Bioinformatics and Biometrics for the Australian Grains Industry

CAIGE Durum Wheat

2020 Analysis Report

GRDC Project No: UOS1901-002RTX

Ky Mathews[1], Robin Wilson[2], Deepak Baranwal[3], Julie M Nicol[3], Richard Trethowan[3]
email:kmathews@uow.edu.au

email:julie.nicol@sydney.edu.au

[1]Centre for Bioinformatics and Biometrics
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong

[2]Integrated Breeding Platform

[3]Plant Breeding Institute
Faculty of Science
University of Sydney

August 2, 2021

Contents

CONTENTS

Acknowledgements: Thanks to Brian Cullis for his guidance in this analysis.

1 Background

The CIMMYT Australian ICARDA Germplasm Evaluation project (CAIGE) is a Grains Research and Development Corporation (GRDC) funded project (UOS1901-002RTX) which commenced in 2005 and is managed by the University of Sydney in collaboration with the University of Queensland and Australian wheat breeders. From 2015 onwards statistical support has been provided by the University of Wollongong (UOW) employed on the GRDC projects: Statistics for the Australian Grains Industry (SAGI2) and Bioinformatics and Biometrics for the Australian Grains Industry (BBAGI) (UW00009). The key objective of CAIGE is to evaluate germplasm developed by CIMMYT (International Maize and Wheat Improvement Centre) and ICARDA (International Centre for Agricultural Research in the Dry Areas) for potential introgression into Australian wheat breeding programs.

Entries (varieties) are evaluated at different sites across the Australian wheat belt and selected by breeding companies for introgression into their breeding programs to ultimately release improved varieties to Australian wheat growers. Each year a new cohort of entries is evaluated in a multi-environment trial (MET) series. The relationship between entries, pedigree information, is available in the CAIGE Breeding Management System (BMS) database, (http://www.caigeproject.org.au/breeding-management-system). The inclusion of pedigree information enhances the analysis of the CAIGE dataset in two ways. There is a significant lack of entry connectivity between years in the CAIGE durum wheat dataset and until now the CAIGE MET analyses have been within year only. It is well known that the seasonal variation in Australia is large and selecting entries for use as parents or direct release based on a multi-year analysis is advised. Pedigree information enables an across year analysis through the ancestral relationships across years as the incoming germplasm from CIMMYT and ICARDA come from their breeding programs which have a high level of co-ancestry from year to year. Further, the use of pedigree information in the analysis allows for the partitioning of the total genetic (variety) effects into additive and non-additive effects. The additive effects are equivalent to estimating the general combining ability of a line and are the most appropriate effects to use when selecting entries to use as parents in a breeding program. This meets a key aim of the CAIGE project - the identification of lines for use as parents in Australian breeding programs. Furthermore, the total genetic effects can be estimated from this analysis and can be used to identify entries that could potentially be directly released to Australian growers - subject to meeting disease and quality criteria.

A factor analytic linear mixed model (FA-LMM) framework is used for the analysis. This one-stage approach, following Smith et al. (2001) and Gogel et al. (2018), allows for the inherent imbalance in the dataset, individual trial spatial modelling and appropriate modelling of the variety by environment interaction (VEI). In addition, the genetic variance-covariance matrix is modelled using the numerator relationship matrix derived from the pedigree information (Oakey et al., 2006, 2007; Burgueno et al., 2007; Cullis

et al., 2010).

The CAIGE project requested an across year analysis for the durum dataset, in line with bread wheat and barley datasets. However, the data curation of the pedigree information was significant and due to time pressures it was decided to perform a MET analysis on the yield and agronomic traits for the 2020 trials only. The purpose of this report is to describe the 2020 dataset, the statistical methods used and present the results for yield and the agronomic traits. This report is organised as follows: the genetic material is described in Section 2, the phenotypic data and experimental designs are described in Section 3, the statistical methods are described in Section 4 and the results presented in Section 5. Finally, Section 6 indicates the scripts, their purpose and location within the BBAGI project.

2 Genetic material

The genetic material consists of CIMMYT, CIMMYT-TURKEY and ICARDA late stage breeding lines and Australian checks. The non-Australian lines are generally those ready to be deployed in international nurseries or were selected by Australian, CIMMYT or ICARDA breeders at CIMMYT and/or ICARDA breeding locations. The material is imported into Australia under strict quarantine arrangements in small amounts (30g) and a seed increase is undertaken before the material is ready for yield evaluation.

2.1 Entry list

Each year a new cohort of entries is released from quarantine and is evaluated in yield trials in Australian environments. There were 12 Australian released varieties included in two plots in each trial in 2020. There were 434 unique entries (excluding fillers) evaluated in 2020, their distribution by year and material source is presented in Table 1. The ratio of CIMMYT to ICARDA entries varies from year to year. There are 2019 entries in the 2020 dataset because the 2019 Breeza trial was re-evaluated in 2020 at Spring Ridge (D20SPRR2) alongside the 2020 entry list trial. The 2019 trial used the same experimental design provided in 2019 and is a subset of the full 2019 entry list. The D19SPRR2 trial was sown in late August 2019 and not considered representative of this region, hence the re-evaluation in 2020.

Each entry has passport information which help to uniquely identify it. This passport information includes the Quarantine Code (QCode) for the international nursery and Quarantine Number (QNo), the entry number within a nursery; Genotype Identifier (GID) from the CIMMYT/ICARDA databases or the CAIGE Breeding Management System (BMS); Cross Id (CID) and Sister Id (SID) from the CIMMYT IWIS2 database; accession identifier (acc_id) from the Australian Grain Genebank; the germplasm source (Source = CIMMYT, CIMMYT-TURKEY, ICARDA and Australia), the pedigree string and selection history. GID is assigned based on the selection history. Each year there is a flow of germplasm from CIMMYT to CIMMY-TURKEY where entries are evaluated for soil borne pathogens traits. Some of these entries that are sent to CIMMYT-TURKEY

are also sent to Australia in the CIMMYT durum selection nursery (ZDG18). The following year some of the same entries may also be sent to Australia in the Soil Borne Pathogen nursery(SBP19, say). Hence, there are duplicate entries that have entered Australian quarantine and been evaluated in different years. These are identified by GID which is based on an entrie's selection history. A new field, PreferredName, is created to manage these duplicates where the two EntryCD names are concatenated together, e.g. 11:SBP19_44:ZDG18. GID assures that duplicate entries are identified, it is then a discussion between the biometrician and CAIGE staff to determine that they are true duplicates (see Mathews (2021) for further discussion). There are 35 entries in this dataset which are considered duplicates, i.e. been imported in more than one nursery. These can be identified by searching for "_" in the PreferredName field.

Entry names in CAIGE are usually a concatenation of the QCode and QNo, such that QNo:QCode. It is important to understand that QNo is nested within QCode, so an entry with QNo=1 with QCode=ZDG18 is different from an entry with QNo=1 and QCode=SBP19. If the line is a check line from CIMMYT or ICARDA then the QNo:QCode designation is replaced with the check name. Since there were some duplicate entries in the CAIGE durum dataset, GID was used as the unique identifier and the term used in the analysis to link the pedigree and phenotypic data together.

Table 1: Number of entries (excluding fillers) in the CAIGE durm wheat 2020 dataset by entry year and material source (Australia, CIMMYT, CIMMYT-ICARDA, ICARDA). The source totals are the unique number of entries across years. Note this includes entries from the 2019 entry list due to a re-evaluation in the W20SPRR2 environment

Year	Australian	CIMMYT	CIMMYT-TURKEY	ICARDA	Total
2020	9	95	31	79	214
2019	9	159	4	58	230
Total	12	250	35	137	434

There is 1 filler in the dataset. It is not included in the estimation of genetic variance but is retained in the dataset and modelled following the process described in Appendix C of Tolhurst et al. (2019).

2.2 Pedigree information

Pedigree information describes the relationships between all entries in the dataset. Following Henderson (1976) this information is converted to a numerator relationship matrix, $\boldsymbol{A}$, which can be used in a linear mixed model to partition the total genetic variance into additive and non-additive effects (Oakey et al., 2006, 2007; Burgueno et al., 2007; Cullis et al., 2010). Estimated breeding values for parental selection can then be obtained from this analysis.

The pedigree information was extracted by CAIGE staff from the CAIGE BMS database. Whilst there is some uncertainty about how far back in the ancestral tree it is required for accurate estimation of the coefficient of parentage it is reasonable to assume that the greater the depth the more accurate the estimate. Thus, it was decided to use the

pedigree strings, not GID, for the ancestors in the pedigree file. This required a significant curation process of the entry selection histories and pedigree strings, and the parent and grandparent GID and pedigree strings which is documented in Mathews (2021). Drs. Karim Ammar (CIMMYT breeder) and Filippo Bassi's (ICARDA) collaborative assistance with this process was invaluable.

