Bioinformatics and Biometrics for the Australian Grains Industry Technical Report Series

CAIGE Bread Wheat MET Analysis, 2018

*** Stakeholders Report ***

Daniel Tolhurst & Ky Mathews
National Institute for Applied Statistics Research Australia
School of Mathematics and Applied Statistics
University of Wollongong
email: kmathews@uow.edu.au

January 31, 2019

This report describes the yield analysis of the CIMMYT Australia ICARDA Germplasm Evaluation (CAIGE) Bread Wheat MET dataset for 2018. Additional analyses were conducted on phenology traits, however these were performed on the individual-environment level so we restrict this report to the complex trait of yield. The report is structured as follows: an Executive Summary is provided in Section 1 presenting the key results, an Analysis Summary in Section 2 outlining the methods and results in more detail and a Glossary of terms to which the reader can refer in Section 3.

1 Executive Summary

The MET analysis for **Bread Wheat** was conducted on the 2018 dataset. There were two plots at each of Edgeroi and Horsham where the entry was unknown, so these were considered filler plots and removed from the genetic analysis. The final, cleaned dataset contained **8 environments (with a single CAIGE trial in each), 315 varieties and 2232 plots (records)**. The number of varieties (partitioned into CIMMYT, ICARDA, ISR, SBP, checks and total), plots, p-rep level and mean yield (t/ha) are presented in Table 1. The geographic location of these trials are presented in Figure 1.

Table 1: Summary of the CAIGE Bread Wheat MET dataset for 2018. *Total number of unique varieties across environments. +Does not include fillers.

		Varieties+								Mean
State	Environment	CIMMYT	ICARDA	ISR	SBP	Checks	Total	Records+	p-rep	Yield
NSW	1 Edgeroi	99	55	17	11	14	196	238	0.19	3.29
	2 Narrabri	130	133	20	12	14	309	552	0.70	5.71
	3 Junee	108	57	18	12	13	208	240	0.13	1.61
VIC	4 Horsham	97	60	11	12	14	194	238	0.16	2.59
SA	5 Roseworthy	101	61	18	11	15	206	240	0.16	2.71
	6 Balaklava	101	64	17	12	14	208	240	0.15	0.78
WA	7 Toodyay	107	52	17	12	15	203	240	0.18	4.54
	8 Dandaragan	95	55	18	12	14	194	240	0.21	3.74
Total	-	130*	134*	20*	12*	19*	315*	2228	-	3.48

The connectivity across environments is shown in Figure 2. The axis labels of this figure are arranged according to a highway one order (i.e. environments 1 through 8). **The variety connectivity between all environments is excellent** (≥ 147).

For this dataset a factor analytic model of order four (FA4) was fitted, corresponding to a total genetic variance accounted for (VAF) of 78.2%. There were two trials where the VAF was less than 50%, namely `Edgeroi` (43.5%) and `Toodyay` (33.3%); see Section 2.4 for further details.

The between environment genetic correlations for all environments is presented as a heatmap in Figures 3. The axis labels of this figure are arranged according to the dendrogram. **This heatmap shows that there is substantial cross-over VEI present in Bread Wheat.**

1 Executive Summary

The empirical best linear unbiased predictions (EBLUPs) of the common variety by environment (CVE) effects, environment loadings, variety scores, between environment genetic correlation matrix and environment variety connectivity matrix are provided in the accompanying excel spreadsheet, CAIGE_BreadWheat_2018-METresults. In addition, these results are available in the Production Value Plus (PV-PLUS) APP, which is the recommended way to interrogate the CAIGE MET analysis results. **Please contact Dr. Ky Mathews if you would like access to this APP.**

1 Executive Summary

Figure 1: Geographic locations of envrionments in CAIGE Bread Wheat 2018, namely 1- Edgeroi, 2- Narrabri, 3- Junee, 4- Horsham, 5- Roseworthy, 6- Balaklava, 7- Toodyay and 8- Dandaragan.

