Bioinformatics and Biometrics for the Australian Grains Industry Technical Report Series

Analysis Report: CAIGE Durum Wheat Yield Trial MET 2017

Ramethaa Pirathiban[1], Jesse Rand and Ky Mathews
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong
[1]ramethaa@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

February 26, 2018

1 Executive Summary

This report describes the analysis of the CAIGE Durum Wheat yield Multi Environment Trial (MET) analysis for 2017. 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 (http://www.caigeproject.org.au/).

Trials were designed as p-rep trials by BBAGI staff for seven locations across three states of Australia - NSW, VIC and SA: Narrabri, Rowena, Lockhart, Kaniva, Roseworthy, Kapunda and Tamworth. The trial at Lockhart was not harvested leaving six trials for inclusion in the MET analysis. In addition, the Rowena trial was grown at North Star due to the original site being unsuitable.

A one-stage Factor Analytic approach (Smith et al., 2001) was applied to the raw plot data for the MET analysis. The final factor analytic model was of order 3 (FA3) and accounted for 91% of the variance. The common variety effects (CVEs) from the final model were estimated and provided to the breeders/trial managers via the CAIGE website. PV-Plus plots (Smith et al., 2015) were produced for the top performing non-Australian varieties and Australian checks for all trial sites. In addition, these results are available in an interactive APP, https://nvtsagi.shinyapps.io/caigepvplus/, which is the recommended way to interrogate the MET analysis results. The heatmap of between environment genetic correlations showed that there was significant genotype by environment interaction in this MET data set.

2 Introduction

CAIGE is a GRDC (Grains Research and Development Corporation) funded project to evaluate bread wheat, durum wheat and barley germplasm developed by the International Maize and Wheat Improvement Center (CIMMYT) and the International Center for Agricultural Research in the Dry Areas (ICARDA).

The key objective of CAIGE is to evaluate germplasm developed by CIMMYT and ICARDA for inclusion in Australian Wheat Breeding Programs. The germplasm is trialed in different environments across the Australian wheat-belt and selected by breeding companies to be included in their breeding programs and ultimately released to the Australian wheat growers.

This report describes the Multi-Environment Trial (MET) analysis for the Durum Wheat trials conducted by breeding companies for the CAIGE project in 2017. The key trait of interest is yield. This was the second yield trialing season for Durum Wheat in the CAIGE project. There were seven trials sown for Durum Wheat across the wheat belt in Australia, six trials were harvested and used in the final MET analysis.

3 Description of Data

In 2017, a total of 125 varieties (synonymous with entries) were evaluated at the seven locations across Australia. The variety list consists of 25 entries from CIMMYT, 92 from ICARDA and 8 Australian checks. These varieties were distributed as evenly as possible across seven locations in the Australian wheat-belt (Table 1).

Table 1: Trial location, state, management organisation, total number of varieties, number of CIMMYT varieties, number of ICARDA varieties, number of Australian check varieties, number of plots, trial mean yield (TMY) and percentage variance accounted for (%vaf) for CAIGE Durum Wheat trials 2017.

Location	State	Organisation	Number of Varieties	Number of CIMMYT	Number of ICARDA	Number of Checks	Number of Plots	TMY (t/ha)	%vaf
Kaniva	SA	Uni Adelaide	88	3	81	4	132	5.64	100.00
Kapunda	SA	Uni Adelaide	94	6	84	4	144	5.15	99.99
Narrabri	NSW	AGT	92	6	82	4	144	3.97	47.49
Northstar	NSW	AGT	91	6	81	4	144	3.75	91.76
Roseworthy	SA	Uni Adelaide	93	5	84	4	144	4.11	79.66
Tamworth	NSW	NSWDPI	96	5	83	8	144	3.6	62.45
Lockhart	NSW	AGT	92	4	84	4	132	-	-

Connectivity of varieties between trials is an important point to consider in MET analyses. Due to seed limitations it was not possible to evaluate all varieties in all trials. Losing the trial at Lockhart resulted in four varieties being removed from the MET analysis: 1 ICARDA ('53ZDL16') and 3 CIMMYT ('97ZDG16', '77ZDG16', '143ZDG16') varieties. 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 2017. The diagonal represents the number of unique varieties in each Location.

