2023

Information based diagnostics for the optimal construction of Multi-environment trial datasets

Christopher Lisle

Information based diagnostics for the optimal construction of Multi-environment trial datasets.

Christopher Lisle

School of Mathematics & Applied Statistics

University of Wollongong

A thesis submitted for the degree of

Doctor of Philosophy (PhD)

May 2023

Abstract

The objectives of this thesis are to present novel approaches for optimising the construction of multi environment trial (MET) datasets from a series of plant variety trials. These include evaluating varieties in designed trials at various locations and typically across many years. The MET datasets are then analysed to evaluate how well each variety performs in each environment. Although sophisticated and relevant statistical analyses have been proven to increase the reliability of predicted variety by environment (VE) effects, there has been little research into how to construct an appropriate dataset. This thesis fills a void in the literature by providing information-based diagnostics for the optimal construction of the MET dataset.

The approaches are demonstrated using two motivating datasets: the first is a Oat (*Avena sativa*) dataset and the other is a Durum wheat (*Triticum durum* L. ssp. Durum Desf.) dataset. The former is used as an example of a dataset with independent variety effects, whereas the latter is used as an example of a dataset with related variety effects. These are also used for their attributes in the development of real-world grounded simulation studies to examine the performance of the proposed diagnostics and also to investigate established methodologies and concerns.

The breeding process is a progressive system that revolves around the evaluation and selection of superior varieties. This naturally results in datasets with varied levels of balance in terms of the number of varieties in common between environments, a metric known as "variety connectivity". Previously, it was considered that variety connectivity was a primary driver of the reliability of the prediction of the variety by environment (VE) effects. These concerns have traditionally been used in the construction of the MET dataset. This thesis addresses these concerns first through real-world grounded

ABSTRACT

simulation studies that demonstrate the sometimes intricate linkages between genetic scenarios and variety connectivity. As a result, typical variety connectivity measurements are proven to be insufficient as a method to construct MET datasets.

I then provide a systematic approach for the construction of MET datasets for selection in plant breeding programs. Where I first discuss the structure of MET datasets, with the focus on identifying groups of varieties that entered the first stage of testing in the same year, which is denoted as contemporary groups (CGs), and also the establishment of data bands, which are related to trials. This enables a thorough, and complete listing of the trials in which the varieties of interest were grown and their progression between years and stage. The approach for increasing the amount of data available for the varieties under consideration is simple and straightforward, with only a few steps.

To quantify different MET datasets, I employ from model-based design theory the $\mathcal{A}$-optimality criterion, since this aligns with minimising the probability of an incorrect selection decision. Then I propose the use of the $\mathcal{D}$-optimality criterion to assess the variance of the residual maximum likelihood (REML) estimated variance parameters. In comparison to traditional connectivity type measures, the $\mathcal{D}$-optimality diagnostic is shown to provide a superior diagnostic since it encapsulates not only variety connectivity but also other structural features such as trial size, replication, and genetic relatedness. Thus, is shown to result in better forecasting of the uncertainty of genetic variance parameter estimates in the construction of a MET dataset.

This thesis is shown to address a void in the literature, by providing a rigorous and formal framework for the optimal construction of MET datasets for selection in plant breeding programs. This is accomplished by striking the balance between maximising the variety information through the use of the CG methodology, and maximising the reliability of variance parameter estimates through the use of the diagnostic $\mathcal{D}$-value.

This thesis is dedicated to my wife, Erin, who has been a continuous source of love, encouragement, and support in my endeavours to balance job, studies, and life. I am very grateful to have you in my life. She has discussed ideas and avoided various detours. She also took charge of our family at a period of my studies that was extremely challenging in recent years owing to the COVID-19 pandemic. Without her support I believe I would not have been able to accomplish all that I have.

I love you with all my heart.

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me. It would have not been possible without all the support I received from many people. Highlights of this journey include a trip to France and a sharing/co-win of the JB Douglas award.

Firstly, I want to thank my supervisors, Alison Smith, Carole Birrell, and Brian Cullis, for giving me this opportunity to improve and for their encouragement throughout. There are no words to express how grateful I am. It was a long and difficult journey juggling studies, job, and life, not to mention COVID and homeschooling, but it was well worth it. During this time, my self-assurance and comprehension have increased, allowing me to become a very capable and well-rounded statistical consultant.

Foremost, a heartfelt thank you to Alison; I will be eternally grateful. Thank you for your patience, since I know there have been times when I have gone astray, done something dodgy, or have misunderstood concepts until you have laid things out in black and white. It has been a trip full of twists and turns, and I would not have crossed the finish line if it wasn't for your consistent encouragement and motivation. I am proud of what I have accomplished in the previous several years, and I know you are as well. However, this is a heartfelt thanks to your supervision, advice, and friendship.

To Carole, thank you for your unwavering support and for serving as my sounding board. I've enjoyed our chats over a Mocha on campus or in Jamberoo during the countless COVID lockdowns. Your sharp eye in reading my thesis was fabulous, catching a number of concepts that were not obvious, as well as odd spelling mistakes and symbols that did not fit.

ACKNOWLEDGEMENTS

To Brian, I don't have enough room here to express my gratitude. I can't believe it's been 20 years since I first started working with you! You have always been there for me and pushed me to do my best. I know I still have a lot to learn and that I may never be the statistician you genuinely encourage me to be, but you have always been there for me. We have not always agreed, but the older and wiser I am, the more I realise you are (generally) always correct. You have been more than a supervisor, colleague, or boss; you have been family.

My PhD journey would not have been possible without the continual support of my friends and colleagues, particularly Dini Ganesalingam and Beverley Gogel, and their extended commitments to take care of most of the consulting burden during much of my study. Thank you to David Hughes and Nicholas Lambert, the CBBoys, for making my time at work more enjoyable. Sorry for disappearing for the past few years, but I will be back, and will continue crushing you both on the squash court.

Finally, I would like to thank my family, particularly my wife Erin and our four children, Grace, Olivia, Evan, and Sierra. My studies have pulled me away from day-to-day activities at times, but I have always done my best to balance work, study, and home life, even if it appears that I have mostly locked myself up in my office. I swear that now that my studies are nearly through, I will be back to taking the kids to the park, the beach, and walking Daisy. I would not have been able to complete, or even attempt to complete, my studies if it hadn't been for my families sacrifice.

Declaration

I, Christopher James Lisle, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Mathematics and Applied Statistics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Christopher James Lisle.
May, 2023

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Publications, conference and local presentations

Publications

Lisle, C., Smith, A., Birrell, C,. & Cullis, B. R. (2021). Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials. *Frontiers in Plant Science*. **12**, 2856. doi: 10.3389/fpls:2021.785430.

Abstract: 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 chapter we have provided the link between the objectives of this work and those in model-based experimental design. I 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. I 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.

Chapters 9 reproduces the publication above with notational changes consistent with the nomenclature used in this thesis and additional results added.

PUBLICATIONS, CONFERENCE AND LOCAL PRESENTATIONS

Smith, A., Ganesalingam, A., Lisle, C., Smith, A., Kadkol, G., Hobson, K. & Cullis, B. R. (2021). Use of Contemporary Groups in the Construction of Multi-Environment Trial Datasets for Selection in Plant Breeding Programs. *Frontiers in Plant Science*. **11**, 2325. doi: 10.3389/fpls:2020.623586

Abstract: Plant breeding programs use multi-environment trial (MET) data to select superior varieties, with the ultimate aim of increasing genetic gain. Selection accuracy can be improved with the use of advanced statistical analysis methods that employ informative models for variety by environment interaction, include information on genetic relatedness and appropriately accommodate within-experiment error variation. The gains will only be achieved, however, if the methods are applied to suitable MET datasets. In this chapter I present an approach for constructing MET datasets that optimizes the information available for selection decisions. This is based on two new concepts that characterize the structure of a breeding program. The first is that of "contemporary groups" which are defined to be groups of varieties that enter the initial testing stage of the breeding program in the same year. The second is that of "data bands" which are sequences of experiments that correspond to the progression through stages of testing from year to year. MET datasets are then formed by combining bands of data in such a way as to trace the selection histories of varieties within contemporary groups. Given a specified dataset, we use the $\mathcal{A}$-optimality criterion from the model-based design literature to quantify the information for any given selection decision. I demonstrate the methods using a motivating examples from a Durum breeding program. Datasets constructed using contemporary groups and data bands are shown to be superior to other forms, in particular those that relate to a single year alone.

Chapter 7 presents a reproduction of key sections of this publication, with notational changes consistent with the nomenclature used in this thesis. Note that although the candidate is not the first author of this paper, the candidate had a key role in dataset curation, application of the methodology to the examples, and manuscript preparation and revision.

Kadkol, G., Sissons, M., Lambert, N., & Lisle, C. (2022). Genetic improvement in grain yield and quality of Australian Durum wheat over six decades of breeding. *Cereal Chemistry*. 1–22. doi: 10.1002/cche.10562

Background and Objectives: Durum wheat breeding commenced in Australia in the 1930s by the New South Wales Department of Primary Industries (NSW DPI), Australia. Dural was developed in 1956 from a cross between North African landraces. Since then, another 20 varieties have been released in the following six decades by NSW DPI and the University of Adelaide. These were evaluated for agronomic and detailed quality traits including pasta-making quality over three seasons. This study aimed to quantify progress achieved in Australian Durum breeding for yield, quality, and some agronomic traits since the release of Dural.

Findings: A demonstrated grain yield improvement at the rate of 27.8 kg ha-1 year-1 up to 2017 has been achieved when generally maintaining low screenings and high thousand grain weight. The recent varieties are medium early in their maturity relative to Dural and possess improved lodging tolerance which has resulted in better adaptation to seasonal conditions under dryland cultivation, and also, adaptation to high input irrigated cropping. Breeding has also resulted in higher technological quality, specifically, improved semolina yellow color, higher dough strength, improved pasta brightness and yellowness but with slightly declining grain protein content. Strong genetic gain was observed in semolina color traits, namely, reduction in a* and increase in b*.

Conclusion: The progress in Australian Durum breeding over the last six decades compares well to the progress achieved in other countries for yield, agronomic, and quality traits. The improvements in yield and quality are expected to continue with future developments focussing on adaptation to meet the changing climate and improving tolerance to a major disease, crown rot, while maintaining high technological quality.

Significance and Novelty: This is the first study quantifying the progress made in Australian Durum breeding efforts and the first study that includes pasta-making quality traits from an historical breeding perspective.

This paper summarises the historic successes made in the Durum breeding program, which is the source of the Durum dataset presented in Chapter 4.

Conference and local presentations

*Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2022). A new diagnostic to assess information available for genetic variance parameter estimation in multi-environment trials. *23rd Annual JB Douglas Postgraduate Awards Day. Aerial Function Centre, Sydney.*
*Joint winner https://www.statsoc.org.au/News-and-media-releases/13017830.

Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2022). A new diagnostic to assess information available for genetic variance parameter estimation in multi-environment trials. *XVIIIth Eucarpia Biometrics in Plant Breeding Conference. Paris, France.*

Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2022). A new diagnostic to assess information available for genetic variance parameter estimation in multi-environment trials. *Australian Plant Breeding Conference 2022. QT Gold Coast, Australia.*

Lisle, C. (2021). Use of Contemporary Groups in the Construction of Multi-Environment Trial Datasets for Selection in Plant Breeding Programs. *SAGI Symposium. Mercure Clear Mountain Lodge, Brisbane, Australia.*

Lisle, C. (2021). A new diagnostic to assess information available for variance parameter estimation in Multi-Environment Trials (METs). *NIASRA Presentation Series.*

Lisle, C. (2020). A new diagnostic to assess information available for variance parameter estimation in METs. *CBB Seminar Series.*

Lisle, C. (2019). Varietal Connectivity: does it affect the accuracy of variety predictions from factor analytic multi-environment trial analyses? *Wheat Breeding Assembly. Adelaide Oval, Adelaide, Australia.*

Lisle, C., Hughes, D., & Mathews, K. L. (2019). Evaluating the Effect of the Nugget Variance in the Spatial Analysis of National Variety Trials. *International Biometric Society Australasian Region Conference. Adelaide Botanic Gardens, Australia.*

Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2018). Varietal connectivity: Does it affect the accuracy of results from a multi-environment trial analysis? *Australasian Applied Statistics Conference. Rotorua, New Zealand.*

PUBLICATIONS, CONFERENCE AND LOCAL PRESENTATIONS

Lisle, C., Dungey, H. S., Jefferson, P., Smith, A., Birrell, C., & Cullis, B. R. (2016). Analysis of longitudinal data in a MET framework for *Pinus Radiata*. *Australasian Applied Statistics Conference. Barragga Bay, Australia.*

Glossary

Symbols

m	Number of varieties.
p	Number of environments.
b	Number of field replicates in a trial.
c	Number of field columns in a trial.
r	Number of field rows in a trial.
n	Number of field plots in a trial.
M	Total number of unique varieties across environments.
d	Number of variety by environment combinations.
t	Generation interval.
t/ha	Tonnes per hectare.
$\boldsymbol{y}$	Vector of responses.
$\boldsymbol{y_2}$	Vector of predicted responses.
$\boldsymbol{H}$	Variance of the data vector $\boldsymbol{y}$.
$\boldsymbol{H_2}$	Variance of the data vector $\boldsymbol{y_2}$.
$\boldsymbol{X}$	Design matrix for fixed effects.
$\boldsymbol{\tau}$	Vector of fixed effects.

$\boldsymbol{Z_g}$	Design matrix for random genetic effects.
$\boldsymbol{u_g}$	Vector of random genetic effects.
$\boldsymbol{u_a}$	Vector of random additive genetic effects.
$\boldsymbol{u_e}$	Vector of random non-additive genetic effects.
$\boldsymbol{Z_p}$	Design matrix for random non-genetic (peripheral) effects.
$\boldsymbol{u_p}$	Vector of random non-genetic (peripheral) effects.
$\boldsymbol{e}$	Vector of errors.
$\boldsymbol{\eta}$	Full $mp \times 1$ vector of variety mean parameters for individual environments.
$\boldsymbol{\eta}_d$	$d \times 1$ sub-vector of the VE combinations present in the data.
$\hat{\boldsymbol{\eta}}_d$	Vector of predicted VE means.
$\boldsymbol{D}$	$d \times mp$ Indicator matrix.
$\boldsymbol{\Omega}$	Variance matrix for $\hat{\boldsymbol{\eta}}_d$.
$\boldsymbol{\xi}$	Error variance matrix for two-stage approach.
$\boldsymbol{\Sigma}$	Error variance matrix.
$\boldsymbol{\Sigma_c}$	Correlation matrix in the column direction.
$\boldsymbol{\Sigma_r}$	Correlation matrix in the row direction.
$\boldsymbol{G_g}$	Genetic variance matrix.
$\boldsymbol{G_a}$	Additive genetic variance matrix.
$\boldsymbol{G_e}$	Non-additive genetic variance matrix.
$\boldsymbol{G_p}$	Non-genetic variance matrix.
σ^2	Error variance.
σ_v^2	Variety variance.

σ^2_{ve}	Variety by environment variance.
σ^2_g	Genetic variance.
σ^2_a	Additive genetic variance.
σ^2_e	Non-additive genetic variance.
ρ_c	Correlation parameter for the column dimension.
ρ_r	Correlation parameter for the row dimension.
$\boldsymbol{A}$	The numerator relationship matrix (NRM).
$\bar{a}$	Average inbreeding coefficient.
F_i	Inbreeding coefficient.
f_{ij}	Coefficient of parentage between varieties i and j.
ℓ_R	Residual log-likelihood.
$\boldsymbol{\mathfrak{I}_A}$	Average information matrix.
$\boldsymbol{\mathfrak{I}_E}$	Expected information matrix.
$\boldsymbol{\mathfrak{I}_{2E}}$	Expected information matrix for a two-stage approach.
$\boldsymbol{\mathfrak{I}_O}$	Observed information matrix.
k	Number of factors fit in the FA model.
k_a	Number of additive factors fit in the FA model.
k_e	Number of non-additive factors fit in the FA model.
$\boldsymbol{\Lambda}$	Matrix of environment loadings.
$\boldsymbol{\Lambda_a}$	Matrix of additive environment loadings.
$\boldsymbol{\Lambda_e}$	Matrix of non-additive environment loadings.
$\boldsymbol{\Psi}$	Diagonal matrix of specific environment variances.

$\boldsymbol{\Psi_a}$	Diagonal matrix of specific additive environment variances.
$\boldsymbol{\Psi_e}$	Diagonal matrix of specific non-additive environment variances.
$\boldsymbol{f}$	Vector of variety scores.
$\boldsymbol{f_a}$	Vector of variety additive scores.
$\boldsymbol{f_e}$	Vector of variety non-additive scores.
$\boldsymbol{\delta}$	Vector of VE lack of fit effects.
$\boldsymbol{\delta_a}$	Vector of additive VE lack of fit effects.
$\boldsymbol{\delta_e}$	Vector of non-additive VE lack of fit effects.
$\boldsymbol{\beta}$	VE regression component.
$\boldsymbol{\beta_a}$	Additive VE regression component.
$\boldsymbol{\beta_e}$	Non-additive VE regression component.
S	Total number of simulations.
$x_{1,2}$	Number of varieties in common between two trials.
$\bar{R}$	Model-based estimate of trial reliability.
R_k^D	Design-based reliability value for variety k.
R_{kc}^D	Design-based reliability value for variety k and connectivity level c.
$\bar{R}^D$	Design-based trial reliability.
R_k^S	Simulated reliability value for variety k.
R_{kc}^S	Simulated reliability value for variety k and connectivity level c.
$\bar{R}^S$	Simulated trial reliability.
$\bar{R}_{loss}$	Difference between design-based and simulated trial reliabilities.
$\bar{R}_X$	Maximum potential trial reliability.

$\bar{R}_{X_p}$	Proportion of maximum potential trial reliability achieved.
$\boldsymbol{\kappa}$	Variance parameters.
$\boldsymbol{\kappa_g}$	Genetic variance parameters.
$\boldsymbol{\kappa_{g0}}$	Chosen genetic variance parameters.
n_{κ_g}	Number of genetic variance parameters.
$n_{\kappa_{g_j}}$	Number of genetic variance parameters for environment j.
$\mathcal{A}$	$\mathcal{A}$-value.
$\mathcal{D}$	Diagnostic $\mathcal{D}$-value.
$\mathcal{D}_j$	Diagnostic $\mathcal{D}$-value for environment j.
$\mathcal{D}$(A)	Diagnostic $\mathcal{D}$-value for additive VE variance parameters.
$\mathcal{D}_j$(A)	Diagnostic $\mathcal{D}$-value for additive variance parameters for environment j.
$\mathcal{D}$(I)	Diagnostic $\mathcal{D}$-value for non-additive (independent) VE variance parameters.
$\mathcal{D}_j$(I)	Diagnostic $\mathcal{D}$-value for non-additive (independent) variance parameters for environment j.
$\mathcal{D}$(A+I)	Diagnostic $\mathcal{D}$-value for additive and non-additive VE variance parameters.
$\mathcal{D}_j$(A+I)	Diagnostic $\mathcal{D}$-value for additive and non-additive variance parameters for environment j.

Terminology and acronyms

AHDB: Agriculture and Horticulture Development Board.

AIC: Akaike information criteria. Used to compare different models and determine which one is the best fit for the data.

Accuracy: The correlation between the true and predicted effects.

ADD%: Percentage of additive variance to total.

ANOVA: Analysis of variance. A traditional collection of statistical models and their associated estimation procedures.

AR1: Separable auto regressive of order 1. The spatial correlation model generally used in the analysis of field trials.

ARAM: Approximate reduced animal model.

ASReml-R: The linear mixed model `R` package used to analyse LMMs.

BLUE: Best Linear Unbiased Estimate.

BLUP: Best Linear Unbiased Predictor.

Breeder trial: A comparative variety trial managed by a public breeding program.

CC: Compound covariance. A model which assumes a common variance and covariance within and between trials.

CG: Contemporary Groups. A cohort of test varieties produced jointly at the crossing block stage.

CRD: Completely randomised design. A design which allows treatments to be completely randomised to experimental units.

CVE: Common Variety by environment effect. The component of the VE effect that corresponds to the linear combinations of the common factors in the FA model.

Data band: A group of trials aligning with the testing system, namely the progression through stages from year to year.

Durum: A type of spring wheat generally used for semolina, and to make pasta and dough.

Diagonal model: A model that can be used in the MET analysis to provide an independent structure for the between trials genetic variance matrix.

Design matrix: A matrix which maps the incidence of a corresponding effects to the data.

Design based values: Theoretical based values derived using `ASReml-R`. These are generally derived reliability type measures on the BLUPs.

EBLUE: Empirical Best Linear Unbiased Estimate. All fixed effects in the linear mixed model analysis are estimated using the method of best linear unbiased estimation, but with variance parameter estimates replaced with their REML estimates.

EBLUP: Empirical Best Linear Unbiased Predictor. All random effects in the linear mixed model analysis are predicted using the method of best linear unbiased prediction, but with variance parameter estimates replaced with their REML estimates.

Environment: A year and location combination, often comprises of multiple trials.

Environment loading: Factor loading for an environment from a FALMM.

FAk: Factor analytic model of order k.

FAk_a, k_e: Factor analytic model of order k_a and k_e for additive and non-additive effects.

FALMM: Factor analytic linear mixed model.

FAST: Factor analytic selection tools. Methodology derived from Smith & Cullis (2018); Smith et al. (2021b) to summarise VE effects from a FALMM.

GC: Genetic correlation.

GRDC: Grains Research and Development Corporation. The funding body for NVT, and global leader in grains industry research and development.

GV: Genetic variance.

IID: Independent and identically distributed. Defining a structure with no covariance.

LMM: Linear Mixed Model.

MET: Multi-environment Trials. A series of comparative variety trials grown in different trials (typically indexed by year and geographic location).

MSE: Mean squared error.

MSEP: Mean squared error of prediction.

NDBA: Northern Durum Breeding Australia.

NOBP: National Oat Breeding Program.

NRM: Numerator Relationship Matrix. This is a symmetrical matrix that represents the genetic relationships between varieties, which assumes inheritance laws for correlated genetic (additive) effects.This is built from a pedigree file.

NVT: National Variety Trials. A comparative variety trial managed by GRDC. Each trial comprises a single randomisation of varieties to a set of field plots.

ODW: The model-based optimal design `R:` package used to design plant breeding trials.

One-stage analysis: A linear mixed model analysis of a dataset that comprises individual plot yield data combined across all trials.

Pedigree file: A structured representation of a varieties ancestral links. Normally in the form of 'Me', 'Mum', and 'Dad.

PEV: Prediction error variance.

p-rep: Partially replicated. A standard design used for the early stages of plant breeding trials where some varieties are replicated and others with just one replicate.

RCB: Randomised complete blocks. A standard design used for plant breeding trials where similar experimental units are grouped into blocks or replicates.

Reliability: The square of the accuracy value.

REML: Residual Maximum Likelihood. The method of estimation for variance parameters in the linear mixed model analysis.

REMLRT: Residual maximum likelihood ratio test.

RGG: Rate of genetic gain. The variable in the breeders equation.

RL: Recommended list.

SEM: Standard error of the mean.

Specific variance: The error term in the FALMM model.

TMY: The mean yield of all field plots at a trial.

Trial: Each trial comprises a single randomisation of varieties to a set of field plots.

Tsize: Trial size. A term used in the simulation studies to distinguish how many varieties are in a trial.

Unstructured model: A model that can be used in the MET analysis to provide a structure for the between trials genetic variance matrix. This is a completely saturated form so that for p environment there contains $p(p+1)/2$ parameters.

VAF%: Variance accounted for. The percentage of VE variance for a trial that is accounted for by the common factors in an FA model.

Variety: An entry in a trial or environment.

Variety connectivity: Number of unique varieties in common between pairs of trials or environments.

VE: Variety by Environment.

VEI: Variety by Environment Interaction.

Variety score: Factor scores for a variety from a FALMM.

Chapter 1

Introduction

Plant breeding is a branch of agriculture that focuses on manipulating plant heredity to develop new and improved plant varieties for use by society (Acquaah, 2013). It consists of methods for the creation, selection, and fixation of superior plants in terms of productivity or quality (Moose & Mumm, 2008). During this process, the ability to select the best varieties and discard others is critical in constantly improving the breeding gene pool (Zamir, 2001). The new varieties might have a greater yield, better grain quality, stronger disease resistance, better agronomical qualities, and better quality attributes. Ideally, they will have a new set of characteristics that are superior to the existing varieties (Luckett & Halloran, 1995).

Breeding of new crop varieties has been, and continues to be, a vital means for meeting global food demands. Increased productivity is largely achieved by creating new varieties with higher yields. The plant breeding process that targets grain yield is a lengthy one and comprises a number of successive stages in which the newly developed breeding lines are grown in designed field trials. Note that the term "breeding line" is technically correct in this context, but to be consistent with the majority of literature, it will be replaced with the term "variety" for the remainder of this thesis. In the first stage (often called Stage 1, S1), a large number of varieties is grown in a small number of trials at different geographic locations and with low levels of within trial replication. A proportion of varieties from S1 are then selected, on the basis of superior yields, to progress to the next stage of testing (Stage 2, S2) in the following year. The process continues through to Stage 3 (S3), and Stage 4 (S4) with decreases

in the number of varieties being grown in each subsequent trial, but with higher levels of within-trial replication and larger numbers of trials spread across the target growing region. The culmination of the testing process is the identification of the best new varieties that may be suitable for use by growers. In Australia, the elite varieties which are considered for commercial release at the end of S4 are evaluated in the Grains Research and Development Corporation (GRDC) funded National Variety Trials (NVT) program (www.grdc.com.au/research/trials,-programs-and-initiatives/national-variety-trials). In this system, varieties from all major plant breeding companies are tested together in large numbers of trials with standardised management in order to provide independent information for growers.

Identification of superior varieties at any stage in a plant breeding program or in a testing program such as NVT, is achieved using an analysis of yield data combined across a series of trials, also known as a multi-environment trial (MET). In this context, "environments" are defined as the combinations of the geographic locations and years in which the trials were conducted. The analysis of MET data is an essential component of variety evaluation as it allows the investigation of varietal yield performance across a range of locations and seasonal conditions.

1.1 Historical methods for multi-environment trial analysis

Historically, MET data were analysed using a two-stage process in which the first stage consisted of separate analyses of individual trials to obtain variety means for each trial (and possibly an associated measure of uncertainty). The resultant variety by trial table of means was then used as "pseudo" data in a second stage, across trial analysis. One of the earliest approaches for the second stage analysis was the analysis of variance (ANOVA) for a series of experiments as described in Cochran & Cox (1950). This method partitions sources of variation associated with the main effects of varieties, the main effects of environments, variety by environment interaction (VEI) and within trial error variation (a pooled estimate obtained from the first stage individual trial analyses). Implicitly, the main effects and interactions are regarded as fixed effects so, in particular, the variety effects and interactions are estimated using least squares.

A major limitation with the ANOVA approach is that it requires complete data, that is, observations from all varieties in all environments. This is rarely the case so that the application of this method, or indeed any method that requires complete data, typically involves sub-optimal practices such as taking subsets of data or substituting imputed values in the missing cells of the variety by trial table. The problem of incomplete variety by trial tables was addressed in several key papers led by H. D. Patterson (Patterson et al., 1977; Patterson, 1978; Patterson & Silvey, 1980). These papers describe approaches for analysing incomplete MET datasets using models in which variety main effects are regarded as fixed effects but environment main effects and VEI are regarded as random effects. Thus, these are examples of linear mixed models (LMMs) although this terminology was not used by the authors. In the LMMs, both the environment effects and VEI are assumed to comprise independent and identically distributed sets of effects, each with an associated variance component. Patterson et al. (1977) were the first to consider partitioning VEI into several sources associated with the geographic location ("centre") and year of the trial. Thus, their model includes random effects for variety by centre, variety by year and variety by centre by year interactions, each with its own variance component. They are also the first to introduce the use of residual maximum likelihood (REML) (Patterson & Thompson, 1971) for the estimation of variance components in unbalanced MET data. Using a similar model, and with reference to incomplete variety by trial data, Patterson (1978) details a least squares method for estimating variety means (across environments). Finally, Patterson & Silvey (1980) provides a comprehensive overview of the methods of MET analysis used in the United Kingdom crop variety testing system (an analogous system to the NVT). The models and estimation methods are as given in the earlier two papers, but, importantly, even though they still regard variety effects as fixed effects in their model, (Patterson & Silvey, 1980, p. 230) compute an estimate of genetic variance ("the variance among the population of variety means") which they comment "enables us to calculate the gains in varieties selected for recommendation." They also raise the issue of "selection bias", noting that this "arises because a variety is more likely to be recommended if its trial mean yield exceeds its true mean yield". They quantify this bias using the ratio of genetic variance to total variance. These concepts align more closely with the assumption of random, rather than fixed, variety effects. In fact, the bias adjustment provided by Patterson & Silvey (1980) is consistent with the phenomenon of "shrinkage" associated

with the prediction of random variety effects using the technique of best linear unbiased prediction (BLUP) (Robinson, 1991).

Following on from this work, Cullis et al. (1996a,b) consider the analysis of a large and highly unbalanced Australian MET dataset. They use a LMM similar to Patterson & Silvey (1980) in the sense that there are variety and environment main effects, and VEI is partitioned into sources associated with location and year. They also include some genetic covariates so that the variety main effects and VEI are partitioned further. Cullis et al. (1996b) assume fixed environment main effects and fixed effects for the genetic covariates. All other effects, including the variety main effects (adjusted for the covariates) are assumed to be random. They used REML for the estimation of variance components. Furthermore, they used BLUP to obtain predictions of variety main effects and key interactions and showed how these could be used for selection. This is arguably the first published use of BLUP for selection in a MET. The predictions were calculated using REML estimates of the variance parameters so that they were, in fact, *empirical* best linear unbiased predictions (EBLUPs). This distinction between BLUPs, which are based on known variance parameters, and EBLUPs, which are based on estimates of variance parameters, is an important theme in this thesis.

The next major step change in the analysis of MET data was the work of Cullis et al. (1998) who advocated replacing the two-stage approach with a one-stage analysis of individual plot data combined across trials. They argued that with incomplete data and non-orthogonal within-trial analyses, the efficiency loss in the two-stage approach could be substantial, particularly with low levels of within-trial replication. Their reference to non-orthogonal within-trial analyses was particularly important, since, by this time, it was widely accepted that spatial methods of analysis provided significant gains in accuracy for field trials. In their fully efficient one-stage approach for MET data, Cullis et al. (1998) used a LMM with fixed environment main effects, random variety main effects and random VEI. They recognised that VEI variance often differed between environments, so in their model they included a separate VEI variance parameter for each environment. At the within-trial level, Cullis et al. (1998) adopted the spatial modelling ideas of Cullis & Gleeson (1991) and Gilmour et al. (1997). These are based on the fact that field trials are typically arranged as rectangular arrays of plots indexed

by rows and columns. A two-dimensional (row × column) separable spatial correlation model is used for the errors. The correlation structure for each dimension is given by an autoregressive process of order one which is a function of a single autocorrelation parameter and reflects the fact that the correlation between plot errors decreases as the distance between plots increases. The Cullis et al. (1998) approach proceeds by first analysing each trial separately in order to assess the adequacy of the spatial models, to diagnose the existence of additional fixed or random terms (typically associated with rows and/or columns) that may need to be added to the model and to identify potential outliers. Once satisfactory within-trial models have been identified, they are carried through to the LMM for the MET analysis where they are re-estimated. Thus, the variance parameters to be estimated in the one-stage analysis include not only the genetic variance parameters (associated with the variety main effects and VEI) but also variance parameters associated with individual trial designs (for example, associated with replicate block factors) and individual trial spatial variances and autocorrelations.

Cullis et al. (1998) was a landmark paper in raising awareness of the serious deficiencies in the two-stage approach for the analysis of MET data. The efficiency losses associated with this approach have also been demonstrated (Gogel, 1997; Welham et al., 2010; Gogel et al., 2018) so that the one-stage approach is accepted as being necessary to achieve accurate selection in a plant breeding program.

1.2 Factor analytic linear mixed models for multi environment trial analysis

In all the historic methods of MET analysis discussed in Section 1.1, the logic follows the standard statistical approach for a factorial experiment in which the effects are partitioned into main effects and interactions. Smith et al. (2001b) moved away from this framework and adopted the quantitative genetics view in which yields in different environments are synonymous with different traits (Falconer, 1952). It is therefore natural to consider estimation of a genetic variance matrix for the environments. Such a matrix comprises genetic variances for each environment (reflecting the magnitude of variation in yield between varieties in individual environments) and genetic covariances between pairs of environments (which, when expressed as correlations, reflect the

agreement/disagreement in variety rankings). Smith et al. (2001b) achieved this by proposing a LMM that included environment main effects and the variety effects for individual environments (henceforth called the VE effects). Note that no variety main effects are included in the model and that VE effects represent the nested effects of varieties within environment and not VEI. Given the multi-trait analogy, Smith et al. (2001b) fit the environment main effects as fixed effects and the VE effects as random effects with a two-dimensional (Variety $\times$ Environment) separable variance structure. If there are p environments and m unique varieties across the entire MET dataset then there are mp VE effects (even if the data are unbalanced). If they are ordered as varieties within-environments, the variance structure can be written as $\boldsymbol{G_e} \otimes \boldsymbol{G_v}$, where $\boldsymbol{G_e}$ is a $p \times p$ variance matrix for the environment dimension (the between environment genetic variance matrix) and $\boldsymbol{G_v}$ is an $m \times m$ matrix for the variety dimension.

Smith et al. (2001b) note that there are many possible forms for $\boldsymbol{G_e}$, the most parsimonious of which is the compound symmetric (or uniform) structure that arises from the historic MET model with random variety effects (with variance component σ^2_v) and VEI (with variance component σ^2_{ve}). In this case $\boldsymbol{G_e}$ only involves 2 parameters, with every genetic variance (diagonal element) being given by $\sigma^2_v + \sigma^2_{ve}$ and every genetic covariance (off-diagonal element) by σ^2_v. Partitioning VEI into sources associated with locations and years (Cullis et al., 1996b, as in), or allowing for VEI variance heterogeneity (as in Cullis et al., 1998) provides slightly more general, but still restrictive forms for $\boldsymbol{G_e}$, none of which regularly provide a good fit to MET data. The most general form for $\boldsymbol{G_e}$ is the unstructured form as is used in multi-trait analyses and which involves $p(p+1)/2$ variance parameters. Multi-trait applications typically involve only a few traits (less than 10) whereas METs can involve numbers of environments in excess of $p = 50$. With such large numbers, estimation of the variance parameters in an unstructured matrix will likely be unstable and result in inaccurate estimates (Kelly et al., 2007).

Smith et al. (2001b) propose the use of factor analytic (FA) forms for $\boldsymbol{G_e}$ which arise by assuming a multiplicative model for the VE effects. A factor analytic model of order k, denoted FAk, comprises k multiplicative terms and a residual. Each of the multiplicative terms is a product of a set of variety "scores" and environment "loadings" which

are all estimated from the data. The variety scores are assumed to be random effects and the environment loadings are variance parameters. The residuals are assumed to be random effects with a separate variance (known as a "specific" variance) for each environment. The loadings and specific variances combine to form a quite general form for $\boldsymbol{G_e}$ that allows for both heterogeneity of genetic variance and covariance. It is much more parsimonious than the unstructured form with $p(k+1)-k(k-1)/2$ parameters to be estimated. The Smith et al. (2001b) analysis is similar to Cullis et al. (1998) in that it is a one-stage approach that incorporates within-trial error variance modelling as previously described. This so-called factor analytic linear mixed model (FALMM) provides EBLUPs of the VE effects that can then be summarised across environments in meaningful ways to facilitate variety selection (Smith & Cullis, 2018; Smith et al., 2021b, see).

In the original FALMM paper of Smith et al. (2001b), the VE effects were assumed independent between varieties so that $\boldsymbol{G_v} = \boldsymbol{I}_m$. More recently it has been shown that the accuracy of predicted VE effects can be greatly improved by including information on the genetic relatedness of varieties in the FALMM. This can be achieved using pedigree information for varieties, which is a structural representation of an individual's ancestral links. This information is represented using a Numerator Relationship Matrix (NRM), often denoted as $\boldsymbol{A}$ (Henderson, 1976). A FALMM that incorporates pedigree information has been shown to improve selection accuracy (Oakey et al., 2007; Beeck et al., 2010). The associated model partitions the VE effects into additive and non-additive effects. Separable variance models are assumed for each set and are given by $\boldsymbol{G_a} \otimes \boldsymbol{A}$ and $\boldsymbol{G_e} \otimes \boldsymbol{I}_m$ where $\boldsymbol{G_a}$ and $\boldsymbol{G_e}$ are known as the additive and non-additive between environment genetic variance matrices, respectively. In the FALMM, these variance matrices each have their own factor analytic form. Information on genetic relatedness can also be obtained using genomic (marker) data, in which case it is represented using a Genomic Relationship Matrix (GRM), often denoted as $\boldsymbol{K}$ (VanRaden, 2008). The FALMM proceeds in a similar manner to Oakey et al. (2007) except that the NRM is replaced by the GRM in the variance model for the additive effects (Tolhurst et al., 2019).

The FALMM approach has been found to regularly provide a good fit to MET data and to provide key information on variety performance for selection. It is now widely used for the analysis of Australian MET data and is regarded as the current "gold standard"

method. Most of these analyses are conducted as part of the routine selection process in plant breeding programs so are confidential and rarely published. However, some key articles involving the application of FALMMs to MET data are summarised in Table 1.1. This table provides an insight into the structure of MET datasets that have been analysed using FALMMs. The numbers of environments (p) included in the datasets range from 4 to 196, whilst the number of varieties (m) ranges from 26 to 6951 (excluding the Radiata Pine studies). The degree of incompleteness in each dataset is measured by the "percentage fill-in", that is, the number of variety by environment combinations present in the data expressed as a percentage of the total number of combinations (mp). These range from 6% to 100% (excluding the Radiata Pine studies). Also shown (in the last two columns) are the orders of FA models fitted, and whether the models included information on genetic relatedness.

A key paper that provides direct evidence of the merit of FALMMs for variety selection is Kelly et al. (2007). They consider eight example datasets from Australian plant breeding programs (see Table 1.1 for summary information) and fit one-stage MET analyses following Smith et al. (2001b). They assume independence between varieties so that the variance model for the VE effects is given by $\boldsymbol{G_e} \otimes \boldsymbol{I}_m$. They used a range of models for $\boldsymbol{G_e}$ including a compound symmetric and unstructured model together with FA models of order $k = 1$ up to the maximum order possible for the dataset. The goodness of fit of the models was investigated using AIC which showed that an FA model was the best for 6 datasets and an unstructured for 2 datasets (NSW barley and Qld wheat). This is an interesting result considering that most of these datasets involved relatively small numbers of environments so that the unstructured model might be expected to perform well. Note that one of the datasets for which unstructured was best had the smallest number of environments (4 for NSW barley). Kelly et al. (2007) also consider the ability of the FALMM to provide accurate VE predictions for the purposes of selection. Clearly, selection errors will be minimised when the correlation between the true and predicted VE effects is maximised and the mean squared error of prediction (MSEP) is minimised. (Kelly et al., 2007, p. 1064) comment that this is achieved with the use of BLUP but the proviso is that "the BLUPs are calculated on the basis of the true form for the genetic variance matrix". There is the additional potential loss of accuracy associated with the fact that variance parameters must be estimated so that selection

will be based on EBLUPs rather than BLUPs.

Kelly et al. (2007) investigate the accuracy of VE EBLUPs via a simulation study. They generated data for 12 scenarios comprising the factorial combinations of three numbers of varieties ($m = 80,\ 200,\ 500$) and four genetic variance matrices ($\boldsymbol{G_e}$ with variance parameters taken from the FA2 model fitted to the Qld wheat dataset, the unstructured model fitted to this dataset, the FA2 model fitted to the SA barley dataset, the unstructured model fitted to this dataset). Importantly, the datasets were complete with all varieties present in all trials. In each simulation run, the data were analysed using six models for $\boldsymbol{G_e}$, including diagonal, compound symmetric, unstructured, FA1, FA2 and FA3 models. The accuracy of the VE EBLUPs was investigated using MSEP. When data were generated using the FA2 model, and when the FA2 model was used for analysis, the MSEP was smaller than that when the unstructured model was used. However, when data were generated using an unstructured model, the results differed depending on the number of varieties. For the largest number, that is, $m = 500$, the MSEP was smallest when the unstructured model was used for analysis. In the case of $m = 80$ or $m = 200$ varieties, FA models had MSEP values that were lower or equal to those from an unstructured model. The implication is that, unless the number of varieties is large, there will be insufficient information to reliably estimate the variance parameters in an unstructured model and that this will have a negative impact on the accuracy of the associated VE EBLUPs and thence selection. This is in agreement with Sales & Hill (1976a,b) who demonstrated, in an animal breeding context, that poorly estimated genetic variance parameters reduces genetic gain. The accuracy of the EBLUPs from an FA model were shown to be superior to those from an unstructured model for small numbers of varieties, the implication being that the FA variance parameters are more reliably estimated. The proposal that the reliability of genetic variance parameter estimates will have an impact on the accuracy of VE predictions is a theme that will be explored in this thesis. This will be done in a broad framework that encompasses incomplete MET data.

Table 1.1: Summary of key studies based on the analysis of METs using FALMM, and their dataset composition, including: number of years, environments (Envs), and varieties (Vars); the minimum, maximum and median numbers of varieties in each environment; the percentage of variety by environment combinations observed (%fill-in); and the forms of the separable $\boldsymbol{G_e} \otimes \boldsymbol{G_v}$ variance matrices fitted to the VE effects.

		Number of			Varieties per Env			%	Models for	
References	Dataset descriptions	Years	Envs	Vars	Min	Max	Median	fill-in	$\boldsymbol{G_e}$ [a]	$\boldsymbol{G_v}$ [b]
Smith et al. (2001b)	SA Stage 3 Barley trials conducted in 1997	1	7	172	172	172	172	100	FA3	I
Thompson et al. (2003)	NSW Stage 3 and 4 Barley trials conducted 1999-2001	3	62	216	29	118	32	25	FA2	I
Kelly et al. (2007)	NSW Stage 2 Barley trials conducted in 2004	1	4	321				75	FA3	I
	Qld Stage 2 Barley trials conducted in 2004	1	6	720				75	FA3	I
	Qld Stage 2 Sorghum trials conducted in 2004	1	5	644				69	FA3	I
	Qld Stage 2 Wheat trials conducted in 2004	1	7	1160	1160	1160	1160	100	FA3	I
	SA Stage 2 Barley trials conducted in 2004	1	10	480	480	480	480	100	FA3	I
	Vic Stage 2 Barley trials conducted in 2004	1	6	202	202	202	202	99	FA3	I
	Vic Stage 2 Green Lentil trials conducted in 2004	1	6	50				83	FA3	I
	Vic Stage 2 Red Lentil trials conducted in 2004	1	9	231				91	FA3	I
Beeck et al. (2010)	Australian Early stage Canola trials conducted 2007-2008 for grain yield	2	19	332	153	260	154	38	FA3,3	A,I
	Australian Early stage Canola trials conducted 2007-2008 for oil	2	13	332	153	260	183	40	FA2,3	A,I
Welham et al. (2010)	Australian Late stage Wheat trials conducted in 1998	1	14	34	26	34	29	86	FA3	I
	UK Late stage Wheat trials conducted in 1998	1	12	26	20	26	25	92	FA3	I
Cullis et al. (2014)[c]	Australia and New Zealand Radiata Pine genetic trials conducted 1968-2005	37	77	2733	17	588	86	4	FA3	A
Smith et al. (2015)	NVT Wheat trials from Southern region conducted 2009-2013	5	196	200	36	63	47	24	FA5	I
Gogel et al. (2018)	NVT Wheat trials from Southern region conducted 2011-2015	5	192	188	29	59	47	24	FA2	I
Smith & Cullis (2018)[c]	Australia and New Zealand Radiata Pine genetic trials conducted 1968-2007	39	92	3061	17	588	105	4	FA3	A
Smith et al. (2019)	Australian Canola blackleg trials conducted 2011-2016	6	70	357	20	138	90	21	FA4	I
Tolhurst et al. (2019)	Australian Stage 2 Wheat trials conducted in 2015	1	8	2868	609	2845	1320	21	FA4,2	K,I
Cocks et al. (2019)	Australian Frost expression Wheat trials conducted 2010-2016	7	17	238	28	108	54	14	FA3	I
Smith et al. (2021b)	Australian Stage 3 and 4 Wheat trials conducted 2014-2017	4	73	622	96	320	214		FA4	I
Ferrante et al. (2021)	Australian Frost expression Wheat trials conducted 2010-2019	10	26	264	28	106	53		FA3	I
	Australian Frost expression Barley trials conducted 2012-2019	8	24	66	19	48	35		FA3	I
Chapter 5	Australian Stage 4 Oat trials conducted 2012-2016	5	41	163	48	65	52	33	FA5	I
Chapter 8	Australian Stage 1-4 Durum trials conducted 2014-2018	5	30	6951	60	1836	101	6	FA4,3	A,I
NVT-online	NVT Lentil trials conducted 2017-2021	5	59	31	14	17	15	49	FA4	I

[a] Factor analytic model of order k for independent variety effects; and of order k_a, k_e for those models with variety relationship matrices.
[b] Model for variety effects: Independent (I), numerator relationship matrix (A), and genomic relationship matrix (K).
[c] Reported summaries are by parents, rather than by variety.

1.3 Multi-environment trial dataset construction

MET datasets for variety selection are unique in the sense that there is no single definitive dataset for any given selection decision so that the plant breeder and statistician must determine the data to combine for analysis. Given that the breeding process involves trials that span both geographic locations and years, the decisions as to which trials to include in the dataset can be quite complex. Importantly, the potential gains of using the gold standard FALMM will only be realised if the models are applied to appropriately constructed MET datasets (see Smith et al., 2021a, for example). Despite the importance of this issue, there has been little research and there is a lack of consensus in the literature on how MET datasets should be constructed for variety selection. However, several authors, including Cullis et al. (2000) and Arief et al. (2015), have stressed the importance of including VEI associated with seasonal conditions in the models, so there is a clear need to include multiple years of data in a MET to enable accurate selection. Thus, most of the recent publications in Table 1.1 involve datasets that include trials from several years. Typically, this also involves the combination of trials from different stages of testing. Most of these datasets were compiled using informal approaches, but with the premise that the aim should be to include as many trials as possible to capture all available yield data on the varieties under consideration for selection. This suggests that trials should be combined in order to maximise information (data) on varieties. However, combining data from trials in different years and stages leads to unbalanced (incomplete) data with low percentage fill-in (see Table 1.1). It has been a major concern for some years that incomplete data may adversely affect the reliability of estimation of the genetic variance parameters in the FALMM. As discussed in Section 1.2, poor estimates of genetic variance parameters will reduce the accuracy of the VE EBLUPs and thence selection. It is therefore important to establish whether incomplete data, or other structural features of a MET dataset, may impact on the reliability of genetic variance parameter estimates. This may then provide another criteria for MET dataset construction, namely to maximise the information for estimation of genetic variance parameters. These two aspects of MET dataset construction are the focus of this thesis and are introduced in the following two sections to provide context for this research.

1.3.1 Maximising information on varieties

Smith et al. (2021a) formalise a method for constructing MET datasets to maximise variety information. They introduce two new concepts relating to the two fundamental components in a breeding program, namely the varieties and the trials. On the variety side, they define "contemporary groups" which are groups of varieties that commence their yield testing in the same year. On the trial side, they define "data bands" which are groups of trials across years that reflect the sequence of stages of selection. Smith et al. (2001b) show how to combine data bands to form MET datasets for given selection decisions. Their method maximises the amount of information available for variety selection where information may be "direct" (observed data on the varieties) or "indirect" (gained from genetically related varieties through the inclusion of pedigree or marker data in the analysis). Smith et al. (2001b) also provide a measure to quantify this information for a given MET dataset. They use the $\mathcal{A}$-optimality criterion from the model-based experimental design literature. This criterion was chosen since, when calculated with respect to random treatment (variety) effects, the $\mathcal{A}$-value reflects the probability of correctly ranking varieties and thence being able to select the best (Bueno Filho & Gilmour, 2003, 2007).

1.3.2 Maximising information for genetic variance parameter estimation

The work of Smith et al. (2021a) was invaluable in formalising a method for constructing MET datasets and for providing a criterion to compare and thence choose a suitable dataset for a given set of variety selection decisions. A key consequence of this approach is that it typically results in unbalanced data. Within the framework of the original FALMM in which varieties were assumed unrelated, the percentage fill-in was thought to be 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., 2001b, 2015; Ward et al., 2019). Thus, a preliminary step prior to analysis was the examination of "variety connectivity", that is, the number of varieties in common between pairs of environments. Many of the papers listed in Table 1.1 therefore include this information, which is typically presented graphically as a heatmap. As an example, a heatmap

presenting the number of common varieties between environments for the dataset described in Smith et al. (2015) is presented in Figure 1.1. This dataset comes from the NVT program and consists of wheat trials from the Southern region conducted between 2009-2013. The full dataset comprises 200 varieties with 9462 variety by environment combinations observed out of the possible 38416 (%fill-in of 24%). In general in this dataset, all environments for a given year had a similar set of varieties, and much fewer numbers of varieties in common between years.

In more recent applications of the FALMM in which information on genetic relatedness of varieties has been included, the heatmap has continued to be used but may be formed for "parental connectivity", that is, the number of parents in common between environments, since this was seen as important for the reliability of the additive genetic variance parameter estimates (Cullis et al., 2014; Smith & Cullis, 2018). Often, where "poor" connectivity was identified, either in terms of varieties or parents, this information was used to remove individual environments or even years from a MET dataset prior to analysis, as it was thought they would adversely affect genetic variance parameter estimation. Note that this was still deemed necessary even after a MET dataset was identified using the Smith et al. (2021a) approach as having a good (low) $\mathcal{A}$-value because this criterion is calculated with reference to an underlying LMM in which the variance parameters are assumed known.

There has been little in the literature to establish whether these connectivity methods are the most appropriate for forecasting the reliability of variance parameter estimation that will result from the analysis of a given MET dataset. The lack of research on this topic is addressed in this thesis and has resulted in the publication of Lisle et al. (2021). In a similar manner to Smith et al. (2001b), a criterion from the model-based experimental design literature is proposed as a diagnostic tool that can be applied to a MET dataset prior to analysis. In the case of Lisle et al. (2021), the criterion is that of $\mathcal{D}$-optimality, since this enables an assessment of the generalised variance of, or equivalently the information for, the genetic variance parameter estimates. This thesis therefore offers a diagnostic approach for constructing optimal MET datasets that aims to balance the amount of variety information with the information available for genetic variance parameter estimation.

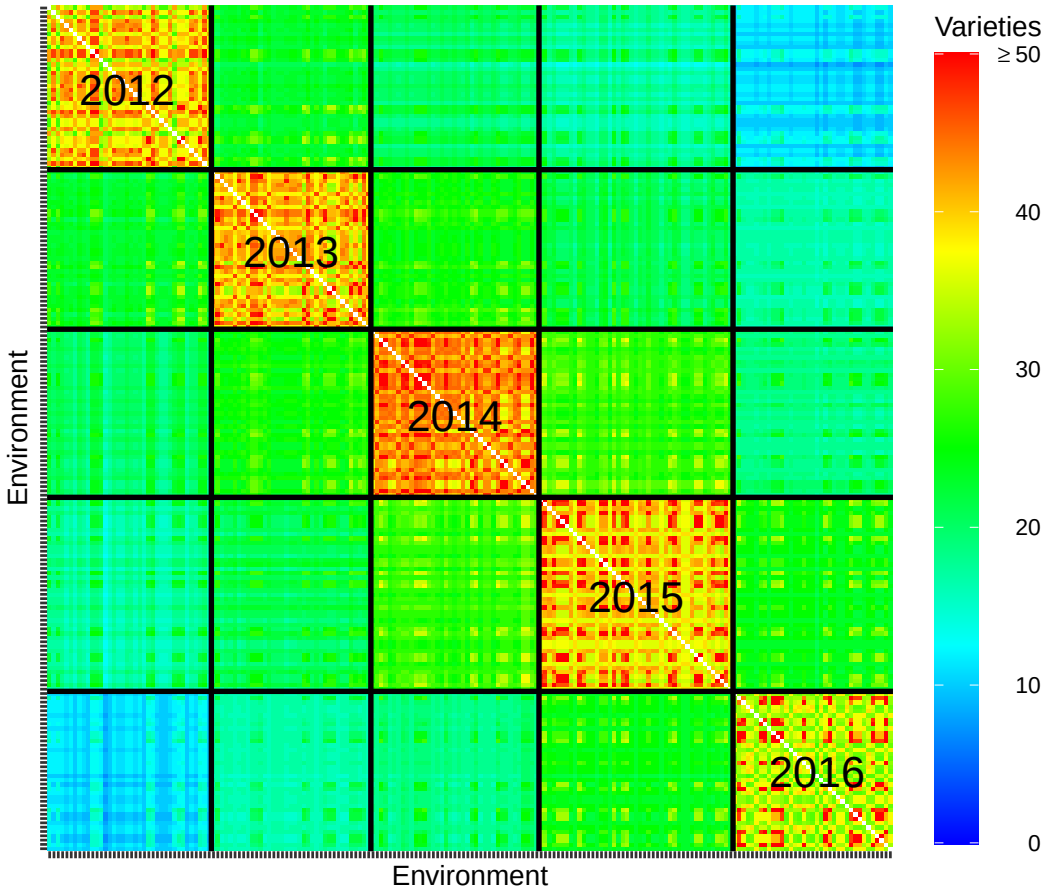

Figure 1.1: Variety connectivity across environments for the Wheat dataset described in Smith et al. (2015). The colours on the off-diagonals indicate the number of varieties in common between pairs of environments. Boundaries for years are indicated by the black lines (2012 - 2016 inclusive from left to right and top to bottom). Right and bottom ticks represent the 196 environments.

1.4 Structure of thesis

The objectives of this thesis are to present novel approaches for optimising the construction of MET datasets from a series of plant variety trials. Although sophisticated and relevant statistical analyses have been proven to increase the reliability of predicted VE effects, there has been little research into how to construct an appropriate dataset. This thesis fills a void in the literature by providing information-based diagnostics for the optimal construction of the MET dataset. The remaining chapters in this thesis are arranged as follows.

Chapter 2 This chapter provides details about the estimation of linear mixed models, including the residual maximum likelihood (REML) estimation of variance parameters and best linear unbiased prediction (BLUP) of random effects. It is limited to theory of direct relevance to the thesis so it covers concepts of particular importance to the analysis of data from plant breeding trials.

Chapter 3 This chapter discusses general aspects about simulation studies and provides fundamental concepts that are used for the studies presented in later chapters. This includes the definition of performance measures of bias and mean squared error (MSE) for REML estimates of variance parameters, and also the reliability of both BLUPs and empirical best linear unbiased predictions (EBLUPs).

Chapter 4 This chapter is devoted to the description of two motivating datasets that are used throughout the thesis. They relate to two distinct types of data which will allow two types of analysis to be investigated. The first, the so-called Oat dataset, is a late-stage variety evaluation dataset in which only elite varieties are considered and which will be analysed as if the varieties are unrelated, that is, with the assumption of independence between the effects for different varieties. The second, the Durum wheat dataset, is a full plant breeding dataset in which varieties from all stages of testing are considered and for which pedigree information is available. The analysis of these data will include the pedigree information via a Numerator Relationship Matrix so that genetic relatedness between the varieties will be accommodated.

Chapter 5 The process of fitting factor analytic linear mixed models (FALMMs), including the spatial modelling of individual trials, is important to the themes of this thesis. This process is demonstrated in this chapter for models in which independence is assumed between the effects for different varieties. The Oat dataset is used for this purpose.

Chapter 6 In this chapter, a simulation study with a range of treatments is used to investigate the effect of variety connectivity on the reliability of genetic variance parameter estimates in a MET analysis and also on the reliability of variety predictions. This is conducted within the framework of models in which independence is assumed between varieties so the analysis of the Oat dataset from the previous chapter provides the trial structural elements and variance component values for the simulation study.

Chapter 7 This chapter demonstrates how MET datasets may be created using the methodology of Smith et al. (2021a). The chapter includes some general results about $\mathcal{A}$-optimality and a reproduction of key sections of Smith et al. (2001b). Note that although the candidate is not the first author of this paper, the candidate had a key role in dataset curation, application of the methodology to the examples, and manuscript preparation and revision.

Chapter 8 This chapter is analogous to Chapter 5 but demonstrates the model fitting process when pedigree information is included via the Numerator Relationship Matrix. The Durum dataset is used for this purpose.

Chapter 9 This chapter develops the $\mathcal{D}$-optimality diagnostic to assess the information in a MET dataset for the REML estimation of genetic variance parameters. The chapter includes some general results about $\mathcal{D}$-optimality and a reproduction of Lisle et al. (2021).

Chapter 10 Concluding remarks are provided in this chapter.

Chapter 2

Linear mixed models: key results for the analysis of plant breeding trials

Linear Mixed Models (LMMs) are widely used for the analysis of plant breeding data. Two key reasons for this are their capacity to deal with incomplete data and their allowance for correlated effects with complex variance structures. Such variance structures include those required to model the variety effects in a multi-environment trial analysis and so-called spatial models that accommodate correlation in the errors in field trials. These models will be fully described and applied to examples in later chapters. The aim of the current chapter is to present LMM theory that is directly relevant to this thesis and allows development of the material presented in later chapters.

This chapter is arranged as follows: I first describe the general form of the LMM in Section 2.1; I then provide the derivation, and distributional properties of the residual maximum likelihood (REML) estimates of variance parameters in Section 2.2; then in Section 2.3 I derive best linear unbiased predictions (BLUPs) of random effects, and present a measure of reliability and inference for random effects; in Section 2.4 I introduce empirical BLUPs (EBLUPs) given that prediction of the random effects are typically calculated using REML estimates of variance parameters; and finally in Section 2.5 I have concluding remarks.

2.1 The linear mixed model

I begin with a general LMM for the $(n \times 1)$ data vector $\boldsymbol{y} = (y_1, ..., y_n)^\top$ which is written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{2.1}$$

where $\boldsymbol{\tau}$ is the $(t \times 1)$ vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}$ is the $(b \times 1)$ vector of random effects with associated design matrix $\boldsymbol{Z}$; and $\boldsymbol{e}$ is the $(n \times 1)$ vector of errors. It is assumed that

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where the variance matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are functions of variance parameters. The vector of random effects $\boldsymbol{u}$ comprises q sub vectors $\boldsymbol{u}_i$ so that $\boldsymbol{u} = (\boldsymbol{u}_1^\top \dots \boldsymbol{u}_q^\top)^\top$. It is assumed that $\boldsymbol{u}_i$ comprises b_i effects so that there are $b = \Sigma_{i=1}^q b_i$ effects for $\boldsymbol{u}$. The sub-vectors are assumed to be independent so that $\text{var}(\boldsymbol{u}) = \boldsymbol{G}$ is a block diagonal matrix given by $\oplus_{i=1}^q \boldsymbol{G}_i$. In many cases, a sub-vector $\boldsymbol{u}_i$ will be assumed to be independent and identically distributed (IID) so it will have a simple variance component structure given by $\boldsymbol{G}_i = \sigma_i^2 \boldsymbol{I}_{b_i}$.

As discussed in Chapter 1, data from plant breeding trials are analysed in order to rank the varieties so that the best can be selected for continued testing or for commercial release. This aim is best achieved with the assumption of random effects so that variety effects comprise one of the sub-vectors in $\boldsymbol{u}$ in the LMM. In the case of the analysis of a single trial, one approach is to assume IID variety effects. However, varieties are created as part of a crossing program so there are familial relationships between them. These can be accommodated in the LMM by including a known relationship matrix as part of their variance structure which means that the variety effects are now correlated. In the case of the analysis of a series of trials, that is, a MET, the variance structure relates to the variety effects in individual environments so the associated variance structure will be required to accommodate correlations between environments. The forms of variance structures for variety effects, both for a single trial analysis and a MET analysis will be discussed in detail in later chapters.

The variance matrix $\boldsymbol{\Sigma}$ for the errors may take many forms. In the simplest case, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, where IID errors are assumed. In the spatial analysis of a field trial the errors may have a correlated structure, so that $\boldsymbol{\Sigma}$ is non-identity. This form will be described in later chapters.

Under these assumption the distribution of $\boldsymbol{y}$ is Gaussian with mean $\boldsymbol{X\tau}$ and variance

$$\text{var}(\boldsymbol{y}) = \boldsymbol{H} = \boldsymbol{ZGZ}^\top + \boldsymbol{\Sigma}$$

2.2 Residual Maximum Likelihood (REML) estimation of variance parameters

Let $\boldsymbol{\kappa}$ denote the vector of unknown variance parameters associated with $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$. In this thesis, the estimation of these parameters is achieved using the residual maximum likelihood (REML) method of Patterson & Thompson (1971). This is based on a residual (rather than full) likelihood function for the data vector. Verbyla (1990) presents a derivation of the residual likelihood function that involves a transformation of the data vector, $\boldsymbol{y}$, to obtain a reduced vector, $\boldsymbol{y_2}$ that represents a set of $n - t$ linear functions that have zero mean. Specifically, he considers the transformation

$$\boldsymbol{L}^\top \boldsymbol{y} = \begin{bmatrix} \boldsymbol{L_1}^\top \boldsymbol{y} \\ \boldsymbol{L_2}^\top \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} \tag{2.2}$$

where $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ are $n \times t$ and $n \times (n-t)$ matrices respectively, both of full column rank, and satisfying $\boldsymbol{L_1}^\top \boldsymbol{X} = \boldsymbol{I}_t$ and $\boldsymbol{L_2}^\top \boldsymbol{X} = \boldsymbol{0}$. Verbyla (1990) shows that the residual log-likelihood for estimating the variance parameters in Equation (2.1) is the marginal log-likelihood based on $\boldsymbol{y_2}$ which is given by

$$\ell_R = -\frac{1}{2}\Big\{ \log|\boldsymbol{H}| + \log|\boldsymbol{X}^\top \boldsymbol{H}^{-1} \boldsymbol{X}| + \boldsymbol{y}^\top \boldsymbol{P} \boldsymbol{y} \Big\}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{H}^{-1}\boldsymbol{X})^- \boldsymbol{X}^\top \boldsymbol{H}^{-1}$ with $(\boldsymbol{X}^\top \boldsymbol{H}^{-1}\boldsymbol{X})^-$ being any generalised inverse of $(\boldsymbol{X}^\top \boldsymbol{H}^{-1}\boldsymbol{X})$.

