Bioinformatics and Biometrics for the Australian Grains Industry

Technical Report Series

GRDC Project Code:US00073

Analysis Report: CAIGE Durum 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 Durum 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 managed in seven locations across two states: NSW and SA. Three trials were excluded for environmental reasons: water logging, lodging and crown rot inoculation.

The trial deliberately inoculated with crown rot (Tamworth) was analysed separately for response to crown rot, although a control treatment (i.e. no inoculum) was not applied.

A factor Analytical approach (Smith et al., 2001) was used for the MET analysis. The final factor analytic model was of order 1 (FA1) and accounted for 59.3% of the variance. An FA1 model is the highest order that can be performed on a MET dataset with four 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) performs 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 seven durum 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 201 varieties (synonymous with entries) were evaluated at 7 locations across Australia. The variety list consists of 106 entries from CIMMYT (ZDG15), 16 from ICARDA's 38th IDYN (ZDK15), 71 from ICARDA's 38th IDON (ZDL15) and 8 Australian checks. These varieties were distributed as evenly as possible across 7 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), percentage variance accounted for (%vaf) and breeder comments for CAIGE Durum Wheat trials 2016

Location	State	Organisation	no. Varieties	no. plots	%p-rep	TMY	%vaf	Comments
Kaniva	SA	Uni of Adelaide	140	192	37.10	6.84		severe lodging
Kapunda	SA	Uni of Adelaide	157	192	22.30	6.54	51.2	good
Narrabri	NSW	Uni of Sydney	154	192	24.70	5.09	34.1	good
Roseworthy	SA	Uni of Adelaide	167	192	15.00	6.77	89.9	good
Tamworth	NSW	NSWDPI	180	240	33.30	3.40		crown rot inoculated
Wagga Wagga	NSW	AGT	141	192	36.20	NA		water logged
Walgett	NSW	AGT	165	192	16.40	5.97	0.5	low %vaf

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.

Table 2: Connectivity of Varieties across Locations - Durum Wheat 2016

	Kaniva	Kapunda	Narrabri	Roseworthy	Tamworth	Walgett	Wagga Wagga
Kaniva	140	122	116	118	140	111	135
Kapunda	122	157	144	131	145	127	121
Narrabri	116	144	154	128	140	127	118
Roseworthy	118	131	128	167	152	152	115
Tamworth	140	145	140	152	180	151	141
Walgett	111	127	127	152	151	165	115
Wagga Wagga	135	121	118	115	141	115	141

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 in 2016 (similar to Chong et al. (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`).

4 Statistical Analysis

4.1 Yield multi-environment trial

Three of the trials were excluded from the MET dataset as the main aim was to assess yield potential in these environments. Kaniva experienced severe lodging and the breeder reported that no seed was kept for subsequent use (*pers. comm.* Jason Able 28 Feb 2017). The Tamworth trial was deliberately inoculated with crown rot (*Fusarium roseum* Graminearum.) and hence was excluded from this MET for yield potential, and analysed separately. The Wagga Wagga trial experienced severe water logging and was not harvested.

A one-stage MET analysis commenced using the four trials with available yield 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.

4.1.1 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 within 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).

4.1.2 Analysis

There were no covariates reported for these trials and so the analysis commenced by identifying any outliers and confirming these with the researchers. Next, 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 preceding steps 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.

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

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

Location	linear Range	linear Row	random Range	random Row	ColRep
Kapunda	1	0	0	1	1
Narrabri	1	1	0	1	1
Roseworthy	0	0	0	0	1
Walgett	0	0	0	0	1

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=4 trials is k=1 factors because an increase in the number of factors will result in more parameters being estimated than are possible in the fully unstructured model ($p(p+1)/2 = 10$). For the FA1 model, there are $pk + p - k(k-1)/2 = 8$ parameters to estimate. Table 4 shows the loglikelihood and percent variance explained from each model fitted and Table 1 shows the percent variance accounted for (%vaf) at each trial in the FA1 model.