	Kaniva	Kapunda	Narrabri	Northstar	Roseworthy	Tamworth	Lockhart
Kaniva	88						
Kapunda	79	94					
Narrabri	75	81	92				
Northstar	76	79	85	91			
Roseworthy	82	86	79	78	93		
Tamworth	77	84	85	82	83	96	
Lockhart	80	79	82	84	80	80	92

The experimental design accommodated this imbalance through the partial replication (p-rep) paradigm of Cullis et al. (2006). These experimental designs were generated by BBAGI staff in 2017 (Pirathiban & Mathews, 2017) using the `od` software (Butler, 2016) in `R` (R Development Core Team, 2015) for seven locations. A total of six trials were included in the MET analysis for Durum Wheat as the Lockhart trial was not harvested. The experimental design contained two replicate blocks (`Block`) in either the row (`Row`) or column (`Range`) direction and the experimental unit (EU) is the intersection between the `Rows` and `Ranges`, i.e. the `Plot`.

4 Statistical Analysis

A one-stage multi-environment trial (MET) analysis was conducted for the six trials with available yield plot data. Data for all six trials, denoted in the analysis as `Experiments`, were collated and included in the MET analysis. For this analysis we model the spatial trends at each `Experiment` 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: `Block, Row, Range, Plot`. The blocking structure for the MET analysis is then

`Experiment/Block/Plot`

which, following Wilkinson & Rogers (1973) expands to

`Experiment + Experiment:Block + Experiment:Block:Plot`.

The final term (`Experiment:Block:Plot`) 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 VEI 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` × `Experiment`

in order to model the VEI.

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 VEI, hence the final mixed model formula is

`fixed=` ∼ `1 + Experiment`

`random=` ∼ `Variety:Experiment + Experiment:Block + Experiment:Block:Plot`

where `Experiment:Block:Plot` represents the residual variation. Spatial variation at each experiment was accounted for by using the separable autoregressive spatial structure

(AR1 × AR1) to model the residual variation within each site (Gilmour et al., 1997) except at Narrabri. A separable identical error variance (ID × ID) was used for the site Narrabri as the AR1 × AR1 structure yielded small negative correlations in both directions: `Range` and `Row`.

4.2 Analysis

The MET analysis was carried out in `R` (R Development Core Team, 2015) using `ASReml-R` (Butler et al., 2015). There were no covariates reported for these trials and so the analysis commenced by identifying any outliers and inspecting whether they should be retained or removed from the analysis. 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. This allows us to investiagte the spatial trends at each `Experiment` and fit trends that are significant for each individual `Experiment`. The spatial terms fitted to the final model are given in Table 3 below.

Table 3: Final Spatial Models fitted to each Experiment: 1 = fitted, 0 = not fitted, aa = AR1 × AR1, ii = ID × ID.

Experiment	linear Row	linear Range	random Row	random Range	Block	Residual
Kaniva	0	0	0	1	1	aa
Kapunda	0	0	0	1	1	aa
Narrabri	0	1	1	1	1	ii
North Star	0	1	0	1	1	aa
Roseworthy	0	0	0	0	1	aa
Tamworth	1	0	1	0	1	aa

Table 4: Residual maximum likelihood (REML) loglikelihood and percentage variance accounted for (%vaf) the models fitted to CAIGE Durum Wheat MET dataset 2017.

Model	REML Log likelihood	%vaf
DIAG	162.49	-
FA1	182.64	57
FA2	189.34	67
FA3	193.54	90

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 $p = 6$ 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 ($p(p + 1)/2 = 21$). For the FA3 model, there are $pk + p - k(k - 1)/2 = 21$ parameters to estimate. Table 4 shows

the loglikelhood and percent variance explained from each model fitted and Table 1 shows the percent variance accounted for (%vaf) at each `Experiment` in the FA3 model.