The REML estimation of $\boldsymbol{\kappa}$ requires the calculation of the REML scores. The form

for the score for κ_i is given by

$$\begin{aligned} U_R(\kappa_i) &= \frac{\partial \ell_R}{\partial \kappa_i} \\ &= -\frac{1}{2}\left\{ tr(\boldsymbol{P}\dot{\boldsymbol{H}}_i) - \boldsymbol{y}^\top \boldsymbol{P}\boldsymbol{H}_i\boldsymbol{P}\boldsymbol{y} \right\} \end{aligned}$$

where the "dot" notation indicates the derivative, so that $\dot{\boldsymbol{H}}_i = \partial \boldsymbol{H}/\partial \kappa_i, i = 1 \ldots n_\kappa$, where n_κ is the number of elements in $\boldsymbol{\kappa}$. REML estimates of κ_i are obtained by setting $\boldsymbol{U}_R(\boldsymbol{\kappa}) = \boldsymbol{0}$, which typically requires a numerical solution. Gradient methods are useful and involve the first term in a Taylor's series expansion. If I expand the score equation about the value of $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(m)}$ I find,

$$\boldsymbol{U}_R(\boldsymbol{\kappa}) = \boldsymbol{U}_R\left(\boldsymbol{\kappa}^{(m)}\right) + \left.\frac{\partial \boldsymbol{U}_R(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}^\top}\right|_{\kappa=\kappa^{(m)}} \left(\boldsymbol{\kappa} - \boldsymbol{\kappa}^{(m)}\right)$$

Equating the right hand side to zero and re-arranging gives

$$\begin{aligned} \boldsymbol{\kappa} &= \boldsymbol{\kappa}^{(m)} - \left[\frac{\partial \boldsymbol{U}_R(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}^\top}\right]^{-1}_{\kappa=\kappa^{(m)}} \boldsymbol{U}_R\left(\boldsymbol{\kappa}^{(m)}\right) \\ &= \boldsymbol{\kappa}^{(m)} + \left[\boldsymbol{\mathcal{I}_O}^{(m)}\right]^{-1} \boldsymbol{U}_R\left(\boldsymbol{\kappa}^{(m)}\right) \end{aligned} \tag{2.3}$$

where $\boldsymbol{\mathcal{I}_O}^{(m)}$ is the observed information matrix for $\boldsymbol{\kappa}$ at $\boldsymbol{\kappa}^{(m)}$. This equation provides the updated value which is denoted $\boldsymbol{\kappa}^{(m+1)}$. This is known as the Newton-Raphson algorithm. Elements of the observed information matrix are,

$$\begin{aligned} \boldsymbol{\mathcal{I}_O}(\kappa_i, \kappa_j) &= -\frac{\partial U_R(\kappa_i)}{\partial \kappa_j} \\ &= \frac{1}{2}\left(\frac{\partial tr\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right)}{\partial \kappa_j} - \frac{\partial \boldsymbol{y}^\top \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{\partial \kappa_j} \right) \\ &= \frac{1}{2} tr\left(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\right) - \frac{1}{2} tr\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) + \boldsymbol{y}^\top \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y} \end{aligned} \tag{2.4}$$

where $\dot{\boldsymbol{H}}_{ij} = \partial^2 \boldsymbol{H}/\partial \kappa_i \partial \kappa_j$.

As an example, given an initial estimate $\boldsymbol{\kappa}^{(0)}$, an update is,

$$\boldsymbol{\kappa}^{(1)} = \boldsymbol{\kappa}^{(0)} + \mathcal{I}_{\boldsymbol{O}}(\kappa^{(0)}, \kappa^{(0)})^{-1} \boldsymbol{U}_R \left(\boldsymbol{\kappa}^{(0)}\right)$$

An alternative to using $\mathcal{I}_{\boldsymbol{O}}$ is the use of the expected information matrix $\mathcal{I}_{\boldsymbol{E}}$, which has elements given by

$$\mathcal{I}_{\boldsymbol{E}}(\kappa_i, \kappa_j) = \mathrm{E}\left[-\frac{\partial U_R(\kappa_i)}{\partial \kappa_j}\right] \tag{2.5}$$

$$= \frac{1}{2} tr\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i \boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \tag{2.6}$$

When $\mathcal{I}_{\boldsymbol{O}}$ is replaced by $\mathcal{I}_{\boldsymbol{E}}$ in Equation 2.3 this is known as the Fisher-scoring algorithm.

In practice, the trace terms in Equations 2.4 and 2.6 involving matrices of order n are often not feasible to calculate. To combat this Gilmour et al. (1995) derive the so-called average information matrix $\mathcal{I}_{\boldsymbol{A}}$, which is formed by averaging $\mathcal{I}_{\boldsymbol{O}}$ and $\mathcal{I}_{\boldsymbol{E}}$ and approximating $\boldsymbol{y}^\top \boldsymbol{P}\dot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y}$ by its expectation $tr(\boldsymbol{P}\dot{\boldsymbol{H}}_{ij})$ in those cases when $\dot{\boldsymbol{H}}_{ij} \neq 0$. For other variance models the average information matrix is an exact average. The elements of the average information matrix are,

$$\mathcal{I}_{\boldsymbol{A}}(\kappa_i, \kappa_j) = \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{P}\dot{\boldsymbol{H}}_i \boldsymbol{P}\dot{\boldsymbol{H}}_j \boldsymbol{P}\boldsymbol{y} \tag{2.7}$$

2.2.1 Distributional properties of REML estimates

It is well known that if variance parameters are estimated using (full) maximum likelihood, then the resultant estimates are asymptotically normal with zero mean and variance matrix given by the inverse of the information matrix (see Mardia & Marshall, 1984, for example). Cressie & Lahiri (1993) discuss the distributional properties of variance parameter estimates obtained via residual maximum likelihood. Let $\hat{\boldsymbol{\kappa}}$ denote the REML estimate of $\boldsymbol{\kappa}$. Also let $\mathcal{I}_{\boldsymbol{E}}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top)$ denote the full $(n_k \times n_k)$ expected information matrix with elements as given in Equation 2.6. Then Cressie & Lahiri (1993) show that asymptotically,

$$\hat{\boldsymbol{\kappa}} \sim \mathcal{N}\left(\boldsymbol{0}, \mathcal{I}_{\boldsymbol{E}}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top)^{-1}\right)$$

This result about the asymptotic variance of $\hat{\boldsymbol{\kappa}}$ is used as the basis for a novel diagnostic as provided in Chapter 9.

2.2.2 Residual maximum likelihood ratio test

The residual maximum likelihood ratio test (REMLRT) is a generic approach for comparing the fit of nested models. To compare models 1 and 2 which have residual log-likelihoods of ℓ_1 and ℓ_2, respectively, and where model 2 contains an extra d variance parameters, the REMLRT statistic is given as

$$D = 2(\log(\ell_2) - \log(\ell_1)) \tag{2.8}$$

This has a distribution which is approximately chi-square on d df. If the model is obtained by constraining variance parameters to be non-negative, the probability is computed using a mixture of chi-square distributions as described in Self & Liang (1987); Stram & Lee (1994).

2.3 Best Linear Unbiased Predictions (BLUPs) of random effects

In this thesis the main emphasis is the prediction of random effects, in particular the variety effects, and the reliability of those predictions. In terms of prediction of the random effects, it is well known (see Robinson, 1991; Searle, 1997, for example) that the best linear unbiased prediction (BLUP) of $\boldsymbol{u}$ is given by

$$\begin{aligned} \tilde{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{y} \\ &= \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}) \end{aligned} \tag{2.9}$$

where $\hat{\boldsymbol{\tau}} = \left(\boldsymbol{X}^\top \boldsymbol{H}^{-1}\boldsymbol{X}\right)^- \boldsymbol{X}^\top \boldsymbol{H}^{-1}\boldsymbol{y}$ so that $\boldsymbol{X}\hat{\boldsymbol{\tau}}$ is the best linear unbiased estimate (BLUE) of $\boldsymbol{X}\boldsymbol{\tau}$. The prediction error variance (PEV) for the random effects is given by

$$\boldsymbol{\Omega} = \mathrm{var}(\boldsymbol{u} - \tilde{\boldsymbol{u}}) = \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \tag{2.10}$$

Searle (1997) presents an approach for the prediction of $\boldsymbol{u}$ based on the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{y}$. Using standard results concerning multivariate Normal dis-

tributions he shows that

$$\mathrm{E}(\boldsymbol{u}|\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau}) \tag{2.11}$$

$$\mathrm{var}(\boldsymbol{u}|\boldsymbol{y}) = \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \tag{2.12}$$

Searle (1997) then notes that Equation (2.11) has the same form as the BLUP in Equation (2.9) but with $\boldsymbol{\tau}$ rather than $\hat{\boldsymbol{\tau}}$. Similarly Equation (2.12) has the same form as the PEV in Equation (2.10) but with $\boldsymbol{H}^{-1}$ rather than $\boldsymbol{P}$. In what follows, I show that the exact forms for the BLUP and PEV of the random effects can be found using a conditional approach but replacing the data vector $\boldsymbol{y}$ with the reduced vector, $\boldsymbol{y_2}$ from Equation 2.2. It is noted that Diffey et al. (2017) also condition on $\boldsymbol{y_2}$ rather than $\boldsymbol{y}$ in their development of a REML expectation-maximisation (EM) algorithm for variance parameter estimation. The joint distribution of $\boldsymbol{y_2}$ and $\boldsymbol{u}$ is given by

$$\begin{bmatrix} \boldsymbol{y_2} \\ \boldsymbol{u} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{L_2}^\top \boldsymbol{H} \boldsymbol{L_2} & \boldsymbol{L_2}^\top \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{L_2} & \boldsymbol{G} \end{bmatrix} \right)$$

I then consider the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{y_2}$ which, using results in Verbyla (1990), leads to

$$\begin{aligned} \mathrm{E}(\boldsymbol{u}|\boldsymbol{y_2}) &= \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{L_2}\left(\boldsymbol{L_2}^\top \boldsymbol{H}\boldsymbol{L_2}\right)^{-1} \boldsymbol{y_2} \\ &= \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{y} \\ &= \tilde{\boldsymbol{u}} \\ \mathrm{var}(\boldsymbol{u}|\boldsymbol{y_2}) &= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{L_2}\left(\boldsymbol{L_2}^\top \boldsymbol{H}\boldsymbol{L_2}\right)^{-1} \boldsymbol{L_2}^\top \boldsymbol{Z}\boldsymbol{G} \\ &= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= \boldsymbol{\Omega} \end{aligned}$$

as required. Finally, I therefore have that

$$\boldsymbol{u}|\boldsymbol{y_2} \sim \mathcal{N}(\tilde{\boldsymbol{u}}, \boldsymbol{\Omega}) \tag{2.13}$$

2.3.1 Reliability of BLUPs

A measure of reliability for an individual BLUP can be obtained by considering the joint distribution of $\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}$ which is normal with zero mean and variance matrix

$$\begin{aligned} \text{var}\begin{pmatrix}\boldsymbol{u}\\ \tilde{\boldsymbol{u}}\end{pmatrix} &= \begin{bmatrix}\boldsymbol{G} & \boldsymbol{G}\boldsymbol{Z}^\top\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{Z}^\top\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} & \boldsymbol{G}\boldsymbol{Z}^\top\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{G} & \boldsymbol{G}-\boldsymbol{\Omega}\\ \boldsymbol{G}-\boldsymbol{\Omega} & \boldsymbol{G}-\boldsymbol{\Omega}\end{bmatrix} \end{aligned} \tag{2.14}$$

Mrode & Thompson (2005) define the accuracy of a prediction as the correlation between the prediction and the true value of the random effect. Equation 2.14 shows that the accuracy for the BLUP of the i^{th} random effect is given by

$$\text{cor}(u_i, \tilde{u}_i) = \sqrt{1 - \frac{\boldsymbol{d}^\top\boldsymbol{\Omega}\boldsymbol{d}}{\boldsymbol{d}^\top\boldsymbol{G}\boldsymbol{d}}} \tag{2.15}$$

where $\boldsymbol{d}$ is a vector of length b containing all zeros apart from a "1" in position i. Note that this could also be written as

$$\text{cor}(u_i, \tilde{u}_i) = \sqrt{1 - \frac{\boldsymbol{\Omega}_{ii}}{\boldsymbol{G}_{ii}}}$$

where $\boldsymbol{\Omega}_{ii}$ is the i^{th} diagonal element of $\boldsymbol{\Omega}$ so represents the prediction error variance of $\tilde{u}_i$ and $\boldsymbol{G}_{ii}$ is the i^{th} diagonal element of $\boldsymbol{G}$ so represents the variance of u_i. Mrode & Thompson (2005) also comment that the accuracy of prediction is often presented in terms of reliability, which is the square of the accuracy value.

2.3.2 Inference on random effects

In the case of fixed effects, inference is typically formulated using hypothesis tests in which the null hypothesis relates to "no difference" (that is, equality) between effects or simply "no effect" (equality with zero). The concept and approach for such tests of hypothesis is well understood. In the case of random effects, any tests involving equality are inappropriate because the effects relate to a random variable. It is possible, however, to make meaningful probability statements about $\boldsymbol{u}$ by using the conditional distribution of $\boldsymbol{u}|\boldsymbol{y_2}$ as derived in the previous section. In the plant breeding trial application it is of interest to determine the probability that, given the data, the true (genetic) effect

for variety i is higher than that for variety j. If I let u_i and u_j be the effects for variety i and j respectively, then formally I consider

$$\begin{aligned}
\Pr(u_i > u_j|\boldsymbol{y_2}) &= \Pr(u_i - u_j > 0|\boldsymbol{y_2}) \\
&= \Pr((u_i - \tilde{u}_i) - (u_j - \tilde{u}_j) > -(\tilde{u}_i - \tilde{u}_j)|\boldsymbol{y_2}) \\
&= \Pr\left(\frac{(u_i - \tilde{u}_i) - (u_j - \tilde{u}_j)}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}} > -\frac{(\tilde{u}_i - \tilde{u}_j)}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}}|\boldsymbol{y_2}\right) \\
&= 1 - \Phi\left(-\frac{(\tilde{u}_i - \tilde{u}_j)}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}}\right)
\end{aligned}$$

where $\boldsymbol{d}$ is a vector of length b containing all zeros apart from a "1" in position i and "-1" in position j and Φ is the standard normal cumulative distribution function. Note that $\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}$ is therefore var $((u_i - \tilde{u}_i) - (u_j - \tilde{u}_j))$, so represents the prediction error variance (PEV) of the difference $\tilde{u}_i - \tilde{u}_j$. The general framework about probability statements for random effects is used as the basis for the diagnostic presented in Chapter 7.

2.4 Empirical Best Linear Unbiased Predictions (EBLUPs) of random effects

In the previous section I presented results about BLUPs of random effects. These assume that the variance parameters in $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are known. In practice, this is typically not the case and the parameters must be estimated from the data. When variance parameters must be estimated, it is common practice to obtain predictions of random effects using Equation 2.9 but replacing the parameters in $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ with their REML estimates. If the resultant matrices are denoted by $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{\Sigma}}$ (and thence $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{P}}$) the expression

$$\tilde{\boldsymbol{u}}^* = \hat{\boldsymbol{G}} \boldsymbol{Z}^\top \hat{\boldsymbol{P}} \boldsymbol{y} \tag{2.16}$$

can be calculated, but it is not the BLUP. It is instead called the Empirical Best Linear Unbiased Prediction (EBLUP). Also note that the PEV for the EBLUP is often calculated using the form in Equation 2.10, that is

$$\boldsymbol{\Omega} = \hat{\boldsymbol{G}} - \hat{\boldsymbol{G}} \boldsymbol{Z}^\top \hat{\boldsymbol{P}} \boldsymbol{Z} \hat{\boldsymbol{G}} \tag{2.17}$$

However, this is not correct because no account has been taken of the variability associated with the estimation of the variance parameters (Searle, 1997; Sales & Hill, 1976a). It is, however, the standard measure provided in LMM software.

The properties of EBLUPs and the impact on their reliability as a result of the REML estimation of the variance parameters is a focus of this thesis. To avoid clumsy notation, predicted random effects will always be denoted as $\tilde{\boldsymbol{u}}$ and where necessary, the distinction between BLUPs and EBLUPs will be made in words.

2.5 Concluding remarks

The objectives of this thesis are to optimise the construction of MET datasets by using a LMM statistical approach to critically assess the structure of the dataset and thus improve the reliability of the predicted VE effects. This chapter has presented key LMM theory for the analysis of plant breeding trials, which are are used and demonstrated in Chapters 5 and 8 for the analysis of two motivating datasets, as well as provide the theoretical framework for novel diagnostics presented in Chapters 7 and 9.

Chapter 3

Simulation study methodology

This thesis includes simulation studies with objectives of addressing several broad concerns with the construction of MET datasets (see Chapters 6 and 9). Concepts for the preparation and presentation of the findings from these studies are presented in this chapter. The simulation studies were created within the `R` statistical computing environment (R Core Team, 2020), and all statistical analyses were completed with the `ASReml-R` (Butler et al., 2017) package.

Simulation studies are used to generate empirical data about how statistical approaches perform in different settings. These are in contrast to algebraic solutions, which are not always attainable or maybe difficult to obtain. The ability to understand the behaviour of statistical approaches is a fundamental strength of simulation studies since the parameters of interest are known from the data generation process. This allows the examination of characteristics like bias and mean square error (MSE), and importantly to the aims of this thesis, the reliability of predicted genetic effects (see Section 2.3.1, which is defined as the (squared) correlation between the prediction and the true value of the random effect). Because simulation studies create data from known distributions, they must be anchored in real-world circumstances so that the simulated results appropriately represent the aims and objectives of the study. The plant breeding trials detailed in Chapter 4 from Oat and Durum breeding programs inspired the investigations presented in this thesis.

This chapter is arranged as follows: In Section 3.1 I demonstrate how to simulate a

completely randomised designed (CRD) field trial, which is used as the motivation throughout the remainder of this chapter. In Section 3.2 I describe the methodology to form the performance measures of bias and mean squared error (MSE) for the residual maximum likelihood (REML) estimates of the variance parameters, and also the reliability of Empirical Best Linear Unbiased Predictions (EBLUPs) of the random effects. In Section 3.3 I look more closely at the reliability of BLUPs and EBLUPs to develop several novel concepts of importance to the simulation studies in Chapters 6 and 9. I describe additional simulation study methodology which is important for running a successful simulation study in Section 3.4. Concluding remarks are presented in Section 3.5.

3.1 Simulation of a trial with a completely randomised design

In this section, I demonstrate how to simulate grain yield data from a CRD field trial. Here I consider a trial with $m = 24$ varieties with $b = 3$ replicates each, which are completely randomised to $n = 72$ plots. The statistical model for the (72×1) data vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_{72})^\top$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{e} \tag{3.1}$$

where $\boldsymbol{\tau}$ is the overall mean with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the (24×1) vector of variety effects with associated design matrix $\boldsymbol{Z_g}$; and $\boldsymbol{e}$ is the (72×1) vector of errors. It is assumed

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_g} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where simple variance structures are assumed for $\{\boldsymbol{G_g}, \boldsymbol{\Sigma}\}$ so that, $\boldsymbol{G_g} = \sigma_g^2 \boldsymbol{I}_{24}$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{72}$. Under these assumptions, the distribution of $\boldsymbol{y}$ is Gaussian with mean $\boldsymbol{X}\boldsymbol{\tau}$ and variance

$$\text{var}(\boldsymbol{y}) = \boldsymbol{H} = \sigma_g^2 \boldsymbol{Z_g} \boldsymbol{Z_g}^\top + \sigma^2 \boldsymbol{I}_{72}$$

For demonstration purposes, it is assumed that $\sigma_g^2 = 0.2$ and $\sigma^2 = 1$. In terms of the fixed effects, without loss of generality I choose $\tau = 0$.

3.1.1 Simulation and analysis of the data vector

Here I provide the techniques and R code to simulate and analyse the data vector $\boldsymbol{y}$ using the model in Equation 3.1. I simulate $S = 200$ times, but note that this is inspected in Sections 3.4.3 to investigate how many simulations are required to make sure the results are precise. The R code for generating $\boldsymbol{y}$ and the corresponding analysis is shown below

```
#variance parameter values.
sg2 <- 0.2; s2 <- 1
#number of varieties and replication
m <- 24; b <- 3
#number of simulations
S <- 200
#Variety names V1-V24
Varname <- rep(paste("V",1:m,sep=''),b)

#Set up base datafile. Rectangular grid with no.columns = b.
data <- data.frame(Column=rep(1:b,each=m), Row=rep(1:m,b),
Variety=sample(Varname))
data$Variety <- factor(data$Variety,
levels = mixedsort(as.character(unique(data$Variety))))

#Setup matrices to store results. True and predicted ug's
Tug.test <- matrix(data=NA,ncol=S,nrow=m) #True
Pug.test <- matrix(data=NA,ncol=S,nrow=m) #Pred (EBLUPs)
#REML estimate for variance parameters.
sg2.pred <- matrix(data=NA,ncol=1,nrow=S) #REML Genetic var
s2.pred <- matrix(data=NA,ncol=1,nrow=S) #REML Error var

for(i in 1:S){ #run S times
        u_g <- rnorm(m,0,sqrt(sg2))
        e <- rnorm(m*b,0,s2)
        data$y <- u_g[data$Variety] + e
        RD.as <- asreml(y ~ 1, random=~Variety, data = data)

        Tug.test[,i] <- u_g
        Pug.test[,i] <- RD.as$coefficients$random[1:m]
        sg2.pred[i] <- summary(RD.as)$varcomp[1,1]
        s2.pred[i] <- summary(RD.as)$varcomp[2,1] }
```

where `u_g` is the (24×1) vector of (true) variety effects, which are stored in the `Tug.test`

object; `e` is the (72×1) vector of errors; `data` is the datafile with 72 rows of data; `Variety` is a factor with 24 levels, each occurring randomly three times; and finally `y` is the (72×1) data vector. Note, `R` library packages `asreml` and `gtools` are required for the code above.

For the `ASReml-R` code, `1` is the term reflecting the overall mean; `Variety` is the factor representing the 24 levels of varieties which are fitted as random effects; `sg2.pred` and `s2.pred` are objects containing the REML estimates of the genetic and error variance parameters respectively; and `Pug.test` is an object containing the EBLUPs of the variety effects. I note that an error term is not specified in the model call, which defaults to a simple error term representing independent and identically distributed (IID) effects.

3.2 Simulation study performance measures

3.2.1 Bias and mean square error

In this section I demonstrate the methodology to calculate bias and MSE performance measures of the REML estimates of the variance parameters. I use the study described in Section 3.1 and I recall that the values of the variance parameters used for data generation were $\sigma_g^2 = 0.2$ and $\sigma^2 = 1$, and ran $S = 200$ simulations.

For each of the variance parameters $\{\sigma_g^2, \sigma^2\}$, let T_s denote the REML estimate from the sth simulated dataset $s = \{1, 2, \ldots, S\}$, and let μ denote the true value that was used for generating data for either parameter. I define

$$\text{Mean} = \bar{T} = \frac{1}{S}\sum_{s=1}^{S} T_s \tag{3.2}$$

$$\text{Bias} = \bar{T} - \mu \tag{3.3}$$

$$\text{MSE} = \frac{1}{S}\sum_{s=1}^{S}(T_s - \mu)^2 \tag{3.4}$$

Table 3.1 presents the mean, bias, and MSE for $\{\sigma_g^2, \sigma^2\}$ given the estimates. I see that there is a small amount of negative bias for σ_g^2, and small positive bias for σ^2. The MSE for σ_g^2 is shown to be half of that shown for σ^2.

Table 3.1: Bias and MSE of the REML estimates of the variance parameters $\{\sigma_g^2, \sigma^2\}$ from the simulation study of Section 3.1 with $S = 200$ simulations.

Parameter	Mean	Bias	MSE
σ_g^2	0.195	-0.005	0.021
σ^2	1.006	0.006	0.043

3.2.2 Reliability of EBLUPs

This section describes the methods for calculating the reliabilities of EBLUP from a simulation study. I define $\{\boldsymbol{u}_{\boldsymbol{g}ks}, \tilde{\boldsymbol{u}}_{\boldsymbol{g}ks}\}$ to be the true (simulated) effects and EBLUPs respectively, for variety $k = \{1 \dots m\}$ and simulation $s = \{1 \dots S\}$. The reliability for the EBLUPs is computed as the square of the sample correlation between the true effects and the EBLUPs. Thus, the reliability for variety k is given by

$$\begin{aligned} R_k^S =& \text{cor}\left(\boldsymbol{u}_{\boldsymbol{g}ks}, \tilde{\boldsymbol{u}}_{\boldsymbol{g}ks}\right)^2 \\ =& \frac{\left(\sum_{s=1}^{S}\left(u_{gks} - \bar{u}_{gk}\right)\left(\tilde{u}_{gks} - \bar{\tilde{u}}_{gk}\right)\right)^2}{\sum_{s=1}^{S}\left(u_{gks} - \bar{u}_{gk}\right)^2 \sum_{s=1}^{S}\left(\tilde{u}_{gks} - \bar{\tilde{u}}_{gk}\right)^2} \end{aligned} \tag{3.5}$$

where $\bar{u}_{gk}$ denotes the mean across S simulations of the true effects for variety k, and $\bar{\tilde{u}}_{gk}$ is the mean for the predicted effects. I note that the superscript 'S' in the term R_k^S refers to the fact that the values are derived from the simulated results, with this identification becoming crucial in later sections of this chapter. I also denote a trial reliability as the across varieties average of R_k^S, which is defined as

$$\bar{R}^S = \frac{1}{m}\sum_{k=1}^{m} R_k^S \tag{3.6}$$

The R_k^S for each variety for our CRD simulation study (with $m = 24$) are presented in Table 3.2. The R_k^S values range from 0.150 (V17) to 0.395 (V2). This is also depicted graphically in Figure 3.1 for V1 which shows a positive linear relationship between $\boldsymbol{u}_{\boldsymbol{g}s}$ and $\tilde{\boldsymbol{u}}_{\boldsymbol{g}s}$, which has a $R_k^S = 0.276$. Thus, the correlation between the $y-$ and $x-$values in Figure 3.1 is $\sqrt{0.276} = 0.525$. I also find that $\bar{R}^S = 0.264$.

Table 3.2: Reliability of EBLUPs R_k^S for each variety $k = \{1 \dots m\}$ using Equation 3.5, from the CRD simulation study with $S = 200$ simulations.

Variety	R_k^S	Variety	R_k^S	Variety	R_k^S	Variety	R_k^S
V1	0.276	V7	0.246	V13	0.264	V19	0.284
V2	0.395	V8	0.282	V14	0.231	V20	0.275
V3	0.221	V9	0.317	V15	0.213	V21	0.231
V4	0.187	V10	0.252	V16	0.308	V22	0.299
V5	0.358	V11	0.287	V17	0.150	V23	0.354
V6	0.244	V12	0.218	V18	0.266	V24	0.179

3.3 More on the reliability of BLUPs and EBLUPs

This section examines the methodology and principles underpinning the reliability values for the BLUPs (see Section 2.3.1), that is, those based on known variance parameters. I also present their relationship to the reliabilities for the EBLUPs, that is, those derived with REML estimates of the variance parameters.

The more detailed examination of reliability is needed for two key reasons. First is that in the simulation studies in later chapters I require real-world values for the variance parameters for data generation. Thus, I need to examine analyses of a number of field trials and summarise parameters across trials in some manner to obtain "typical" values for the variance parameters. Because variance parameters are scale-dependent, and because our main focus is on the variety effects, a sensible approach is to instead consider trial reliability of the predicted variety effects. I must first develop a relative measure of reliability so that averages across trials can be sensibly taken. I then need to "calibrate" these relative reliabilities of BLUPs against values of variance parameters. In this way I can obtain values of variance parameters that provide typical levels of trial reliability. The second reason is that a major aim is to examine the impact of the accuracy of variance parameter estimation on the reliability of variety predictions. This involves a comparison of the reliability of BLUPs against the reliability of EBLUPs.

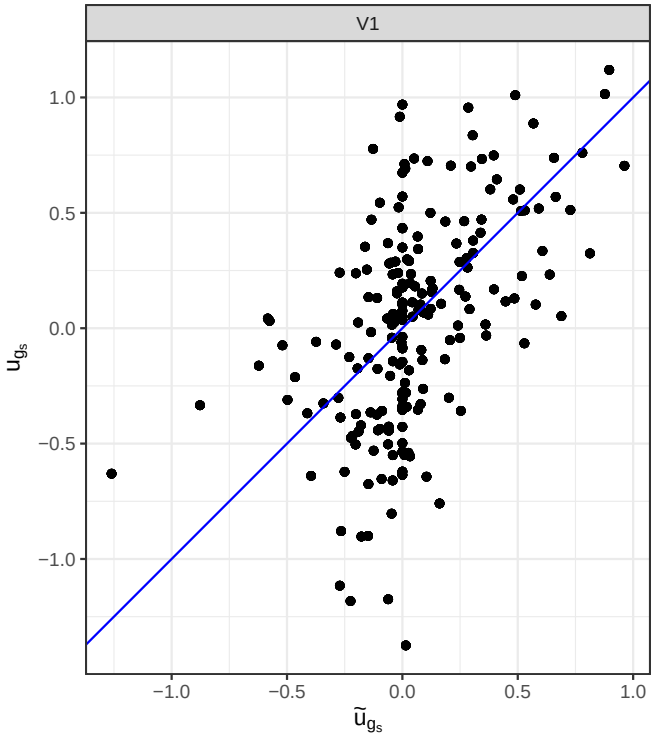

Figure 3.1: True (simulated) effects $\boldsymbol{u}_{\boldsymbol{g}_s}$ against the EBLUPs $\tilde{\boldsymbol{u}}_{\boldsymbol{g}_s}$ for V1 from the CRD simulation study with $S = 200$ simulations. Solid blue line represents a 1:1 line.

3.3.1 Maximum reliability of BLUPs

In this section I use the CRD example, for which an algebraic form for reliability is easily obtained, to show that reliability has an upper limit that depends on structural elements including the number of varieties.

I again consider a CRD variety trial, containing b replicates of m varieties ($n = bm$). I assume, without loss of generality, that the data are ordered as replicates within varieties, so that $\boldsymbol{Z}_{\boldsymbol{g}} = \boldsymbol{I}_m \otimes \boldsymbol{1}_b$. Assuming $\boldsymbol{G} = \sigma_g^2 \boldsymbol{I}_m$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$ then

$$\begin{aligned}\boldsymbol{H} &= \sigma_g^2 \boldsymbol{Z}_{\boldsymbol{g}} \boldsymbol{Z}_{\boldsymbol{g}}^\top + \sigma^2 \boldsymbol{I}_n \\ &= \zeta_1 \boldsymbol{I}_m \otimes \boldsymbol{A}_b + \zeta_2 \boldsymbol{I}_m \otimes \boldsymbol{B}_b\end{aligned}$$

where $\zeta_1 = b\sigma_g^2 + \sigma^2$; $\zeta_2 = \sigma^2$; $\boldsymbol{A}_b = \mathbf{1}_b\mathbf{1}_b^\top/b$, and $\boldsymbol{B}_b = \boldsymbol{I}_b - \boldsymbol{A}_b$. It then follows that

$$\boldsymbol{H}^{-1} = \frac{\boldsymbol{I}_m \otimes \boldsymbol{A}_b}{\zeta_1} + \frac{\boldsymbol{I}_m \otimes \boldsymbol{B}_b}{\zeta_2} \qquad \boldsymbol{P} = \frac{\boldsymbol{B}_m \otimes \boldsymbol{A}_b}{\zeta_1} + \frac{\boldsymbol{I}_m \otimes \boldsymbol{B}_b}{\zeta_2}$$

From Equation 2.10, it is shown that the prediction error variance (PEV) is given by

$$\begin{aligned}\text{PEV}\left(\tilde{\boldsymbol{u}}_{\boldsymbol{g}}\right) &= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}_{\boldsymbol{g}}^\top \boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{g}}\boldsymbol{G} \\ &= \sigma_g^2\boldsymbol{I}_m - \sigma_g^2 h^2(\boldsymbol{I}_m - \boldsymbol{A}_m)\end{aligned}$$

where $\boldsymbol{A}_m = \mathbf{1}_m\mathbf{1}_m^\top/m$, $\boldsymbol{B}_m = \boldsymbol{I}_m - \boldsymbol{A}_m$, and $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma^2/b)$. The quantity h^2 is widely referred to in plant breeding as the line mean heritability. Therefore, for a single variety $k = \{1 \dots m\}$ it is found that

$$\text{PEV}\left(\tilde{u}_{g_k}\right) = \sigma_g^2 - \sigma_g^2 h^2(1 - 1/m) \tag{3.7}$$

Finally, the reliability R_k (Equation 2.15) of the prediction for variety $k = \{1 \dots m\}$ is

$$\begin{aligned}R_k &= 1 - \frac{\text{PEV}\left(\tilde{u}_{g_k}\right)}{\text{var}\left(u_{g_k}\right)} \\ &= 1 - \frac{\left[\sigma_g^2 - \sigma_g^2 h^2(1 - 1/m)\right]}{\sigma_g^2} \\ &= h^2(1 - 1/m)\end{aligned} \tag{3.8}$$

so that in this simple case, the reliability R_k is the same for all varieties. Thus, when averaged across varieties, a trial reliability is

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_k \tag{3.9}$$

Again, for the CRD example, R_k is the same for all varieties and therefore also equal to $\bar{R}$. For other examples this is not the case, such as unequal replication and more complex variance structures. I can show that in Equation 3.8 as m gets large, R_k (and $\bar{R}$) tends to h^2, that is

$$\lim_{m\to\infty} \bar{R} = h^2 \tag{3.10}$$

Also, when $h^2 = 1$ in Equation 3.8, this defines the maximum potential $\bar{R}$ to be proportional to m. The maximum potential trial reliability is therefore defined as

$$\bar{R}_X = 1 - 1/m \tag{3.11}$$

Furthermore, I define the relationship with h^2, as $\bar{R}$ divided by $\bar{R}_X$, which is also defined as the proportion of the maximum potential trial reliability

$$\bar{R}_{X_p} = h^2 = \frac{\bar{R}}{\bar{R}_X} \tag{3.12}$$

Furthermore, rearranging Equation 3.12 for σ_g^2 for a particular $\bar{R}_{X_p}$, I show that

$$\sigma_g^2 = \frac{\sigma^2 \bar{R}_{X_p}}{b(1 - \bar{R}_{X_p})} \tag{3.13}$$

3.3.2 Design-based reliability of BLUPs

As shown in Section 3.3.1 for the example of a CRD field trial with simple variance structures, it is reasonably straightforward to obtain algebraic forms for the reliabilities of BLUPs. However, an algebraic solution may not always be attainable for more complicated scenarios. Examples may include variance structures for errors other than the assumption of IID, covariance structures for random effects, and unequal variety replication.

In this section I compute analogous reliability based values that assume known variance parameters (not REML estimates), which I define as 'design-based' reliability values. Here the prediction error variances (PEVs) of the variety effects from an `ASReml-R` model fit by constraining variance parameters to the known values. I then use Equation 2.15 and denote the resultant reliability for variety k as R_k^D, where the superscript 'D' refers to the fact these are identified as design-based reliability values. Similarly, I then use Equation 3.9 and denote the trial reliability as $\bar{R}^D$.

The code and output to obtain design-based reliability values for the CRD example using the model in Equation 3.1 is shown below

```
#Set variance parameter values.
sg2 <- 0.2; s2 <- 1
#Set number of varieties and replication
m <- 24; b <- 3
#Variety names V1-V24
vars <- rep(paste("V",1:m,sep=""),b)

#Set up base datafile. Rectangular grid with no.columns =b
data <- data.frame(Column=factor(rep(1:b,each=m)),
Row=factor(rep(1:m,b)), Variety=sample(Varname))
data$Variety <- factor(data$Variety,
levels=mixedsort(as.character(unique(data$Variety))))
data$y <- 1

#Create starting value file.
v24.sv <- asreml(y~1, random=~Variety, residual=~units,
data=data, start.values=T)

#Constrain variances parameters to equal specific values.
sv <- v24.sv$vparameters.table
sv$Value <- c(sg2,s2)
sv$Constraint <- 'F'

#Run model with the starting values file.
v24.as <- asreml(y~1, random=~Variety, residual=~units,
data=data, G.param=sv, R.param=sv)

#obtain pev's and calculate reliabilities for 24 varieties.
vcoeff <- v24.as$vcoeff$random
round(1-vcoeff/0.2,3)
# [1] 0.359 0.359 0.359 0.359 0.359 0.359 0.359 0.359
# [9] 0.359 0.359 0.359 0.359 0.359 0.359 0.359 0.359
#[17] 0.359 0.359 0.359 0.359 0.359 0.359 0.359 0.359

#algebraic solution
round(0.2/(0.2+(1/3))*(1-1/24),3)
#[1] 0.359
```

where `start.values` is the variance table object, which is given the variance parameter values of 0.2 and 1 for σ_g^2 and σ^2 respectively, and constrained to these values (`'F'`); and `vcoeff` from `v24.as$vcoeff$random` contains the PEV (see Equation 2.10) for the varieties. I see that the reliabilities from the design-based and algebraic solution both give the same reliability value of 0.359.

3.3.3 Reliability response of BLUPs

The simulation study given in Chapter 6 utilizes the properties of Equation 3.12 to obtain values for σ_g^2 for a given $\bar{R}_{X_p}$, with varying levels of m. To demonstrate the reliability identities as shown in the previous sections, I present graphically in Figure 3.2 the relationships between $\{\sigma_g^2, m, \bar{R}, \bar{R}_{X_p}\}$, given $b = 3$ for four levels of $m = \{12, 24, 48, 96\}$ (Tsize). For illustration I have highlighted $\bar{R}_{X_p} = \{0.85, 0.90, 0.95, 0.99\}$ values on each of the profile curves.

In Figure 3.2(a) I present the relationship of the genetic variance σ_g^2 to $\bar{R}$. As shown, there are different responses for the four Tsizes with all tending towards their corresponding $\bar{R}_X = \{0.92, 0.96, 0.98, 0.99\}$ values respectively, which are represented by the coloured horizontal dashed lines.

In Figure 3.2(b) I present the relationship of the genetic variance σ_g^2 to $\bar{R}_{X_p}$. For this example of a CRD, the four Tsizes have the same profile. I show for the $\bar{R}_{X_p}$ values of $\{0.85, 0.90, 0.95, 0.99\}$, that $\sigma_g^2 = \{1.89, 3.0, 6.33, 33.0\}$. Thus, for example, with $m = 24$ to achieve an $\bar{R}_{X_p} = 0.85$ with $b = 3$, a value of $\sigma_g^2 = 1.89$ is required. This strategy is followed in the simulation study detailed in Chapter 6. In other words this calibration procedure is used to choose genetic variances that reflected a range of reliabilities observed given a dataset with unequal numbers of varieties.

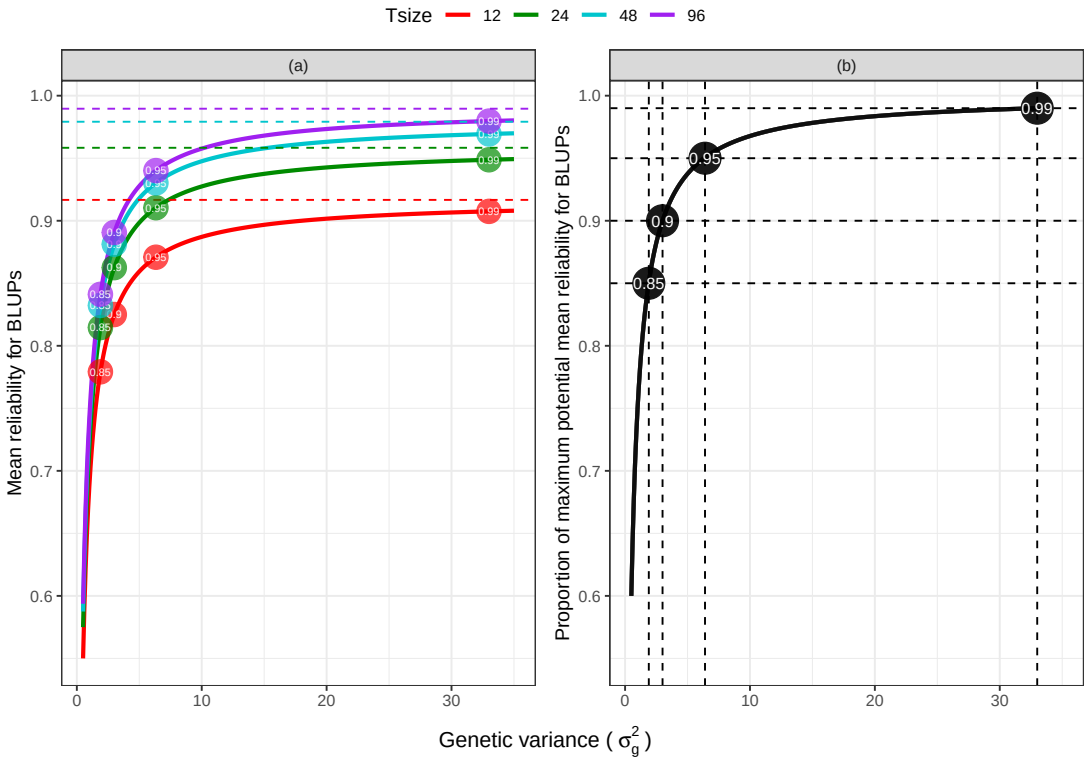

Figure 3.2: Response profiles for (a) mean reliability for BLUPs ($\bar{R}$), (b) proportion of maximum potential mean reliability for BLUPs ($\bar{R}_{X_p}$) over varying levels of genetic variance σ_g^2 from completely randomised designed trials with 12, 24, 48, and 96 varieties (Tsize), each with $b = 3$ replicates. Highlighted are $\bar{R}_{X_p} = \{0.85, 0.90, 0.95, 0.99\}$. Dashed lines in (a) represent the $\bar{R}_{X_p}$ values for each Tsize; and (b) the $\bar{R}_{X_p} = \{0.85, 0.90, 0.95, 0.99\}$ values and their corresponding σ_g^2 values.

3.3.4 Mean loss in reliability of EBLUPs

The significance of the design-based reliability values R_k^D and $\bar{R}^D$ are that these provide a general methodology for computing the reliabilities of BLUPs. As a result, these give an upper bound to the simulated reliability values of the EBLUPs (see Equation 3.5), with observed disparities between the design-based and simulated reliability values attributed as losses owing to the REML estimation of the variance parameters. I define loss as

$$\bar{R}_{\text{loss}} = \bar{R}^D - \bar{R}^S \tag{3.14}$$

For illustration, I now vary $b = \{2, 3, 4, 6, 8, 10, 15, 20, 25, 50\}$ for the CRD field trial simulation study example and complete the same steps as outlined in Section 3.1 for

each level of b. The mean simulation reliability of the EBLUPs and the design-based reliability of the BLUPs are given in Table 3.3, along with their associated losses. These are also presented visually in Figure 3.3, where (a) shows the relationship of the two reliability values and (b) their associated losses. Loss is shown to exhibit a clear non-linear response, with low replication presenting a larger loss in EBLUP reliability.

Table 3.3: Simulated ($\bar{R}^S$) and design-based ($\bar{R}^D$) trial reliability and loss ($\bar{R}_{\text{loss}}$) for the variety effects for the scenario with $m = 24$, with simulated parameters $\sigma_g^2 = 0.2$ and $\sigma_g^2 = 1$, over varying levels of variety replication b.

b	$\bar{R}^S$	$\bar{R}^D$	$\bar{R}_{\text{loss}}$
2	0.185	0.274	0.089
3	0.285	0.359	0.074
4	0.368	0.426	0.057
6	0.490	0.523	0.033
8	0.564	0.590	0.026
10	0.624	0.639	0.015
15	0.711	0.719	0.007
20	0.763	0.767	0.004
25	0.796	0.799	0.003
50	0.870	0.871	0.001

3.4 Additional simulation study information

This section provides additional information that is important for running a successful simulation study. In Section 3.4.1 I discuss the use of random number generation; then in Section 3.4.2 I provide the procedures to find model and variance parameter convergence of `ASReml-R` model fits; and finally in Section 3.4.3 I define the methodology in determining how many simulations are necessary to achieve a specified level of precision.

3.4.1 Random number generation

Simulation studies are often referred to as Monte Carlo simulations due to their connections with the casino and the roll of the dice. That is, they rely on a number of dice rolls, or in this case, the development of a number of random numbers. The insertion

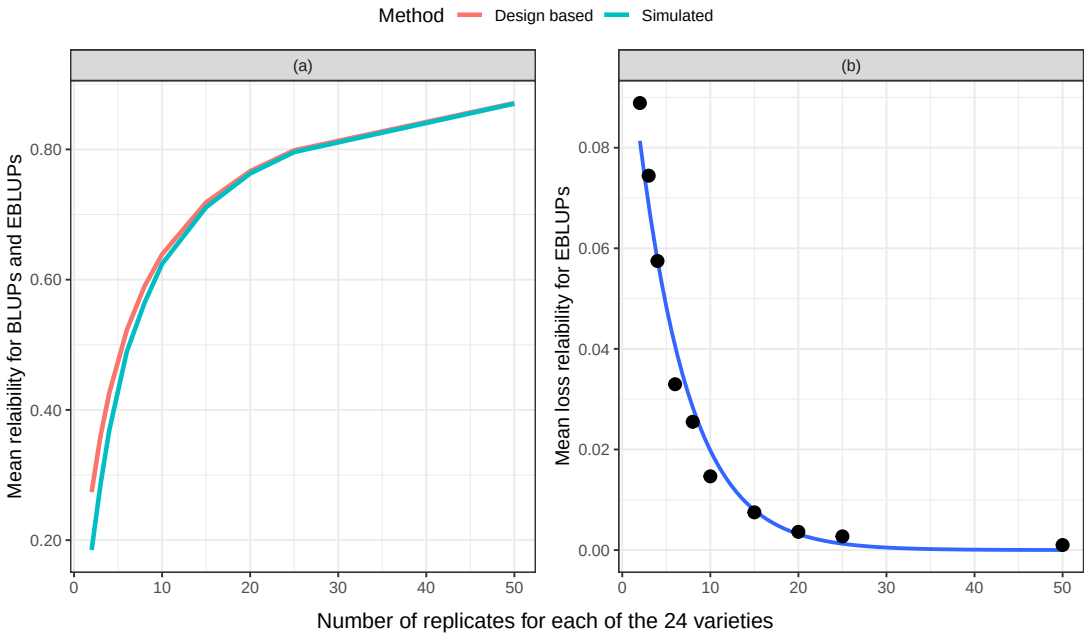

Figure 3.3: (a) Mean reliability for variety effects for design-based (BLUPs reliability) and simulated (EBLUPs reliability), against number of replicates. (b) Mean loss in reliability of the EBLUPs of variety effects, against number of replicates. The solid blue line in (b) represents a smooth line through the points.

of bias into the system, similar to the casino situation, is unacceptable since it may lead to erroneous outcomes. Random number generators (RNG) are the essential tool for simulation studies, with much research detailing the consequences of using them incorrectly (L'Ecuyer, 1990; Hellekalek, 1998; Ewald et al., 2008). "Bad random number generators may ruin a simulation" (Hellekalek, 1998).

In `R`, pseudo-random numbers created from programmed equations are used to generate random numbers. Despite the fact that these numbers appear to be random, they are not. Eventually, the numbers will repeat in the same sequence. It is critical to have repeatable outcomes in simulation research so that fascinating or troublesome results can be reproduced and questioned. The `set.seed` function in `R` is used to do this, which allows random numbers to be duplicated given a seed. It is critical that the seed numbers are not duplicated, as this will result in identical outcomes. As a result, a set of

seeds with the same length as the number of simulations is necessary. Using successive numbers is also risky since they include a correlation structure (Hellekalek, 1998), which can lead to false conclusions if correlations concealed in the random numbers and in the simulated system interact constructively (Compagner, 1995).

3.4.2 Model fitting

`ASReml-R` (Butler et al., 2017) was used to fit all statistical models in this thesis. To enable faster convergence, the `ASReml-R` model was given the true variance parameter values as the initial REML estimates of the variance parameters. The number of iterations necessary for convergence is recorded and utilised as a key outcome in the simulation studies that follow.

Each model was given a maximum of 10 updates (`asreml.update`), with a total of 13 iterations per update to allow for model convergence. Both convergence of the residual log-likelihood and parameter estimates changing by less than 1% from the previous iteration were used to determine convergence. Model fits that did not reach convergence after the ten updates were considered as failures. The convergence code is shown below.

```
asreml.options(ai.sing=TRUE, fail='soft')
for(z in 1:10){ #allow for up to 10 updates
if(!model.asr\$converge |
      max(summary(model.asr)\$varcomp\$'%ch',na.rm=T) > 1)
{ model.asr <- update(model.asr) }}
```

The `ai.sing=TRUE` and `fail='soft'` options in `asreml.options` are set, so that error messages do not halt the simulation run abruptly, but are converted to warning messages, thereby enabling the simulation to continue. It is noted that in the compilation of results, those runs containing convergence issues were excluded.

3.4.3 Number of simulations

Defining how many simulations S are required to achieve a specific level of precision is not always considered (Burghout, 2004; Truong et al., 2015). Typically, the number of simulations is determined by reviewing published simulation studies or past studies. Every simulation study, however, has a specific amount of variability that should be considered.

The number of simulations S in this thesis is determined by analysing the $\bar{R}^S$ values, with the objective of simulating to within 1% (E=0.01). This is accomplished by running a large pilot simulation study and partitioning into smaller sets denoted here as simblocks, and then examining the range of $\bar{R}^S$ results.

To illustrate this methodology, I again use the model in Equation 3.1, and similar to Section 3.3.4 I vary the levels of varietal replication $b = \{2, 4, 6, 8, 10\}$. For each replication scenario I then simulate $S = 20000$ times, and partition S into simblocks $= \{100, 200, 500, 800, 1000, 2000\}$. Hence, each simblock contains a different number of sets. For example, where simblock=100 contains 200 sets, and where simblock=2000 contains 10 sets. Table 3.4 shows the difference (E) between the top and bottom 2.5% quantiles for $\bar{R}^S$ for each simblock and replication level b combination. This suggests $S = 2000$ are required for all scenarios apart from $b = 2$ which would require a greater number of simulations.

Table 3.4: Difference (E) between the top and bottom 2.5% quantiles for $\bar{R}^S$ for each simblock and replication level b combination. For example for simblock 500 with $b = 2$, the correlations quantiles were 0.172 and 0.209, with a difference of 0.037. Yellow cells show those scenarios where an E=0.01 has been achieved.

	Simblock					
b	100	200	500	800	1000	2000
2	0.073	0.057	0.037	0.030	0.028	0.018
4	0.074	0.047	0.030	0.021	0.018	0.008
6	0.061	0.041	0.020	0.021	0.010	0.005
8	0.056	0.038	0.021	0.013	0.019	0.005
10	0.049	0.039	0.018	0.016	0.017	0.010

3.5 Concluding remarks

The simulation study approaches, and reliability based principles of BLUPs provided in this chapter serve as the foundation for the simulation studies described in Chapters 6 and 9. The former examines the impact of variety connectivity on the reliability of varietal predictions from a Factor Analytic Multi Environment Trial analysis. The later simulation study investigates the performance of a new diagnostic measure which offers an alternative to the traditional variety connectivity methodology.

Chapter 4

Motivating datasets

This chapter describes the datasets used as motivation throughout this thesis. These are used for demonstration of statistical techniques; and for their key attributes which are used in forthcoming simulation studies. Two distinct types of motivating datasets are considered, namely a late-stage variety evaluation dataset in which only elite varieties are considered and a full plant breeding dataset in which varieties from all stages of testing are considered.

Many countries generate and analyse late stage evaluation datasets on an annual basis, including the Australian National Variety Trials (NVT) system (NVT online) and the United Kingdom Agriculture and Horticulture Development Board (AHDB) recommended list (RL) system (AHDB RL system). In particular, the NVT system examines near-release varieties submitted by private plant breeding companies with the goal of providing growers with independent information. Each year, around 700 trials covering 10 different crops are conducted. For each crop, MET datasets are created, and the analysis approach is a FALMM (see Chapter 2), in which the variety effects are considered to be independent (that is, the varieties are assumed to be unrelated). The late-stage evaluation dataset included in this thesis was compiled using Stage 4 (S4) trials undertaken by the National Oat Breeding Program (NOBP) between 2012 and 2016. It should be emphasised that this is not the type of dataset recommended for breeding program variety selections (see Chapter 7), however, it is considered here to resemble typical late-stage evaluation MET datasets, similar to those used in the NVT and AHDB RL systems. This dataset is here-after referred to as the Oat dataset.

The second motivating dataset reflects the full structure of a plant breeding program, that is, multiple stages across years. The Stage 1 (S1) to S4 trials conducted between 2013 and 2018 by the Durum Breeding Australia (DBA) program are used for this purpose. In contrast to the first motivating example, the method of analysis for such data will incorporate pedigree information in order to accommodate relationships between varieties. This dataset is here-after referred to as the Durum dataset.

This Chapter is arranged as follows: the Oat dataset is described in Section 4.1; and then in Section 4.2 the Durum dataset is described.

4.1 Oat dataset

The NOBP has objectives of developing high quality export hay varieties along with improved milling oats. The program is based in Adelaide at the University of Adelaide, Waite Campus with operations run by breeders Dr. Pamela Zwer and Dr. Sue Hoppo.

Oats (*Avena sativa*) are grown in Australia across the grain cropping regions of south-west Western Australia (WA), South Australia's (SA) Eyre and York Peninsulas, western and northeastern Victoria (Vic), and the Riverina and central New South Wales (NSW). The NOPB-derived varieties account for around 85% of oat production in South Eastern Australia.

The NOBP S4 trials sown between 2012 and 2016 are considered here. Within S4, varieties are often tested for multiple years, whereas in earlier stages, varieties were either culled or promoted after one season. A small elite subset of the varieties are submitted to the NVT program for further evaluation. Furthermore, following testing in both S4 and NVT, top performing varieties are considered for commercial release.

4.1.1 Description of data

Table 4.1 provides a summary of the Oat dataset. There were 49 trials spread throughout 43 environments. An environment is defined in this context as a year and location

combination exposed to comparable management practices. When there are multiple trials within an environment, these are classified as co-located trials (Smith et al., 2021a). There are 12 trials classified as co-located in the Oat dataset. Trials in the Oat dataset were sown in WA, SA, Vic, and central NSW, as illustrated in Figure 4.1.

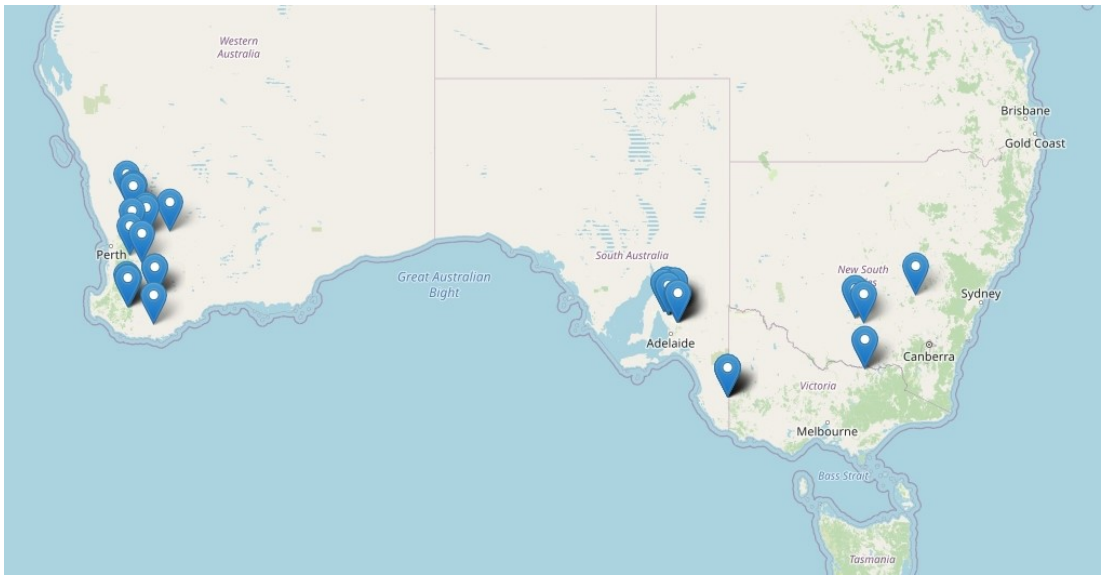

Figure 4.1: Map showing the locations of the trials in the Oat dataset.

The number of trials per year ranged from 7 (2014) to 12 (2012 and 2013). Each trial was laid out in a contiguous array of plots with between 3 and 30 columns and 6 to 60 rows. Varieties were randomised in complete replicate blocks with three replicates each, with replicate blocks aligned with columns, in so-called randomised complete block (RCB) designs. As an example, Figure 4.2 presents the randomisation for OMaB15CUND6. As shown, the trial is sown in a rectangular array consisting 12 columns and 14 rows, with replicate blocks spanning four columns. It is noted that this trial is used for demonstration of a single trial analysis in Chapter 5.

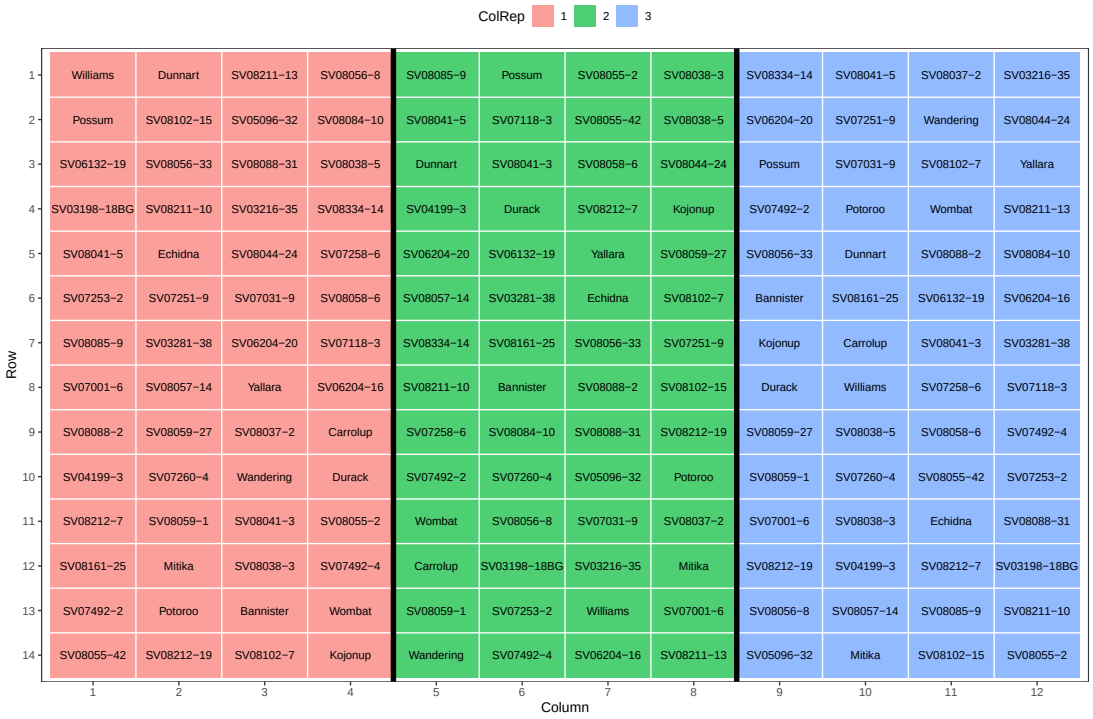

Figure 4.2: Spatial layout for OMaB15CUND6. Colours as represented by the legend identify the replicate blocks (ColRep). Text represents the varieties sown in each plot. Plot dimensions are 10m in length by 1.75m in width. Note: Plots as depicted are not drawn to scale.

The number of varieties in each environment ranged from 24 to 60, with a median of 52 varieties. Within a year, the same cohort of varieties generally appeared in all environments. There were 163 varieties across all environments, which include 15 commercial varieties and 148 test varieties. These are the varieties that are being evaluated for commercial release, retention in S4 next year, or considered for inclusion in NVT.

Figure 4.3 presents the variety connectivity across environments, displaying the number of varieties in common between all pairs of environments. This demonstrates that the number of common varieties ranges from 13 to 60. This is also presented in tabular format in Table 4.2 by year, where the number of varieties per year ranged from 48 (2014) to 65 (2013), with at least 16 (between 2012 and 2016) varieties in common between years.

The trial mean yield (TMY) for each trial (see Table 4.1) varied from 0.62 to 6.19 t/ha, corresponding to a tenfold increase in yield. The change is often related to the amount of accessible soil moisture. For example, winter 2016 was the second wettest since records began in 1900, with several regions exceeding prior records. In comparison to previous years, these rains resulted in high TMY values for 2016.

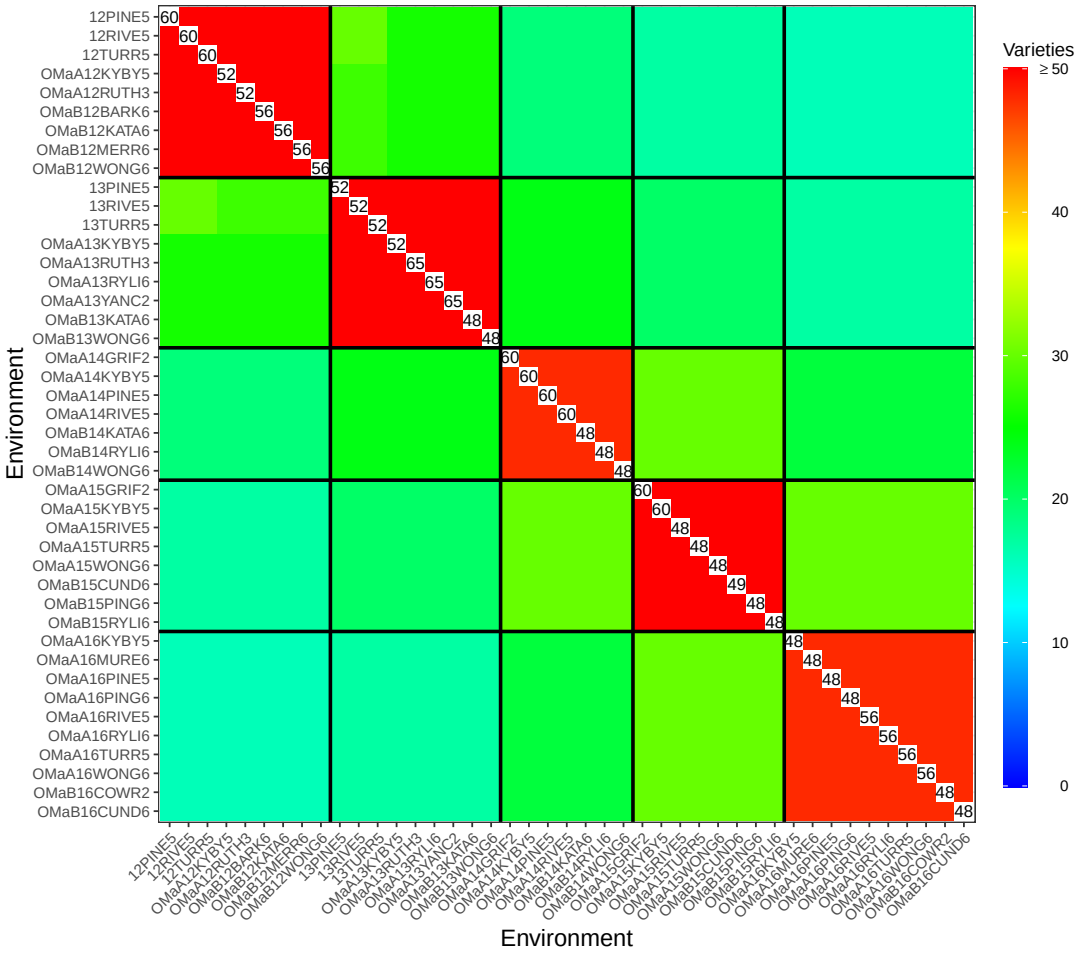

Figure 4.3: Variety connectivity across environments for the Oat dataset. The colours on the off-diagonals indicate the number of common varieties between pairs of environments. Numbers on the diagonals show the number of varieties in each environment. Boundaries for years are indicated by the black lines (2012 - 2016 inclusive from left to right and top to bottom).

