



Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials

Chris Lisle^{1*}, Alison Smith¹, Carole L. Birrell² and Brian Cullis¹

¹ Centre for Biometrics and Data Science for Sustainable Primary Industries, School of Mathematics and Applied Statistics, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, NSW, Australia, ² School of Mathematics and Applied Statistics, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, NSW, Australia

OPEN ACCESS

Edited by:

Nunzio D'Agostino,
University of Naples Federico II, Italy

Reviewed by:

Edoardo Pasolli,
University of Naples Federico II, Italy
Zitong Li,
Commonwealth Scientific and
Industrial Research Organisation
(CSIRO), Australia
Rodrigo S. Alves,
Universidade Federal de Viçosa, Brazil

*Correspondence:

Chris Lisle
clisle@uow.edu.au

Specialty section:

This article was submitted to
Plant Bioinformatics,
a section of the journal
Frontiers in Plant Science

Received: 29 September 2021

Accepted: 15 November 2021

Published: 07 December 2021

Citation:

Lisle C, Smith A, Birrell CL and Cullis B
(2021) Information Based Diagnostic
for Genetic Variance Parameter
Estimation in Multi-Environment Trials.
Front. Plant Sci. 12:785430.
doi: 10.3389/fpls.2021.785430

Plant breeding programs evaluate varieties in series of field trials across years and locations, referred to as multi-environment trials (METs). These are an essential part of variety evaluation with the key aim of the statistical analysis of these datasets to accurately estimate the variety by environment (VE) effects. It has previously been thought that the number of varieties in common between environments, referred to as “variety connectivity,” was a key driver of the reliability of genetic variance parameter estimation and that this in turn affected the reliability of predictions of VE effects. In this paper we have provided the link between the objectives of this work and those in model-based experimental design. We propose the use of the \mathcal{D} -optimality criterion as a diagnostic to capture the information available for the residual maximum likelihood (REML) estimation of the genetic variance parameters. We demonstrate the methods for a dataset with pedigree information as well as evaluating the performance of the diagnostic using two simulation studies. This measure is shown to provide a superior diagnostic to the traditional connectivity type measure in the sense of better forecasting the uncertainty of genetic variance parameter estimates.

Keywords: multi-environment trials, linear mixed models, \mathcal{D} -optimality, variety connectivity, simulation study

1. INTRODUCTION

The objective of plant breeding is to breed superior plant varieties for various traits of economic importance. Selection of superior varieties is a result of the data analysis from a series of plant variety trials at a number of locations and possibly over several years, which are known as Multi-Environment Trials (METs). The breeding process is a progressive system involving the selection of superior varieties for further rounds of testing, removal of poor performing varieties and the inclusion of new varieties. This leads to datasets with varying levels of balance with respect to the number of varieties in common between environments, a measure known as “varietal connectivity.” Unbalanced datasets are common when the MET spans multiple breeding stages and seasons.

Given that most MET datasets are unbalanced and exhibit variance heterogeneity (Cullis et al., 2000; Chapman et al., 2003), it is critical that appropriate statistical methodology is used for analysis. Various authors have recommended the use of factor analytic (Smith et al., 2001b) linear mixed models (FALMM) to model the variety effects in different environments (VE effects),

with a one-stage approach referred to as the gold standard of analysis (Gogel et al., 2018). The FALMM approach has been shown to provide a parsimonious and computationally stable approximation to the fully saturated unstructured model (Kelly et al., 2007). It is used in many plant breeding programs within Australia and overseas due to its flexibility and ability to handle unbalanced datasets. Additionally, it provides reliable predictions of the VE effects which can then be summarised in an informative manner using the selection tools of Smith and Cullis (2018) and Smith et al. (2021b).

Not only is it important to use an appropriate method of analysis for MET data, it is also crucial to construct a suitable dataset. The latter has only recently been addressed in the literature. Smith et al. (2021a) outlined an approach for constructing MET datasets that optimises the information available for selection decisions. This is based on new concepts that characterise the structure of a breeding program, defining groups of varieties that enter the initial testing stage of the breeding program in the same year to have come from the same “contemporary group” (CG). MET datasets are formed by combining bands of data to trace the selection histories of varieties within CG to maximise the amount of direct information. For a given dataset the \mathcal{A} -optimality criterion from the model-based design literature is used to quantify the information for any given selection decision. The criterion is based on the \mathcal{A} -value which is the average pairwise variance of elementary variety contrasts. This is appropriate for assessing variety effects because a small average pairwise variance is synonymous with a low probability of making incorrect selection decisions (Bueno Filho and Gilmour, 2003). \mathcal{A} -values are computed using a specified linear mixed model (LMM), requiring the specification of the fixed and random effects and values for the variance parameters. The Smith et al. (2021a) approach assumes known variance parameters, whereas in practise they are required to be estimated. Hence, the purpose of this paper is to develop a diagnostic that measures the likely reliability of genetic variance parameter estimates for a given MET dataset. This could then be used in conjunction with Smith et al. (2021a).

It had previously been thought that variety connectivity was a key driver of the reliability of genetic variance parameter estimation and that this in turn affected the reliability of predictions of VE effects (Smith et al., 2001a, 2015; Ward et al., 2019). To combat these concerns, problematic environments were often removed from MET datasets if they appeared to have insufficient numbers of varieties in common with other environments. However, there has been little work to establish whether variety connectivity is the most appropriate measure to use for this purpose. Lisle (pers. comm) found that although variety connectivity was influential, there appeared to be other factors at play. Additionally, the number of varieties in common between environments may not be relevant for analyzes in which information on genetic relatedness is included since “connectivity” in this case is a more general concept.

In this paper we address the issue in a formal manner by considering the information for estimation of genetic variance parameters. As in Smith et al. (2021a) we use model-based

experimental design concepts and assess information using a pre-specified LMM. In this setting our interest lies in the reliability of genetic variance parameter estimates across all environments, and for individual environments. We therefore consider the \mathcal{D} -optimality criterion which measures the generalised variance of parameter estimates (Butler, 2013; Russell, 2018). This measure is typically used in the context of design searches for estimating (fixed) treatment effects (Atkinson et al., 2007), whereas the application to the problem of variance parameter estimation is not common in the literature. However, Loeza-Serrano and Donev (2014) and Nuga et al. (2017) considered \mathcal{D} -optimality to search for experimental designs for estimating variance components. As in Smith et al. (2021a) our application does not involve a design search but rather the quantification of information for a given design (dataset).

In this paper we consider the impact of MET dataset structure on the reliability of residual maximum likelihood (REML) estimates of genetic variance parameters by proposing \mathcal{D} -optimality values as a diagnostic measure. The paper is arranged as follows. Section 2 describes the statistical methodology, including a one-stage and two-stage approach for computing the diagnostic. The methods allow for the inclusion of information on genetic relatedness. In section 3, we apply the diagnostic to a durum wheat MET dataset, then examine the performance of the diagnostic using two simulation studies. Some concluding remarks are given in section 4.

2. STATISTICAL METHODS

2.1. Model for Analysis

Let y_j denote the n_j -vector of data for the j^{th} environment, $j = 1, \dots, p$. We then let \mathbf{y} denote the n -vector of data combined across all environments in the MET, so write $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_p^T)^T$. Note that $n = \sum_{j=1}^p n_j$. The LMM for \mathbf{y} can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e} \quad (1)$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix \mathbf{X} ; \mathbf{u}_g is the vector of random genetic effects with associated design matrix \mathbf{Z}_g ; \mathbf{u}_p is a vector of random non-genetic (or peripheral) effects with associated design matrix \mathbf{Z}_p and $\mathbf{e} = (\mathbf{e}_1^T, \mathbf{e}_2^T, \dots, \mathbf{e}_p^T)^T$ is the combined vector of residuals from all environments. The vector of fixed effects includes mean parameters for individual environments. The vector of random peripheral effects includes effects associated with the experimental designs within environments. It is assumed:

$$\begin{bmatrix} \mathbf{u}_g \\ \mathbf{u}_p \\ \mathbf{e} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_g & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \right) \quad (2)$$

where the matrices \mathbf{G}_g , \mathbf{G}_p and \mathbf{R} are variance matrices for \mathbf{u}_g , \mathbf{u}_p and \mathbf{e} , respectively. \mathbf{G}_g is known as the between environments variance/covariance matrix and is described in later sections. \mathbf{G}_p is assumed to be block diagonal given by $\mathbf{G}_p = \oplus_{i=1}^b \sigma_{p_i}^2 \mathbf{I}_{q_i}$ where b is the number of components in \mathbf{u}_p and q_i is the number of effects in (length of) \mathbf{u}_{p_i} . \mathbf{R} is assumed to be block diagonal, so

that $\mathbf{R} = \bigoplus_{j=1}^p \mathbf{R}_j$ where $\mathbf{R}_j = \text{var}(\mathbf{e}_j)$ is the variance matrix for the residuals for the j^{th} environment. In the LMM of Smith et al. (2001b), spatial models are used for the residuals and the matrices \mathbf{R}_j correspond to separable autoregressive processes.

The random genetic effects \mathbf{u}_g comprise the variety effects nested within environments, and will be referred to as the VE effects. If we let m denote the total number of unique varieties across all environments, then the vector \mathbf{u}_g has length mp , which is ordered as varieties within environments. In this paper we allow for the use of pedigree information, so we partition the VE effects into additive and non-additive (residual VE) effects (Oakey et al., 2007) as follows:

$$\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$$

It is assumed that $\text{var}(\mathbf{u}_a) = \mathbf{G}_a \otimes \mathbf{A}$ where \mathbf{A} is the numerator relationship matrix and \mathbf{G}_a is a $p \times p$ symmetric positive (semi)-definite matrix that will be referred to as the between environment additive genetic variance matrix. In terms of the non-additive effects, it is assumed that $\text{var}(\mathbf{u}_e) = \mathbf{G}_e \otimes \mathbf{I}_m$ where \mathbf{G}_e is a $p \times p$ symmetric positive (semi)-definite matrix that will be referred to as the between environment non-additive genetic variance matrix. The variance matrix of the total VE effects (that is, additive plus non-additive) is therefore given by:

$$\text{var}(\mathbf{u}_g) = \mathbf{G}_g = \mathbf{G}_a \otimes \mathbf{A} + \mathbf{G}_e \otimes \mathbf{I}_m \quad (3)$$

Note that if no pedigree information is included in the analysis then $\mathbf{u}_g = \mathbf{u}_e$ and $\mathbf{G}_g = \mathbf{G}_e \otimes \mathbf{I}_m$. Finally, the variance matrix for the data vector is given by:

$$\text{var}(\mathbf{y}) = \mathbf{H} = \mathbf{Z}_g \mathbf{G}_g \mathbf{Z}_g^\top + \mathbf{Z}_p \mathbf{G}_p \mathbf{Z}_p^\top + \mathbf{R} \quad (4)$$