The final `ASReml-R` call was

```
asr.rr3ar <- asreml(Yield ~ Experiment +
                      at(Experiment, mt$Experiment$lrow):lin(Row) +
                      at(Experiment, mt$Experiment$lrange):lin(Range),
                    random = ~ rr(Experiment, 3):Variety +
                      diag(Experiment):Variety +
                      at(Experiment, mt$Experiment$blk):Block +
                      at(Experiment, mt$Experiment$rrow):Row +
                      at(Experiment, mt$Experiment$rrange):Range,
                    residual = ~ dsum(~ar1(Range):ar1(Row)| Experiment,
                        levels = mt$Experiment$resid$aa) +
                      dsum(~id(Range):id(Row)| Experiment,
                        levels = mt$Experiment$resid$ii),
                    na.action = na.method(y='include', x='include'),
                      data=metdata, 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 `Experiment` names at which the term will be fitted. For example, `mt$lrow` contains the `Experiment` name "Tamworth" as a linear trend at Tamworth in the `Row` 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 from the analysis with all six `Experiments` is shown in Figure 1. Kaniva and Kapunda, sites in Victoria and South Australia are highly correlated, indicating that the variety ranks were similar for these two trials. In contrast, Kapunda and Tamworth have no correlation and the variety rankings are not similar, indicating that with respect to yield these environments are dissimilar. It is clear from this heatmap that Narrabri had zero correlation with the remaining `Experiments` such as Kaniva, Kapunda and Tamworth and together with the low genetic variance for this `Experiment` will account for the low %vaf for this location. This suggests that this `Experiment` was compromised in some way, so that it shows potentially poor performance. The feedback from the researchers confirmed that this trial was not irrigated and thus shows the poor performance.