Table 4.1: Summary of the trials in the Oat dataset. Horizontal dashed lines separates the years. Trial Mean Yield (TMY) in tonnes per hectare is shown in the last column. [1]The final digit in the trial and environment acronyms denotes the Australian state where it was sown, where 2=NSW, 3=Vic, 5=SA, 6=WA.

Year	Trial[1]	Environment[1]	Number of			TMY
			Columns	Rows	Varieties	
2012	OMaB12BARK6	OMaB12BARK6	12	13	52	1.53
	OMaB12KATA6	OMaB12KATA6	12	13	52	2.58
	OMaA12KYBY5	OMaA12KYBY5	12	13	52	3.65
	OMaB12MERR6	OMaB12MERR6	12	13	52	0.62
	OMaA12PINE5	12PINE5	12	13	52	2.64
	OMaB12PINE5	12PINE5	12	9	36	2.57
	OMaA12RIVE5	12RIVE5	12	13	52	3.67
	OMaB12RIVE5	12RIVE5	12	9	36	4.45
	OMaA12RUTH3	OMaA12RUTH3	3	52	52	2.80
	OMaA12TURR5	12TURR5	12	13	52	2.71
	OMaB12TURR5	12TURR5	12	9	36	3.00
	OMaB12WONG6	OMaB12WONG6	12	13	52	4.21
2013	OMaB13KATA6	OMaB13KATA6	12	15	60	4.20
	OMaA13KYBY5	OMaA13KYBY5	12	15	60	3.80
	OMaA13PINE5	13PINE5	12	6	24	2.33
	OMaB13PINE5	13PINE5	12	15	60	2.44
	OMaA13RIVE5	13RIVE5	12	6	24	3.49
	OMaB13RIVE5	13RIVE5	12	15	60	3.56
	OMaA13RUTH3	OMaA13RUTH3	3	60	60	3.63
	OMaA13RYLI6	OMaA13RYLI6	12	15	60	4.71
	OMaA13TURR5	13TURR5	12	6	24	3.33
	OMaB13TURR5	13TURR5	12	15	60	3.00
	OMaB13WONG6	OMaB13WONG6	12	15	60	4.54
	OMaA13YANC2	OMaA13YANC2	30	6	60	5.63
2014	OMaA14GRIF2	OMaA14GRIF2	6	24	48	6.06
	OMaB14KATA6	OMaB14KATA6	12	12	48	2.44
	OMaA14KYBY5	OMaA14KYBY5	12	12	48	1.76
	OMaA14PINE5	OMaA14PINE5	12	12	48	3.16
	OMaA14RIVE5	OMaA14RIVE5	12	12	48	4.36
	OMaB14RYLI6	OMaB14RYLI6	12	12	48	2.57
	OMaB14WONG6	OMaB14WONG6	12	12	48	2.77
2015	OMaB15CUND6	OMaB15CUND6	12	14	56	2.27
	OMaA15GRIF2	OMaA15GRIF2	6	28	56	5.84
	OMaA15KYBY5	OMaA15KYBY5	12	14	56	3.30
	OMaB15PING6	OMaB15PING6	12	14	56	2.36
	OMaA15RIVE5	OMaA15RIVE5	12	14	56	2.92
	OMaB15RYLI6	OMaB15RYLI6	12	14	56	3.16
	OMaA15TURR5	OMaA15TURR5	12	14	56	2.50
	OMaA15WONG6	OMaA15WONG6	12	14	56	0.84
2016	OMaB16COWR2	OMaB16COWR2	6	24	48	4.41
	OMaB16CUND6	OMaB16CUND6	12	12	48	3.29
	OMaA16KYBY5	OMaA16KYBY5	12	12	48	4.18
	OMaA16MURE6	OMaA16MURE6	12	12	48	2.41
	OMaA16PINE5	OMaA16PINE5	12	12	48	6.19
	OMaA16PING6	OMaA16PING6	12	12	48	4.25
	OMaA16RIVE5	OMaA16RIVE5	12	12	48	3.89
	OMaA16RYLI6	OMaA16RYLI6	12	12	49	5.95
	OMaA16TURR5	OMaA16TURR5	12	12	48	5.13
	OMaA16WONG6	OMaA16WONG6	12	12	48	5.84

Table 4.2: Number of varieties in common within and between years for the Oat dataset. Diagonal elements shown in bold represent the number of varieties present in each year. Off-diagonals represent the number of common varieties between pairs of years.

Year	2012	2013	2014	2015	2016
2012	**60**	30	19	17	16
2013	30	**65**	24	20	17
2014	19	24	**48**	30	22
2015	17	20	30	**56**	30
2016	16	17	22	30	**49**

4.2 Durum dataset

Durum wheat (*Triticum Durum* L. ssp. Durum Desf.) is a spring wheat that is ground into semolina, used to make pasta, and a finer flour used to make bread or pizza dough (Kadkol et al., 2022). It is cultivated in Australia's northern grains areas, Northern and Southern NSW, Southern and Central Queensland (Qld), SA's mid north and Yorke Peninsula regions, and areas of Victoria's (Vic) Wimmera region. Dr. Gururaj Kadkol directs the Northern Durum Breeding Australia (NDBA) program, which is located at the Tamworth Agricultural Institute.

Considered here are the Durum wheat breeding trials from the NDBA program sown between 2013 and 2018. Trials in the Durum dataset were sown in Northern NSW, Southern and Central Qld, as illustrated in Figure 4.4. Unlike the Oat dataset, I consider all stages of breeding. To show the full set of selections, Figure 4.5 shows the BL progression and retention across stages and years. The S1 trials in any year contains, on average, 1120 BL evaluated in one environment. As the program progresses the number of BL decrease and the number of environments increase, until S4 which, on average, contains 48 BL evaluated in eight environments. As with the oats, the Durum program sends a small set of elite varieties to the NVT program.

As shown in Figure 4.5, there are four selection decisions in 2018 for S1 through to S4. The number and direction of the arrows show the progression and retention of the BL. As shown in greater detail the progression paths of the selection decisions for S4, the progression path for the 56 BL of interest in Figure 4.6. This shows that 13 of the BL

have progressed from the S1 cohort in 2013, 12 from the S1 cohort in 2014, and 31 from the S1 cohort in 2015. Similarly, the progression and retention of the selection decisions for S3 BL are shown in Figure 4.7. This is similar for the 93 BL of interest, where 5, 22, and 66 BL advanced from the S1 cohorts in 2014, 2015, and 2016, respectively.

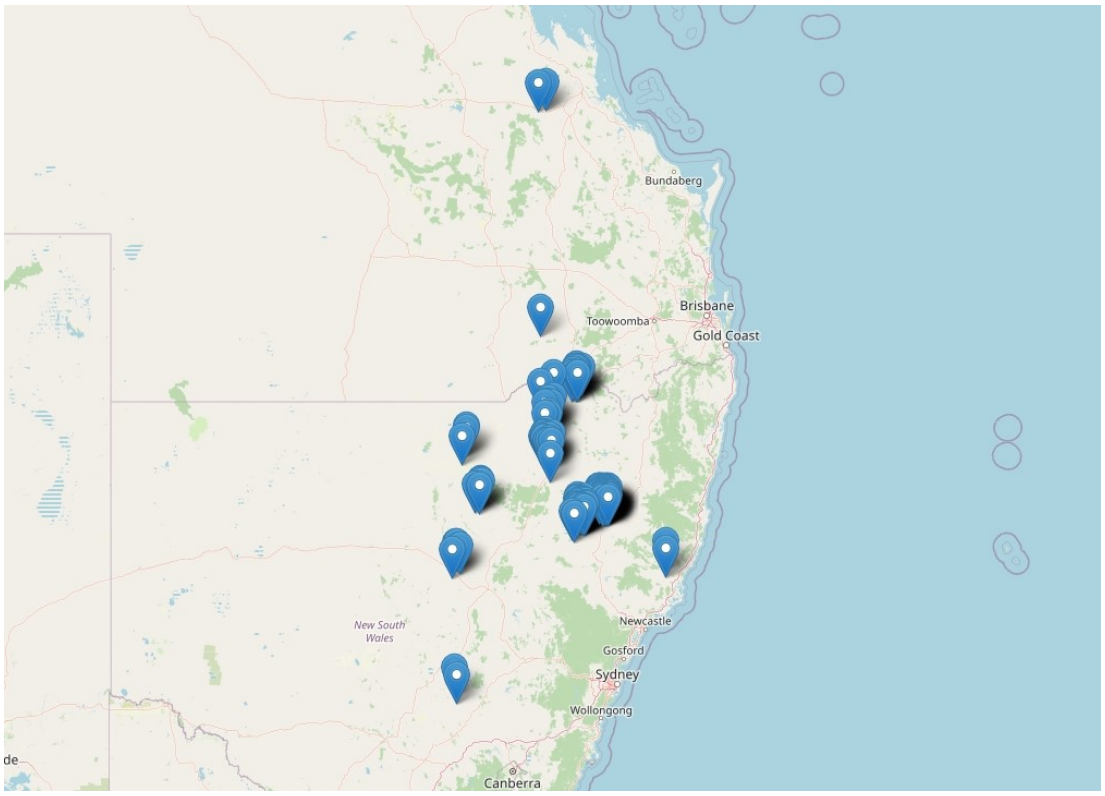

Figure 4.4: Map of Eastern Australia showing locations of the environments in the Durum dataset.

4.2.1 Description of data

Table 4.3 summarises the Durum environments conducted between 2013 and 2018. There were 49 environments and 139 trials, with a total of 7158 varieties evaluated throughout four breeding stages. In contrast to the Oats dataset, the Durum dataset is dominated by co-located trials, with 22 environments including multiple trials. Several of these environments consist of more than two trials, with as many as 13 trials (2017-Tworth) conducted, typically spanning several breeding stages.

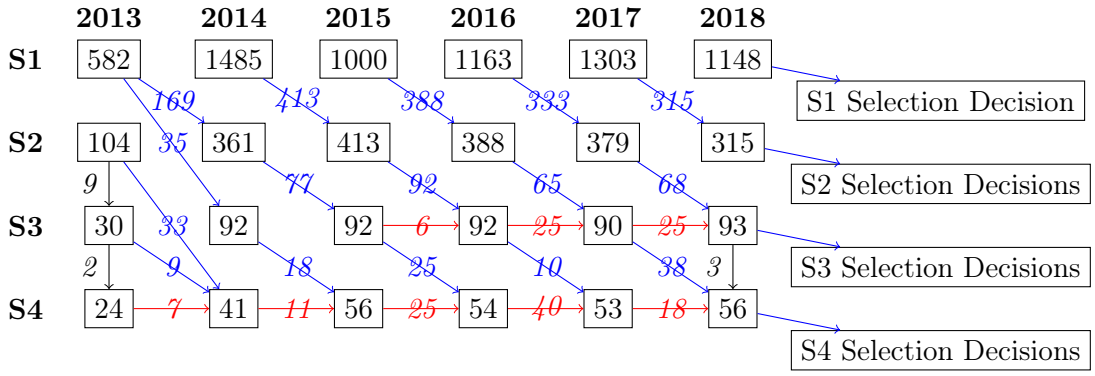

Figure 4.5: Test varieties progression and retention across stages in the Durum dataset. Numbers and arrows indicate the number and direction of BL respectively that have: progressed to the next stage (blue), retained in the same stage (red), fast tracked to the next stage (grey).

The trials were laid out in contiguous arrays of plots, and designed as grid-plot, partially replicated (p-rep) (Cullis et al., 2006), or RCB with two or three replicates. Figure 4.8 shows the randomisation for the 2016-Breeza environment, which includes both an S3 and an S4 trial as an example of co-located trials. They are stacked in the column direction, with the S4 trial sown above the S3 trial. Each were sown into rectangular arrays with 12 columns and 15 or 16 rows. S3 and S4 trials were designed as RCB with three and two replicates respectively. It is noted that this co-located environment is used for demonstration for a single co-located trial pedigree analysis in Chapter 8.

Table 4.4 records the number of varieties within and between years. This shows the number of common varieties between successive years is roughly one-third from the previous year with similar decreases over further years. The variety connectivity between pairs of environments is shown in the lower triangle of Figure 4.9. As shown, there is poor variety connectivity with many environments, which is typically seen between environments which are two or more years apart. As the dataset comprises environments across breeding stages, this is to be expected given the objectives of the progressive breeding cycle.

Table 4.3: Summary of environments in the Durum dataset (2013 - 2018): number of trials for each stage of testing (S1, S2, S3, S4); the total number of trials, plots, varieties; and mean yield (tonnes per hectare). Horizontal dashed lines separate the years.

Year	Environment	Number of trials					Number of		Mean
		S1	S2	S3	S4	Total	Plots	Varieties	yield (t/ha)
2013	2013-Breeza	2	0	1	1	4	888	642	3.14
	2013-Coonamble	0	0	0	1	1	96	32	3.28
	2013-Duaringa	0	0	0	1	1	96	32	3.06
	2013-Moree	0	2	1	1	4	408	151	3.22
	2013-Nstar	0	0	1	1	2	168	60	3.31
	2013-Spridge	0	0	0	1	1	96	32	3.63
	2013-Trangie	0	0	0	1	1	96	32	2.47
	2013-Tworth	0	2	1	1	4	408	151	2.02
	2013-Walgett	0	2	0	1	3	336	123	3.62
2014	2014-Breeza	3	0	0	1	4	1440	982	4.33
	2014-Coonable	0	0	0	1	1	144	48	4.41
	2014-Edgeroi	0	4	0	1	5	912	409	2.80
	2014-Moree	0	0	1	1	2	336	140	3.01
	2014-Nstar	0	0	1	1	2	336	140	3.84
	2014-Spridge	0	0	0	1	1	144	48	7.15
	2014-Trangie	0	0	0	1	1	144	48	2.42
	2014-Tworth	2	4	1	1	8	1804	1052	3.86
2015	2015-Breeza	0	0	2	2	4	744	152	5.04
	2015-Bribbaree	0	0	0	1	1	180	60	2.80
	2015-Coonamble	0	0	0	1	1	180	61	2.36
	2015-Duaringa	0	0	0	1	1	180	60	3.09
	2015-Edgeroi	0	4	1	1	6	1236	554	1.75
	2015-Nbri	0	0	0	1	1	180	60	5.47
	2015-Nstar	0	0	1	2	3	552	152	4.74
	2015-Trangie	0	0	0	1	1	180	60	2.93
	2015-Tulloona	0	0	0	1	1	180	60	4.37
	2015-Tworth	6	4	1	1	12	2424	1555	4.00
2016	2016-Breeza	0	0	1	1	2	372	152	4.35
	2016-Edgeroi	0	0	0	1	1	180	60	4.79
	2016-Gurley	0	0	0	1	1	180	60	5.62
	2016-Nstar	0	0	1	2	3	552	152	5.49
	2016-Tworth	6	3	1	1	11	2628	1704	4.81
2017	2017-Blbgra	0	0	0	1	1	180	60	1.12
	2017-Breeza	0	0	1	1	2	384	158	5.31
	2017-Bribbaree	0	0	0	1	1	180	60	1.20
	2017-Coonamble	0	0	0	1	1	180	60	1.61
	2017-Edgeroi	0	0	0	1	1	180	60	3.93
	2017-Garah	0	0	0	1	1	180	60	1.84
	2017-Gurley	0	0	0	1	1	180	60	2.12
	2017-Nstar	0	0	1	1	2	384	158	3.41
	2017-Tworth	7	3	1	2	13	3014	1836	4.26
	2017-Westmar	0	0	0	1	1	180	60	2.24
2018	2018-Blbgra	0	0	0	1	1	198	66	1.24
	2018-Breeza	6	3	1	1	11	2502	1629	5.53
	2018-Coonamble	0	0	0	1	1	198	66	1.56
	2018-Gurley	0	0	1	0	1	210	105	2.23
	2018-Moree	0	0	0	1	1	198	66	1.51
	2018-Trangie	0	0	0	1	1	198	66	1.02
	2018-Tworth	0	3	1	1	5	1074	481	2.24

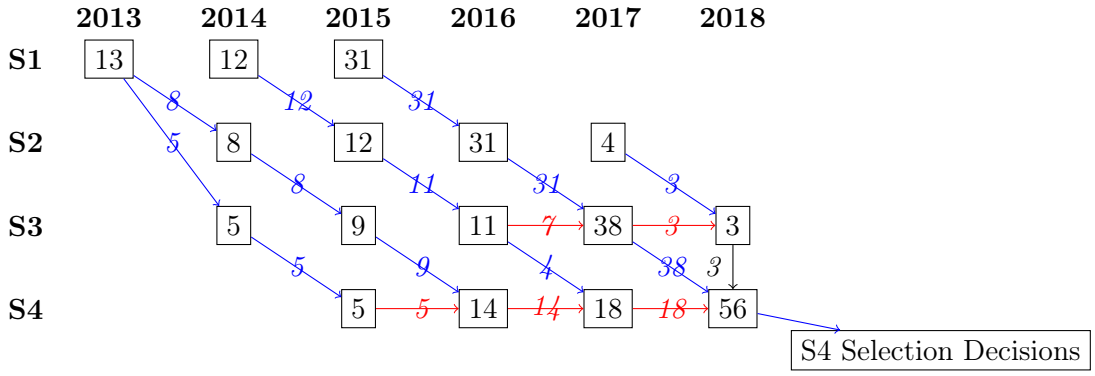

Figure 4.6: Test varieties progression and retention across stages for the 56 S4 varieties considered for selection decisions in 2018 for the Durum dataset. Numbers and arrows indicate the number and direction of BL respectively that have: progressed to the next stage (blue), retained in the same stage (red), fast tracked to the next stage (grey).

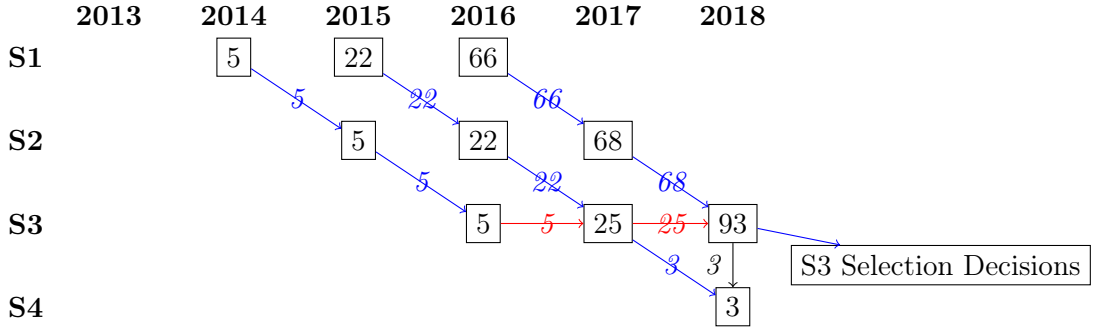

Figure 4.7: Test varieties progression and retention across stages for the 93 S3 varieties considered for selection decisions in 2018 for the Durum dataset. Numbers and arrows indicate the number and direction of BL respectively that have: progressed to the next stage (blue), retained in the same stage (red), fast tracked to the next stage (grey).

4.2.2 Description of pedigree data

The pedigree information associated with the Durum dataset contained 7623 records. This relates to all 7158 varieties grown in the environments and 465 parental varieties that were not grown in the environments. The numerator relationship matrix (NRM) (see Chapter 2) was formed using the `pedicure` package (Butler, 2016) in `R` (R Core Team, 2020). The inbreeding coefficients of varieties with phenotypic data ranged from 1.750 to 1.998 with a mean of 1.903. Similar to the number of common varieties between years, the upper triangle of Table 4.5 shows the number of parents in common within and between years. This shows reasonable connections between years, with a gentle decrease in common parents in successive years.

Table 4.4: Number of varieties in common within and between years for the Durum dataset. Diagonal elements shown in bold represent the number of varieties present in each year. Off-diagonals represent the number of common varieties between years.

Year	2013	2014	2015	2016	2017	2018
2013	**733**	252	90	39	31	20
2014	252	**1986**	534	140	74	36
2015	90	534	**1556**	540	145	83
2016	39	140	540	**1704**	479	150
2017	31	74	145	479	**1837**	477
2018	20	36	83	150	477	**1629**

Table 4.5: Number of parents in common within and between years for the Durum dataset. Diagonal elements shown in bold represent the number of parents present in each year. Off-diagonals represent the number of common parents between years.

Year	2013	2014	2015	2016	2017	2018
2013	**58**	43	32	20	21	17
2014	43	**53**	36	23	24	20
2015	32	36	**50**	35	35	30
2016	20	23	35	**36**	34	29
2017	21	24	35	34	**43**	33
2018	17	20	30	29	33	**40**

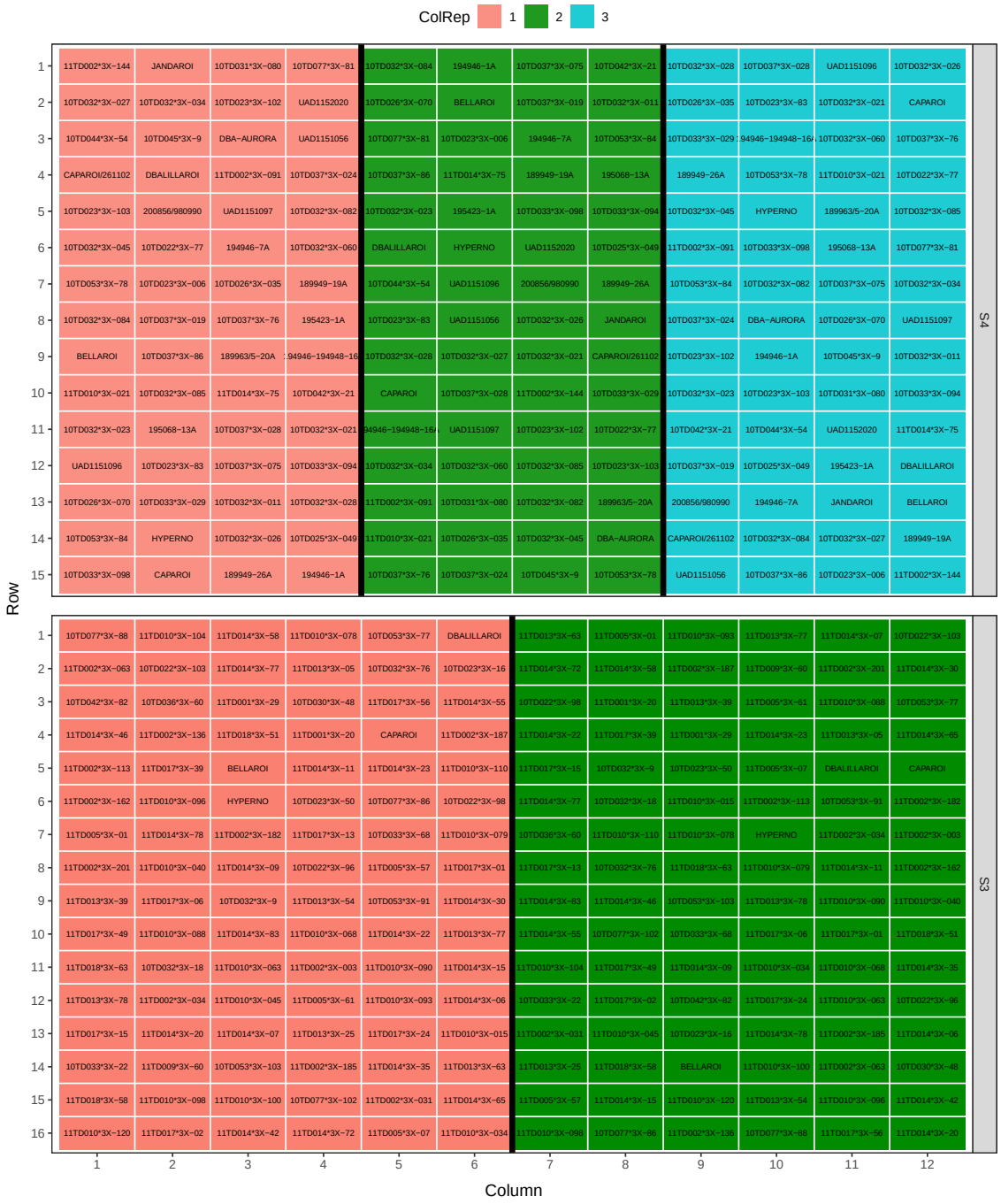

Figure 4.8: Spatial layout for 2016-Breeza. The S4 trial (top) is sown above the S3 trial (bottom) with no gaps. Colours as represented by the legend identify the replicate blocks (ColRep). Text represents the varieties sown in each plot. Plot dimensions are 10m in length and 1.75m in width. Plots as depicted are not drawn to scale.

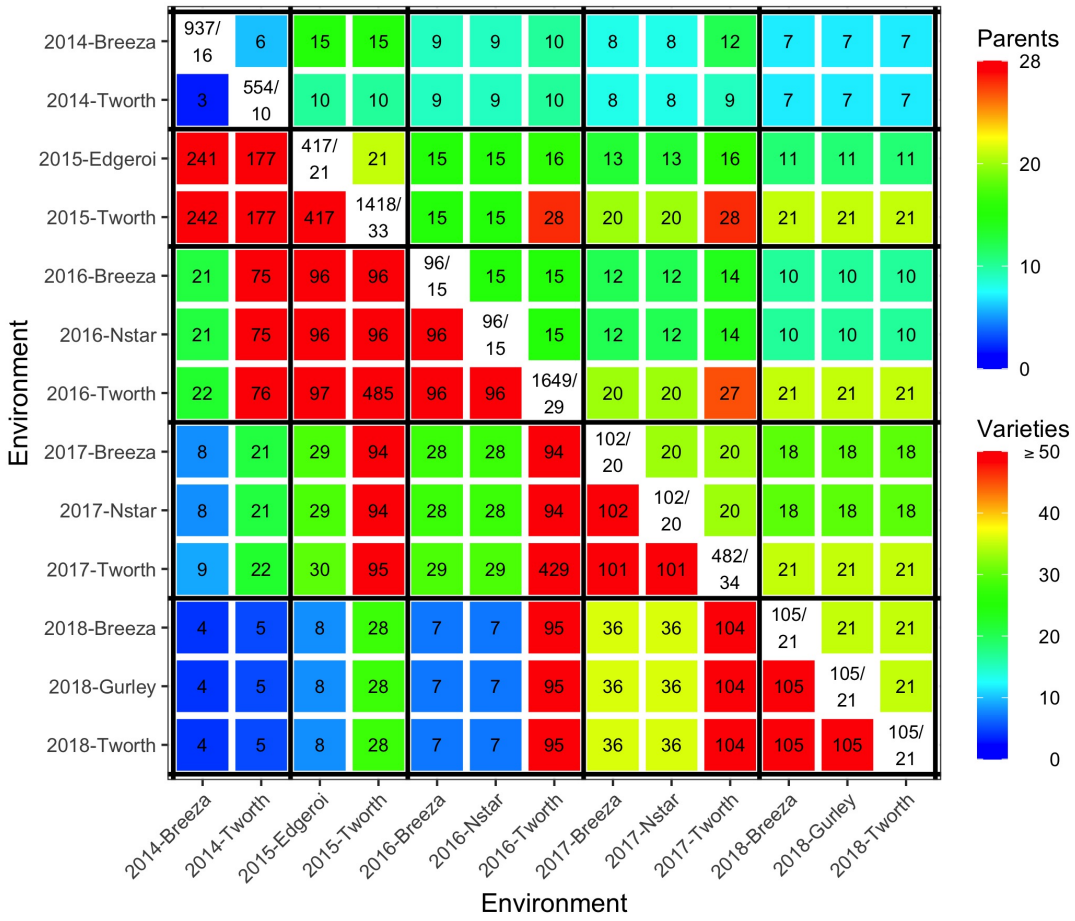

Figure 4.9: Heatmap of the number of common varieties (lower triangle) and parents (upper triangle) between all pairs of environments in the Durum dataset. The numbers in the diagonal boxes represent the number of varieties/parents in each environment; the colours as referenced in the legends. Boundaries for years are indicated by the black horizontal and vertical lines.

4.3 Concluding remarks

This thesis heavily relies on the Oat and Durum datasets described in this chapter. The Oat dataset is used to resemble standard late-stage variety testing procedures used across the world with analyses involving independent variety effects. The Durum dataset, on the other hand, depicts the entire structure of a plant breeding program, and with the availability of pedigree information enables the analysis to allow for the relationships between varieties.

The Oat dataset is first used to demonstrate single trial and MET analyses in Chapter 5. The structural elements of trial size and dimensions of the Oat dataset, as well as their estimated variance parameters, are then used in a comprehensive simulation study in Chapter 6. Finally, in Chapter 9, the Oat dataset is used to illustrate the methods of a new diagnostic.

The Durum dataset provides the motivation in Chapter 7, which outlines an approach for generating MET datasets that maximises the information available for selection decisions. Using this methodology, an appropriate Durum dataset is constructed, and with the corresponding pedigree file, is subsequently utilised in Chapter 8 to illustrate co-located trial and MET pedigree analyses. Then, in Chapter 9, a subset of the Durum dataset is used as the motivation to both test the performance of the approach in a simulation study and to demonstrate the methodologies.

Chapter 5

Statistical analysis of the Oat dataset

This chapter uses the Oat dataset presented in Chapter 4 to demonstrate the statistical methodology described in Chapter 2 with analyses involving independent variety effects. This dataset and corresponding analysis is used to resemble standard late-stage variety testing procedures used across the world. The structural elements, results and summaries contained in this chapter are used as the motivation to the simulation study in Chapter 6 to investigate the effects of variety connectivity on the reliability of varietal predictions from a FA MET analysis.

Firstly, single and co-located trial analyses are illustrated in Sections 5.1 and 5.2, then a summary of all single environment analyses is given in Section 5.3. Finally, the MET analysis for the full Oat dataset is presented in Section 5.4.

5.1 Spatial analysis of a single trial

The spatial analysis of a single trial is demonstrated using the S4 trial in Cunderdin in 2015 (OMaB15CUND6). Grain yield is the trait of interest, which is measured in tonnes per hectare (t/ha), with plot yields ranging from $\sim$ 1.4 to 3.4 t/ha, with no missing plot yields. This trial was sown in a rectangular grid with $c = 12$ columns by $r = 14$ rows, ($n = 168$ plots), and $m = 56$ varieties replicated $b = 3$ times, as illustrated in Figure 4.2. Complete replicates were aligned to blocks of four columns.

5.1.1 Statistical analysis

The methods outlined in Chapter 2 were used for the analysis of OMaB15CUND6. An initial model (M1) comprises terms reflecting the trial design and an AR1×AR1 process (see Chapter 2) for the errors was used. The LMM for the (168×1) data vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_{168})^\top$ may be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_b}\boldsymbol{u_b} + \boldsymbol{e} \tag{5.1}$$

where $\boldsymbol{\tau}$ is the overall mean with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the (56×1) vector of random Variety effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_b}$ is the (3×1) vector of random replicate block effects with associated design matrix $\boldsymbol{Z_b}$; and $\boldsymbol{e}$ is the (168×1) vector of errors. It is assumed that

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{u_b} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_g} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where,

- $\boldsymbol{G_g} = \sigma_g^2 \boldsymbol{I}_{56}$ (Genetic variance matrix)
- $\boldsymbol{G_b} = \sigma_b^2 \boldsymbol{I}_3$ (Replicate block variance matrix)
- $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Sigma}_c(\rho_c) \otimes \boldsymbol{\Sigma}_r(\rho_r)$ (Error variance matrix)

where $\boldsymbol{\Sigma}_c$ (12×12) and $\boldsymbol{\Sigma}_r$ (14×14) correlation matrices for columns and rows respectively.

The associated `ASReml-R` code for Model M1 is given as

```
M1 <- asreml(yield ~ 1, random=~ Variety + ColRep,
residual =~ ar1(Column):ar1(Row), data=o15CUND6.df)
```

where `yield` is the data vector of plot yields (t/ha); `1` denotes the overall mean; `Variety, ColRep` are terms representing the random variety and replicate block effects respectively; the errors are modelled using an AR1×AR1 structure as given by `ar1(Column):ar1(Row)`; and `o15CUND6.df` is the data object.

Potential outliers were inspected using studentised conditional residuals (Smith & Cullis, 2021), which can be obtained from `ASReml-R` using the code below.

```
aom.df <- update(M1, aom=TRUE)
```

Inspection of these residuals did not identify any outliers as shown in Figure 5.1, which presents a normal probability plot of the studentised conditional residuals.

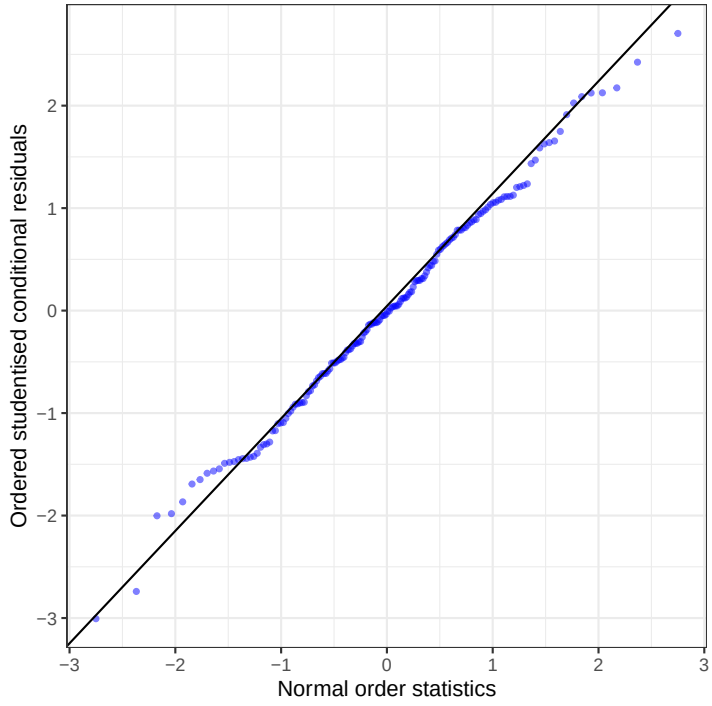

Figure 5.1: Normal probability plot for studentised conditional residuals from M1 of the analysis of OMaB15CUND6. Black line represents a 1-1 line.

Upon inspection of the residual and variogram plots from Model M1 (Model 5.1) as presented in Figures 5.2 and 5.4(a), there is clear evidence of a column effect. This is evident in Figure 5.2 as the points in each panel are not evenly spread about zero. For example, mostly all points are below zero in column 1 and all are above zero in column 6. The variogram in Figure 5.4(a) also shows this as the face along the column direction is not smoothly approaching a plateau but has distinct jumps.

Model M1 is adjusted to include random `Column` effects $\boldsymbol{u_c}$ with $\text{var}(\boldsymbol{u_c}) = \sigma_c^2 \boldsymbol{I}_{12}$, which is denoted as Model M2. There is a significant change to the log likelihood between

Models M1 and M2 with a REMLRT = 12.80 (see Table 5.1). Both Figure 5.3 and 5.4(b) from Model M2 show improvements to Model M1 as represented by the smooth profile in the column direction. However, there is some indication of row effects, with the variogram not reaching a plateau and with noticeable surges in the row direction.

I make an adjustment to Model M2 by including random `Row` effects $\boldsymbol{u_r}$ with $\text{var}(\boldsymbol{u_r}) = \sigma_r^2 \boldsymbol{I}_{14}$, which is denoted as Model M3. This shows a significant improvement between Models M2 to M3 with a REMLRT = 9.96 (see Table 5.1). The variogram in Figure 5.4(c) from Model M3 appears more like the theoretical and therefore it is chosen as the final model. Summaries of the three models are presented in Table 5.1, which gives the number of variance parameters, the residual log-likelihood (ℓ_R), and the REMLRT statistics between models.

Table 5.1: OMaB15CUND6: Summary of models fitted: sources of variation; residual log-likelihood (ℓ_R); number of variance parameters; and likelihood ratio test (REMLRT).

	Source of variation		Variance			
Model	Global/Extraneous[a]	local[b]	parameters	ℓ_R	REMLRT[c]	
M1		AR1×AR1	5	99.38		
M2	ran(Column)	AR1×AR1	6	105.78	12.80	***
M3	ran(Row)	AR1×AR1	7	110.76	9.96	***

[a] ran(Column) and ran(Row) represent the random effects for the column and row factors.
[b] Correlation structures for separable spatial process: AR1 = autoregressive of order 1.
[c] Comparison of model i with model $i-1$; *** $p < 0.001$.

The REML estimates of the variance parameters for the three models are given in Table 5.2. Furthermore, the average trial reliability $\bar{R} = 0.644$ (see Section 2.3.1 and Equation 3.9) and corresponding proportion to the maximum potential reliability is $\bar{R}_{X_p} = 0.656$ (see Equation 3.12). These values will be further explored in Section 5.3 with the analysis of all environments.

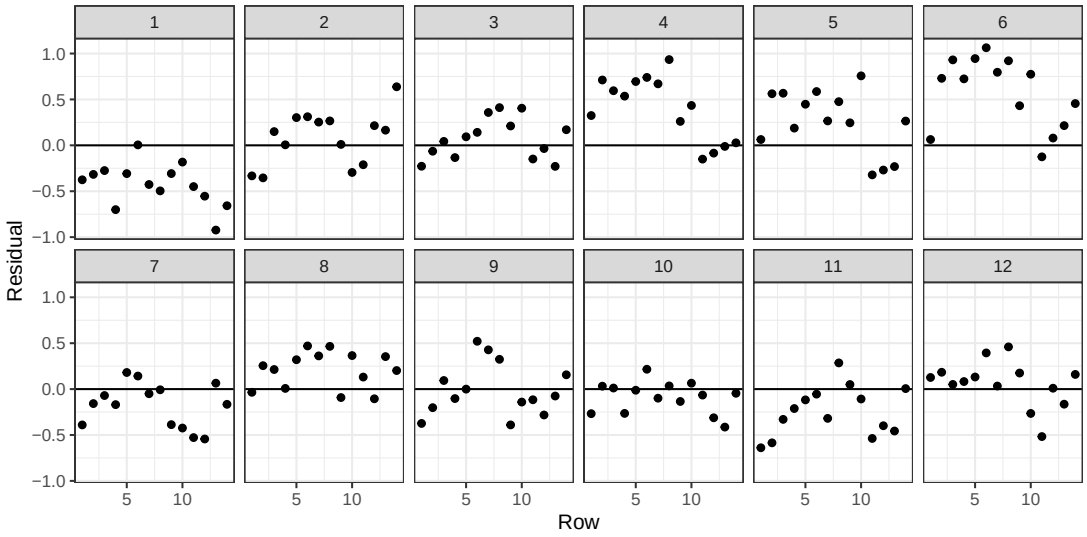

Figure 5.2: OMaB15CUND6 for Model M1. Residuals against row number for each column.

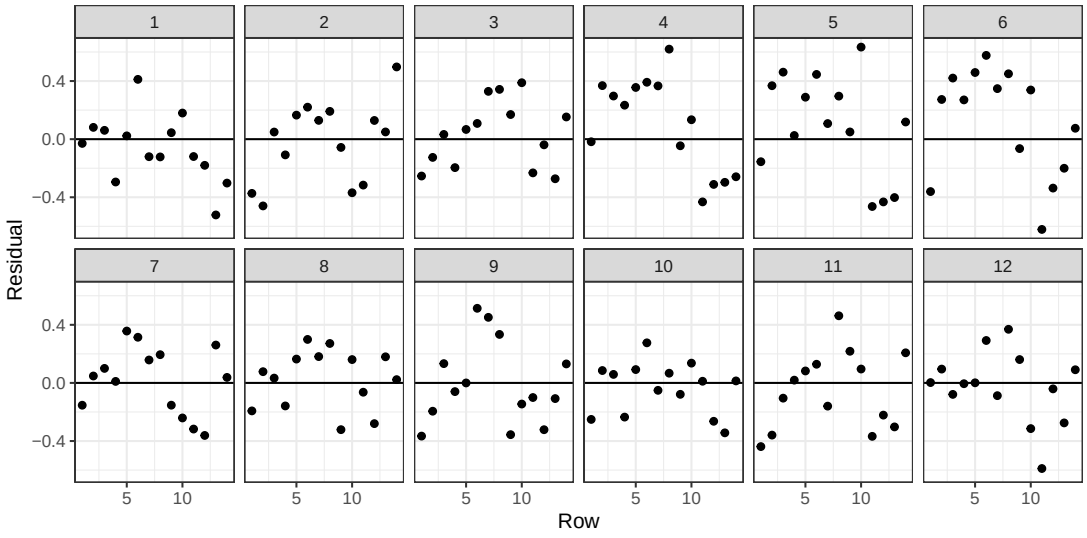

Figure 5.3: OMaB15CUND6 for Model M2. Residuals against row number for each column.

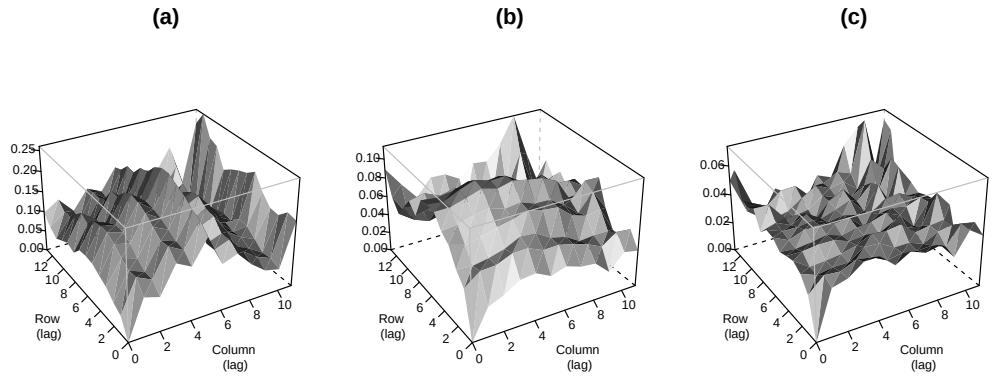

Figure 5.4: OMaB15CUND6: Variogram plots for (a) Model M1: AR1×AR1, (b) Model M2: AR1×AR1 + ran(Column), and (c) Model M3: AR1×AR1 + ran(Column) + ran(Row). x and y ordinates are displacement in the row and column directions respectively, measured as difference in row/column numbers.

Table 5.2: OMaB15CUND6: REML estimates of variance parameters for: Variety ($\hat{\sigma}_g^2$); ColRep ($\hat{\sigma}_b^2$); Column ($\hat{\sigma}_c^2$); and Row ($\hat{\sigma}_r^2$); spatial variance ($\hat{\sigma}^2$); column spatial correlation ($\hat{\rho}_c$); and row spatial correlation ($\hat{\rho}_r$), for Models M1, M2, and M3. B denotes variance parameter is estimated at the boundary value (0 for variance components).

	Genetic	Non-genetic			Error		
Model	$\hat{\sigma}_g^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_r^2$	$\hat{\sigma}^2$	$\hat{\rho}_c$	$\hat{\rho}_r$
M1	0.037	B	-	-	0.196	0.512	0.736
M2	0.033	B	0.063	-	0.094	0.507	0.393
M3	0.029	B	0.073	0.031	0.056	0.156	0.305

5.1.1.1 Empirical best linear unbiased predictions

From the analysis of M3, the variety EBLUPs $\tilde{\boldsymbol{u}}_g$ are obtained (see Equation 2.9). These are given in Table 5.3 for the first and last four varieties in alphabetical order, along with their reliability (r^2) values (see Equation 2.15), these ranged from 0.58 to 0.63 with a mean of 0.62 ($\bar{R}$). Figure 5.5 presents the relationship between $\tilde{\boldsymbol{u}}_g$ and the raw means for individual varieties, where the dashed line represents a 1-1 relationship. The generally low reliabilities results in substantial shrinkage so that the points in Figure 5.5 centre around a line with a slope much less than 1.

The remaining EBLUPs are presented below,

$\tilde{\boldsymbol{u}}_b = \{0, 0, 0\}$

$\tilde{\boldsymbol{u}}_c = \{-0.451, 0.044, 0.005, 0.307, 0.180, 0.463, -0.237, 0.142, -0.069, -0.122, -0.291, 0.029\}$

$\tilde{\boldsymbol{u}}_r = \{-0.203, 0.021, 0.096, -0.065, 0.108, 0.269, 0.064, 0.221, -0.053, 0.080, -0.266, -0.179,$
$-0.160, 0.066\}$

Note that since the estimate of the variance for the replicate block term `ColRep` ($\hat{\sigma}^2_b$) was on the boundary the corresponding EBLUPs ($\tilde{\boldsymbol{u}}_b$) were zero.

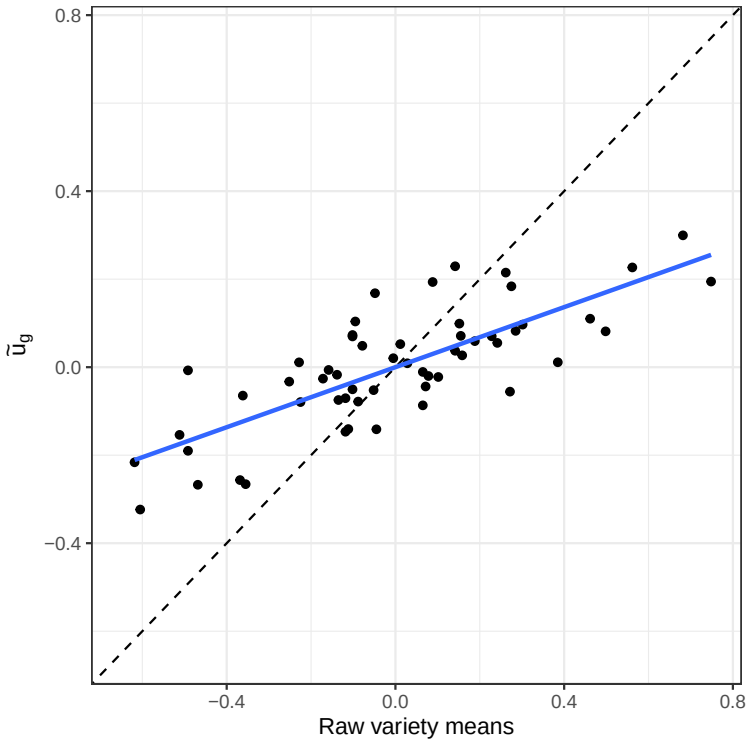

Figure 5.5: OMaB15CUND6: $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ against raw centred varietal means. Dashed line represents a 1-1 line. A regression line through the points is shown by the solid blue line.

Table 5.3: OMaB15CUND6: $\tilde{u}_g$, prediction error variance (PEV), and r^2 estimates for the first and last four varieties in alphabetical order.

Variety	$\tilde{u}_g$	PEV	r^2
Bannister	0.110	0.011	0.622
Carrolup	-0.257	0.011	0.625
Dunnart	0.184	0.011	0.615
Durack	0.081	0.011	0.623
⋮	⋮	⋮	⋮
Wandering	0.215	0.011	0.612
Williams	0.168	0.011	0.611
Wombat	0.070	0.011	0.623
Yallara	-0.044	0.011	0.617

5.2 Spatial analysis of a co-located trial

I now examine the analysis of a co-located trial. As in Smith et al. (2021a), co-located trials are only deemed to comprise a single environment when they are all managed in the same way, that is, they are sown and harvested within a similar time frame and subjected to the same agronomy practices including fertilizer, herbicide, and pathway regimes.

To illustrate this type of analysis, the two S4 trials (OMaA12PINE5 and OMaB12PINE5) sown in Pinery in 2012, with the environment name '12PINE5'. The trials will henceforth be numbered as trials 1 and 2. They were both sown and managed similarly, however their spatial arrangement in relation to each other is unknown. The trait of interest is grain yield which is measured in tonnes per hectare (t/ha), with plot yields ranging from $\sim$ 1.2 to 3.8 t/ha, with no missing plot yields.

Each trial was sown into rectangular arrays with $c = 12$ columns and $r_1 = 13$ and $r_2 = 9$ rows, with $n_1 = 156$ and $n_2 = 108$ ($\Sigma_{j=1}^{2} n_j = 264$) plots respectively. There were $m_1 = 52$ and $m_2 = 36$ varieties with $m = 60$ unique varieties across trials, with 28 varieties in common between trials. Both trials were designed as RCB with $b = 3$ replicates, with replicate blocks aligned with four columns in both trials.

5.2.1 Statistical analysis

Similar to the statistical analysis of a single trial I fit an initial model (M1.c) that comprises terms reflecting the co-location of the trials within the environment and the individual trial designs (Jordan, 2022). The LMM for the (264×1) data vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_{264})$ is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_t}\boldsymbol{u_t} + \boldsymbol{Z_b}\boldsymbol{u_b} + \boldsymbol{e} \tag{5.2}$$

where $\boldsymbol{u_g}$ is the (60×1) vector of random variety effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_t}$ is the (2×1) vector of random trial effects with associated design matrix $\boldsymbol{Z_t}$; $\boldsymbol{u_b}$ is the (6×1) vector of random replicate block effects for each trial with associated design matrix $\boldsymbol{Z_b}$; and $\boldsymbol{e}$ is the (264×1) vector of errors across both trials.

It is assumed that

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{u_t} \\ \boldsymbol{u_b} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_g} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_t} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where for the genetic and non-genetic effects it is assumed that

- $\boldsymbol{G_g} = \sigma_g^2 \boldsymbol{I}_{60}$ (Genetic variance matrix)
- $\boldsymbol{G_t} = \sigma_t^2 \boldsymbol{I}_2$ (Trial variance matrix)
- $\boldsymbol{G_b} = \sigma_b^2 \boldsymbol{I}_6$ (Replicate block variance matrix)

As the spatial arrangement of the two trials in relation to each other is unknown, I follow the 'equal constrained' approach of Jordan (2022), where the spatial variance and spatial correlations for the two trials are constrained to be equal. I let $\boldsymbol{e} = (\boldsymbol{e}_1^\top, \boldsymbol{e}_2^\top)^\top$ and write

$$\begin{aligned} \boldsymbol{\Sigma} = \text{var}(\boldsymbol{e}) = \text{var}\begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix} &= \begin{bmatrix} \sigma_1^2 \boldsymbol{\Sigma}_{c_1}(\rho_{c_1}) \otimes \boldsymbol{\Sigma}_{r_1}(\rho_{r_1}) & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 \boldsymbol{\Sigma}_{c_2}(\rho_{c_2}) \otimes \boldsymbol{\Sigma}_{r_2}(\rho_{r_2}) \end{bmatrix} \\ &= \oplus_{j=1}^{2} \sigma_j^2 \boldsymbol{\Sigma}_{c_j}(\rho_{c_j}) \otimes \boldsymbol{\Sigma}_{r_j}(\rho_{r_j}) \end{aligned}$$

where $\boldsymbol{\Sigma}_{c_1}$ (12×12), $\boldsymbol{\Sigma}_{r_1}$ (13×13), $\boldsymbol{\Sigma}_{c_2}$ (12×12), and $\boldsymbol{\Sigma}_{r_2}$ (9×9) are the correlation matrices for column and rows for both trials. I then have the constraints of: $\sigma_1^2 = \sigma_2^2$;

$\rho_{c_1} = \rho_{c_2}$; and $\rho_{r_1} = \rho_{r_2}$. There are six variance parameters to be estimated for Model M1.c.

The associated `ASReml-R` code is given as

```
M1.c <- asreml(yield ~ 1, random=~ Variety + Trial + Trial:ColRep,
residual =~ dsum(ar1(Column):ar1(Row)|Trial),
data=o12PINE5.df, vcc=Mcc)
```

where `yield` is the data vector of plot yields (t/ha); `1` denotes the overall mean; `Variety, Trial, Trial:ColRep` are terms representing the random variety, trial and within trials replicate blocks effects respectively. The errors are modelled using an AR1×AR1 structure for each trial as given by `dsum(ar1(Column):ar1(Row)|Trial)`. The constraints are implemented within the constraint matrix `Mcc` which is detailed in Table 5.4.

Table 5.4: The `Mcc` matrix of constraints for the analysis of 12PINE5. The second column (V1) defines the grouping of variance parameters by assigning the same number to each parameter within a group, and the third column (V2) contains the scaling coefficient. As an example, the spatial variances for trials OMaA12PINE5 and OMaB12PINE5 are constrained to variance parameter number 3.

`Mcc` matrix			Variance
Term	V1	V2	parameter
Trial_OMaA12PINE5!R	3	1	σ_1^2
Trial_OMaB12PINE5!R	3	1	σ_2^2
Trial_OMaA12PINE5!Column!cor	1	1	ρ_{c_1}
Trial_OMaB12PINE5!Column!cor	1	1	ρ_{c_2}
Trial_OMaA12PINE5!Row!cor	2	1	ρ_{r_1}
Trial_OMaB12PINE5!Row!cor	2	1	ρ_{r_2}

Following the same procedures as shown in Section 5.1, I added both random column (Model M2.c) and row effects (Model M3.c) with significant changes in ℓ_R as shown in Table 5.5. The REML estimates of the variance parameters for the three models (M1.c, M2.c, and M3.c) are given in Table 5.6. In particular this shows the results of the constraints. For example the estimates from M3.c are $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = 0.030$, $\hat{\rho}_{c_1} = \hat{\rho}_{c_2} = 0.103$, and $\hat{\rho}_{r_1} = \hat{\rho}_{r_2} = 0.166$.

Table 5.5: 12PINE5: Summary of models fitted: sources of variation; residual log-likelihood (ℓ_R); number of variance parameters; and likelihood ratio test (REMLRT).

	Source of variation		Variance			
Model	Global/Extraneous[a]	local[b]	parameters	ℓ_R	REMLRT[c]	
M1.c		AR1×AR1	6	192.44		
M2.c	ran(Column)	AR1×AR1	7	200.27	15.66	***
M3.c	ran(Row)	AR1×AR1	8	205.39	10.24	***

[a] ran(Column) and ran(Row) represent the random effects for the column and row factors within trials.
[b] Correlation structures for separable spatial process: AR1 = autoregressive of order 1.
[c] Comparison of model i with model $i-1$; *** $p < 0.001$.

Table 5.6: 12PINE5: REML estimates of variance parameters for: Variety ($\hat{\sigma}^2_g$); Trial ($\hat{\sigma}^2_t$); within trials ColRep ($\hat{\sigma}^2_b$); within trials Column ($\hat{\sigma}^2_c$); and within trials Row ($\hat{\sigma}^2_r$); spatial variance ($\hat{\sigma}^2$); column spatial correlation ($\hat{\rho}_c$); and row spatial correlation ($\hat{\rho}_r$), for Models M1.c, M2.c, and M3.c.

	Genetic	Non-genetic				Error					
Model	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_t$	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_c$	$\hat{\sigma}^2_r$	$\hat{\sigma}^2_1$	$\hat{\rho}_{c_1}$	$\hat{\rho}_{r_1}$	$\hat{\sigma}^2_2$	$\hat{\rho}_{c_2}$	$\hat{\rho}_{r_2}$
M1.c	0.167	0.008	0.009			0.052	0.234	0.437	0.052	0.234	0.437
M2.c	0.158	0.005	0.008	0.014		0.039	0.325	0.148	0.039	0.325	0.148
M3.c	0.157	0.004	0.009	0.015	0.008	0.030	0.103	0.166	0.030	0.103	0.166

5.2.1.1 Empirical best linear unbiased predictions

From the analysis of M3.c, the variety EBLUPs $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ are given in Table 5.7 for the first and last four varieties in alphabetical order, along with their r^2 values (see Equation 2.15), these ranged from 0.91 to 0.95 with $\bar{R} = 0.93$. Figure 5.6 presents the relationship between $\tilde{\boldsymbol{u}}_g$ and the raw means for individual varieties, where the dashed line represents a 1-1 relationship. The high reliabilities (mean of 0.93) results in little shrinkage so that the points in Figure 5.6 centre around a line with a slope close to 1.

Table 5.7: 12PINE5: $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$, prediction error variance (PEV), and r^2 estimates for the first and last four varieties in alphabetical order.

Variety	$\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$	PEV	r^2
Bannister	0.353	0.008	0.951
Carrolup	-0.091	0.008	0.951
Dunnart	0.335	0.008	0.951
Durack	0.136	0.008	0.951
⋮	⋮	⋮	⋮
SV05302-19	0.344	0.012	0.921
SV05305-49	0.080	0.013	0.920
WA01Q265-1	-0.002	0.013	0.920
Wandering	0.478	0.008	0.951

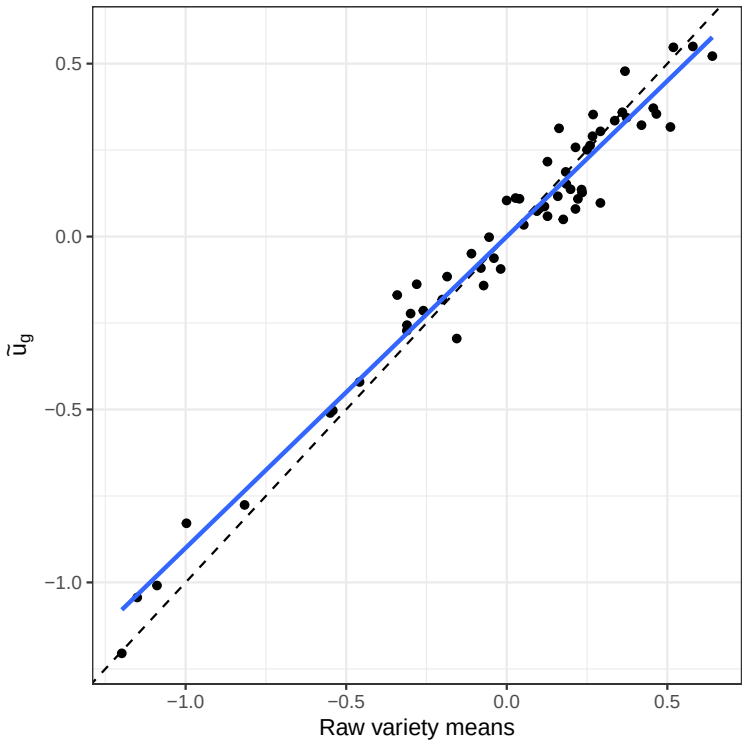

Figure 5.6: 12PINE5: $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ against raw centred varietal means. Dashed line represents a 1-1 line. A regression line through the points is shown by the solid blue line.

5.3 Results from all single environment analyses

I now consider the analyses for all 43 environments in the Oat dataset. The same procedures as shown in the previous sections for the analysis of OMaB15CUND6 and 12PINE5 were used. There were in total 51 outliers identified and removed from the dataset. A summary of all unique models fitted to the non-genetic effects is presented in Table 5.8. `ColRep` is the random model term for replicate block effects; `Column` and `Row` are random model terms for column and row effects respectively; the spatial models are defined as AR1×AR1 for spatial correlation in both directions, ID×AR1 for independent columns, and spatial correlation in the row direction, and ID×ID for independent and identically distributed (IID) errors. Note that for co-located trials, a random `Trial` term is also fitted and `ColRep, Column` and `Row` denote replicate block, column, and row effects within trials. The most common non-genetic model was M1.o with 13 environments.

Table 5.8: Set of unique models fitted to the non-genetic effects for the 43 environments in the Oat dataset. `ColRep` is the random model term for replicate block effects; `Column` and `Row` are random model terms for column and row effects; the spatial models are defined as AR1×AR1 for spatial correlation in both directions, ID×AR1 for spatial in the row direction only, and ID×ID for IID errors.

Model	`ColRep`	`Column`	`Row`	Spatial	Environments
M1.o	✓	-	-	AR1×AR1	13
M2.o	✓	-	-	ID×AR1	6
M3.o	✓	-	✓	AR1×AR1	2
M4.o	✓	-	✓	ID×ID	1
M5.o	✓	-	✓	ID×AR1	6
M6.o	✓	✓	-	AR1×AR1	9
M7.o	✓	✓	✓	AR1×AR1	6

Note that for co-located trials, a random `Trial` term is also fitted and `ColRep, Column` and `Row` denote replicate block, column, and row effects within trials.

The REML estimates for the genetic and non-genetic variance parameters are shown for each environment in Table 5.9. There were two environments OMaA14GRIF2 and OMaA16PING6 which did not have a positive genetic variance, with the variance components estimated at 0 (boundary). The trial reliability ($\bar{R}$) (Equation 2.15) ranged from 0.25 to 0.94 with a median of 0.81 (OMaA14GRIF2 and OMaA16PING6 removed).

When adjusted for the number of varieties, the proportion of maximum potential reliability ($\bar{R}_{X_p}$) (Equation 3.11) ranged from 0.26 to 0.95 with a median of 0.82. The last row of Table 5.9 reports median results for M1.o and the non-genetic effects: the median variance component estimates were $\hat{\sigma}^2_b = 0.019$, $\hat{\sigma}^2 = 0.301$, $\hat{\rho}_c = 0.248$, $\hat{\rho}_r = 0.629$ respectively. The median for the genetic variance was $\hat{\sigma}^2_g = 0.083$.

Summaries of these results are used in the formation of the simulation study described in Chapter 6. In particular, the 10%, 50%, and 90% quantiles of the $\bar{R}_{X_p}$ values of 0.54, 0.82, 0.94 were adopted for the three levels: Low (L), Medium (M), and High (H), in the calibration for the genetic variances. Finally, the non-genetic parameters median values from M1.o (last row of Table 5.9) were used also in the simulation study to simulate data, as well as to create trial designs.