The first step in fitting the model in Equation (1) is the estimation of the variance parameters associated with the random effects and residuals. We let $\boldsymbol{\kappa}$ denote the vector of (unknown) variance parameters and let n_κ be the associated number of parameters. We use residual maximum likelihood (REML) estimation which requires calculation of the REML scores for the elements of $\boldsymbol{\kappa}$. These are given by:

$$U(\kappa_i) = -\frac{1}{2} [\text{tr}(\mathbf{P}\dot{\mathbf{H}}_i) - \mathbf{y}^\top \mathbf{P}\dot{\mathbf{H}}_i \mathbf{P}\mathbf{y}] \quad (5)$$

where $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{X} (\mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{H}^{-1}$ with $(\mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X})^{-1}$ being any generalised inverse of $(\mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X})$. The “dot” notation indicates a derivative so that $\dot{\mathbf{H}}_i = \partial \mathbf{H} / \partial \kappa_i$, $i = 1 \dots n_\kappa$. The REML estimate of $\boldsymbol{\kappa}$ is obtained by equating the scores to zero and will be denoted by $\hat{\boldsymbol{\kappa}}$. This typically requires an iterative scheme. A computationally efficient scheme is the average information algorithm of Gilmour et al. (1995) which is a Fisher scoring algorithm in which the average information matrix, \mathcal{I}_A , is used instead of the expected information matrix, \mathcal{I}_E . The elements of these matrices are given by:

$$\begin{aligned} \mathcal{I}_A(\kappa_i, \kappa_j) &= \frac{1}{2} \mathbf{y}^\top \mathbf{P}\dot{\mathbf{H}}_i \mathbf{P}\dot{\mathbf{H}}_j \mathbf{P}\mathbf{y} \\ \mathcal{I}_E(\kappa_i, \kappa_j) &= \frac{1}{2} \text{tr}(\mathbf{P}\dot{\mathbf{H}}_i \mathbf{P}\dot{\mathbf{H}}_j) \end{aligned} \quad (6)$$

Given the REML estimates of the variance parameters we can then compute empirical best linear unbiased estimates (EBLUEs) of the fixed effects and empirical best linear unbiased predictions (EBLUPs) of the random effects in Equation (1). In particular, the EBLUPs of the VE effects are given by $\tilde{\mathbf{u}}_g = \mathbf{G}_g \mathbf{Z}_g^\top \mathbf{P}\mathbf{y}$ and these have an associated prediction error variance of $\text{var}(\tilde{\mathbf{u}}_g - \mathbf{u}_g) = \mathbf{G}_g - \mathbf{G}_g \mathbf{Z}_g^\top \mathbf{P}\mathbf{Z}_g \mathbf{G}_g$. Note that in these equations the matrices \mathbf{G}_g and \mathbf{P} are formed using the REML estimate $\hat{\boldsymbol{\kappa}}$ of $\boldsymbol{\kappa}$. We can then compute a model based reliability (Mrode and Thompson, 2005) for an individual VE effect prediction as the square of the correlation between the true effect and the EBLUP. For the k^{th} VE effect, this is obtained as:

$$\text{cor}(\tilde{u}_{gk}, u_{gk})^2 = 1 - \frac{(\mathbf{G}_g - \mathbf{G}_g \mathbf{Z}_g^\top \mathbf{P}\mathbf{Z}_g \mathbf{G}_g)_{kk}}{(\mathbf{G}_g)_{kk}} \quad (7)$$

where the subscript “ kk ” indicates the k^{th} diagonal element of the associated matrix.

2.2. Information Based Diagnostic for Genetic Variance Parameter Estimation

An asymptotic variance matrix for the REML estimates of the variance parameters can be obtained as the inverse of the information matrix. This could either be the average information matrix or, more traditionally, the expected information matrix. For the purposes of developing a diagnostic, we use the latter as it does not depend on the data, the elements of which are given in Equation (6). In this paper the interest lies in the estimation of genetic variance parameters, so we partition the variance parameters as $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_g^\top, \boldsymbol{\kappa}_{\bar{g}}^\top)^\top$, where $\boldsymbol{\kappa}_g$ are the genetic variance parameters associated with \mathbf{G}_a and \mathbf{G}_e and $\boldsymbol{\kappa}_{\bar{g}}$ are the remaining variance parameters, that is, associated with the peripheral random effects and residuals. We partition the full expected information matrix accordingly and write as:

$$\mathcal{I}_E(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top) = \begin{bmatrix} \mathcal{I}_E(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top) & \mathcal{I}_E(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_{\bar{g}}^\top) \\ \mathcal{I}_E(\boldsymbol{\kappa}_{\bar{g}}, \boldsymbol{\kappa}_g^\top) & \mathcal{I}_E(\boldsymbol{\kappa}_{\bar{g}}, \boldsymbol{\kappa}_{\bar{g}}^\top) \end{bmatrix} \quad (8)$$

The asymptotic variance matrix for $\hat{\boldsymbol{\kappa}}_g$ can then be obtained as:

$$\boldsymbol{\Sigma}(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top) = [\mathcal{I}_E(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top) - \mathcal{I}_E(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_{\bar{g}}^\top) (\mathcal{I}_E(\boldsymbol{\kappa}_{\bar{g}}, \boldsymbol{\kappa}_{\bar{g}}^\top))^{-1} \mathcal{I}_E(\boldsymbol{\kappa}_{\bar{g}}, \boldsymbol{\kappa}_g^\top)]^{-1} \quad (9)$$

Smith and Cullis (2018) recommend the use of factor analytic models for \mathbf{G}_a and \mathbf{G}_e . Other possibilities include compound symmetric and unstructured forms. Irrespective of the form used, the parameters of interest are the variances and covariances in \mathbf{G}_a and \mathbf{G}_e . The aim in this paper is to develop a diagnostic that reflects the information available to estimate these parameters but which does not require the fitting of the full LMM. In order to achieve this we apply some of the concepts from model-based design in which the aim is to search a design space for a configuration which is optimal in some sense under a pre-specified LMM. The latter includes specification of the terms in the model and also values for the variance parameters. Although the aim here is not to search a design space but rather to assess a particular design (dataset) we can proceed in a similar manner by considering a pre-specified LMM. In order to simplify

computations but enable wide applicability we use a LMM that has a relatively simple structure for the non-genetic effects. In terms of the model in Equation (1) we assume that the fixed effects comprise a mean parameter for each environment (so that $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_p)^\top$) and we assume there are no peripheral effects so write:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g(\mathbf{u}_a + \mathbf{u}_e) + \mathbf{e} \quad (10)$$

where the design matrices are given by $\mathbf{X} = \bigoplus_{j=1}^p \mathbf{1}_{n_j}$ and $\mathbf{Z}_g = \bigoplus_{j=1}^p \mathbf{Z}_{g_j}$ where \mathbf{Z}_{g_j} is the $n_j \times m$ design matrix for the VE effects for environment $j (= 1, \dots, p)$. The genetic variance matrices, \mathbf{G}_a and \mathbf{G}_e , are assumed to have unstructured forms with $p(p+1)/2$ unique variance parameters in each that are denoted by $\sigma_{a_{js}}$ and $\sigma_{e_{js}}$ ($j \leq s = 1, \dots, p$), respectively.

Finally, we assume that the residual variance matrices are given by $\mathbf{R}_j = \sigma_j^2 \mathbf{I}_{n_j}$ so that $\mathbf{R} = \bigoplus_{j=1}^p \sigma_j^2 \mathbf{I}_{n_j}$. The variance matrix for the data vector is then given by:

$$\mathbf{H} = \mathbf{Z}_g(\mathbf{G}_a \otimes \mathbf{A} + \mathbf{G}_e \otimes \mathbf{I}_m) \mathbf{Z}_g^\top + \mathbf{R}$$

and the unknown variance parameters are $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_g^\top, \boldsymbol{\kappa}_g^\top)^\top$ where $\boldsymbol{\kappa}_g$ comprises $\sigma_{a_{js}}$ and $\sigma_{e_{js}}$ ($j \leq s = 1, \dots, p$) and $\boldsymbol{\kappa}_g^\top$ comprises σ_j^2 ($j = 1, \dots, p$). We then use pre-specified values of these parameters to compute the information matrix in Equation (8) and thence the variance matrix in Equation (9). The chosen variance parameters will be denoted $\boldsymbol{\kappa}_{g_0}$ and $\boldsymbol{\kappa}_{g_0}^\top$ and the resultant variance matrix denoted by $\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_0}, \boldsymbol{\kappa}_{g_0}^\top)$. We then consider the \mathcal{D} -optimality criterion of model-based design because it is used to search for designs that minimise the generalised variance of parameter estimates. In our setting we wish to measure the generalised variance of the genetic variance parameter estimates for a given dataset. This can be obtained for the complete set of genetic variance parameters as:

$$\mathcal{D} = \frac{\log |\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_0}, \boldsymbol{\kappa}_{g_0}^\top)|}{n_{\boldsymbol{\kappa}_g}} \quad (11)$$

where the vertical bar represents the determinant and $n_{\boldsymbol{\kappa}_g}$ is the number of genetic variance parameters which is used as a divisor to provide a scaling for comparisons across models and/or datasets.

Although the overall \mathcal{D} -value is of interest, our focus is on individual environments and their relative contribution to the reliability of genetic variance parameter estimation. We therefore also compute a \mathcal{D} -value for environment $j (= 1, \dots, p)$ as:

$$\mathcal{D}_j = \frac{\log |\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_{0j}}, \boldsymbol{\kappa}_{g_{0j}}^\top)|}{n_{\boldsymbol{\kappa}_{g_j}}} \quad (12)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_{0j}}, \boldsymbol{\kappa}_{g_{0j}}^\top)$ is the partition of $\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_0}, \boldsymbol{\kappa}_{g_0}^\top)$ that relates to environment j and $n_{\boldsymbol{\kappa}_{g_j}}$ is the associated number of genetic variance parameters. In the case of models in which information on genetic relatedness is not used we have $n_{\boldsymbol{\kappa}_{g_j}} = p$, and the parameters are the genetic variance for the environment and all

$p - 1$ genetic covariances with other environments. In models in which the genetic effects are partitioned into additive and non-additive effects we have $n_{\boldsymbol{\kappa}_{g_j}} = 2p$. To distinguish between these different genetic models we label the diagnostic values as $\mathcal{D}_j(\text{A+I})$ if they correspond to a LMM with both additive (A) and non-additive (or independent, I) VE effects; $\mathcal{D}_j(\text{I})$ if they correspond to the LMM with independent VE effects alone (that is, genetic relatedness is not used) or $\mathcal{D}_j(\text{A})$ if they correspond to the LMM with additive VE effects alone. Irrespective of the genetic model used, the diagnostic values for all p environments can then be scrutinised in various ways in order to check for “problem” environments with large values, which represent those environments with large variance.