The results generated by this MET analysis included the common variety $\times$ environment effects (CVE effects, t/ha) for each `Variety` and each `Experiment` and a measure of the

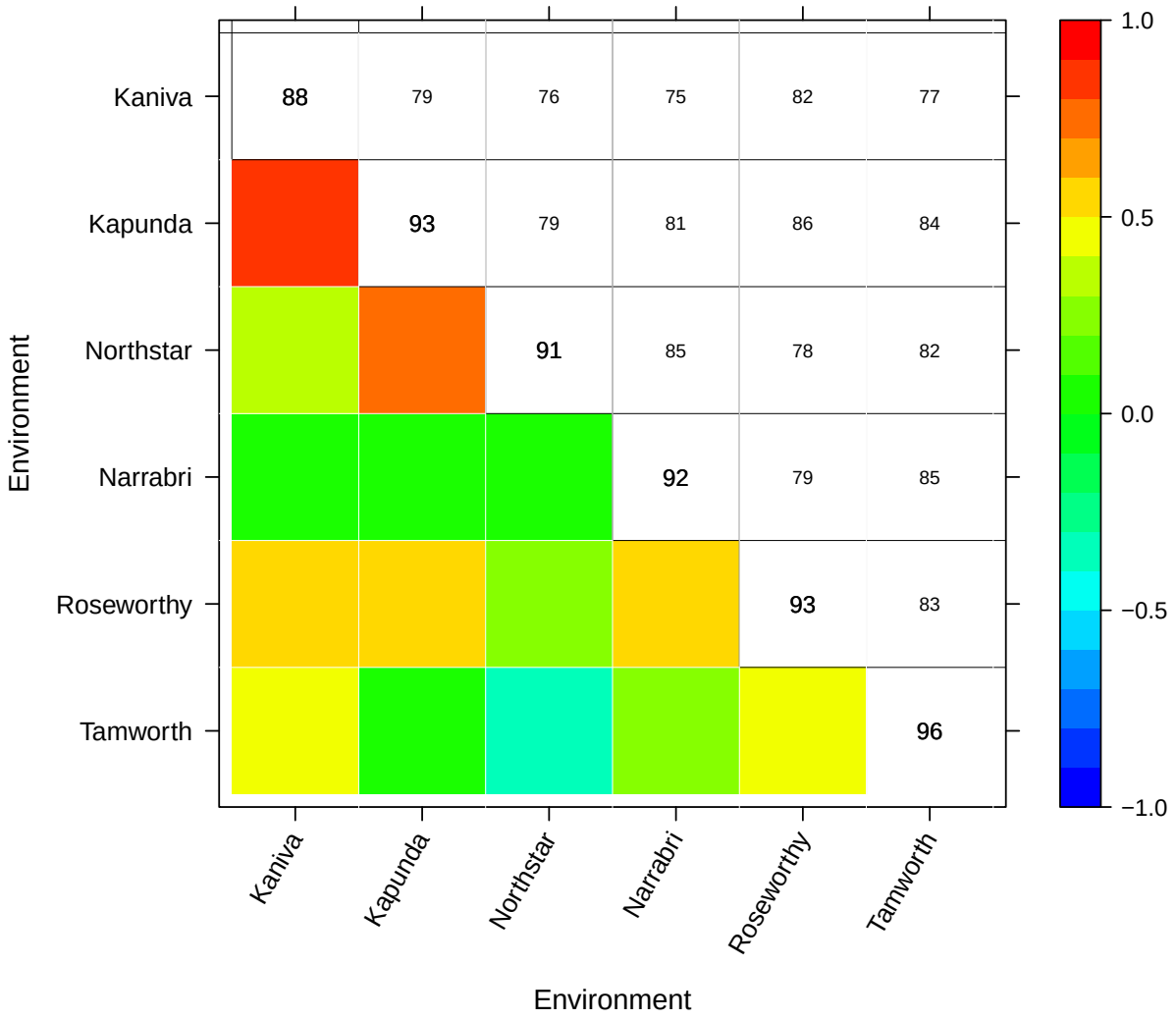

Figure 1: Heatmap of the REML estimates of the between environment genetic correlation matrix for all environments in the 2017 CAIGE Durum Wheat MET dataset (lower triangle), and number of varieties in common between a pair of environments (upper triangle). Axis labels are in dendrogram order.

accuracy of the estimation. The CVE effect 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 VEI. Further, caution should also be exercised when investigating those CVE effects with low accuracy (see the accompanying Excel workbook, worksheet *CVEeffects*). There are 44 varieties in Durum Wheat MET dataset that have accuracies less than 0.5 for some environments. There are 3 varieties, namely "71ZDG16", "75ZDG16" and "94ZDG16" with accuracies less than 0.05. These varieties were grown only in Narrabri so there is little information to accurately predict the variety response to each of the factors in the FA model as the responses of these varieties to all factors are based on one environment only.

Factor analytic selection tools (FAST) developed by Smith & Cullis (2018) are proposed to summarise the predictions in a concise yet informative manner. These tools can be derived from the final factor analytic model, which are measures of overall performance (OP) and stability (root mean square deviation, RMSD) across the environments in the MET dataset. As the (rotated) estimated loadings for the first factor in the final factor analytic model are positive (see the accompanying