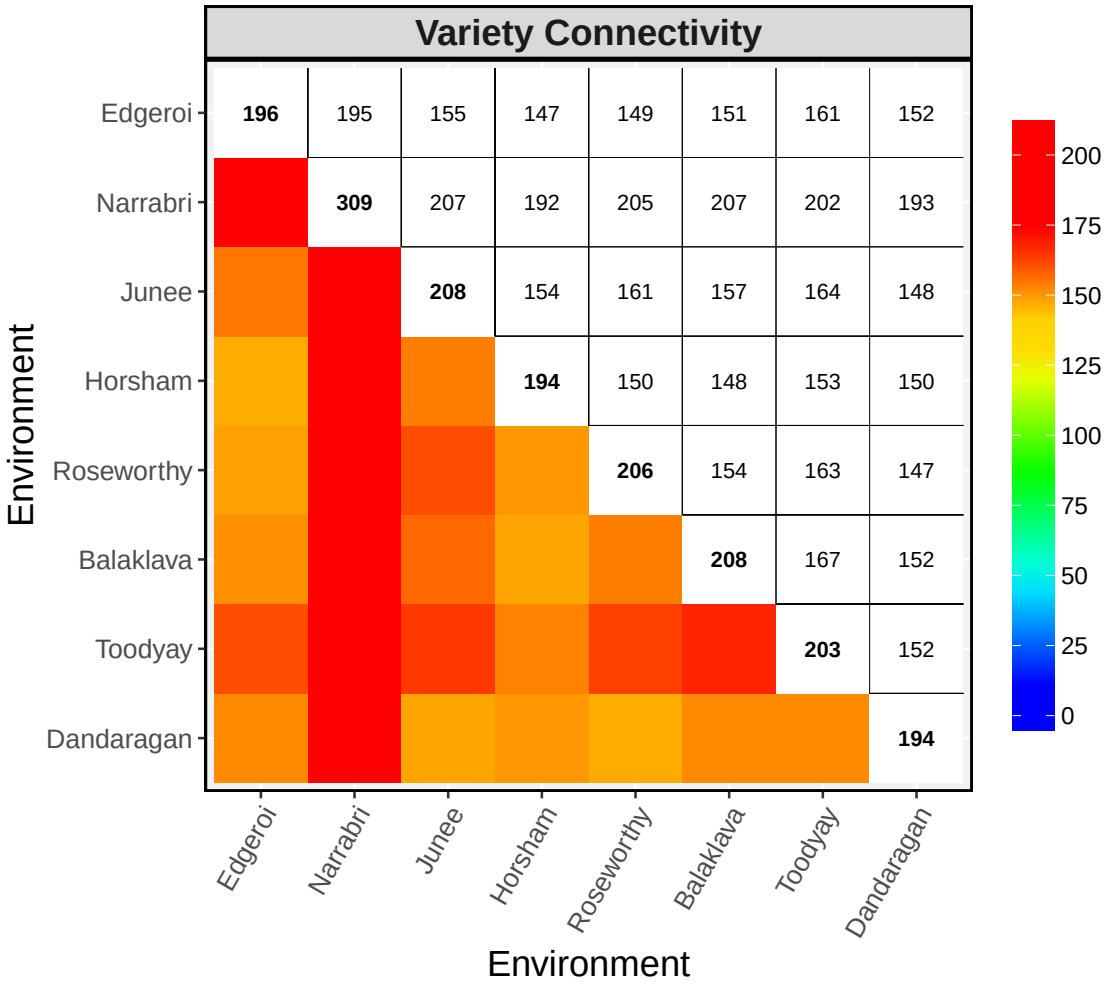


Figure 2: Variety connectivity between all 8 environments in CAIGE Bread Wheat 2018. The axis labels are arranged according to a highway one order.