Finally, a further computational simplification can be made in the calculation of $\boldsymbol{\Sigma}(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top)$ by using the marginal variance matrix for the genetic variance parameters rather than the conditional matrix as given in Equation (9). Thus, we can use:

$$\boldsymbol{\Sigma}(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top) = [\mathcal{I}_E(\boldsymbol{\kappa}_g, \boldsymbol{\kappa}_g^\top)]^{-1} \quad (13)$$

This is a reasonable simplification given that the non-genetic variance parameters in the pre-specified LMM are simply the residual variances so that the uncertainty associated with their estimation is likely to be small.

2.3. A Two-Stage Procedure

We first let d_j be the number of varieties in environment $j (= 1, \dots, p)$ and define $d = \sum_{j=1}^p d_j$ to be the number of VE combinations present in the data. Then note that formation of $\boldsymbol{\Sigma}(\boldsymbol{\kappa}_{g_0}, \boldsymbol{\kappa}_{g_0}^\top)$ using Equations (9) or (13) involves calculating traces of matrices of dimension n . The dimensionality of the problem can be reduced by considering a two-stage approximation to the LMM as described in Gogel et al. (2018). Given the simple form for the model in Equation (10) and the associated variance matrices, we may expect little loss in using this approach and the benefit is a reduction in dimensionality from n (total number of plots in the dataset) to d (total number of VE combinations present).

In the first stage of the two-stage approach, a separate analysis is conducted for each environment in order to obtain predicted variety means and a measure of their uncertainty. In these analyses the variety effects are regarded as fixed effects. The predicted means are combined across environments to form the data for the second stage analysis. We adopt the notation of Gogel et al. (2018) so let $\boldsymbol{\eta}$ denote the full $mp \times 1$ vector of variety mean parameters for individual environments and let $\boldsymbol{\eta}_d$ be the $d \times 1$ sub-vector corresponding to the VE combinations present in the data. Thus, we can write $\boldsymbol{\eta}_d = \mathbf{D}\boldsymbol{\eta}$ where \mathbf{D} is a $d \times mp$ indicator matrix that selects the appropriate elements. We let $\hat{\boldsymbol{\eta}}_d$ be the vector of predicted variety means for individual environments from the first stage. In our setting the individual environment analyses are particularly simple, involving only a single set of effects, namely the fixed variety effects, and the residual variance for environment j is simply $\sigma_j^2 \mathbf{I}_{n_j}$. This means that the variance matrix of $\hat{\boldsymbol{\eta}}_d$ from the first stage is given by $\boldsymbol{\Omega} = \bigoplus_{j=1}^p \boldsymbol{\Omega}_j$ where

TABLE 1 | Durum example: Summary of environments in the durum MET dataset.

No.	Environment	Number of trials				Number of	
		S1	S2	S3	Total	Plots	Varieties
1	2014-Breeza	3	0	0	3	1,296	937
2	2014-Tworth	2	0	0	2	700	554
3	2015-Edgeroi	0	4	0	4	864	417
4	2015-Tworth	6	4	0	10	2,052	1,418
5	2016-Breeza	0	0	1	1	192	96
6	2016-Nstar	0	0	1	1	192	96
7	2016-Tworth	6	3	1	10	2,448	1,649
8	2017-Breeza	0	0	1	1	204	102
9	2017-Nstar	0	0	1	1	204	102
10	2017-Tworth	0	3	1	4	1,004	482
11	2018-Breeza	0	0	1	1	210	105
12	2018-Gurley	0	0	1	1	210	105
13	2018-Tworth	0	0	1	1	210	105
Total		17	14	9	40	9,786	3,708

Number of: trials for each stage of testing (S1, S2, and S3), total trials, plots and varieties.

Ω_j is the $d_j \times d_j$ diagonal matrix given by $\sigma_j^2 \text{diag}(1/r_{ji})$, where r_{ji} is the number of plots of variety i in environment j .

The LMM for the second stage combined analysis of the p environments can then be written as:

$$y_2 = X_2 \tau + D(u_a + u_e) + \xi \quad (14)$$

where $y_2 = \hat{\eta}_d$ from the first stage, and $X_2 = D(I_p \otimes I_m)$. In terms of the variance structures, $\text{var}(u_a) = G_a \otimes A$ and $\text{var}(u_e) = G_e \otimes I_m$ (as in the one-stage analysis) and $\text{var}(\xi) = \Omega$ where this is known from the first stage. The variance matrix for the data vector in the second stage is then given by:

$$H_2 = D(G_a \otimes A + G_e \otimes I_m) D^T + \Omega \quad (15)$$

Elements of the expected information matrix for the variance parameters in the second stage LMM are then given by:

$$\mathcal{I}_{2E}(\kappa_i, \kappa_j) = \frac{1}{2} \text{tr}(P_2 \dot{H}_2 P_2 \dot{H}_2) \quad (16)$$

where $P_2 = H_2^{-1} - H_2^{-1} X_2 (X_2^T H_2^{-1} X_2)^{-1} X_2^T H_2^{-1}$. This now involves matrices of dimension d rather than n . As in the previous section 2.2 we do not actually conduct the two-stage analysis but compute the expected information matrix using Equation (16) for a given choice of variance parameter values. We then form the (marginal) variance matrix for the genetic variance parameter estimates using Equation (13) and denote the resultant matrix by $\Sigma_2(\kappa_g, \kappa_g^T)$. This is substituted into Equations (11) and (12) to compute the diagnostic.

A β -version of the R script file to calculate \mathcal{D} -values is available upon request to the corresponding author.

3. RESULTS

3.1. Application of Diagnostic to a MET Dataset

The motivating MET dataset of this paper comes from the Durum Breeding Australia North program. Here we present some key summary information and apply the diagnostic procedure. Summary information for the durum wheat dataset is given in **Table 1**. The dataset comprised $m = 3,708$ varieties from $n = 9,786$ plots corresponding to 40 trials from breeding stages Stage 1 (S1) to Stage 3 (S3) across $p = 13$ environments sown between 2014 and 2018. The number of varieties per environment ranged from 96 to 1,649 with a median of 105. We note that there are $d = 6,168$ variety by environment combinations present in the data, representing a nearly 40% reduction when using a 2-stage approach for computing the diagnostic. The pedigree information comprised 3,959 records and included all the varieties in the MET dataset. The inbreeding coefficients of the latter ranged from 0.750 to 0.998 with a mean of 0.911. The number of varieties in common between pairs of environments (displayed in a heatmap in **Figure 1**) ranged from 3 to 485 with a median of 36.

Given that the analysis of the durum data for the purposes of variety selection would proceed using a LMM with the partitioning of the VE effects into additive and non-additive effects, we computed the \mathcal{I}_{2E} based on this model (so that $n_{\kappa_g} = 182$). The values of the variance parameters for calculation of the diagnostic were set to:

$$\begin{aligned} \sigma_{a_{js}} &= 0.10; & j = s = 1, \dots, p \\ &= 0.08; & j < s = 1, \dots, p \end{aligned}$$

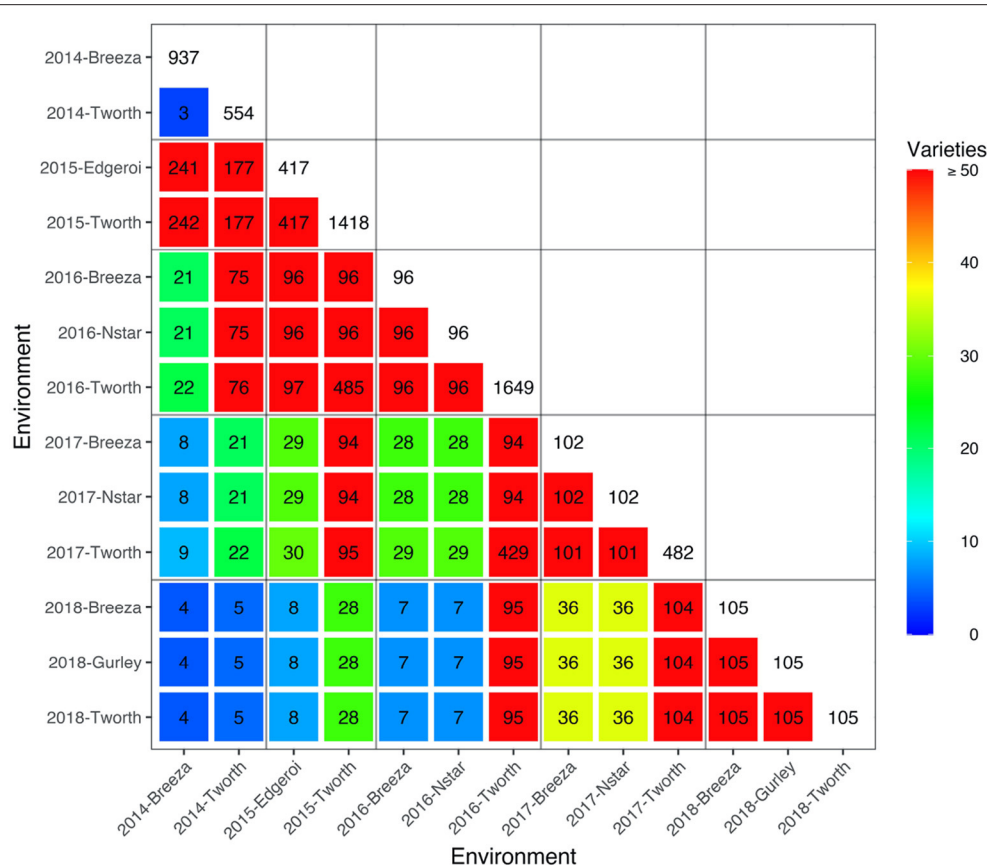


FIGURE 1 | Durum example: Heatmap of the number of varieties in common between all pairs of environments in the durum MET dataset. The colors as referenced in the legends. The boxes along the diagonal show the number of unique varieties in individual environments. Boundaries for years are indicated by the grey horizontal and vertical lines.

$$\begin{aligned}\sigma_{e_{js}} &= 0.05; & j = s = 1, \dots, p \\ &= 0.04; & j < s = 1, \dots, p \\ \sigma_j^2 &= 0.15; & j = 1, \dots, p\end{aligned}$$

These values were chosen as being both representative of actual estimates from historical analyzes that are often encountered in practise and of a magnitude that could allow the diagnostic to provide good discrimination between environments. In particular, we have set the additive genetic variance to 80% of the total genetic variance (see Equation 3) for each environment and therefore 20% for the non-additive genetic variance), and a between environments correlation of 0.8 for both additive and non-additive VE effects.