5.4 One-stage multi-environment trial analysis

I now examine the MET analysis of the Oat dataset using the methods described in Chapter 2. This dataset includes $p = 43$ environments and $m = 163$ varieties; however, as demonstrated in the preceding section, the genetic variances for OMaA14GRIF2 and OMaA16PING6 were estimated on the boundary. Therefore these environments were excluded from the MET dataset examined here, resulting in $p = 41$ environments, $t = 47$ trials, $m = 163$ varieties, and $n = 7068$ field plots.

I now let $\boldsymbol{y}_j$ be the $(n_j \times 1)$ vector of yield data for environment $(j = 1, 2, \ldots, 41)$, and let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_{41}^\top)^\top$ be the combined (7068×1) vector of yield data across environments. The LMM for $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{5.3}$$

where $\boldsymbol{\tau}$ is a (41×1) vector of fixed effects which comprise solely of separate environment means, with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is a (6683×1) vector of random VE effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_p}$ is a vector of random non-genetic (peripheral) effects consisting of trial, replicate blocks, column and row effects, as established by the models in the single environment analyses (see Table 5.9), with an accompanying

Table 5.9: REML estimates for genetic and non-genetic variance parameters, and model based estimates of reliability ($\bar{R}$) and proportion of maximum potential reliability ($\bar{R}_{X_p}$) for each environment. B denotes variance parameter is estimated at the boundary value (0 for variance components). The final column represents the non-genetic model chosen (M1.o-M7.o). The final two rows present the median variance parameter estimates across all environments, and the median across only those environments which used the M1 model. Horizontal dashed lines separate years (2012-2016). The two environments illustrated in this chapter for analysis are represented by grey rows.

	Genetic	Non-genetic				Error					
Environment	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_t$	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_c$	$\hat{\sigma}^2_r$	$\hat{\sigma}^2$	$\hat{\rho}_c$	$\hat{\rho}_r$	$\bar{R}$	$\bar{R}_{X_p}$	Model
OMaB12BARK6	0.056		0.169			0.199		0.259	0.481	0.490	M2.o
OMaB12KATA6	0.271		B		0.017	0.361	0.011	0.834	0.880	0.897	M3.o
OMaA12KYBY5	0.210		0.081			0.476	0.118	0.718	0.766	0.781	M1.o
OMaB12MERR6	0.014		B			0.101	0.072	0.479	0.342	0.348	M1.o
12PINE5	0.157	0.004	0.009	0.015	0.008	0.030	0.103	0.166	0.934	0.950	M7.o
12RIVE5	0.353	0.163	0.119	0.297		0.092	0.177	0.109	0.917	0.932	M6.o
OMaA12RUTH3	0.275		0.016			0.068		0.295	0.914	0.932	M2.o
12TURR5	0.210	0.041	B	0.054	0.025	0.115	0.079	0.216	0.846	0.860	M7.o
OMaB12WONG6	0.168		0.047			0.132	0.450	0.673	0.899	0.917	M1.o
OMaB13KATA6	0.219		0.003			0.164	0.015	0.323	0.811	0.825	M1.o
OMaA13KYBY5	0.067		B			0.911	0.683	0.454	0.365	0.371	M1.o
13PINE5	0.107	B	B	0.088	0.008	0.045	0.037	0.287	0.880	0.893	M7.o
13RIVE5	0.314	B	B		0.015	0.198		0.567	0.886	0.900	M5.o
OMaA13RUTH3	0.211		0.008			0.060		0.536	0.928	0.944	M2
OMaA13RYLI6	0.426		0.022	0.090		0.207	0.012	0.273	0.855	0.869	M6.o
13TURR5	0.362	0.043	0.003	0.092		0.119	0.236	0.410	0.922	0.936	M6.o
OMaB13WONG6	0.131		0.096			0.188	0.248	0.624	0.802	0.815	M1.o
OMaA13YANC2	0.187		B	0.442		0.227	0.224	0.113	0.686	0.698	M6.o
OMaA14GRIF2	0.000		0.058			0.776		0.004	0	0	M2.o
OMaB14KATA6	0.021		B	0.481		0.338	0.672	0.420	0.252	0.257	M6.o
OMaA14KYBY5	0.018		0.022		0.016	0.113		0.248	0.326	0.333	M5.o
OMaA14PINE5	0.086		B	0.105		0.097	0.355	0.086	0.741	0.757	M6.o
OMaA14RIVE5	0.123		0.011		0.021	0.311		0.833	0.806	0.823	M5.o
OMaB14RYLI6	0.051		B			0.411	0.197	0.638	0.451	0.461	M1.o
OMaB14WONG6	0.151		0.008		0.010	0.070		0.624	0.898	0.917	M5.o
OMaB15CUND6	0.029		B	0.073	0.031	0.056	0.156	0.305	0.616	0.627	M7.o
OMaA15GRIF2	1.513		0.016		0.082	0.233			0.926	0.943	M4.o
OMaA15KYBY5	0.053		0.018			0.122	0.193	0.629	0.723	0.736	M1.o
OMaB15PING6	0.069		0.471			0.244	0.278	0.645	0.649	0.660	M1.o
OMaA15RIVE5	0.866		0.031	0.071		0.136	0.180	0.183	0.935	0.952	M6.o
OMaB15RYLI6	0.059		0.019			0.723	0.331	0.883	0.646	0.658	M1.o
OMaA15TURR5	0.329		0.033	0.117	0.050	0.109	0.067	0.323	0.888	0.904	M7.o
OMaA15WONG6	0.016		0.026	0.014	0.006	0.022	0.083	0.070	0.640	0.652	M7.o
OMaB16COWR2	0.184		0.022			0.242		0.281	0.708	0.723	M2.o
OMaB16CUND6	0.180		0.089	0.056		0.123	0.226	0.303	0.819	0.837	M6.o
OMaA16KYBY5	0.173		0.203			0.344	0.250	0.483	0.693	0.708	M1.o
OMaA16MURE6	0.094		0.079			0.301	0.238	0.707	0.708	0.723	M1.o
OMaA16PINE5	0.387		B			0.268		0.751	0.906	0.926	M2.o
OMaA16PING6	0.000		0.049	0.464		0.376	0.145	0.169	0	0	M6.o
OMaA16RIVE5	0.308		B		0.031	0.369		0.663	0.815	0.833	M5.o
OMaA16RYLI6	0.083		B			0.524	0.299	0.228	0.366	0.374	M1.o
OMaA16TURR5	0.375		B		0.075	0.378	0.154	0.680	0.848	0.866	M3.o
OMaA16WONG6	0.422		B		0.173	0.388		0.545	0.810	0.828	M5.o
Median	0.168	0.023	0.011	0.090	0.021	0.199	0.187	0.415	0.806	0.823	All
Median	0.083		0.019			0.301	0.248	0.629	0.693	0.708	M1.o

design matrix $\boldsymbol{Z_p}$; and $\boldsymbol{e}$ is the (7068×1) vector of errors. It is assumed that

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{u_p} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_e} \otimes \boldsymbol{I}_{163} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where $\boldsymbol{G_e}$ is a (41×41) symmetric positive (semi)-definite matrix known as the between environments genetic variance matrix, which is modelled here using an FA structure of order k denoted as FAk and given as

$$\boldsymbol{G_e} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$$

where $\boldsymbol{\Lambda}$ is the $(41 \times k)$ matrix of environment loadings; and $\boldsymbol{\Psi}$ is a (41×41) diagonal matrix with elements referred to as the specific environment variances. It is therefore assumed for the VE effects that

$$\boldsymbol{u_g} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_{163})\boldsymbol{f} + \boldsymbol{\delta}$$

where $\boldsymbol{f}$ is the $(163k \times 1)$ vector of variety scores and $\boldsymbol{\delta}$ is the (6683×1) vector of VE lack of fit effects. It is assumed that

$$\text{var}(\boldsymbol{f}) = \boldsymbol{I}_k \otimes \boldsymbol{I}_{163} \tag{5.4}$$

$$\text{var}(\boldsymbol{\delta}) = \boldsymbol{\Psi} \otimes \boldsymbol{I}_{163} \tag{5.5}$$

so that

$$\text{var}(\boldsymbol{u_g}) = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_{163} \tag{5.6}$$

Note that $\boldsymbol{u_g}$ can be further simplified to

$$\boldsymbol{u_g} = \boldsymbol{\beta} + \boldsymbol{\delta} \tag{5.7}$$

where $\boldsymbol{\beta} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_{163})\boldsymbol{f}$ is the so called VE regression component. Therefore, I consider $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ to be the building blocks of $\boldsymbol{u_g}$.

For the non-genetic effects it is assumed that

$$\boldsymbol{G_p} = \oplus_{k=1}^{v} \sigma^2_{p_k} \boldsymbol{I}_{q_k}$$
$$\boldsymbol{\Sigma} = \oplus_{j=1}^{47} \sigma^2_j \begin{cases} \boldsymbol{\Sigma_{c}}_j(\rho_{c_j}) \otimes \boldsymbol{\Sigma_{r}}_j(\rho_{r_j}) \\ \boldsymbol{I_{c}}_j \otimes \boldsymbol{\Sigma_{r}}_j(\rho_{r_j}) \\ \boldsymbol{I_{c}}_j \otimes \boldsymbol{I_{r}}_j \end{cases}$$

where v is the number of components in $\boldsymbol{u_p}$ and q_k is the number of effects in (length of) $\boldsymbol{u}_{\boldsymbol{p}_i}$ and the errors are modelled using one of the spatial structures AR1×AR1, ID×AR1 or ID×ID. The non-genetic and error terms fitted are those established for single environment analyses, as indicated in Table 5.9.

As an example, the `ASReml-R` code to fit the LMM with an FA1 model for $\boldsymbol{u_g}$ (Equation 5.7) is given below

```
rr1.asr <- asreml(yield ~ Environment,
random = ~ rr(Environment, 1):Variety + diag(Environment):Variety +
at(Environment, col.env):Trial +
at(Environment, crep):Trial:ColRep +
at(Environment, rcol):Trial:Column +
at(Environment, rrow):Trial:Row,
residual = ~ dsum(~ar1(Column):ar1(Row)| Trial, levels = aa) +
dsum(~id(Column):ar1(Row)| Trial, levels =ia) +
dsum(~id(Column):id(Row)| Trial, levels =ii),
vcc=Mcc, na.action = na.method(y='include', x='include'),
data=oatsMET.df)
```

where `col.env` is a vector of environment names containing co-located trials; `crep, rcol, rrow` are vectors of environment names containing those fitted with `ColRep, Column, Row` random effects respectively; `aa, ia, ii` are vectors of environment names which have an AR1×AR1, ID×AR1, and ID×ID spatial structures for errors respectively; `Mcc` is the matrix which constrains spatial parameters equal for the co-located environments (see Table 5.10); and `oatsMET.df` is the dataset containing the full data object for the Oat dataset.

The FA1 model for $\boldsymbol{u_g}$ has been fitted by splitting into the two constituent parts, namely the regression part associated with $\boldsymbol{\beta}$ and the lack of fit part associated with $\boldsymbol{\delta}$.

The term `rr(Environment, 1):Variety` relates to $\boldsymbol{\beta}$ and fits a so-called reduced rank variance structure of order 1 for the environment dimension, namely

$$\text{var}(\boldsymbol{\beta}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top} \otimes \boldsymbol{I}_{163}$$

where $\boldsymbol{\Lambda}$ is a (41×1) matrix of loadings. The term `diag(Environment):Variety` relates to $\boldsymbol{\delta}$ and fits a diagonal variance structure for the environment dimension with the variance structure shown in Equation 6.6.

Table 5.10: First 11 rows of the `Mcc` matrix for the MET analysis for the Oat dataset. The second column (V1) defines the grouping of variance parameters by assigning the same number to each parameter within a group, and the third column (V2) contains the scaling coefficient. As an example, the spatial variances for trials OMaA12PINE5 and OMaB12PINE5 are constrained to variance parameter number 83.

Term	V1	V2
Trial_OMaB12BARK6!R	110	1
Trial_OMaB12KATA6!R	111	1
Trial_OMaA12KYBY5!R	89	1
Trial_OMaB12MERR6!R	112	1
Trial_OMaA12PINE5!R	83	1
Trial_OMaB12PINE5!R	83	1
Trial_OMaA12RIVE5!R	84	1
Trial_OMaB12RIVE5!R	84	1
Trial_OMaA12RUTH3!R	90	1
Trial_OMaA12TURR5!R	85	1
Trial_OMaB12TURR5!R	85	1

A series of FALMM were fit to the data with increasing numbers of factors (values of k), as shown in Table 5.11. The Akaike information criteria (AIC) showed significant improvements of successive models up to and including an FA5 model (see also Figure 5.7) and hence the FA5 model was chosen as the final model. The FA5 model variance accounted for (VAF%) by all five factors was 85.3%.

Table 5.11: Summary of number of variance parameters, residual log-likelihood (ℓ_R), and AIC for the seven variance models fitted to the 41 environments in the Oat dataset. Grey row corresponds to the model with the smallest AIC.

Model	Parameters	ℓ_R	AIC
diag	227	1587.06	-2720.11
FA1	268	2050.42	-3564.85
FA2	308	2151.44	-3686.87
FA3	347	2202.80	-3711.60
FA4	385	2253.86	-3737.72
FA5	422	2293.98	-3743.96
FA6	458	2327.86	-3739.72

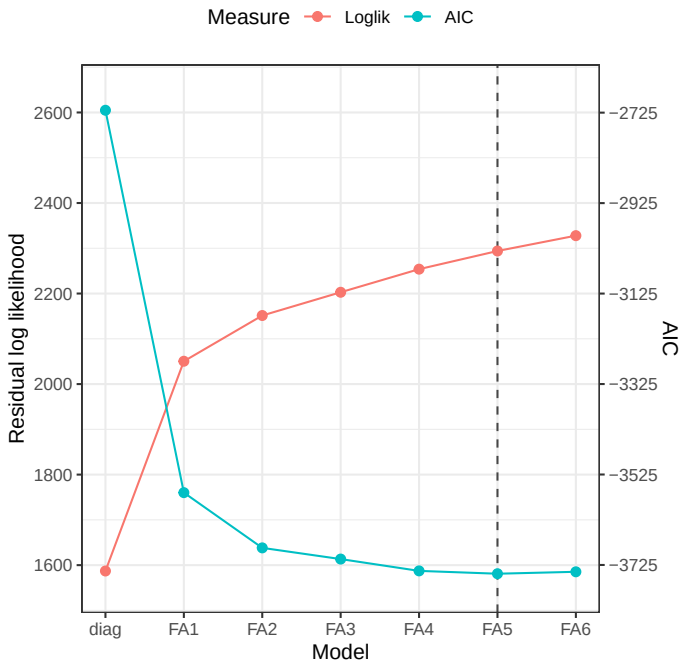

Figure 5.7: Residual log-likelihood (left-hand side y-axis) and AIC (right-hand side y-axis) for each model fitted to the Oat dataset. Colors as referenced in the legend. Dotted vertical line represents the FA5 model as it has the lowest AIC value.

Table 5.12 provides summaries of the environment information from the FA5 model fitted to VE effects: REML estimates of loadings for each factor, specific variances, and percentage variance accounted for (VAF%) by all five factors. On an individual environment basis, all but two environments had greater than 50% explained by the

regression part of the FA model, and 30 environments had greater than 80% explained.

It is noted that the algorithm in `ASReml-R` (Butler et al., 2017) fixes all $k(k-1)/2$ elements in the upper triangle of $\boldsymbol{\Lambda}$ to zero as shown in Table 5.12. This matrix may be rotated as desired for interpretative purposes, where it is usually most meaningful to rotate the estimated loadings to a principal component solution (Smith et al. (2001b, 2015, 2021b) for examples). As the emphasis of this thesis is on the reliability of VE effects, the loadings do not need to be interpreted, and hence the loadings provided in this thesis are unrotated.

The REML estimates of the loadings and specific variances can be used to form the REML estimate of the between environments genetic variance matrix, denoted $\hat{\boldsymbol{G}}_{\boldsymbol{e}}$. In Table 5.13, a subset of this matrix based on the FA5 model is shown for the nine environments in 2012. The diagonal elements represent the genetic variances which are also presented in Table 5.12. The $\hat{\boldsymbol{G}}_{\boldsymbol{e}}$ matrix is transformed to the correlation parametrisation with the resultant $\hat{\boldsymbol{G}}_{\boldsymbol{e}}^{(c)}$ matrix graphically presented in Figure 5.8 by way of a heatmap. The rows and columns of the matrix have been ordered as environments within years. The pairwise between environments genetic correlations ranged between -0.71 and 0.97, with an average of 0.49. The simulation study detailed in Chapter 6 uses summaries of these correlations, specifically the 20%, 50%, and 90% quantiles of 0.2, 0.5, and 0.8 for the three levels of L, M, and H, respectively.

Table 5.12: Summary of environment information from FA5 model fitted to VE effects: REML (unrotated) estimates of loadings for each factor, specific variances ($\hat{\psi}$), genetic variances ($\hat{\sigma}_g^2$), and percentage variance accounted for (VAF%) by all five factors. Horizontal dashed lines separates years (2012-2016).

	Environment loadings							
Environment	1	2	3	4	5	$\hat{\psi}$	$\hat{\sigma}_g^2$	VAF%
OMaB12BARK6	0.210	0	0	0	0	0.005	0.049	90.1
OMaB12KATA6	0.348	0.308	0	0	0	0.022	0.239	90.6
OMaA12KYBY5	0.251	0.228	0.268	0	0	0	0.187	100
OMaB12MERR6	0.001	0.106	-0.046	0.044	0	0	0.016	100
12PINE5	0.243	0.231	0.046	0.077	-0.119	0.016	0.151	89.2
12RIVE5	0.397	0.340	0.167	0.012	-0.131	0	0.318	100
OMaA12RUTH3	0.048	0.144	0.207	-0.107	-0.008	0.177	0.254	30.5
12TURR5	0.289	0.240	-0.055	0.140	-0.104	0.031	0.205	85.1
OMaB12WONG6	0.321	0.155	-0.067	-0.084	-0.111	0	0.151	100
OMaB13KATA6	0.453	0.175	0.145	-0.083	0.092	0	0.272	100
OMaA13KYBY5	0.102	0.080	-0.086	-0.014	0.110	0.030	0.067	54.6
13PINE5	0.222	0.126	0.108	0.125	-0.147	0.002	0.116	98.1
13RIVE5	0.318	0.252	0.315	0.086	-0.170	0.040	0.341	88.2
OMaA13RUTH3	-0.035	0.158	0.315	-0.078	-0.000	0.088	0.219	59.9
OMaA13RYLI6	0.266	0.421	0.219	-0.024	0.311	0.105	0.498	78.9
13TURR5	0.427	0.115	0.131	0.182	-0.245	0.072	0.378	80.8
OMaB13WONG6	0.203	0.191	0.154	-0.154	-0.081	0.008	0.139	94.6
OMaA13YANC2	0.187	0.071	0.258	0.050	-0.101	0.082	0.201	59.4
OMaB14KATA6	0.016	0.027	0.012	-0.100	-0.088	0	0.019	100
OMaA14KYBY5	0.047	-0.071	-0.008	0.034	-0.000	0.008	0.017	50.3
OMaA14PINE5	0.166	0.153	0.156	-0.070	-0.164	0	0.107	100
OMaA14RIVE5	0.316	0.159	0.018	-0.067	-0.180	0	0.162	100
OMaB14RYLI6	0.007	0.053	0.070	-0.222	0.025	0	0.058	100
OMaB14WONG6	0.138	0.030	-0.024	-0.178	-0.102	0.092	0.154	40.5
OMaB15CUND6	0.099	-0.038	0.047	-0.075	-0.145	0	0.040	100
OMaA15GRIF2	0.994	0.566	0.414	0.200	-0.524	0.345	2.139	83.9
OMaA15KYBY5	0.167	-0.035	0.096	0.017	-0.084	0.014	0.060	76.2
OMaB15PING6	0.089	-0.092	0.162	-0.173	-0.080	0.011	0.090	87.8
OMaA15RIVE5	0.574	0.345	0.611	0.299	-0.318	0.086	1.097	92.2
OMaB15RYLI6	0.061	-0.007	0.135	-0.163	-0.052	0.004	0.055	92.8
OMaA15TURR5	0.384	0.209	0.361	0.192	-0.253	0.011	0.433	97.6
OMaA15WONG6	0.048	-0.087	-0.029	-0.078	-0.005	0.001	0.017	97.1
OMaB16COWR2	0.342	0.163	0.038	-0.029	0.157	0.047	0.217	78.4
OMaB16CUND6	0.202	0.169	0.258	-0.023	-0.302	0.038	0.265	85.7
OMaA16KYBY5	0.305	-0.072	0.247	0.167	-0.045	0.016	0.205	92.3
OMaA16MURE6	0.098	0.004	0.109	-0.158	-0.225	0.007	0.105	92.9
OMaA16PINE5	0.565	-0.009	0.267	0.016	-0.112	0.104	0.507	79.5
OMaA16RIVE5	0.405	0.148	0.366	0.082	-0.144	0.114	0.461	75.3
OMaA16RYLI6	0.050	-0.167	0.124	-0.201	0.057	0	0.089	100
OMaA16TURR5	0.593	0.010	0.343	0.121	-0.185	0.059	0.577	89.7
OMaA16WONG6	0.367	0.288	0.414	-0.314	-0.287	0.023	0.593	96.1

Table 5.13: Subset of the REML estimate of the $\boldsymbol{G_e}$ matrix for the 2012 environments only. Diagonal elements and those shown in bold represent the genetic variances ($\hat{\sigma}_g^2$) and the off-diagonal elements the covariances.

Environment	OMaB12BARK6	OMaB12KATA6	OMaA12KYBY5	OMaB12MERR6	12PINE5	12RIVE5	OMaA12RUTH3	12TURR5	OMaB12WONG6
OMaB12BARK6	**0.049**	0.073	0.053	0.000	0.051	0.083	0.010	0.061	0.067
OMaB12KATA6	0.073	**0.239**	0.158	0.033	0.155	0.243	0.061	0.175	0.160
OMaA12KYBY5	0.053	0.158	**0.187**	0.012	0.126	0.222	0.101	0.112	0.098
OMaB12MERR6	0.000	0.033	0.012	**0.016**	0.026	0.029	0.001	0.035	0.016
12PINE5	0.051	0.155	0.126	0.026	**0.151**	0.199	0.047	0.146	0.117
12RIVE5	0.083	0.243	0.222	0.029	0.199	**0.318**	0.103	0.202	0.182
OMaA12RUTH3	0.010	0.061	0.101	0.001	0.047	0.103	**0.254**	0.023	0.034
12TURR5	0.061	0.175	0.112	0.035	0.146	0.202	0.023	**0.205**	0.134
OMaB12WONG6	0.067	0.160	0.098	0.016	0.117	0.182	0.034	0.134	**0.151**

5.4.1 Variety predictions

The EBLUPs of the VE effects for a subset of six varieties and nine environments are shown in Table 5.14. The $\tilde{u_{g_{ij}}}$ and their building block components $\tilde{\beta}_{ij}$ and $\tilde{\delta}_{ij}$ are provided. Note that there is an EBLUP for the common VE effect ($\tilde{\beta}_{ij}$) for a variety for every environment, even if the variety was not grown there. This is because it is associated with the factors (the regression part of the model). In contrast, the EBLUP for the specific VE effect ($\tilde{\delta}_{ij}$) is 0 for a variety when it is not grown in an environment. Note also that $\tilde{\delta}_{ij}$ will be 0 if the VAF% for environment j was 100 (that is, the specific variance was estimated on the boundary). Consequently, in these cases the EBLUP of the common and total VE effects are identical.

There is considerable literature on the interpretation and presentation of FALMM results, such as Smith et al. (2015); Smith & Cullis (2018); Smith et al. (2021b). However, again as the goals of this thesis are to study the reliability of VE effects, the presentation of results are not addressed further here.

Table 5.14: Summary set of results for a subset of six varieties in nine environments from the FA5 MET analysis of the Oat dataset. Tick-marks indicate where the varieties were grown. The EBLUPs of VE effects, that is $\tilde{u_{g}}_{ij}$ and their building block components $\tilde{\beta}_{ij}$ and $\tilde{\delta}_{ij}$.

	Variety	OMaB12BARK6	OMaB12KATA6	OMaA12KYBY5	OMaB12MERR6	12PINE5	12RIVE5	OMaA12RUTH3	12TURR5	OMaB12WONG6
Presence	Bannister	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Carrolup	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Mortlock					✓	✓		✓	
	SV02020-21	✓	✓	✓	✓	✓	✓	✓	✓	✓
	SV05175-26	✓	✓	✓	✓	✓	✓	✓	✓	✓
	SV06071-16									
$\tilde{\beta}_{ij}$	Bannister	0.299	0.544	0.631	-0.085	0.354	0.794	0.429	0.237	0.577
	Carrolup	-0.376	-0.261	-0.066	0.128	-0.184	-0.323	0.108	-0.254	-0.542
	Mortlock	-0.145	-0.462	-0.659	0.007	-0.318	-0.688	-0.455	-0.196	-0.289
	SV02020-21	0.065	0.388	-0.042	0.133	0.095	0.118	-0.061	0.222	0.273
	SV05175-26	0.095	-0.035	-0.019	-0.018	0.216	0.162	-0.173	0.280	0.098
	SV06071-16	-0.140	-0.263	-0.138	-0.047	-0.290	-0.346	0.056	-0.373	-0.251
$\tilde{\delta}_{ij}$	Bannister	-0.011	0.027	0.000^{+}	0.000^{+}	-0.009	0.000^{+}	0.122	-0.063	0.000^{+}
	Carrolup	-0.028	0.053	0.000^{+}	0.000^{+}	0.047	0.000^{+}	0.373	0.115	0.000^{+}
	Mortlock	0	0	0	0	0.070	0.000^{+}	0	-0.038	0
	SV02020-21	0.012	-0.237	0.000^{+}	0.000^{+}	0.089	0.000^{+}	-0.223	0.090	0.000^{+}
	SV05175-26	0.006	-0.037	0.000^{+}	0.000^{+}	0.056	0.000^{+}	-0.092	0.213	0.000^{+}
	SV06071-16	0	0	0	0	0	0	0	0	0
$\tilde{u_{g}}_{ij}$	Bannister	0.288	0.571	0.631	-0.085	0.344	0.794	0.552	0.174	0.577
	Carrolup	-0.403	-0.207	-0.066	0.128	-0.137	-0.323	0.481	-0.139	-0.542
	Mortlock	-0.145	-0.462	-0.659	0.007	-0.248	-0.688	-0.455	-0.234	-0.289
	SV02020-21	0.076	0.151	-0.042	0.133	0.184	0.118	-0.284	0.313	0.273
	SV05175-26	0.101	-0.072	-0.019	-0.018	0.272	0.162	-0.265	0.494	0.098
	SV06071-16	-0.140	-0.263	-0.138	-0.047	-0.290	-0.346	0.056	-0.373	-0.251

$^{+}$ $\tilde{\boldsymbol{\delta}}_{ij} = 0.000$ because VAF%=100 for this environment.

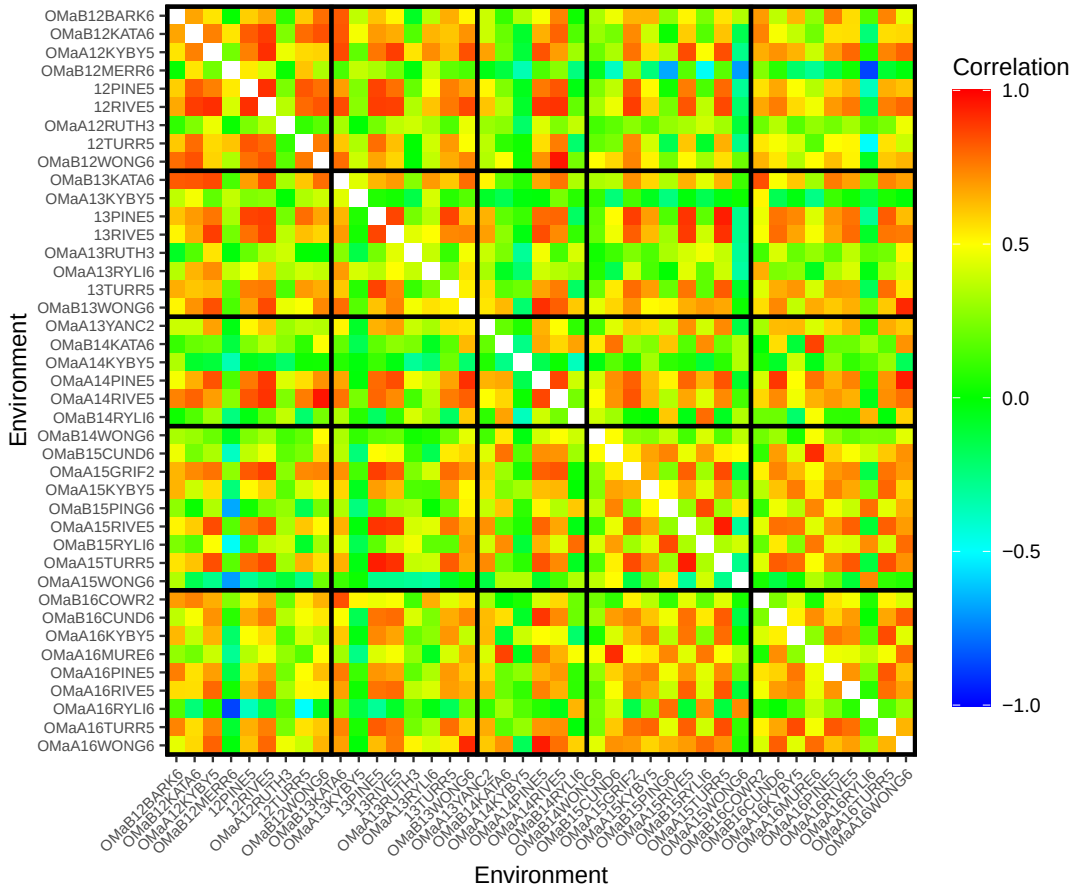

Figure 5.8: Heatmap of $\boldsymbol{G}_e^{(c)}$ from the FA5 model for the MET analysis of the Oat dataset. Key depicts the correlation colour scale. Boundaries for years are indicated by the black lines (2012 - 2016 inclusive from left to right and top to bottom).

5.5 Concluding remarks

This chapter uses the Oat dataset presented in Chapter 4 to demonstrate the statistical methodology described in Chapter 2 with analyses involving independent variety effects. It should be reiterated that this is not the sort of dataset suggested for breeding program variety selections (see Chapter 7), but it resembles typical late-stage evaluation MET datasets, such as those used in the NVT and AHDB RL systems. I first illustrate a single and co-located trial analysis using subsets of this data, and then the MET analysis utilising the entire Oat dataset.

The estimated variance parameters from these analyses are used in the simulation study presented in Chapter 6. Average parameter estimates are used to simulate data and create bespoke trial designs. Furthermore, quantiles for estimated genetic variances and genetic correlations between trials are employed in the development of the three level (low, medium, high) factorial structure. For the three levels of genetic variance, the calibration methods outlined in Chapter 3 are used given levels of $\bar{R}_{X_p} = \{0.54, 0.82, 0.94\}$ and their structural elements of trial size and variance parameter values. For the three levels of genetic correlations between trials the values $\{0.2, 0.5, 0.8\}$ are used.

Chapter 6

The effect of variety connectivity on the reliability of varietal predictions from a factor analytic multi-environment trial analysis

The techniques of variety selection entail a cyclic process of variety evaluation revolving: re-evaluating the best, removing the poor, and including new varieties as part of the breeding objectives. As a result, between environments and years, a necessary and frequently complex selection history of varieties is observed. This leads to datasets with varying numbers of common varieties between pairs of environments, a measure known as "variety connectivity". Due to the rapid turnover of varieties in the early stages of breeding, the tendency for poor variety connectivity is particularly observed, as demonstrated by the datasets in Chapter 4, and in particular the Durum dataset consisting of multiple breeding stages.

A 1-stage FALMM (Smith et al., 2001b) to model VE effects is regarded as the gold standard (Gogel et al., 2018) method of analysis for MET data. The reader is directed to Chapter 2, which discusses the statistical approach for these models, as well as Chapters 5 and 8, which illustrates with the analysis of the Oat and Durum datasets, respectively.

Historically, the majority of MET analyses did not include information on genetic relat-

edness, so that variety effects were assumed to be independent. Within this framework, it was assumed that variety connectivity was a significant driver of the reliability of genetic variance parameter estimates in a MET analysis, which affected the reliability of predicted VE effects (Smith et al., 2001a, 2015; Ward et al., 2019). To address these concerns, problematic environments were frequently omitted from MET datasets if they shared too few varieties with other environments. However, there is no scientific evidence supporting this strategy. Therefore in this chapter, a simulation study with a range of treatments is used to investigate the effect of variety connectivity on the reliability of genetic variance parameter estimates in a MET analysis and also on the reliability of VE predictions.

Without loss of generality, and in order to focus purely on variety connectivity between individual pairs of environments, the study involved two trials (environments) only. The number of varieties in common between the two trials is then varied across a range of genetic scenarios. To keep the simulation studies anchored in real-world circumstances, trial structural elements and variance component values from the Oat dataset (Chapters 4 and 5) were used to simulate datasets. This dataset is used to resemble standard late-stage variety testing procedures used across the world with analyses involving independent variety effects.

This chapter is arranged as follows: the treatments assessed in the simulation study are described in Section 6.1; in Section 6.2 the trial plot structure and trial design procedures are described; in Section 6.3 I discuss the methods to simulate the MET datasets; then I describe the statistical procedures for the analysis of the simulated datasets in Section 6.4; I then present in Section 6.5 the results from the simulation study; and finally in Section 6.6 I have concluding remarks.

6.1 Treatments assessed in simulation study

The simulation is based on datasets with two trials, labelled as $Env1$ and $Env2$. The treatments considered involve the factorial combinations of two trial sizes, three levels of genetic variance, three levels of genetic correlation between trials, and varying levels of variety connectivity. For the three level treatments, these are classified into low (L),

medium (M), and high (H) levels, based on the range of values reported in Chapter 5 from the analysis of the Oat dataset. For each of these combinations a range of variety connectivities are investigated. The resulting components of the treatment structure are fully described in the following sections.

6.1.1 Trial size

Two trial sizes ($Tsize$) consisting of $m = 24$ or 48 varieties with three replicates ($b = 3$) each are considered. These are referred to as T24 and T48 respectively. As a result the trials have $n = 72$ or 144 plots respectively. Table 6.1 presents the summary of the two $Tsize$ treatments. In particular, the size of m were chosen to represent the range of trial sizes seen in the Oat dataset described in Chapter 4.

Table 6.1: Summaries of the two trial size ($Tsize$) scenarios.

Tsize	Varieties (m)	Columns	Rows	Plots (n)
T24	24	12	6	72
T48	48	12	12	144

6.1.2 Levels of genetic variance

The genetic variance levels for each trial ($\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ for $Env1$ and $Env2$, respectively) were calculated using the calibration procedure explained in Chapter 3. The aim was to choose genetic variances that reflected a range of reliabilities observed in the analyses of the Oat dataset (Chapter 4). In order to allow for the unequal numbers of varieties in the Oat trials, I used the proportion of maximum potential reliability ($\bar{R}_{X_p}$, see Chapter 3, Equation 3.12) as the measure to summarise across those trials. The resultant 10%, 50% and 90% quantiles of $\bar{R}_{X_p}$ were $\{0.54, 0.82, 0.94\}$. Using the profile curves in Figure 3.2 of Chapter 3, these mapped to genetic variances of $\{0.2, 0.75, 2.5\}$ which were then used as the L, M and H values for data generation in the simulation. In the simulated study, I maintain the M level of genetic variance for $Env1$ ($\sigma^2_{g_1} = 0.75$), but vary the genetic variance for $Env2$ ($\sigma^2_{g_2} = \{0.2, 0.75, 2.5\}$).

6.1.3 Levels of genetic correlation between trials

The three levels of genetic correlation between trials (ρ_{12}) are taken as the 20%, 50%, and 80% rounded quantile range values of $\{0.2, 0.5, 0.8\}$ respectively from the analysis of the Oat dataset given in Chapter 5 (see Section 5.4). These are translated to the between trials covariance (σ_{12}) given values of $\{\sigma^2_{g_1}, \sigma^2_{g_2}\}$ as

$$\sigma_{12} = \rho_{12}\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}} \tag{6.1}$$

6.1.4 Levels of variety connectivity

The levels of variety connectivity between $Env1$ and $Env2$ ($x_{1,2}$) were varied from one variety in common between $Env1$ and $Env2$, through to all varieties in common ($1 : m$). As shown in Table 6.2, not all levels of $x_{1,2}$ were tested and instead increments between 1 and m were chosen to be proportional to the size of $x_{1,2}$, with greater increments for the larger scenarios. As a result, the T24 and T48 scenarios feature 13 and 19 treatment levels of variety connectivity, respectively.

Table 6.2: Levels and increments of variety connectivity ($x_{1,2}$).

$x_{1,2}$	Increment	Levels
1	1	$\{1\}$
2:24	2	$\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24\}$
28:48	4	$\{28, 32, 36, 40, 44, 48\}$

As an example, the T24 scenario has treatment levels of $x_{1,2} = \{1, 2, 4, \ldots, 22, 24\}$, with the maximum total number of varieties across both $Env1$ and $Env2$ being 47, which occurs for the scenario when there is only 1 variety in common ($x_{1,2} = 1$). I label these varieties as V1 through V47. The same 24 varieties (V1 - V24) are held in $Env1$ across all scenarios, and vary the number of common varieties in $Env2$. Examples of the varieties present in $Env2$ include: all 24 varieties are in common ($x_{1,2} = 24$) where I have varieties V1 - V24 in $Env2$; if I then consider 22 varieties in common ($x_{1,2} = 22$) I have varieties V3 - V26 in $Env2$ (V25, and V26 are only present in $Env2$); finally if I then consider the most extreme case with only 1 variety in common ($x_{1,2} = 1$), I have varieties V24 - V47 present in $Env2$ (only V24 is in common with $Env1$).

6.1.5 Overall treatment structure of simulation study

The treatment structure employed in the simulation study is summarised in Table 6.3. There are 288 scenarios, that reflect the factorial combination of two trial sizes, three levels of genetic variance, three levels of genetic correlation between trials, and either 13 or 19 levels of variety connectivity.

Table 6.4 lists the genetic parameter values for each of the nine genetic scenarios. Each of the scenarios has been assigned a three-character acronym, indicating the genetic variance for $Env1$ ($\sigma^2_{g_1} = \{0.75\}$), the level of genetic variance for $Env2$ ($\sigma^2_{g_2} = \{0.20, 0.75, 2.50\}$), and the level of genetic correlation between the two trials ($\rho_{12} = \{0.2, 0.5, 0.8\}$). For example, the acronym 'MLL' represents medium $\sigma^2_{g_1} = 0.75$, low $\sigma^2_{g_2} = 0.20$, and low $\rho_{12} = 0.2$. It is worth noting once more that $\sigma^2_{g_1} = 0.75$ (the medium level) for each of the nine genetic scenarios.

Table 6.3: Overview of factorial structure in simulation study.

	Number of levels			Total
Tsize	Genetic variance	Genetic correlation	Variety connectivity	Scenarios
T24	3	3	13	117
T48	3	3	19	171
Total				288

Table 6.4: Structure and parameter values for the nine genetic scenarios. The three-character acronym indicates the level of genetic variance for $Env1$ ($\sigma^2_{g_1}$), the level of genetic variance for $Env2$ ($\sigma^2_{g_2}$), and the level of genetic correlation between the two trials (ρ_{12}). Note, that the covariance parameter σ_{12} has been translated using Equation 6.1.

Scenario	$\sigma^2_{g_1}$	$\sigma^2_{g_2}$	σ_{12}	ρ_{12}
MLL	0.75	0.20	0.08	0.2
MLM	0.75	0.20	0.19	0.5
MLH	0.75	0.20	0.31	0.8
MML	0.75	0.75	0.15	0.2
MMM	0.75	0.75	0.38	0.5
MMH	0.75	0.75	0.60	0.8
MHL	0.75	2.50	0.27	0.2
MHM	0.75	2.50	0.68	0.5
MHH	0.75	2.50	1.10	0.8

6.2 Trial plot structure and trial design used in simulation study

The trial structure for $Env1$ and $Env2$ were created to reflect trials observed in the Oat dataset described in Chapter 4, that is, in terms of: the number of varieties; replication; number of rows and columns; and replicate block alignment. The framework and design of trials are fully discussed in the sections that follow.

6.2.1 Trial plot structure

Recall from 6.1.1 that I have two trial sizes, T24 and T48, with $m = 24$ or 48 varieties, $b = 3$ replicates and hence $n = 72$ or 144 plots, as illustrated in Table 6.1. Because of the levels of $x_{1,2}$ I define M to be the maximum number of varieties across $Env1$ and $Env2$, which corresponds to $x_{1,2} = 1$. Thus, I have $M = 47$ for T24 and $M = 95$ for T48. Each trial is arranged in rectangular arrays with 12 columns and either 6 (T24) or 12 (T48) rows, with replicate blocks aligned with four columns.

6.2.2 Trial design

Trial designs were created using the model-based designs software `ODW` (Butler, 2013) within the `R` statistical computing environment (R Core Team, 2020). These designs comprised random Variety and ColRep effects, and a AR1×AR1 structure for the errors. The median parameter values from Chapter 5 (see Table 5.9) were used for the parameter values. An example of the `ODW` code is given by

```
design.base <- odw(fixed=~ 1, random=~ Variety + ColRep,
residual=~ ar1(Column):ar1(Row),
permute=~ Variety, swap=~ ColRep,
search="tabu+rw",
G.param = sv, R.param = sv, maxit=100,
data=T24.trials )
```

where `permute` defines the model term to permute, which here is `Variety`; `swap` defines where legal treatment exchanges are allowed, which for this example are within `ColRep`; `sv` are the values of pre-specified variance parameters, which is detailed in Table 6.5; `search="tabu+rw"` specifies the search strategy uses a TABU search (Taillard, 1991)

with random walk; and `maxit=100` specifies that 100 TABU loops are completed. The errors are modelled using an AR1×AR1 structure as given by `ar1(Column):ar1(Row)`.

Table 6.5: The `sv` object contains the starting values for the `ODW` trial design code for the simulation study. This example is for $Env1$, however the same parameter values are also used for $Env2$.

Component	Value	Parameter
Variety	0.75	$\sigma^2_{g_1}$
ColRep	0.10	$\sigma^2_{b_1}$
Column:Row!R	1.00	σ^2_1
Column:Row!Column!cor	0.20	ρ_{c_1}
Column:Row!Row!cor	0.60	ρ_{r_1}

Separate trial designs were created for each trial ($Env1$ and $Env2$) within the treatment framework of $Tsize$ and connectivity level, corresponding to 64 trial designs in total.

6.3 Methods to simulate MET datasets

This section provides the techniques used to simulate the data vector, which corresponds to a MET dataset with $p = 2$ trials. I use the framework of the treatment and plot structure provided in Sections 6.1 and 6.2.

6.3.1 Simulation of the MET datasets

The LMM for data generation for the combined across trials $(2n \times 1)$ data vector $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top)^\top$ may be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_b}\boldsymbol{u_b} + \boldsymbol{e} \tag{6.2}$$

where the fixed effects $\boldsymbol{\tau}$ comprise a mean for each environment; $\boldsymbol{u_g}$ is the $(2M \times 1)$ vector of random VE effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_b}$ is the (6×1) vector of replicate block effects for each environment, with associated design matrix $\boldsymbol{Z_b}$; and

$\boldsymbol{e}$ is the combined $(2n \times 1)$ vector of errors across trials. It is assumed that

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{u_b} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_e} \otimes \boldsymbol{I}_M & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where the structures of $\{\boldsymbol{G_e}, \boldsymbol{G_b}, \boldsymbol{\Sigma}\}$ are described in the following sections. Under these assumptions, the distribution of $\boldsymbol{y}$ is Gaussian with mean $\boldsymbol{X\tau}$ and variance

$$\text{var}(\boldsymbol{y}) = \boldsymbol{H} = \boldsymbol{Z_g}(\boldsymbol{G_e} \otimes \boldsymbol{I}_M)\boldsymbol{Z_g}^\top + \boldsymbol{Z_b}\boldsymbol{G_b}\boldsymbol{Z_b}^\top + \boldsymbol{\Sigma} \tag{6.3}$$

Without loss of generality, for data simulation it is assumed that $\tau_1 = \tau_2 = 0$.

6.3.1.1 Simulated variety by trial genetic effects

Given that the simulation study includes two trials, an unstructured matrix would be a reasonable model for the VE effects. However, the FALMM of Smith et al. (2001b) is widely used in many plant breeding programs and is the focus of this thesis. Therefore, an FA structure (see Section 5.4) of order 1 denoted as FA1 was used to simulate $\boldsymbol{u_g}$. It is assumed that

$$\boldsymbol{G_e} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \tag{6.4}$$

where $\boldsymbol{\Lambda}$ is the (2×1) matrix of environment loadings with elements $\{\lambda_{11}, \lambda_{12}\}$; and $\boldsymbol{\Psi}$ is a (2×2) diagonal matrix with elements $\{\psi_1, \psi_2\}$ referred to as the specific environment variances. It is therefore assumed for the VE effects that

$$\boldsymbol{u_g} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_M)\boldsymbol{f} + \boldsymbol{\delta}$$

where $\boldsymbol{f}$ is the $(M \times 1)$ vector of variety scores and $\boldsymbol{\delta}$ is the $(2M \times 1)$ vector of VE lack of fit effects. It is assumed that

$$\text{var}(\boldsymbol{f}) = \boldsymbol{I}_M \tag{6.5}$$

$$\text{var}(\boldsymbol{\delta}) = \boldsymbol{\Psi} \otimes \boldsymbol{I}_M \tag{6.6}$$

so that

$$\text{var}(\boldsymbol{u_g}) = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_M \tag{6.7}$$

As there are two trials, a constraint of $\psi = \psi_1 = \psi_2$ is employed to prevent the model from being over-parametrised. The values for $\{\lambda_{11}, \lambda_{12}, \psi\}$ for data generation using this model are obtained from the parameters $\{\sigma^2_{g_1}, \sigma^2_{g_2}, \sigma_{12}\}$ as follows

$$\boldsymbol{G_e} = \begin{bmatrix} \lambda_{11}^2 + \psi & \lambda_{11}\lambda_{12} \\ \lambda_{12}\lambda_{11} & \lambda_{12}^2 + \psi \end{bmatrix} = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{12} \\ \sigma_{12} & \sigma^2_{g_2} \end{bmatrix}$$

Thus, solving for λ_{11}, λ_{12}, and ψ the solution is

$$\begin{aligned} \lambda_{11} &= \left(\frac{-(\sigma^2_{g_2} - \sigma^2_{g_1}) + \sqrt{(\sigma^2_{g_2} - \sigma^2_{g_1})^2 + 4\sigma^2_{12}}}{2} \right)^{\frac{1}{2}} \\ \lambda_{12} &= \frac{\sigma_{12}}{\lambda_{11}} \\ \psi &= \sigma^2_{g_1} - \lambda^2_{11} \end{aligned}$$

The parameter values as given in Table 6.4 for $\{\sigma^2_{g_1}, \sigma^2_{g_2}, \sigma_{12}\}$ are re-parametrised to their respective FA1 parameter form as given in Table 6.6 for each scenario.

Table 6.6: True FA1 parameter values of the loadings for *Env*1 (λ_{11}), and *Env*2 (λ_{12}), and the combined across trial specific variance (ψ) for each of the nine genetic scenarios.

Scenario	λ_{11}	λ_{12}	ψ
MLL	0.103	0.749	0.189
MLM	0.248	0.782	0.139
MLH	0.373	0.830	0.061
MML	0.387	0.387	0.600
MMM	0.612	0.612	0.375
MMH	0.775	0.775	0.150
MHL	1.339	0.205	0.708
MHM	1.409	0.486	0.514
MHH	1.509	0.726	0.223

6.3.1.2 Simulated non-genetic effects

The replicate block effects ($\boldsymbol{u_b}$), and errors ($\boldsymbol{e}$) are simulated given the variance matrices

- $\boldsymbol{G_b} = \oplus_{j=1}^{2} \sigma^2_{b_j} \boldsymbol{I}_3$ (replicate block variance matrix)
- $\boldsymbol{\Sigma} = \oplus_{j=1}^{2} \sigma^2_j \boldsymbol{\Sigma_{c}}_j(\rho_{c_j}) \otimes \boldsymbol{\Sigma_{r}}_j(\rho_{r_j})$ (error variance matrix)

where $\boldsymbol{G_b}$ is assumed to have a block diagonal structure with variances $\sigma^2_{b_j}$ ($j = 1, 2$ for $Env1$, $Env2$), and $\boldsymbol{\Sigma}$ is assume to have a separable autoregressive process of order 1 (denoted AR1×AR1), where $\boldsymbol{\Sigma}_{c_j}$ and $\boldsymbol{\Sigma}_{r_j}$ are correlation matrices, containing parameters ρ_{c_j} and ρ_{r_j} for the column and row dimensions respectively, with spatial variances σ^2_j.

The parameter variance values $\{\sigma^2_{b_1}, \sigma^2_{b_2}, \sigma^2_1, \sigma^2_2, \rho_{c_1}, \rho_{c_2}, \rho_{r_1}, \rho_{r_2}\}$ for each simulation dataset were derived from the average variance parameters from the analysis of the Oat dataset given in Chapter 4. I consider $\sigma^2_{b_1} = \sigma^2_{b_2} = 0.1$; $\sigma^2_1 = \sigma^2_2 = 1$; $\rho_{c_1} = \rho_{c_2} = 0.2$; $\rho_{r_1} = \rho_{r_2} = 0.6$.

6.3.2 Number of simulations

Using the methodologies described in Section 3.4.3, the T24 MLL scenario was utilised to determine how many simulations are required to achieve an acceptable error of 1% ($E = 0.01$) level of accuracy for the reliabilities of VE effect predictions with 95% confidence. This scenario was chosen because it was projected to be the most unpredictable, resulting in the worst case scenario. The findings indicated that $S = 3000$ simulations were required to attain $E = 0.01$ accuracy, with an additional set of 200 simulations to accommodate for convergence problems. As a consequence, in all scenarios, $S = 3200$ simulations were conducted but trimmed to 3000 valid simulations. Therefore, a total of 921600 datasets were simulated across all treatments.

6.4 Statistical procedures used in simulation study

This section describes the methodologies and procedures employed in the statistical analysis of the simulated MET datasets and the corresponding capture of results. The simulation study was completed within the `R` statistical computing environment (R Core Team, 2020) and analyses were completed using `ASReml-R` (Butler et al., 2017). The framework for these procedures is discussed in the sections that follow.

6.4.1 Steps used in simulation study

In the simulation study, the steps for simulation $s = \{1, \ldots, S = 3200\}$ within a *Tsize* {T24, T48} are as follows

1. Generate the random VE effects $\boldsymbol{u_g}$, replicate block effects $\boldsymbol{u_b}$, and errors $\boldsymbol{e}$ as per the LMM in Equation 6.2 and for the pre-specified variance parameters as described in Sections 6.3.1.1 and 6.3.1.2. In terms of the fixed effects, without loss of generality I choose $\tau_1 = \tau_2 = 0$. As noted previously, I have generated $2M$ VE effects, which correspond to the maximum total number of varieties across all connectivity levels. I denote the resultant vectors for simulation s by $\boldsymbol{u_{g}}_s$, $\boldsymbol{u_{b}}_s$, and $\boldsymbol{e}_s$.

2. For the connectivity level $x_{1,2}$, I subset the appropriate elements of $\boldsymbol{u_{g}}_s$, which I label as $\boldsymbol{u_{g}}_{s_c}$, where c corresponds to the levels of $x_{1,2}$ (see Table 6.2). I then form the $\boldsymbol{y}$ data vector as per Equation 6.2 and then fit the LMM in Equation 6.2. I then save the REML estimates of the FA variance parameters, denoting as $\{\hat{\lambda}_{11_{sc}}, \hat{\lambda}_{12_{sc}}, \hat{\psi}_{sc}\}$ and then convert back to the corresponding genetic variance parameters denoted as $\{\hat{\sigma}^2_{g_{1sc}}, \hat{\sigma}^2_{g_{2sc}}, \hat{\sigma}_{12_{sc}}\}$; the REML estimates for the non-genetic variance parameters which is denoted as $\{\hat{\sigma}^2_{b_{1sc}}, \hat{\sigma}^2_{b_{2tc}}, \hat{\sigma}^2_{1_{sc}}, \hat{\sigma}^2_{2_{sc}}, \hat{\rho}_{c_{1sc}}, \hat{\rho}_{c_{2sc}}, \hat{\rho}_{r_{2sc}}, \hat{\rho}_{r_{2sc}}\}$. I then save the EBLUPs of the VE effects, denoted as $\tilde{\boldsymbol{u}}_{\boldsymbol{g}sc}$. Additionally, I save model convergence parameters denoted as $\{i_{sc}, u_{sc}\}$ to represent the number of iterations and updates required for model convergence.

3. Repeat step 2. for each level of $x_{1,2}$.

I focus on the set of results from $Env1$ since this contains the same varieties across all $x_{1,2}$ levels, thus allowing a fair comparison across scenarios.

6.4.2 Statistical analyses

In the analysis of the MET datasets as described in step 2, I fit a FALMM of order $k = 1$ for $\boldsymbol{u_g}$ (Equation 5.7). As an example, the `ASReml-R` code and updating procedure is given below.

```
FA1.asr <- asreml(y ~ Trial,
random=~ rr(Trial,1):Variety + Trial:Variety +
at(Trial):ColRep,
residual=~ dsum(~ar1(Column):ar1(Row) | Trial),
data=MET.df,R.param = sv, G.param = sv)

#update - up to 10 times.
for(z in 1:10){
if(!FA1.asr$converge |
max(summary(FA1.asr)$varcomp$'%ch',na.rm=T) > 1)
{
FA1.asr <- update.asreml(FA1.asr)
if(is.null(dim(FA1.asr$trace)[2]))
                {iters <- iters+1}
if(!is.null(dim(FA1.asr$trace)[2]))
                {iters <- iters+dim(FA1.asr$trace)[2]}
up.no <- up.no+1
}
```

Note that the FA1 model for $\boldsymbol{u_g}$ has been fitted by splitting into the two constituent parts, namely the regression part associated with $\boldsymbol{\beta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ and the lack of fit part associated with $\boldsymbol{\delta}$ (also see Chapter 5). The term `rr(Trial, 1):Variety` relates to $\boldsymbol{\beta}$ and fits a so-called reduced rank variance structure of order 1 for the environment dimension, and `Trial:Variety` relates to $\boldsymbol{\delta}$. `sv` is the starting values for the residual and random parameters (see Table 6.7 for an example); `update.asreml` provides the updating of the `FA1.asr` ASReml object using the last set of REML variance parameter estimates. `iters` and `up.no` are objects which count how many iterations and updates were completed.

Table 6.7: Example starting values (sv) for `ASReml-R` model for MML scenario.

Component	Value	Constraint*	Parameter
`rr(Trial, 1):Variety!Trial1!fa1`	0.103	U	λ_{11}
`rr(Trial, 1):Variety!Trial2!fa1`	0.749	U	λ_{12}
`Trial:Variety`	0.189	P	ψ
`at(Trial, Trial1):ColRep`	0.100	P	$\sigma^2_{b_1}$
`at(Trial, Trial2):ColRep`	0.100	P	$\sigma^2_{b_2}$
`Trial_Trial1!R`	1	P	σ^2_1
`Trial_Trial1!Column!cor`	0.200	U	ρ_{c_1}
`Trial_Trial1!Row!cor`	0.600	U	ρ_{r_1}
`Trial_Trial2!R`	1	P	σ^2_2
`Trial_Trial2!Column!cor`	0.200	U	ρ_{c_2}
`Trial_Trial2!Row!cor`	0.600	U	ρ_{r_2}

*The parameter constraints are U (unconstrained) and P (positive).

6.4.3 Reliability of EBLUPs

I calculate two main performance measures or metrics in relation to the varying levels of $x_{1,2}$, namely a measure of the genetic variance parameter estimates and a measure of the reliability of the predicted variety effects for $Env1$. For each level of $x_{1,2}$, the reliability for variety $k = \{1 \dots m\}$ in $Env1$ was computed as the square of the sample correlation between the true (simulated) effects (element of $\boldsymbol{u}_{\boldsymbol{g}_{sc}}$ for the variety and in $Env1$) and the EBLUP (element of $\tilde{\boldsymbol{u}}_{\boldsymbol{g}_{sc}}$ for the variety and in $Env1$). This is denoted as R^S_{kc} (see Equation 3.5), and when averaged across varieties it is denoted as $\bar{R}^S_c$ (see Equation 3.6).

6.5 Results of simulation study

The reliability-based metrics based on VE effects are the main focus in this thesis. These metrics are interpreted such that better reliability provides superior variety selection performance. These are compared to design-based values (see Chapter 3), with the disparities showing the loss of reliability associated with estimation of the variance parameters.

This section is structured in the following way: model convergence is examined in Section 6.5.1; variance component summaries including bias and MSE are given in Section 6.5.2; reliability based values for the predicted genetic VE effects are presented in Section 6.5.3; and then in Section 6.5.4 I examine the associated loss in the reliability of the predicted VE effects.

6.5.1 Model convergence

Model convergence is determined within `ASReml-R` by inspection of the REML log-likelihood, which defines convergence as "REML log-likelihood changes less than 0.002*current iteration number" Butler et al. (2017). I employ the same strategy to find convergence using the above definition as well as individual variance parameter estimates changing by less than 1%, as described in Chapter 3.

Based on past experience analysing MET datasets, model convergence issues are commonly observed, which is thought to be attributed to the amount of information, or lack thereof to estimate the unknown variance parameters. To achieve REML log-likelihood and variance parameter estimate convergence, model updates (`asreml.update`) are frequently necessary (Butler et al., 2017). In the majority of cases, the extra iterations obtained using `asreml.update` leads to convergence. However, failure to converge is noticed in some extreme conditions. The effect of variety connectivity on `ASReml-R` model convergence is investigated in this section.

The number of model convergence problems for each scenario in the simulation study ranged from 0 to 61 out of 3200 model fits. This is demonstrated to be far less than the 200 additional simulations specified as buffers. The number of model convergence problems for each scenario is depicted in Figure 6.1. In addition, there were two singularity problems, both for the MMM scenario. A total of 1343 model fits failed out of a total of 921600 model fits (0.15%).

Figure 6.1 shows non-linear decreasing responses for the average number of iterations, and convergence issues across the range of $x_{1,2}$ values. These patterns alone suggest that variety connectivity influences the reliability of variance parameter estimates.

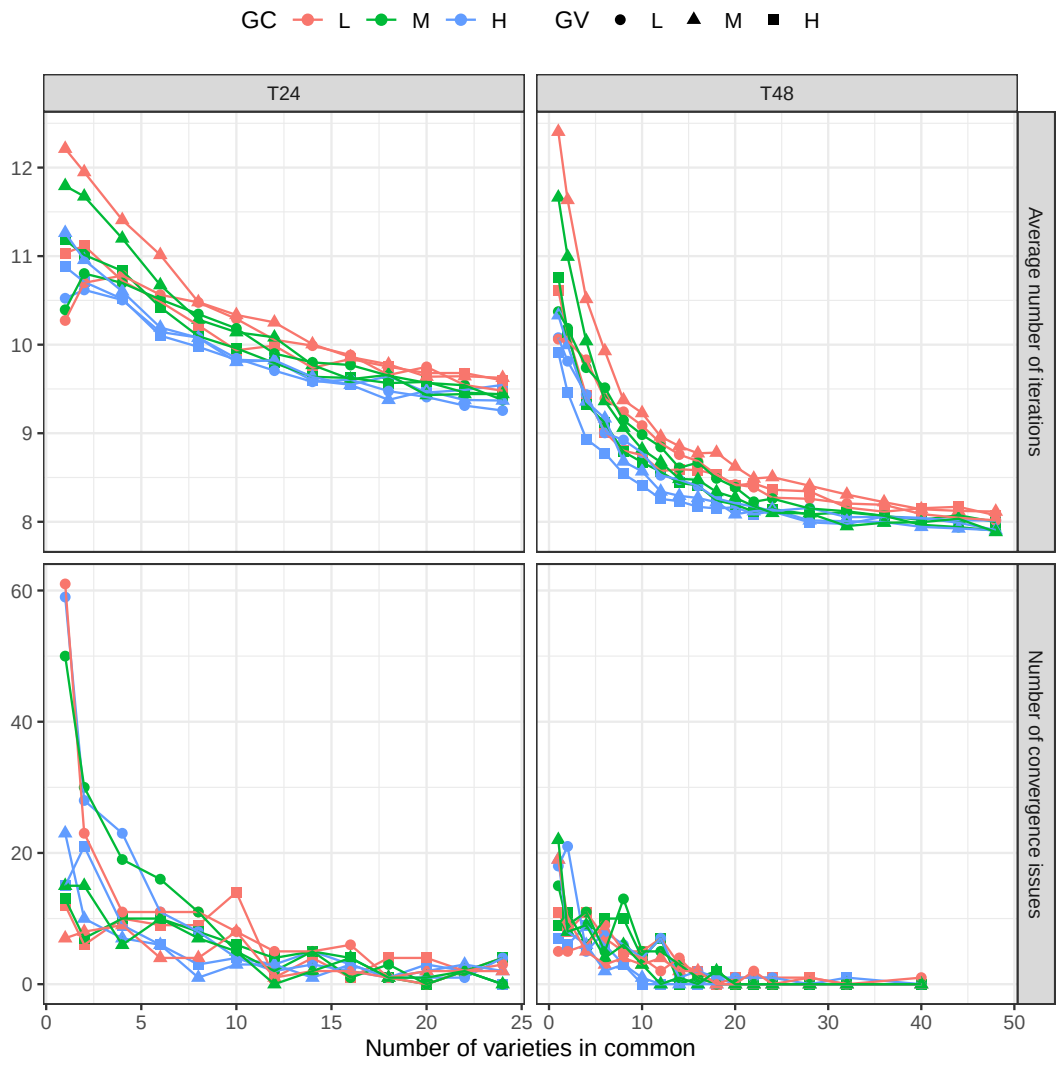

Figure 6.1: Number of iterations required for model convergence, and number of convergence issues from the 3200 simulations for each scenario. Colours and shapes as depicted in legend, which reference the three levels of genetic variance (GV), and between trials genetic correlations (GC).

