Bioinformatics and Biometrics for the Australian Grains Industry

Technical Report Series

GRDC Project Code:US00073

Analysis Report: CAIGE Bread Wheat Yield Trial MET 2016

Jodine O'Connor[1] and Ky Mathews
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong
[1]jodine@uow.edu.au

Richard Trethowan
Professor Plant Breeding
Director of the IA Watson Grain Research Centre
Sydney Institute of Agriculture
University of Sydney
richard.trethowan@sydney.edu.au

November 7, 2017

1 Executive Summary

This report describes the analysis of the CAIGE Bread Wheat yield Multi-Environment Trial (MET) dataset for 2016.

The data were provided in a timely manner and in a consistent format following the ICIS database conventions. The data are available via the CAIGE website www.caigeproject.org.

The trials were designed as *p*-rep trials by SAGI-II staff, similar to those described in Chong et al. (2016).

The trials were located in eight locations across four states: NSW, VIC, SA and WA. Due to extensive frost damage at the trial site at Cadoux (WA) the yield data was not provided and hence the MET dataset contained the seven remaining trials.

A factor Analytical approach (Smith et al., 2001) was used for the MET analysis. The final factor analytic model was of order 3 (FA3) and accounted for 69.5% of the variance. An FA3 model is the highest order that can be performed on a MET dataset with seven trials.

The common variety effects (CVEs) from the final model were estimated and provided to the CAIGE community via the CAIGE website www.caigeproject.org.au. A PV-Plus plot (Smith et al., 2015) was produced for the lines that most frequently occurred in the top 10 across all trials.

The heatmap of between environment genetic correlations showed that there was significant genotype by environment interaction in this data set.

2 Introduction

The CIMMYT-Australian-ICARDA Germplasm Evaluation (CAIGE) project (US00073), funded by the Grains Research and Development Corporation (GRDC) imports and evaluates bread and durum wheat germplasm developed by the International Maize and Wheat Improvement Center (CIMMYT) and the International Center for Agricultural Research in the Dry Areas (ICARDA) in Australian environments. A companion project (UQ00043) peforms a similar role with barley. The key objective of the CAIGE project is to evaluate germplasm developed by CIMMYT and ICARDA in Australian environments and thus enable Australian Breeding companies to have access to novel sources of germplasm for disease and adaptation. The germplasm is trialed in different environments across Australia's wheat growing regions and potentially selected by breeding companies for inclusion in their breeding programs.

This report describes the Multi-Environment Trial (MET) analysis for the eight bread wheat trials conducted by breeding companies for the CAIGE project in 2016. The key trait of interest is yield.

3 Description of Data

In 2016, a total of 239 varieties (synonymous with entries) were evaluated at 8 locations across Australia. The variety list consists of 100 entries from ICARDA (CWA15), 75 from CIMMYT (RAV15), 50 from Dr. Rajaram (ZIG14) and 14 Australian checks. These varieties were distributed as evenly as possible across 8 locations in the Australian wheat growing region, Table 1.

Table 1: Number of varieties, number of plots, *p*-rep percentage and trial mean yield (TMY) for CAIGE Bread Wheat trials 2016

Location	State	Organisation	#Entries	#Plots	p%	TMY (t/ha)
Balaklava	SA	LPB	201	288	43.3	6.42
Cadoux	WA	Intergrain	228	338	38.2	NA
Horsham	VIC	Bayer	195	252	29.2	6.94
Junee	NSW	LPB	201	288	43.3	4.42
Narrabri	NSW	USYD	239	391	61.1	5.61
North Star	NSW	AGT	238	360	51.3	4.29
Roseworthy	SA	AGT	198	264	33.3	5.56
Toodyay	WA	Edstar	200	288	44.0	5.13

Connectivity of varieties between trials is an important point to consider in multi-environment trial (MET) analyses. Due to seed limitations it was not possible to evaluate all varieties in all trials, however, the degree of connectivity between trials is sufficient to provide accurate REML estimates, see Table 2.