The result of this data curation is a data frame in the so-called "Me Mum Dad" format which is described in the `R pedicure` library (Butler, 2019) as a *data frame with (at least) three columns that correspond to the individual, female parent and male parent, respectively. The row giving the pedigree of an individual must appear before any row where that individual appears as a parent. Founders use 0 (zero) or NA in the parental columns.* The "Me Mum Dad" data frame used in the generation of the numerator relationship matrix had with 2276 records. An example "Me Mum Dad" file is presented in Table 2. The individual in the last row, 800039930 which has QNo:QCode 41:ZDL19, has a 'Mum' (female parent) with pedigree ICAJOUDY1/ATLAS1 and a 'Dad' (male parent) of ICARASHA2. These parents are in rows 5 and 6, respectively, in this table, *above* the individual that will be tested in the field.

Table 2: Example of a *Me Mum Dad* file for the CAIGE Durum Wheat dataset. Me is the individual being tested or an ancestor of an entry being tested, Mum and Dad contain the name of the parents. A value of zero for Mum or Dad indicates unknown pedigree or base individuals. Fn is the filial generation number and is a single number for single plant selections (SPP) or of the form $n_b : n_f$ indicating the level of SPP and number of bulked generations. $\mathcal{F}$ is the inbreeding coefficient.

Me	Mum	Dad	Fn	$\mathcal{F}$
TERBOL 97-3	0	0	6	0.96875
ATLAS1	MGNL3	AGHRASS-2	7	0.98438
MIKI	0	0	6	0.96875
ICAJOUDY1	ATLAST1/961081	ICASYR1	7	0.98438
ICAJOUDY1/ATLAS1	ICAJOUDY1	ATLAS1	7	0.98438
ICARASHA2	MIKI	TERBOL 97-3	7	0.98438
800039930	ICAJOUDY1/ATLAS1	ICARASHA2	7:8	0.98445

Functions `chkPed` and `trimPed` from the `pedicure` (Butler, 2019) library on the `R` statistical computing platform (R Core Team, 2020) are used to curate the pedigree (i.e. "Me Mum Dad") file. `Pedicure` is freely available from www.mmade.org/pedicure. A function specifically written for this project, `pedigree.cut` was used to cut the pedigree strings back to founders and include them in the "Me Mum Dad" file. The pedigree data frame for the 2020 dataset contained 2276 records; the 434 entries evaluated in 2020 (includes the 2019 entries evaluated at D20SPRR2) and 1842 ancestors.

The fourth column, Fn, in the pedigree file (Table 2) is the filial generation number of the line. This can take the form of an integer (1 = simple cross, 2 indicates the progeny of this cross selfed once, 3 is selfed twice, and so on). If the breeding program uses bulk selections before a single plant selection then Fn takes the form $n_b : n_f$, where n_b is the Fn of the single plant selection and n_f indicates the final generation of bulking. The difference between n_f and n_b is the number of generations of bulking. For example, an $F_{2:5}$ line, F_2 derived F_5 line, would have an Fn of 2:5, indicating a single plant selection

at F2 followed by 3 generations of bulking. The fifth column self represents the level of selfing where

$$self = \begin{cases} Fn - 1, & \text{where } Fn \text{ is an integer} \\ n_f - 1, & \text{otherwise} \end{cases} \quad (1)$$

The selection histories of the CAIGE test entries were used to define the Fn values for these lines. This is described in Mathews (2021). Australian checks are considered fixed and assigned a Fn=8, whilst all parents and grandparents are assigned a Fn=6.

The inbreeding coefficient $\mathcal{F}$ (Henderson, 1976) for each record in the pedigree file is then calculated as

$$\mathcal{F} = 1 - \frac{1}{2}^{self}, \text{ where } self \text{ is defined in equation (1).} \quad (2)$$

The diagonals of the numerator relationship matrix ($\boldsymbol{A}$) are equal to one plus the inbreeding coefficient for an individual (i.e 1 + $\mathcal{F}$), and the off-diagonals are equal twice the coefficient of parentage (ancestry) between any two individuals (Wright, 1922). Jordan & Cullis (2020) have developed an improved algorithm for calculating the inbreeding coefficient of bulk selected lines. In essence it is a weighted average of the inbreeding coefficients of an F_{n_f} and an F_{n_b} individual adjusted for the coefficient of parentage. This leads to inbreeding coefficients such as the last row in Table 2. In practice, the inverse of $\boldsymbol{A}$ is used in the design and analysis and this is calculated using the `ainverse` function in the linear mixed model software `asreml` (Butler et al., in prep.) following the process outlined in Gilmour et al. (2004) and Meuwissen & Luo (1992). `asreml` is licensed software available from www.vsni.co.uk.

Table 3: Numerator relationship (**A**) matrix for GID 800039930 (41:ZDL19) and ancestors in the CAIGE durum wheat 2020 pedigree file.

GID	TERBOL 97-3	ATLAS1	MIKI	ICAJOUDY1	ICAJOUDY1/ATLAS1	ICARASHA2	800039930
TERBOL 97-3	1.969	0.000	-0.000	0.000	0.000	0.984	0.492
ATLAS1	0.000	1.984	0.000	-0.000	0.992	0.000	0.496
MIKI	-0.000	0.000	1.969	0.000	0.000	0.984	0.492
ICAJOUDY1	0.000	-0.000	0.000	1.984	0.992	0.000	0.496
ICAJOUDY1/ATLAS1	0.000	0.992	0.000	0.992	1.984	0.000	0.992
ICARASHA2	0.984	0.000	0.984	0.000	0.000	1.984	0.992
800039930	0.492	0.496	0.492	0.496	0.992	0.992	1.984

A summary of the resulting inbreeding coefficients by genetic source (Table 4) shows that the level of inbreeding in the durum data from all sources is very similar, with CIMMYT and CIMMYT-TURKEY having slightly lower inbreeding coefficients than ICARDA.

2 Genetic material

Table 4: Summary of the inbreeding coefficients for the 434 entries by genetic source (Australian, CIMMYT, CIMMYT-TURKEY and ICARDA) in the CAIGE Durum Wheat 2020 dataset.

Source	Min	Mean	Max
Australian	0.992	0.992	0.992
CIMMYT	0.938	0.974	0.992
CIMMYT-TURKEY	0.938	0.951	0.996
ICARDA	0.969	0.994	1.000

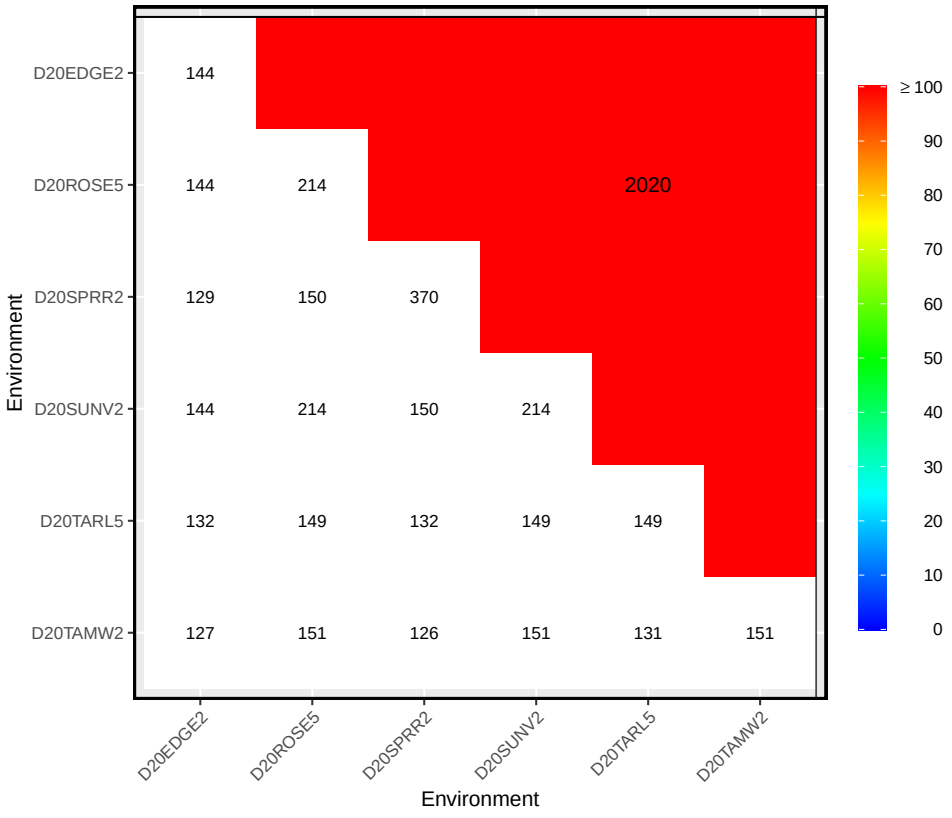

Figure 1: Heatmap of the variety connectivity matrix between all environments in the CAIGE durum wheat 2020 dataset

The variety connectivity matrix (Figure 1) illustrates that the between environment variety connectivity within 2020 is very good (minimum 126, maximum 214). The parental connectivity matrix (Figure 2) indicates that there this connectivity is improved at the parent level with a minimum number of parents in common between environments of 176.