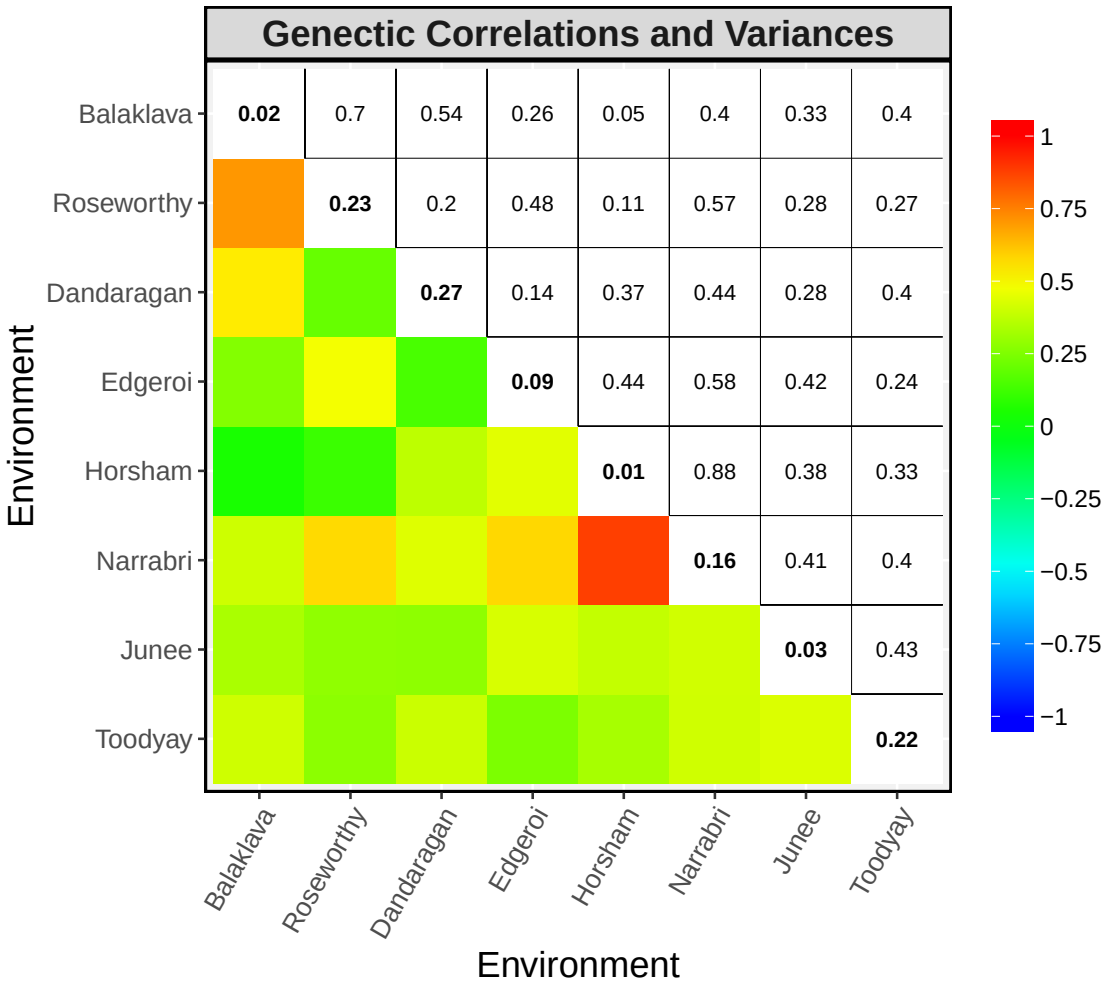


Figure 3: Genetic correlations between all 8 environments in CAIGE Bread Wheat 2018. The genetic variance of individual environments are provided on the diagonal. The axis labels are ordered according to the dendrogram.

2 Analysis Summary

2.1 Introduction

CAIGE is a GRDC funded project to evaluate Bread Wheat, Durum Wheat and Barley germplasm developed by the International Maize and Wheat Improvement Centre (CIMMYT) and the International Centre 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 trialled 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.

Trials were designed as *p*-rep trials by BBAGI staff for eight locations across four states of Australia; NSW (Edgeroi, Narrabri and Junee), VIC (Horsham), SA (Roseworthy and Balaklava) and WA (Toodyay and Dandaragan).

This report describes the yield analysis of the CAIGE Bread Wheat Multi-Environment Trial (MET) analysis for 2018. The data are available via the CAIGE website (http://www.caigeproject.org.au/).

The fully efficient one-stage Factor Analytic (FA) approach of Smith et al. (2001) and Gogel et al. (2018) has been adopted for the analysis of CAIGE yield MET datasets. The results of this analysis are reported according to Smith et al. (2015) and Smith & Cullis (2018).

2.2 Description of Data

The final dataset contained 8 environments (trials) and 315 varieties resulting in 2232 records. Summaries of the final dataset are provided by environment in Table 1.

2.3 Statistical Methods

For full documentation on the method and theory of the FA approach used in the one-stage CAIGE MET analysis, readers are referred to Gogel et al. (2018), although a brief summary is provided here. The MET analysis involved two steps: (i) individual trial analyses were performed in order to determine the most adequate spatial models that accounted for the presence of global and extraneous trends exhibited in the field. In this step variety effects were fitted as random and outliers identified and removed from all subsequent analyses. The spatial models fitted in all analyses are presented in Table 2. All models fitted here were saved and stored for inclusion during the following step. An independent structure was then fitted to the between-environment genetic variance matrix to examine whether any environments had zero genetic variance. Next, (ii) FA structures

Table 2: Summary of design, linear and random terms fitted to the 2018 Bread Wheat MET dataset, including a two-dimensional separable autoregressive process for the residual variance.

