Bioinformatics and Biometrics for the Australian Grains Industry

CAIGE Bread Wheat Designs 2020

GRDC Project No.: US00073

Jess Meza and Ky Mathews[1]
email:[1]kmathews@uow.edu.au

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

Julie Nicol, Sandra Micallef and Richard Trethowan
email:julie.nicol@sydney.edu.au

The Plant Breeding Institute
The University of Sydney

September 16, 2020

Contents

CONTENTS

Acknowledgements: Thanks to Brian Cullis for technical advice in generating these designs and clarifying ideas in the report.

This report describes the experimental design of the CIMMYT Australia ICARDA Germplasm Evaluation (CAIGE) bread wheat multi-environment trial (MET) series for 2020. The report is structured as follows: Section 1 provides Background information on the CAIGE project and purpose of this trial series; Sections 2 and 3 describe the Genetic Material and Trial Information, respectively. The experimental design process and results from the designs is described in Section 4. Section 5 describes the input, processing and output files used to generate these designs and report.

1 Background

The CIMMYT Australian ICARDA Germplasm Evaluation (CAIGE) project is a Grains Research and Development Corporation (GRDC) funded project managed by the University of Sydney in collaboration with the University of Queensland and Australian wheat breeders. 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 (environments) across the Australian wheat belt and selected by breeding companies to be included in their breeding programs and ultimately released to Australian wheat growers. The terms 'entries' and 'varieties' are used interchangeably.

2 Genetic Material

The CAIGE bread wheat 2020 entry list consists of 504 entries from 10 CAIGE imported nurseries and 14 Australian checks (Table 1). Entry names in CAIGE are a concatenation of the Quarantine Code (QCode) for the international nursery and Quarantine Number (QNo), the entry number within that nursery, such that QNo:QCode. It is important to understand that QNo is nested within QCode, so an entry with QNo=1 with QCode=ZWB19 is different from an entry with QNo=1 and QCode=ZIF19.

Table 1: CAIGE bread wheat entry list for 2020 described by the 10 CIMMYT and ICARDA nurseries with the QCode, nursery description, year of import and number of entries.

QCode	Description	ImportYear	nEntries
BWT19	CIMMYT-Turkey soil borne pathogen resistant lines	2019	22
Checks	Australian Checks	2020	14
SBP19	CIMMYT-Turkey soil borne pathogen resistant lines	2019	7
ZIF19	ICARDA selections	2019	151
ZIG18	RSI Dr Rajaram selections	2018	2
ZIG19	RSI Dr Rajaram selections	2019	50
ZIZ18	ICARDA selections	2018	2
ZIZ19	ICARDA breeder selections	2019	55
ZSA19	CIMMYT Stress Adapted Trait Yield Nurseries	2019	30
ZWB19	CIMMYT elite selections	2019	140
ZWY19	CIMMYT Wheat Yield Consortium Yield Trial	2019	31

2 Genetic Material

This year, for the first time, the pedigree information for these entries and the checks will be used to perform the designs following Cullis et al. (2018) and Ganesalingam et al. (2019a). The pedigree information was extracted by CAIGE staff from the BMS database which is used to manage CIMMYT and ICARDA germplasm by the CAIGE project. For the purposes of these designs only parent and grandparent information was used. It may be possible to use more generations given sufficient time to extract and audit the data. There are three numerical identifiers for a breeding line in the BMS system: Genotype ID (GID), Cross ID (CID) and Sister ID (SID). The latter two jointly identify a unique line. GID is the identifier of choice for the pedigree information. The pedigree information was provided in wide format with the parental GID and pedigree and grandparent GID to the right of the individual GID with the grandparent pedigree information in a separate worksheet. These data were manipulated into a format called "Me Mum Dad" 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.* An example of the pedigree file is provided in Table 2. The individual in the last row, 135:ZWB19, has a 'Mum' (female parent) with GID 7039324 and a 'Dad' (male parent) of 6334696. These parents are in rows 6 and 3, respectively, in this table, *above* the individual that will be tested in the field.

Table 2: Example of a 'Me Mum and Dad' file for the CAIGE Bread Wheat 2020 trials. Me is the individual being tested or an ancestor or an entry being tested, Mum and Dad contain the pedigree (GID in this dataset) for the parents. A value of zero for Mum or Dad indicates unknown pedigree or base individuals. Fn is the filial generation number and self is the level of selfing.