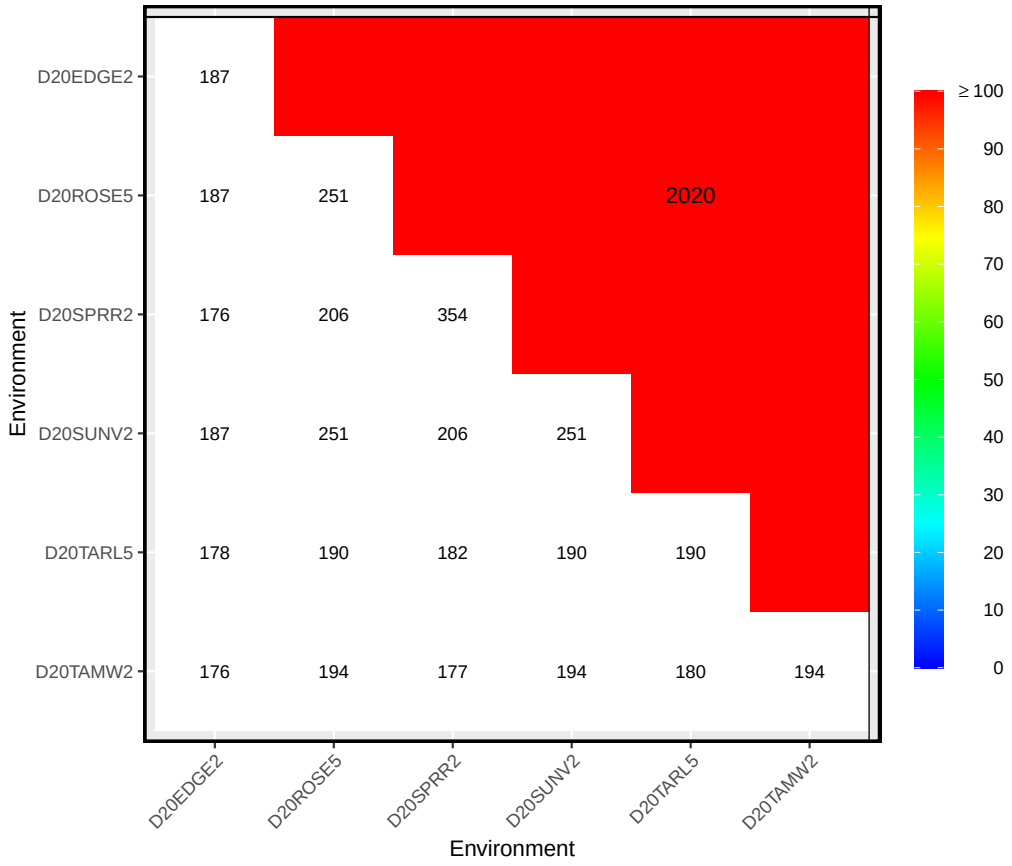

Figure 2: Heatmap of the parental connectivity matrix between all environments in the CAIGE durum wheat 2020 dataset

3 Phenotypic data

The key trait of interest for this analysis was yield (t/ha). Other traits such as yield under crown rot inoculation, flowering and maturity dates, plant height, have also been collated but only sparsely across the dataset.

3.1 Trial information

The complete CAIGE durum yield dataset is from 2016-2020, however, the pedigree information for the 2016-2019 dataset was not completed in time for the delivery of 2020 results and a MET analysis of the 2020 trials only was conducted. An across year MET analysis is underway.

There are 7 trials in the 2020 yield dataset evaluated at 6 locations. An environment is

defined as a year-location combination. The terms trial and environment are synonymous for all but the D20SPRR2 environment where the 2019 and 2020 entries were evaluated in adjacent blocks in co-located trials B19 and A20, respectively. Yield was also measured on a trial conducted by NSW DPI at Tamworth (D20TAMW2) but all plots in this trial were inoculated with crown rot and hence it was analysed separately. It is noted that all the CAIGE durum wheat trials conducted at Tamworth from 2016 onwards were inoculated with crown rot. Hence, it is recommended that a separate across year MET analysis be considered for these crown rot inoculated trials. One reason is that the treatment, crown rot inoculate, is confounded with the environment and any comparison of yield performance between lines in a different environment to the Tamworth trials is confounded by this treatment, i.e. one cannot tell if the yield difference is due to the environment or the treatment. This is an example of pseudo-replication.

The specific details for the 2020 5 yield *per se* trials and the crown rot trial are provided in Table 5. For each trial the experimental design plot factors, the number of entries (varieties) and their level of replication and the trial mean yield (t/ha) and reliability estimated from single environment analyses. Note, the genetic variance and reliability of the co-located trials at D20SPRR2 are the same because they are estimated at the environment, not trial, level.

The number of entries in each trial ranges from 144 in Edgeroi to 214 in Roseworthy and Sunville. Except for the re-evaluated 2019 entries at Spring Ridge these trials were designed together with pedigree information as detailed in Mathews et al. (2020). All entries were evaluated at Roseworthy and Sunville, and where possible replicated in two plots. The trial mean yield for the yield *per se* trials ranges from 2.24-6.43t/ha which is representative of Australian wheat production region. The reliability, equivalent to the line mean broad sense heritability, ranges from 0.46 to 0.81. It is of interest to note that the percentage of additive genetic variance in the total genetic variance ranges from 14 to 53% for the yield *per se* trials, but is much lower for the crown rot inoculated trial (D20TAMW2) at 9% (Table 5).

At the advent of the 3-year project (US00073) commencing in 2019, GRDC advised the CAIGE team that there was now a requirement to track and manage publicly and privately funded trials in the project. Ky Mathews, from BBAGI, made it very clear to GRDC and the CAIGE team that separate analyses for publicly and privately funded trials would not be entertained as this is a commercial, not scientific, decision. However, the tracking of public and private trials is required to ensure results from privately funded trials are not released publicly and is an extra burden on the statistical team as the CAIGE raw plot data is not stored in BMS and hence the management of this is not a simple filter as one would anticipate. There are 3 privately funded trials in the current dataset as indicated in italics in Table 5.

3 Phenotypic data

Table 5: Experimental design information for each trial in the 2020 CAIGE durum wheat dataset. The number of blocks, ranges, rows, plots and entries, trial mean yield (TMY, t/ha) and reliability (r^2), total genetic variance and additive variance expressed as a percentage of the total genetic variance (%Additive Variance) from single environment analyses are provided. The number of entries allocated to one, two and more that two plots is indicated and the resulting percentage partial replication (%-p-rep). Filler plots are not included in the entry number figures. *Italicised* environment names are privately funded trials. The crown rot inoculated trial, D20TAMW2, is separated by a solid horizontal line as it is not included in the MET analysis.

Env	Block	Range	Row	Plot	Entry	Entries replicated			%p-rep	TMY	r^2	Genetic	%Additive
						once	twice	> 2		(t/ha)		Variance	Variance
D20EDGE2	2	24	8	192	144	101	38	5	29.9	2.24	0.491	0.084	53
D20ROSE5	2	12	24	288	214	149	56	9	30.4	4.47	0.681	0.229	14
D20SPRR2-A20	2	12	17	204	147	94	49	4	36.1	6.02	0.457	0.180	39
D20SPRR2-B19	2	12	21	288	230	199	28	3	13.5	6.43	0.457	0.180	39
D20SUNV2	2	24	12	288	214	151	55	8	29.4	4.19	0.814	0.269	31
D20TARL5	2	12	16	192	149	110	35	4	26.2	5.83	0.534	0.196	53
D20TAMW2	2	12	19	228	151	83	59	9	45.0	2.27	0.750	0.180	9

3.2 Agronomic traits

In 2020 the CAIGE collaborators agreed to collate phenology and height data to augment the yield and disease data. However, it was not measured in all environments and phenology is measured as either days to heading or zadoks, depending on the location. Height was measured as a height score with values from from 5 to 9. These secondary traits were only measured at Sunville (D20SUNV2) and Tamworth (D20TAMW2).