The experimental design accommodated this imbalance through the partial replication (p-rep) paradigm of Cullis et al. (2006). The experimental designs were performed by SAGI-II

Table 2: Connectivity of Entries across Locations - Bread Wheat 2016

Location	Balaklava	Horsham	Junee	Narrabri	North Star	Roseworthy	Toodyay
Balaklava	201	190	200	200	200	188	199
Horsham	190	195	191	195	195	181	191
Junee	200	191	201	201	201	189	200
Narrabri	200	195	201	239	238	197	200
North Star	200	195	201	238	238	197	200
Roseworthy	188	181	189	197	197	198	189
Toodyay	199	191	200	200	200	189	200

(Chong et al., 2016) in 2016 using the od software (Butler, 2016) in R (R Development Core Team, 2015). The experimental design contained 2 replicate blocks in the column direction (Replicate) and the experimental unit (EU) is the intersection between the columns (Range) and rows (Row).

The trial at Cadoux in Western Australia experienced a severe frost and the yield data were not collated. Thus, there were yield data for seven trials available for the MET analysis.

4 Statistical Analysis

A one-stage MET analysis was conducted for the seven trials with available yield plot data. For this analysis we model the spatial trends at each trial as per Gilmour et al. (1997) and use the factor analytic (FA) approach of Smith et al. (2001) to model the genotype by environment (G×E) variance matrix.

4.1 Linear Mixed Model

In a randomisation based model there are both blocking and treatment factors and the experimental design and purpose of analysis dictates the structure of those factors. The blocking factors in each of the CAIGE trials was the same, Replicate, Row, Range the intersection of Range and Row with in a Trial is the smallest experimental unit (EU), Plot. The blocking structure for the MET analysis is then

```
Trial/[Replicate/Plot)
```

which, following Wilkinson & Rogers (1973) expands to

```
Trial + Trial:Replicate + Trial:Replicate:Plot.
```

The final term indexes both the EUs and the observational units (OUs) defined as the smallest unit on which a response will be measured and is equivalent to the residual. This model formula is used to define the *random* model formula in the **ASReml-R** package (Butler et al., 2015).

The treatment factor, based on a randomisation based model is `Variety` only. However, typically in MET experiments the aim is to model the V×E interaction and the main effect of Variety is often not explicitly fitted in the MET analysis, see for example (Smith et al., 2001). Hence the treatment structure is given by `Variety:Trial`.

For a randomisation based model, blocking factors are generally fitted as *random* and treatment factors are fitted as *fixed*. However, the aim of this MET analysis is to predict the genetic effects of the `Varieties` on yield (t/ha) and model the V×E interaction, hence the final mixed model formula is

```
fixed= ~ 1 + Trial
random= ~ Trial:Variety + Trial:Replicate + Trial:Replicate:Plot
```

where `Trial:Replicate:Plot` represents the residual variation. Spatial variation at each `Trial` is accounted for by using the separable autoregressive spatial structure (AR1×AR1) to model the residual variance of each trial (Gilmour et al., 1997).

4.2 Analysis

The analysis commenced by identifying if any recorded covariates were suitable for inclusion in the model. Next, any outliers were identified and queried with the researcher(s). In the third step, the spatial variation in the individual trials was modelled and once these were determined the analysis proceeded using factor analytic models for the V×E variance matrix.

For the first three steps in this process the V×E matrix is modelled with a diagonal (`DIAG`) structure which effectively fits all trials in the dataset but allows for separate genetic and residual variances for each trial.

Two covariates, shattering and lodging, were reported for the Roseworthy trial, however both were deemed to be genetically driven and were not included in subsequent models. The spatial terms fitted to the final model are given in Table 3.