Table 4: Summary of models, Residual maximum likelihood (REML) loglikelihood and percentage variance accounted for (%vaf) fitted to CAIGE Durum Wheat MET dataset 2016

Model	REML Log likelihood	%vaf
DIAG	-219.84	-
FA1	-169.23	56.3

The final `asreml-R` call was

```
asr <- asreml(yield ~ Trial +
              at(Trial, mt$lrange):lin(Range) +
              at(Trial, mt$lrow):lin(Row),
             random = ~fa(Trial, 1):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)
```

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 "Kapunda" as a global trend at Kapunda 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.

4.1.3 Results

The between environment genetic correlation matrix from the analysis with all four trials is shown in Figure 1. The highest correlation is between Kapunda and Roseworthy, both southern sites, whereas the correlations are lower between Kapunda and Narrabri and Roseworthy and Narrabri. It is clear from this heatmap that Walgett had zero correlation with the remaining trials and together with the low genetic variance for this trial will account for the low %vaf for this location. This suggests that this trial was compromised in some way but there are no trial notes to indicate this and no feedback from the researchers regarding its potentially poor performance.

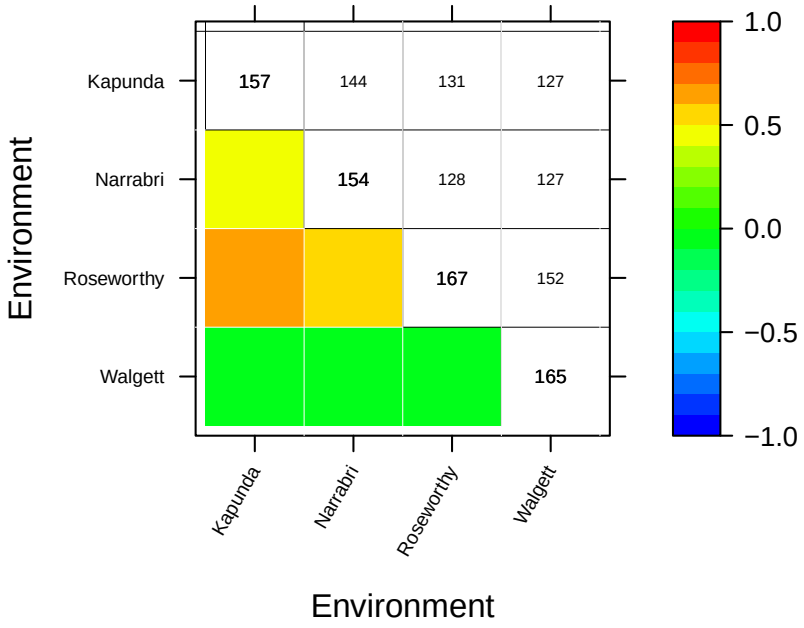

Figure 1: Genetic correlations between trials in the lower triangle and number of varieties in common in upper triangle and number of varieties in each trial on diagonal.

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 Fig 2. 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 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 the Australian commercial check varieties, DBA Aurora and Saintly, across all the locations. The top performing lines, at all locations, were all from CIMMYT (ZDG15).