Table 6: Mean (standard deviation) for days to heading, maturity zadoks and height (score) data for CAIGE durum wheat trials evaluated in 2020.

Env	Days to Heading	Maturity Zadoks	Height (cm)
D20SUNV2		50.3 (3.63)	7.1 (0.8)
D20TAMW2	117.8 (3.16)		

3.3 Experimental design

The experimental designs for the CAIGE durum wheat 2020 dataset is implicitly unbalanced. The amount of seed available per entry post-quarantine and multiplication varies such that some entries have enough seed for allocation to 2 plots per location and others do not. Further, in order to accommodate the known large variety by environment interaction in Australian environments the entries are evaluated in as many locations as possible within a year. The partially replicated designs proposed by Cullis et al. (2006) are well suited to this scenario. Cullis et al. (2020) extended the partially-replicated model based designs to include pedigree information thereby ensuring appropriate allocation of entries at the ancestral level across a location.

In partially-replicated designs, entries with sufficient seed (including test and check entries) are replicated across the field trial whilst the remaining entries are allocated to one plot only. The allocation of entries across locations within years can ensure that across the MET dataset the entries are equally replicated, depending on seed and plot availability. The extension of the pedigree partially-replicated designs of Cullis et al. (2020) to include

allocation of entries across environments is presented in Cullis et al. (2018); Ganesalingam et al. (2019). These so-called incomplete MET designs using pedigree were implemented in the design of the 2020 CAIGE durum wheat trials, Mathews et al. (2020).

Crop variety trials are usually laid out in a rectangular array of rows and column. A model-based design approach was implemented using the Optimal Design (`od`) library (Butler & Cullis, 2019) on the `R` statistical platform (R Core Team, 2020). The designs were partially replicated block designs which take into account the row and column (range) arrangement. Replicated entries were allocated across the field layout such that the blocks at each location were near-resolvable. The percentage replication ranged from 13.5% for D20SPRR2-B19 to 45% for D20TAMW2, with a median of 29.9%; all 7 trials had a %p-rep less than 50%. It is understood that a low level of entry replication per trial would be offset by the pedigree relationship information, however, research is required to determine whether this understanding is justifiable.

The re-evaluation of the 2019 Breeza trial (D19BREE2) at Spring Ridge in 2020 (D20SPRR2), called D20SPRR2-B19 added some complexity to the modelling of the D20SPRR2 environment. This complexity is described here to better understand the final model fitted (Section 4). Firstly, an optimal design could have been generated which included both the 2019 (B19) and 2020 (A20) entries allocated to plots across the available field area. The aim of the CAIGE yield analysis is to identify entries that perform well, it does not matter what year they entered Australia or were previously evaluated, so it would have been better to have one design with one residual variance in this environment. Secondly, the design provided by BBAGI in 2019 for the B19 entries was re-used in 2020. The design provided in 2019 was performed as a MET design taking into account the other allocation of the 2019 entries to other plots in oher 2019 environments. This design on its own is probably not optimal. Further, the same field area was not available in 2020 and hence the 2019 trial was split across two field blocks by 3 rows, Figure 3. To be clear, the problem is not that the trials had to be split across two distant field blocks, this is a practical fact of conducting field trials, it is simply that a better design could have been generated to accommodate these known constraints.

4 Statistical Methods

In this section the statistical methods applied to the 2020 dataset are described. For the purposes of the analysis we defined the term Variety to describe entries and Environment defined a year by location combination.

A one-stage analysis of trials from a MET dataset combines the individual plot data across trials. This allows appropriate modelling of the individual trial variation and is achieved in the linear mixed model framework (Smith et al., 2001; Gogel et al., 2018).

The base model for each environment was a random block effect (where available, Table 7) and the residual model was a separable auto-regressive correlation model fitted in both the column and row directions. Spatial models were then determined following Gilmour

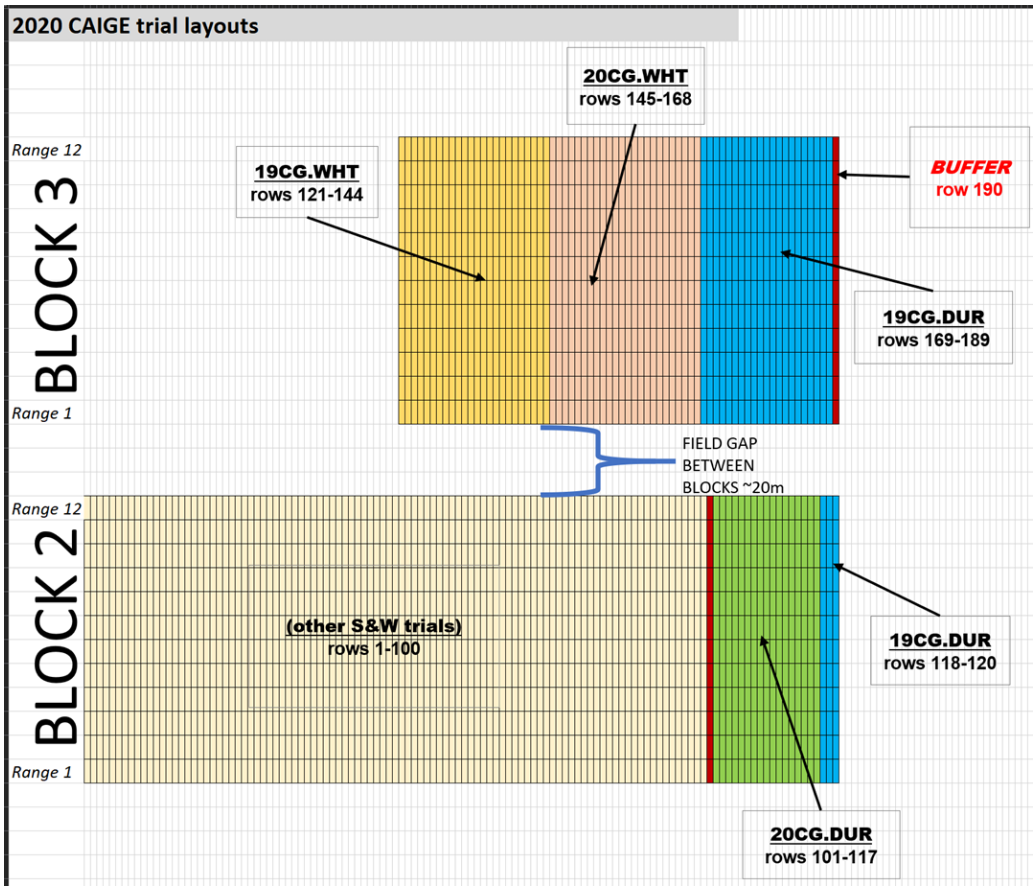

Figure 3: Layout of CAIGE durum trials at D20SPRR2. The blue shaded area is the B19 trial and the green the A20 trial. *Courtesy: SWS Seeds*

et al. (1997). Table 7 provides a summary of the models fitted by year. It can be seen that *column* is fitted more frequently than *row* as a random effect to accommodate extraneous spatial variation - this fits with the trial management practices mostly occurring across columns. Even though there are 6 environments, the terms Block and the ar1Row are fitted 7 times in this dataset. This is the effect of the two different field blocks at D20SPRR2, see Figure 3, where there is a Block effect within each field block as per the experimental designs and a separate residual is fitted for each field block. A separable spatial auto-correation model accommodating the row and column dimensions is fitted as per Gilmour et al. (1997). In the case of the D20SPRR2 environment, there are so-called co-located trials. In this case, the co-located entity is the field block, not the trial, as the D20SPRR2-B19 trials is split across the two field blocks. The field block residuals are fitted separately but the row and column auto-correation parameters are constrained to be the same for each field block. The assumption in such cases is that the management practices and field variation are the same across the environment and can be considered to be the same for the co-located entities (usually trials, field blocks in this case). This method of accommodating non-adjacent co-located trials has been practiced in the analysis of National Variety Trials (NVT) since 2015, also analysed by BBAGI at UOW (*pers. comm.* B Cullis).