The factor analytic modelling process commences with one factor (k=1) and continues until either the limit of the data is reached or the overall percentage variance accounted for reaches 80%. For example, the limit for this dataset with p=7 trials is k=3 factors because an increase in the number of factors will result in more parameters being estimated than are possible in the fully unstructured model ($p(p+1)/2 = 28$). For the FA3 model, there are $pk + p - k(k-1)/2 = 25$ parameters to estimate. Table 4 shows the loglikelhood and percent variance explained from each model fitted and Table 5 shows the percent variance accounted for (%vaf) at each `Trial` in the FA3 model.

Table 3: Final Spatial Models fitted to each Trial, 1=fitted, 0=not fitted

Location	linear Range	linear Row	random Range	random Row	Replicate
Balaklava	0	0	1	1	1
Horsham	0	1	1	1	1
Junee	0	0	1	1	1
Narrabri	0	1	0	0	1
North Star	1	0	1	0	1
Roseworthy	0	0	1	0	1
Toodyay	0	1	1	0	1

Table 4: Summary of models fitted to CAIGE Bread Wheat MET dataset 2016

Model	REML Loglikelhood	%vaf
DIAG	-62.74	-
FA1	238.44	51.5
FA2	254.65	64.9
FA3	265.77	69.5

The low percentage of variance accounted for in the Horsham Trial (32.5%) was of concern and the breeder and trial management team asked if there could be any explanation of this (CAIGE Annual General Meeting (6th March 2017)). Additional information was received indicating that there were issues at the time of planting due to very wet soil and blocked tynes. This information was included in the model but was not significant, and subsequently dropped.

The final asreml-R call was

```
asr.rr3ar  < −  asreml(yield ∼ Trial +
                 at(Trial, mt$lrange):lin(Range) +
                 at(Trial, mt$lrow):lin(Row),
                random = ∼rr(Trial, 3):Variety + diag(Trial):Variety
                 at(Trial):Replicate +
                 at(Trial, mt$rrange):Range +
                 at(Trial, mt$rrow):Row,
                residual = ∼dsum(∼ar1(Range):ar1(Row) | Trial,
                levels = mt$resid$aa),
                na.action = na.method(y='include', x='include'),
                data=nvtdata,G.param = gam, R.param = gam)
```

Table 5: Mean Yield (t/ha) and Percent % variance accounted for at each Trial in the final FA3 model of the CAIGE Bread Wheat MET analysis 2016

Trial	Mean Yield (t/ha)	%vaf
Balaklava	6.45	75.5
Horsham	6.94	32.5
Junee	4.42	66.1
Narrabri	5.61	100.0
North Star	4.29	77.4
Roseworthy	5.56	81.3
Toodyay	5.13	92.2

where `mt` is a list formed in `R` by a function `model.fit()` which converts the information in Table 3) to a list with a component for each term to be fitted in the model (i.e. non-zero for Table 3). Each component of `mt` is a vector of `Trial` names at which the term will be fitted. For example, `mt$lrange` contains the trial name "North Star" as a global trend at North Star in the range direction was considered significant. In practice, the factor analytic models were fitted using reduced rank (`rr`) and diagonal (`diag`) terms, in order for the appropriate REML estimates of the common variety by environment effects (CVE) to be obtained easily.

4.3 Results

The between environment genetic correlation matrix is shown in Figure 1. Toodyay and North Star are highly correlated, indicating that the variety ranks were similar for these two trials. In contrast, Horsham and Balaklava have no correlation and the variety rankings are not similar, indicating that with respect to yield these environments are dissimilar.

The results generated by this MET analysis included the common variety by environment effects (CVE effects, t/ha) for each `Trial` and `Variety`combination and a measure of the accuracy of the estimation. The CVE effect (previously referred to as `regblup`) is the empirical best linear unbiased prediction of the common variety by environment effects (t/ha). They are the predicted values for that part of the total V×E variety effects attributed to the common V×E interaction.