6.5.2 Variance parameter estimates

Presented here are the summaries of bias and MSE of the estimated variance parameters from the simulation study. This includes,

- $\{\lambda_{11}, \lambda_{12}, \psi\}$ (FA variance parameters)
- $\{\sigma^2_{g_1}, \sigma^2_{g_2}, \rho_{12}\}$ (Genetic variance parameters)
- $\{\sigma^2_{b_1}, \sigma^2_{b_2}, \sigma^2_1, \sigma^2_2, \rho_{c_1}, \rho_{c_2}, \rho_{r_1}, \rho_{r_2}\}$ (Non-genetic variance parameters)

6.5.2.1 Bias

Figures 6.2, 6.3, 6.4, and 6.5 present the bias (see Chapter 3, Equation 3.3) for each variance parameter estimate across varying levels of $x_{1,2}$ for non-genetic, FA, and genetic variance/correlation components. For the non-genetic variance parameters (Figures 6.2, 6.3) I see as expected, there are no bias relationships in relation to the levels of $x_{1,2}$. When comparing T24 to T48, there is a greater bias with the non-genetic components for T24. For the trial genetic variances (Figure 6.5) both *Tsize* show very little bias. For FA parameters (Figure 6.4), as well as the genetic correlation on the other hand, reveal a clear decrease in bias as connectivity is increased.

6.5.2.2 Mean squared error for factor analytic parameters

Figure 6.6 present the MSE (see Chapter 3, Equation 3.4) for each of the FA parameter estimates across varying levels of $x_{1,2}$. The MSE reveals a non-linear declining relationship as connectivity is increased for all FA parameters, with almost identical profiles across *Tsize*. However, for scenarios with high levels of genetic correlation between trials (GV=H), flat responses are shown, particularly for ψ.

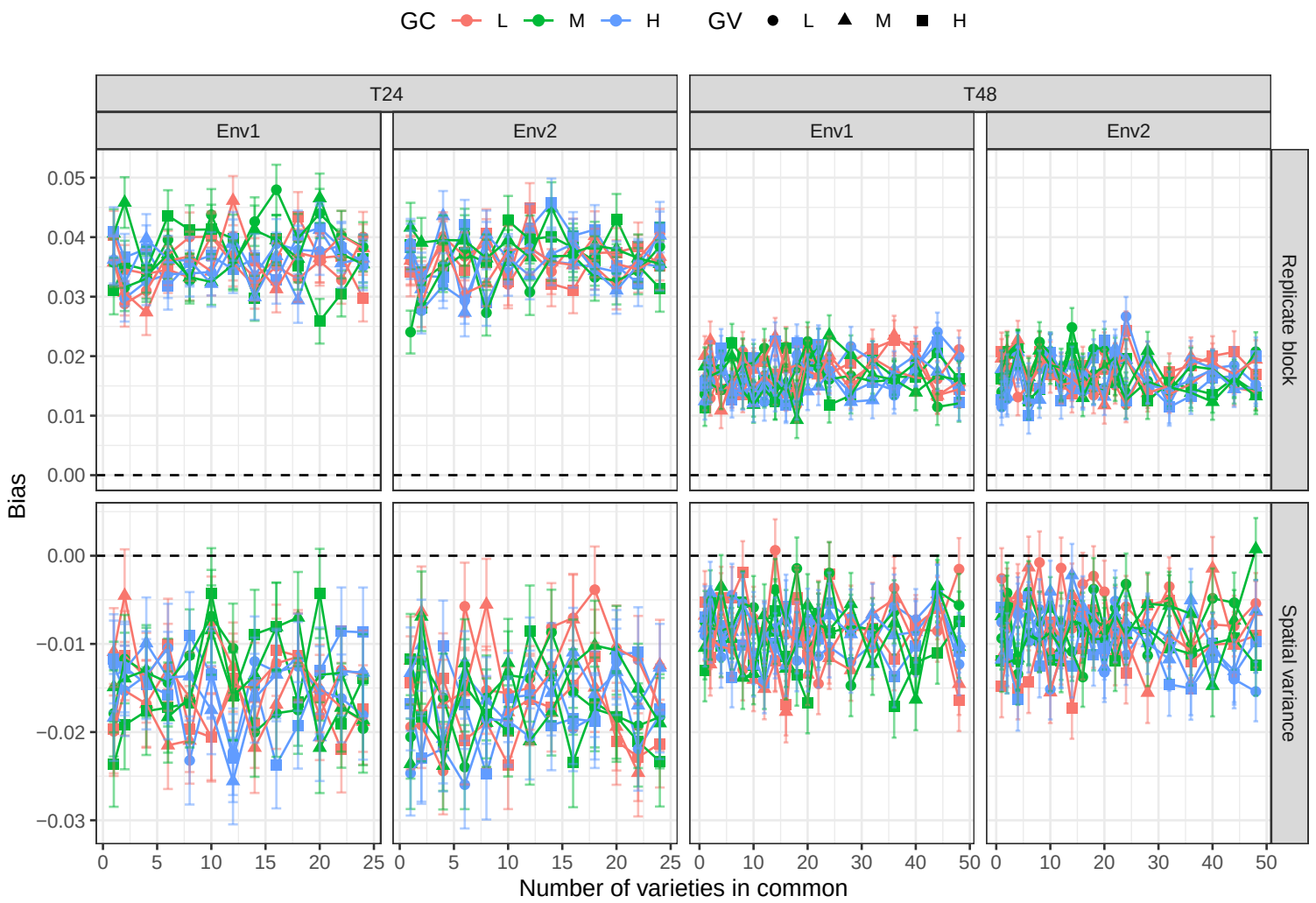

Figure 6.2: Bias with standard error bars for non-genetic variance components for each trial: replicate block $\{\sigma^2_{b_1}, \sigma^2_{b_2}\}$; and spatial variance $\{\sigma^2_1, \sigma^2_2\}$; against number of varieties in common ($x_{1,2}$). Colours and shapes corresponding to the legend which identifies each scenario, denoting the three levels of genetic correlation between trials (GC); and the three levels of genetic variance (GV). Dashed horizontal line represents no bias.

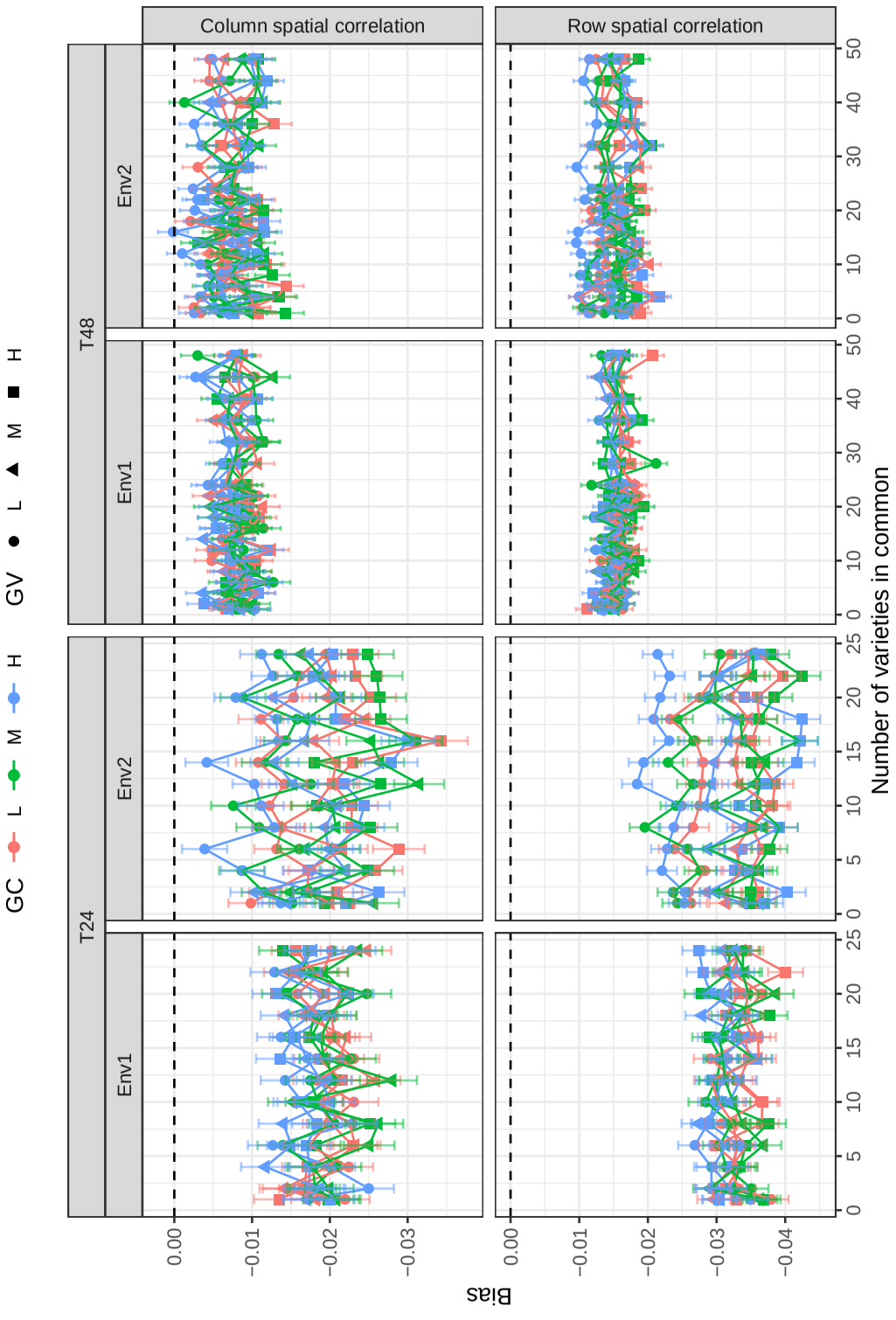

Figure 6.3: Bias with standard error bars for non-genetic variance components for each trial: spatial correlation in the column direction $\{\rho_{c_1}, \rho_{c_2}\}$; and spatial correlation in the row direction $\{\rho_{r_1}, \rho_{r_2}\}$ against number of varieties in common $(x_{1,2})$. Colours and shapes corresponding to the legend which identifies each scenario, denoting the three levels of genetic correlation between trials (GC); and the three levels of genetic variance (GV). Dashed horizontal line represents no bias.

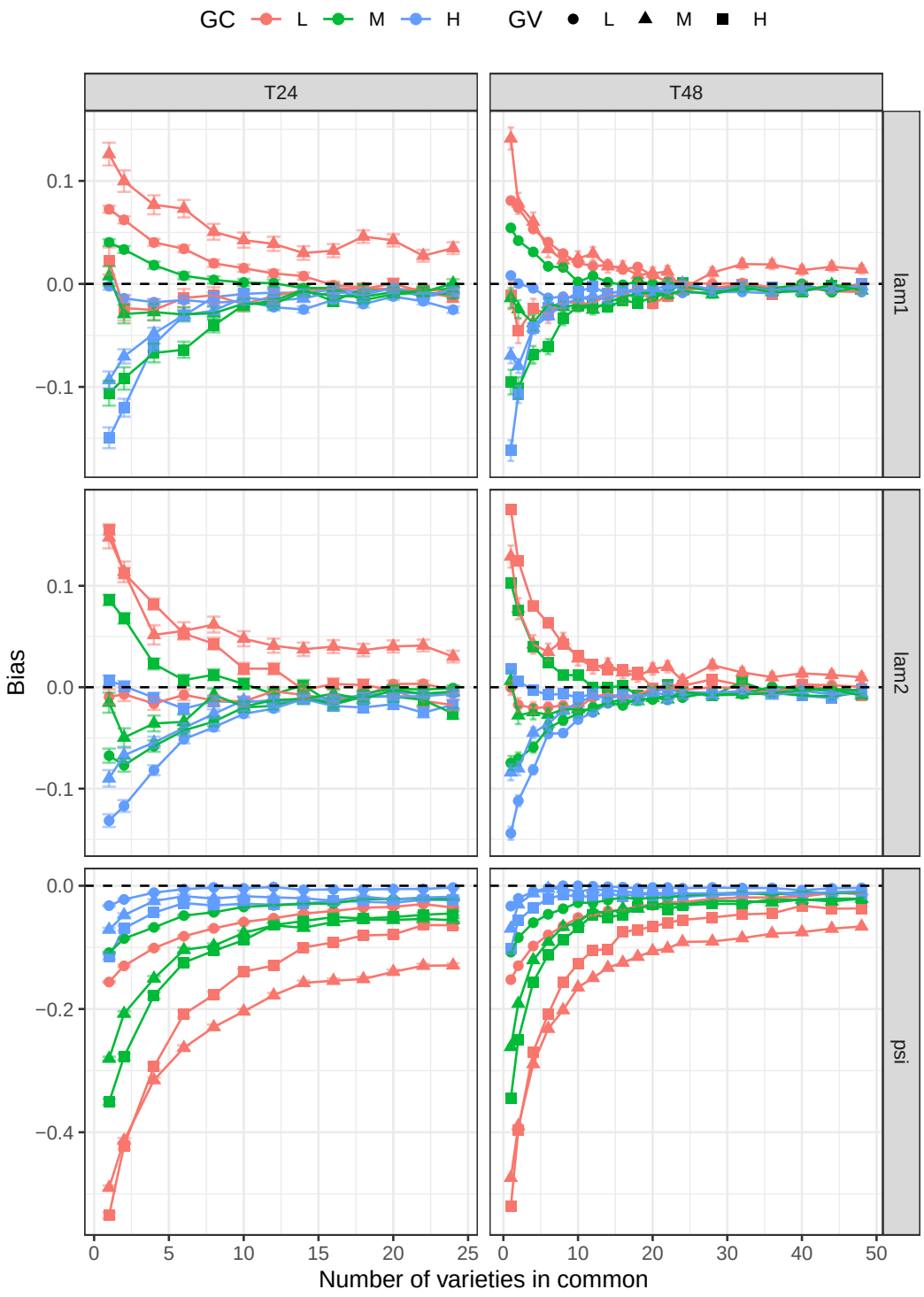

Figure 6.4: Bias with standard error bars for FA variance components against number of varieties in common ($x_{1,2}$). Loading for $Env1$ (lam1 $= \lambda_{11}$); loading for $Env2$ (lam2 $= \lambda_{12}$); and combined across trial specific variance (psi $= \psi$). Colours and shapes corresponding to the legend which identifies each scenario, denoting the three levels of genetic correlation between trials (GC); and the three levels of genetic variance (GV). Dashed horizontal line represents no bias.

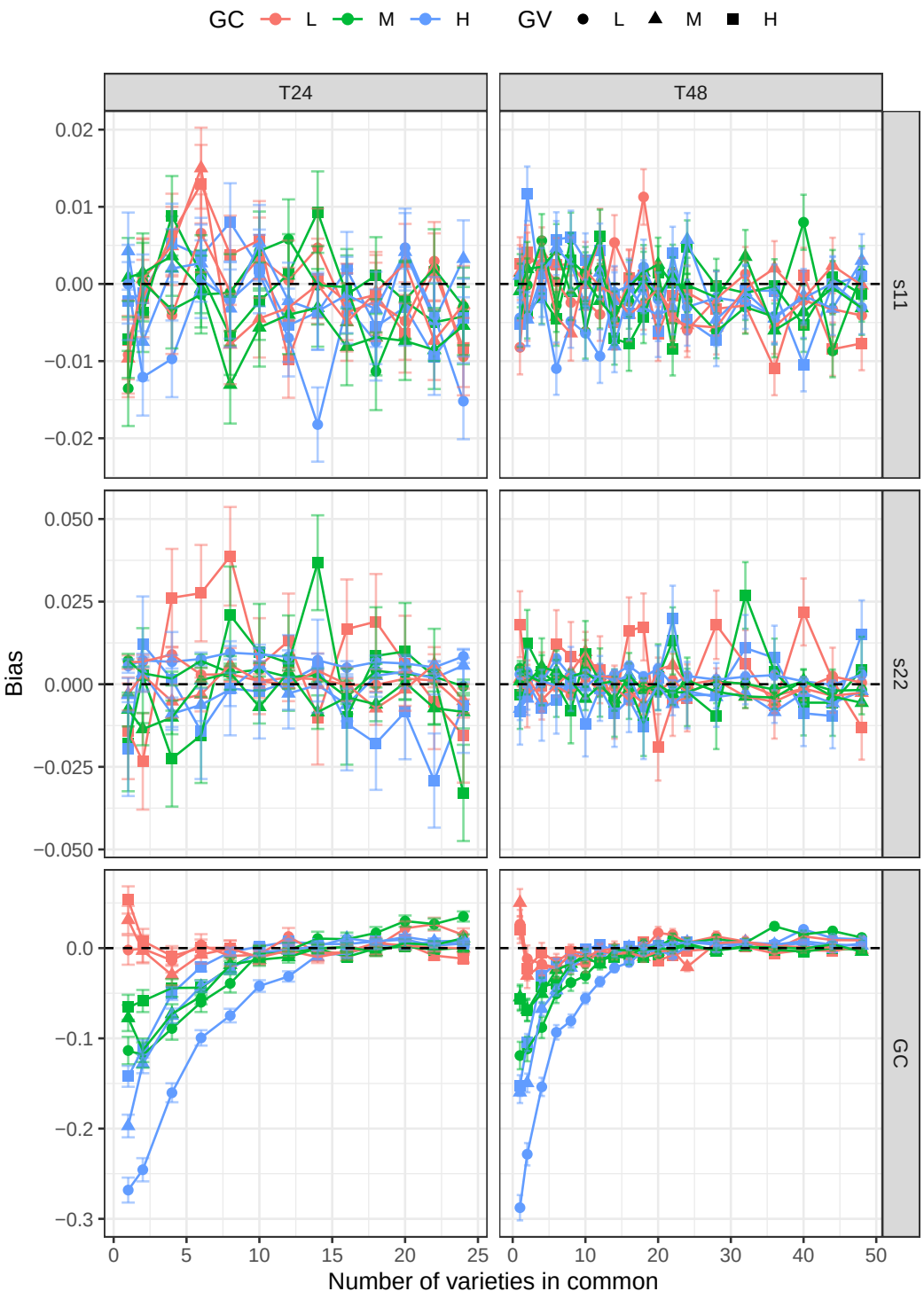

Figure 6.5: Bias with standard error bars for genetic variance components and genetic correlations against number of varieties in common ($x_{1,2}$). Genetic variance for $Env1$ (s11 $= \sigma^2_{11}$); genetic variance for $Env2$ (s22 $= \sigma^2_{12}$); and genetic correlation between $Env1$ and $Env2$ (GC $= \rho_{12}$). Colours and shapes corresponding to the legend which identifies each scenario, denoting the three levels of genetic correlation between trials (GC); and the three levels of genetic variance (GV). Dashed horizontal line represents no bias.

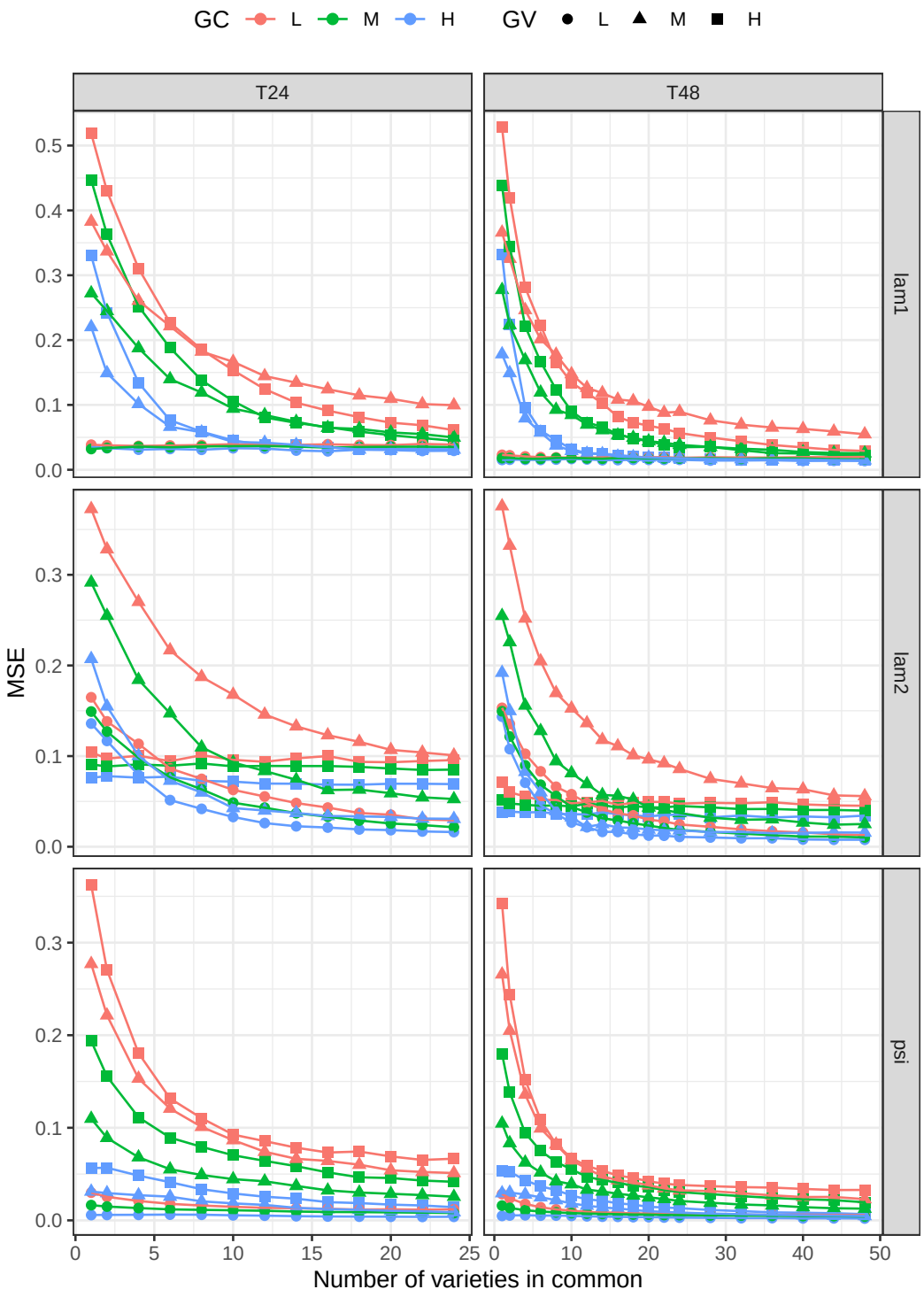

Figure 6.6: MSE for FA variance parameters against numbers of varieties in common ($x_{1,2}$). Loading for $Env1$ (lam1 = λ_{11}); loading for $Env2$ (lam2 = λ_{12}); and combined across trial specific variance (psi = ψ). Colours and shapes corresponding to the legend which identifies each scenario, denoting the three levels of genetic correlation between trials (GC); and the three levels of genetic variance (GV).

6.5.3 Reliability of predicted VE effects

The ultimate purpose of analysing a MET dataset is to select the (true) best varieties from a cohort of varieties. This section examines the reliability of predicted VE effects in respect to variety connectivity. I begin by looking at the results of the T24 MMM scenario, and then extend to investigate further scenarios. I again reiterate that the focus is on the reliability of predicted VE effects from the varieties in $Env1$ alone, since this contains the same varieties across all connectivity levels, thus allowing a fair comparison.

As discussed in Chapter 3 (Section 3.3.2), I compute analogous reliability based values that assume known variance parameters (not REML estimates), which I call design-based reliability values. These reflect the maximum possible reliabilities, which I denote as R^D_{kc} for variety k, and again when averaged across all varieties I denote $\bar{R}^D_c$. I recall that, R^D relates to BLUPs, and R^S to EBLUPs, that is, those derived using the REML estimates of the variance components.

Table 6.8 and Figure 6.7(a) present both simulated and design-based reliability values $\bar{R}^S_c$ and $\bar{R}^D_c$ for the T24 MMM scenario. I observe $\bar{R}^S_c$ values are typically increasing with the number of common varieties, but the $\bar{R}^D_c$ values are flat. As a result, the difference between the two reliability values ($\bar{R}_{\text{loss}}$) narrow as the number of common varieties increases. This is discussed more in the next section in measures of associated loss.

6.5.4 Mean loss in reliability of EBLUPs of VE effects

I calculate the mean loss in reliability of EBLUPs of VE effects (see Chapter 3, Equation 3.14) within each level of variety connectivity c, as the difference between the (across varieties average) simulated ($\bar{R}^S_c$) and design-based ($\bar{R}^D_c$) reliabilities. As previously stated, since the $\bar{R}^S$ and $\bar{R}^D$ refer to the reliabilities of the EBLUPs and BLUPs, respectively, and consider the losses owing to the REML estimation process of the variance components and prediction of the EBLUPs of VE effects.

Table 6.8: T24 MMM: Simulated (R_c^S) and design-based (R_c^D) reliability and loss ($\bar{R}_{\text{loss}}$) for the predicted VE effects for $Env1$ against number of varieties in common ($x_{1,2}$).

$x_{1,2}$	$\bar{R}_c^S$	$\bar{R}_c^D$	$\bar{R}_{\text{loss}}$
1	0.743	0.796	0.053
2	0.763	0.801	0.038
4	0.759	0.797	0.038
6	0.765	0.796	0.031
8	0.757	0.792	0.035
10	0.763	0.793	0.030
12	0.770	0.794	0.024
14	0.768	0.794	0.026
16	0.770	0.793	0.023
18	0.774	0.794	0.020
20	0.773	0.793	0.021
22	0.770	0.793	0.024
24	0.775	0.792	0.017

For the T24 MMM scenario, Table 6.8 and Figure 6.7(b) show the mean losses in reliability of the predicted VE effects for $Env1$ for varieties V1 - V24 that were present in both $Env1$ and $Env2$. The losses are shown to decrease at a non-linear rate with the losses ranging from 0.053 for $x_{1,2} = 1$ to 0.017 for $x_{1,2} = 24$.

The mean loss in reliability of the EBLUPs of VE effects for $Env1$ against $x_{1,2}$ for T24 and T48 are shown in Figures 6.8 and 6.9, respectively. Across scenarios, they exhibit similar complex relationships for both $Tsize$, however with greater losses for T24 than for T48. These results clearly show that variety connectivity alone does not fully explain losses in reliability of predicted VE effects.

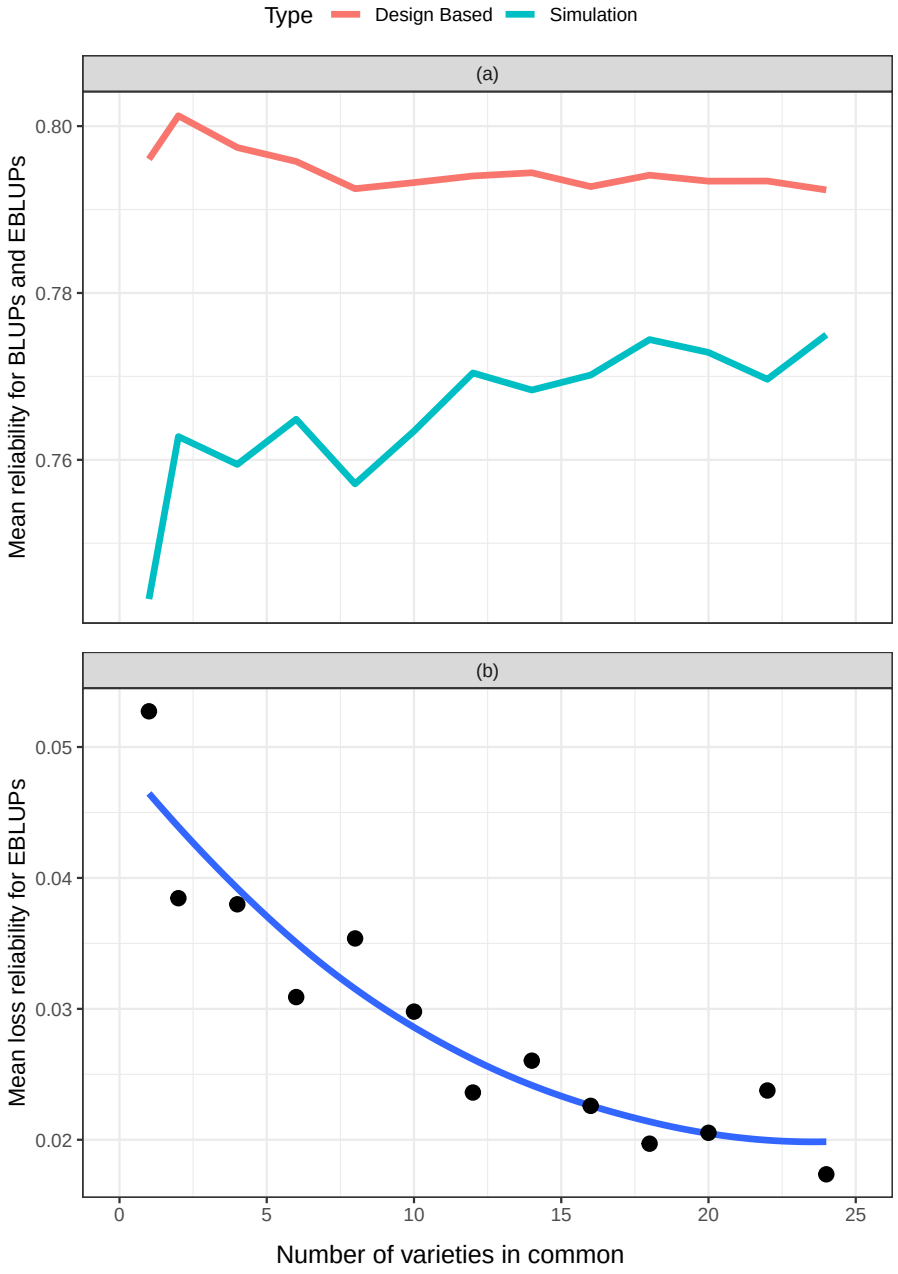

Figure 6.7: T24 MMM: (a) Mean reliability for VE effects for $Env1$ for design-based values (BLUPs) and simulated predicted values (EBLUPs), against number of varieties in common ($x_{1,2}$). Colours as represented in the legend represent design or simulated-based values. (b) Mean loss in reliability (difference between design and simulation-based reliability values) of the EBLUPs of VE effects for $Env1$, against number of varieties in common ($x_{1,2}$). The solid blue line in (b) represents a smooth line through the points.

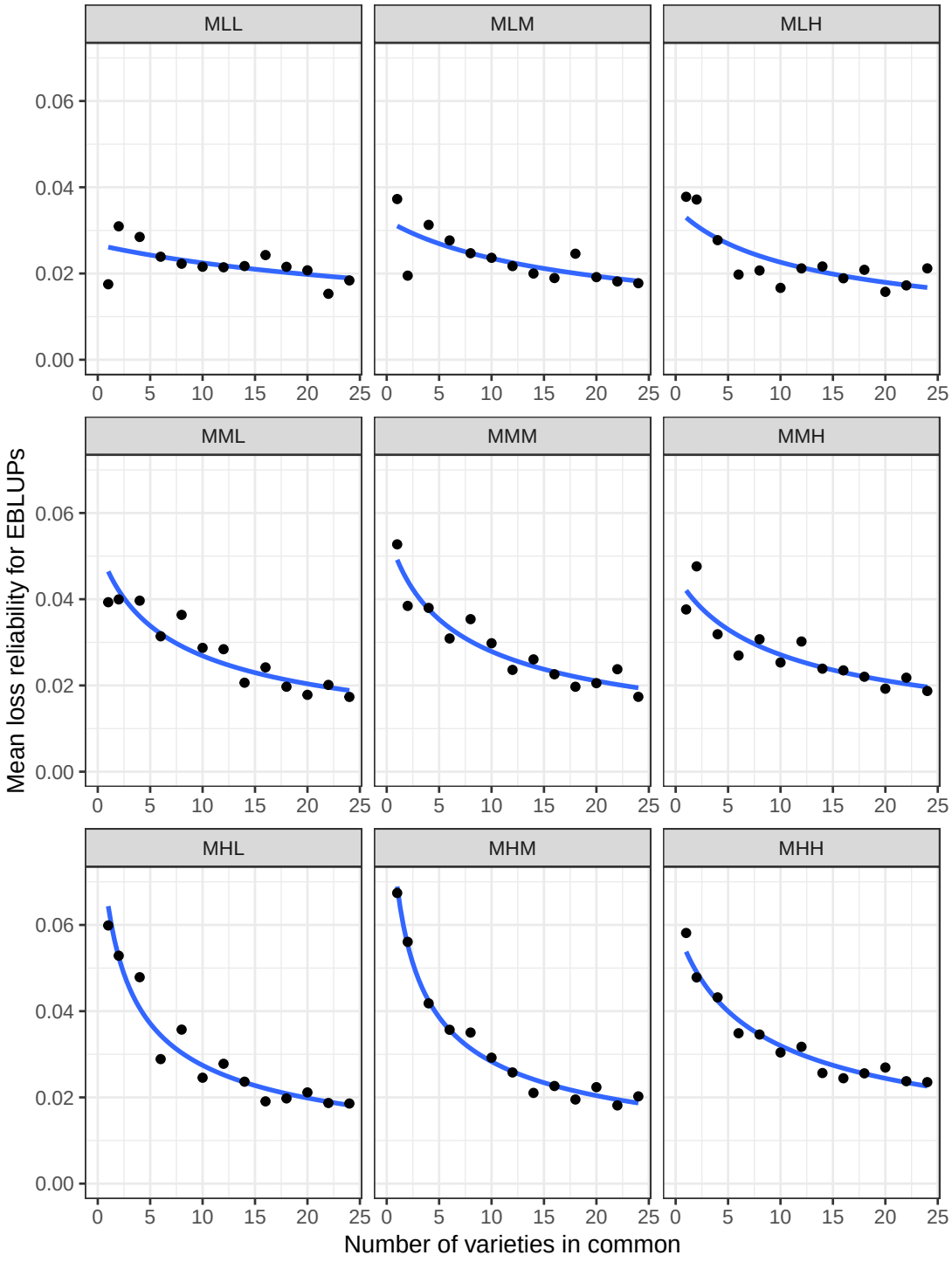

Figure 6.8: T24: Mean loss in reliability of the EBLUPs of VE effects for $Env1$, against number of varieties in common ($x_{1,2}$) for all scenarios. The solid blue line represents a smooth line through the points.

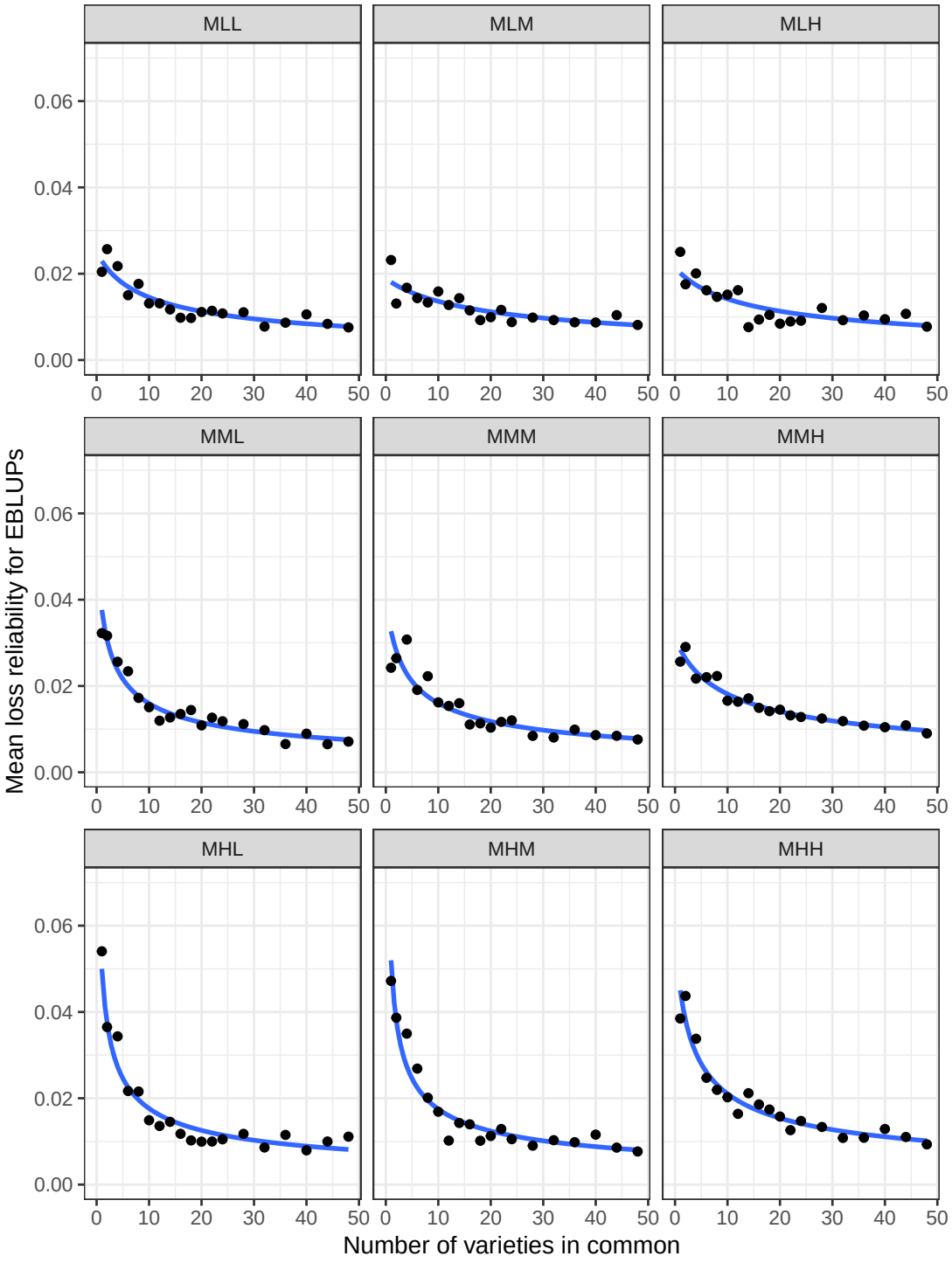

Figure 6.9: T48: Mean loss in reliability of the EBLUPs of VE effects for $Env1$, against number of varieties in common ($x_{1,2}$) for all scenarios. The solid blue line represents a smooth line through the points.

6.6 Concluding remarks

In this chapter I investigate the impact of variety connectivity on the reliability of predicted VE effects from a FALMM. The simulated scenarios are based on the Oat dataset (see Chapter 4), which corresponds to the late stages of a breeding cycle with a low to moderate number of varieties. It is also thought to represent standard late-stage evaluation MET datasets used internationally with analyses involving independent variety effects.

The results have shown complex relationships between genetic scenarios and variety connectivity. However, I have observed differences in reliabilities and mean losses of predicted VE effects in relation to trial size, with smaller trial sizes exhibiting greater losses. These results indicate that traditional variety connectivity measures do not capture the whole story and should be used with caution to evaluate the inclusion or exclusion of environments in a MET dataset.

This work has inspired new methodology, which is described in Chapter 9. The objective of this new method is to give a superior diagnostic than the traditional connectivity type measure in the sense of better forecasting the uncertainty of genetic variance parameter estimates. In addition, rather than just the basic assumption of independent variety effects used in this chapter, I present a technique for also investigating the reliabilities of predicted VE effects with datasets containing related variety effects.

Chapter 7

Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs

Because of the intrinsic structure of a breeding program, and consequently selection on a stage basis, analysis of plant breeding datasets frequently consists of a succession of single year based analyses (Arief et al., 2019). According to Bernal-Vasquez et al. (2017), datasets are frequently analysed on a year basis rather than spanning years for two reasons: it is simpler and faster; and it is difficult to quantify variation across years due to the lack of common varieties across breeding stages. We emphasize that, because plant breeding operations are commercial, there is frequently a lack of literature on the methods used to create a MET dataset within the context of a commercial breeding program.

The objective of this chapter is to demonstrate how to create MET datasets using the contemporary grouping (CG) methodology of Smith et al. (2021a). The fundamental concept is to include sufficient trials to gather as much information as possible on the varieties under evaluation for selection. The impact of this method is evaluated using the $\mathcal{A}$-optimality criterion from model-based design theory.

This chapter is arranged as follows: Section 7.1 contains some general results about

$\mathcal{A}$-optimality. Following this, Section 7.2 presents a reproduction of key sections of the publication listed below. Note that although the candidate is not the first author of this paper, the candidate had a key role in dataset curation, application of the methodology to the examples, and manuscript preparation and revision. In Section 7.3 we demonstrate how the CG methodology is used in practice when consulting with a plant breeder. Finally, in Section 7.4 we have concluding remarks.

Smith, A., Ganesalingam, A., Lisle, C., Smith, A., Kadkol, G., Hobson, K. & Cullis, B. R. (2021). Use of Contemporary Groups in the Construction of Multi-Environment Trial Datasets for Selection in Plant Breeding Programs. *Frontiers in Plant Science*. **11**, 2325. doi: 10.3389/fpls:2020.623586

7.1 Preliminary remarks about $\mathcal{A}$-optimality

In the experimental design literature, $\mathcal{A}$-optimality is used to search for designs that minimise the average variance of treatment contrasts (Butler, 2013). The criterion is often referred to as the "average pairwise variance" and is widely used for fixed treatment effects in comparative experiments (John & Williams, 1996). In the context of MET dataset construction, we wish to know whether the trials that have been compiled provide sufficient information to limit the probability of errors in selecting varieties to progress to the next stage of testing. Bueno Filho & Gilmour (2003, 2007) show that if the $\mathcal{A}$-optimality criterion for fixed treatment effects is generalised for the case of random treatment (variety) effects, then it aligns with the probability of incorrect selection decisions. If we let u_i and u_j be the random effects for variety i and j respectively (for $i, j = 1 \dots m$) and let $\tilde{u}_i$ and $\tilde{u}_j$ be the associated BLUPs, then the $\mathcal{A}$-optimality criterion in this context is given by

$$\mathcal{A} = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \text{var}\left((u_i - \tilde{u}_i) - (u_j - \tilde{u}_j)\right) \tag{7.1}$$

Thus in the random effects setting the $\mathcal{A}$-optimality criterion is the "average pairwise prediction error variance" (Cullis et al., 2006).

In Bueno Filho & Gilmour (2003) it is commented that "a good design is strongly associated with the precision of pairwise comparisons because from these comparisons an appropriate ranking can be carried out" and thence selections can be made. In Bueno Filho & Gilmour (2007) a Bayesian justification is given for the use of the criterion in Equation 7.1. Their argument, using our framework and notation, proceeds by defining the problem as selecting s out of the m varieties for progression to the next stage of testing. If we let S be the set of the s best varieties then interest focuses on comparisons of varieties in this set with those not in the set. Formally, we are interested in differences $\Delta_{ij} = u_i - u_j$ for $i \in S$ and $j \notin S$. Then to minimise the probability of any single error in selection, given the data, we must minimise $\Pr(\Delta_{ij} < 0 | \boldsymbol{y_2})$. In Section 2.3.2 it was shown that this probability is a function of the prediction error variance of the difference, that is $\text{var}\left((u_i - \tilde{u}_i) - (u_j - \tilde{u}_j)\right)$ so that this variance must be minimised. Bueno Filho & Gilmour (2007) then state that "any such mistake leads to an incorrect selection" so that the design criterion should minimise $\sum_{i \in S} \sum_{j \notin S} \text{var}\left((u_i - \tilde{u}_i) - (u_j - \tilde{u}_j)\right)$. Finally, Bueno Filho & Gilmour (2007) state that if "any treatment [variety] is equally likely to be amongst the best s", then the criterion required is to minimise the total (or average) of all $m(m-1)$ pairwise prediction error variances. This is exactly the $\mathcal{A}$-optimality criterion of Equation 7.1.

7.2 Reproduction of Smith, Ganesalingam, Lisle, Kadkol, Hobson and Cullis (2021)

In the reproduction, notational changes have been made to be consistent with the nomenclature used in this thesis. It is worth noting that the Chickpea dataset is not utilised here. Parts of the introduction and discussion sections of the paper have been omitted.

7.2.1 Methods for MET dataset construction

In Chapter 4 it was demonstrated that breeding programs are supported by a group of varieties produced jointly at the crossing block stage. These are said to be 'born' together and assessed successively in the advanced assessment stages. These are referred here as contemporary groupings (CGs). For illustration we again consider the Durum dataset, which is fully described in Chapter 4, in which there are four stages of testing,

which are denoted S1 to S4. From 2015 to 2018, the following four CGs were created: CG15, CG16, CG17, and CG18, which corresponded to varieties in S1 trials during that time frame. Then, for example, a subset of the CG15 varieties was progressed to S2 trials in 2016, then to S3 in 2017, and eventually to S4 in 2018. As varieties progress through stages, their number decreases, as illustrated in the rows of Figure 4.5. Four selection decisions would be made yearly on S1, S2, S3, and S4 varieties as they move to the next stage of testing.

The CG concept for MET data construction is first illustrated using a hypothetical breeding program with four stages of testing (S1 to S4) and in which varieties progress through stages without fast-tracking (skipping stages) or retention (remaining in a stage for more years of testing). The aim is to construct a dataset to enable accurate selection decisions for 2018.

We first consider the decisions on varieties in S4 in 2018. These varieties commenced their testing in S1 trials in 2015 (so are all members of CG15), were selected to be tested in S2 trials in 2016, then S3 trials in 2017, and finally S4 trials in 2018. Thus, in order to capture all of the data on the varieties under consideration for selection, we would combine data from all of the trials in this sequence. Overall, this would suggest a separate analysis for each of the selection decisions, based on combining data from the following trials:

- **S1 selection decisions:** S1 trials 2018.
- **S2 selection decisions:** S2 trials 2018 + S1 trials 2017.
- **S3 selection decisions:** S3 trials 2018 + S2 trials 2017 + S1 trials 2016.
- **S4 selection decisions:** S4 trials 2018 + S3 trials 2017 + S2 trials 2016 + S1 trials 2015.

It is instructive to illustrate this compilation of trials across stages and years using tables such as Table 7.1. In this table the diagonal bands of stages across years are labelled as A to I, with the labels A to D being assigned in such a way that they align with the S1 to S4 trials in the year of selection (here 2018). The datasets described above correspond to the diagonal bands of trials labelled as A to D. Thus, for example, band D comprises data from S1 trials in 2015, S2 trials in 2016, S3 trials in 2017, and

S4 trials in 2018. It is important to note that, for any given trial, the data from all of the harvested plots is included and not just the data on the varieties of interest. Thus, for example, the data from S1 trials in 2015 relates to all of the varieties tested in those trials. We describe bands A to F in Table 7.1 as being "complete" in the sense that they trace back to the first stage of testing, namely S1. In contrast, bands G to I are incomplete, with band G missing S1 trials, band H missing S1 and S2 trials and band I missing S1, S2, and S3 trials. This has implications in terms of selection bias which will be discussed in Section 7.2.3.

In the absence of retention or fast-tracking, all the varieties in S1, S2, S3, and S4 in 2018 are members of CG18, CG17, CG16, and CG15, respectively. Thus, all the varieties within a stage in 2018 belong to a single CG only and the entire selection history for any of these varieties is captured in the associated data band. The generalization to a more complex scenario will be discussed in the context of the Durum dataset example (see Section 7.2.3).

In terms of information available for each of the four selection decisions it is instructive to differentiate between "direct" and "indirect" information. The former relates to observed data so is maximized by including all trials in which the varieties of interest have been grown. In the hypothetical example this corresponds to the bands so suggests the conduct of four analyses each based on a separate band (A, B, C, and D). However, the use of a FALMM for analysis creates the possibility of also using indirect information derived from genetically related varieties in other trials. We would therefore recommend undertaking a single analysis using data combined across these bands. This recommendation can be justified by applying the method described in Section 7.2.2 to quantify information for selection. Finally, we note that in the MET analysis, V×E is modelled with reference to environments which are defined to be combinations of trial locations and years. Combining across bands may lead to the presence of multiple trials at a single location within a year. For example, in any given year, locations with S1 trials also typically include S2, S3, and S4 trials. We refer to such trials as 'co-located'.

Table 7.1: Data bands for potential inclusion in a MET dataset for selection decisions in 2018 from a breeding program with four stages of selection (S1 to S4).

	Year					
Stage	2013	2014	2015	2016	2017	2018
S1	F	E	D	C	B	A
S2	G	F	E	D	C	B
S3	H	G	F	E	D	C
S4	I	H	G	F	E	D

7.2.2 Quantifying information for selection in MET datasets

In order to discriminate among possible MET datasets in terms of the amount of information available for selection decisions, we note that the problem has strong links with optimal (model-based) design. As Butler et al. (2014) state, "The goal of optimal design is to discriminate among competing designs in an effort to maximize the treatment information from a fixed number of experimental units." This requires the use of an optimality criterion, and, in the context of plant breeding trials in which the treatments are varieties and the aim is selection, the $\mathcal{A}$-optimality criterion (see Section 7.1) is appropriate since this aligns with minimizing the probability of an incorrect selection decision (Bueno Filho & Gilmour, 2003). $\mathcal{A}$-optimality is based on the so-called $\mathcal{A}$-value which is the average pairwise variance of elementary treatment contrasts. We therefore propose to use $\mathcal{A}$-values to quantify the treatment (variety) information available in any given MET dataset.

In model-based design, $\mathcal{A}$-values are computed under a pre-specified LMM which we will term the design model. Specification of the design model requires specification of the fixed and random effects, the variance models for the random effects and errors and the values of the associated variance parameters. The design model is usually chosen to be as close as possible to that expected for the analysis. Additionally, the variance parameter values are chosen as being "typical" so may be based on historic analyses. The model proposed in this thesis for the analysis of MET data is the FALMM with the inclusion of pedigree information (when available). Variety selections using this model are typically focused on the measure of overall variety performance (across environments) as presented in Smith & Cullis (2018). However, the factor analytic variance

parameters are specific to the individual environments in the dataset so that typical values do not exist. Therefore a more generic, but still realistic design model is required for assessing MET dataset information. We have chosen a variance component model that involves random variety main effects and random V×E effects, both of which are partitioned into additive and non-additive effects. This is, in fact, a sub-model of the FALMM. The $\mathcal{A}$-values are then computed for the total (additive plus non-additive) variety main effects since these provide a measure of average performance of varieties across environments.

In order to determine reasonable values for the variance parameters in this design model we consider Cullis et al. (2000) who conducted variance component analyses of grain yield in 22 MET datasets from Australian crop variety evaluation programs. The environments in those datasets were classified according to the year, the geographic region and possibly location within region so Cullis et al. (2000) partitioned V×E accordingly. In the Durum dataset example we do not have regional information nor are trials typically located in identical positions from year to year. However, we recognize the importance of variety by year interaction so maintain this as a separate source in the design model. Thus, we have used the variety main effects (V), variety by year interaction (V×Y), and error sources of variation from Cullis et al. (2000), and have added together the remaining sources to form residual variety by environment interaction. The mean percentage contributions for each of these sources across all 22 datasets was 13.77% (V), 8.59% (V×Y), 37.91% (residual V×E), and 39.73% (Error) (see Table 7.2). In the model-based design literature, and without loss of generality, a value of one is typically assumed for the error variance. We adopt the same approach here (see second row in Table 7.2).

In contrast to our approach, the analyses in Cullis et al. (2000) do not involve information on genetic relatedness. We therefore make the further assumption that additive variance comprises 80% of total variance, a value that is often encountered in practice. The final values for the variance parameters in the design model are given in the third and fourth rows of Table 7.2. All $\mathcal{A}$-values in this thesis were computed using `ASReml-R` (Butler et al., 2017).

Table 7.2: Variance parameter values for design model for variety main effects (V), variety by year interaction (V×Y), residual variety by environment interaction (V×E), and Error.

	V	V×Y	V×E	Error
Mean % from Cullis et al. (2000)	13.77	8.59	37.91	39.73
Total variance parameter	0.35	0.22	0.95	1.00
Additive variance parameter	$0.28/\bar{a}$	$0.18/\bar{a}$	$0.76/\bar{a}$	
Non-additive variance parameter	0.07	0.04	0.19	

Rows in the table are: means of percentages in Cullis et al. (2000); associated (total) variance parameter values assuming error variance of one; additive variance parameter values (numerator is 80% of total values and denominator, $\bar{a}$, is the mean of the diagonal elements of NRM); non-additive variance parameter values (20% of total values).

7.2.3 Application to the Durum dataset

In this section, we demonstrate how to use the CG methods to create an appropriate MET dataset from the Durum dataset described in Chapter 4 for the selection and progression to the next stage of testing of the 2018 S1, S2, S3, and S4 test varieties. Table 7.3 shows the number of test varieties for each stage and year (2013 to 2018). It is important to note that test varieties refer exclusively to those under consideration for selection, so excludes check varieties.

At any point of selection, a test variety may be chosen to go to the next level of testing, kept in the current stage, or rejected. We clearly observe that Durum test varieties are often held within stages for extra years of testing. The distribution of the 2018 varieties across CGs is shown in the last columns of Table 7.3. As an example, for the S3 test varieties (also see Figure 4.7), the majority (66) correspond to CG16, thus following the straightforward progression along band C, as shown in Table 7.1. However, a significant number (22) correspond to CG15 and progressed from S1 trials in 2015 to S2 trials in 2016, to S3 trials in 2017, and finally to S3 trials in 2018. Finally, five test varieties correspond to CG14 which advanced from S1 trials in 2014 to S2 trials in 2015 to S3 trials in 2016, and were then kept in S3 in 2017 and 2018. We have also observed a similar pattern for the S4 test varieties (see Figure 4.6), however with a higher level of

retention. This has ramifications for the MET dataset's construction.

The starting point for MET dataset construction for selection decisions on the 2018 test varieties (S1 to S4) includes all trials in bands A-D. With the retention of test varieties, it is evident that most of the data on the 30 S3 and S4 test varieties that belonged to CG14 and CG13 in 2018 will be lost. Table 7.4, for example, indicates that there are nine test varieties in S4 for which five years of data would be missing if the dataset only included bands A-D; another nine test varieties for which four years would be missing; and seven test varieties lacking two or three years of data. This is, in our opinion, unacceptable. As a result, we look at the inclusion of bands E and F to the data. The improvement in collecting more information on the test varieties of interest is seen in Table 7.4. Full data on the test varieties of interest may be acquired by adding band G, but we advise against doing so because band G is incomplete as it lacks S1 trials (see Table 7.1), and we do not have the whole selection history for many of the test varieties in band G. The inclusion of band G, as well as the two incomplete bands H and I, may result in 'selection bias', that is, bias in the estimates of the genetic variance parameters (Thompson, 1973) and hence we do not recommended. As a result, bands A-F are chosen for the final dataset, where only five S4 test varieties are missing data in this dataset. The final MET dataset for analysis (bands A-F) included yield data on 6951 varieties from 21660 plots, comprising 97 trials across 30 environments.

Finally, to compare the MET datasets we use the $\mathcal{A}$-optimality criterion (see Section 7.1) as a diagnostic for each selection and for each of the five MET datasets: 2018 data for each stage, the diagonal band of data for each stage, combined data bands A-D and A-E, and the final dataset (bands A-F). Figure 7.1 illustrates the resultant $\mathcal{A}$-values, which clearly show the final dataset's advantage in each scenario. The decrease in $\mathcal{A}$-values is mostly driven by an increase in the amount of direct information (as indicated in the mean number of environments per variety), but indirect information also plays a significant role. For example, the S1 varieties under consideration for selection in 2018

were only grown in one environment, therefore there is no clear information difference between utilising the 2018 data alone for this stage vs the final dataset. The $\mathcal{A}$-value for the final dataset is substantially smaller, suggesting the influence of indirect information from relatives of the S1 variety.

Table 7.3: Number of test varieties in each Stage (S1-S4) and year (2013-2018) in the Durum dataset.

	Number of test varieties						Number of 2018 test varieties					
Stage	2013	2014	2015	2016	2017	2018	CG18	CG17	CG16	CG15	CG14	CG13
S1	582	1485	1000	1163	1303	1148	1148	0	0	0	0	0
S2	105	361	413	388	379	315	0	315	0	0	0	0
S3	30	92	92	92	90	93	0	0	66	22	5	0
S4	25	41	57	55	53	56	0	0	0	31	12	13

The final columns give the number of test varieties in each contemporary group (CG13 - CG18) for the varieties for selection in 2018.

Table 7.4: MET dataset construction for 2018 selection decisions in the Durum dataset.

	#years	Bands in dataset			
Stage	missing	A-D	A-E	A-F	A-G
S1	0	1148	1148	1148	1148
S2	0	315	315	315	315
S3	0	88	93	93	93
	2	5	0	0	0
S4	0	31	42	51	56
	1	0	1	3	0
	2	6	0	2	0
	3	1	0	0	0
	4	9	13	0	0
	5	9	0	0	0

Number of test varieties missing data in datasets comprising bands A-D, A-E, A-F, and A-G. Total number of test varieties for selection: 1148, 315, 93 and 56 (for stages S1 to S4).

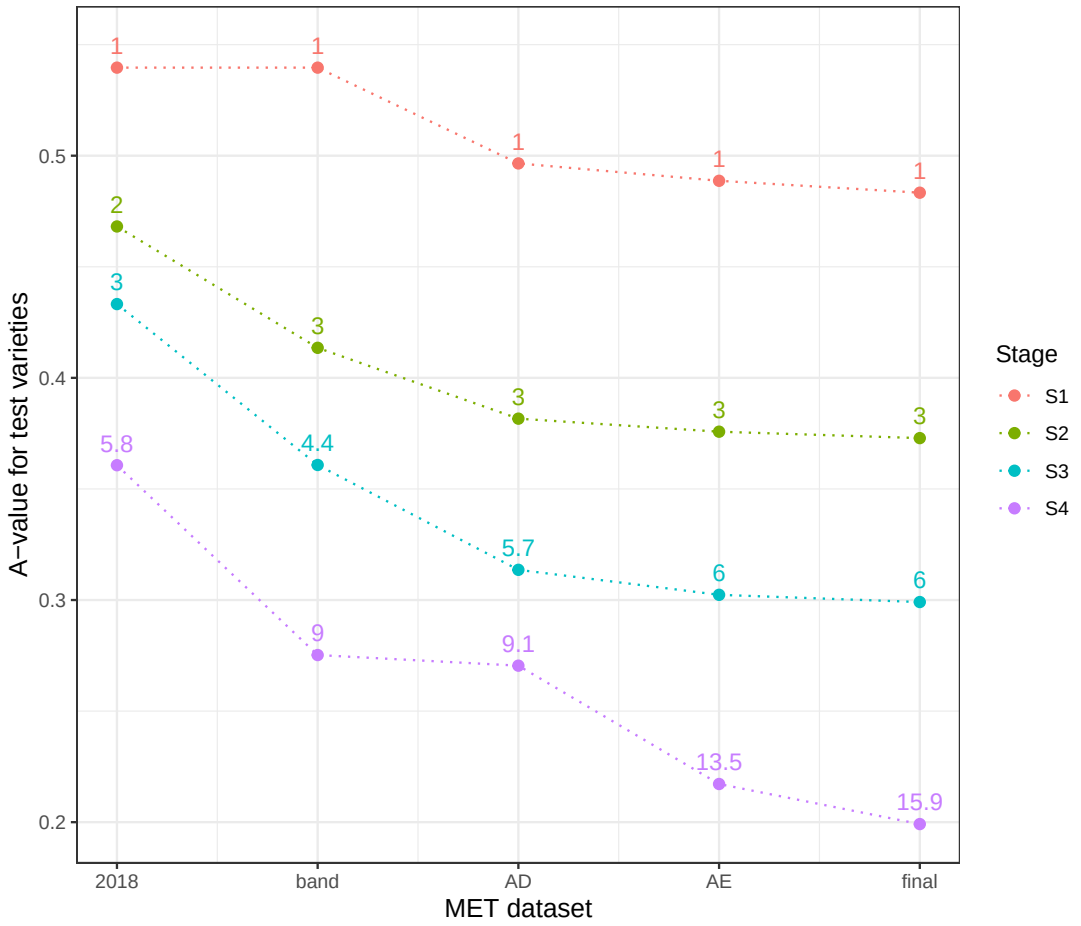

Figure 7.1: $\mathcal{A}$-values for 2018 test varieties under consideration for selection in the Durum dataset (S4: 56 varieties, S3: 93 varieties, S2: 315 varieties, and S1: 1148 varieties). $\mathcal{A}$-values are given for five MET datasets: 2018 data alone for each stage (2018), diagonal bands for each stage (band), data bands A-D (AD), data bands A-E (AE), and for the final (final) dataset as described in Section 7.2.3. For each MET dataset, each point is labelled with the average number of environments for the test varieties. It is worth noting that the S1 MET datasets for 2018 and band are same.

7.3 Contemporary groups in practice

The CG methodology has been frequently used to construct MET datasets in Australian plant breeding programs since the introduction of the publication. In this section, I describe how the CG techniques are applied in these programs to provide breeders a proposal for the MET datasets for analysis. A summary of the output provided to breeders is shown in this section, referred to here after as the 'CG spreadsheet', as well as the `R` code to generate this output.

I again use the Durum dataset as described in Chapter 4 as our example, and consider the varieties grown in the S3 trials in 2018, for which there are 93 test varieties for evaluation. The CG approach is simple and straightforward, with only a few steps, which are as follows.

1. Define varieties as either test varieties or checks. As previously defined, test varieties are those under consideration for selection, whereas checks are normally released varieties intended for comparison. This process is frequently more complicated than it appears, because some test varieties are utilised both as a check and as a candidate for selection.

2. For each current selection decision I then determine the CG for each test variety. The `R` code and output are below.

```
gg <- unique(alldata$Variety[alldata$YearStage=='2018S3'])
vS3.2018.df <- data.frame(Variety=gg)
vS3.2018.df$Vtype <- 'Test'
vS3.2018.df$Vtype[is.element(vS3.2018.df$Variety,
genos.checks)] <- 'Check'
length(gg) # 105
vvy <- with(alldata,tapply(as.numeric(as.character(Year)),
list(Variety,Year),mean))

yy <- vvy[vS3.2018.df$Variety,]
fy <- apply(yy,1,function(x) x[!is.na(x)][1])
vS3.2018.df$year1 <- fy
vvs <- with(alldata,tapply(as.numeric(Stage),
list(Variety,Year),min))
ss <- vvs[vS3.2018.df$Variety,]
fs <- apply(ss,1,function(x) x[!is.na(x)][1])
```

```
vS3.2018.df$stage1 <- fs
ss <- as.data.frame(ss)
names(ss) <- paste('S',names(ss),sep='-')
ss$Variety <- row.names(ss)
ss <- ss[!is.element(ss$Variety,genos.checks),]
vS3.2018.df <- merge(vS3.2018.df,ss,by='Variety',all.x=T)
vS3.2018.df$CG <- 'Unknown'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2013'] <- 'CG13'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2014'] <- 'CG14'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2015'] <- 'CG15'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2016'] <- 'CG16'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2017'] <- 'CG17'
vS3.2018.df$CG[vS3.2018.df$stage1=='1'&
vS3.2018.df$year1=='2018'] <- 'CG18'
vS3.2018.df$CG[vS3.2018.df$Vtype=='Check'] <- 'Check'
vS3.2018.df <- vS3.2018.df[order(vS3.2018.df$Vtype,
vS3.2018.df$CG),]

print(head(vS3.2018.df,4), row.names = FALSE)
#Variety Vtype year1 stage1 S-13 S-14 S-15 S-16 S-17 S-18   CG
#10TD032  Test  2014      1   NA    1    2    3    3    3 CG14
#10TD036  Test  2014      1   NA    1    2    3    3    3 CG14
#11TD010  Test  2014      1   NA    1    2    3    3    3 CG14
#11TD014  Test  2014      1   NA    1    2    3    3    3 CG14

with(vS3.2018.df,table(CG))
# CG
# CG14   CG15   CG16  Check
#    5     22     66     10
```

where `alldata` is the Durum dataset, `YearStage` is the concatenation of `Year` and `Stage`, and `geno.checks` is the list of check varieties determined in Step 1. As already shown in Table 7.3, the S3 varieties include 66 varieties derived from S1 in 2016 (CG16), 22 varieties from CG15, 5 varieties from CG14, and 10 are checks. This step is also particularly critical as a data checking mechanism, as any varieties with unknown origins may indicate that there are missing trials or problems with variety naming.