Table 7: Model summary indicating the number of trials where a term was fitted for the 2020 yield CAIGE durum wheat yield analysis. Vertical lines delineate between plot structure and residual terms. ar1 refers to the autoregressive correlation model of order 1 fitted at the residual level; an identity model is fitted if an ar1 is not.

Year	Environment	Block	Column	Row	ar1Column	ar1Row
2020	6	7	3	2	5	7

The first step in the analysis of a MET dataset is to analyse each environment individually, this can be performed in one step where the genetic variance-covariance matrix is modelled with a diagonal structure, that is, independence between environments and different genetic variances for each environment. This model was fitted with and without the numerator relationship matrix, $\boldsymbol{A}$, to the 2020 dataset to demonstrate to the CAIGE collaborators that the ranks of the total variety effects within each environment with and without pedigree were comparable.

A key feature of MET datasets is the re-ranking of variety performance in different environments, known as variety by environment interaction (VEI). The factor analytic linear mixed model (FA-LMM) parsimoniously models the between environment variance-covariance matrix and usually results in a good fit to the data (Smith et al., 2001; Gogel et al., 2018). FA-LMMs account for the covariances of the VE effects between environments using a small number, k, of (unknown) common factors, which are estimated from the data. An FAk model has order k.

An FA-LMM was fitted to the yield *per se* dataset. The yield under crown rot inoculum and agronomic traits were only measured in one trial each (Table 6) and individual analyses were performed.

Further, the genetic variance-covariance matrix can also be incorporated into a FA-LMM using the numerator relationship matrix derived from pedigree information (Oakey et al., 2006, 2007). This enables the total VE effects to be partitioned into additive and non-additive VE effects, and their respective between environment genetic variances matrices can be modelled with separate FA models (Oakey et al., 2007; Smith & Cullis, 2018). The order, k_a or k_e, of the respective additive and non-additive FA models are increased independently until a good fit to the data is found. Logliklihood ratio tests for nested models, Akaike Information Criterion (AIC) for non-nested models and descriptive statistics, such as the percentage genetic variance accounted for (%vaf) (Smith et al., 2015), are utilised in this decision making process.

Recently, factor analytic selection tools (FAST) were derived to summarise the large amount of information generated from FA-LMM (Smith & Cullis, 2018). They consist of measures of overall performance (OP), stability (root mean square deviation, RMSD) and responsiveness and summarise variety performance across the dataset. OP and RMSD are based on the first factor, which by design, accounts for the most genetic variance and responsiveness is based on the remaining k factors. Importantly, all measures are based

on the same scale as the data, i.e. t/ha for yield in this dataset.

In practice, the agronomic traits can be analysed before the yield data as they can be used as a data curation step to ensure that the process of naming entries is appropriate, see Mathews et al. (2021) for details. The agronomic data was inspected carefully to determine if the 35 duplicate entries were justifiable. Unfortunately, with only one trait measured in one environment it was difficult to make any conclusive decisions and the duplicates were maintained.

5 Results

In this section, we present a summary of the models fitted for each trait: yield, days to heading, maturity zadoks and height. For the key trait, yield, the results from an individual analysis (i.e. no between environment covariances considered) with and without pedigree information is presented to demonstrate the similarity in total variety effects predicted from these two models (Section 5.1). This section includes the results from the single crown rot trial (D20TAMW2). Next, the results from the MET analysis of the yield *per se* trials, where the between environment variance-covariance matrix is modelled using a factor analytic linear mixed model (Section 5.2). The results for the agronomic traits are presented in Section 5.3.

5.1 Yield: individual trials 2020

An independent genetic variance model, where no relationship between environments (i.e. a diagonal variance structure) was fitted, without and with considering the relationship between varieties, to the 2020 yield *per se* dataset. The pedigree (DIAG-DIAG) model had a significantly larger REML loglikelihood (80.8) than the non-pedigree (DIAG) model (52.5), resulting a significant difference when tested formally with a REML loglikelihood ratio test (LRT) (Table 8). The percentage of additive and non-additive to total genetic variances were 77% and 23%, respectively - this is a typical ratio for yield MET datasets. Note, these models do not include the D20TAMW2 trial.

Variety by environment empirical Best Linear Unbiased Predictions (VE-eBLUPs) were formed from both models. We compare the VE-BLUPS from each model to demonstrate that the more complex pedigree model provides predictions that are highly correlated with those from the simplest model where between environment correlations and between entry relationships are not considered. From the pedigree model both the additive and non-additive predictions were formed and these were summed to produce the total VE-eBLUPs. The correlations between the total VE-eBLUPS from both models and the additive VE-eBLUPS from the pedigree model were calculated overall and at the environment level. The correlation between the total VE-eBLUPs from both models was 0.968, the correlation between the total VE-eBLUPs from the identity model and the additive VE-eBLUPs from the pedigree model is 0.831. The correlations between the non-pedigree and pedigree VE-eBLUPs for individual environments ranged from 0.878 (D20SPRR2) to 0.999 (D20TAMW2) across the full dataset. The correlations between the

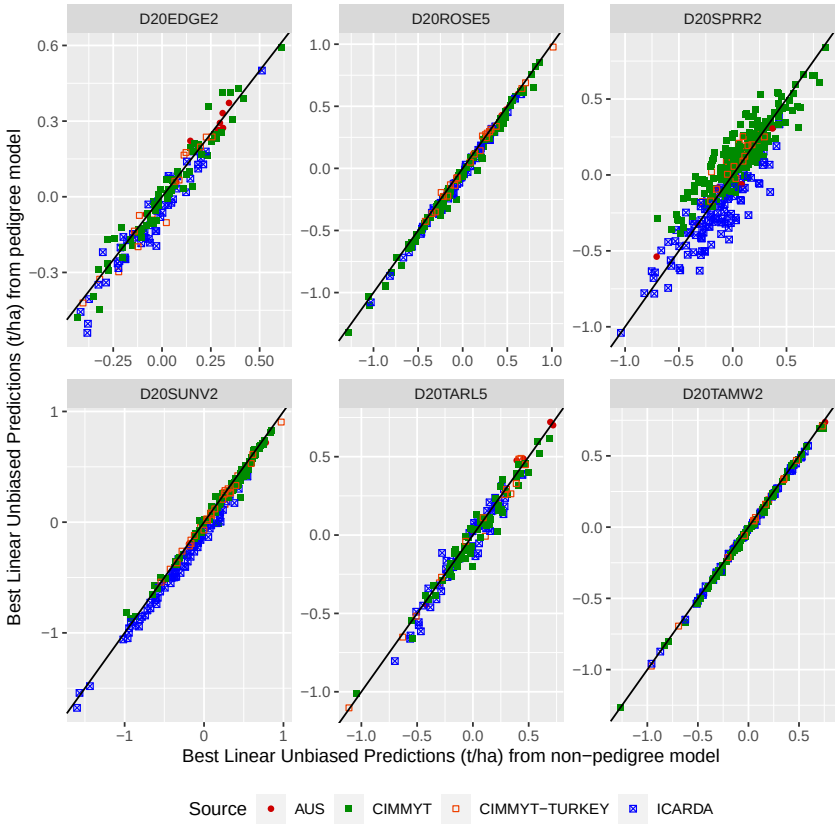

Figure 4: Total Variety by Environment Best Linear Unbiased Predictions for the pedigree versus the non-pedigree model for durum wheat trials conducted in 2020 where each environment is considered independently. The legend indicates the germplasm source. Australian and ICARDA checks are labelled.

non-pedigree and pedigree VE-eBLUPS for the 2020 environments is presented graphically in Figure 4. D20TAMW2, the crown rot inoculated trial, has the lowest additive genetic variance (Table 5) which explains the lack of variation around the 1:1 line for this environment compared to D20SPRR2 and D20EDGE2 where the additive genetic variance is 39 and 53% of the total genetic variance, respectively. The inclusion of the 2019 entries at D20SPRR2 where the ratio of CIMMYT to ICARDA entries is higher than in the 2020 entry list (Table 1) also contributes to the greater variability for this environment compared to other environments. Figure 4 demonstrates that the predictions from the simple model are highly correlated with those from the more complex, informative model.

5.2 Yield: Multi-environment trial 2020

The modelling of the 5 environments in this dataset commenced with an independent genetic variance model, where no relationship between varieties or environments was considered (DIAG, Table 8). It then progressed incrementally with increasing complexity and resulted in a final model where k_a=2 and k_e=1. This model had the maximum log-likelihood and the lowest AIC. The factor %VAF for the additive effects for the 2 additive

factors are 60% and 24%, and for the one non-additive factor is 66%. An FA2-FA2 model was also fitted, and although the total %VAF was higher than the FA2-FA1 model (89 versus 78%), the formal statistical LRT test indicated that these models were not significantly different from each other and the FA2-FA1 model has the lowest AIC, (Table 8).