A very useful and effective interpretation tool to display the CVE effects developed by Smith et al. (2015) is the production value (PV)-Plus plot. This plot is commonly used to present the results of the National Variety Trial (NVT) system. The top 10 `Varieties` that performed in the top 10 most often were selected for demonstration of the PV-Plus plot. The PV-Plus plot is shown in Fig2. The dashed line represents the expected average yield, adjusted to zero for all varieties included in the dataset at that site. A positive production value (CVE) indicates that the variety is expected to yield higher than the average and a negative production value (CVE) indicates that the variety is expected

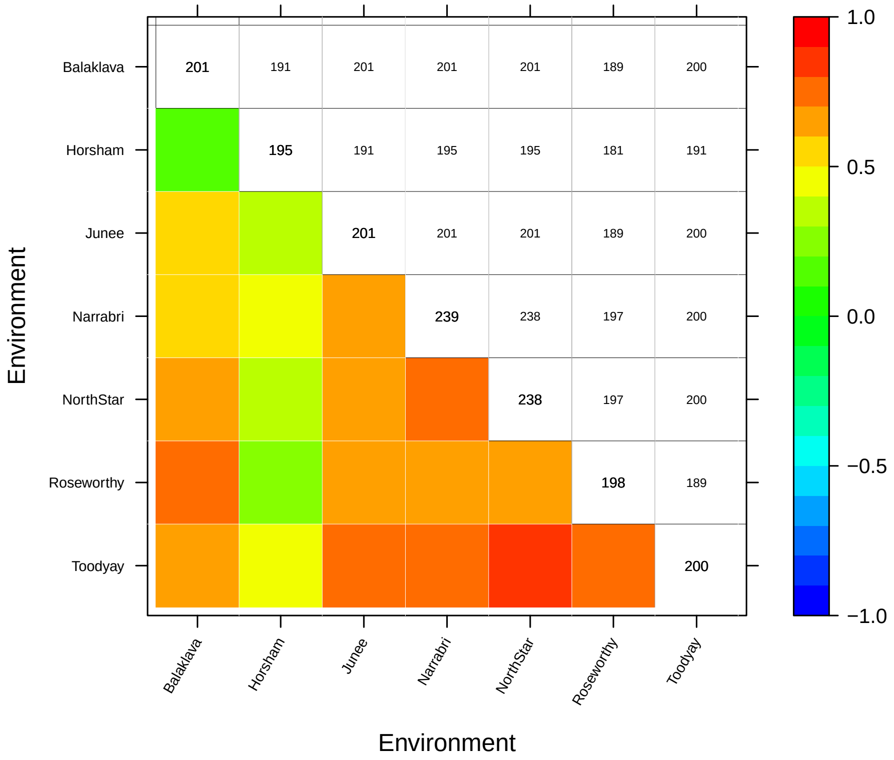

Figure 1: Genetic correlations between trials in the lower triangle and number of varieties in common in upper triangle and number of varieties in each trial on diagonal.

to yield lower than the average and a production value (CVE) of zero indicates that the variety is expected to yield on average. It is clear that there are varieties that are performing better than two of the Australian commercial check varieties, Axe and Mace, across all the locations.

4.4 File Management

Analysis was carried out by both Jodine O'Connor and Ky Mathews. Raw data files, analysis files, R script files and results files are located on the hard drive of Jodine O'Connor's computer located in */home/brian/jodine/CAIGE/Wheat/3. Analysis/2016* and backed up to external hard drive.

An Excel workbook *CAIGE-BW2016-METresults-FINAL.xlsx* containing all results was sent to Dr Richard Threthowan and associates of CAIGE on 17th March, 2017.