The $\mathcal{D}_j(\text{A+I})$ -values for each environment from this pre-specified LMM are given in **Table 2**. The environments 2016-Tworth and 2015-Tworth had the smallest $\mathcal{D}_j(\text{A+I})$ -values and therefore the greatest information to estimate genetic variance parameters. Whereas, 2018-Tworth, 2018-Gurley, and 2018-Breeza had the largest $\mathcal{D}_j(\text{A+I})$ -values and therefore the least information to estimate genetic variance parameters.

Given the high percentage of additive genetic variance we also computed the simpler diagnostic, based on a LMM with additive

VE effects alone. For this diagnostic $n_{\kappa_g} = 91$ representing a four-fold reduction in the number of elements in $\mathcal{I}_{2E}(\kappa_i, \kappa_j)$ and hence a significant savings in computation. The resultant $\mathcal{D}_j(\text{A})$ -values are presented in **Figure 2A** and **Table 2**, which show little difference compared with the $\mathcal{D}_j(\text{A+I})$ -values. In particular, **Figure 2A** shows an almost 1:1 relationship.

To investigate the robustness of the diagnostic we also inspect the $\mathcal{D}_j(\text{A+I})$ - and $\mathcal{D}_j(\text{A})$ -values with parameters set to those which are at the lower end of those seen in practise. The values of the variance parameters for calculation of the diagnostic were set to:

$$\begin{aligned}\sigma_{a_{js}} &= 0.05; & j = s = 1, \dots, p \\ &= 0.02; & j < s = 1, \dots, p \\ \sigma_{e_{js}} &= 0.15; & j = s = 1, \dots, p \\ &= 0.06; & j < s = 1, \dots, p \\ \sigma_j^2 &= 0.15; & j = 1, \dots, p\end{aligned}$$

In particular, we have set the additive genetic variance to 40% of the total genetic variance (see Equation 3) for each environment and therefore 60% for the non-additive genetic variance, and a

TABLE 2 | Durum example: Diagnostic \mathcal{D}_j -values based on LMMs with additive and non-additive VE effects [$\mathcal{D}_j(A+I)$] and those based on LMMs with additive VE effects alone [$\mathcal{D}_j(A)$].

Environment	High		Low	
	$\mathcal{D}_j(A+I)$	$\mathcal{D}_j(A)$	$\mathcal{D}_j(A+I)$	$\mathcal{D}_j(A)$
2016-Tworth	-8.28	-9.25	-7.82	-8.60
2015-Tworth	-8.27	-9.27	-7.82	-8.61
2017-Tworth	-8.03	-9.04	-7.45	-8.25
2015-Edgeroi	-7.95	-8.94	-7.30	-8.09
2017-Breeza	-7.65	-8.69	-6.95	-7.80
2017-Nstar	-7.65	-8.69	-6.95	-7.80
2014-Breeza	-7.56	-8.61	-6.74	-7.47
2016-Breeza	-7.51	-8.53	-6.76	-7.61
2016-Nstar	-7.51	-8.53	-6.76	-7.61
2014-Tworth	-7.43	-8.51	-6.72	-7.53
2018-Breeza	-7.39	-8.55	-6.67	-7.59
2018-Gurley	-7.39	-8.55	-6.67	-7.59
2018-Tworth	-7.39	-8.55	-6.67	-7.59

High and low parameter values of additive variance (80 and 40%, respectively) and between environments genetic correlation (0.8 and 0.4, respectively) are used. Environments are ordered in ascending order on their $\mathcal{D}_j(A+I)$ -values (High).

between environments correlation of 0.4 for both additive and non-additive VE effects. The resultant \mathcal{D}_j -values are presented in both **Figure 2** and **Table 2**. Once again there is little difference in the rankings of environments for $\mathcal{D}_j(A+I)$ compared with $\mathcal{D}_j(A)$ (**Figure 2D**). Additionally the rankings were robust to the two choices of variance parameters (high and low) used in forming the diagnostic, with the only noticeable change being associated with 2014-Breeza (**Figures 2B,C**).

As a comparison with the historical measure of variety connectivity the $\mathcal{D}_j(A+I)$ values computed using the high set of parameters have been plotted against mean connectivity in **Figure 3**. Two other structural characteristics of the environments are indicated on this plot, namely the number of varieties grown and the mean replication per variety. The figure shows that the diagnostic values encompass numerous structural elements of the environments.

3.2. Simulation Studies to Investigate the Performance of the Diagnostic

3.2.1. LMM Without Pedigree Information

Within the framework of the LMM with independent VE effects it was previously thought that variety connectivity was a key driver of the reliability of variance parameter estimation and that this in turn affected the reliability of predictions of VE effects. We therefore consider a simulation study in which a range of connectivity levels is examined and assess the performance of both variety connectivity and the \mathcal{D} -optimality diagnostic. For simplicity, and without loss of generality, we use $p = 2$ environments and label these as Env1 and Env2. Each environment has the same number of varieties (so that $d_1 = d_2$), and we vary the number of varieties in common (which is given by $c = d - m$). We assume the trials in Env1 and Env2 comprise

3 replicates and consider 4 sizes (Tsize) of trial corresponding to different numbers of varieties, namely $d_1 (= d_2) = 12, 24, 48$ and 96 so that $n_1 (= n_2) = 36, 72, 144$ and 288.

The simulation study for the first trial size (Tsize=12) is described in the following. We consider the connectivity levels $c = 2, 4, \dots, 12$ (increments of 2). The maximum total number of varieties across Env1 and Env2 is $m = 22$, corresponding to $c = 2$. We label these varieties as V1 - V22. We assume that the 12 varieties in Env1 are always V1 - V12. The 12 varieties in Env2 are then V1 - V12 for $c = 12$; V2 - V13 for $c = 11$ and so on to V11 - V22 for $c = 2$. Our focus is on Env1 because this contains the same varieties across all connectivity levels so allows a fair comparison across these levels. The underlying LMM is as in Equation (10) but with independent VE effects alone (that is, without the additive VE effects) so that $n_{\kappa_g} = 3$. Given the data structure for each value of c and some pre-specified variance parameters, we can compute the diagnostic for Env1 which will be denoted $\mathcal{D}_{1c}(I)$. For the purposes of both the calculation of $\mathcal{D}_{1c}(I)$ and of data generation in the simulation study we chose the values of the variance parameters to be $\kappa_{g_0} = (\sigma_{e_{11_0}} = 0.2, \sigma_{e_{12_0}} = 0.16, \sigma_{e_{22_0}} = 0.2)^\top$ and $\kappa_{g_0} = (\sigma_{1_0}^2 = 0.15, \sigma_{2_0}^2 = 0.15)^\top$. The diagnostic is calculated using the two-stage formula for expected information, that is, as in Equation (16) and is given by:

$$\mathcal{D}_{1c}(I) = \frac{\log \left| \Sigma_c \left((\sigma_{e_{11_0}}, \sigma_{e_{12_0}}), (\sigma_{e_{11_0}}, \sigma_{e_{12_0}})^\top \right) \right|}{2} \quad (17)$$

In the simulation study, the steps for the t^{th} simulation ($t = 1 \dots N$) are as follows:

1. Generate the random genetic effects \mathbf{u}_e and residuals \mathbf{e} as per the LMM in Equation (10) and for the pre-specified variance parameters κ_{g_0} and κ_{g_0} . In terms of the fixed effects, without loss of generality we choose $\tau_1 = \tau_2 = 0$. Note that we generate $2m$ genetic effects, where $m = 22$ which corresponds to the maximum total number of varieties across all connectivity levels. We denote the resultant vector for simulation t by \mathbf{u}_{e_t} . The residuals for simulation t are denoted by \mathbf{e}_t which is a vector of length $n_1 + n_2 = 72$.
2. For the connectivity level c , we subset the appropriate 24 elements of \mathbf{u}_{e_t} . We will label the associated vector as $\mathbf{u}_{e_{tc}}$. We then form the data vector and fit the LMM as in Equation (10), without the inclusion of pedigree information. We save the REML estimates of the genetic variance parameters, denoting these as $\hat{\sigma}_{e_{11_{tc}}}, \hat{\sigma}_{e_{12_{tc}}}, \hat{\sigma}_{e_{22_{tc}}}$ and save the EBLUPs of the genetic effects, denoting these as $\hat{\mathbf{u}}_{e_{tc}}$.
3. Repeat step 2. for each value of c

A total of $N = 2,000$ simulations was conducted for each trial size by variety connectivity combination. The simulation based diagnostics and reliabilities were only computed for the LMMs in Step 2. that achieved convergence (with one update if required) and resulted in a positive definite form for the REML estimate of \mathbf{G}_e . All models in this paper were fitted using the ASRem1-R package (Butler et al., 2017) within R (R Core Team, 2020).