Table 8: Summary of models fitted to the CAIGE durum wheat 2020 yield dataset for the additive and non-additive effects; information includes the model, the number of additive and non-additive genetic variance parameters, the loglikelihood (Loglik), loglikelihood ratio test (LRT), Akaike Information Criteria (AIC) and additive, non-additive and total genetic variance accounted for (%). The FA2-FA1 model is the final model.

Model	No. parameters					% Variance Accounted For		
Add-Non	Additive	NonAdditive	Loglik	LRT	AIC	Additive	Non-additive	Total
DIAG			52		-41			
DIAG-DIAG	5	5	81	1.63E-13	-94	77	23	
FA1-DIAG	10	5	116	9.21E-15	-153	58	22	
FA1-FA1	10	10	123	2.47E-03	-158	59	78	69
FA2-FA1	14	10	126	3.87E-02	-157	84	70	78
FA2-FA2	14	14	127	3.57E-01	-151	82	98	89

Selection decisions change depending on the purpose. Figure 5 shows that for the lines evaluated in 2020 the selection purpose will change which lines are selected. To select breeding lines for direct release then the total overall performance (vertical axis, Figure 5) is used, however to select lines for use as parents then their additive overall performance (horizontal axis, Figure 5) is of interest. For example, the points above the horizontal black line indicating the top 10% cut-off represent lines that would be selected for direct release. However, without the additive information the lines in the top-left quadrat may also be selected as parents when they not be suited to this purpose. Furthermore, it is clear that there are other lines (bottom right quadrat) which would better suited for use as parents but do not perform well on an individual line basis.

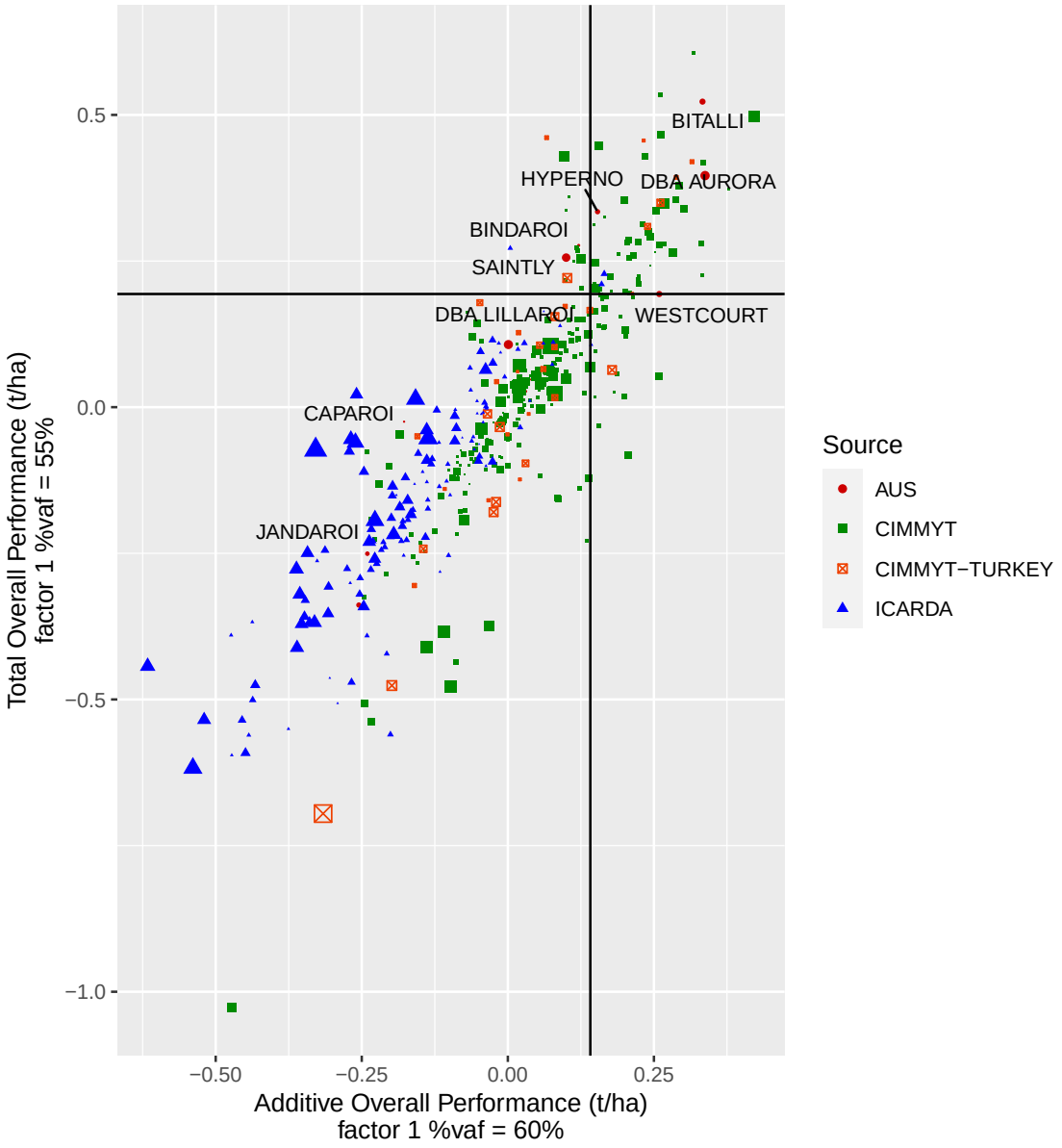

Figure 5: Overall Performance for Total and Additive effects for the 2020 entries in the 2020 yield dataset. Points to the top and right of the black horizontal and vertical lines, respectively, represent the top 10% cut-off. The legend indicates the germplasm source and Australian checks are labelled. The size of the points indicates the stability (RMSD) of the line with smaller points indicating greater stability.

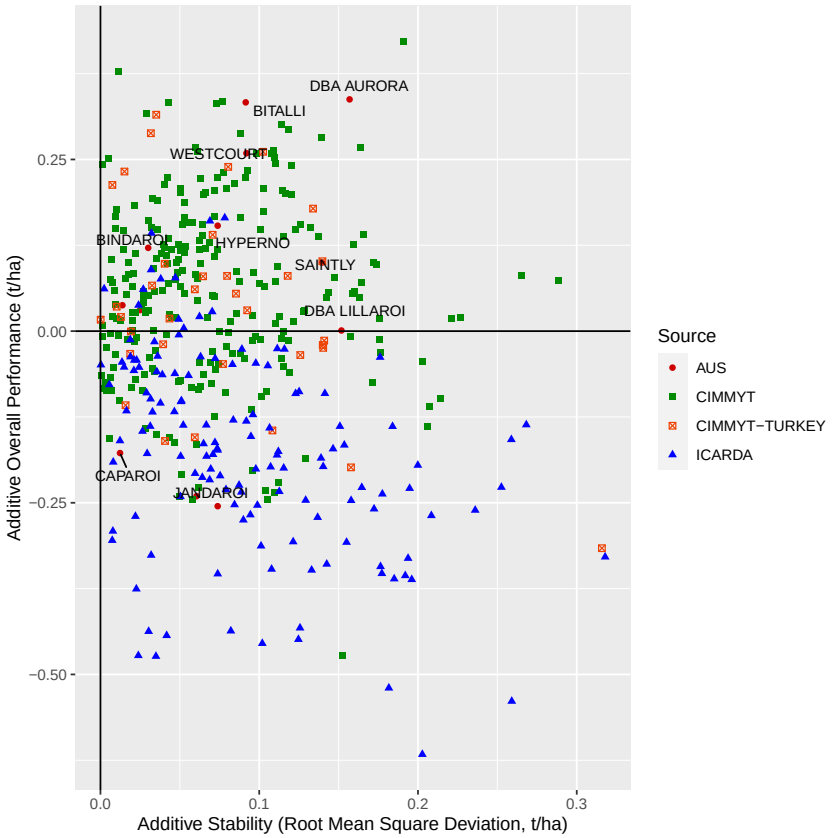

Figure 6: Factor Analytic Selection Tools: Overall Performance versus Stability (RMSD) for the additive genetic effects of the lines evaluated in 2020. The legend indicates the germplasm source. Australian checks are labelled.