Me	Mum	Dad	Fn	self
5848468	0	0	6	5
6176013	4985684	4905277	6	5
6334696	5129008	5398047	6	5
6682480	5848468	5398424	6	5
6687200	6341425	6176013	6	5
7039324	6334696	6176013	6	5
10:ZWB19	6687200	6176013	6	5
116:ZWB19	6682480	6176013	6	5
135:ZWB19	7039324	6334696	6	5

Functions `chkPed` and `trimPed` from the `pedicure` (Butler, 2019) library on the R statistical computing platform (R Core Team, 2019) are used to curate the pedigree (i.e. "Me Mum Dad") file. `Pedicure` is freely available from www.mmade.org/pedicure. The final pedigree data frame contained 1430 records; the 504 entries under evaluation and 926 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 $n_f : n_b$, where n_f is the Fn

of the single plant selection 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

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

Entries from CIMMYT nurseries were assigned an Fn=6 (self=5) as the usual practice is to have F_6 derived lines submitted to international nurseries, whilst entries from ICARDA nurseries were assigned an Fn=4 (self=3) (R. Trethowan *pers. comm.*). Australian checks are considered fixed and assigned a self=7, whilst all parents and grandparents are assigned a self=5.

The inbreeding coefficient $\mathcal{F}$ (Henderson, 1976) is then calculated as

$$\mathcal{F} = 1 - \frac{1}{2}^{self}. \quad (2)$$

The diagonals of the numerator relationship matrix ($\boldsymbol{A}$) are equal to the 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). 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., 2018) following the process outlined in Gilmour & Dutkowski (2004) and Meuwissen & Luo (1992). `asreml` is licensed software available from www.vsni.co.uk.

To illustrate, the pedigrees for the three entries (10:ZWB19, 116:ZWB19 and 135:ZWB19) and two parents in Table 2 are provided for reference in Table 3, and the $\boldsymbol{A}$ matrix for these entries is presented in Table 4. A grandparent, GID = 6341425, is also included Table 4.

Table 3: Pedigrees for the 3 entries and two parents in Table 2 the CAIGE bread wheat 2020 pedigree file. Note that BORL14 is a synonym for REEDLING #1.

Entry	Pedigree
6176013	REEDLING #1
7039324	WBLL4/KUKUNA//WBLL1/3/WBLL1*2/BRAMBLING/4/REEDLING #1
10:ZWB19	WBLL1*2/BRAMBLING//WBLL1*2/BRAMBLING/3/2*BORL14
116:ZWB19	FRNCLN*2/KINGBIRD #1//REEDLING #1
135:ZWB19	WEEBILL4/KUKUNA//WEEBILL1/3/WEEBILL1*2/BRAMBLING*2/4/REEDLING #1

It is clear from the "Me Mum and Dad" file (Table 2) and the $\boldsymbol{A}$ matrix (Table 4) that entry 10:ZWB19 is more closely related to 116:ZWB19 (0.7383) than 135:ZWB19 (0.3691). This is because 135:ZWB19 does not have the same Dad as an ancestor.

Table 4: Numerator relationship (**A**) matrix for 3 entries, selected parents and grandparents in the CAIGE bread wheat 2020 pedigree file.

GID	10:ZWB19	116:ZWB19	135:ZWB19	7039324	6176013	6341425
10:ZWB19	1.9841	0.7383	0.3691	0.7383	1.4766	0.4922
116:ZWB19	0.7383	1.9688	0.2461	0.4922	0.9844	0.0000
135:ZWB19	0.3691	0.2461	1.9841	1.4766	0.4922	0.0000
7039324	0.7383	0.4922	1.4766	1.9688	0.9844	0.0000
6176013	1.4766	0.9844	0.4922	0.9844	1.9688	0.0000
6341425	0.4922	0.0000	0.0000	0.0000	0.0000	1.9688

3 Trial Information

There are fourteen trials planned for the 2020 CAIGE bread wheat MET series. They are distributed across the Australian wheat belt and managed by ten institutions. The type (public or commercial), site, institution responsible and trial design specifications are presented in Table 5. The trials will be designed as a partially-replicated (p-rep) multi-environment trial series following Cullis et al. (2020) and Ganesalingam et al. (2019b). Narrabri and Roseworthy are considered home sites and at least one replicate of each entry is assigned for evaluation in each of these environments. The number of plots in the remaining trials were determined such that they did not exceed the allowable dimensions, all available seed was utilised, and the number of plots for each trial with similar allowable dimensions is as similar as possible.

Table 5: Trial site and design specifications for CAIGE bread wheat trials 2020.