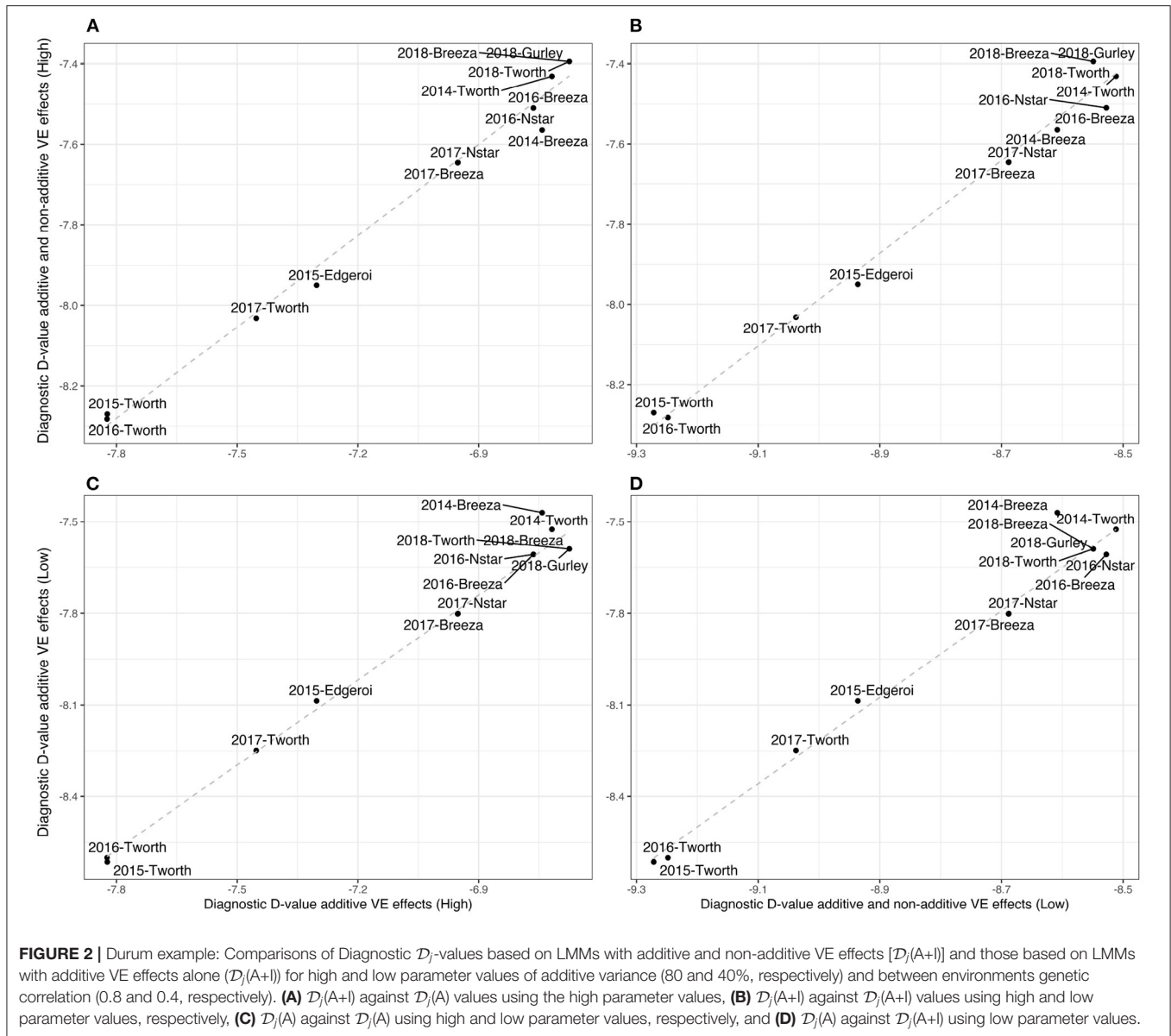


FIGURE 2 | Durum example: Comparisons of Diagnostic \mathcal{D}_j -values based on LMMs with additive and non-additive VE effects [$\mathcal{D}_j(A+)$] and those based on LMMs with additive VE effects alone ($\mathcal{D}_j(A)$) for high and low parameter values of additive variance (80 and 40%, respectively) and between environments genetic correlation (0.8 and 0.4, respectively). **(A)** $\mathcal{D}_j(A+)$ against $\mathcal{D}_j(A)$ values using the high parameter values, **(B)** $\mathcal{D}_j(A+)$ against $\mathcal{D}_j(A)$ values using high and low parameter values, respectively, **(C)** $\mathcal{D}_j(A)$ against $\mathcal{D}_j(A)$ using high and low parameter values, respectively, and **(D)** $\mathcal{D}_j(A)$ against $\mathcal{D}_j(A+)$ using low parameter values.

The simulations were conducted in order to obtain two main quantities of interest for each value of c , namely a measure of the reliability of the genetic variance parameter estimates and a measure of the reliability of the predicted variety effects for Env1. For the former, we computed a simulation based equivalent of the diagnostic in Equation (17), namely:

$$\mathcal{D}_{1c}^s(I) = \frac{\log |V_c((\hat{\sigma}_{e_{11c}}, \hat{\sigma}_{e_{12c}}), (\hat{\sigma}_{e_{11c}}, \hat{\sigma}_{e_{12c}})^T)|}{2} \quad (18)$$

where the determinant is with respect to the sample variance/covariance matrix of the REML estimates of the genetic variance parameters for Env1. In terms of the variety predictions, we computed the reliability of the EBLUPs for the 12 varieties that were always present in Env1, namely V1 - V12. For each value of c , the reliability for variety $k(= 1 \dots 12)$ in Env1

was computed as the square of the sample correlation between the true (generated) effects (element of $\mathbf{u}_{e_{1c}}$ for the variety and Env1) and the EBLUPs (element of $\tilde{\mathbf{u}}_{e_{1c}}$ for the variety and Env1). This will be denoted R_{kc}^s .

Noting that the simulation based reliabilities (R_{kc}^s) of the variety predictions take into account the uncertainty in having to estimate the variance parameters, we compute analogues values that assume known variance parameters. These reliabilities therefore reflect the maximum possible values against which we can measure the loss attributable to variance parameter estimation. This was achieved by fitting, for each value of c , the LMM as per Equation (10) but with the variance parameters fixed at the value κ_0 . We then computed the model based reliability for variety $k(= 1 \dots 12)$ in Env1 using Equation (7) but because this has been computed with respect to known variance parameters (not REML estimates) we will call it the design based

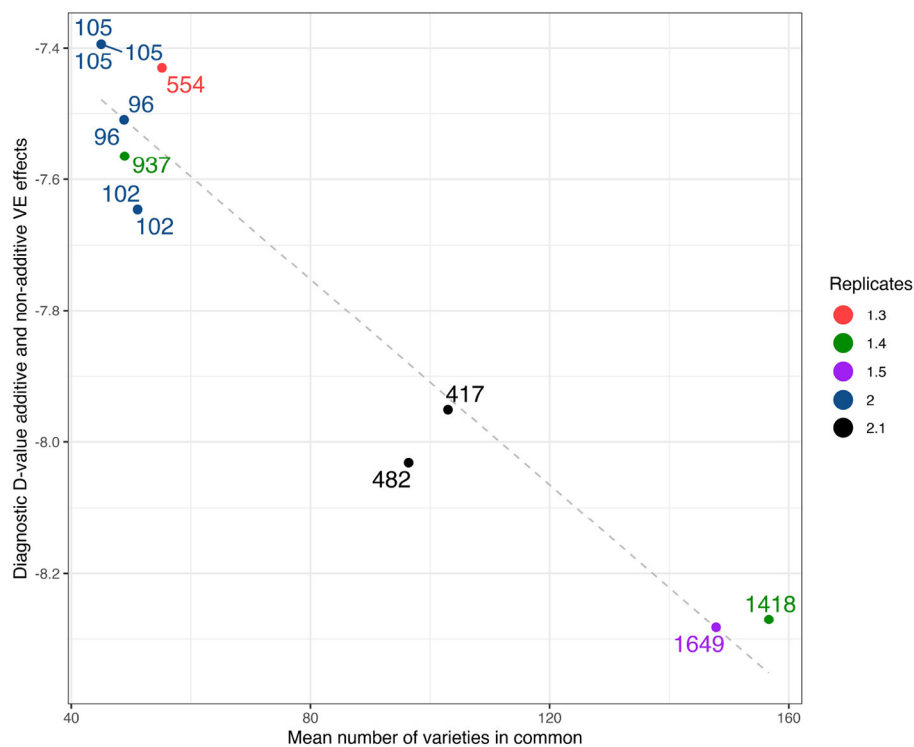


FIGURE 3 | Durum example: Diagnostic $\mathcal{D}_{(A+I)}$ -values based on LMM with additive and non-additive VE effects plotted against mean number of varieties in common. Labels show the total number of varieties, colors as represented in legend show the mean number of replicates.

reliability (R_{kc}^D). We calculated the associated loss for the EBLUP reliabilities as:

$$R_{kc}^D - R_{kc}^S \quad (19)$$

Finally we summarise these by taking means across the varieties in Env1 that were also present in Env2. We restrict the results to this set of varieties because in any MET analysis, there is a fundamental difference between varieties that were present in multiple environments (so-called connected varieties) and those that were present in a single environment only. The MET analysis, compared with separate analyzes of individual environments, has the potential to improve the reliability of predictions for connected varieties through the use of additional data. This is not the case for varieties present in a single environment only. Hence our focus is the connected varieties.

3.2.2. Results of Simulation Without Pedigree Information

First we note that the number of simulations in which the model fitting was successful (as defined in section 3.2.1) was strongly related to the number of varieties in common between the two environments (see **Figure 4**). The number of successful model fits for the connectivity level of $c = 2$ was particularly low and additionally the results were found to be unreliable.

Therefore, in what follows, the results for this connectivity level have been excluded.

The relationship between the diagnostic $\mathcal{D}_{1c}(I)$ -values (from Equation 17) and the simulation based equivalent $\mathcal{D}_{1c}^S(I)$ -values (from Equation 18) is shown in **Figure 5A**. This good agreement shown in **Figure 5A** clearly indicates that the diagnostic performs well in terms of forecasting the level of uncertainty in genetic variance parameter estimation. **Figure 5B** shows that there is a decreasing linear relationship between (log) variety connectivity and the uncertainty in genetic variance parameter estimation, but this is only within a given trial size, that is, for a given number of varieties. The connectivity measure fails for comparisons involving trials with different number of varieties.

Figure 6 shows the mean losses in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. These are plotted against the diagnostic $\mathcal{D}_{1c}(I)$ -values, with a separate panel for each trial size. Each point has been supplemented with a standard error of the mean (SEM) which was based on a pooled estimate of error across all trial sizes and connectivity levels. Thus differences in SEM reflect differences in the numbers of varieties used to compute the means (that is, differences in connectivity). The panels in this figure show that, for a given trial size, the loss in reliability of EBLUPs is well predicted by the diagnostic $\mathcal{D}_{1c}(I)$ -values. This also holds across trial sizes, although the relationship is more variable (**Figure 7**).

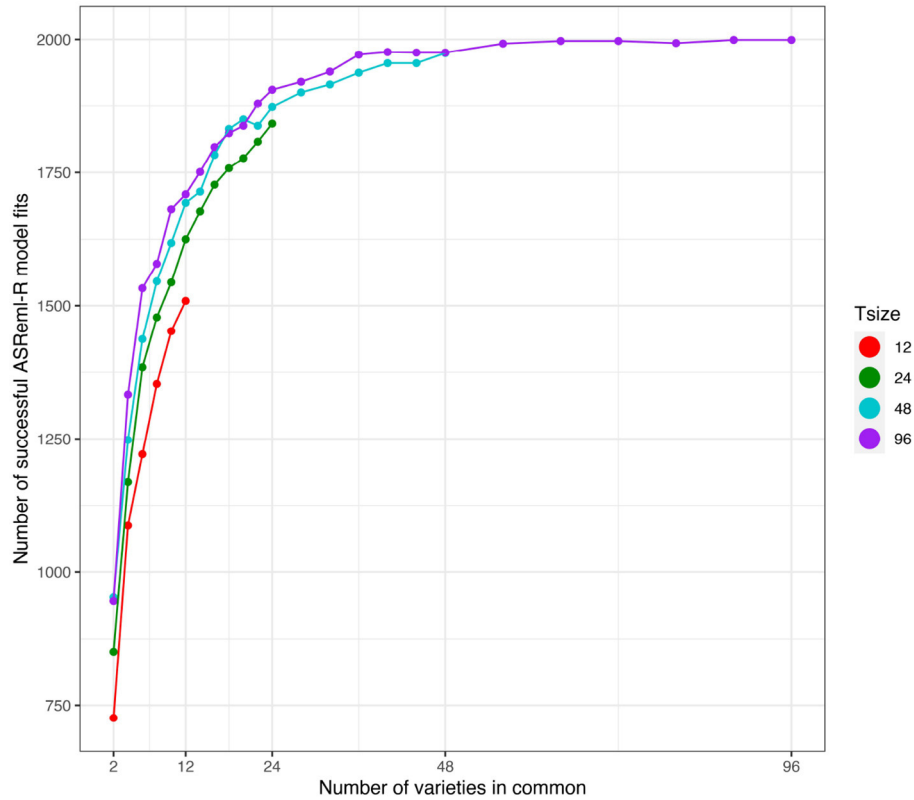


FIGURE 4 | Independent VE effects simulation study: Number of successful model fits from $N = 2,000$ simulations plotted against number of varieties in common for four trial sizes (trials with 12, 24, 48, and 96 varieties). Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $c = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