		Linear		Random		
Environment	Block	Column	Row	Column	Row	Resid
Edgeroi	✓	✓	-	✓	-	AR1⊗AR1
Narrabri	✓	✓	-	✓	-	AR1⊗AR1
Junee	✓	-	-	✓	-	AR1⊗AR1
Horsham	✓	-	-	✓	✓	AR1⊗AR1
Roseworthy	✓	-	-	✓	-	AR1⊗AR1
Balaklava	✓	✓	-	✓	-	AR1⊗AR1
Toodyay	✓	✓	-	✓	-	AR1⊗AR1
Dandaragan	✓	-	-	✓	✓	AR1⊗AR1

of order 1 through k were fitted to the between environment genetic variance-covariance matrix, which models heterogeneous environment variances and heterogeneous covariance between environments. The order of the FA model, i.e. k is increased incrementally in order to achieve a satisfactory level of the total VE variance accounted for by the k common factors.

All analyses were conducted in the R software environment R Core Team (2019) with the ASReml-R (v4.0) package (Butler et al., 2017) used for fitting the mixed models.

2.4 Results

2.4.1 Modelling and percentage of variance accounted for

The final model fitted was an **FA4**, corresponding to a total genetic variance accounted for **(VAF) of 78.2%**. Table 4 presents the individual %VAFs for each environment and FA model.

The variety predictions output from the FA MET analysis are empirical best linear unbiased predictions (EBLUPs) of the common variety by environment (CVE) effects as they correspond to that part of the variety by environment (VE) effects associated with the common factors in the FA model. By definition, the CVE effects for one environment are correlated with the CVE effects in at least one other environment. Whilst the use of the CVE effects have numerous advantages, care must be exercised when examining those corresponding to environments with a low %VAF. Such environments are generally uncorrelated with others and if these are of particular interest (and considered reproducible) then other prediction methods may be more appropriate. Recall there were two trials where the VAF was less than 50%, namely `Edgeroi` (43.5%) and `Toodyay` (33.3%).

Care must also be exercised when examining those CVE effects with low accuracy. There were three varieties with an accuracy less than 0.2 for some trials, `BARLEY`, `TINCURRIN` and

Table 3: Summary of models fitted to the 2018 Bread Wheat MET dataset. Included is the percentage of genetic variance explained by the common factors in terms of the number of environments below 50%, above 80% and overall mean.

	Genetic			%VAF		
Model	Params	LogLik	AIC	<50	>80	
diag	8	729.1	-1362.3	-	-	-
FA1	16	829.1	-1546.2	5	1	47.0
FA2	23	844.5	-1562.9	3	2	58.0
FA3	29	854.4	-1570.8	2	3	67.4
FA4	**34**	**859.7**	**-1571.4**	**2**	**5**	**78.2**

Table 4: Summary of the percentage of variance explained for (%VAF) each factor analytic model along with the total genetic variance (VAR) for the final model.

	%VAF				
Environment	FA1	FA2	FA3	FA4	VAR
Edgeroi	33.9	50.2	50.1	43.5	0.09
Narrabri	59.2	74.7	74.5	100.0	0.16
Junee	29.2	29.7	32.5	100.0	0.03
Horsham	100.0	100.0	100.0	100.0	0.01
Roseworthy	53.2	57.0	92.0	100.0	0.23
Balaklava	46.8	100.0	100.0	85.4	0.02
Toodyay	28.9	25.4	34.4	33.3	0.22
Dandaragan	24.8	27.3	55.9	63.5	0.27

`WAXWING`. These varieties were tested in only one environment each (`Horsham`, `Roseworthy` and `Balaklava`, respectively).

2.4.2 Predictions and Correlations

Overall variety performance and stability

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 Environments), 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 4 presents the OP plotted against the stability of the varieties in the Bread 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 stability across the sampled environments.

PV-PLUS system

Another 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 in the National Variety Trial (NVT) system.

The production values (PVs), EBLUPS of the CVE effects, reported in the accompanying spreadsheet, can be interpreted as positive or negative differences relative to a baseline, which reflects the expected average yield of all (315) varieties in the current CAIGE Bread Wheat MET dataset, if grown in that particular environment. Consequently, varieties may be viewed as having expected yields that are below (PV <0), equal (PV = 0) or above (PV >0) the baseline for a particular environment. Note that dashed lines are used in the PVPLUS-APP to pass through those environments in which the accuracy is less than 0.8. This default can be altered via the user. The standard error bars are plus/minus the square root of the prediction error variance (PEV) for the EBLUP of the CVE effect. To assist with interpretation of the VEI, the heatmap of the REML estimate of the between environment genetic correlation matrix is also provided.