Type	State	Site	Institution	Contact Person	No.Range	No.Row	No.Plots	Block Direction	g/plot
Commercial	SA	Balaklava	LongreachPB	Bertus Jacobs	12	35	420	Range	50
Commercial	NSW	Breeza	S&W Seeds	Chris Moore	12	24	288	Range	50
Commerical	WA	Corrigin	Intergrain	Allan Rattay	12	24	288	Range	50
Commercial	QLD	Dalby	Rebel Seeds	Derrick Mickelborough	12	24	288	Range	50
Public	QLD	Gatton	UQ	Mark Dieters	12	24	288	Range	50
Commercial	WA	Goomalling	LongreachPB	Bertus Jacobs	12	35	420	Range	30
Commercial	NSW	Junee	LongreachPB	Bertus Jacobs	12	35	420	Range	30
Commerical	VIC	Longerenong	BASF	Maqbool Ahmad	12	35	420	Range	55
Commercial	WA	Mingenew	Intergrain	Allan Rattay	12	24	288	Range	50
Public	NSW	Narrabri	USYD	Annette Tredea	24	30	720	Row	50
Commercial	NSW	North Star	AGT	Meiqin Lu	24	18	432	Range	50
Commercial	SA	Roseworthy	AGT	James Edwards	24	30	720	Range	50
Public	VIC	Swan Hill	USyd (Kalyx)	Bridget Doyle	12	24	288	Range	50
Public	WA	York	Living Farm	Ian Edwards	12	24	288	Row	50

Annette Tredea (USYD) and Mark Dieters (UQ) requested alternate designs be provided for Narrabri and Gatton such that the Range and Row dimensions were reversed but the blocking direction remained the same. This was to allow for some flexibility in arranging the trials at these sites.

Table 6: Alternate design specifications for CAIGE bread wheat trials 2020 at Gatton and Narrabri

Type	State	Site	Institution	Contact Person	No.Range	No.Row	No.Plots	Block Direction	g/plot
Public	QLD	Gatton	UQ	Mark Dieters	24	12	288	Range	50
Public	NSW	Narrabri	USYD	Annette Tredea	30	24	720	Row	50

4 Experimental Design

The experimental design process allocating entries to plots within sites consists of two steps. The first allocates entries to sites and the second is allocates entries to plots within sites. Each step is described in detail.

The designs were performed using the `od` library (Butler & Cullis, 2019) on the `R` statistical platform (R Core Team, 2019). This software allows considerable flexibility including modelling of variance matrices, such as the $\boldsymbol{A}$ matrix and use of realistic starting parameter estimates.

4.1 Allocation of entries to sites

The allocation of entries to sites needs to consider the amount of seed available for each entry, the number of available plots within and across sites and any design constraints or logistical requirements of the researcher. In this dataset, the following constraints were accommodated:

- the 14 Australian checks are allocated to 2 plots at each site;
- Roseworthy and Narrabri are considered "home" sites and hence at least one replicate of each entry needs to be allocated to each of these sites;
- the remaining 12 sites should have trial sizes (i.e. number of total plots) as similar as possible.

There were 5,325 50g seed packets in total. Each of the fourteen checks had 30 packets and test entry distribution ranged from six entries with two packets through to one entry with 29 packets. Not all trials required 50 g/plot (Table 5). The initial allocation of entries to sites aimed to allocate all available seed and resulted in a total of 5,568 plots across all sites. This process was complex and multi-staged, with seed packets allocated in batches and remaining seed amounts calculated based on actual site seed requirements, which was then converted into remaining 50g seed packets for further allocation. For instance 120:ZIZ19 had 198.7 g seed available, or three 50g plots, which are sown at Narrabri, Roseworthy and North Star. These three sites require 50 g/plot each leaving 48.7 g of seed enabling for a fourth 30g plot to be sown at Junee, thus maximising use of all available seed. Seed was allocated in such batches until all trials were rectangular and seed use maximised.

We implement the Design Tableau approach of Smith & Cullis (2019) to determine the linear mixed model required for the allocation of entries to sites. The approach defines the plot and treatment factors and structure, describes the design function and the progression from the randomisation-based model to the final model used to generate the design.

Step 1 plots and associated factors

- plots (observational units) = packets (5325 units initially)
- plot factors = {U, Site (14), Packet (5325)}

Step 2 treatments and associated factors

- treatments = entries (504)
- treatment factors = {1, Entry (504)}

Step 3 design function

- 14 checks have 2 packets in each site
- at least one packet of all entries is at Narrabri and Roseworthy
- the remaining 12 sites to a similar number of entries

Step 4 anatomical variables

- pedigree information in the form of an $\boldsymbol{A}$ inverse matrix

Step 5 No extraneous variables

Step 6 plot structure

- U/Site/Packet = U + Site + Site:Packet

Step 7 treatment structure

- 1 + Entry

Step 8 aliased terms: "1" and "U".

Step 9 Design Tableau

Working model 0 (WM0, in Table 7 is the randomisation based model where treatment factors are considered as fixed and plot factors are considered as random (Bailey, 2008). However, the purpose of the CAIGE yield trials is to select entries which requires estimation of entry ranks obtained from using the empirical best linear unbiased predictions (E-BLUPs), thus entry is fitted as random (WM1, in Table 7). The numerator relationship matrix, $\boldsymbol{A}$, was used to ensure related individuals are well distributed across all sites. Similarly, the term `Site`, a plot factor, is consider as random effect in WM0 but we choose to fit it as fixed in WM1 because we wish to provide E-BLUPs which are

Table 7: Design tableau for the CAIGE bread wheat 2020 design phase I: allocation of entries to sites. Terms of the linear mixed model are given as well as their state, fixed (F) or random (R), for the sequence of working models (WM0–WM1). Variance models for random terms in the final model (WM1) are also given. The dashed line separates treatment and plot factors.