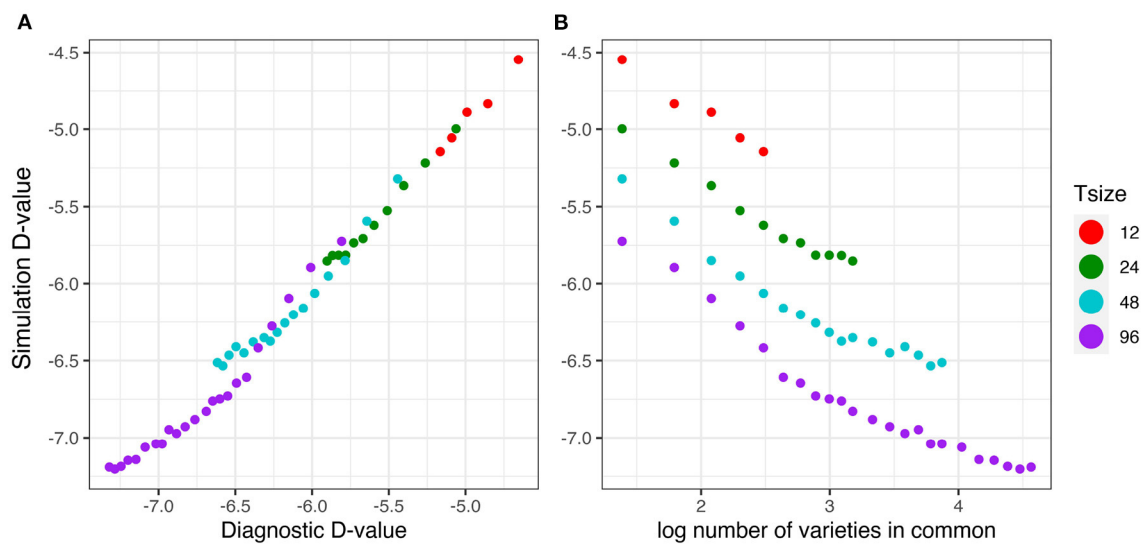


FIGURE 5 | Independent VE effects simulation study: Simulation based $\mathcal{D}_{1c}^S(l)$ -values plotted against **(A)** diagnostic $\mathcal{D}_{1c}(l)$ -values and **(B)** log number of varieties in common for four trial sizes (trials with 12, 24, 48, and 96 varieties) and a sequence of connectivity levels. Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $c = 4$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

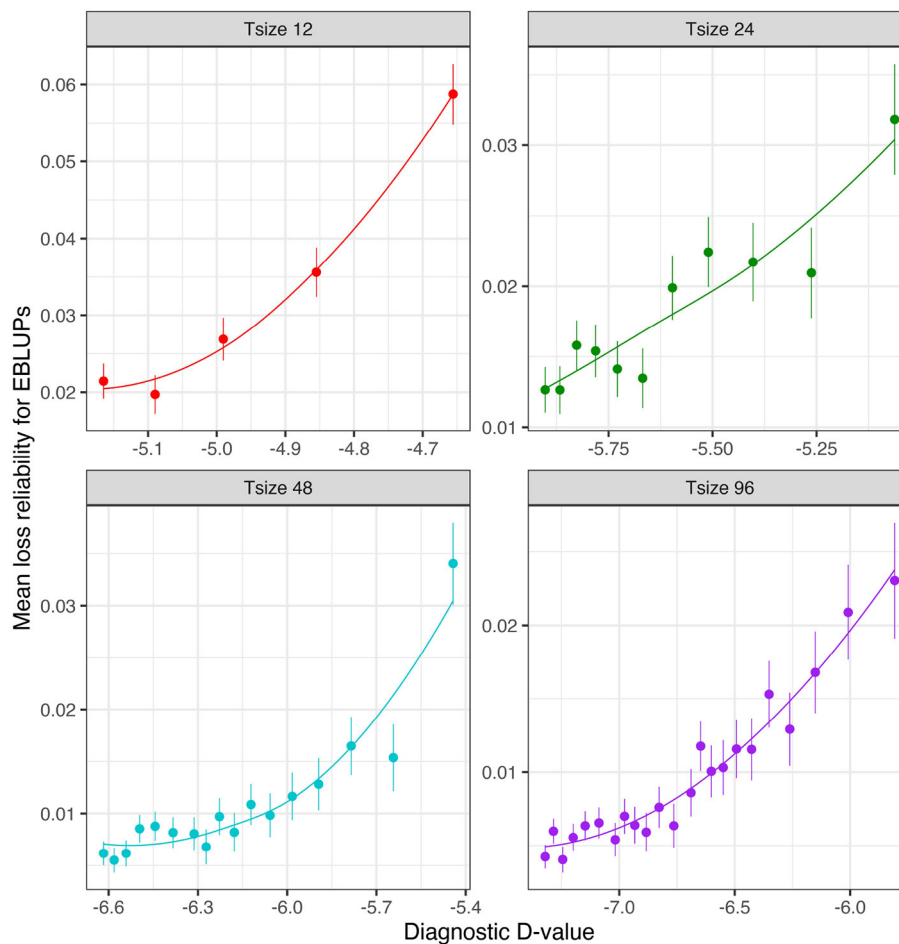


FIGURE 6 | Independent VE effects simulation study: Mean loss in reliability of the EBLUPS of VE effects for Env1 for those varieties that were present in both environments. Each panel corresponds to a different trial size (trials with 12, 24, 48, and 96 varieties) and the points correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through the means for each Tsize.

Results displayed in **Figures 4, 5, 7** have been extracted for the “best case” scenario of 100% connectivity for each Tsize and are presented in **Table 3**. This again shows the good agreement between the diagnostic $\mathcal{D}_{1c}(I)$ -values and the simulation based $\mathcal{D}_{1c}^s(I)$ -values and the relationship between the diagnostic and the loss in reliability of VE predictions. It also shows that, even with 100% connectivity, there were substantial problems with the smallest trial size in terms of all criteria (number of successful model fits, reliability of genetic variance parameter estimates and reliability of VE effect predictions).

3.2.3. LMM With Pedigree Information

We then extended the simulation study in order to assess the performance of the diagnostic in terms of correlated VE effects. To simplify the simulations, and without loss of generality, we considered the LMM as in Equation (1) but without the non-additive VE effects, so that as in section 3.2.1, $n_{\kappa_g} = 3$. The set-up for the study is the same as in section 3.2.1 but we only consider the two larger trial sizes, namely $d_1 (= d_2) = 48$ and 96 so that $n_1 (= n_2) = 144$ and 288. Across the range of connectivity levels

the total number of varieties required for the simulation was 191 (corresponding to $c = 2$ for the trial size of 96) and we label these as V1-V191. The simulation study requires a numerator relationship matrix (\mathbf{A}) for these varieties. We therefore chose V1-V191 from the actual lines in Stage 3 (S3) and Stage 4 (S4) in 2018 and 2017 in the durum data and computed \mathbf{A} from the associated pedigree information. For the chosen subset of varieties, the inbreeding coefficient ranged from 0.938 to 0.969 with a mean of 0.957.

For the purposes of both the calculation of $\mathcal{D}_{1c}(\mathbf{A})$ and of data generation in the simulation study we chose the values of the variance parameters to be $\kappa_{g_0} = (\sigma_{a_{11_0}} = 0.1, \sigma_{a_{12_0}} = 0.08, \sigma_{a_{22_0}} = 0.1)^\top$ and $\kappa_{g_0} = (\sigma_{1_0}^2 = 0.15, \sigma_{2_0}^2 = 0.15)^\top$.

3.2.4. Results of Simulation With Pedigree Information

As in the independent VE effects study, the number of simulations in which the model fitting was successful was

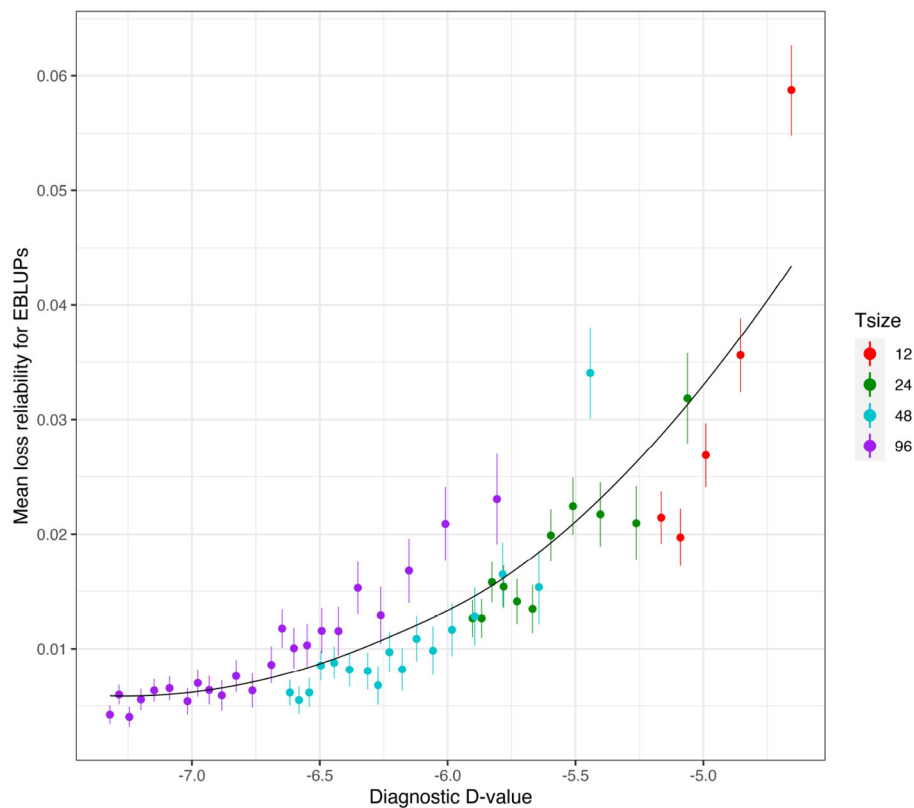


FIGURE 7 | Independent VE effects simulation study: Mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. The colours correspond to different trial sizes (trials with 12, 24, 48, and 96 varieties) and the points for each colour correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through all the means.