3. I then compile all the CGs to form a Table similar to that shown in Table 7.3. Below is the R code.

```
vS1.2018.df$Stage <- 'S1'
vS2.2018.df$Stage <- 'S2'
vS3.2018.df$Stage <- 'S3'
vS4.2018.df$Stage <- 'S4'
durumALL.2018.df <- rbind(vS1.2018.df,vS2.2018.df,
vS3.2018.df,vS4.2018.df)

tmp <- droplevels(alldata[!alldata$Variety %in%
genos.checks & !is.na(alldata$Variety),])
tmp$StageVar <- paste(tmp$Variety,tmp$YearStage,sep='_')
tmp <- tmp[!duplicated(tmp$StageVar),]
ttv <- table(tmp$Stage,tmp$Year)

durumALL.2018.df$Stage <- factor(durumALL.2018.df$Stage,
levels=c('S1','S2','S3','S4'))
durumALL.2018.df$CG <- factor(durumALL.2018.df$CG,
levels=c('CG18','CG17','CG16','CG15','CG14','CG13','Check'))

ttg <- table(durumALL.2018.df$Stage,durumALL.2018.df$CG)
cbind(ttv,ttg)

#    2013 2014 2015 2016 2017 2018 CG18 CG17 CG16 CG15 CG14 CG13 Check
#S1   582 1485 1000 1163 1303 1148 1148    0    0    0    0    0     7
#S2   105  361  413  388  379  315    0  315    0    0    0    0     6
#S3    30   92   92   92   90   93    0    0   66   20    5    0    12
#S4    25   41   57   55   53   56    0    0    0   31   12   13    12
```

4. For each selection decision I then form a table presenting the flow from stage to stage across years. Below is the R code. Note that this is how Figure 4.7 was formed.

```
tmp.S3 <- tmp[tmp$Variety %in% vS3.2018.df$Variety,]
tmp.S3$uniq <- paste(tmp.S3$Variety,tmp.S3$YearStage,sep="_")
tmp.S3 <- tmp.S3[!duplicated(tmp.S3$uniq),]
#Example for the 93 S3 varieties
table(tmp.S3$Stage,tmp.S3$Year)

#     2013 2014 2015 2016 2017 2018
# S1     0    5   22   66    0    0
# S2     0    0    5   22   68    0
# S3     0    0    0    5   25   93
# S4     0    0    0    0    0    3
```

5. Finally, to decide upon which dataset, I calculate the $\mathcal{A}$-values for the possible MET datasets. As an example, the R code below calculates the $\mathcal{A}$-values for each of the four selection decisions using the 'final' dataset. Note that they match the values on Figure 7.1.

```
final.df$Environment <- factor(final.df$Environment)
final.df$Gkeep <- factor(final.df$Gkeep)
final.df$Year <- factor(final.df$Year)

final.sv <- asreml(y ~ Environment,
random =~ vm(Gkeep,ped.giv) + idv(Year):vm(Gkeep,ped.giv) +
idv(Environment):vm(Gkeep,ped.giv) +
ide(Gkeep) + Year:ide(Gkeep) + Environment:ide(Gkeep),
residual=~idv(units), data=final.df, start.values=T,
na.action = na.method(x='include'), maxit=1)

gams <- final.sv$vparameters.table
gams$Value <- c(0.28/abar, 0.18/abar, 0.76/abar,
0.07, 0.04, 0.19,1,1)
gams$Constraint <- 'F'

final.asr <- asreml(y ~ Environment,
random =~ vm(Gkeep,ped.giv) + idv(Year):vm(Gkeep,ped.giv) +
idv(Environment):vm(Gkeep,ped.giv) +
ide(Gkeep) + Year:ide(Gkeep) + Environment:ide(Gkeep),
residual=~idv(units), data=final.df,
R.param=gams, G.param=gams,
na.action = na.method(x='include'), maxit=1,workspace='1gb')

final.s1.pvs <- predict(final.asr, classify='Gkeep',
only=c("vm(Gkeep, ped.giv)","ide(Gkeep)"),
maxit=1, levels=vars.s1,pworkspace='4gb')
(final.A.s1 <- final.s1.pvs$avsed^2)
#  overall
#0.4833488

final.s2.pvs <- predict(final.asr, classify='Gkeep',
only=c("vm(Gkeep, ped.giv)","ide(Gkeep)"),
maxit=1, levels=vars.s2,pworkspace='4gb')
(final.A.s2 <- final.s2.pvs$avsed^2)
#  overall
#0.3728789
```

```
final.s3.pvs <- predict(final.asr, classify='Gkeep',
only=c("vm(Gkeep, ped.giv)","ide(Gkeep)"),
maxit=1, levels=vars.s3,pworkspace='4gb')
(final.A.s3 <- final.s3.pvs$avsed^2)
#  overall
#0.2990924

final.s4.pvs <- predict(final.asr, classify='Gkeep',
only=c("vm(Gkeep, ped.giv)","ide(Gkeep)"),
maxit=1, levels=vars.s4,pworkspace='4gb')
(final.A.s4 <- final.s4.pvs$avsed^2)
#  overall
#0.1991485
```

This information is then compiled into the CG spreadsheet, and discussed with the breeder in terms of the composition of the MET datasets. Discussions may cover varieties with unknown origins, ensuring that the MET datasets contain as much data as feasible for the key test varieties of interest, and estimating the time required for the corresponding MET analyses.

The final Durum MET dataset, as demonstrated in the earlier parts of this chapter and detailed in Smith et al. (2021a), has trials in data bands A-F (see Table 7.1), with 30 environments, 97 trials, and 7628 varieties. This MET dataset has been analysed in Chapter 8, to demonstrate analyses including pedigree information.

7.4 Concluding remarks

In this chapter I have demonstrated how to create MET datasets using the CG methodology of Smith et al. (2021a). This methodology has filled a gap in the literature by providing a systematic approach for the construction of MET datasets for selection in plant breeding programs. The approach for increasing the amount of data available for the varieties under consideration is simple and straightforward, with only a few steps. To quantify different MET datasets, we employ from model-based design theory the $\mathcal{A}$-optimality criterion as described in Section 7.1. When applied to the Durum dataset (see Chapter 4), we clearly demonstrate the superiority of the MET datasets generated

using the CG approach, especially when contrasted to more frequently used strategies. The information benefits were connected with both direct and indirect information collected from trials where genetically similar varieties were grown. We have demonstrated that by sequentially building the MET dataset using data bands, we may create MET datasets that capture as much information as feasible for the varieties of interest.

The final MET dataset derived in Section 7.2.3 has been used in Chapter 8 to demonstrate a MET analysis with pedigree information. I note however, that combining trials using the CG approach may result in datasets with poor variety connectivity between environments. Whilst one of the advantages of using FALMM (as discussed in Chapter 1) is the ability to handle unbalanced data, there has been concern about the reliability of the estimated genetic variance parameters when variety connectivity is poor. I have already shown in Chapter 6 that variety connectivity influences the reliability of genetic effects, it is also well-known that poorly estimated genetic variance parameters will result in a reduction in genetic gain (Sales & Hill, 1976b,a). The $\mathcal{A}$-value approach does not take this into account since the variance parameters are assumed known. In Chapter 9 a formal information based diagnostic using the $\mathcal{D}$-optimality criterion will be developed for this purpose. This may therefore be applied jointly with the $\mathcal{A}$-value approach in order to balance variety information and reliability of variance parameter estimation in the search for an optimum MET dataset.

Chapter 8

Statistical analysis of the Durum dataset

This chapter uses the Durum dataset originally introduced in Chapter 4 to demonstrate the statistical procedures described in Chapter 2, with analyses including pedigree information. In order to construct a suitable dataset, the final dataset was constructed using the contemporary group (CG) methodology (see Chapter 7). Smith et al. (2021a) demonstrated that this approach clearly showed superiority of the MET datasets constructed in comparison to traditional techniques. The information gains were linked with both direct information, acquired from trials in which the varieties of interest were grown, and indirect information derived from trials in which genetically similar varieties were grown. The CG method is also shown to reduce the influence of selection-based bias on datasets produced from a single year of data and affected by seasonally exceptional years.

The analyses in this chapter differ from those provided in Chapter 5 for the Oat dataset in that I use pedigree information to partition VE effects into additive and non-additive (residual) VE effects (Oakey et al., 2007). The superiority of these models has been widely demonstrated, as evidenced in Oakey et al. (2007); Beeck et al. (2010).

This chapter is arranged as follows: Section 8.1 contains some general information about the Numerator Relationship Matrix (NRM), which is used in the analysis to allow for related variety effects. Following this a co-located pedigree trial analysis is illustrated in

Section 8.2, and then a summary of all single environment pedigree analyses is provided in Section 8.3. Then in Section 8.4 the MET analysis of the full Durum dataset (see Chapter 7) is presented. Finally, in Section 8.5 I have concluding remarks.

8.1 Numerator relationship matrix

In this chapter I now assume related variety effects, which are accommodated in the statistical model through the Numerator Relationship Matrix (NRM), commonly known as the $\boldsymbol{A}$ matrix. This is a symmetric matrix that represents the genetic relationships between varieties, which assumes inheritance laws for correlated genetic (additive) effects. The NRM is built from a pedigree file, which is a structured representation of an individual's ancestral links. The inclusion of this information in the LMM allows for the links between varieties both within and across environments. The elements of $\boldsymbol{A} = \{a_{ij}\}$ are given as

$$a_{ii} = 1 + F_i$$
$$a_{ij} = 2f_{ij}$$

where F_i represents the inbreeding coefficient, which represents twice the probability that two alleles at a given locus are identical by descent, or in other words, the extent an individual is more likely to be homozygous rather than heterozygous because of related parents; and f_{ij} is the coefficient of parentage between varieties i and j.

In practice, the inverse of the NRM ($\boldsymbol{A}^{-1}$) is preferred in the LMM due to it computational properties given that it is generally more sparse. In this thesis, the creation, manipulation and calculation of the pedigree file and $\boldsymbol{A}^{-1}$ are constructed using the `pedicure` `R` package (Butler, 2016).

8.2 Spatial analysis of a co-located trial with pedigree information

To illustrate the spatial co-located trial analysis with pedigree information, the S4 (16S4BRZ) and S3 trials (16S3BRZ) sown in Breeza in 2016, with the environment name '2016-Breeza' is used. They were both sown and managed similarly, with the S4

trial sown above the S3 trial as shown in Figure 4.8. The trait of interest is grain yield which is measured in tonnes per hectare (t/ha), with plot yields ranging from ~ 2.6 to 5.6 t/ha, with one missing plot yield. The trials will henceforth be numbered as trials 1 (S3) and 2 (S4). Each trial was sown into rectangular arrays with $c = 12$ columns and $r_1 = 15$ or $r_2 = 16$ rows, with $n_1 = 180$ and $n_2 = 192$ ($\Sigma_{j=1}^{2} n_j = 372$) plots respectively. There are $m_1 = 60$ and $m_2 = 96$ varieties with 152 unique varieties across trials, with four check varieties in common between trials. Both trials were designed as a RCB with $b_1 = 3$ and $b_2 = 2$ replicates, with replicate blocks aligned with four and six columns for trials 1 and 2 respectively.

The pedigree information contained 254 records, relating to the 152 varieties grown and 102 ancestral varieties that were not grown. This information is used to form a numerator relationship matrix (NRM, see Section 8.1), denoted $\boldsymbol{A}$. This is a $(m \times m)$ matrix where $m = 254$.

8.2.1 Statistical analysis

The methods outlined in Chapter 2, with extensions for the co-located aspect as described in Chapter 5, were used for the analysis of 2016-Breeza. An initial model (M1) comprises terms reflecting the co-location of the trials within environment, the trial designs and an AR1×AR1 process for the errors was used. With the allowance of the pedigree information the vector of variety effects ($\boldsymbol{u_g}$), which is denoted as the total variety effects, can be decomposed into additive ($\boldsymbol{u_a}$) and non-additive ($\boldsymbol{u_e}$) variety effects, so that

$$\boldsymbol{u_g} = \boldsymbol{u_a} + \boldsymbol{u_e} \tag{8.1}$$

The LMM labelled Model 1 (M1) for the (372×1) data vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_{372})^\top$ may be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g}(\boldsymbol{u_a} + \boldsymbol{u_e}) + \boldsymbol{Z_t u_t} + \boldsymbol{Z_b u_b} + \boldsymbol{e} \tag{8.2}$$

where $\boldsymbol{\tau}$ is the overall mean with associated design matrix $\boldsymbol{X}$; $\{\boldsymbol{u_a}, \boldsymbol{u_e}\}$ are the (254×1) vectors of random additive and non-additive variety effects respectively with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_t}$ is the (2×1) vector of random trial effects with associated design matrix $\boldsymbol{Z_t}$; $\boldsymbol{u_b}$ is the (5×1) vector of random replicate block effects for each trial with

associated design matrix $\boldsymbol{Z_b}$; and $\boldsymbol{e}$ is the combined (372×1) vector of errors across both trials. It is assumed that

$$\begin{bmatrix} \boldsymbol{u_a} \\ \boldsymbol{u_e} \\ \boldsymbol{u_t} \\ \boldsymbol{u_b} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_a} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_e} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G_t} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where

- $\boldsymbol{G_a} = \sigma_a^2 \boldsymbol{A}$ (Additive genetic variance matrix)
- $\boldsymbol{G_e} = \sigma_e^2 \boldsymbol{I}_{254}$ (Non-additive genetic variance matrix)
- $\boldsymbol{G_t} = \sigma_t^2 \boldsymbol{I}_2$ (Trial variance matrix)
- $\boldsymbol{G_b} = \sigma_b^2 \boldsymbol{I}_5$ (Replicate block variance matrix)
- $\boldsymbol{\Sigma} = \oplus_{j=1}^{2} \sigma_j^2 \boldsymbol{\Sigma}_{c_j}(\rho_{c_j}) \otimes \boldsymbol{\Sigma}_{r_j}(\rho_{r_j})$ (Error variance matrix)

where $\boldsymbol{\Sigma}_{c_j}$ and $\boldsymbol{\Sigma}_{r_j}$ are the (12×12) and $(r_j \times r_j)$ correlation matrices for columns and rows respectively. Similar to the methods outlined in Chapter 5 for the analysis of the Oat dataset, I use the 'equal constrained' approach of Jordan (2022), where the spatial variance and spatial correlations for the two trials are constrained to be equal.

The associated `ASReml-R` code for M1 is given as

```
M1.asr <- asreml(yield ~ 1,
random =~ vm(Variety,breeza16.ainv) + ide(Variety) +
Trial + Trial:ColRep,
residual =~ dsum(~ar1(Column):ar1(Row)|Trial),
vcc=Mcc, data=breeza16.df)
```

where `yield` is the data vector of plot yields (t/ha); `1` denotes the overall mean; `Trial`, `Trial:ColRep` are terms representing the random trial and within trials replicate block effects respectively. The errors are modelled using an AR1×AR1 structure for each trial as given by `dsum(ar1(Column):ar1(Row)|Trial)`; `Mcc` is the matrix of constraints (see Table 5.4 for an example); and `breeza16.df` is the data object.

The variety effects as shown by Equation 8.1 have been partitioned into additive and

non-additive effects. The term `vm(Variety,breeza16.ainv)` relates to the additive effects $\boldsymbol{u_a}$ which uses `breeza16.ainv` which is the (sparse form) $\boldsymbol{A}^{-1}$ for the 2016-Breeza environment. The `vm()` is `ASReml` syntax referring to a known relationship structure, with the form `vm(obj, source)` where, `obj=Variety, source=breeza16.ainv`. The term `ide(Variety)` relates to the non-additive effects $\boldsymbol{u_e}$, where `ide()` creates a partner term to `vm()` with the same levels.

Following the same procedures as shown in Chapter 5 for the analysis of the Oat dataset, random column and row effects were added with a significant ($p < 0.001$) increase in ℓ_R after their inclusion. The adjusted M1 is denoted as Model 2 (M2). There were no outliers identified. Table 8.1 displays the REML estimates of the variance parameters for M2.

In relevance to this thesis, I use $\hat{\sigma}^2_a = 0.054$ and $\hat{\sigma}^2_e = 0.074$ from Table 8.1. Thus, $\widehat{\text{var}}(\boldsymbol{u_g}) = \hat{\sigma}^2_a\boldsymbol{A} + \hat{\sigma}^2_e\boldsymbol{I}_{254}$ and for an individual variety I have $\widehat{\text{var}}(u_{g_i}) = a_{ii}\hat{\sigma}^2_a + \hat{\sigma}^2_e$ where a_{ii} is the i^{th} diagonal element of $\boldsymbol{A}$. For the 152 varieties with phenotypic data in 2016-Breeza, the values of a_{ii} range from 1.750 to 1.998 with a mean of $\bar{a} = 1.879$. I then define an (average) total genetic variance for the varieties grown in this environment as $\hat{\sigma}^2_g = \bar{a}\hat{\sigma}^2_a + \hat{\sigma}^2_e = 0.175$. Finally, I compute the percentage of total genetic variance accounted for by the additive component as $\bar{a}\hat{\sigma}^2_a/\hat{\sigma}^2_g \times 100\% = 57.6\%$.

Table 8.1: REML estimates of variance parameters for the co-located trial analysis for 2016-Breeza for Model M2.

Term in model	REML estimate	Parameter
Trial	0.000	σ_t^2
Trial:ColRep	0.001	σ_b^2
Trial:Column	0.026	σ_c^2
Trial:Row	0.002	σ_r^2
vm(Variety, breeza16.ainv)	0.054	σ_a^2
ide(Variety)	0.074	σ_e^2
Trial_16S3BRZ!R	0.064	σ_1^2
Trial_16S3BRZ!Column!cor	0.060	ρ_{c_1}
Trial_16S3BRZ!Row!cor	0.211	ρ_{r_1}
Trial_16S4BRZ!R	0.064	σ_2^2
Trial_16S4BRZ!Column!cor	0.060	ρ_{c_2}
Trial_16S4BRZ!Row!cor	0.211	ρ_{r_2}

8.2.1.1 Empirical best linear unbiased predictions

From the analysis of M2, the variety EBLUPs for the total variety effects $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ (see Equation 8.1) are obtained. These are presented in Table 8.2 for the first and last four varieties in alphabetical order, along with their PEV and r^2. I also show their additive ($\tilde{\boldsymbol{u}}_{\boldsymbol{a}}$) and non-additive ($\tilde{\boldsymbol{u}}_{\boldsymbol{e}}$) EBLUPs. Figure 8.1 presents the relationship between $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ and the raw means for individual varieties, where the dashed line represents a 1-1 relationship. The reasonably high reliabilities (mean of 0.82) results in little shrinkage so that the points centre around a line with a slope close to 1. The r^2 for the total variety effects ranged from 0.70 to 0.89 with $\bar{R} = 0.82$.

Table 8.2: 2016-Breeza: $\tilde{\boldsymbol{u}}_{\boldsymbol{a}}$, $\tilde{\boldsymbol{u}}_{\boldsymbol{e}}$, $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$, prediction error variance (PEV) for $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$, and r^2 estimates for the first and last four varieties in alphabetical order.

Variety	$\tilde{\boldsymbol{u}}_{\boldsymbol{a}}$	$\tilde{\boldsymbol{u}}_{\boldsymbol{e}}$	$\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$	PEV	r^2
10TD022*3X-103	-0.044	-0.036	-0.080	0.053	0.699
10TD022*3X-77	-0.005	0.022	0.017	0.027	0.847
10TD022*3X-96	-0.042	-0.033	-0.075	0.035	0.800
10TD022*3X-98	-0.087	-0.099	-0.186	0.035	0.802
⋮	⋮	⋮	⋮	⋮	⋮
UAD1151056	-0.416	-0.308	-0.724	0.026	0.851
UAD1151096	-0.098	-0.072	-0.170	0.026	0.851
UAD1151097	-0.159	-0.118	-0.277	0.026	0.851
UAD1152020	-0.166	-0.122	-0.288	0.026	0.852

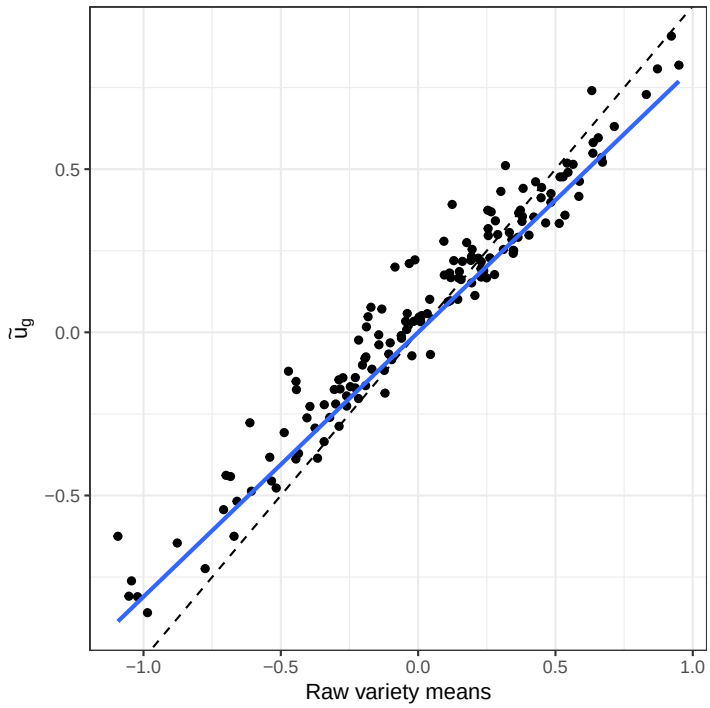

Figure 8.1: 2016-Breeza: $\tilde{\boldsymbol{u}}_{\boldsymbol{g}}$ against raw centred varietal means. Dashed line represents a 1-1 line. A regression line through the points is shown by the solid blue line.

8.3 Summary of results from all single environment analyses

I now consider the analyses for all 30 environments in the Durum dataset. The same procedures as shown in the previous section for the analysis of 2016-Breeza were used. There were 12 outliers identified and removed from the Durum dataset. A summary of all unique models fitted to the non-genetic effects is presented in Table 8.3. These models include the non-genetic effects of `ColRep` which is the random model term for the replicate block effect; `Column` and `Row` are random model terms for column and row effects; the spatial models for all environments was an AR1×AR1 for spatial correlation in both directions. The inclusion of the non-genetic model terms are denoted by the tick marks in Table 8.3 with the four unique models denoted as Models M1.d, M2.d, M3.d, and M4.d. Note that for co-located trials, a random `Trial` term is also fitted and `ColRep, Column` and `Row` denote replicate block, column, and row effects within trials. M4.d was the most commonly selected non-genetic model, with 13 environments, which include random effects for `ColRep, Column` and `Row`.

Table 8.4 provides the REML parameter estimates for the genetic, non-genetic, and error variance parameters. The $\bar{a}$ is shown for each environment, which is the mean of the diagonal elements for $\boldsymbol{A}$ for the subset of varieties grown in each environment, the percentage of additive variance (ADD%) is also provided. The average ADD% across all environments was 58%, but note that this increases to ∼80% when the MET analysis is used (see Section 8.4).

Table 8.3: Set of unique models fitted to the non-genetic effects for the 30 environments in the final Durum dataset. `ColRep` is the random model term for replicate block effects; `Column` and `Row` are random model terms for column and row effects; the spatial model AR1×AR1 denotes spatial correlation in both directions.

Model	`ColRep`	`Column`	`Row`	Spatial	Environments
M1.d	✓	-	-	AR1xAR1	4
M2.d	✓	✓	-	AR1xAR1	9
M3.d	✓	-	✓	AR1xAR1	4
M4.d	✓	✓	✓	AR1xAR1	13

Note that for co-located trials, a random `Trial` term is also fitted and `ColRep, Column` and `Row` denote replicate block, column, and row effects within trials.

Table 8.4: REML estimates for genetic $\{\hat{\sigma}^2_a, \hat{\sigma}^2_e, \hat{\sigma}^2_g\}$, non-genetic $\{\hat{\sigma}^2_t, \hat{\sigma}^2_b, \hat{\sigma}^2_c, \hat{\sigma}^2_r\}$, and error $\{\hat{\sigma}^2, \hat{\rho}_c, \hat{\rho}_r\}$ variance parameters for the analysis of the Durum environments. $\bar{a}$ shows the mean of the diagonal elements of $\boldsymbol{A}$ for the subset of varieties grown in each environment. ADD% shows the percentages of additive genetic variance to the total genetic variance. The final column represents the non-genetic model chosen (M1.d-M4.d). B denotes variance parameter is estimated at the boundary value (0 for variance components). Horizontal dashed lines separate years (2013-2018). The environment 2016-Breeza illustrated in this chapter for analysis is represented by the grey row.

	Genetic			Non-genetic				Error					
Environment	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_e$	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_t$	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_c$	$\hat{\sigma}^2_r$	$\hat{\sigma}^2$	$\hat{\rho}_c$	$\hat{\rho}_r$	$\bar{a}$	ADD%	Model
2013-Breeza	0.106	0.038	0.236	0.007	B	0.002	0.001	0.133	0.278	0.325	1.877	84.0	M4.d
2014-Breeza	0.070	0.011	0.143	B	0.021	0.005	0.007	0.083	0.596	0.429	1.881	92.1	M4.d
2014-Edgeroi	0.027	0.010	0.060	0.027	0.006	0.002		0.057	0.426	0.629	1.877	83.0	M2.d
2014-Tworth	0.037	0.001	0.071	0.118	0.031	0.003		0.056	0.524	0.370	1.876	98.0	M2.d
2015-Breeza	0.051	B	0.097	0.008	B	0.007	0.046	0.147	0.129	-0.073	1.882	100	M4.d
2015-Edgeroi	0.033	0.035	0.097	0.121	0.001	0.006	0.000	0.031	0.169	0.618	1.877	64.0	M4.d
2015-Nstar	0.019	0.086	0.122		0.059	0.031		0.030	0.045	0.012	1.882	29.6	M2.d
2015-Tworth	0.012	0.019	0.042	0.155	0.006	0.002	0.004	0.078	0.131	0.301	1.916	56.0	M4.d
2016-Breeza	0.054	0.074	0.175	B	0.001	0.026	0.002	0.064	0.060	0.211	1.878	57.6	M4.d
2016-Edgeroi	0.006	0.218	0.229		0.004			0.026	0.357	0.035	1.883	4.8	M1.d
2016-Gurley	0.056	0.189	0.294		0.012		0.026	0.069	-0.112	0.461	1.883	35.6	M3.d
2016-Nstar	0.086	0.034	0.196	0.031	0.005	0.015	0.004	0.096	0.052	0.630	1.878	82.5	M4.d
2016-Tworth	0.071	0.096	0.232	1.258	0.014	0.027	0.013	0.157	0.353	0.399	1.923	58.6	M4.d
2017-Blbgra	0.038	0.021	0.092		0.028	0.002	0.028	0.038	0.193	0.119	1.881	77.0	M4.d
2017-Breeza	0.041	0.183	0.261	B	0.014	0.007		0.158	0.269	0.379	1.907	29.9	M2.d
2017-Bribbaree	0.009	0.026	0.042		B	0.020		0.140	0.031	0.829	1.881	38.8	M2.d
2017-Coonamble	0.018	0.024	0.057		B			0.086	0.247	0.545	1.881	58.4	M1.d
2017-Edgeroi	0.005	0.009	0.020		0.009			0.071	-0.056	0.732	1.881	51.8	M1.d
2017-Garah	B	0.020	0.020		0.028		0.004	0.079	0.005	0.434	1.881	0	M3.d
2017-Gurley	0.010	0.069	0.087		B	0.011	0.005	0.081	-0.012	0.722	1.881	20.6	M4.d
2017-Nstar	0.010	0.014	0.034	0.030	0.005		0.004	0.069	0.338	0.288	1.907	58.3	M3.d
2017-Tworth	0.021	0.017	0.058	0.341	0.021	0.032	0.012	0.142	0.283	0.469	1.932	71.1	M4.d
2017-Westmar	0.008	0.020	0.035		B	0.002		0.019	0.145	0.404	1.881	44.5	M2.d
2018-Blbgra	0.005	0.001	0.010		0.001	0.001		0.016	0.261	0.717	1.916	94.6	M2.d
2018-Breeza	0.136	0.046	0.304	0.171	0.017	0.031	0.023	0.095	0.062	0.064	1.892	85.0	M4.d
2018-Coonamble	0.001	0.002	0.004		0.010	0.003		0.064	-0.197	0.084	1.916	42.6	M2.d
2018-Gurley	0.001	0.021	0.023		B		0.001	0.153	0.065	0.888	1.927	6.4	M3.d
2018-Moree	0.003	0.022	0.028		B	0.041		0.053	0.488	0.839	1.916	19.5	M2.d
2018-Trangie	0.006	B	0.012		B			0.048	0.149	0.596	1.916	100	M1.d
2018-Tworth	0.015	B	0.029	0.472	0.170	0.026	0.006	0.110	0.189	0.767	1.930	100	M4.d

8.4 One-stage multi-environment trial analysis

I now consider the MET analysis of the Durum dataset using the methods described in Chapter 2. This dataset consists of $p = 30$ environments, $t = 97$ trials, and $n = 21660$ plots. The pedigree information associated with the Durum dataset contained $m = 7628$ varieties, relating to all 6951 varieties grown across all environments and 677 ancestral varieties that were not grown.

I now let $\boldsymbol{y}_j$ be the $(n_j \times 1)$ vector of yield data for environment $j(= 1, 2, \ldots, 30)$, and let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_{30}^\top)^\top$ be the combined (21660×1) vector of yield data across environments. The LMM for $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}(\boldsymbol{u_a} + \boldsymbol{u_e}) + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{8.3}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects which comprise solely of separate environment means, with associated design matrix $\boldsymbol{X}$; $\{\boldsymbol{u_a}, \boldsymbol{u_e}\}$ are (228840×1) vectors of random additive and non-additive VE effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_p}$ is a vector of random non-genetic (peripheral) effects consisting of trial, replicate blocks, column and row effects as established by the models in the single environment analyses (see Table 8.4), with associated design matrix $\boldsymbol{Z_p}$; and $\boldsymbol{e}$ is the (21660×1) vector of errors. It is assumed that

$$\begin{bmatrix} \boldsymbol{u_a} \\ \boldsymbol{u_e} \\ \boldsymbol{u_p} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_a} \otimes \boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_e} \otimes \boldsymbol{I}_{7628} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G_p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

where $\{\boldsymbol{G_a}, \boldsymbol{G_e}\}$ are (30×30) symmetric positive (semi)-definite matrices known as the between environments additive and non-additive genetic variance matrices respectively. Similar to the Oat MET analysis described in Chapter 5, these matrices are modelled using (separate) FA structures of order $\{k_a, k_e\}$ respectively given as

$$\boldsymbol{G_a} = (\boldsymbol{\Lambda_a}\boldsymbol{\Lambda_a}^\top + \boldsymbol{\Psi_a}) \qquad\qquad \boldsymbol{G_e} = (\boldsymbol{\Lambda_e}\boldsymbol{\Lambda_e}^\top + \boldsymbol{\Psi_e})$$

where $\{\boldsymbol{\Lambda_a}, \boldsymbol{\Lambda_e}\}$ are the $(30 \times k_a)$ and $(30 \times k_e)$ matrices of additive and non-additive loadings respectively; $\{\boldsymbol{\Psi_a}, \boldsymbol{\Psi_e}\}$ are the (30×30) diagonal matrices where the diagonal elements are the specific additive and non-additive environment variances respectively;

and $\boldsymbol{A}$ is the (7628 × 7628) NRM for the Durum dataset. It is then assumed for the additive and non-additive VE effects that

$$\boldsymbol{u_a} = (\boldsymbol{\Lambda_a} \otimes \boldsymbol{I}_{7628})\boldsymbol{f_a} + \boldsymbol{\delta_a} \qquad \boldsymbol{u_e} = (\boldsymbol{\Lambda_e} \otimes \boldsymbol{I}_{7628})\boldsymbol{f_e} + \boldsymbol{\delta_e}$$

where $\{\boldsymbol{f_a}, \boldsymbol{f_e}\}$ are the $(7628k_a \times 1)$ and $(7628k_e \times 1)$ vectors of additive and non-additive variety scores respectively, and $\{\boldsymbol{\delta_a}, \boldsymbol{\delta_e}\}$ are the (228840 × 1) vectors of additive and non-additive VE lack of fit effects respectively. It is assumed that

$$\begin{aligned} \text{var}(\boldsymbol{f_a}) &= \boldsymbol{I}_{k_a} \otimes \boldsymbol{A} & \text{var}(\boldsymbol{f_e}) &= \boldsymbol{I}_{k_e} \otimes \boldsymbol{I}_{7628} \\ \text{var}(\boldsymbol{\delta_a}) &= \boldsymbol{\Psi_a} \otimes \boldsymbol{A} & \text{var}(\boldsymbol{\delta_e}) &= \boldsymbol{\Psi_e} \otimes \boldsymbol{I}_{7628} \end{aligned}$$

so that

$$\text{var}(\boldsymbol{u_a}) = (\boldsymbol{\Lambda_a}\boldsymbol{\Lambda_a}^\top + \boldsymbol{\Psi_a}) \otimes \boldsymbol{A} \qquad \text{var}(\boldsymbol{u_e}) = (\boldsymbol{\Lambda_e}\boldsymbol{\Lambda_e}^\top + \boldsymbol{\Psi_e}) \otimes \boldsymbol{I}_{7628}$$

It should also be noted, that the total VE effects $\boldsymbol{u_g}$ is the simple addition of $\boldsymbol{u_a}$ and $\boldsymbol{u_e}$ (see Equation 8.1). These can be further simplified to

$$\boldsymbol{u_a} = \boldsymbol{\beta_a} + \boldsymbol{\delta_a} \qquad \boldsymbol{u_e} = \boldsymbol{\beta_e} + \boldsymbol{\delta_e}$$

where $\boldsymbol{\beta_a} = (\boldsymbol{\Lambda_a} \otimes \boldsymbol{I}_{7628})\boldsymbol{f_a}$ and $\boldsymbol{\beta_e} = (\boldsymbol{\Lambda_e} \otimes \boldsymbol{I}_{7628})\boldsymbol{f_e}$ are the so called additive/non-additive VE regression components. Finally, I denote the full genetic model in which an FAk_a has been used for $\boldsymbol{u_a}$ and an FAk_e has been used for $\boldsymbol{u_e}$ as FAk_a,k_e.

For the non-genetic effects it is assumed that

$$\begin{aligned} \boldsymbol{G_p} &= \oplus_{k=1}^{v} \sigma^2_{p_k} \boldsymbol{I}_{q_k} \\ \boldsymbol{\Sigma} &= \oplus_{j=1}^{97} \sigma^2_j \boldsymbol{\Sigma}_{c_j}(\rho_{c_j}) \otimes \boldsymbol{\Sigma}_{r_j}(\rho_{r_j}) \end{aligned}$$

where v is the number of components in $\boldsymbol{u_p}$ and q_k is the number of effects in (length of) $\boldsymbol{u_{p_k}}$; and the errors are modelled using AR1×AR1 spatial structures. The non-genetic and error terms fitted are those established for the single environment analyses, as indicated in Table 8.4.

As an example, the `ASReml-R` code to fit the LMM with an FA1,1 model for $\boldsymbol{u_a}$ and $\boldsymbol{u_e}$ is given below

```
rr1rr1.asr <- asreml(Yield ~ Environment,
random =~ rr(Environment,1):vm(Variety,durum.ainv) +
diag(Environment):vm(Variety,durum.ainv) +
rr(Environment,1):ide(Variety) + diag(Environment):ide(Variety) +
at(Environment,col.env):Trial +
at(Environment,crep):Trial:ColRep +
at(Environment,rcol):Trial:Column +
at(Environment,rrow):Trial:Row,
residual=~dsum(~ar1(Column):ar1(Row)|Trial),data=final.df,
vcc=Mcc, na.action = na.method(x='include'))
```

where `col.env` is a vector of environment names containing co-located trials; `crep, rcol, rrow` are vectors of environment names containing those fitted with `ColRep, Column, Row` random effects respectively; `Mcc` is the matrix which constrains spatial parameters equal for the co-located environments; and `final.df` is the dataset containing the full data object for the Durum dataset.

The FA1,1 model for $\{\boldsymbol{u_a}, \boldsymbol{u_e}\}$ have both been fitted by splitting them into the two constituent parts, namely the regression part associated with $\{\boldsymbol{\beta_a}, \boldsymbol{\beta_e}\}$, and the lack of fit part associated with $\{\boldsymbol{\delta_a}, \boldsymbol{\delta_e}\}$. The terms `rr(Environment,1):vm(Variety,durum.ainv)` and `rr(Environment,1):ide(Variety)` relate to $\boldsymbol{\beta_a}$ and $\boldsymbol{\beta_e}$ respectively, where I again fit reduced rank variance structures of order 1 for the environment dimension, namely

$$\text{var}(\boldsymbol{\beta_a}) = \boldsymbol{\Lambda_a}\boldsymbol{\Lambda_a}^\top \otimes \boldsymbol{A} \qquad \text{var}(\boldsymbol{\beta_e}) = \boldsymbol{\Lambda_e}\boldsymbol{\Lambda_e}^\top \otimes \boldsymbol{I}_{7628}$$

where $\{\boldsymbol{\Lambda_a}, \boldsymbol{\Lambda_e}\}$ are (30×1) matrices of additive and non-additive loadings respectively. The terms `diag(Environment):vm(Variety,durum.ainv)` and `diag(Environment):ide(Variety)` relate to $\boldsymbol{\delta_a}$ and $\boldsymbol{\delta_e}$ respectively, which fit diagonal variance structures for the environment dimension, that is

$$\text{var}(\boldsymbol{\delta_a}) = \boldsymbol{\Psi_a} \otimes \boldsymbol{A} \qquad \text{var}(\boldsymbol{\delta_e}) = \boldsymbol{\Psi_e} \otimes \boldsymbol{I}_{7628}$$

where $\{\boldsymbol{\Psi_a}, \boldsymbol{\Psi_e}\}$ are the (30×30) diagonal matrices for the additive and non-additive specific variances respectively.

A series of FALMM were fit to the data with increasing numbers of additive factors (values of k_a) and non-additive factors (values of k_e), as shown in Table 8.5. The Akaike information criteria (AIC) showed significant improvements of successive models up to and including an FA4,3 model (see also Figure 8.2) and hence the FA4,3 model was chosen as the final model. The FA4,3 model variance accounted for (VAF%) by the factors was 92.1%, 86.6%, and 90.7% for additive, non-additive, and total VE effects, respectively.

Table 8.5: Summary of number of variance parameters, residual log-likelihood (ℓ_R), and AIC for the eight variance models fitted to the 30 environments in the Durum dataset. Grey row corresponds to the model with the smallest AIC.

Model	k_a	k_e	Parameters	ℓ_R	AIC
diag	-	-	234	11037.51	-21607.02
FA1,1	1	1	294	11470.71	-22353.42
FA2,1	2	1	323	11582.12	-22518.24
FA2,2	2	2	352	11645.82	-22587.64
FA3,2	3	2	380	11696.16	-22632.31
FA3,3	3	3	408	11726.49	-22636.99
FA4,3	4	3	435	11755.17	-22640.33
FA5,3	5	3	461	11777.84	-22633.67

Table 8.6 presents summaries of the environment information from the FA4,3 model fitted to the additive and non-additive VE effects: REML estimates of loadings for each factor, specific variances, genetic variances, and percentage variance accounted for by the four and three additive and non-additive factors respectively, as well as across all seven factors. On an individual environment basis, all environments had greater than 50% explained by the regression part of the FA model, and 24 environments had greater than 80% explained. The average percentage additive variance across environments was 74.9%. This was computed using $\bar{a} = 1.905$ which was the mean of the diagonal elements of $\boldsymbol{A}$ for all 6951 varieties with phenotypic data.

The REML estimates of the loadings and specific variances can be used to form the REML estimates of the additive and non-additive between environments genetic variance matrices, denoted $\{\hat{\boldsymbol{G}}_{\boldsymbol{a}}, \hat{\boldsymbol{G}}_{\boldsymbol{e}}\}$, respectively. I also define an (average) estimated

total genetic variance matrix as $\hat{\boldsymbol{G}}_{\boldsymbol{g}} = \bar{a}\hat{\boldsymbol{G}}_{\boldsymbol{a}} + \hat{\boldsymbol{G}}_{\boldsymbol{e}}$. Note that the diagonals of this matrix are provided in Table 8.6. The $\hat{\boldsymbol{G}}_{\boldsymbol{g}}$ matrix is transformed to the correlation parametrisation with the resultant $\hat{\boldsymbol{G}}_{\boldsymbol{g}}^{(c)}$ matrix graphically presented in Figure 8.3 by way of a heatmap. The rows and columns of the matrix have been ordered as environments within years. The pairwise between environments genetic correlations ranged between -0.64 and 0.91, with an average of 0.24.

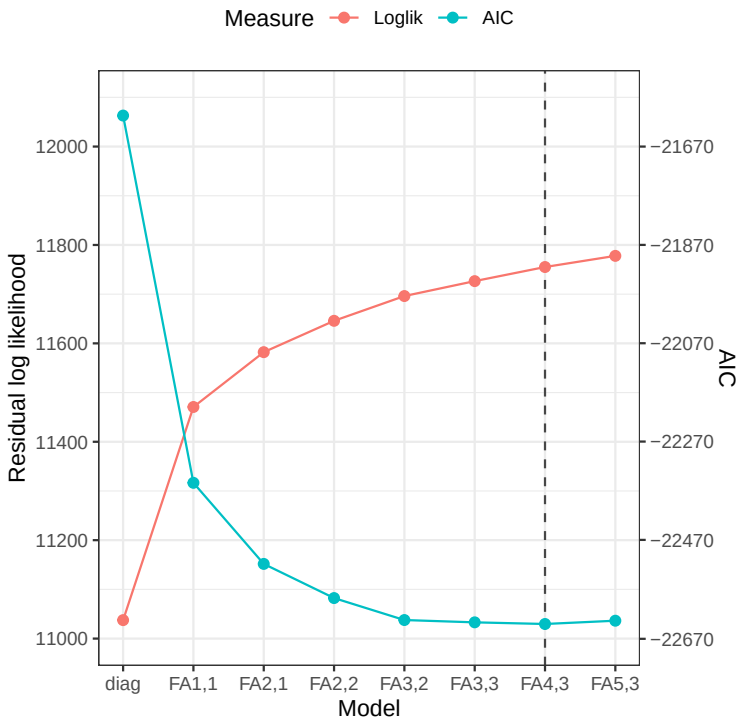

Figure 8.2: Residual log-likelihood (left-hand side y-axis) and AIC (right-hand side y-axis) for each model fitted to the motivating Durum MET dataset. Colours as referenced in the legend. Dotted vertical line represents the FA4,3 model as it has the lowest AIC value.

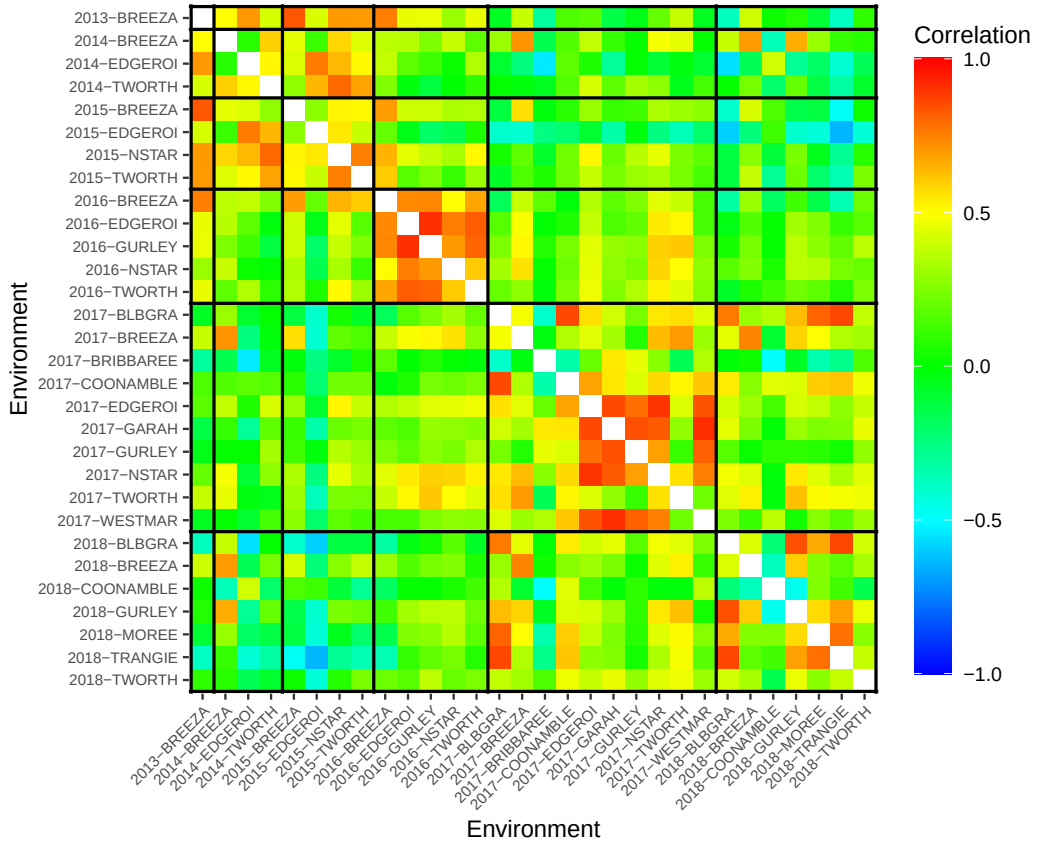

Figure 8.3: Heatmap of $\hat{\boldsymbol{G}}_g^{(c)}$ from the FA4,3 model for the MET analysis of the Durum dataset. Key depicts the correlation colour scale. Boundaries for years are indicated by the black lines (2013 - 2018 inclusive from left to right and top to bottom).

Table 8.6: Summary of environment information from the FA4,3 model fitted to the additive and non-additive VE effects: REML (unrotated) estimates of loadings for each factor, specific variances $\{\hat{\psi}_a, \hat{\psi}_e\}$, genetic variances $\{\hat{\sigma}^2_a, \hat{\sigma}^2_e\}$, and percentage variance accounted for {VAF$_a$%, VAF$_e$%} by the four and three additive and non-additive factors respectively. Along with the genetic variance ($\hat{\sigma}^2_g$) and percentage variance accounted for (VAF$_g$%) for the total VE effects. ADD% shows the percentages of additive variance to the total. Horizontal dashed lines separate years (2013-2018).

	Additive							Non-Additive						Total		
	Environment loadings							Environment loadings								
Environment	1	2	3	4	$\hat{\psi}_a$	$\hat{\sigma}^2_a$	VAF$_a$(%)	1	2	3	$\hat{\psi}_e$	$\hat{\sigma}^2_e$	VAF$_e$(%)	$\hat{\sigma}^2_g$	VAF$_g$(%)	ADD%
2013-Breeza	0.328	0	0	0	0	0.205	100	0.237	0	0	0	0.056	100	0.261	100	78.4
2014-Breeza	0.106	0.223	0	0	0	0.116	100	0.119	0.082	0	0	0.021	100	0.137	100	84.8
2014-Edgeroi	0.106	-0.070	-0.105	0	0	0.052	100	0.115	0.022	-0.049	0	0.016	100	0.068	100	76.3
2014-Tworth	0.083	0.095	-0.128	-0.009	0	0.062	100	0.017	-0.024	-0.077	0	0.007	100	0.068	100	90.0
2015-Breeza	0.186	0.024	-0.001	0.011	0	0.067	100	0.100	-0.059	0.167	0	0.041	100	0.109	100	61.9
2015-Edgeroi	0.103	-0.033	-0.187	-0.100	0.003	0.113	95.1	0.060	0.013	-0.035	0.013	0.018	27.7	0.132	85.7	86.1
2015-Nstar	0.264	0.094	-0.146	0.048	0.011	0.216	90.1	0.030	0.072	-0.093	0	0.015	100	0.231	90.7	93.6
2015-Tworth	0.133	0.037	-0.030	0.015	0.004	0.047	82.1	0.010	-0.063	-0.081	0	0.011	100	0.057	85.4	81.6
2016-Breeza	0.287	0.000	0.045	0.030	0.009	0.178	90.9	-0.051	0.077	0.040	0	0.010	100	0.188	91.4	94.6
2016-Edgeroi	0.199	0.004	0.071	0.097	0	0.103	100	-0.050	0.341	0.100	0	0.129	100	0.232	100	44.4
2016-Gurley	0.234	-0.025	0.143	0.180	0	0.207	100	-0.068	0.269	0.058	0.021	0.102	79.1	0.308	93.1	67.0
2016-Nstar	0.098	0.049	0.047	0.101	0.026	0.097	48.0	-0.015	0.186	0.092	0	0.043	100	0.140	64.1	69.1
2016-Tworth	0.207	-0.050	0.010	0.152	0.014	0.158	82.9	-0.060	0.243	0.013	0.023	0.086	73.3	0.244	79.5	64.8
2017-Blbgra	-0.055	0.062	0.010	0.166	0.001	0.069	96.1	0.103	0.058	-0.031	0	0.015	100	0.083	96.8	82.2
2017-Breeza	0.124	0.246	0.201	0.150	0	0.264	100	0.174	0.064	0.226	0	0.085	100	0.350	100	75.6
2017-Bribbaree	0.014	0.023	0.023	-0.008	0.001	0.005	50.9	-0.160	-0.055	0.034	0	0.030	100	0.035	93.0	14.3
2017-Coonamble	-0.002	0.018	-0.017	0.156	0	0.047	100	0.080	-0.033	-0.043	0	0.009	100	0.057	100	83.5
2017-Edgeroi	0.040	0.052	-0.033	0.103	0.001	0.032	95.1	-0.036	-0.000	0.016	0	0.002	100	0.033	95.4	95.2
2017-Garah	0.018	0.057	-0.012	0.125	0	0.037	100	-0.118	-0.062	0.030	0	0.019	100	0.055	100	66.5
2017-Gurley	0.053	0.035	-0.080	0.146	0	0.060	100	-0.170	-0.051	0.025	0.018	0.050	64.0	0.111	83.7	54.6
2017-Nstar	0.054	0.075	0.001	0.115	0	0.041	100	-0.050	0.025	0.035	0.003	0.007	61.4	0.049	94.4	85.4
2017-Tworth	0.056	0.047	0.081	0.097	0.002	0.043	93.2	0.065	0.050	-0.011	0.016	0.022	30.6	0.066	71.9	66.0
2017-Westmar	0.014	0.026	-0.043	0.134	0	0.039	99.7	-0.070	-0.063	0.092	0	0.017	100	0.057	99.8	69.6
2018-Blbgra	-0.034	0.051	0.021	0.040	0	0.011	100	0.002	0.011	-0.031	0	0.001	100	0.012	100	91.1
2018-Breeza	0.127	0.252	0.151	0.057	0.024	0.247	81.5	0.141	-0.093	-0.001	0.026	0.055	52.3	0.301	76.2	81.9
2018-Coonamble	-0.009	-0.023	-0.017	0.020	0	0.002	100	0.025	0.003	0.025	0	0.001	100	0.004	100	65.1
2018-Gurley	0.005	0.091	0.050	0.050	0	0.025	100	0.013	0.056	-0.081	0	0.010	100	0.035	100	71.8
2018-Moree	-0.038	0.039	0.023	0.081	0.005	0.029	66.5	0.058	0.065	0.003	0	0.008	100	0.037	73.4	79.4
2018-Trangie	-0.044	0.019	0.028	0.053	0	0.011	100	0.019	0.036	-0.027	0	0.002	100	0.014	100	82.1
2018-Tworth	0.020	0.011	0.048	0.064	0.007	0.026	50.3	-0.017	-0.026	-0.044	0	0.003	100	0.029	55.2	90.1

8.4.1 Variety predictions

The EBLUPs of the additive, non-additive and total VE effects for a subset of six varieties and eight environments are shown in Table 8.7. The building block components of $\tilde{\beta}_{ij}$ and $\tilde{\delta}_{ij}$ for the additive and non-additive VE effects are provided, similar to the working example in Chapter 5 for the MET analysis of the Oat dataset. Additionally, the building blocks for the total VE effects are presented (see Equation 8.1). Again, when a variety is not grown in an environment, the EBLUP for the specific non-additive VE effect for $\tilde{\delta}_{eij}$ is 0. However, because the NRM is used in calculation of the variance of the additive specific VE effects, this does not apply to the additive VE effects.

8.5 Concluding remarks

The Durum dataset originally introduced in Chapter 4 is used in this chapter to demonstrate the statistical procedures outlined in Chapter 2, with analyses including pedigree information. Unlike the Oat dataset (see Chapter 4) the final Durum dataset used for analysis in this chapter is appropriately constructed for the selection decisions of interest, as shown in Chapter 7. I provide a single trial analysis based on subsets of this data, followed by a MET analysis based on the entire Durum dataset. I show that when pedigree information is included, the variety effects are partitioned into additive and non-additive variety effects.

Summaries of the results presented in this chapter serve as the motivation and structure for the simulation study presented in Chapter 9, which is used to examine the performance of the new diagnostic also detailed in that chapter.

Table 8.7: Summary set of results for a subset of six varieties in eight environments from the FA4,3 MET analysis of the Durum dataset. Presence shown by tick-marks shows where the varieties were grown. The EBLUPs of additive/non-additive VE effects, that is $\{\tilde{u_{aij}}, \tilde{u_{eij}}\}$ and their building block components $\{\tilde{\beta}_{aij}, \tilde{\beta}_{eij}\}$ and $\{\tilde{\delta}_{aij}, \tilde{\delta}_{eij}\}$. Also presented are the EBLUPs of the total VE effects for $\tilde{\beta}_{g_{ij}}$ and $\tilde{u_{g_{ij}}}$.

		Variety	2013-Breeza	2014-Breeza	2014-Edgeroi	2014-Tworth	2015-Breeza	2015-Edgeroi	2015-Nstar	2015-Tworth
Presence		Jandaroi	✓	✓	✓	✓	✓	✓	✓	✓
		DBA Lillaroi						✓		✓
		Vitron		✓						
		Neodur		✓						
		Bellaroi					✓	✓	✓	✓
		DBA Lillaroi-16A			✓	✓	✓	✓	✓	
Additive VE effects	$\tilde{\beta}_{aij}$	Jandaroi	0.025	0.494	0.025	0.449	0.028	0.596	0.286	0.084
		DBA Lillaroi	0.244	0.126	0.214	0.261	0.148	0.308	0.437	0.153
		Vitron	0.255	0.447	-0.031	0.220	0.184	0.026	0.359	0.164
		Neodur	-0.408	-0.715	0.050	-0.352	-0.294	-0.042	-0.574	-0.262
		Bellaroi	0.296	-0.279	0.058	-0.268	0.119	-0.064	-0.167	0.004
		DBA Lillaroi-16A	0.585	-0.083	0.406	0.189	0.307	0.435	0.548	0.233
	$\tilde{\delta}_{aij}$	Jandaroi	0.000	-0.000	-0.000	0.000	-0.000	0.030	-0.132	0.057
		DBA Lillaroi	-0.000	0.000	0.000	0.000	0.000	-0.053	0.081	0.019
		Vitron	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Neodur	0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Bellaroi	-0.000	0.000	-0.000	-0.000	0.000	0.030	0.040	0.075
		DBA Lillaroi-16A	-0.000	0.000	0.000	-0.000	0.000	-0.041	0.030	0.020
	$\tilde{u_{aij}}$	Jandaroi	0.025	0.494	0.025	0.449	0.028	0.625	0.154	0.141
		DBA Lillaroi	0.244	0.126	0.214	0.261	0.148	0.255	0.518	0.173
		Vitron	0.255	0.447	-0.031	0.220	0.184	0.026	0.359	0.164
		Neodur	-0.408	-0.715	0.050	-0.352	-0.294	-0.042	-0.574	-0.262
		Bellaroi	0.296	-0.279	0.058	-0.268	0.119	-0.033	-0.127	0.079
		DBA Lillaroi-16A	0.585	-0.083	0.406	0.189	0.307	0.394	0.578	0.253
Non-additive VE effects	$\tilde{\beta}_{eij}$	Jandaroi	-0.270	-0.102	-0.156	-0.083	-0.021	-0.088	-0.071	-0.094
		DBA Lillaroi	0.064	0.004	0.039	0.037	-0.006	0.023	0.013	0.049
		Vitron	0.104	0.077	0.057	0.000	0.026	0.030	0.035	-0.014
		Neodur	-0.166	-0.123	-0.091	-0.000	-0.042	-0.048	-0.056	0.023
		Bellaroi	0.540	0.318	0.233	-0.040	0.336	0.115	0.031	-0.081
		DBA Lillaroi-16A	0.326	0.167	0.142	-0.005	0.193	0.071	0.013	-0.018
	$\tilde{\delta}_{eij}$	Jandaroi	0.000^{+}	0.000^{+}	0.000^{+}	0.000^{+}	0.000^{+}	0.066	0.000^{+}	0.000^{+}
		DBA Lillaroi	0	0	0	0	0	0.104	0	0.000^{+}
		Vitron	0	0.000^{+}	0	0	0	0	0	0
		Neodur	0	0.000^{+}	0	0	0	0	0	0
		Bellaroi	0	0	0	0	0.000^{+}	-0.045	0.000^{+}	0.000^{+}
		DBA Lillaroi-16A	0	0	0.000^{+}	0.000^{+}	0.000^{+}	-0.046	0.000^{+}	0
	$\tilde{u_{eij}}$	Jandaroi	-0.270	-0.102	-0.156	-0.083	-0.021	-0.022	-0.071	-0.094
		DBA Lillaroi	0.064	0.004	0.039	0.037	-0.006	0.127	0.013	0.049
		Vitron	0.104	0.077	0.057	0.000	0.026	0.030	0.035	-0.014
		Neodur	-0.166	-0.123	-0.091	-0.000	-0.042	-0.048	-0.056	0.023
		Bellaroi	0.540	0.318	0.233	-0.040	0.336	0.070	0.031	-0.081
		DBA Lillaroi-16A	0.326	0.167	0.142	-0.005	0.193	0.025	0.013	-0.018
Total VE effects	$\tilde{\beta}_{g_{ij}}$	Jandaroi	-0.245	0.393	-0.132	0.366	0.007	0.507	0.215	-0.010
		DBA Lillaroi	0.308	0.130	0.253	0.298	0.142	0.331	0.450	0.203
		Vitron	0.358	0.524	0.026	0.220	0.210	0.057	0.394	0.149
		Neodur	-0.574	-0.838	-0.041	-0.352	-0.336	-0.091	-0.630	-0.239
		Bellaroi	0.836	0.039	0.291	-0.308	0.455	0.051	-0.136	-0.077
		DBA Lillaroi-16A	0.911	0.084	0.548	0.184	0.499	0.506	0.561	0.215
	$\tilde{u_{g_{ij}}}$	Jandaroi	-0.245	0.393	-0.132	0.366	0.007	0.603	0.084	0.047
		DBA Lillaroi	0.308	0.130	0.253	0.298	0.142	0.383	0.531	0.222
		Vitron	0.358	0.524	0.026	0.220	0.210	0.057	0.394	0.149
		Neodur	-0.574	-0.838	-0.041	-0.352	-0.336	-0.091	-0.630	-0.239
		Bellaroi	0.836	0.039	0.291	-0.308	0.455	0.036	-0.096	-0.003
		DBA Lillaroi-16A	0.911	0.084	0.548	0.184	0.499	0.419	0.591	0.236

$^{+}$ $\tilde{\delta}_{eij} = 0.000$ because VAF$_e$%=100 for this environment.

Chapter 9

Information based diagnostic for genetic variance parameter estimation in multi-environment trials

In Chapter 7 it was shown how to construct a MET dataset, that is, to determine the trials to be compiled, in order to maximise the amount of data available for selection decisions on varieties of interest. The criterion of $\mathcal{A}$-optimality was used to assess any given dataset since it reflects the probability of making selection errors (see Section 7.1). $\mathcal{A}$-optimality is based on known variance parameter values whereas in practice, these must be estimated from the data. In construction of a MET dataset it is therefore also important to assess whether the structure of the dataset supports accurate estimation of the variance parameters, in particular the genetic variance parameters.

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. In Chapter 6

I found that although variety connectivity was influential, there appeared to be other factors at play. In response to this, Lisle et al. (2021), proposed the criterion of $\mathcal{D}$-optimality as a diagnostic to assess the information available for the REML estimation of genetic variance parameters in a MET analysis.

This chapter is arranged as follows: Section 9.1 contains some general results about $\mathcal{D}$-optimality. Following this, Section 9.2 presents an adaptation of the publication listed below. In Section 9.3 I demonstrate how to calculate $\mathcal{D}$-values in practice. Finally, in Section 9.4 I have concluding remarks.

Lisle, C., Smith, A., Birrell, C,. & Cullis, B. R. (2021). Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials. *Frontiers in Plant Science*. **12**, 2856. doi: 10.3389/fpls:2021.785430.

9.1 Preliminary remarks about $\mathcal{D}$-optimality

In the experimental design literature, $\mathcal{D}$-optimality is used to search for designs that minimise the generalised variance of parameter estimates (Butler, 2013; Russell, 2018). The parameters are usually fixed effects, for example regression coefficients, within an ordinary (or generalised) linear model (Russell et al., 2009). Applications for linear mixed models are less common, but Ankenman et al. (2003) present an interesting example in which $\mathcal{D}$-optimality is used in order to minimise the variance of the estimates of both fixed effects and variance components.

The $\mathcal{D}$-optimality criterion is usually defined in terms of the determinant of the information matrix for the parameter estimates. In the experimental design context, a design is said to be $\mathcal{D}$-optimal if it maximises this determinant. Equivalently it is $\mathcal{D}$-optimal if it minimises the determinant of the variance matrix of the parameter estimates. The latter is preferred as it is more consistent with the $\mathcal{A}$-optimality criterion in the sense that "smaller is better". In our setting, the parameters of interest are the variance parameters, $\boldsymbol{\kappa}$, in a LMM, and we wish to assess a MET dataset for the ability to provide accurate REML estimates $\hat{\boldsymbol{\kappa}}$. The asymptotic variance matrix of $\hat{\boldsymbol{\kappa}}$ is given

by the inverse of the information matrix as given in Section 2.2. I formally define the variance matrix here as

$$\boldsymbol{V}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top) = \boldsymbol{\mathcal{I}}_{\boldsymbol{E}}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top)^{-1}$$

so the $\mathcal{D}$-optimality value is given by

$$\mathcal{D} = \log |\boldsymbol{V}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top)|$$

where the vertical bar represents the determinant.