Term	WM0	WM1	Variance model
1[U]	F	F	-
Entry	F	R	$\sigma_g^2 \boldsymbol{A}_{504}$
U[1]	-	-	
Site	R	F	
Site:Packet	R	R	$\sigma^2 \boldsymbol{I}_{5297}$

comparable across all sites and fitting `Site` as fixed enables this.

Step 10 Optimal design (`od`) call

The variance parameter estimates from the CAIGE bread wheat 2019 analysis (Lambert et al., 2020) were used to provide starting values for the 2020 designs. Pedigree information was not used in the 2019 analysis and hence only an empirical estimate for the *total* genetic variance was available. However, the estimates of the additive and non-additive genetic variances can be derived by partitioning the total genetic variance appropriately. We describe this for one site. First, the total genetic effects, $\boldsymbol{u}_g$, for m varieties (entries) can be partitioned into additive ($\boldsymbol{u}_a$) and non-additive ($\boldsymbol{u}_e$) components as per Oakey et al. (2006).

$$\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e. \tag{3}$$

The variance matrix for $\text{var}(\boldsymbol{u}_a) = \sigma_a^2 \boldsymbol{A}$, where σ_a^2 is the additive genetic variance and $\boldsymbol{A}$ is the numerator relationship matrix. The variance matrix for $\text{var}(\boldsymbol{u}_e) = \sigma_e^2 \boldsymbol{I}_m$, where σ_e^2 is the non-additive genetic variance and $\boldsymbol{I}_m$ an identity matrix of size m. Thus, the total variances for the first 5 entries in $\boldsymbol{u}_g$ can be written as

$$\begin{aligned}
\text{var}(u_{g1}) &= \sigma_{g1}^2 = \sigma_{a1}^2 a_{11} + \sigma_{e1}^2 \\
\text{var}(u_{g2}) &= \sigma_{g2}^2 = \sigma_{a2}^2 a_{22} + \sigma_{e2}^2 \\
\text{var}(u_{g3}) &= \sigma_{g3}^2 = \sigma_{a3}^2 a_{33} + \sigma_{e3}^2 \\
\text{var}(u_{g4}) &= \sigma_{g4}^2 = \sigma_{a4}^2 a_{44} + \sigma_{e4}^2 \\
\text{var}(u_{g5}) &= \sigma_{g5}^2 = \sigma_{a5}^2 a_{55} + \sigma_{e5}^2.
\end{aligned} \tag{4}$$

Now to estimate the additive and non-additive genetic variance parameters for the 2020 designs we back-calculate from the estimate of total genetic variance at the site level for a typical entry using the following formulae,

$$\hat{\sigma}_g^2 = \bar{a}\hat{\sigma}_a^2 + \hat{\sigma}_e^2 \tag{5}$$

where $\bar{a}$ is the mean of the diagonal of the elements of $\boldsymbol{A}$. For this dataset $\bar{a} = 1.9307$.

The total genetic variance parameter estimates from the ten 2019 sites ranged from 0.010 to 0.269, with a mean of 0.088. Assuming that the additive variance is q (0.8 for this dataset) of the total genetic variance, the starting values for $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ for the allocation of entries to sites were determined to be 0.036 and 0.018, respectively, using the following formulae

$$\begin{aligned} \hat{\sigma}_a^2 &= q\hat{\sigma}_g^2/\bar{a} \\ \hat{\sigma}_e^2 &= \hat{\sigma}_g^2 - \bar{a}\hat{\sigma}_a^2 \end{aligned} \tag{6}$$

The `od` model call is:

```
allo1 <- od(fixed=~Site,
            random=~ ric(Entry, ainv),
            residual=~ units,
            permute=~ ric(Entry, ainv),
            swap=~ SwapBlk,
            data = base.df, search='tabu', maxit=2)
```

where

- `ainv` is the inverse **A** matrix
- `ric()` is the Ricardo model which allows use of `ainv`
- `SwapBlk` is a factor with 6 levels where `checks` = 14 checks with 2 packets each allocated to each site; `Narrabri` and `Roseworthy` = allocation of one packet for each entry exactly once to Narrabri and Roseworthy, respectively; `A` = allocation of all remaining packets across all sites with 50 g/plot; `B` = allocation of all remaining packets across all sites with 30 g/plot; and `C` = allocation of all remaining packets across all sites with 55 g/plot. `B` and `C` were necessary to ensure that no swaps were made which violate seed availability.
- `Entries` are swapped within `SwapBlk` levels which are defined across `Sites`
- `base.df` is a data frame 5,568 records long including the 14 `Site`, `Entry` with all possible available packets, `Range`, `Row` and `Block` specifications.
- the search algorithm was a tabu algorithm as described in the `od` manual (Butler & Cullis, 2019).