TABLE 3 | Independent VE effects simulation study: Summary of key results for each trial size and the case of 100% connectivity between the two trials: Simulation based $\mathcal{D}_{1c}^s(l)$ -values; diagnostic $\mathcal{D}_{1c}(l)$ -values; mean loss in reliability of EBLUPs of VE effects for Env1 (with associated standard error); number of successful model fits out of $N = 2,000$ simulations.

Tsize	Varieties in common	$\mathcal{D}_{1c}^s(l)$	$\mathcal{D}_{1c}(l)$	EBLUP reliability		Successful model fits
				loss	se	
12	12	-5.15	-5.16	0.021	0.0023	1,509
24	24	-5.85	-5.90	0.013	0.0016	1,842
48	48	-6.51	-6.62	0.006	0.0011	1,974
96	96	-7.19	-7.32	0.004	0.0008	1,999

related to the number of varieties in common between the two environments (see **Figure 8**). However, a key difference was that the number of successful model fits for the connectivity level of $c = 2$ was reasonable so these results have been included in what follows.

The results are presented in the same format as in section 3.2.1. The good agreement shown in **Figure 9A** clearly indicates that the diagnostic performs well in terms of forecasting the level of uncertainty in genetic variance parameter estimation in the presence of pedigree information. **Figure 9B** shows that there is a decreasing linear relationship between (log) variety connectivity and the uncertainty in genetic variance parameter

estimation, but this only holds for trials with the same number of varieties.

The mean loss in reliability of the EBLUPs of the additive VE effects for Env1 for those varieties that were present in both environments is well predicted by the diagnostic $\mathcal{D}_{1c}(A)$ -values, both for individual trial sizes (**Figure 10**) and across trial sizes (**Figure 11**).

We again investigate the robustness to a change in variance parameters. The same low bound values used in the durum example are used and applied to the same set of simulation studies. These results are provided in **Supplementary Materials** and highlight similar trends to the results above.

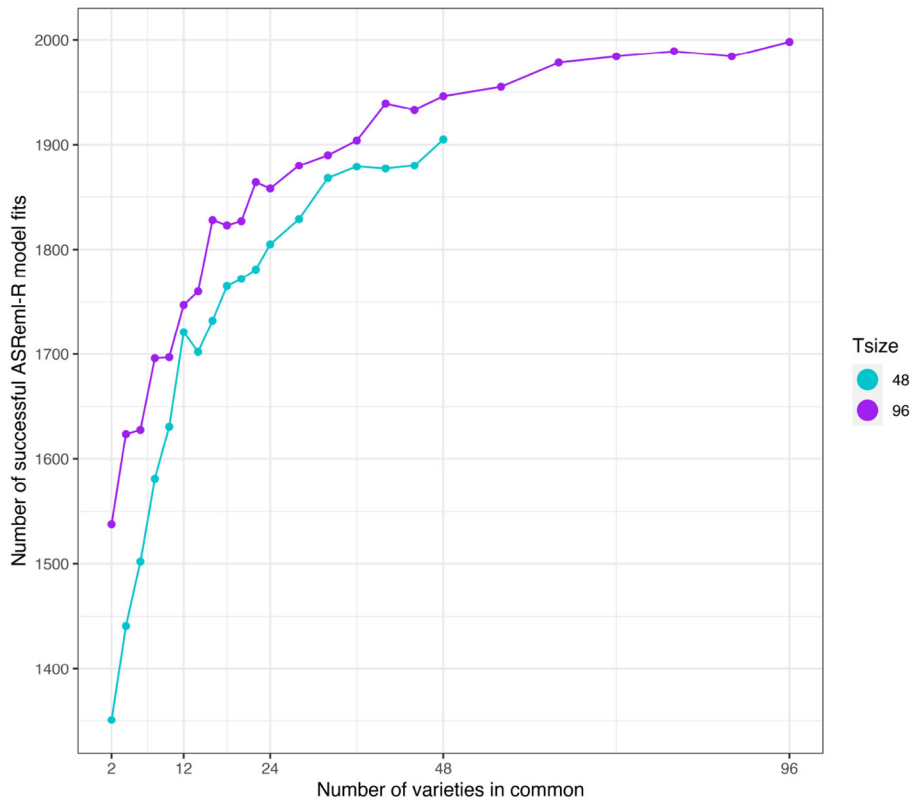


FIGURE 8 | Additive VE effects simulation study: Number of successful model fits from $N = 2,000$ simulations plotted against number of varieties in common for two trial sizes (trials with 48 and 96 varieties). Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $c = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

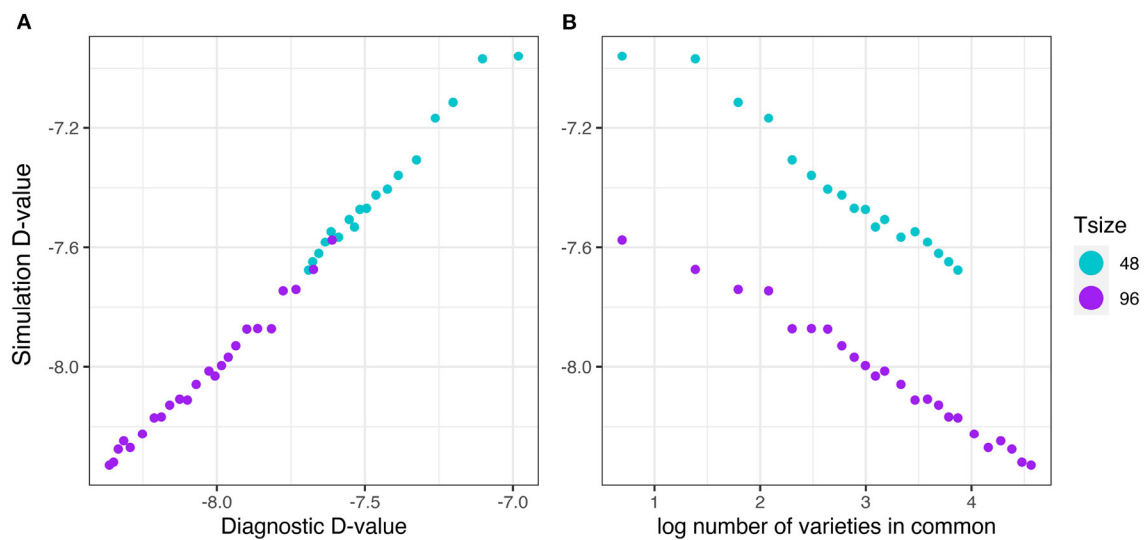
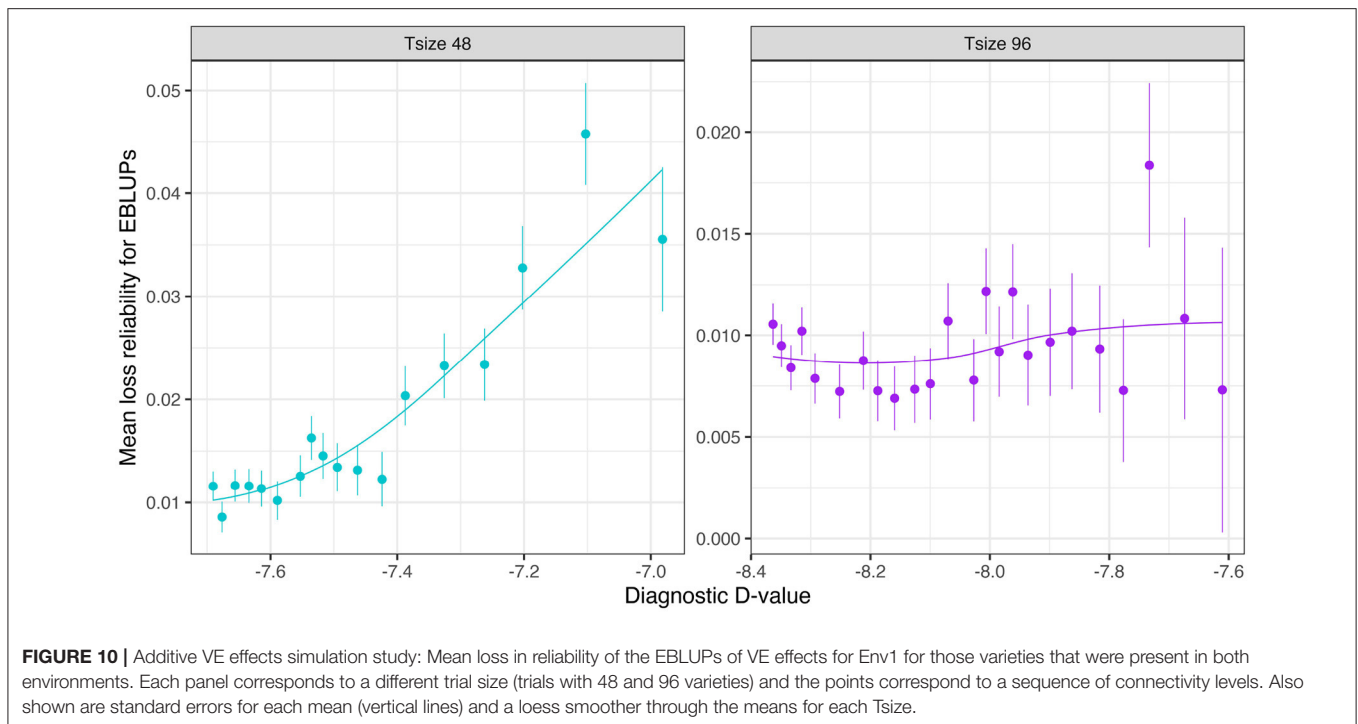


FIGURE 9 | Additive VE effects simulation study: Simulation based \mathcal{D}_{1c}^s (A)-values plotted against (A) diagnostic \mathcal{D}_{1c} (A)-values and (B) log number of varieties in common for two trial sizes (trials with 48 and 96 varieties) and a sequence of connectivity levels. Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $c = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).



4. DISCUSSION

Despite the fact that Multi-environment trials (METs) are a key aspect of plant breeding, there is little in the literature that addresses the “design” of METs in the sense of determining which trials should be combined. Within the paradigm of the factor analytic linear mixed model (FALMM) approach for MET analysis, the aim of both the design and analysis is the reliable prediction of the variety effects for individual environments (VE effects). These are the baseline predictions which can then be summarised across environments in meaningful ways, for example, using the interaction class approach of Smith et al. (2021b).