Excel workbook, worksheet *Trials*), the fitted values associated with the first factor represent non-crossover interaction and this characteristic can be exploited to obtain measures of both overall performance and stability. Figure 2 presents the OP plotted against the stability of the varieties in the Durum Wheat MET dataset. Varieties which are in the top left hand side of the plot can be interpreted as high performing and stable whilst those in the bottom right hand side are low performing and sensitive to environmental influences. It is clear that in this dataset there are ICARDA and CIMMYT varieties which demonstrate high performance and are more stable than the Australian commercial checks.

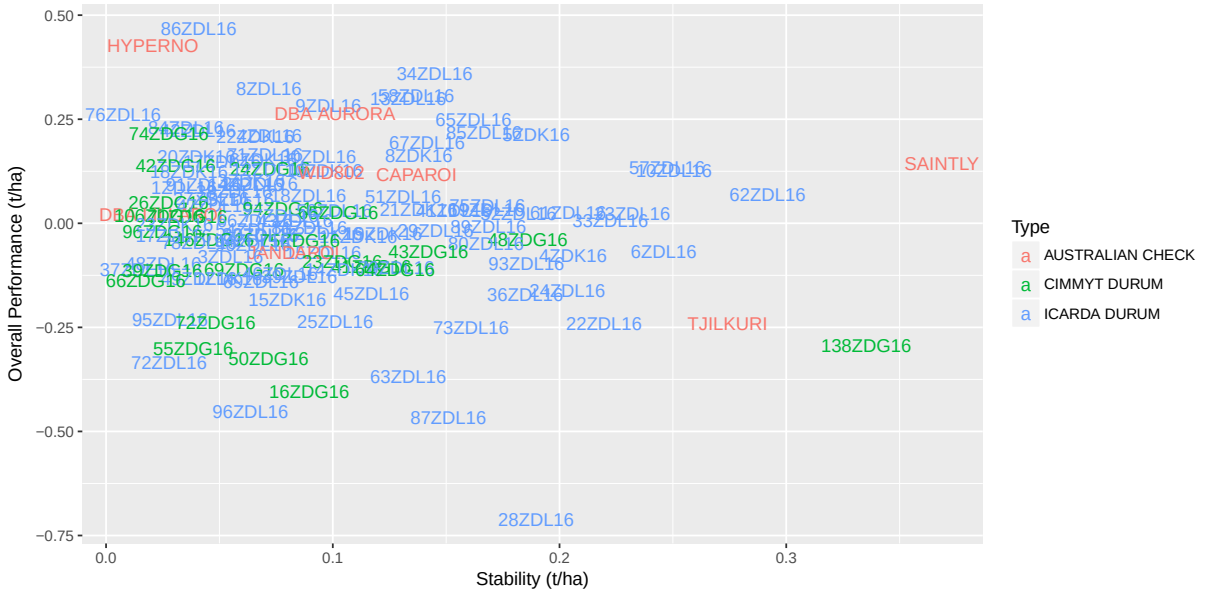

Figure 2: Overall performance vs Stability (RMSD) plot of the `Varieties` across all environments in the CAIGE Bread Wheat MET Analysis 2017

Another very useful and effective interpretation tool to display the CVE effects has been developed by Smith et al. (2015) called a production value (PV)-Plus plot. This plot is commonly used with the National Variety Trial (NVT) system. Six top performing non-Australian `Varieties` together with the six top performing Australian check `Varieties` at all six `Experiments` were shown in the PV-Plus plots, see Figure 3. The horizontal dashed line represents the average yielding variety. A positive production value indicates that the `Variety` is expected to yield higher than the average and a negative production value indicates that the `Variety` is expected to yield lower than the average and a production value of zero indicates that the `Variety` is expected to yield on average. The CIMMYT lines were under-represented in this MET dataset due to lack of seed and were not selected for presentation on the PV-Plus plots. The PV-Plus results are available in an interactive APP, https://nvtsagi.shinyapps.io/caigepvplus/, which allows the user to select the lines of interest to them for viewing as provided in the examples below.