4 Experimental Design

The result of this allocation was an optimal spread of entries across sites (Table 8). For example, there are 6 entries with 2 packets and each of them was allocated to 2 sites, there are 8 entries with 3 packets and each of them was allocated to 3 sites etc.

Table 8: Original number entries allocated to sites summarised by the number of packets by number of sites.

Number	Number of packets																										
of Sites	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	30
2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	45	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	70	11	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	55	7	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	42	14	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	14	10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	5	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	13	10	14	7	13	7	5	3	4	3	14	1

Table 9: Original number of plots and entries allocated to each site, level of replication in terms of the number of entries with 1, 2, 3 and 4 packets per site and the overall percentage partial replication.

Site	nPlots	nEntry	$p = 1$	$p = 2$	$p = 3$	$p = 4$	%p-rep
Balaklava	420	403	386	17	0	0	4
Breeza	288	271	254	17	0	0	6
Corrigin	288	232	176	56	0	0	24
Dalby	288	253	218	35	0	0	14
Gatton	288	246	204	42	0	0	17
Goomalling	420	318	216	102	0	0	32
Junee	420	356	292	64	0	0	18
Longerenong	420	336	252	84	0	0	25
Mingenew	288	271	254	17	0	0	6
Narrabri	720	504	379	49	61	15	25
North Star	432	415	398	17	0	0	4
Roseworthy	720	504	390	42	42	30	23
Swan Hill	288	265	242	23	0	0	9
York	288	270	252	18	0	0	7

4.2 Reduction of entries allocated to sites

On 31st March 2020, Julie Nicol informed Ky Mathews that the allocations of entries to sites in Table 9 were not feasible for Narrabri, Roseworthy, North Star and Longerenong, as they were too large. The confusion had arisen as we misunderstood an instruction to "use all available seed packets". A review of the designs was undertaken and the following actions taken:

1. **Reduction of plot numbers for Narrabri, Roseworthy, North Star and Longerenong.**

2. **An inspection of the original allocation of entries to plots within sites, for all sites, indicated that they were not optimal and new randomisations were provided for all trials except York which had already been sown.**

3. **On 7 April 2020, Julie Nicol advised Ky Mathews that the trials allocated to Dalby and Swan Hill were now to be grown at Dulacca and Lalbert, respectively, and the design files and tables updated accordingly.**

The seed had already been sent to collaborators and hence it was not possible to reallocate entries across sites. The following reductions within sites were made

- Narrabri reduced from 720 plots to 624 with no change in the number of entries; primarily by reducing the number of any entries with $p > 2$ packets (Table 8) to a maximum of 2 plots. For the 61 entries with p=3 packets all were allocated to 2 plots. For the 15 entries with p=4 packets, 10 were allocated to 2 plots and 5 to 1. The entries selected to be in 1 plot were well represented at other sites.

- Roseworthy reduced from 720 plots to 576 with no change in the number of entries; primarily by reducing the number of any entries with $p > 2$ packets (Table 8) to a maximum of 2 plots. For the 42 entries with p=3 packets, 15 were allocated to 2 plots and 27 to 1. For the 30 entries with p=4 packets, 15 were allocated to 2 plots and 15 to 1.

- North Star reduced from 432 plots and 415 entries to 336 plots and 319 entries; primarily by removing 96 entries with 1 plot because there are only 17 replicated entries (14 of these checks). All 96 entries were present at another site.

- Longerenong reduced from 420 plots and 336 entries to 288 plots and 258 entries; primarily by reducing entries with 2 packets to 1 plot whilst maximising representation of family groups. For 84 entries with p=2 packets 30 were allocated to 2 plots, 14 of these were checks and the remaining 16 were selected based on maximising the presence of family groups at the site. Similarly, 78 entries were selected for removal from the trial based on the number of plots across the full multi-environment trial dataset and their family group representation. This step was done manually.

The result of these reductions can be seen in Table 10. The final allocation of entries to sites is less orthogonal that the original design, as expected with *post hoc* sampling, but is reasonable.

Table 10: Final number of ranges, rows, plots and entries allocated to each site, level of replication in terms of the number of entries with 1 and 2 packets per site and the overall percentage partial replication. Grey shaded rows are sites which changed from the original allocation in Table 9.