To show how this can be interpreted as representing the "generalised variance" of $\hat{\boldsymbol{\kappa}}$ it is helpful to consider the case of $n_k = 2$ so that $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)^\top$. I let

$$\boldsymbol{V}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top) = \text{var} \begin{pmatrix} \hat{\kappa}_1 \\ \hat{\kappa}_2 \end{pmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

then it is shown that

$$\begin{aligned} |\boldsymbol{V}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top)| &= \sigma_{11}\sigma_{22} - \sigma_{12}^2 \\ &= \sigma_{11}(\sigma_{22} - \sigma_{12}^2/\sigma_{11}) \\ &= \text{var}(\hat{\kappa}_1)\text{var}(\hat{\kappa}_2|\hat{\kappa}_1) \end{aligned}$$

so that

$$\mathcal{D} = \log(\text{var}(\hat{\kappa}_1)) + \log(\text{var}(\hat{\kappa}_2|\hat{\kappa}_1))$$

Thus, the $\mathcal{D}$-optimality criterion is the sum of the logarithms of the variance of $\hat{\kappa}_1$ and the variance of $\hat{\kappa}_2$ conditional on $\hat{\kappa}_1$. This result of summing logarithms of variances of estimates conditional on other estimates holds more generally (and does not depend on the ordering) and hence the $\mathcal{D}$-optimality criterion can be interpreted as measuring generalised or "total" variance.

9.2 Reproduction of Lisle, Smith, Birrell and Cullis (2021)

In the reproduction, notational changes have been made to be consistent with the nomenclature used in this thesis. Additional figures, tables and results that were not included in the publication have been added in this chapter and some parts of the in-

troduction and discussion sections of the paper have been omitted.

I note the following additions that were not included in the publication:

- I include the application of the new diagnostic to the Oat dataset in Section 9.2.4. This is used in conjunction with the $\mathcal{A}$-optimality criterion as described in Chapter 7 to investigate the optimal number of years to include in the MET dataset.
- Two larger trial sizes (T192 and T384) are now included in the simulation study for the additive VE effects in Section 9.2.6.2.

9.2.1 Statistical methods

9.2.1.1 Model for analysis

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{9.1}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the vector of random genetic effects with associated design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_p}$ is a vector of random non-genetic (or peripheral) effects with associated design matrix $\boldsymbol{Z_p}$; and $\boldsymbol{e} = (\boldsymbol{e}_1^\top, \boldsymbol{e}_2^\top, \ldots, \boldsymbol{e}_p^\top)^\top$ is the combined vector of errors 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 that

$$\begin{bmatrix} \boldsymbol{u_g} \\ \boldsymbol{u_p} \\ \boldsymbol{e} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_g} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G_p} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix} \right) \tag{9.2}$$

where the matrices $\{\boldsymbol{G_g}, \boldsymbol{G_p}, \boldsymbol{\Sigma}\}$ are variance matrices for $\{\boldsymbol{u_g}, \boldsymbol{u_p}, \boldsymbol{e}\}$ respectively. $\boldsymbol{G_g}$ is known as the between environments variance/covariance matrix and is described in later sections. $\boldsymbol{G_p}$ is assumed to be block diagonal given by $\boldsymbol{G_p} = \oplus_{i=1}^{b} \sigma_{p_i}^2 \boldsymbol{I}_{q_i}$ where b is the number of components in $\boldsymbol{u_p}$ and q_i is the number of effects in (length of) $\boldsymbol{u_{p_i}}$. $\boldsymbol{\Sigma}$ is assumed to be block diagonal, so that $\boldsymbol{\Sigma} = \oplus_{j=1}^{p} \boldsymbol{\Sigma}_j$ where $\boldsymbol{\Sigma}_j = \text{var}(\boldsymbol{e_j})$

is the variance matrix for the errors for the j^{th} environment. In the LMM of Smith et al. (2001b), spatial models are used for the errors and the matrices $\boldsymbol{\Sigma}_j$ correspond to separable autoregressive processes.

The random genetic effects $\boldsymbol{u_g}$ 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 $\boldsymbol{u_g}$ has length mp, which is ordered as varieties within environments. In this chapter 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

$$\boldsymbol{u_g} = \boldsymbol{u_a} + \boldsymbol{u_e}$$

It is assumed that $\text{var}\left(\boldsymbol{u_a}\right) = \boldsymbol{G_a} \otimes \boldsymbol{A}$ where $\boldsymbol{A}$ is the numerator relationship matrix and $\boldsymbol{G_a}$ 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 $\text{var}\left(\boldsymbol{u_e}\right) = \boldsymbol{G_e} \otimes \boldsymbol{I}_m$ where $\boldsymbol{G_e}$ 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

$$\text{var}\left(\boldsymbol{u_g}\right) = \boldsymbol{G_g} = \boldsymbol{G_a} \otimes \boldsymbol{A} + \boldsymbol{G_e} \otimes \boldsymbol{I}_m \tag{9.3}$$

Note that if no pedigree information is included in the analysis then $\boldsymbol{u_g} = \boldsymbol{u_e}$ and $\boldsymbol{G_g} = \boldsymbol{G_e} \otimes \boldsymbol{I}_m$. Finally, the variance matrix for the data vector is given by

$$\text{var}\left(\boldsymbol{y}\right) = \boldsymbol{H} = \boldsymbol{Z_g} \boldsymbol{G_g} \boldsymbol{Z_g}^\top + \boldsymbol{Z_p} \boldsymbol{G_p} \boldsymbol{Z_p}^\top + \boldsymbol{\Sigma} \tag{9.4}$$

The first step in fitting the model in Equation (9.1) 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:

$$U(\kappa_i) = -\frac{1}{2}\left\{ tr(\boldsymbol{P}\dot{\boldsymbol{H}}_i) - \boldsymbol{y}^\top \boldsymbol{P} \boldsymbol{H}_i \boldsymbol{P} \boldsymbol{y} \right\} \tag{9.5}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{X}^\top\boldsymbol{H}^{-1}$ with $\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-}$ being any generalised inverse of $\left(\boldsymbol{X}^\top\boldsymbol{H}^{-1}\boldsymbol{X}\right)$. The "dot" notation indicates a derivative so that $\dot{\boldsymbol{H}}_i = \partial\boldsymbol{H}/\partial\kappa_i, \quad i = 1 \ldots n_\kappa$. 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, $\mathfrak{I}_{\boldsymbol{A}}$, is used instead of the expected information matrix, $\mathfrak{I}_{\boldsymbol{E}}$. The elements of these matrices are given by

$$\begin{aligned} \mathfrak{I}_{\boldsymbol{A}}(\kappa_i, \kappa_j) &= \tfrac{1}{2}\boldsymbol{y}^\top\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y} \\ \mathfrak{I}_{\boldsymbol{E}}(\kappa_i, \kappa_j) &= \tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \end{aligned} \tag{9.6}$$

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 9.1. In particular, the EBLUPs of the VE effects are given by $\tilde{\boldsymbol{u}}_{\boldsymbol{g}} = \boldsymbol{G}_{\boldsymbol{g}}\boldsymbol{Z}_{\boldsymbol{g}}{}^\top\boldsymbol{P}\boldsymbol{y}$ and these have an associated prediction error variance (PEV) of $\mathrm{var}\left(\tilde{\boldsymbol{u}}_{\boldsymbol{g}} - \boldsymbol{u}_{\boldsymbol{g}}\right) = \boldsymbol{G}_{\boldsymbol{g}} - \boldsymbol{G}_{\boldsymbol{g}}\boldsymbol{Z}_{\boldsymbol{g}}{}^\top\boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{g}}\boldsymbol{G}_{\boldsymbol{g}}$. Note that in these equations the matrices $\boldsymbol{G}_{\boldsymbol{g}}$ and $\boldsymbol{P}$ are formed using the REML estimate $\hat{\boldsymbol{\kappa}}$ of $\boldsymbol{\kappa}$. We can then compute a model based reliability (Mrode & 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

$$\mathrm{cor}\left(\tilde{u}_{g_k},\ u_{g_k}\right)^2 = 1 - \frac{\left(\boldsymbol{G}_{\boldsymbol{g}} - \boldsymbol{G}_{\boldsymbol{g}}\boldsymbol{Z}_{\boldsymbol{g}}{}^\top\boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{g}}\boldsymbol{G}_{\boldsymbol{g}}\right)_{kk}}{\left(\boldsymbol{G}_{\boldsymbol{g}}\right)_{kk}} \tag{9.7}$$

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

9.2.2 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, the elements of which are given in Equation (9.6). In this chapter 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 $\{\boldsymbol{G_a}, \boldsymbol{G_e}\}$ and $\boldsymbol{\kappa_{\bar{g}}}$ are the remaining variance parameters, that is, associated with the peripheral random effects and errors. We partition the full expected information matrix accordingly and write as

$$\boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa}, \boldsymbol{\kappa}^\top) = \begin{bmatrix} \boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top) & \boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_{\bar{g}}}^\top) \\ \boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_{\bar{g}}}, \boldsymbol{\kappa_g}^\top) & \boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_{\bar{g}}}, \boldsymbol{\kappa_{\bar{g}}}^\top) \end{bmatrix} \tag{9.8}$$

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

$$\boldsymbol{V}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top) = \left[\boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top) - \boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_{\bar{g}}}^\top)(\boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_{\bar{g}}}, \boldsymbol{\kappa_{\bar{g}}}^\top))^{-1}\boldsymbol{\mathcal{I}_E}(\boldsymbol{\kappa_{\bar{g}}}, \boldsymbol{\kappa_g}^\top)\right]^{-1} \tag{9.9}$$

Smith & Cullis (2018) recommend the use of factor analytic models for $\{\boldsymbol{G_a}, \boldsymbol{G_e}\}$. Other possibilities include compound symmetric and unstructured forms. Irrespective of the form used, the parameters of interest are the variances and covariances in $\{\boldsymbol{G_a}, \boldsymbol{G_e}\}$. The aim in this chapter 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 (9.1) we assume that the fixed effects comprise a mean parameter for each environment (so that $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_p)^\top$) and we assume there are no peripheral effects so write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\left(\boldsymbol{u_a} + \boldsymbol{u_e}\right) + \boldsymbol{e} \tag{9.10}$$

where the design matrices are given by $\boldsymbol{X} = \oplus_{j=1}^{p} \mathbf{1}_{n_j}$ and $\boldsymbol{Z_g} = \oplus_{j=1}^{p} \boldsymbol{Z_{g}}_j$ where $\boldsymbol{Z_{g}}_j$ is the $n_j \times m$ design matrix for the VE effects for environment j $(= 1, \ldots, p)$. The genetic variance matrices, $\{\boldsymbol{G_a}, \boldsymbol{G_e}\}$, are assumed to have unstructured forms with $p(p+1)/2$ unique variance parameters in each that are denoted by $\{\sigma_{a_{js}}, \sigma_{e_{js}}\}(j \leq s = 1, \ldots, p)$, respectively.

Finally, we assume that the error variance matrices are given by $\boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}_{n_j}$ so that $\boldsymbol{\Sigma} = \oplus_{j=1}^{p} \sigma_j^2 \boldsymbol{I}_{n_j}$. The variance matrix for the data vector is then given by

$$\boldsymbol{H} = \boldsymbol{Z_g} \left(\boldsymbol{G_a} \otimes \boldsymbol{A} + \boldsymbol{G_e} \otimes \boldsymbol{I}_m \right) \boldsymbol{Z_g}^\top + \boldsymbol{\Sigma}$$

and the unknown variance parameters are $\boldsymbol{\kappa} = (\boldsymbol{\kappa_g}^\top, \boldsymbol{\kappa_{\bar{g}}}^\top)^\top$ where $\boldsymbol{\kappa_g}$ comprises $\{\sigma_{a_{js}}, \sigma_{e_{js}}\}$ $(j \leq s = 1, \ldots, p)$ and $\boldsymbol{\kappa_{\bar{g}}}$ comprises σ_j^2 $(j = 1, \ldots, p)$. We then use pre-specified values of these parameters to compute the information matrix in Equation (9.8) and thence the variance matrix in Equation (9.9). The chosen variance parameters will be denoted $\boldsymbol{\kappa}_{\boldsymbol{g}_0}$ and $\boldsymbol{\kappa}_{\bar{\boldsymbol{g}}_0}$ and the resultant variance matrix denoted by $\boldsymbol{V}(\boldsymbol{\kappa}_{\boldsymbol{g}_0}, \boldsymbol{\kappa}_{\boldsymbol{g}_0}^\top)$. 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

$$\mathcal{D} = \frac{\log \left| \boldsymbol{V}\left(\boldsymbol{\kappa}_{\boldsymbol{g}_0}, \boldsymbol{\kappa}_{\boldsymbol{g}_0}^\top\right) \right|}{n_{\kappa_g}} \tag{9.11}$$

where the vertical bar represents the determinant and n_{κ_g} is the number of genetic variance parameters and 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 $(= 1, \ldots, p)$ as

$$\mathcal{D}_j = \frac{\log \left| \boldsymbol{V}\left(\boldsymbol{\kappa}_{\boldsymbol{g}_{0_j}}, \boldsymbol{\kappa}_{\boldsymbol{g}_{0_j}}^\top\right) \right|}{n_{\kappa_{g_j}}} \tag{9.12}$$

where $\boldsymbol{V}(\boldsymbol{\kappa}_{\boldsymbol{g}_{0_j}}, \boldsymbol{\kappa}_{\boldsymbol{g}_{0_j}}^\top)$ is the partition of $\boldsymbol{V}(\boldsymbol{\kappa}_{\boldsymbol{g}_0}, \boldsymbol{\kappa}_{\boldsymbol{g}_0}^\top)$ that relates to environment j and $n_{\kappa_{g_j}}$ is the associated number of genetic variance parameters. In the case of models in which information on genetic relatedness is not used we have $n_{\kappa_{g_j}} = p$, 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 $n_{\kappa_{g_j}} = 2p$. To distinguish between these different genetic models we label the diagnostic values as $\mathcal{D}_j$(A+I) if they correspond to a LMM with both additive (A) and non-additive (or independent, I) VE effects; $\mathcal{D}_j$(I) if they correspond to the LMM with independent VE effects alone (that is, genetic relatedness is not used) or $\mathcal{D}_j$(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 scrutinized 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 $\boldsymbol{V}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top)$ by using the marginal variance matrix for the genetic variance parameters rather than the conditional matrix as given in Equation (9.9). Thus we can use

$$\boldsymbol{V}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top) = [\boldsymbol{\mathfrak{I}}_{\boldsymbol{E}}(\boldsymbol{\kappa_g}, \boldsymbol{\kappa_g}^\top)]^{-1} \tag{9.13}$$

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

9.2.3 A two-stage procedure

We first let d_j be the number of varieties in environment j $(= 1, \ldots, p)$ 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 $\boldsymbol{V}(\boldsymbol{\kappa_{g_0}}, \boldsymbol{\kappa_{g_0}}^\top)$ using Equation (9.9) or (9.13) 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 (9.10) 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 $d \times 1$ subvector corresponding to the VE combinations present in the data. Thus, we can write $\boldsymbol{\eta}_d = \boldsymbol{D}\boldsymbol{\eta}$ where $\boldsymbol{D}$ is a $d \times mp$ indicator matrix that selects the appropriate elements. We let $\hat{\boldsymbol{\eta}}_{\boldsymbol{d}}$ 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 error variance for environment j is simply $\sigma_j^2 \boldsymbol{I}_{n_j}$. This means that the variance matrix of $\hat{\boldsymbol{\eta}}_{\boldsymbol{d}}$ from the first stage is given by $\boldsymbol{\Omega} = \oplus_{j=1}^{p} \boldsymbol{\Omega}_j$ where $\boldsymbol{\Omega}_j$ 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

$$\boldsymbol{y_2} = \boldsymbol{X_2}\boldsymbol{\tau} + \boldsymbol{D}\,(\boldsymbol{u_a} + \boldsymbol{u_e}) + \boldsymbol{\xi} \tag{9.14}$$

where $\boldsymbol{y_2} = \hat{\boldsymbol{\eta}}_{\boldsymbol{d}}$ from the first stage, and $\boldsymbol{X_2} = \boldsymbol{D}\,(\boldsymbol{I}_p \otimes \boldsymbol{1}_m)$. In terms of the variance structures, $\text{var}\,(\boldsymbol{u_a}) = \boldsymbol{G_a} \otimes \boldsymbol{A}$ and $\text{var}\,(\boldsymbol{u_e}) = \boldsymbol{G_e} \otimes \boldsymbol{I}_m$ (as in the one-stage analysis) and $\text{var}\,(\boldsymbol{\xi}) = \boldsymbol{\Omega}$ where this is known from the first stage. The variance matrix for the data vector in the second stage is then given by

$$\boldsymbol{H_2} = \boldsymbol{D}\,(\boldsymbol{G_a} \otimes \boldsymbol{A} + \boldsymbol{G_e} \otimes \boldsymbol{I}_m)\,\boldsymbol{D}^\top + \boldsymbol{\Omega} \tag{9.15}$$

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

$$\boldsymbol{\mathcal{I}_{2E}}(\kappa_i, \kappa_j) = \frac{1}{2}\text{tr}\left(\boldsymbol{P_2}\dot{\boldsymbol{H}}_{\boldsymbol{2}i}\boldsymbol{P_2}\dot{\boldsymbol{H}}_{\boldsymbol{2}j}\right) \tag{9.16}$$

where $\boldsymbol{P_2} = \boldsymbol{H_2}^{-1} - \boldsymbol{H_2}^{-1}\boldsymbol{X_2}\left(\boldsymbol{X_2}^\top \boldsymbol{H_2}^{-1}\boldsymbol{X_2}\right)^{-1}\boldsymbol{X_2}^\top \boldsymbol{H_2}^{-1}$. This now involves matrices of dimension d rather than n. As in the previous Section 9.2.2 we do not actually conduct the two-stage analysis but compute the expected information matrix using Equation (9.16) for a given choice of variance parameter values. We then form the (marginal) variance matrix for the genetic variance parameter estimates using Equation (9.13) and denote the resultant matrix by $\boldsymbol{V_2}(\boldsymbol{\kappa}_{\boldsymbol{g}_0}, \boldsymbol{\kappa}_{\boldsymbol{g}_0}^\top)$. This is substituted into Equations (9.11)

and (9.12) to compute the diagnostic.

9.2.4 Application to the Oat dataset

This section shows the application of the diagnostic to the Oat dataset (see Chapter 4). As this dataset does not have pedigree information, it represents a scenario with independent VE effects. Here I present some key information and apply the diagnostic procedure. The dataset comprised $m = 163$ varieties from $n = 7068$ plots corresponding to 47 trials from Stage 4 (S4) across $p = 41$ environments sown between 2012 and 2016. The number of varieties in an environment ranged from 48 to 65 with a median of 52. I note that there are $d = 2216$ variety by environment combinations present in the data, representing a nearly 70% reduction when using a 2-stage approach for computing the diagnostic. The number of varieties in common between pairs of environments ($x_{1,2}$) (see Figure 4.3) ranged from 16 to 65 with a median of 22.

As pedigree information was not available I computed the $\mathfrak{I}_{\boldsymbol{2E}}$ based on the LMM without the partitioning of the genetic effects as in Equation 9.10, so that $n_{\kappa_g} = 861$. The values of the variance parameters for calculation of the diagnostic were set to

$$\sigma_{e_{js}} = \begin{cases} 0.2; & j = s = 1, \ldots, p \\ 0.16; & j < s = 1, \ldots, p \end{cases}$$
$$\sigma_j^2 = \begin{cases} 0.15; & j = 1, \ldots, p \end{cases}$$

These values were chosen as being both representative of actual estimates from historical analyses that are often encountered in practice. In particular, I have set a between environments correlation of 0.8 for the VE effects.

As the Oat dataset does not have pedigree information available, the CG methodology described in Chapter 7 is not appropriate to be used to construct the MET dataset. As such, this is investigated here via the inclusion and exclusion of years in the dataset with examination of the $\mathcal{A}$-values as shown in Table 9.1, and $\mathcal{D}$(I)-values as shown in Table 9.2. The $\mathcal{D}_j$(I)-values for each environment within the yearly datasets, from this pre-specified LMM are given in Table 9.2. This relationship is shown in Figure 9.1, which shows the $\mathcal{A}$-values decreasing over increasing numbers of years, whilst the

$\mathcal{D}$(I)-values increase after two-years. The two-year MET dataset provides the smallest $\mathcal{D}$(I)-value (-7.99) and hence the greatest information to estimate the genetic variance parameters and the five-year dataset provides the highest $\mathcal{D}$(I)-value (-7.87) and hence the least information to estimate the genetic variance parameters. It is also noted that environments in 2013 (mean = -7.34) had the smallest $\mathcal{D}_j$(I)-values and environments in 2012 (mean = -7.28) and 2016 (mean = -7.25) had the largest $\mathcal{D}_j$(I)-values, which represent the greatest and least information to estimate genetic variance parameters, respectively.

Table 9.1: Oat example: Diagnostic $\mathcal{A}$-values based on LMM with independent VE effects over varying numbers of years in the MET dataset. Average number of environments present for the 2016 cohort of test varieties, number of environments, varieties and plots.

Number of			Mean	Number of		
years	Years	$\mathcal{A}$-value	Environments	Environments	Varieties	Plots
5	2012:2016	0.28	16.9	41	163	7068
4	2013:2016	0.28	15.9	32	133	5340
3	2014:2016	0.29	14.7	23	92	3504
2	2015:2016	0.30	13.0	17	74	2640
1	2016	0.36	9.0	9	48	1296

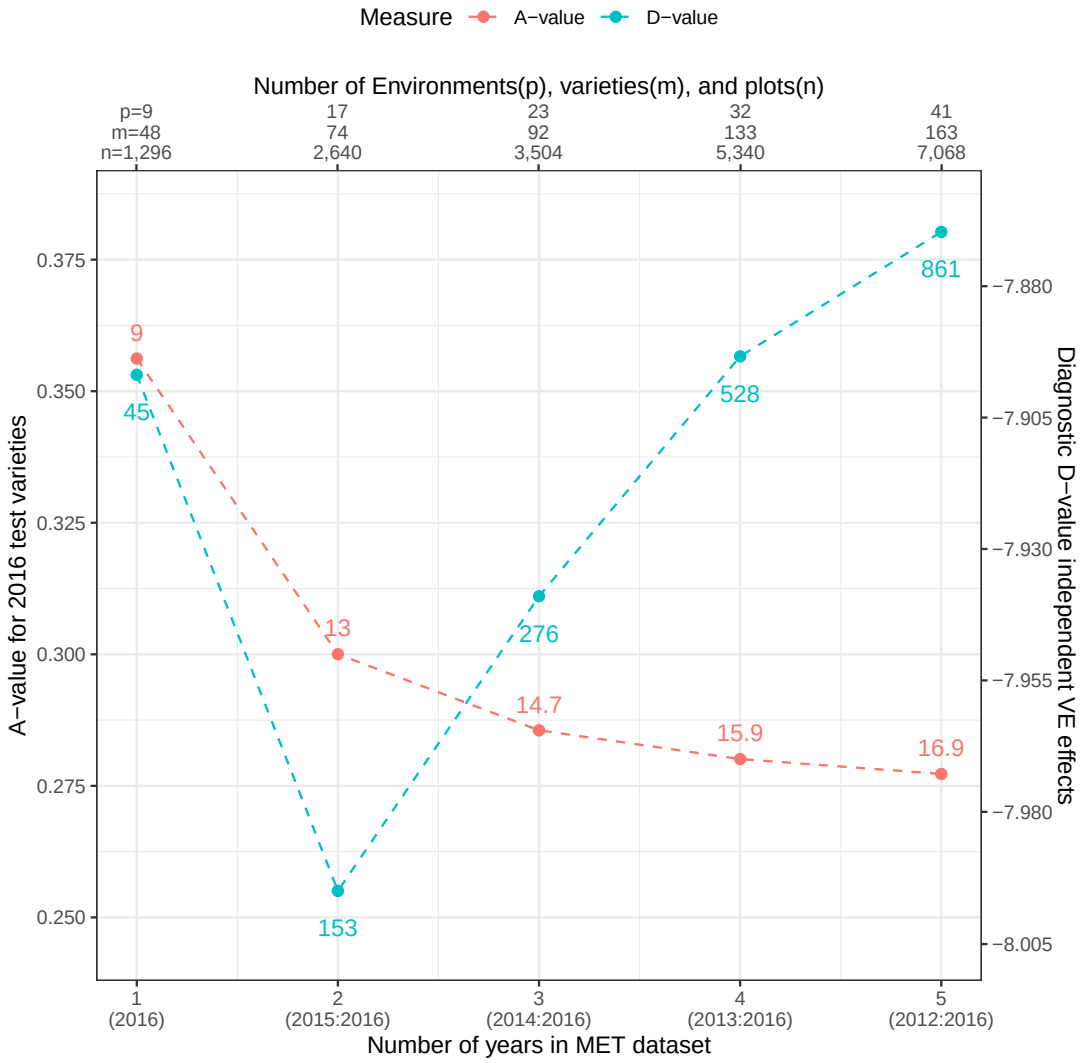

Figure 9.1: Oat example: $\mathcal{A}$- and diagnostic $\mathcal{D}$(I)-values over varying numbers of years in the MET dataset. The points are labelled with the associated mean numbers of environments in which the 2016 test varieties were sown for the $\mathcal{A}$-values and the number of parameters (n_{κ_g}) for the $\mathcal{D}$(I)-values. Left and right y-axes correspond to the $\mathcal{A}$- and diagnostic $\mathcal{D}$(I)-values respectively.

Table 9.2: Oat example: Diagnostic $\mathcal{D}_j$-values based on LMMs with independent VE effects ($\mathcal{D}_j$(I)) over differing numbers of years in the MET dataset. Five year dataset consists of the full five year dataset used as the motivating dataset. Dashed horizontal lines represent year blocks. Bottom two rows present the overall $\mathcal{D}$(I)-values and the number of parameters n_{κ_g}.

Year	No.	Environment	$\mathcal{D}_j$(I) Number of years in dataset 5	4	3	2	1
2012	1	OMaB12BARK6	-7.24				
	2	OMaB12KATA6	-7.24				
	3	OMaA12KYBY5	-7.24				
	4	OMaB12MERR6	-7.24				
	5	12PINE5	-7.37				
	6	12RIVE5	-7.37				
	7	OMaA12RUTH3	-7.24				
	8	12TURR5	-7.37				
	9	OMaB12WONG6	-7.24				
2013	10	OMaB13KATA6	-7.30	-7.26			
	11	OMaA13KYBY5	-7.30	-7.26			
	12	13PINE5	-7.42	-7.37			
	13	13RIVE5	-7.42	-7.37			
	14	OMaA13RUTH3	-7.30	-7.26			
	15	OMaA13RYLI6	-7.30	-7.26			
	16	13TURR5	-7.42	-7.37			
	17	OMaB13WONG6	-7.30	-7.26			
	18	OMaA13YANC2	-7.30	-7.26			
2014	19	OMaB14KATA6	-7.29	-7.30	-7.25		
	20	OMaA14KYBY5	-7.29	-7.30	-7.25		
	21	OMaA14PINE5	-7.29	-7.30	-7.25		
	22	OMaA14RIVE5	-7.29	-7.30	-7.25		
	23	OMaB14RYLI6	-7.29	-7.30	-7.25		
	24	OMaB14WONG6	-7.29	-7.30	-7.25		
2015	25	OMaB15CUND6	-7.31	-7.35	-7.36	-7.36	
	26	OMaA15GRIF2	-7.31	-7.35	-7.36	-7.36	
	27	OMaA15KYBY5	-7.31	-7.35	-7.36	-7.36	
	28	OMaB15PING6	-7.31	-7.35	-7.36	-7.36	
	29	OMaA15RIVE5	-7.31	-7.35	-7.36	-7.36	
	30	OMaB15RYLI6	-7.31	-7.35	-7.36	-7.36	
	31	OMaA15TURR5	-7.31	-7.35	-7.36	-7.36	
	32	OMaA15WONG6	-7.31	-7.35	-7.36	-7.36	
2016	33	OMaB16COWR2	-7.25	-7.27	-7.30	-7.33	-7.25
	34	OMaB16CUND6	-7.25	-7.27	-7.30	-7.33	-7.25
	35	OMaA16KYBY5	-7.25	-7.27	-7.30	-7.33	-7.25
	36	OMaA16MURE6	-7.25	-7.27	-7.30	-7.33	-7.25
	37	OMaA16PINE5	-7.25	-7.27	-7.30	-7.33	-7.25
	38	OMaA16RIVE5	-7.25	-7.27	-7.30	-7.33	-7.25
	39	OMaA16RYLI6	-7.25	-7.27	-7.30	-7.33	-7.25
	40	OMaA16TURR5	-7.25	-7.27	-7.30	-7.33	-7.25
	41	OMaA16WONG6	-7.25	-7.27	-7.30	-7.33	-7.25
		Overall $\mathcal{D}_j$(I)	-7.87	-7.89	-7.94	-7.99	-7.90
		n_{κ_g}	861	528	276	153	45

9.2.5 Application to the Durum dataset

This section shows the application of the diagnostic to the Durum dataset. In this chapter, for reasons of simplicity and clarity, we restrict attention to Stage 3 (S3) selection decisions so we subset the data accordingly from the full Durum dataset as described in Chapter 4. Here we present some key summary information. As this dataset contains pedigree information, it represents a scenario with related genetic effects. Summary information for the Durum dataset is given in Table 9.3. The dataset comprised $m = 3708$ varieties from $n = 9786$ 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 1649 with a median of 105. We note that there are $d = 6168$ 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 3959 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 9.2) ranged from 3 to 485 with a median of 36.

Given that the analysis of the Durum dataset 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 $\mathfrak{J}_{2\boldsymbol{E}}$ based on this model (so that $n_{\kappa_g} = 182$). The values of the variance parameters for calculation of the diagnostic were set to

$$\sigma_{a_{js}} = \begin{cases} 0.1; & j = s = 1, \ldots, p \\ 0.08; & j < s = 1, \ldots, p \end{cases}$$
$$\sigma_{e_{js}} = \begin{cases} 0.05; & j = s = 1, \ldots, p \\ 0.04; & j < s = 1, \ldots, p \end{cases}$$
$$\sigma_j^2 = \begin{cases} 0.15; & j = 1, \ldots, p \end{cases}$$

These values were chosen as being both representative of actual estimates from historical analyses that are often encountered in practice 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

9.3) 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 $\mathcal{D}_j$(A+I)-values for each environment from this pre-specified LMM are given in Table 9.4. The environments 2016-Tworth and 2015-Tworth had the smallest $\mathcal{D}_j$(A+I)-values and therefore the greatest information to estimate genetic variance parameters. Whereas, 2018-Tworth, 2018-Gurley, and 2018-Breeza had the largest $\mathcal{D}_j$(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 $\boldsymbol{\mathcal{I}_{2E}}(\kappa_i, \kappa_j)$ and hence a significant saving in computation. The resultant $\mathcal{D}_j$(A)-values are presented in Table 9.4 and Figure 9.3, which show little difference compared with the $\mathcal{D}_j$(A+I)-values. In particular, Figure 9.3 shows an almost 1:1 relationship.

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

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

In particular, we have set the additive genetic variance to 40% of the total genetic variance (see Equation 9.3) for each environment and therefore 60% for the non-additive genetic variance, and a between environments correlation of 0.4 for both additive and non-additive VE effects. The resultant $\mathcal{D}_j$-values are presented in both Table 9.4 and Figure 9.3. Once again, there is little difference in the rankings of environments for $\mathcal{D}_j$(A+I) compared with $\mathcal{D}_j$(A) (Figure 9.3 (d)). 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 (Figure 9.3(b) and (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 9.4. This figure shows that the diagnostic values encompass numerous structural elements of the environments other than variety connectivity, such as the number of varieties grown and the mean replication per variety.

Table 9.3: Summary of environments in the Durum dataset. Number of: trials for each stage of testing (S1, S2, S3), total trials, plots and varieties. Dashed horizontal lines represent year blocks.

		Number of trials				Number of	
No.	Environment	S1	S2	S3	Total	Plots	Varieties
1	2014-Breeza	3	0	0	3	1296	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	2052	1418
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	2448	1649
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	1004	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	9786	3708

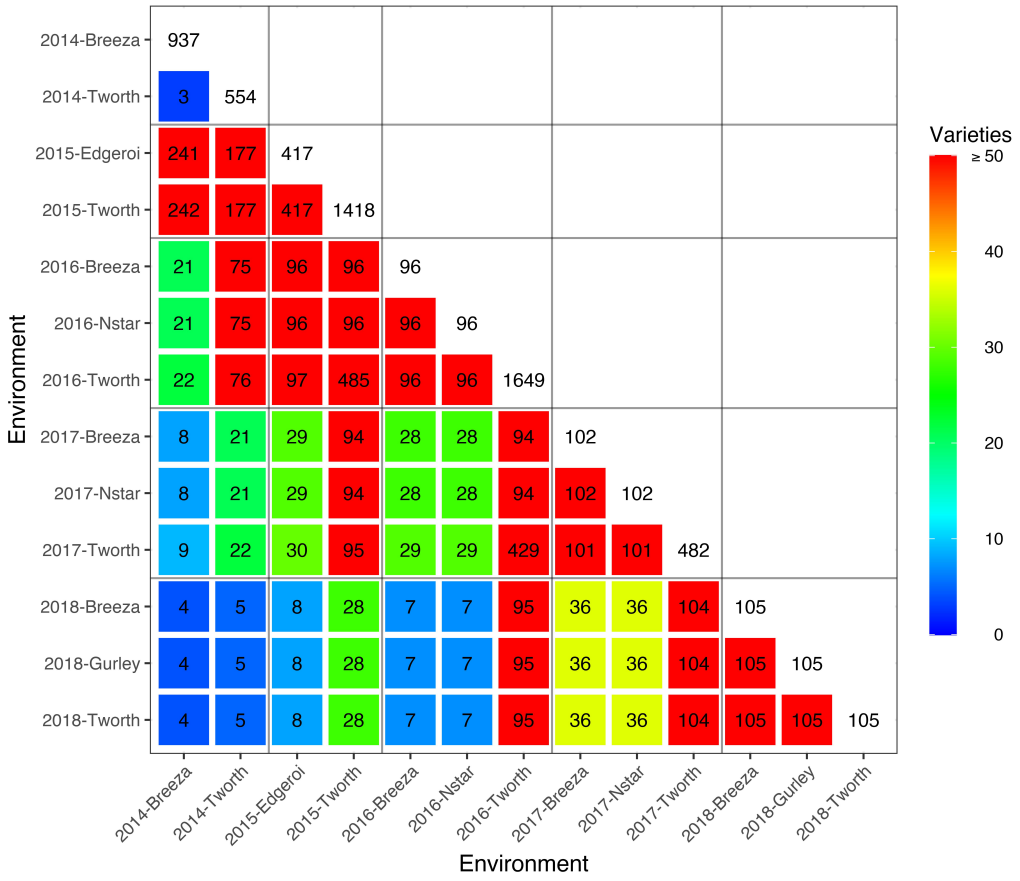

Figure 9.2: Heatmap of the number of varieties in common between all pairs of environments in the Stage 3 Durum MET dataset. The colours are as referenced in the legend. The boxes along the diagonal show the number of unique varieties in individual environments. Boundaries for years are indicated by the grey horizontal and vertical lines.

Table 9.4: 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)). 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).

	High		Low	
Environment	$\mathcal{D}_j$(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

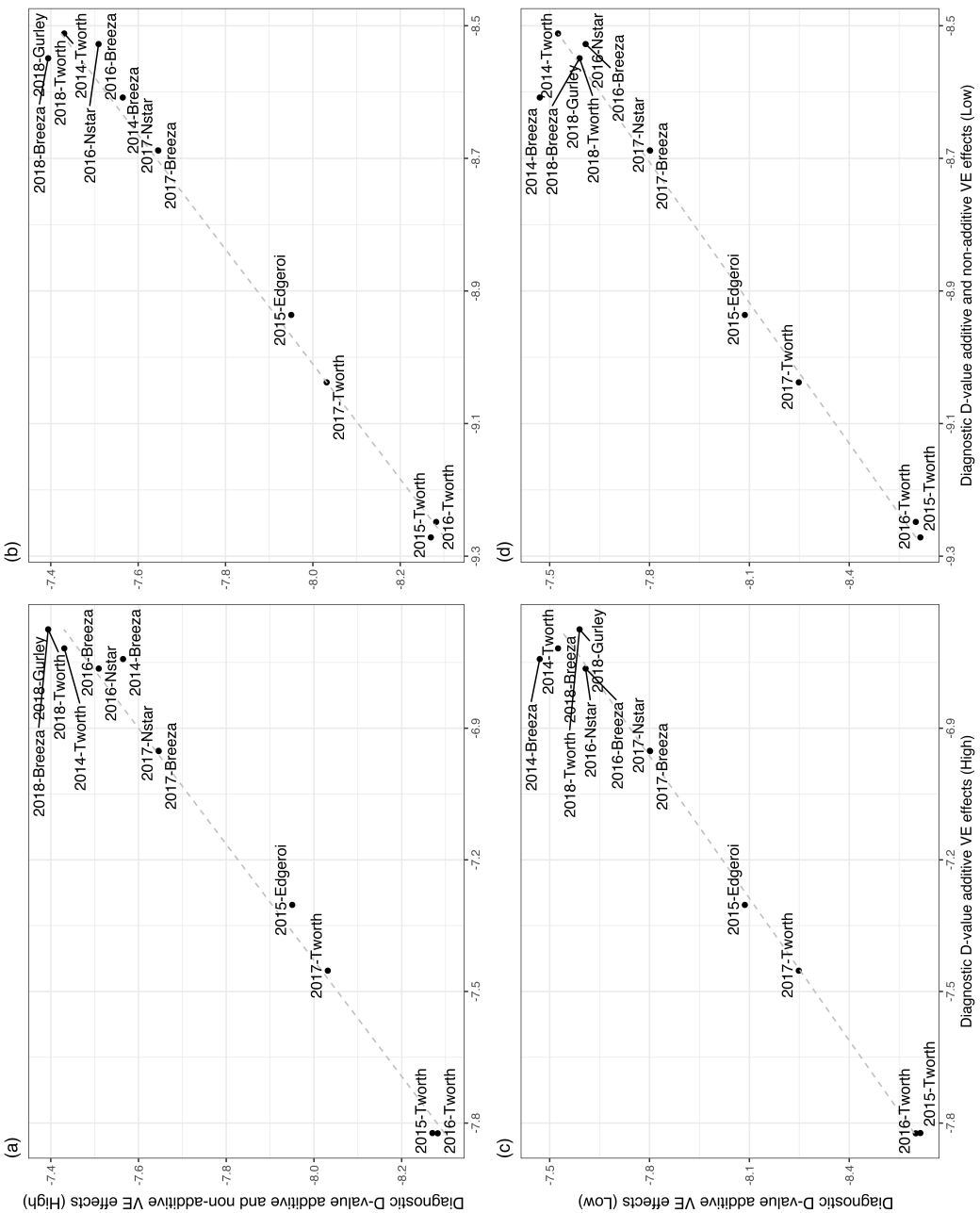

Figure 9.3: Durum example: Comparisons of Diagnostic $\mathfrak{D}_j$-values based on LMMs with additive and non-additive VE effects ($\mathfrak{D}_j$(A+I)) and those based on LMMs with additive VE effects alone ($\mathfrak{D}_j$(A+I)) for high and low parameter values of additive variance (80% and 40% respectively) and between environments genetic correlation (0.8 and 0.4 respectively). (a) $\mathfrak{D}_j$(A+I) against $\mathfrak{D}_j$(A) values using the high parameter values, (b) $\mathfrak{D}_j$(A+I) against $\mathfrak{D}_j$(A+I) values using high and low parameter values respectively, (c) $\mathfrak{D}_j$(A) against $\mathfrak{D}_j$(A) using high and low parameter values respectively, and (d) $\mathfrak{D}_j$(A) against $\mathfrak{D}_j$(A+I) using low parameter values.

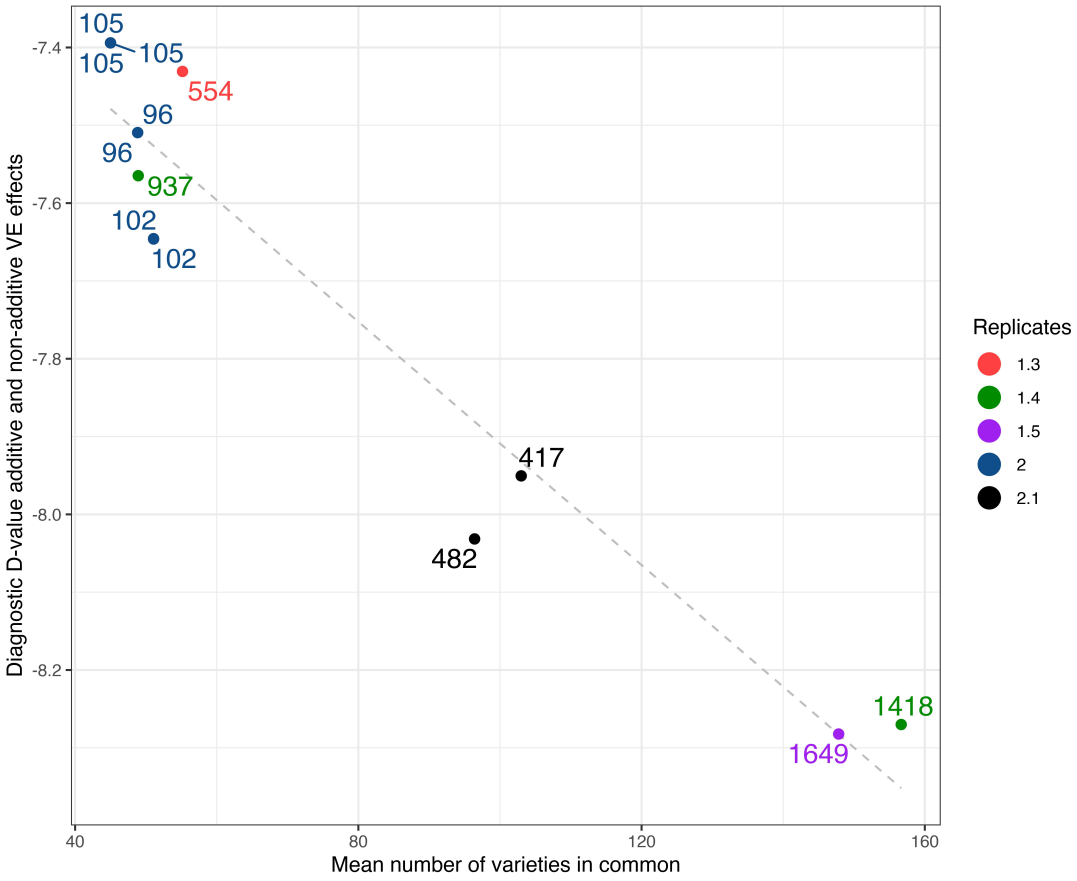

Figure 9.4: Durum example: Diagnostic $\mathcal{D}_j$(A+I)-values based on LMM with additive and non-additive VE effects plotted against mean number of varieties in common. Labels show the total number of varieties, colours as represented in legend show the mean number of replicates.

9.2.6 Simulation studies to investigate the performance of the diagnostic

9.2.6.1 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 $x_{1,2} = d - m$). 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 $d_1(= d_2) = \{12, 24, 48, 96\}$ so that $n_1(= n_2) = \{36, 72, 144, 288\}$.

The simulation study for the first trial size (Tsize=12) is described in the following. We consider the connectivity levels $x_{1,2} = 2, 4 \ldots 12$ (increments of 2). The maximum total number of varieties across Env1 and Env2 is $m = 22$, corresponding to $x_{1,2} = 2$. 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 $x_{1,2} = 12$; V2 - V13 for $x_{1,2} = 11$ and so on to V11 - V22 for $x_{1,2} = 2$. 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 (9.10) 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 $x_{1,2}$ and some pre-specified variance parameters, we can compute the diagnostic for Env1 which will be denoted $\mathcal{D}_{1c}(\mathrm{I})$, where c represents $x_{1,2}$. For the purposes of both the calculation of $\mathcal{D}_{1c}(\mathrm{I})$ and of data generation in the simulation study we chose the values of the variance parameters to be $\boldsymbol{\kappa}_{\boldsymbol{g}_0} = \left(\sigma_{e11_0} = 0.2, \sigma_{e12_0} = 0.16, \sigma_{e22_0} = 0.2\right)^\top$ and $\boldsymbol{\kappa}_{\bar{\boldsymbol{g}}_0} = \left(\sigma^2_{1_0} = 0.15, \sigma^2_{2_0} = 0.15\right)^\top$. The diagnostic is calculated using the two-stage formula for expected information, that is,

as in Equation (9.16) and is given by

$$\mathcal{D}_{1c}(\mathrm{I}) = \frac{\log\left|\boldsymbol{V_c}\left((\sigma_{e11_0}, \sigma_{e12_0}), (\sigma_{e11_0}, \sigma_{e12_0})^\top\right)\right|}{2} \tag{9.17}$$

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

1. Generate the random genetic effects $\boldsymbol{u_e}$ and errors $\boldsymbol{e}$ as per the LMM in Equation (9.10) and for the pre-specified variance parameters $\boldsymbol{\kappa_{g_0}}$ and $\boldsymbol{\kappa_{\bar{g}_0}}$. 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 s by $\boldsymbol{u_e}_s$. The errors for simulation s are denoted by $\boldsymbol{e}_s$ which is a vector of length $n_1 + n_2 = 72$.

2. For the connectivity level $x_{1,2}$, we subset the appropriate 24 elements of $\boldsymbol{u}_{e_s}$. We will label the associated vector as $\boldsymbol{u}_{e_{sc}}$. We then form the data vector and fit the LMM as in Equation (9.10), without the inclusion of pedigree information. We save the REML estimates of the genetic variance parameters, denoting these as $\{\hat{\sigma}_{e11_{sc}}, \hat{\sigma}_{e12_{sc}}, \hat{\sigma}_{e22_{sc}}\}$ and save the EBLUPs of the genetic effects, denoting these as $\tilde{\boldsymbol{u}}_{\boldsymbol{e}_{sc}}$.

3. Repeat Step 2 for each value of $x_{1,2}$

A total of $N = 2000$ simulations was conducted for each trial size. 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 $\boldsymbol{G_e}$. All models in this chapter were fitted using the `ASReml-R` package (Butler et al., 2017) within `R` (R Core Team, 2020).

The simulations were conducted in order to obtain two main quantities of interest for each value of $x_{1,2}$, 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 (9.17), namely:

$$\mathcal{D}^S_{1c}(\mathrm{I}) = \frac{\log\left|\boldsymbol{V_c}\left((\hat{\sigma}_{e11_c}, \hat{\sigma}_{e12_c}), (\hat{\sigma}_{e11_c}, \hat{\sigma}_{e12_c})^\top\right)\right|}{2} \tag{9.18}$$

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 $x_{1,2}$, 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 $\boldsymbol{u}_{\boldsymbol{e}sc}$ for the variety and Env1) and the EBLUPs (element of $\tilde{\boldsymbol{u}}_{\boldsymbol{e}_{sc}}$ for the variety and Env1). This will be denoted R^S_{kc}.

Noting that the simulation based reliabilities (R^S_{kc}) of the variety predictions take into account the uncertainty in having to estimate the variance parameters, we compute analogous 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 $x_{1,2}$, the LMM as per Equation (9.10) but with the variance parameters fixed at the value $\boldsymbol{\kappa}_0$. We then computed the model based reliability for variety k $(= 1 \dots 12)$ in Env1 using Equation (9.7) but because this has been computed with respect to known variance parameters (not REML estimates) we will call it the design based reliability (R^D_{kc}). We calculated the associated loss for the EBLUP reliabilities as

$$\text{loss} = R^D_{kc} - R^S_{kc} \tag{9.19}$$

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 analyses 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.

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 9.2.6.1) was strongly related to the number of varieties in common

between the two environments (see Figure 9.5). The number of successful model fits for the connectivity level of $x_{1,2} = 2$ 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 9.17) and the simulation based equivalent $\mathcal{D}^{S}_{1c}$(I)-values (from Equation 9.18) is shown in Figure 9.6 (a). This good agreement shown in Figure 9.6 (a) clearly indicates that the diagnostic performs well in terms of forecasting the level of uncertainty in genetic variance parameter estimation. Figure 9.6 (b) 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 9.7 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 9.8).

Results displayed in Figures 9.5, 9.6 and 9.8 have been extracted for the "best case" scenario of 100% connectivity for each Tsize and are presented in Table 9.5. This again shows the good agreement between the diagnostic $\mathcal{D}_{1c}$(I)-values and the simulation based $\mathcal{D}^{S}_{1c}$(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).

Table 9.5: Summary of key results for independent VE effects simulation study for each trial size and the case of 100% connectivity between the two trials: simulation based $\mathcal{D}^{s}_{1c}$(I)-values; diagnostic $\mathcal{D}_{1c}$(I)-values; mean loss in reliability of EBLUPs of VE effects for Env1 (with associated standard error); number of successful model fits out of $N = 2000$ simulations.

Tsize	Varieties in common	$\mathcal{D}^{s}_{1c}$(I)	$\mathcal{D}_{1c}$(I)	EBLUP reliability loss	se	Successful model fits
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

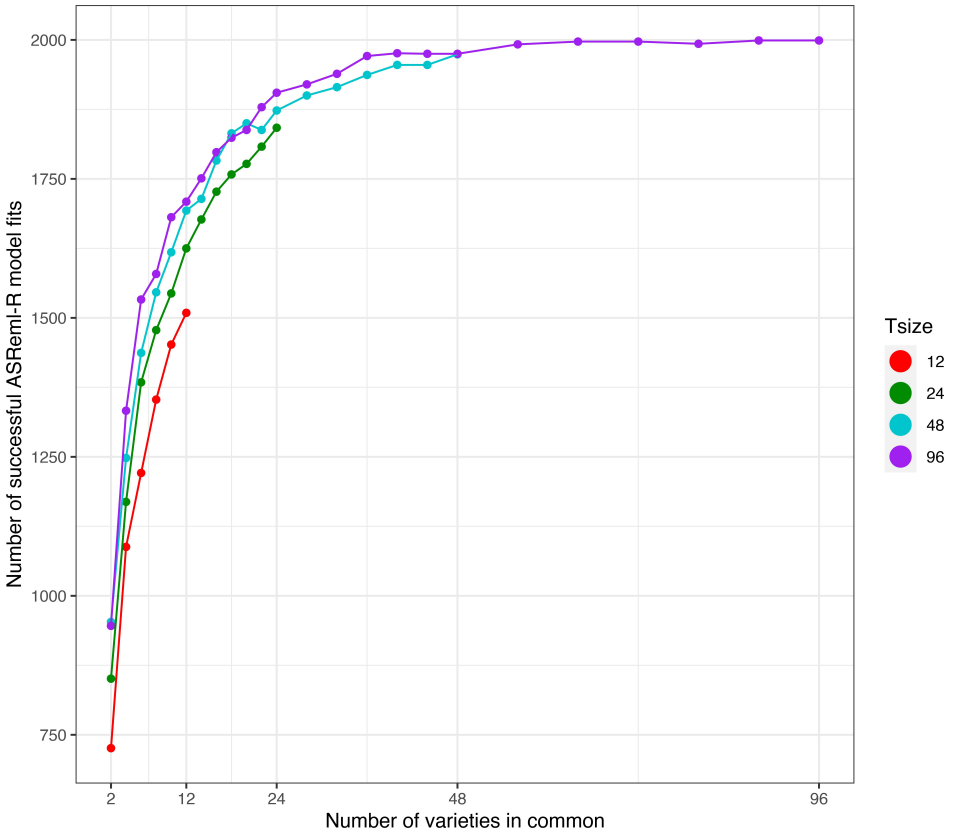

Figure 9.5: Independent VE effects simulation study: number of successful model fits from $N = 2000$ simulations plotted against number of varieties in common for four trial sizes (trials with $\{12, 24, 48, 96\}$ varieties). Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $x_{1,2} = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

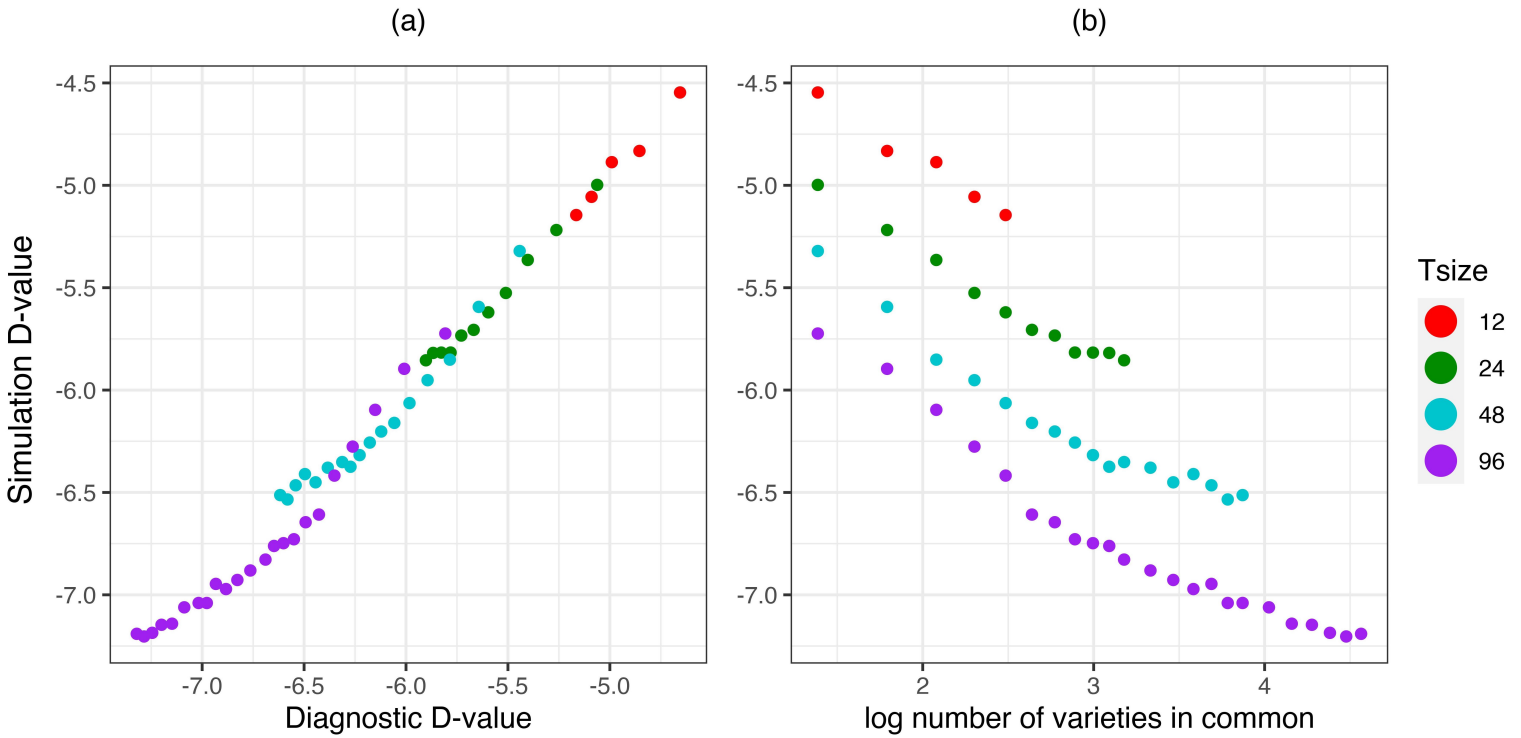

Figure 9.6: Independent VE effects simulation study: simulation based $\mathcal{D}^{s}_{1c}$(I)-values plotted against (a) diagnostic $\mathcal{D}_{1c}$(I)-values and (b) log number of varieties in common for four trial sizes (trials with $\{12, 24, 48, 96\}$ varieties) and a sequence of connectivity levels. Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $x_{1,2} = 4$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

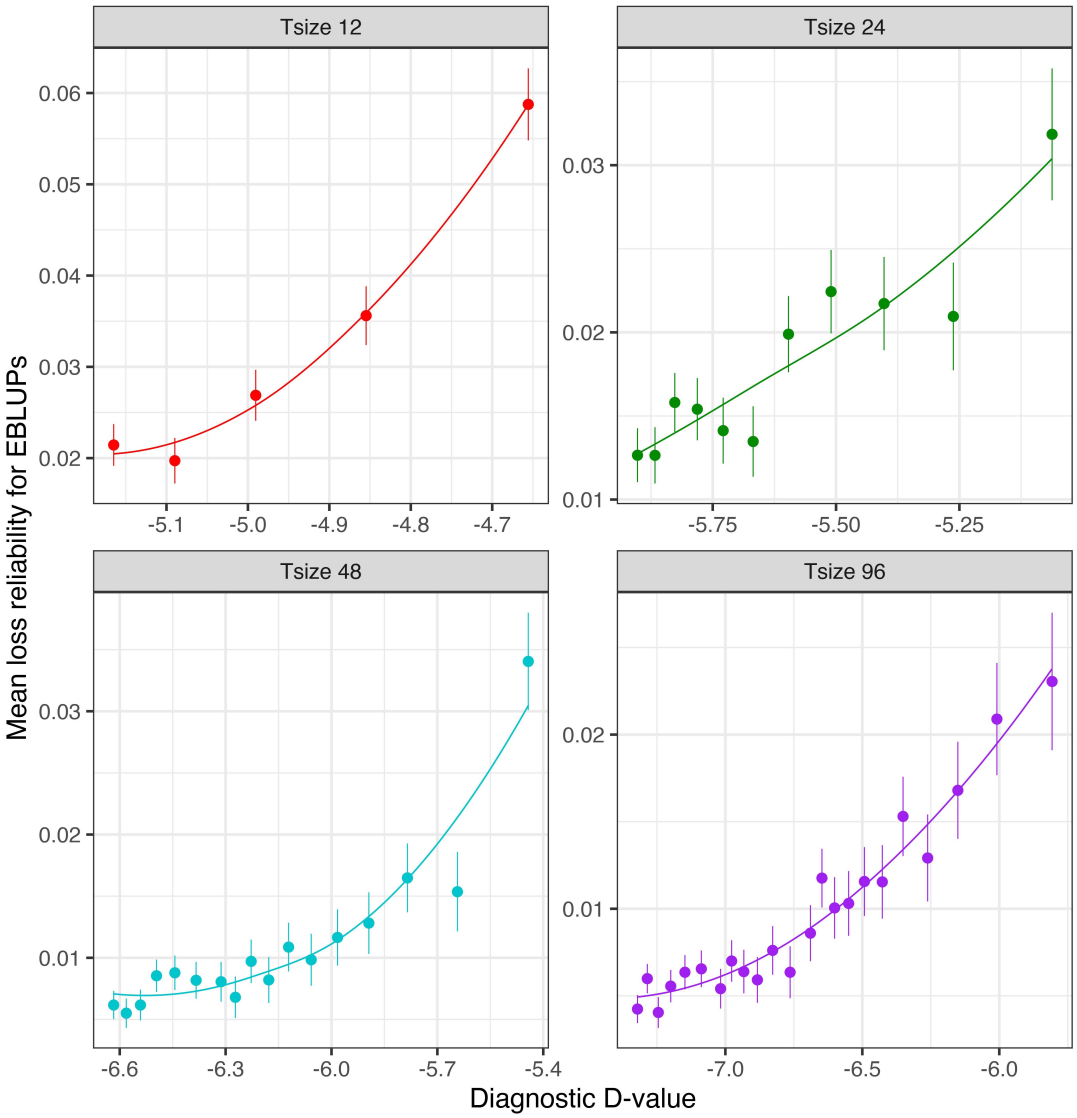

Figure 9.7: Independent VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. Each panel corresponds to a different trial size (trials with $\{12, 24, 48, 96\}$ varieties) and the points correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through the means for each Tsize.

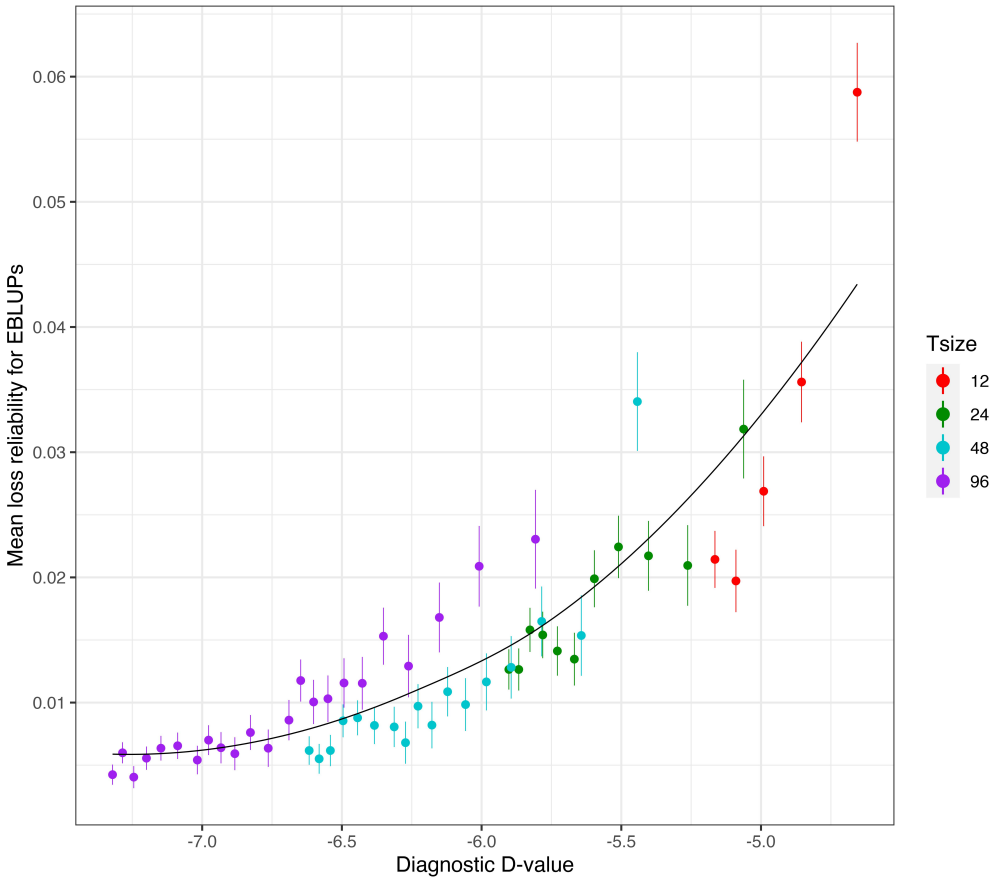

Figure 9.8: Independent VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. The colours correspond to different trial sizes (trials with $\{12, 24, 48, 96\}$ varieties) and the points for each colour correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through all the means.