Overall performance (OP) and root mean square deviation (RMSD) for the additive effects summarise the line performance (Figure 6). The first factor for the additive effects, on which OP and RMSD is based, accounts for 60% of the additive genetic variance. Points in the top left hand corner represent lines with good additive performance and stability across the sampled environments. Figure 6 demonstrates that there are number of imported lines which outperform the Australian varieties, demonstrating the potential value as parents from this material. For this dataset the CIMMYT, CIMMYT-TURKEY and ICARDA lines are similarly stable; the CIMMYT and CIMMYT-TURKEY lines tend to be better performing, based on the first factor, but there is considerable overlap and one should also consider the performance of these lines in individual environments.

The responsiveness statistics from FAST describe the remaining 24% of the 84% variance accounted for in the additive effects. Figure 7 presents the additive overall performance (OP, t/ha) compared to the responsiveness for factor 2. An inspection of the contrasting performance of the Australian check varieties can inform as to the underlying environmental covariates driving these factors. For example, for factor 2 the responsiveness of WESTCOURT and JANDAROI is in contrast with other Australian varieties such as DBA

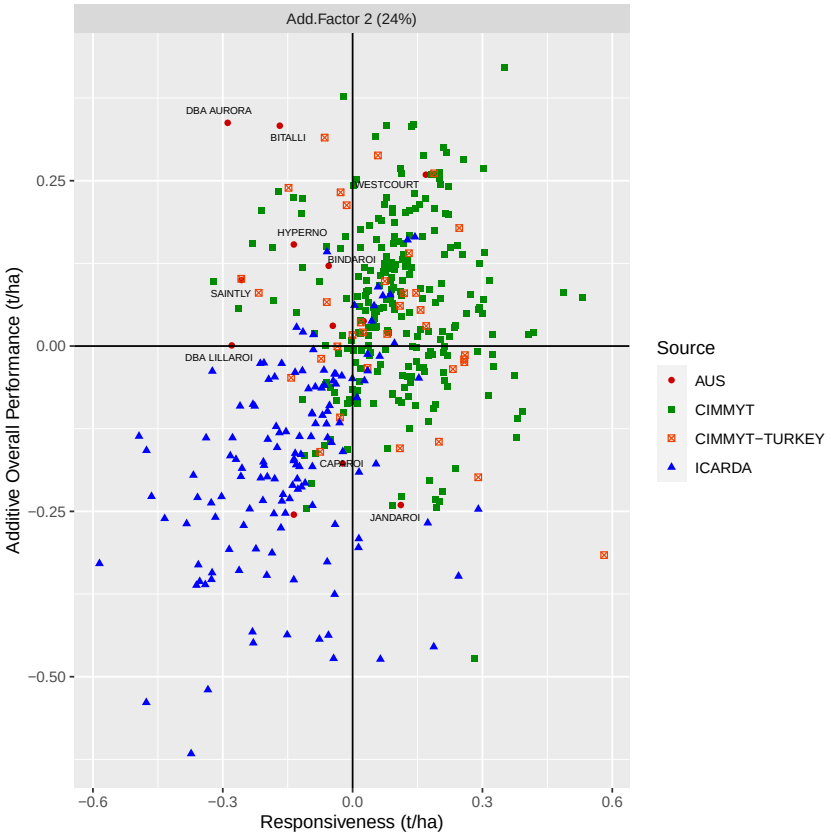

Figure 7: Factor Analytic Selection Tools: Overall Performance versus Responsiveness for the 2020 entries for the 2nd factor fitted to the additive genetic variance-covariance matrix. The legend indicates the germplasm source. Australian checks are labelled.

AURORA and BITALLI. It is also clear from Figure 7 that the second factor is discriminating between the ICARDA (blue triangles) and CIMMYT (green squares) lines. If one can understand the environmental driver of this second factor (e.g. latitude, say) then one can make selections appropriate for the target environment.

The number of lines by germplasm source which occur in the top 10% additive and total overall performers is of interest (Table 9). The numbers in this table have been summarised from the additive and total overall performance calculated for each line from the full dataset. The table shows that there are lines from both CIMMYT and ICARDA which are in the top 10% additive and/or total effects which, based on other traits, may be of interest to the collaborating breeder.

Table 9: Number of lines by germplasm source that were evaluated in 2020, and the number occurring in the top 10% Additive and Total based on overall performance.

Year	Description	Australian	CIMMYT	CIMMYT-TURKEY	ICARDA
2020	Number	12	250	35	137
2020	top 10% Additive	3	33	6	0
2020	top 10% Total	4	31	6	1

5 Results

It is important to understand that the FAST measures OP and RMSD summarise the results from a FA-LMM concisely but they are based on only the first factor estimated from the data. In this dataset the first factor for the additive component explains 60% of the 84% additive genetic variance accounted for by the model. The remaining 24% is the variety by environment interaction effect, which is important to consider when selecting parents for Australian environments. The re-ranking of the lines within environments is clear when the additive overall performance is plotted against the additive common variety by environment best linear unbiased predictions (additive CVE-eBLUPS) for each environment (Figure 8). This plot shows the results for the trials evaluated in 2020. The percentage of additive to total genetic variance for these environments ranged from 14 to 53%, respectively (Table 5). An example of specific additive adaptation is on comparing the Australian varieties WESTCOURT and DBA AURORA. These two varieties have similar additive overall performance (vertical axis, Figure 8). However, DBA AURORA has a higher additive VE-eBLUP in D20EDGE2, D20ROSE5 and D20TARL2 and WESTCOURT performs better in D20SPRR2 and D20SUNV2, (horizontal axis, Figure 8).

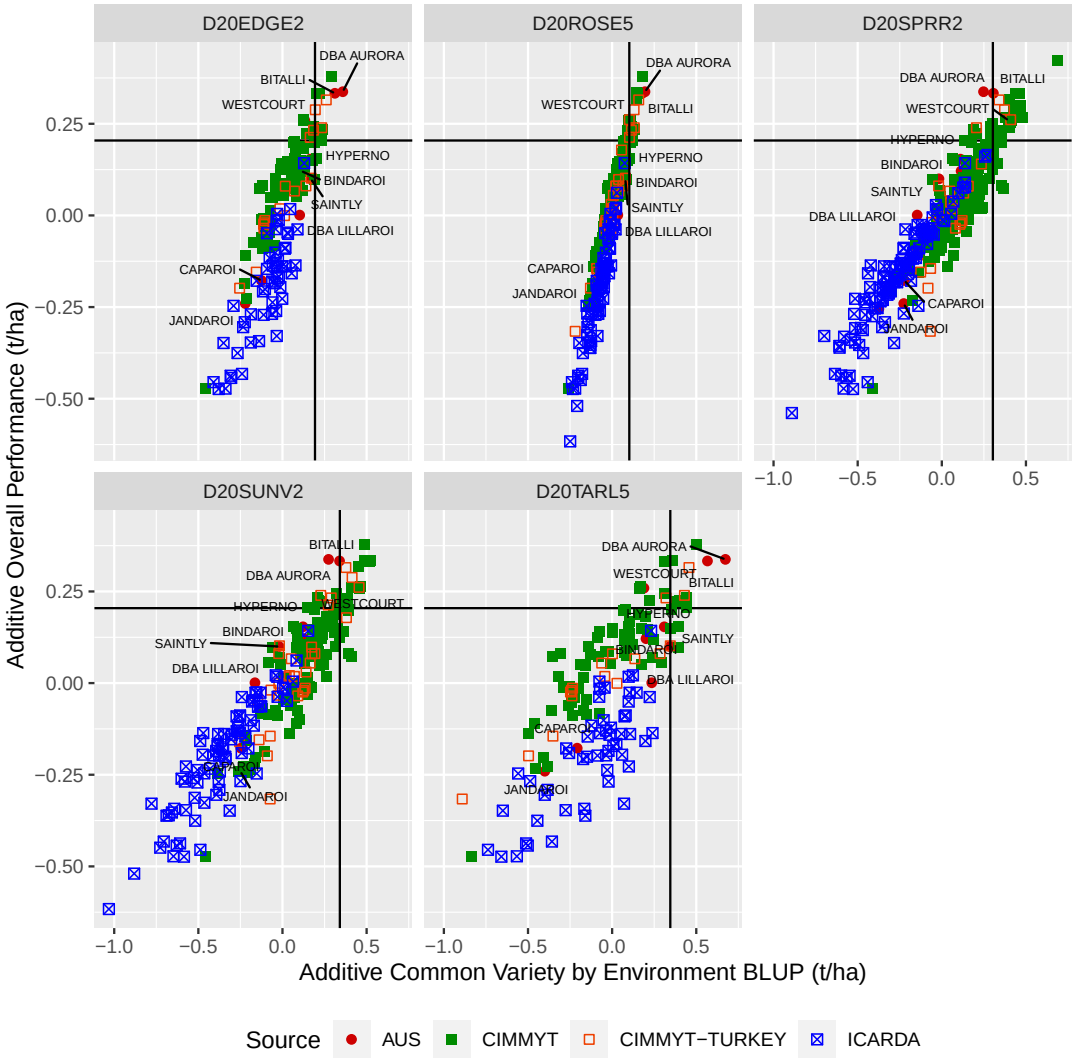

Figure 8: Additive Overall Performance compared to Additive Best Linear Unbiased Predictions (BLUPs) for lines evaluated in **2020**. Points above and to the right of the black horizontal and vertical lines, respectively, represent the top 10% lines based on overall performance (above) and individual environment BLUPs (right). The legend indicates the germplasm source. Australian checks are labelled.