Site	nRange	nRow	nPlots	Block Direction	nEntry	$p = 1$	$p = 2$	%p-rep
Balaklava	12	35	420	Range	403	386	17	4
Breeza	12	24	288	Range	271	254	17	6
Corrigin	12	24	288	Row	232	176	56	24
Dulacca	12	24	288	Range	253	218	35	14
Gatton	12	24	288	Range	246	204	42	17
Goomalling	12	35	420	Range	318	216	102	32
Junee	12	35	420	Range	356	292	64	18
Lalbert	12	24	288	Range	265	242	23	9
Longerenong	12	24	288	Range	258	228	30	12
Mingenew	12	24	288	Row	271	254	17	6
Narrabri	24	26	624	Row	504	384	120	24
North Star	24	14	336	Range	319	302	17	5
Roseworthy	24	24	576	Range	504	432	72	14
York	12	24	288	Range	270	252	18	7

4.3 Allocation of entries to plots within sites

The allocation of packets of entries to plots within sites required the following considerations:

- there are two blocks at each site in the dimension described in Table 5;
- maximise resolvability of entry replicates across blocks, for example, for entries with 2 packets of seed allocated to a site one packet is allocated to each block;
- accommodate spatial variation within each site, where management practices and subsequent variation is in the range direction and extraneous variation is in the row direction.

The Design Tableau approach of Smith & Cullis (2019) to determine the linear mixed model required for the allocation of entries to plots within one site, Narrabri, is as follows. Narrabri has 24 ranges and 26 rows resulting in 624 plots. There are two blocks in the range direction such that Block 1 = Ranges 1:12 and Block 2 = Ranges 13:24, Table **??**.

Step 1 plots and associated factors

- plots (observational units) = plots (624 units)
- plot factors = {U, Block (2), Plots (624)}

Step 2 treatments and associated factors

- treatments = entries (504)
- treatment factors = {1, Entry (504)}

Step 3 design function

- Entries allocated to plots such that blocks are maximally resolvable and the pairwise difference between entries is minimised.

Step 4 anatomical variables

- numerator relationship matrix, $\boldsymbol{A}$
- Range (24) and Row (26) where the intersection of a range and row indexes a plot.

Step 5 No extraneous variables

Step 6 plot structure

- U/Block/Plot = U + Block + Block:Plot

Step 7 treatment structure

- 1 + Entry

Step 8 aliased terms: "1" and "U".

Step 9 Design Tableau

Table 11: Design tableau for the CAIGE bread wheat 2020 design phase II: allocation of entries to plots within a site. Terms of the linear mixed model are given as well as their state, fixed (F) or random (R), for the sequence of working models (WM0–WM1). Variance models for random terms in the final model (WM1) are also given. The dashed line separates treatment and plot factors.

Term	WM0	WM1	Variance model
1[U]	F	F	-
Entry	F	R	$\sigma^2_g \boldsymbol{A}_{504}$
U[1]	-	-	
Block	R	R	$\sigma^2_B \boldsymbol{I}_2$
Range	-	R	$\sigma^2_{Ra} \boldsymbol{I}_{24}$
Row	-	R	$\sigma^2_{Ro} \boldsymbol{I}_{26}$
Block:Plot	residual	R	$\oplus_i \sigma^2 \Sigma(\rho_{Ra}) \otimes \Sigma(\rho_{Ro})$

Working model 0 (WM0, in Table 11) is the randomisation-based model where treatment factors are considered as fixed and plot factors are considered as random (Bailey, 2008). However, the purpose of the CAIGE yield trials is to select entries which requires estimation of entry ranks obtained from using the empirical best linear unbiased predictions (E-BLUPs), thus entry is fitted as random (WM1, in Table 11). The numerator relationship matrix, $\boldsymbol{A}$, was used to ensure related individuals are well distributed across the site. Anatomical terms `Range` and `Row`, with variances σ^2_{Ra} and σ^2_{Ro}, were included as random terms in working model 1 (WM1) to facilitate the spread of entries across the field. In the analysis of field experiments the residual spatial variation is invariably fitted using a separable first-order auto-regressive (AR1) process in both the range and row direction (Gilmour et al., 1997). However, in the design of field experiments using the AR1 × AR1 residual spatial can lead to designs where replicated entries are clustered together and/or allocated to plots on the trial edges (Cullis et al., 2006). The trial at York used the original randomisation provided in mid-March 2020 where the separable AR1 × AR1 residual variance model was applied. This is a statistically efficient but currently unpalatable to CAIGE collaborators who are familiar with designs where duplicated entries are evenly balanced and distributed across an experiment. Hence, for the remaining sites the residual model was independent in both range and row directions.

Step 10 Optimal design (`od`) call

In practice, the allocation of entries to plots within sites occurred for all sites (except Gatton and York) simultaneously, not one site at a time as described above. There was a single call to `od` in this phase. The randomisation of entries to plots for Gatton was performed separately due to time constraints with looming COVID19 restrictions.

This `od` call used appropriate starting values estimated from a linear mixed model analysis of the previous year's dataset (Lambert et al., 2020). The variance parameter estimates for `Block`, `Range` and `Row` from the 2019 analysis were averaged to determine appropriate starting values. The starting values are presented in Table 12.