References

Butler, D. (2016). od: Generate optimal experimantal designs. *R package version 0.75.*

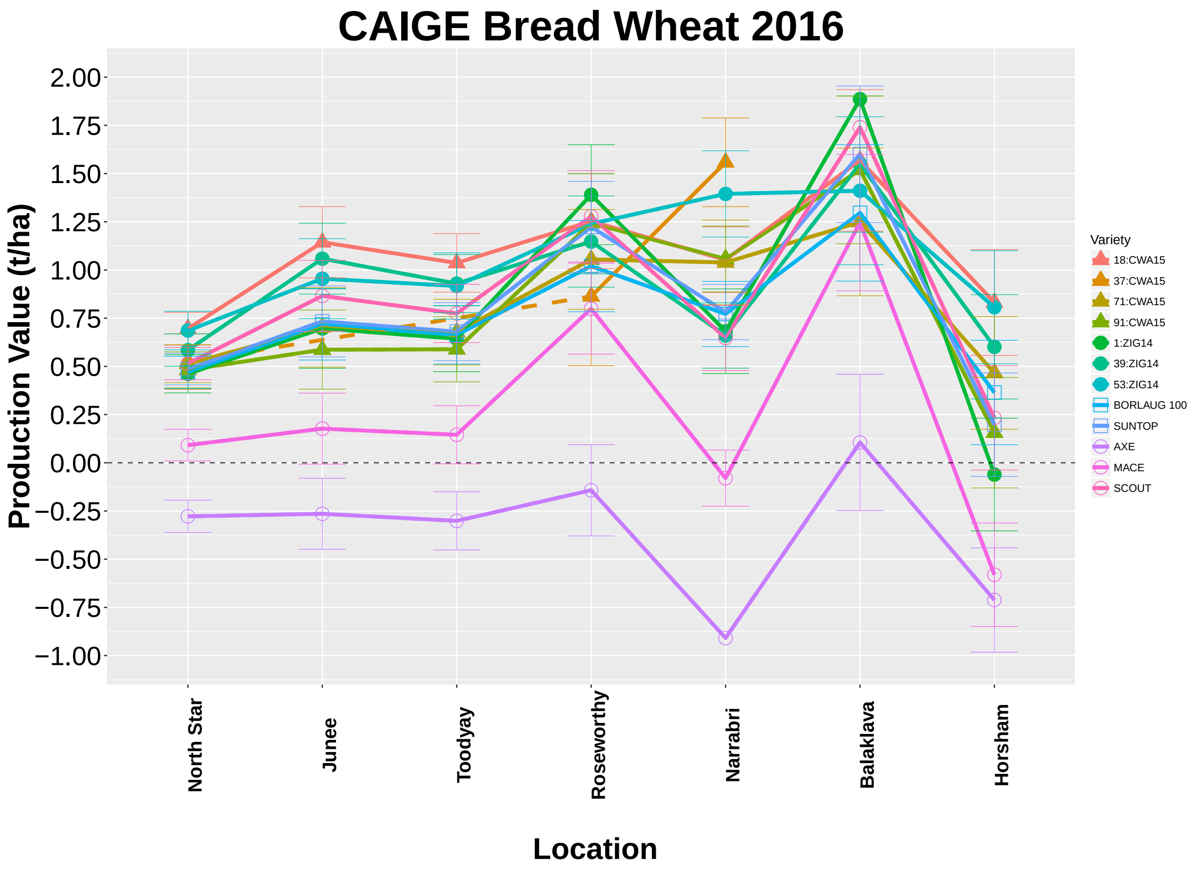

Figure 2: PV-Plus plot of Top 10 performing Varieties across all Trials Production value (t/ha) and standard errors for seven Locations (Trials) in the CAIGE Bread Wheat MET Analysis 2016

Technical report. .

Butler, D., Cullis, B. R., Gilmour, A. R., & Gogel, B. (2015). Asreml: An r package to fit the linear mixed model.

Chong, Y., Tolhurst, D., & Cullis, B. (2016). Design of CAIGE wheat trials 2016. Technical report.

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

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* pages 269–293.

R Development Core Team, R. (2015). R: A Language and Environment for Statistical Computing.

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. 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.

Wilkinson, G. N. & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Appl. Statist.* **22**, 392–399.