The following results are provided in Excel worksheets:

- **VEeffects** - variety by environment effects for all variety by environment combinations. Columns indicating the number of replicates, presence of entry in the current year together with all passport information are provided alongside the additive and total CVE- and VE-eBLUPs, standard errors and ranks.

- **Trials** - environment level statistics including indications of TrialType (public or private), total and additive genetic variance and additive and total loadings for all factors. (Note, this tab name should probably be changed to **Environment**).

- **Varieties** - variety level statistics with passport information including the number of plots, environments and years an entry was evaluated in along side additive and total overall performance (OP), root mean square deviation (RMSD), responsiveness and scores.

- **CrownRotYieldPredictions** - additive and total empirical BLUPS for the D20TAMW2 trial.

- **Correlation matrices** - additive and total between environment correlation matrices

- **Connectivity matrices** - entry and parent level between environment connectivity matrices

The following yield result files were provided to the CAIGE team on 19 February 2021:

- *CAIGE Durum Wheat 2020-METresults-ALL-2021-02-19.xlsx* - all results for all trials

- *CAIGE Durum Wheat 2020-METresults-PUBLIC-2021-02-19.xlsx* - results only for publicly funded trials. FAST not provided because these are based on the whole dataset

- *CAIGE Durum Wheat 2020top 10% and top 10 rank entries-2021-02-19.xlsx* - top 10% and top 10 ranked entries

- *CAIGE-DURUM_DataCompilationFormat-2021-02-19.xlsx* - all results for the 2020 entries only provided in wide format for CAIGE team to add disease and other data too.

5.3 Phenology and height: 2020

The phenology and height data were not consistently measured at all CAIGE Durum wheat trials in 2020 (Table 6). For all three traits the genetic variance model included pedigree information, as for yield, in the form of the numerator relationship matrix to enable prediction of additive and total variety effects. The additive and non-additive genetic variances for these traits is provided in Table 10. These traits are considered highly heritable traits and thus it is curious that for Zadoks measured at D20SUNV2 the ratio of additive to non-additive genetic variance is less than 1. The values for Height and Days to Heading are fairly typical, although the reader should recall that the D20TAMW2 was inoculated with crown rot and this could affect the phenology of lines in this environment.

Table 10: Additive and non-additive genetic variances for the phenology and height traits in the CAIGE Durum wheat 2020 environments.

Env	Trait	Additive	NonAdditive
D20SUNV2	Height	0.136	0.094
D20SUNV2	Zadoks	2.773	4.372
D20TAMW2	Days to Heading	3.437	2.171

For all agronomic traits the individual environment additive and total variety BLUPS were estimated and are available in Excel file, *CAIGE-Durum-AgronomicTraitResults-2021-03-08.xlsx*.

6 Resources

This section is primarily as information for the author and team members of BBAGI.

The folder for this work is found on the BBAGI Dropbox account under `UOW-EIS-NIASRA-BBAGI/Projects/CAIGE/CAIGE20/Durum/Analysis/2-MET`.

The scripts for the analysis are located within the `1Scripts` folder and described below:

- 0-SSanalysis-2016-19.R - analysis of 2016-2019 trials in preparation for the 2016-2020 analysis, determining best spatial models and covariates.
- 1-SSanalysis-2020only.R - analysis of the 2020 trial only.
- 2-METDataprepare.R - adding the 2020 trials to the 2016-2019 dataset
- 2-METped2020-analysis.R - pedigree MET analysis for the 2020 dataset
- 2-METped2020-Results.R - results from the pedigree MET analysis of the 2020 dataset.
- 4-AgronomicTraits-analysis-2020.R - head file for the analysis of the agronomic traits. From this file the 4a-HT-analysis.R, 4b-DTH-analysis.R and 4c-ZADOKS-analysis.R- analysis files are stepped into and then summarised at the end of the script.
- 4a-Height-analysis.R - height analysis
- 4b-DTH-analysis.R - days to heading analysis
- 4c-ZADOKS-analysis.R - maturity zadoks analysis

The purpose of the remaining folders is:

- `1Plots` – plots generated throughout the code
- `1RData` – .RData files saved throughout code
- `1Results` – output files sent to CAIGE staff.

The scripts for this report are found in `UOW-EIS-NIASRA-BBAGI/Projects/CAIGE/CAIGE20/Durum/Analysis/2020-only/Reports/1METped`.

References

Burgueno, J., Crossa, J., Cornelius, P. L., McLaren, G., Trethowan, R. & Krishnamachari, A. (2007). Modeling additve x environment and additive x additve x environment using genetic covariances of relatives of wheat genotypes. *Crop Science* **47**, 311–320.

Butler, D. (2019). *pedicure: pedigree tools*. www.mmade.org. R package version 2.0.0.

Butler, D. & Cullis, B. (2019). *od: Generate optimal experimental designs*. R package version 2.0.0, www.mmade.org.

Butler, D., Cullis, B. & Thompson, R. (in prep.). Asreml: An R package to fit the linear mixed model. *Journal of Statistical Software* .

Cullis, B., Cocks, N., Smith, A. & Butler, D. (2018). Sparse multi-environment trial designs for early stage selection experiments in plant breeding programmes. *National Institute for Applied Statistics Research of Australia Working Paper Series, University of Wollongong, (https://niasra.uow.edu.au/workingpapers/index.html)* .

Cullis, B., Smith, A., Beeck, C. & Cowling, W. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. *Genome* **53**, 1002–1016.

Cullis, B. R., Smith, A. B., Cocks, N. A. & Butler, D. G. (2020). The Design of Early-Stage Plant Breeding Trials Using Genetic Relatedness. *Journal of Agricultural, Biological, and Environmental Statistics* .

Cullis, B. R., Smith, A. B. & E, C. N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics and Computing* **11**, 381–393.

Ganesalingam, A., Smith, A. B., Butler, D. G. & Cullis, B. (2019). Incomplete MET designs for early stage selection. Wheat Breeding Assembly, Adelaide.

Gilmour, A., Cullis, B., Welham, S., Gogel, B. & Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis* **44**, 571–586.

Gilmour, A. R., Cullis, B. R. & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *J. Agric., Biol. and Environ. Statist.* **2**, 269–293.

Gogel, B., Smith, A. & Cullis, B. (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.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* , 69–83.

REFERENCES

Jordan, M. & Cullis, B. (2020). The Numerator Relationship Matrix.

Mathews, K., Nicol, J., Micallef, S. & Trethowan, R. (2020). CAIGE Durum Designs 2020. Tech. rep., University of Wollongong.

Mathews, K., Wilson, R., Baranwal, D., Nicol, J. & Trethowan, R. (2021). CAIGE Bread Wheat 2017-2020 Across Year analysis. Tech. rep., University of Wollongong.

Mathews, K. L. (2021). Curation of CAIGE pedigree datasets. Tech. rep., BBAGI, University of Wollongong.

Meuwissen, T. H. E. & Luo, Z. (1992). Computing inbreeding coefficients in large populations. *Genetics, Selection and Evolution* **24**, 305–315.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. R. & Kuchel, H. (2006). Joint Modelling of Additive and Non-Additive Genetic Line Effects in Single Field Trials. *Theoretical and Applied Genetics* **113**, 809–819.

Oakey, H., Verbyla, A. P., Cullis, B. R., Wei, X. & Pitchford, W. S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics* **114**, 1319–1332.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smith, A., Cullis, B. & Thompson, R. (2001). Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. *Biometrics* **57**, 1138–1147.

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.

Smith, A. B., 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.

Tolhurst, D. J., Mathews, K. L., Smith, A. B. & 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.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist* , 330–8.

REFERENCES