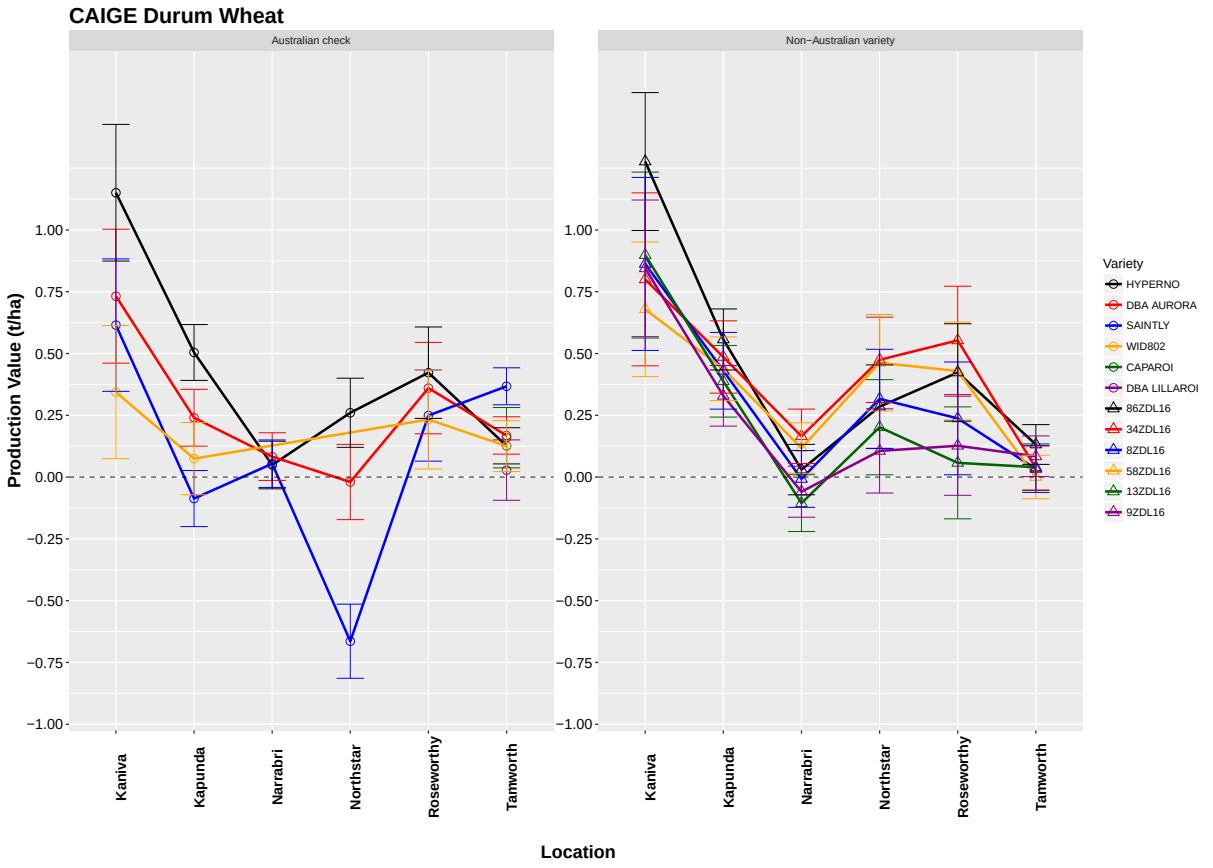

Figure 3: PV-Plus plot of 6 top performing `Varieties` from Australian checks and non-Australian varieties across all `Experiments` in the CAIGE Durum Wheat MET Analysis 2017. Types of the varieties can be distinguished by the shape of the points as shown in the legend (circle: CHECK, triangle: ICARDA). The left plot contains all Australian checks and the right plot contains the 6 top performing non-Australian varieties. Note, CIMMYT varieties are under-represented in this MET dataset due to lack of seed, thus there is no sufficient data to include them on this static plot. Similar plot containing CIMMYT varieties can be viewed using the interactive PV-Plus APP.

5 File Management

Analysis was carried out by both Jesse Rand and Ramethaa Pirathiban under the supervision of Ky Mathews. Raw data files, analysis files, R script files and results files are located on the hard drive of Ramethaa Pirathiban's computer located in */home/ramethaa/Projects/CAIGE/CAIGE2017/DurumWheat/Analysis* and backed up to external hard drive.

An Excel workbook *CAIGEdw-2017-METresults.xlsx* containing all results was sent to Dr Richard Threthowan and associates of CAIGE on 16th January, 2017.

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.

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.

Pirathiban, R. & Mathews, K. (2017). Experimental designs for CAIGE durum 2017. Technical report.

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. & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *submitted for Euphytica* .

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. *Applied Statistics* **22**, 392–399.