Table 12: Starting values for od call to allocate entries to plots within sites. All sites had the same residual variance, hence for brevity Site etc is used to summarise this information for the remaining trials.

Component	Value
ric(Entry, ainv)!Entry_vm	0.036
ric(Entry, ainv)!Entry_ide	0.018
Site:SwapBlock	0.042
Site:nSwapBlock	0.500
Site:Range	0.012
Site:Row	0.005
Site_Junee!R	0.124
Site_Balaklava!R	0.124
Site_ etc	0.124

The `od` call for the final model which allocates entries to plots within sites is:

```
des1 <- od(fixed=~Site,
           random=~ ric(Entry, ainv) + Site:SwapBlock +
             Site:nSwapBlock + Site:Range + Site:Row,
           residual=~units,
           permute=~ ric(Entry, ainv),
           swap=~Site:SwapBlock,
           G.param = des1.sv, R.param = des1.sv,
           reorder = c("Source", "EntryNo", "Pedigree", "Designation",
                       "CID", "SID", "GID", "RepInEnv", "LineRep"),
           data = des0$design, search = "tabu+rw", maxit= 50)
```

where

- `ainv` is the inverse $\mathbf{A}$ matrix
- the permutation is performed on the `Entry` modelled with the $\mathbf{A}$ matrix using a Ricardo model
- the swap is performed such that entries cannot move across combinations of `Site`

and `SwapBlock`. Recall that replicated entries at a site were allocated to one block each in the dataframe setup. `SwapBlock` contains either RangeBlock or RowBlock specific to each site as defined in Table 10. `nSwapBlock` is the complement of `SwapBlock`, i.e. is RangeBlock if `SwapBlock` is RowBlock and enables blocking in both directions.

- `des0$design` is the design data frame resulting from the allocation of entries to sites (Section 4.1).

- `sv` are the starting values provided in Table 12

- the search algorithm was a combined tabu and random walk algorithm as described in the `od` manual (Butler & Cullis, 2019).

The experimental design layout for Narrabri is presented in Figure 1.

The final design files provided to CAIGE staff and collaborators contain:

- entry specific information such as GID, CID, SID, QNo and QCode, Pedigree and Selection History

- design factors such as Site, Range, Row, Block, RepInEnv (number of packets for an entry per site), LineRep (packet number for an entry within a site when in Range, Row order)