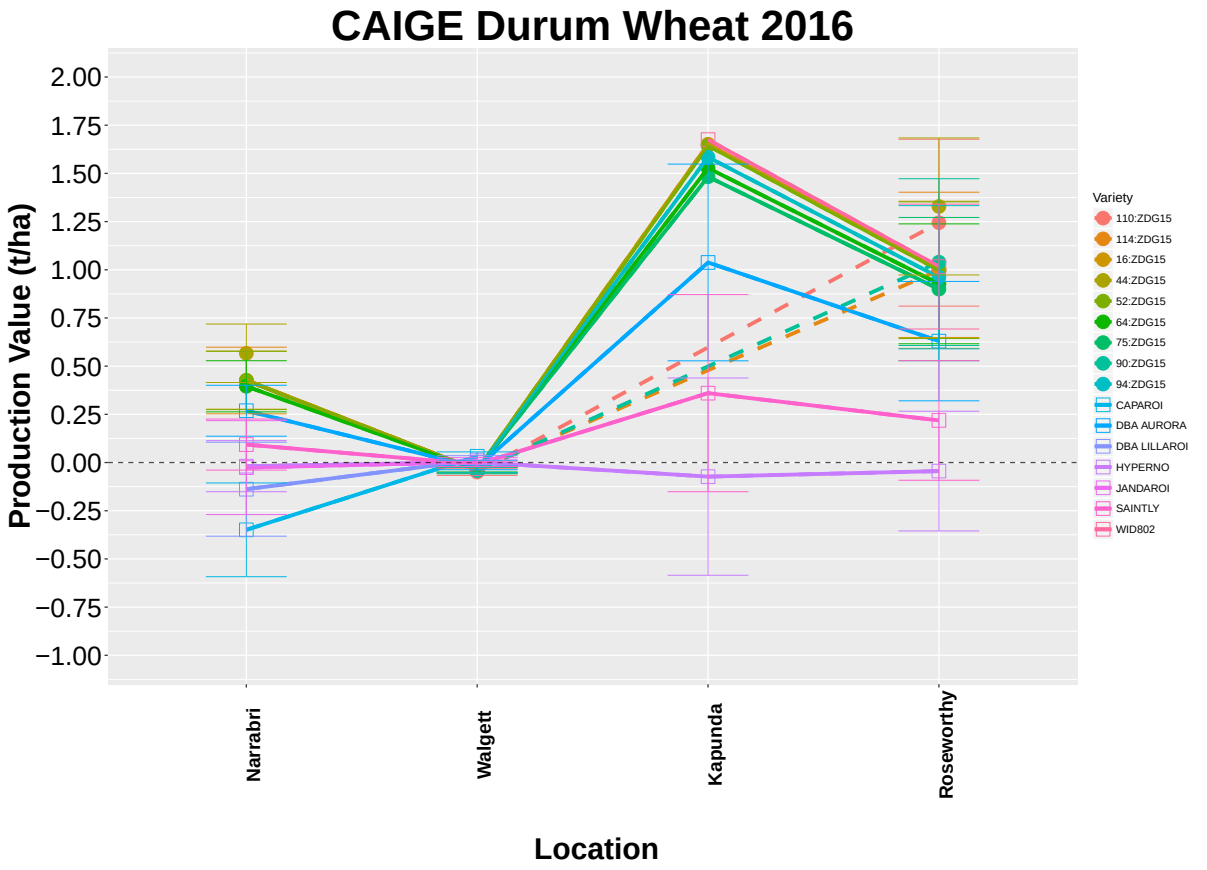

Figure 2: PV-Plus plot of Top 10 performing Varieties across all Trials Production value (t/ha) and standard errors for four Locations (Trials) in the CAIGE Durum Wheat MET Analysis 2016

4.2 Yield response to crown rot

The trial at Tamworth was deliberately inoculated with crown rot and hence for these analyses it was decided to analyse it separately from the other trials. In hindsight, it would be reasonable to include this in the MET analysis as the results for this trial would benefit from using more information in the other trials.

A single-site spatial analysis was performed as per (Gilmour et al., 1997) and the predicted values provided in an Excel spreadsheet to the breeders and the CAIGE team.

This analysis was performed by Ramethaa Pirathiban (UOW).

5 File Management

The MET analysis was carried out by 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 Trethowan and associates of CAIGE on 17th March, 2017. This accompanies this report.

An Excel workbook *CAIGE-durum-tamworth-2016-results.xls* containing the Tamworth trial results was sent to Dr Richard Trethowan and associates of CAIGE on 1st May, 2017. This accompanies this report.

References

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

REFERENCES

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.