9.2.6.2 LMM with pedigree information

We extend 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 consider the LMM as in Equation (9.1) but without the non-additive VE effects, so that as in section 9.2.6.1, $n_{\kappa_g} = 3$. The set-up for the study is the same as in section 9.2.6.1 but we only consider the two larger trial sizes along with two additional larger trial sizes of 192 and 384, namely $d_1(= d_2) = \{48, 96, 192, 384\}$ so that $n_1(= n_2) = \{144, 288, 576, 1152\}$. Across the range of connectivity levels the total number of varieties required for the simulation is 767 (corresponding to $x_{1,2} = 2$ for the trial size of 384) and we label these as V1-V767. The simulation study requires a numerator relationship matrix ($\boldsymbol{A}$) for these varieties. We therefore chose V1-V767 from the actual lines in Stage 3 (S3), Stage 4 (S4) and Stage 2 (S2) in 2018 and 2017 in the Durum data and computed $\boldsymbol{A}$ from the associated pedigree information. For the chosen subset of varieties, the inbreeding coefficient ranged from 0.750 to 0.998 with a mean of 0.925.

For the purposes of both the calculation of $\mathcal{D}_{1c}(\mathrm{A})$ and of data generation in the simulation study we chose the values of the variance parameters to be:
$\boldsymbol{\kappa}_{\boldsymbol{g}_0} = \left(\sigma_{a11_0} = 0.1, \sigma_{a12_0} = 0.08, \sigma_{a22_0} = 0.1\right)^\top$ and $\boldsymbol{\kappa}_{\bar{\boldsymbol{g}}_0} = \left(\sigma^2_{1_0} = 0.15, \sigma^2_{2_0} = 0.15\right)^\top$

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 related to the number of varieties in common between the two environments (see Figure 9.9). However, a key difference was that the number of successful model fits for the connectivity level of $x_{1,2} = 2$ was reasonable so these results have been included in what follows.

The results are presented in the same format as in Section 9.2.6.1. The good agreement shown in Figure 9.10 (a) 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 9.10 (b) 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 9.11) and across trial sizes (Figure 9.12).

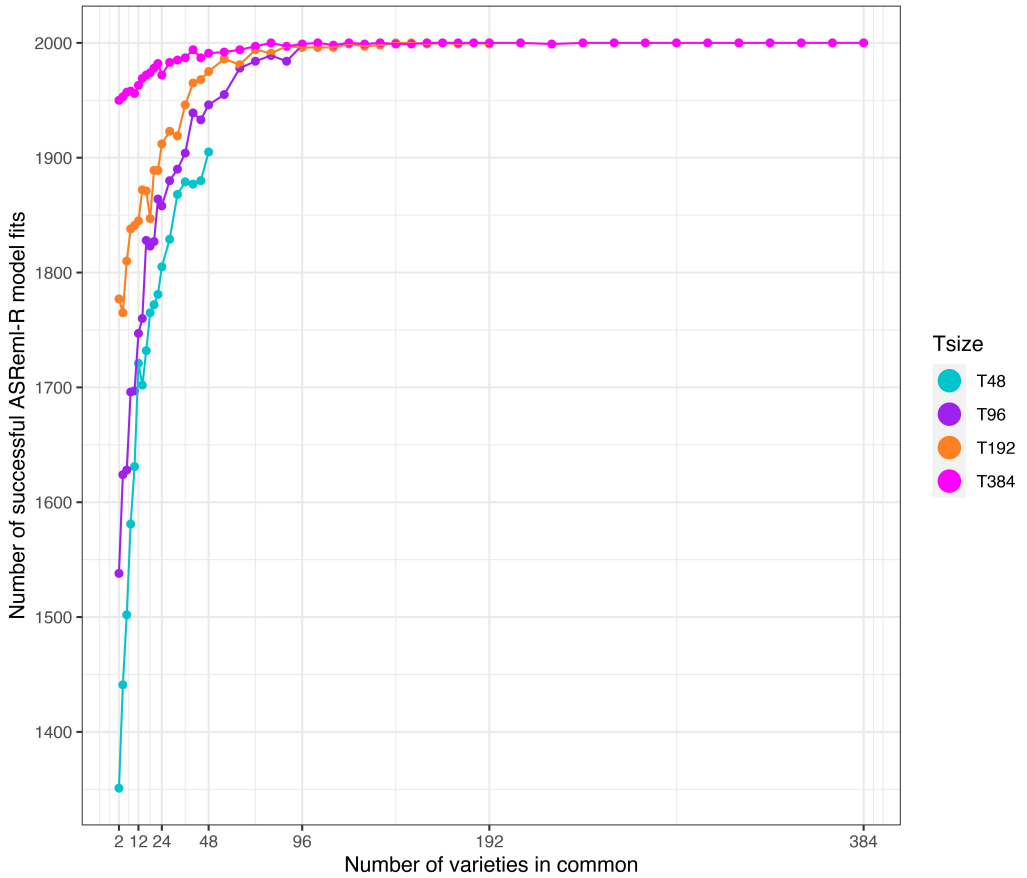

Figure 9.9: Additive VE effects simulation study: number of successful model fits from $N = 2000$ simulations plotted against number of varieties in common for four trial sizes (trials with $\{48, 96, 192, 384\}$ varieties). Trial sizes (Tsize) are represented using different colours. Each point within Tsize corresponds to a different level of variety connectivity which ranges from $x_{1,2} = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

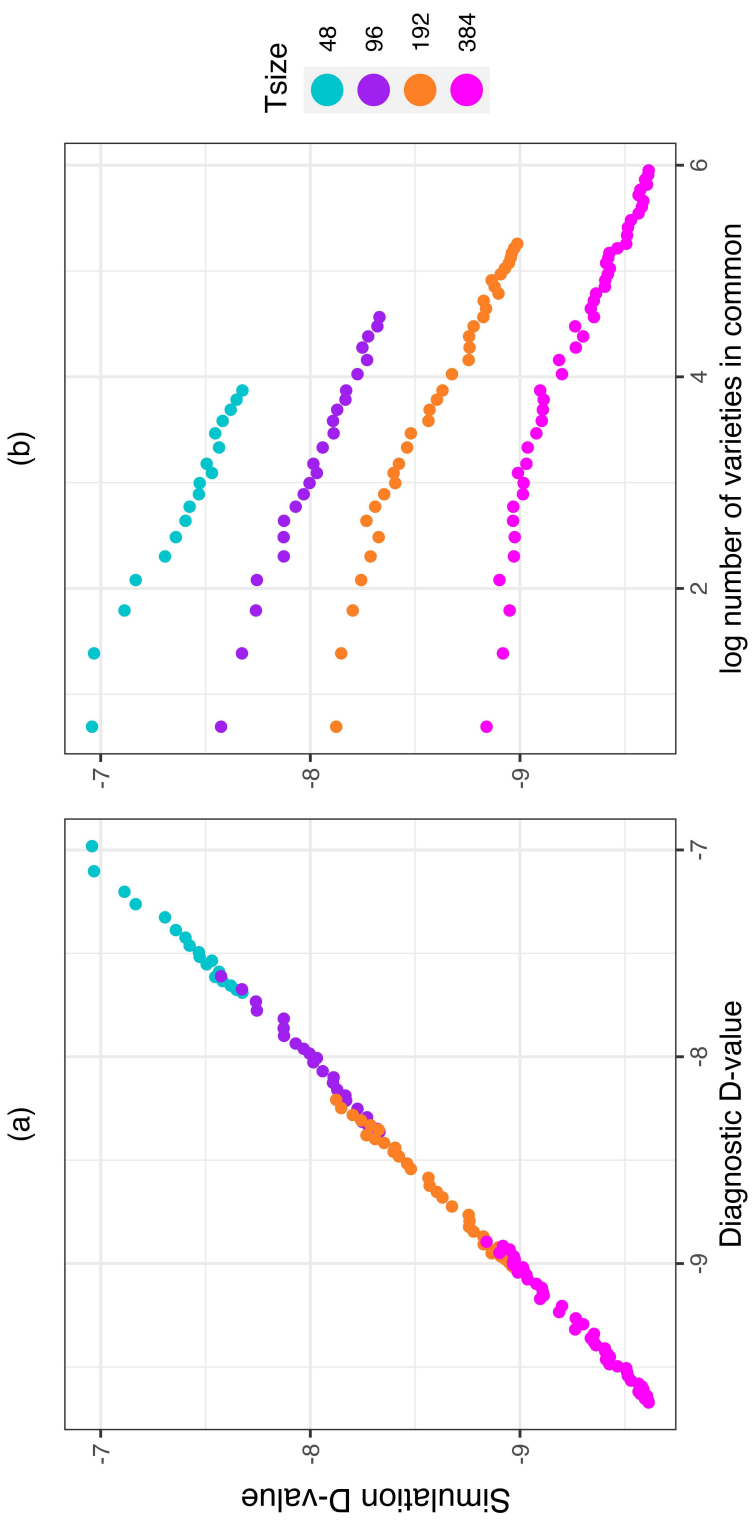

Figure 9.10: Additive VE effects simulation study: simulation based $\mathcal{D}_{1c}^{s}(\mathrm{A})$-values plotted against (a) diagnostic $\mathcal{D}_{1c}(\mathrm{A})$-values and (b) log number of varieties in common for four trial sizes (trials with $\{48, 96, 192, 384\}$ varieties) and a sequence 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 $x_{1,2} = 2$ up to the number of varieties in a trial (representing 100% connectivity between the two trials).

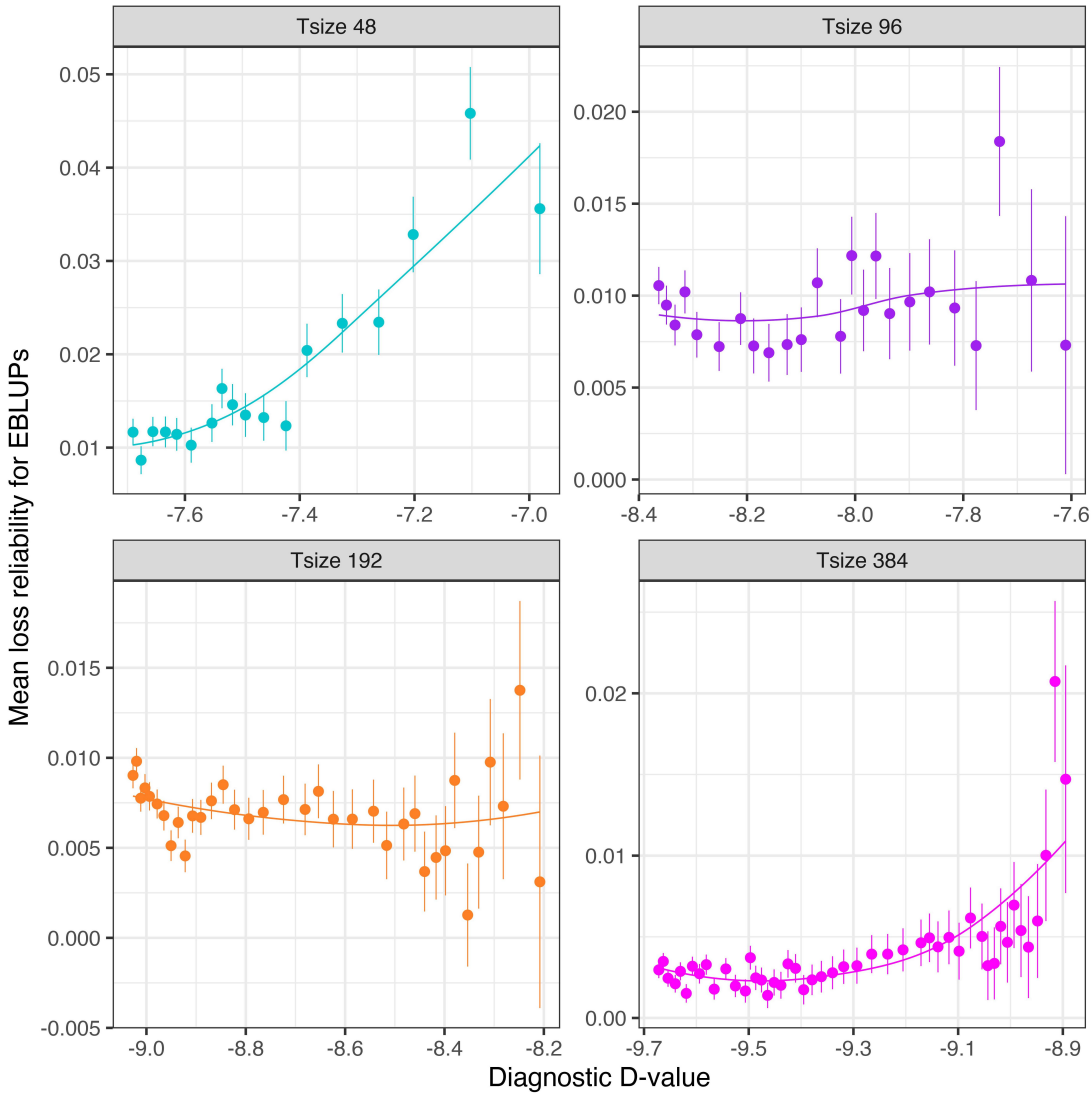

Figure 9.11: Additive VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. Each panel corresponds to a different trial size (trials with $\{48, 96, 192, 384\}$ varieties) and the points correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through the means for each Tsize.

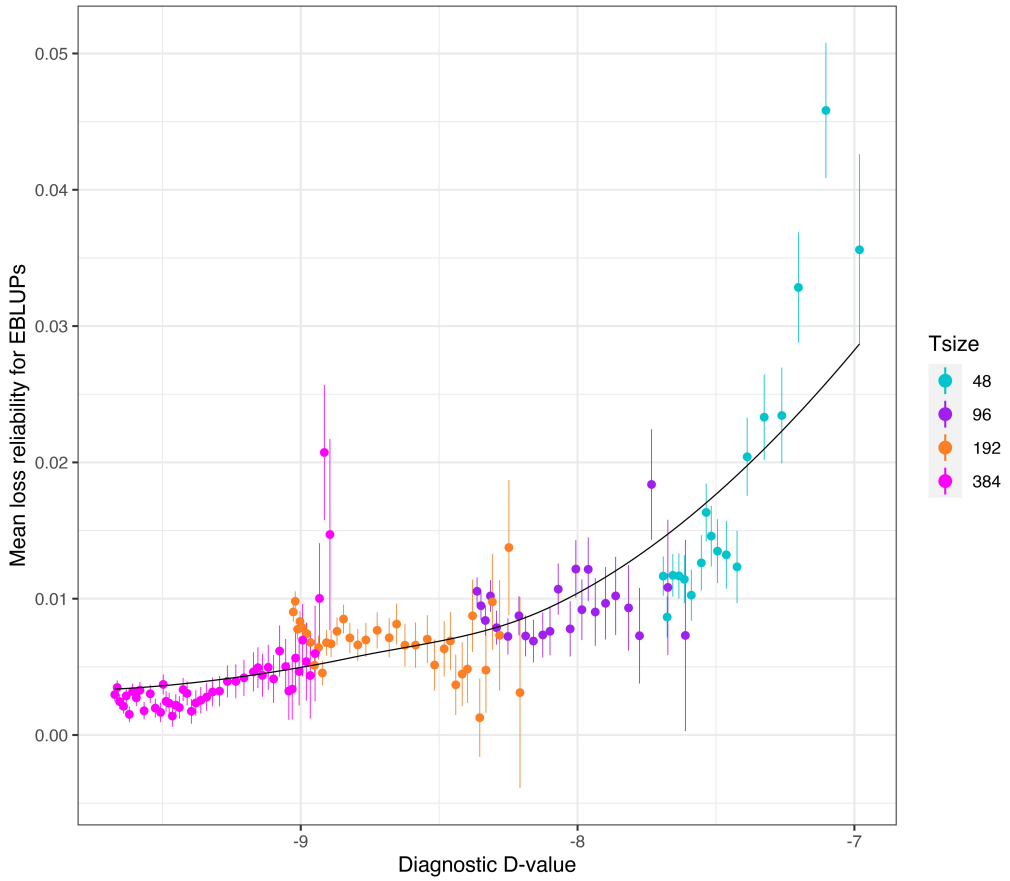

Figure 9.12: Additive VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. The colours correspond to different trial sizes (trials with $\{48, 96, 192, 384\}$ varieties) and the points for each colour correspond to a sequence of connectivity levels. Also shown are standard errors for each mean (vertical lines) and a loess smoother through all the means.

9.2.6.3 Robustness of diagnostic

We again investigate the robustness of the diagnostic to a change in variance parameters. Figures 9.13, 9.14, 9.15 and 9.16 show results from a simulation study where we have set the additive genetic variance to 40% of the total genetic variance for each environment and therefore 60% for the non-additive genetic variance, and a between environments correlation of 0.4 for both additive and non-additive VE effects. These values are at the lower end of those seen in practice and are presented here to show the robustness of the diagnostic. The results highlight similar trends to the original simulation as already presented.

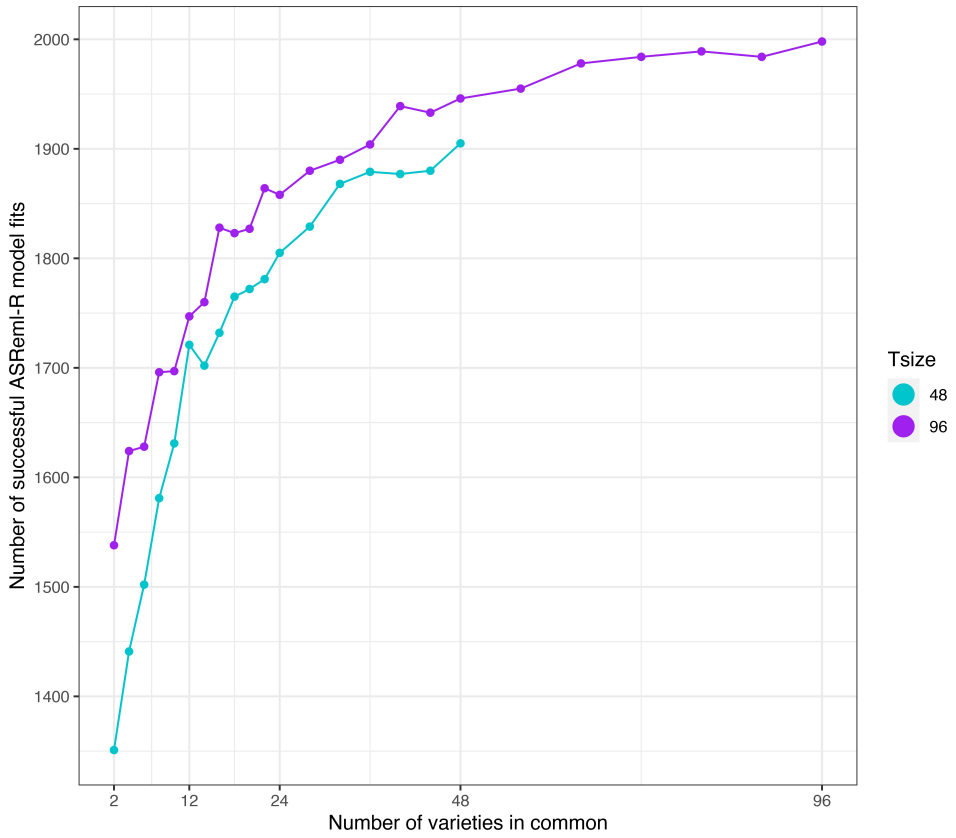

Figure 9.13: Low value scenario: Additive VE effects simulation study: number of successful model fits from $N = 2000$ simulations plotted against number of varieties in common for two trial sizes (trials with 48 and 96 varieties). Trial sizes (Tsize) are represented using different colours. Each point 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).

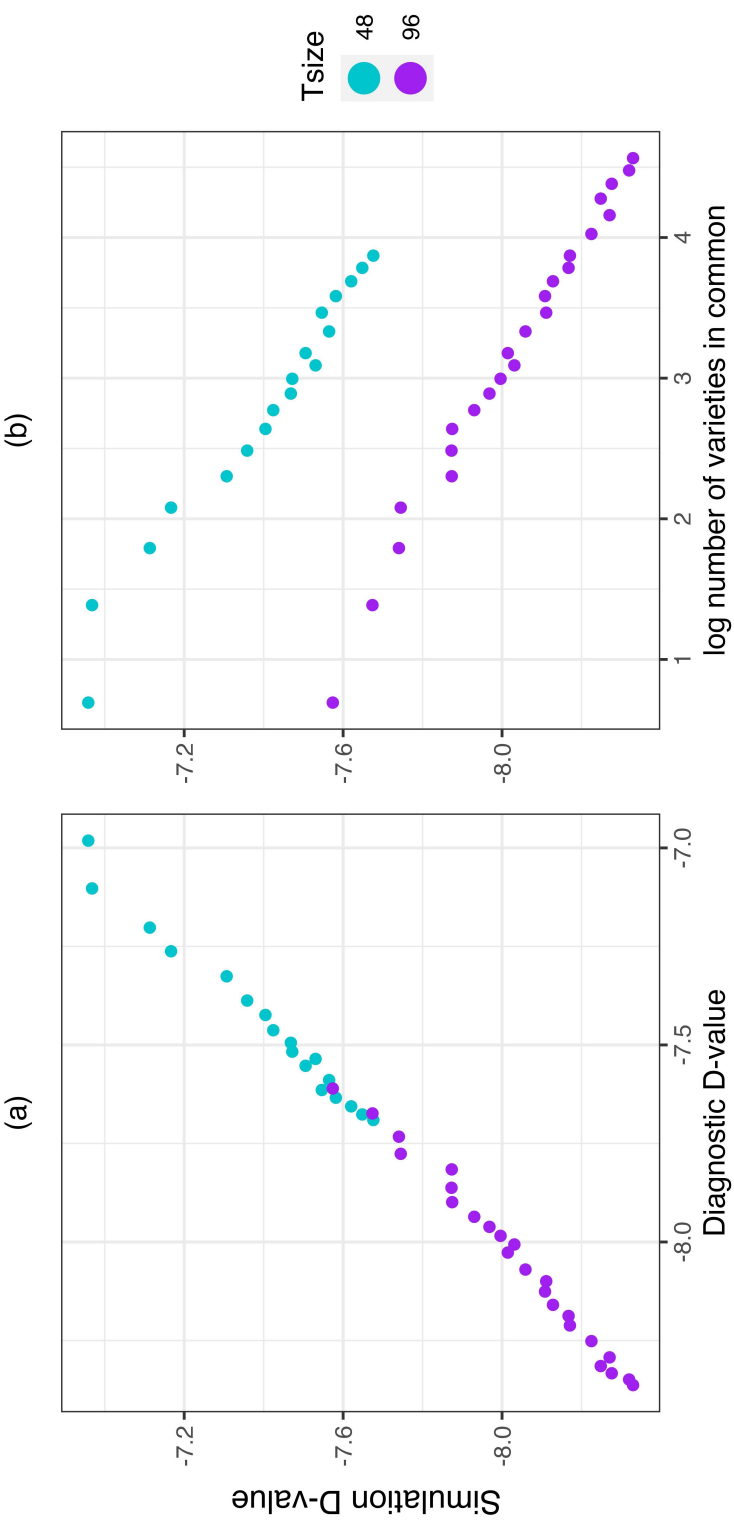

Figure 9.14: Low value scenario: Additive VE effects simulation study: simulation based $\mathcal{D}^s_{1c}(\mathrm{A})$-values plotted against (a) diagnostic $\mathcal{D}_{1c}(\mathrm{A})$-values and (b) log number of varieties in common for two trial sizes (trials with 48 and 96 varieties) and a sequence 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).

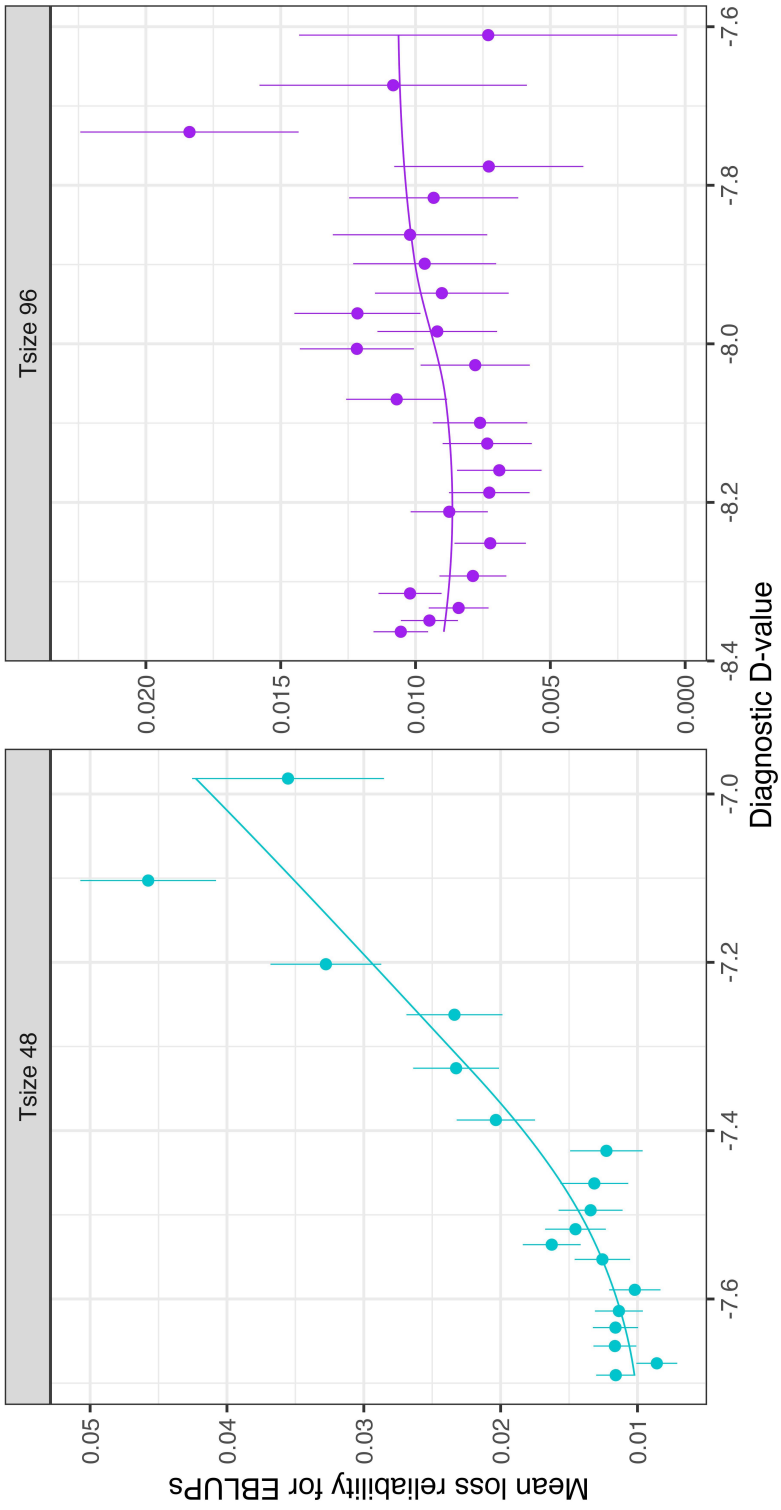

Figure 9.15: Low value scenario: Additive VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. Each panel corresponds to a different trial size (trials with 48 and 96 varieties) and the points correspond to a sequence of connectivity levels. Also shown is a loess smoother through the means for each Tsize.

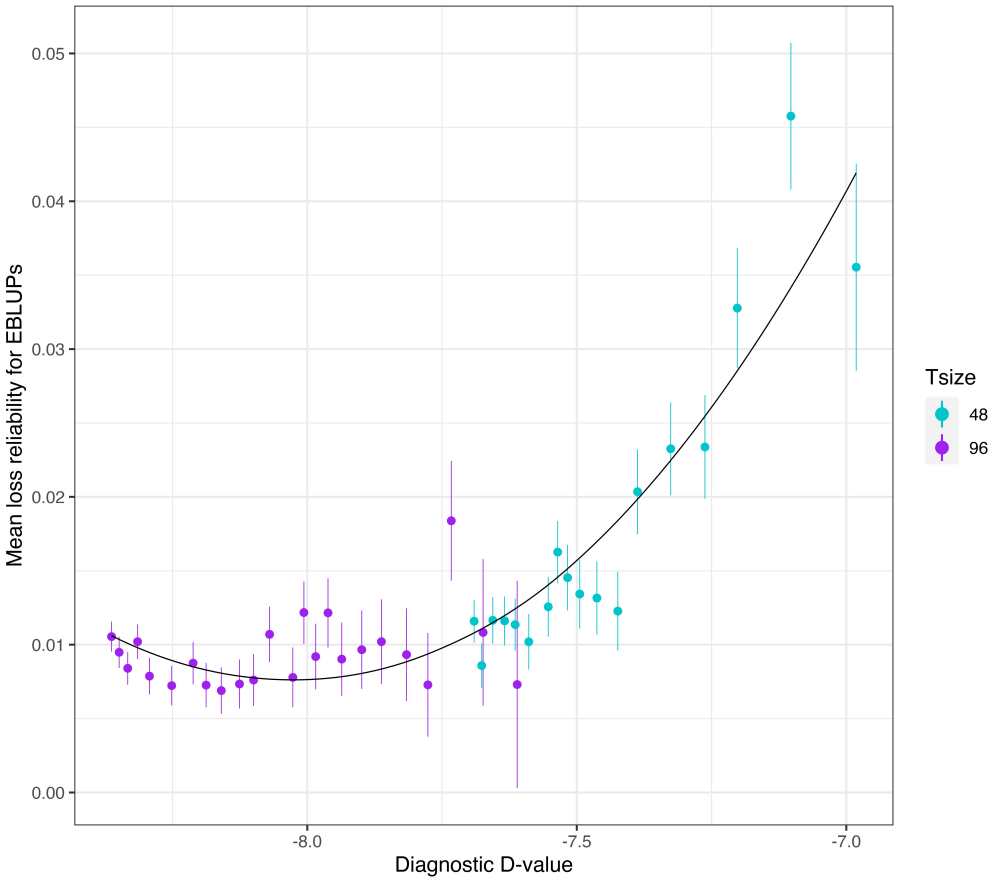

Figure 9.16: Low value scenario: Additive VE effects simulation study: mean loss in reliability of the EBLUPs of VE effects for Env1 for those varieties that were present in both environments. The colours correspond to different trial sizes (trials with 48 and 96 varieties) and the points for each colour correspond to a sequence of connectivity levels. Also shown is a loess smoother through all the means.

9.3 Calculating $\mathcal{D}$-values in practice

For illustration to calculate $\mathcal{D}$-values, I present in this section the `R` code to calculate the $\mathcal{D}_j$(A+I)-values given in the first column of Table 9.4. Note that the `R` function `DoptFull.tot` is given in Appendix A.

```
library(msos)
library(Matrix)
library(gtools)
library(ggplot2)
library(ggrepel)

durumEI.Tot.S3 <- DoptFull.tot(data.df=s3diag.df, sig.vmvar=0.1,
sig.vmcov=0.08, sig.idevar=0.05, sig.idecov=0.04, sigm=0.15,
Efac='Environment', Gfac = 'Variety',Amat=Amat)

M <- solve(durumEI.Tot.S3)
ld.mat.T <- matrix(data=NA, ncol=2,nrow=length(envs))
rownames(ld.mat.T) <- envs
colnames(ld.mat.T) <- c('ld.v1','ld.v2')
for(i in 1:length(envs)){
V11 <- M[grep(envs[i],rownames(M)),grep(envs[i],
colnames(M))];dim(V11)
V12 <- M[grep(envs[i],rownames(M)),grep(envs[i],
colnames(M),invert=T)];dim(V12)
V21 <- M[grep(envs[i],rownames(M),invert=T),grep(envs[i],
colnames(M))];dim(V21)
V22 <- M[grep(envs[i],rownames(M),invert=T),grep(envs[i],
colnames(M),invert=T)];dim(V22)

ld.mat.T[i,2] <- logdet(V22 - V21%*%solve(V11)%*%V12)
ld.mat.T[i,1] <- logdet(V11)
}
round(ld.mat.T[,1]/26,2) #26 parameters
# 2014-BREEZA  2014-TWORTH 2015-EDGEROI  2015-TWORTH
#       -7.56        -7.43        -7.95        -8.27
#2016-BREEZA   2016-NSTAR  2016-TWORTH  2017-BREEZA
#       -7.51        -7.51        -8.28        -7.65
#2017-NSTAR  2017-TWORTH  2018-BREEZA  2018-GURLEY
#       -7.65        -8.03        -7.39        -7.39
#2018-TWORTH
#       -7.39
```

9.4 Concluding remarks

In this chapter we have developed a diagnostic to be applied to a MET dataset prior to analysis in order to assess the reliability of genetic variance parameter estimates, both for the dataset across all environments and for individual environments. Similar to the methods outlined in Chapter 7 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 the 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 calculated prior to the conduct of a MET analysis to investigate the likely reliability of genetic variance parameter estimation. Although this measure is simple to compute and intuitively reasonable, there has been little in the literature to validate its use. This was investigated in Chapter 6 which showed that variety connectivity was influential but there were other factors at play. This is also shown in the simulation studies within this chapter but I note 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). In contrast, the new $\mathcal{D}$-optimality diagnostic predicted reliability across both connectivity levels and trial sizes. The application to the Durum dataset 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, variety replication as well as variety relationships when available.

The simulation study results suggest 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 the 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 practice, 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.

Finally, as demonstrated in Chapter 7, the CG methodology shows the information benefits of using MET datasets that span stages and years. However as a result, this approach may lead to datasets with poor levels of variety connectivity. To determine the best MET dataset, we use the $\mathcal{A}$-value criterion (see Section 7.1) to analyse the range of datasets, as this criterion has been shown to align with the probability of making incorrect selection decisions (Bueno Filho & Gilmour, 2003, 2007). Calculation of the $\mathcal{A}$-values is based on an LMM with known variance parameter values, however in practice these must be estimated from the data. As illustrated in this chapter and in Chapter 6, I have shown the link between the reliability of variance parameter estimation and the reliabilities of the corresponding VE effect predictions. As a consequence, I advocate combining the $\mathcal{A}$ and $\mathcal{D}$-value approaches in order to balance variety information and reliability of variance parameter estimation in the search for an optimum MET dataset.

Chapter 10

Conclusion

Despite the fact that METs are an important part of plant breeding, and although sophisticated and relevant statistical analyses, as detailed in Chapters 1 and 2, have been proven to increase the reliability of the predicted variety effects for individual environments (VE effects), there has been little research into how best to construct an appropriate dataset in terms of deciding which trials should be included. Under the paradigm of the factor analytic linear mixed model (FALMM) methodology of Smith et al. (2001b) for MET analysis, the objective is for the reliable prediction of the VE effects. These predictions can be meaningfully summarised across environments, for example, using the interaction class methodology of Smith & Cullis (2021).

The objectives of this thesis were to present novel approaches for optimising the construction of MET datasets from a series of plant variety trials. Two motivating datasets described in Chapter 4 were utilised to exemplify these objectives. The first is an Oat dataset and the other is a Durum wheat dataset. The former is used as an example of a dataset with independent VE effects, whereas the latter is used as an example of a dataset with related VE effects. These, along with their statistical analyses given in Chapters 5 and 8 respectively, are used for their attributes in the development of real-world grounded simulation studies (see Chapter 3 for their methodology) and diagnostic tools, with the objectives of investigating recognised concerns and providing superior methodology to existing approaches for the construction of MET datasets.

10.1 Summary of research

The breeding process is a progressive system that revolves around the evaluation and selection of superior varieties. This naturally results in datasets with varied levels of balance in terms of the number of varieties in common between environments, a metric known as "variety connectivity". It was long believed that this 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. It is well-known that poorly estimated genetic variance parameters will result in a reduction in genetic gain (Sales & Hill, 1976a,b). Historically, to combat these concerns, environments were often removed from the MET dataset if they appeared to have insufficient numbers of varieties in common with other environments (Smith et al., 2001a, 2015; Ward et al., 2019, for example). However, there has been little research to establish whether this approach is the most appropriate method for this purpose.

These concerns are first investigated through a real-world grounded simulation study in Chapter 6, for the case of independent VE effects. The results from this simulation study demonstrated the intricate linkages between genetic scenarios and variety connectivity. In particular, 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). As a result, the typical variety connectivity approach was demonstrated to be inadequate for constructing MET datasets.

We then provided a systematic approach for the construction of MET datasets for selection in plant breeding programs in Chapter 7. This methodology describes the structure of MET datasets, with the focus on identifying groups of varieties that entered the first stage of testing in the same year, which are denoted as contemporary groups (CGs), and also the establishment of data bands, which are related to trials. This enabled a thorough, and complete listing of the trials in which the varieties of interest were grown and their progression between years and stages. To quantify different MET datasets, we employed the $\mathcal{A}$-optimality criterion from model-based design theory, since this aligns with minimising the probability of an incorrect selection decision (Bueno Filho & Gilmour, 2003, 2007). This measure was applied to variety effects which

demonstrated the importance of adding as many trials as necessary to capture the selection histories of the varieties under consideration. This approach was demonstrated to be superior to other commonly used dataset construction strategies when applied to the Durum dataset presented in Chapter 4, which had four selection decisions. 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 it assumes that the associated variance parameters are known. The approach for increasing the amount of data available for the varieties under consideration was shown to be simple and straightforward, with only a few steps.

It is important to distinguish between 'direct' and 'indirect' information in terms of information accessible for selection decisions. The former is concerned with observable data and is maximised by considering all trials in which the varieties of interest have been grown. The use of a FALMM for analysis with pedigrees, opens up the option of utilising indirect information gained from genetically related varieties. However, it was noted that combining trials across stages and years may result in an unbalanced dataset, as demonstrated for the Durum dataset with 6% of variety by environment combinations observed (see Table 1.1). Whilst one of the advantages of using a FALMM is the ability to handle unbalanced data, I again raised concern about the reliability of estimation of genetic variance parameters in extreme cases when variety connectivity is poor, and thus sought a diagnostic that was superior to traditional connectivity measures.

It was therefore critical to assess the influence of the structure of a MET dataset on the reliability of genetic variance parameter estimation, since this may affect the reliability of variety predictions. As a result, in Chapter 9, a formal diagnostic tool was developed that can be applied to a MET dataset prior to analysis to examine the potential reliability of genetic variance parameter estimates. As in the CG methodology, I again used a model-based design approach, and apply $\mathcal{D}$-optimality to the genetic variance parameters. Computation of the diagnostic measure requires specification of variance parameters, namely the genetic and error variances for individual environments and genetic correlations between pairs of environments. As these values are not known or how they differ between environments prior completing the MET analysis, a pragmatic and sensible approach is used that assumes homogeneity between environments. These

parameter values can be chosen to reflect typical estimates obtained in practice. However, in the application to the Durum dataset, the diagnostic tool is shown to be robust to the specification of these values.

Two simulation studies are conducted in Chapter 9, one employing a LMM with independent VE effects and the other with additive VE effects. They demonstrated that the diagnostic $\mathcal{D}$-values performed well in the sense of predicting the actual reliability of genetic variance parameter estimates, where lower diagnostic $\mathcal{D}$-values represent higher levels of information to estimate the parameters of interest. The diagnostic was shown to encapsulate numerous other structural features of a MET dataset that are influential in determining the reliability of genetic variance parameter estimation. These features include, but are not limited to: variety connectivity; trial size; variety replication; and when available, variety relatedness. Furthermore, the results showed 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 of 12 varieties resulted in larger $\mathcal{D}$-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 scenarios in which there were more varieties in each trial. This is consistent with our experience in analysing MET datasets with small numbers of varieties. Even when the connectivity between these and larger trials is high, there are often computational difficulties in fitting the FALMM.

The problem with environments which have a small number of varieties leads to a broader issue for MET datasets arising from crop variety evaluation programs such as the NVT. In contrast to plant breeding datasets, all trials in evaluation programs have relatively few varieties. Three examples were given in Table 1.1, corresponding to NVT Wheat and Lentil datasets. In the two Wheat datasets (Smith et al., 2015; Gogel et al., 2018), the median number of varieties per trial was 47 in both cases, whereas for Lentils (NVT-online) the median number was 15 with a maximum number of varieties per trial of only 17. In the early years of NVT (2005 – circa 2013), the MET analysis was of the form used in Cullis et al. (1996a,b), namely a variance component model with variety by environment interaction partitioned into sources associated with locations and years. As discussed in this thesis, such a model rarely provides a good fit to the data

in the sense of accounting for variety by environment interaction, but is parsimonious, with very few variance parameters to be estimated. The success of FALMM in plant breeding applications led to its adoption for NVT data (Smith et al., 2015) and this is the current method used for all crops.

10.2 Future direction

In practice, I suggest both the CG and $\mathcal{D}$-optimality methodologies to be used to construct optimal MET datasets for use in plant breeding programs, with environments with large $\mathcal{D}$-values considered for exclusion as these may contribute insufficient information for genetic variance parameter estimates. Furthermore, I note that these approaches outlined in this thesis are presently applied in numerous private breeding programs in Australia and overseas.

The pre-specified LMM for the calculation of the $\mathcal{D}$-optimality diagnostic shown in this thesis utilises a fully saturated form of the genetic variances and covariances within and between environments. It therefore targets MET analyses that employ factor analytic, or unstructured forms for the genetic variance matrices. I note that it is reasonably simple to modify the pre-specified LMM to reflect simpler models, such as a variance component model. Additionally, the partitioning of the genetic effects into additive and non-additive was achieved in this thesis using pedigree information, but the modification to use genomic marker data is straightforward.

The research finding in this thesis concerning the importance of trial size (number of varieties per trial) on the reliability of genetic variance parameter estimates raises concerns about the use of FALMM for datasets such as the NVT Lentils data (see Table 1.1). Thus, leading on from this thesis is the need to investigate whether there is sufficient information in such datasets to reliably estimate the factor analytic variance parameters and whether the resultant variety predictions are more or less accurate compared with those obtained by fitting the simplistic but parsimonious variance component model.

10.3 Final remarks

The main objectives of this thesis were to address a void in the literature, and provide a rigorous and formal framework for the optimal construction of MET datasets for selection in plant breeding programs. These have been accomplished by striking the balance between maximising the variety information through the use of the CG methodology, and maximising the reliability of variance parameter estimates through the use of the diagnostic $\mathcal{D}$-value. It is intended that the improvements in the reliabilities of the VE effects demonstrated in this thesis will lead to more widespread use of these approaches in plant breeding programs. Thus, leading to an increase in genetic gain.

Appendices

Appendix A

R functions

The R code presented here provides the function to calculate the expected information matrix for the total effects, that is, for additive and non-additive effects, as given in Chapter 9.

```
DoptFull.tot <- function(data.df, sig.vmvar, sig.vmcov, sig.idevar,
sig.idecov, sigm, Efac = "env",Gfac = "geno", Amat = Amat) {
####################
#This function calculates EI for the total genetic
effects parameters, that is for additive + non-additive.
####################
#data.df = datafile which contains the structural information,
#environments, varieties.
#sig.vmvar = Additive variance for each environment
#sig.vmcov = Additive covariance between environments
#sig.idevar = Non-additive variance for each environment
#sig.idecov = Non-additive covariance between environments
#sigm = Error variance
#Efac = Name of environment factor in datafile
#Gfac = Name of the variety factor in datafile
#Amat = The NRM for all varieties.

nrep<-table(data.df[[Gfac]],data.df[[Efac]])
dv <- dimnames(nrep)[[1]]
de <- dimnames(nrep)[[2]]
data.df2 <- data.frame(nrep=as.vector(nrep))
data.df2$Env <- factor(rep(de,each=length(dv)))
data.df2$Geno <- factor(rep(dv,length(de)))
data.df2 <- data.df2[data.df2$nrep>0,]
```

```
m <- length(levels(data.df2$Geno)) # 6
e <- length(levels(data.df2$Env))
# genos in each site ...
gnam <- list()
for (i in 1:length(de)) {
gnam[[i]]<-as.character(unique(data.df2$Geno[data.df2$Env==de[i]]))
}
names(gnam) <- de
Gamat <- matrix(sig.vmcov,nrow=length(de),ncol=length(de))
diag(Gamat) <- sig.vmvar
dimnames(Gamat) <- list(de,de)

Gemat <- matrix(sig.idecov,nrow=length(de),ncol=length(de))
diag(Gemat) <- sig.idevar
dimnames(Gemat) <- list(de,de)
Imat <- diag(1,nrow=dim(Amat)[1],ncol=dim(Amat)[2])
rownames(Imat) <- colnames(Imat) <- rownames(Amat)
k <- 0
ZAZ <- ZIZ <- list()
Hmat <- matrix(0,nrow=dim(data.df2),ncol=dim(data.df2))
pind <- pvmtype <- pidetype <- Env1 <- Env2 <- c()
print('Calculating H')
pb <- progress_bar$new(total=length(de))
for (i in 1:length(de)) {
for (j in 1:i) {
k <- k+1
pind[k] <- k
pvmtype[k] <- 'var.vm'
pidetype[k] <- 'var.ide'
Env1[k] <- de[i]
Env2[k] <- de[j]
ZAZ[[paste(de[i],de[j],sep=":")]] <- Amat[gnam[[i]],gnam[[j]]]
ZIZ[[paste(de[i],de[j],sep=":")]] <- Imat[gnam[[i]],gnam[[j]]]
Hmat[data.df2$Env==de[i],data.df2$Env==de[j]] <-
as.matrix(Gamat[de[i],de[j]]*ZAZ[[paste(de[i],de[j],sep=":")]] +
Gemat[de[i],de[j]]*ZIZ[[paste(de[i],de[j],sep=":")]])
if (i!=j) {
pvmtype[k] <- 'cov.vm';   pidetype[k] <- 'cov.ide'
Hmat[data.df2$Env==de[j],data.df2$Env==de[i]] <-
as.matrix(Gamat[de[j],de[i]]*t(ZAZ[[paste(de[i],de[j],sep=":")]]) +
Gemat[de[j],de[i]]*t(ZIZ[[paste(de[i],de[j],sep=":")]]))
```

```
}
}
Hmat[data.df2$Env==de[i],data.df2$Env==de[i]] <-
Hmat[data.df2$Env==de[i],data.df2$Env==de[i]] +
sigm*diag(1/data.df2$nrep[data.df2$Env==de[i]])
pb$tick()
}
p.vm.ind.df <- data.frame(pind=pind,ptype=pvmtype,
Env1=paste0(Env1,".vm"),Env2=paste0(Env2,".vm"))
p.ide.ind.df <- data.frame(pind=pind,ptype=pidetype,
Env1=paste0(Env1,".ide"),Env2=paste0(Env2,".ide"))

pars.n <- c(paste0(names(ZAZ),".vm"),paste0(names(ZIZ),".ide"))
params <- as.data.frame(combinations(n=length(pars.n),r=2,
v=pars.n,repeats.allowed=T))
tmp <- strsplit(params$V1,":")
params$Env1_V1 <- sapply( tmp, "[", 1)
params$Env2_V1 <- sapply( tmp, "[", 2)
params$Env2_V1 <- gsub(".ide","",params$Env2_V1)
params$Env2_V1 <- gsub(".vm","",params$Env2_V1)
tmp <- strsplit(params$V2,":")
params$Env1_V2 <- sapply( tmp, "[", 1)
params$Env2_V2 <- sapply( tmp, "[", 2)
params$Env2_V2 <- gsub(".ide","",params$Env2_V2)
params$Env2_V2 <- gsub(".vm","",params$Env2_V2)

XX <- model.matrix(~data.df2$Env-1)
Hinv <- solve(Hmat)
Pmat <- Hinv - Hinv%*%XX%*%solve(t(XX)%*%Hinv%*%XX)%*%t(XX)%*%Hinv
dimEI <- length(ZAZ)+length(ZIZ)
EI <- matrix(0,nrow=dimEI,dimEI)
dimnames(EI)[[1]] <- dimnames(EI)[[2]] <- pars.n

print('Calculating EI')
pb <- progress_bar$new(total=nrow(params))
for (i in 1:nrow(params)) {
Hdoti <- matrix(0,nrow=dim(data.df2),ncol=dim(data.df2))
if(length(grep('.vm',params$V1[i]))==1){

Hdoti[data.df2$Env==params$Env1_V1[i],
data.df2$Env==params$Env2_V1[i]] <-
```

```
as.matrix(ZAZ[[paste(params$Env1_V1[i],
params$Env2_V1[i],sep=":")]])

Hdoti[data.df2$Env==params$Env2_V1[i],
data.df2$Env==params$Env1_V1[i]] <-
as.matrix(t(ZAZ[[paste(params$Env1_V1[i],
params$Env2_V1[i],sep=":")]]))
}
if(length(grep('.ide',params$V1[i]))==1){
Hdoti[data.df2$Env==params$Env1_V1[i],
data.df2$Env==params$Env2_V1[i]] <-
as.matrix(ZIZ[[paste(params$Env1_V1[i],
params$Env2_V1[i],sep=":")]])

Hdoti[data.df2$Env==params$Env2_V1[i],
data.df2$Env==params$Env1_V1[i]] <-
as.matrix(t(ZIZ[[paste(params$Env1_V1[i],
params$Env2_V1[i],sep=":")]]))
}
PHdoti <- Pmat%*%Hdoti
#####
Hdotj <- matrix(0,nrow=dim(data.df2),ncol=dim(data.df2))
if(length(grep('.vm',params$V2[i]))==1){
Hdotj[data.df2$Env==params$Env1_V2[i],
data.df2$Env==params$Env2_V2[i]] <-
as.matrix(ZAZ[[paste(params$Env1_V2[i],
params$Env2_V2[i],sep=":")]])

Hdotj[data.df2$Env==params$Env2_V2[i],
data.df2$Env==params$Env1_V2[i]] <-
as.matrix(t(ZAZ[[paste(params$Env1_V2[i],
params$Env2_V2[i],sep=":")]]))
}
if(length(grep('.ide',params$V2[i]))==1){
Hdotj[data.df2$Env==params$Env1_V2[i],
data.df2$Env==params$Env2_V2[i]] <-
as.matrix(ZIZ[[paste(params$Env1_V2[i],
params$Env2_V2[i],sep=":")]])

Hdotj[data.df2$Env==params$Env2_V2[i],
data.df2$Env==params$Env1_V2[i]] <-
```

```
as.matrix(t(ZIZ[[paste(params$Env1_V2[i],
params$Env2_V2[i],sep=":")]]))
}
PHdotj <- Pmat%*%Hdotj
EI[params$V1[i],params$V2[i]] <- EI[params$V2[i],
params$V1[i]] <- sum(PHdoti*t(PHdotj))
pb$tick()
}
return(EI.mat=EI/2)}
```

References

Acquaah, G. (2013). *Principles of Plant Genetics and Breeding*. John Wiley and Sons, New York. 1

Ankenman, B., Avilés, A., & Pinheiro, J. (2003). Optimal designs for mixed-effects models with two random nested factors. *Statistica Sinica* **13**, 385–401. 154

Arief, V., DeLacy, I., Crossa, J., Payne, T., Singh, R. S., Braun, H., Tian, T., Basford, K., & Dieters, M. (2015). Evaluating testing strategies for plant breeding field trials: redesigning a CIMMYT international wheat nursery. *Crop Science* **55**, 164–177. 11

Arief, V., Desmae, H., Hardner, C., DeLacy, I., Gilmour, A. R., Bull, J., & Basford, K. (2019). Utilization of multiyear plant breeding data to better predict genotype performance. *Crop Science* **59**, 1–11. 115

Beeck, C. P., Cowling, W. A., Smith, A., & Cullis, B. R. (2010). Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic mixed models with pedigree information. *Genome* **53**, 992–1001. 7, 10, 133

Bernal-Vasquez, A., Gordillo, A., Schmidt, M., & Piepho, H. P. (2017). Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC genetics* **18**, 1–17. 115

Bueno Filho, J. & Gilmour, S. (2003). Planning incomplete block experiments when treatments are genetically related. *Biometrics* **59**, 375–381. 12, 116, 117, 120, 193, 196

Bueno Filho, J. & Gilmour, S. (2007). Block designs for random treatment effects. *Journal of Statistical Planning and Inference* **137**, 1446–1451. 12, 116, 117, 193, 196

REFERENCES

Burghout, W. (2004). A note on the number of replication runs in stochastic traffic simulation models. Technical report. 42

Butler, D. (2013). *On the optimal design of experiments under the linear mixed model.* PhD thesis, The University of Queensland. 92, 116, 154

Butler, D. (2016). Package 'pedicure'. 55, 134

Butler, D., Cullis, B. R., Gilmour, A. R., Gogel, B., & Thompson, R. (2017). ASReml-R Reference Manual Version 4. 27, 41, 80, 96, 100, 121, 175

Butler, D., Smith, A., & Cullis, B. R. (2014). On the Design of Field Experiments with Correlated Treatment Effects. *Journal of Agricultural, Biological, and Environmental Statistics* **19**, 539–555. 120

Cochran, W. G. & Cox, G. M. (1950). *Experimental designs.* John Wiley and Sons, New York. 2

Cocks, N., March, T., Biddulph, T., Smith, A., & Cullis, B. R. (2019). The provision of grower and breeder information on the frost susceptibility of wheat in Australia. *The Journal of Agricultural Science* **157**, 382–398. 10

Compagner, A. (1995). Operational conditions for random-number generation. *Physical Review E* **52**, 5634–5645. 41

Cressie, N. & Lahiri, S. N. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* **45**, 217–233. 21

Cullis, B. R. & Gleeson, A. C. (1991). Spatial analysis of field experiments - an extension to two dimensions. *Biometrics* **47**, 1449–1460. 4

Cullis, B. R., Gogel, B., Verbyla, A., & Thompson, R. (1998). Spatial analysis of multi-environment early generation variety trials. *Biometrics* **54**, 1–18. 4, 5, 6, 7

Cullis, B. R., Jefferson, P., Thompson, R., & Smith, A. (2014). Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. *Theoretical and Applied Genetics* **127**, 2193–2210. 10, 13

Cullis, B. R., Smith, A., & Coombes, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 381–393. 53, 116

Cullis, B. R., Smith, A., Hunt, C., & Gilmour, A. R. (2000). An examination of the efficiency of Australian crop variety evaluation programmes. *Journal of Agricultural Science* **135**, 213–222. 11, 121, 122

Cullis, B. R., Thomson, F. M., Fisher, J. A., Gilmour, A. R., & Thompson, R. (1996a). The analysis of the NSW wheat variety database. I. Modelling trial error variance. *Theoretical and Applied Genetics* **92**, 21–27. 4, 198

Cullis, B. R., Thomson, F. M., Fisher, J. A., Gilmour, A. R., & Thompson, R. (1996b). The analysis of the NSW wheat variety database. II. Variance component estimation. *Theoretical and Applied Genetics* **92**, 28–39. 4, 6, 198

Diffey, S., Smith, A., Welsh, A., & Cullis, B. R. (2017). A new REML (parameter expanded) EM algorithm for linear mixed models. *Australian and New Zealand Journal of Statistics* **59**, 433–448. 23

Ewald, R., Rössel, J., Himmelspach, J., & Uhrmacher, A. (2008). A plug-in-based architecture for random number generation in simulation systems. In *2008 Winter Simulation Conference, Miami, Florido, United States of America*, Miami. 40

Falconer, D. (1952). The problem of environment and selection. *The American Naturalist* **86**, 293–298. 5

Ferrante, A., Cullis, B. R., Smith, A., & Able, J. (2021). A Multi-Environment Trial Analysis of Frost Susceptibility in Wheat and Barley Under Australian Frost-Prone Field Conditions. *Frontiers in Plant Science* **12**. 10

Gilmour, A. R., Cullis, B. R., & Verbyla, A. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 269–293. 4

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450. 21, 158

Gogel, B. (1997). *Spatial analysis of multi-environment variety trials*. PhD thesis, Department of Statistics, University of Adelaide. 5

Gogel, B., Smith, A., & 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**, 1–21. 5, 10, 87, 161, 162, 198

Hellekalek, P. (1998). Good random number generators are (not so) easy to find. *Mathematics and Computers in Simulation* **46**, 485–505. 40, 41

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69–83. 7

John, J. & Williams, E. (1996). *Cyclic and computer generated designs*. Chapman and Hall: London, 2nd edition. 116

Jordan, M. (2022). Spatial models for colocated trials. In *Australasian Plant Breeding Conference, Gold Coast, 09 - 11 May 2022*. 69, 136

Kadkol, G., Sissons, M., Lambert, N., & Lisle, C. (2022). Genetic improvement in grain yield and quality of Australian durum wheat over six decades of breeding. *Cereal Chemistry* **1**, 1–22. 51

Kelly, A., Smith, A., Eccleston, J., & Cullis, B. R. (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Science* **47**, 1063–1070. 6, 8, 9, 10

L'Ecuyer, P. (1990). Random numbers for simulation. *Communications of the ACM* **33**, 86–97. 40

Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2021). Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials. *Frontiers in Plant Science* **12**. 13, 16, 154

Luckett, D. & Halloran, G. (1995). Plant Breeding What Is Plant Breeding and Why Do It? In Oxford University Press, editor, *Plant Breeding*, pages 1–255. Graham Centre for Agricultural Innovation, Charles Sturt University: Wagga Wagga Australia. 1

Mardia, K. & Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146. 21

Moose, S. & Mumm, R. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology* **147**, 969–977. 1

Mrode, R. & Thompson, R. (2005). *Linear models for the prediction of animal breeding models 3rd Edition*. CABI Publishing. 24, 158

Oakey, H., Verbyla, A., Cullis, B. R., Wei, X., & Pitchford, W. S. (2007). Joint modelling of additive and non-additive (genetic line) effects in Mulit-environment trials. *Theoretical and Applied Genetics* **114**, 1319–1332. 7, 133, 157

Patterson, H. D. (1978). Routine least squares estimation of variety means in incomplete tables. *Journal of National Institute of Agricultural Botany* **14**, 401–412. 3

Patterson, H. D. & Silvey, V. (1980). Statutory and Recommended List Trials of Crop Varieties in the United Kingdom. *Journal of the Royal Statistical Society. Series A (General)* **143**, 219–240. 3, 4

Patterson, H. D., Silvey, V., Talbot, M., & Weatherup, S. T. C. (1977). Variability of yields of cereal varieties in U.K. trials. *Journal of Agricultural Science* **89**, 238–245. 3

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554. 3, 19

R Core Team (2020). R: A Language and Environment for Statistical Computing. 27, 55, 92, 96, 175

Robinson, G. K. (1991). That BLUP Is a Good Thing : The Estimation of Random Effects. *Statistical Science* **6**, 15–32. 4, 22

Russell, K. (2018). *Design of Experiments for Generalized Linear Models*. CRC Press, New York. 154

Russell, K., Woods, D., Lewis, S., & Eccleston, J. (2009). D-optimal designs for Poisson regression models. *Statistica Sinica* **19**, 721–730. 154

Sales, J. & Hill, W. G. (1976a). Effect of sampling errors on efficiency of selection indices. 1. Use of information from relatives for single trait improvement. *Animal Production Science* **22**, 1–17. 9, 26, 131, 196

Sales, J. & Hill, W. G. (1976b). Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. *Animal Production Science* **23**, 1–14. 9, 131, 196

Searle, S. R. (1997). The matrix handling of BLUE and BLUP in the mixed linear model. *Linear Algebra and Its Applications* **264**, 291–311. 22, 23, 26

Self, S. & Liang, K. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82**, 605–610. 22

Smith, A., Borg, L., Gogel, B., & Cullis, B. R. (2019). Estimation of Factor Analytic Mixed Models for the Analysis of Multi-treatment Multi-environment Trial Data. *Journal of Agricultural, Biological, and Environmental Statistics* **24**, 1–16. 10

Smith, A. & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* **214**, 1–19. xxi, 7, 10, 13, 82, 120, 159

Smith, A. & Cullis, B. R. (2021). An efficient resampling scheme for outlier detection in linear mixed models. Technical report, NIASRA Working Paper Series. 63, 195

Smith, A., Cullis, B. R., & Gilmour, A. R. (2001a). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics* **43**, 129–145. 88, 153, 196

Smith, A., Cullis, B. R., & Thompson, R. (2001b). Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. *Biometrics* **57**, 1138–1147. 5, 6, 7, 8, 10, 12, 13, 16, 80, 87, 94, 157, 195

Smith, A., Ganesalingam, A., Kuchel, H., & Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics* **128**, 55–72. 10, 12, 13, 14, 80, 82, 88, 153, 196, 198, 199

Smith, A., Ganesalingam, A., Lisle, C., Kadkol, G., Hobson, K., & Cullis, B. R. (2021a). Use of Contemporary Groups in the Construction of Multi-Environment Trial Datasets for Selection in Plant Breeding Programs. *Frontiers in Plant Science* **11**. 11, 12, 13, 16, 47, 68, 115, 130, 133

Smith, A., Norman, A., Kuchel, H., & Cullis, B. R. (2021b). Plant variety selection using interaction classes derived from Factor Analytic Linear Mixed Models: models with independent variety effects. *Frontiers in Plant Science* **12**. xxi, 7, 10, 80, 82

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177. 22

Taillard, E. (1991). Robust taboo search for the quadratic assignment problem. *Parallel Computing* **17**, 433–455. 92

Thompson, R. (1973). The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics* **29**, 527–550. 123

Thompson, R., Cullis, B. R., Smith, A., & Gilmour, A. R. (2003). A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. *Australian and New Zealand Journal of Statistics* **45**, 445–459. 10

Tolhurst, D., Mathews, K., Smith, A., & Cullis, B. R. (2019). Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *Journal of Animal Breeding and Genetics* **136**, 279–300. 7, 10

Truong, L., Sarvi, M., Currie, G., & Garoni, T. (2015). How Many Simulation Runs are Required to Achieve Statistically Confident Results: A Case Study of Simulation-Based Surrogate Safety Measures. In *IEEE Conference on Intelligent Transportation Systems, Gran Canaria, Spain, September 15, 2015*. 42

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy science* **91**, 4414–4423. 7

Verbyla, A. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* **32**, 227–230. 19, 23

REFERENCES

Ward, R., Spohr, L., & Snell, P. (2019). Rice grain quality: an Australian multi-environment study. *Crop and Pasture Science* **70**, 946–957. 12, 88, 153, 196

Welham, S., Gogel, B., Smith, A., Thompson, R., & Cullis, B. R. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Australian and New Zealand Journal of Statistics* **52**, 125–149. 5, 10

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nature reviews genetics* **2**, 983–989. 1