4 Experimental Design

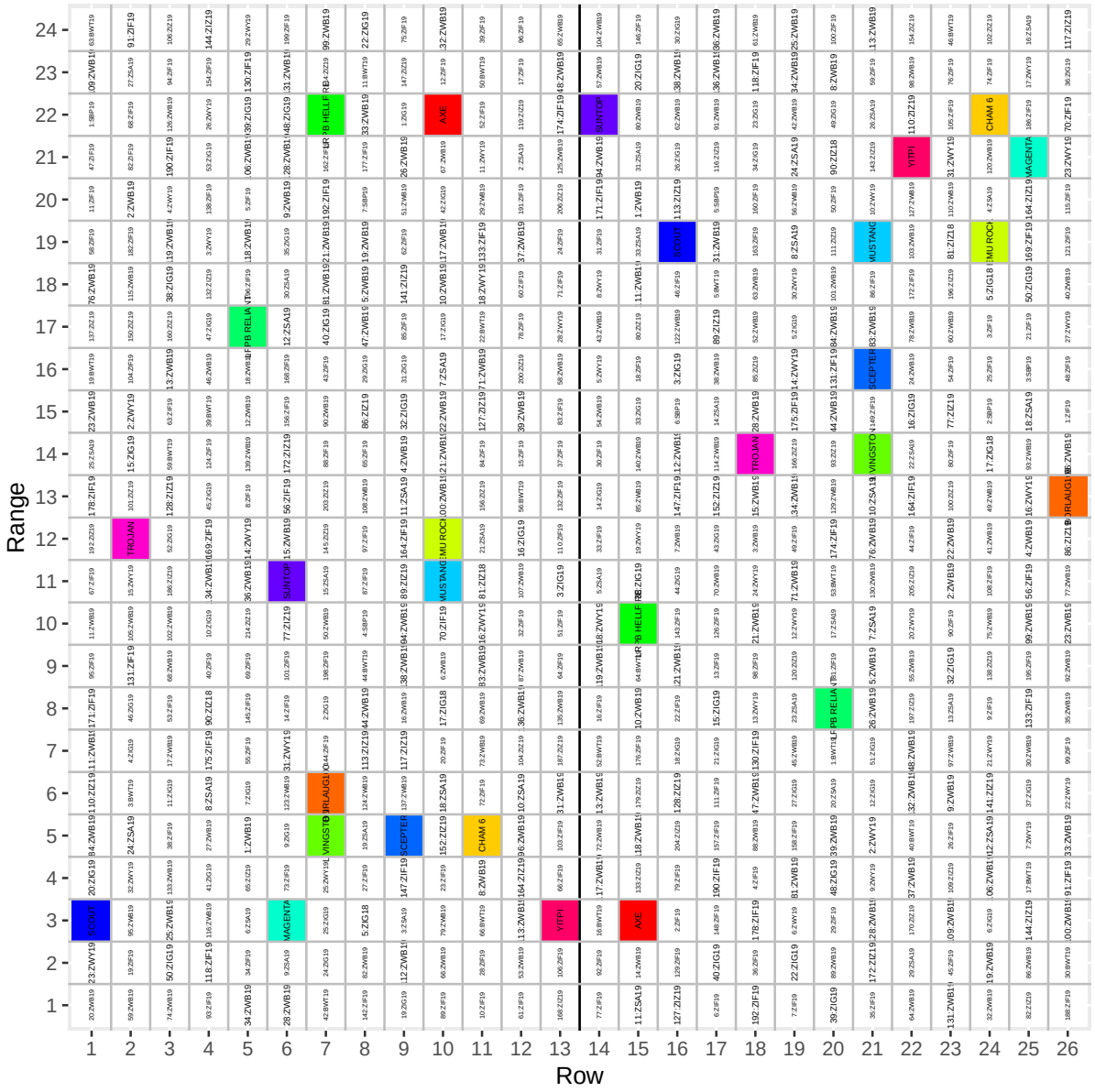

Figure 1: Field layout for CAIGE bread wheat trial at Narrabri, 2020. The solid vertical line delineates the Blocks. Colours represent Australian checks. The size of the font indicates the number of replicates, 1 or 2.

5 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/Bread Wheat/Design`.

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

- 01-CAIGE_2020_BW_DataPrepare.R – prepares the base dataframe (`base.df`) following the specifications dictated by CAIGE staff
- 02-CAIGE_2020_BW_Pedigree.R – cleans and formats the pedigree information provided by CAIGE staff in preparation for use in the designs
- 03-CAIGE_2020_BW_Allocation1.R – the calls to `od` for allocating entries to sites as described above, including design checks. This includes preparing the data based on allocation of entries to sites for packing of seed
- 04-CAIGE_2020_BW_Allocation2.R – the calls to `od` for allocating entries within sites as described above, including design checks. This includes formatting the final design files for sending to breeders and CAIGE staff including design.xlsx and layout.pdf files.
- 05-CAIGE_2020_BW_Alternate.R – the calls to `od` for allocating entries within sites as described above for the alternate configurations for Narrabri and Gatton, including design checks.
- 06-CAIGE-2020-BW-redesign-setup.R – process of subsetting plots from the 4 sites which required reduced plots as described in Section 4.2.
- 07-CAIGE-2020-BW-redesign.R – the call to `od` for allocating entries within sites as described in Section 4.3.
- 08-FieldLayout.R – prepares design data for output in the form of .csv files and layout plots in pdf for each site. This is where the updating of sites names Dalby and Swan Hill to Dulacca and Lalbert.
- metide.R – design functions provided by Brian Cullis to assist in assigning starting values based on previous analyses.

The purpose of the remaining folders is:

5 Resources

- `1DataOrig` – data files provided by CAIGE staff
- `1RData` – .RData files saved throughout code
- `1Report` – files for this report. Includes a subfolder called `sections` where .Rnw for Sections 4.1 and 4.3 are stored.
- `1Results` – output files sent to CAIGE staff.

References

Bailey, R. A. (2008). Design of Comparative Experiments .

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.* www.mmade.org. R package version 2.0.0.

Butler, D., Cullis, B., Gilmour, A. & Thompson, R. (2018). Asreml: An r package to fit the linear mixed model. *NIASRA Working Paper, University of Wollongong* .

Cullis, B., Cocks, N., Smith, A. & Butler, D. (2018). Sparse multi-environment trial designs for early stage selection experiments in plant breeding programmes. Eucarpia, Ghent.

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 Agric., Biol. and Env. Statist.* **11**, 381–393.

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

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

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 Env. Statist.* **2**, 269–293.

Gilmour, A. R. & Dutkowski, G. W. (2004). Pedigree options in ASReml 3 .

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.

Lambert, N., Mathews, K., Nicol, J., Micallef, S. & Trethowan, R. (2020). CAIGE Bread Wheat Analysis 2019. *BBAGI Technical Reports* .

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

Oakey, H., Verbyla, A. P., S, P. 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.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

REFERENCES

Smith, A. & Cullis, B. (2019). Design Tableau: An aid to specifying the linear mixed model for a comparative experiment. Fisher Memorial Lecture.

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

REFERENCES