Smith et al. (2021a) discuss the structure of MET datasets and use a model-based design approach to quantify the amount of information for variety comparisons in a given dataset. They apply the \mathcal{A} -optimality measure to variety effects to show the importance of including as many trials as are necessary to capture the selection histories of the varieties under consideration for selection. As in the case of standard model-based design, the calculation of \mathcal{A} -values is based on a pre-specified linear mixed model (LMM) so assumes that the associated variance parameters are known. It is important, therefore, to also consider the impact of the structure of a MET dataset on the reliability of genetic variance parameter estimation, as this in turn may affect the reliability of variety predictions, so should be used in conjunction with \mathcal{A} -optimality.

In this paper we have developed a diagnostic to be applied to a MET dataset prior to analysis in order to assess the likely reliability of genetic variance parameter estimates, both

for the dataset across all environments and for individual environments. As in Smith et al. (2021a) we use a model-based design approach and apply \mathcal{D} -optimality measures to genetic variance parameters. Two simulation studies, one using a LMM with independent VE effects and the other additive VE effects, showed that these diagnostic \mathcal{D} -values performed well in the sense of predicting the actual reliability of genetic variance parameter estimates.

Historically, variety connectivity between environments was the measure calculated prior to the conduct of a MET analysis to investigate the likely reliability of genetic variance parameter estimation. This measure is simple to compute and intuitively reasonable but there has been nothing in the literature to validate its use. In our simulation studies it was shown that variety connectivity was only able to predict the reliability of genetic variance parameter estimation across connectivity levels for a given trial size (number of varieties in the trial), whereas the new \mathcal{D} -optimality diagnostic predicted reliability across both connectivity levels and trial sizes. The application to the real example also suggested that \mathcal{D} -optimality encapsulates numerous structural features of a MET data-set that are influential in determining the reliability of genetic variance parameter estimation. These features included, but are not limited to, variety connectivity, trial size and variety replication. Computation of the diagnostic requires specification of variance parameters, namely genetic and residual variances for individual environments and genetic correlations between pairs of environments. Clearly we do not know these values or how they differ between environments prior to the MET analysis. In terms of the latter a pragmatic and sensible approach is

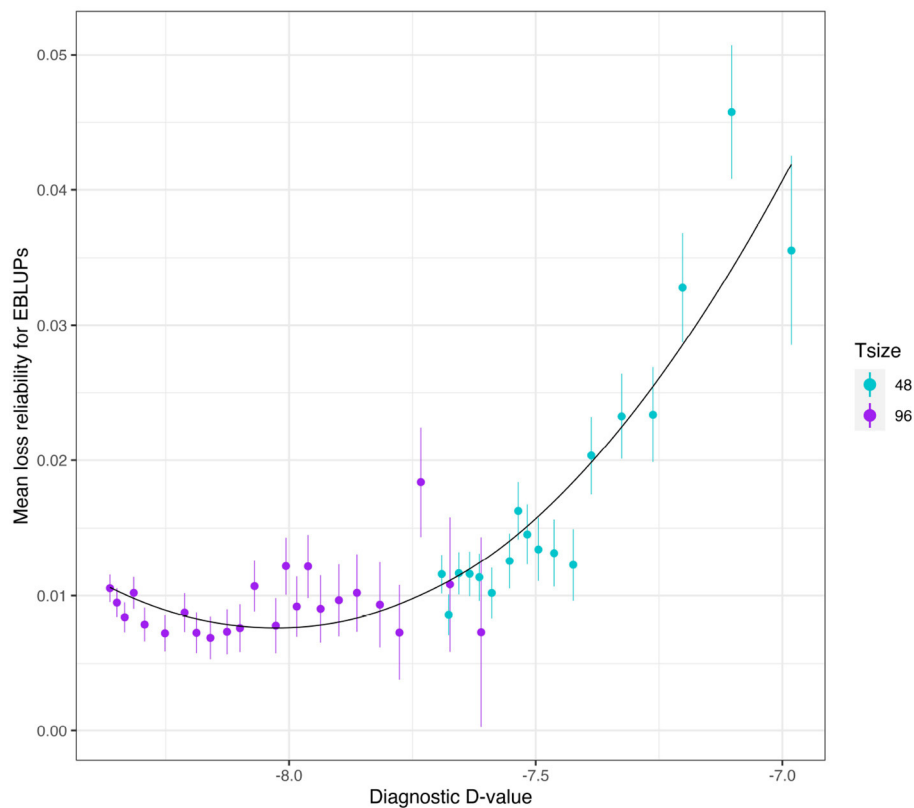


FIGURE 11 | Additive VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. The colours correspond to different trial sizes (trials with 48 and 96 varieties) and the points for each colour correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through all the means.

to assume homogeneity between environments. The parameter values can be chosen to reflect typical estimates obtained in practise. Both the simulation studies and the real example showed that the diagnostic was robust to the specification of these values.

The pre-specified LMM for the \mathcal{D} -optimality diagnostic allowed for the estimation of separate genetic variances for all environments and separate genetic covariances for all pairs of environments. It is therefore targeting MET analyzes that employ factor analytic, or possibly unstructured forms for the genetic variance matrices. It is easy to modify the pre-specified LMM to reflect simpler models, for example, variance component models. Additionally, the partitioning of the genetic effects into additive and non-additive was achieved in this paper using pedigree information but the modification to use genomic (marker) data is straightforward.

The simulation study results suggested that trials with small numbers of varieties will, in general, have larger \mathcal{D} -values when compared with trials with more varieties. Even in the case of 100% connectivity, the smallest trial size considered (12 varieties in each of the two trials) resulted in large \mathcal{D} -values which then translated to substantial losses in the reliability of VE effect predictions. Additionally, the number of successful

model fits was much lower compared with scenario in which there were more varieties in each trial. This is consistent with our experience in analysing MET datasets in which many trials have small numbers of varieties. Even when the connectivity between these and larger trials is high, there are often computational difficulties in fitting the FALMM. In practise, we therefore suggest that individual environment diagnostic values should be examined for a given MET dataset in order to identify environments with large \mathcal{D} -values. These environments may contribute insufficient information for genetic variance parameter estimation so their inclusion in the MET dataset should be carefully considered. Additionally, examination of the overall diagnostic value across all environments may be useful. If the overall \mathcal{D} -value is large this may indicate insufficient information to fit the gold standard FALMM and it may only be possible to fit a simpler model, such as a variance component model.

DATA AVAILABILITY STATEMENT

The data analysed in this study is subject to the following licenses/restrictions the datasets presented in this article are not

readily available because they are owned by New South Wales Department of Primary Industries and the Grains Research and Development Corporation. Requests to access these datasets should be directed to Gururaj Kadkol, gururaj.kadkol@dpi.nsw.gov.au.

AUTHOR CONTRIBUTIONS

CL, BC, and AS conceived the ideas. CL developed and applied the methodology, prepared the data, conducted the simulation study, and prepared the first draft of the manuscript. All authors contributed to the general discussions, manuscript revision, read, and approved the submitted version.

REFERENCES

- Atkinson, A., Donev, A., and Tobias, R. D. (2007). *Optimum Experimental Designs, With SAS*. Oxford: Oxford University Press.
- Bueno Filho, J., and Gilmour, S. (2003). Planning incomplete block experiments when treatments are genetically related. *Biometrics* 59, 375–381. doi: 10.1111/1541-0420.00044
- Butler, D. (2013). *On the optimal design of experiments under the linear mixed model* (Ph.D. thesis). The University of Queensland.
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2017). *ASReml-R Reference Manual Version 4*. Technical report, VSN International Ltd., Hemel Hempstead, HP1 1ES.
- Chapman, S., Cooper, M., Podlich, D., and Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agron. J.* 95, 99–113. doi: 10.2134/agronj2003.0099
- Cullis, B. R., Smith, A., Hunt, C., and Gilmour, A. R. (2000). An examination of the efficiency of Australian crop variety evaluation programmes. *J. Agric. Sci.* 135, 213–222. doi: 10.1017/S0021859699008163
- Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274
- Gogel, B., Smith, A., and Cullis, B. R. (2018). Comparison of a one- and two-stage mixed model analysis of Australia's national variety trial southern region wheat data. *Euphytica* 214:44. doi: 10.1007/s10681-018-2116-4
- Kelly, A., Smith, A., Eccleston, J., and Cullis, B. R. (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci.* 47, 1063–1070. doi: 10.2135/cropsci2006.08.0540
- Loeza-Serrano, S., and Donev, A. (2014). Construction of experimental designs for estimating variance components. *Comput. Stat. Data Anal.* 71, 1168–1177. doi: 10.1016/j.csda.2012.10.008
- Mrode, R., and Thompson, R. (2005). *Linear Models for the Prediction of Animal Breeding Models, 3rd Edn*. Wallingford: CAB International.
- Nuga, O., Amahia, G., and Salami, F. (2017). Optimal designs for the restricted maximum likelihood estimators in a random split-plot model. *Lithuanian J. Stat.* 56, 64–71. doi: 10.15388/LJS.2017.13672
- Oakey, H., Verbyla, A., Cullis, B. R., Wei, X., and Pitchford, W. S. (2007). Joint modelling of additive and non-additive (genetic line) effects in Multit-environment trials. *Theor. Appl. Genet.* 114, 1319–1332. doi: 10.1007/s00122-007-0515-3
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.
- Russell, K. (2018). *Design of Experiments for Generalized Linear Models*. New York, NY: CRC Press.

FUNDING

CL and AS were supported by the Grains Research and Development Corporation (GRDC) through the Statistics for the Australian Grains Industry (SAGI) project (UW00009).

ACKNOWLEDGMENTS

The authors thank the reviewers and editor whose comments have led to an improved manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2021.785430/full#supplementary-material>

- Smith, A., and Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214, 143. doi: 10.1007/s10681-018-2220-5
- Smith, A., Cullis, B. R., and Gilmour, A. R. (2001a). The analysis of crop variety evaluation data in Australia. *Aust. N. Z. J. Stat.* 43, 129–145. doi: 10.1111/1467-842X.00163
- Smith, A., Cullis, B. R., and Thompson, R. (2001b). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006-341X.2001.01138.x
- Smith, A., Ganesalingam, A., Kuchel, H., and Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor. Appl. Genet.* 128, 55–72. doi: 10.1007/s00122-014-2412-x
- Smith, A., Ganesalingam, A., Lisle, C., Kadkol, G., Hobson, K., and Cullis, B. R. (2021a). Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs. *Front. Plant Sci.* 11:2325. doi: 10.3389/fpls.2020.623586
- Smith, A., Norman, A., Kuchel, H., and Cullis, B. R. (2021b). Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. *Front. Plant Sci.* 12:737462. doi: 10.3389/fpls.2021.737462
- Ward, R., Spohr, L., and Snell, P. (2019). Rice grain quality: an Australian multi-environment study. *Crop Pasture Sci.* 70, 946–957. doi: 10.1071/CP19194

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lisle, Smith, Birrell and Cullis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.