of Coombes [2002] is adopted here so that such a block will be called a (spatial) correlation block. Environments may comprise multiple correlation blocks. This may occur due to the presence of multiple trials (associated with different stages, for example) and/or the splitting of individual trials with large numbers of varieties into multiple blocks for management reasons. The correlation blocks may or may not be physically adjacent, but in either case the critical feature is that the plots are indexed using row and column numbers within blocks. In the case of environments with more than one spatial correlation block, the indexing of rows and columns at the correlation block (not environment) level warrants special consideration in the model fitting process. In this case, a separate spatial model is fitted for each correlation block but the associated variance parameters are constrained to be equal across the correlation blocks in the environment. In mathematical terms, the variance matrix for the errors for an environment j that comprises c_j correlation blocks is given by $\mathbf{R}_j = \oplus_{s=1}^{c_j} \mathbf{R}_{js}(\boldsymbol{\phi}_j)$ where \mathbf{R}_{js} is the variance matrix for the errors for correlation block s within environment j . This matrix is a function of a vector of variance parameters $\boldsymbol{\phi}_j$ which is common to all correlation blocks in environment j .

Note that in the motivating example, the majority (134) of trials comprised a single correlation block, with the remaining 26 being split into two or more correlation blocks. Of the 90 environments, 57 had multiple correlation blocks.

References

- N.E. Coombes. *The Reactive Tabu Search for efficient correlated experimental designs*. PhD thesis, Liverpool John Moores University, Liverpool, U.K., 2002.
- B. R. Cullis and A. C. Gleeson. Spatial analysis of field experiments - an extension to two dimensions. *Biometrics*, 47:1449–1460, 1991.
- C.J. Lisle, A.B. Smith, C.L. Birrell, and B.R. Cullis. Information based diagnostic for genetic variance parameter estimation in multi-environment trials. *Frontiers in Plant Science*, 2021. doi: 10.3389/fpls.2021.785430.
- A. Smith, A. Ganesalingam, C. Lisle, G. Kadkol, K. Hobson, and B. Cullis. Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs. *Frontiers in Plant Science*, 2021. doi: 10.3389/fpls.2020.623586.
- A. B. Smith, B. R. Cullis, and R. Thompson. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, 57:1138–1147, 2001.