The following PV-Plus plot is an example of the interactive APP available to breeders and CAIGE managers. This APP is password protected. Please contact Dr. Ky Mathews if you wish to use the PV-PLUS APP to interrogate the data for your own purposes.

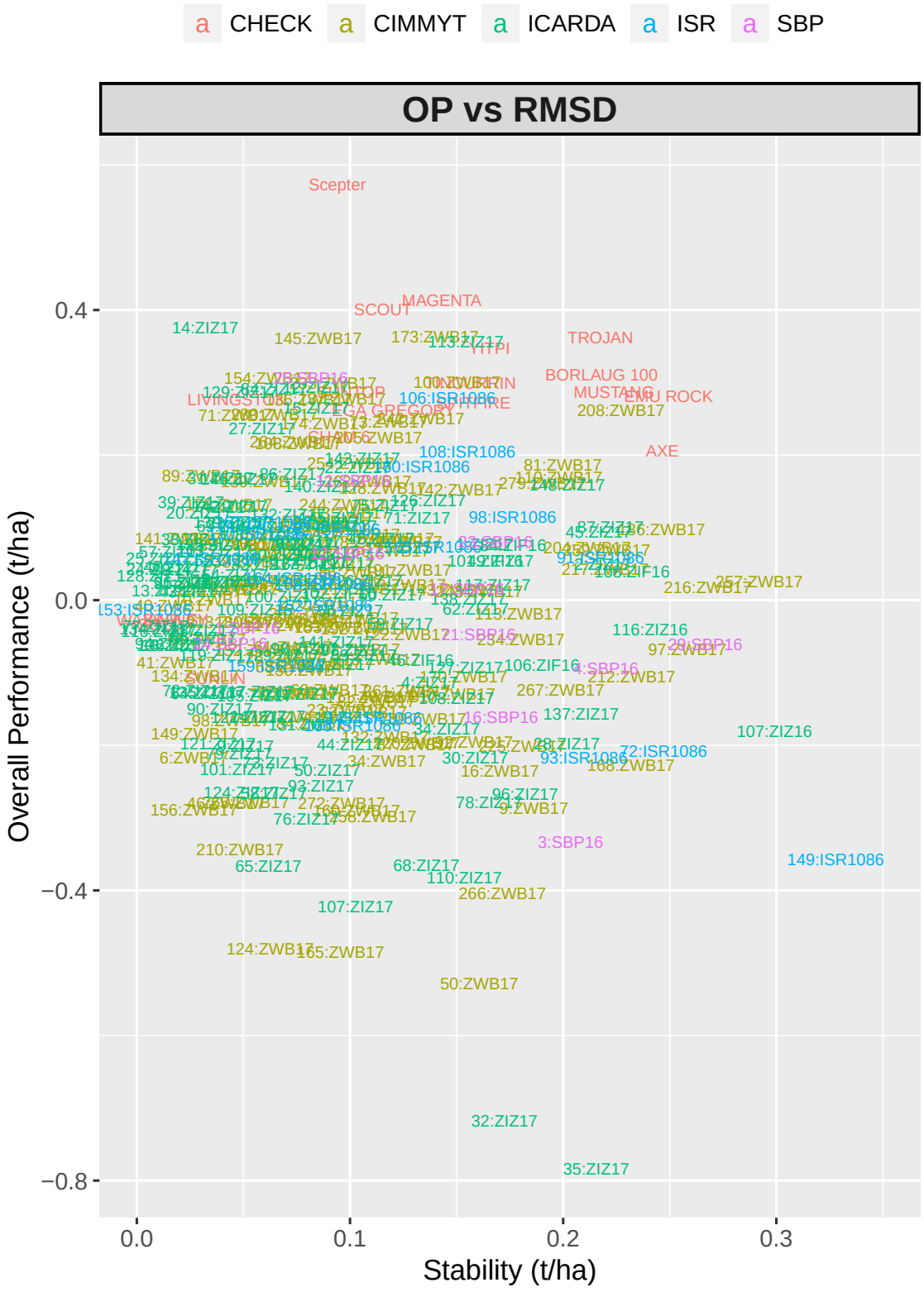
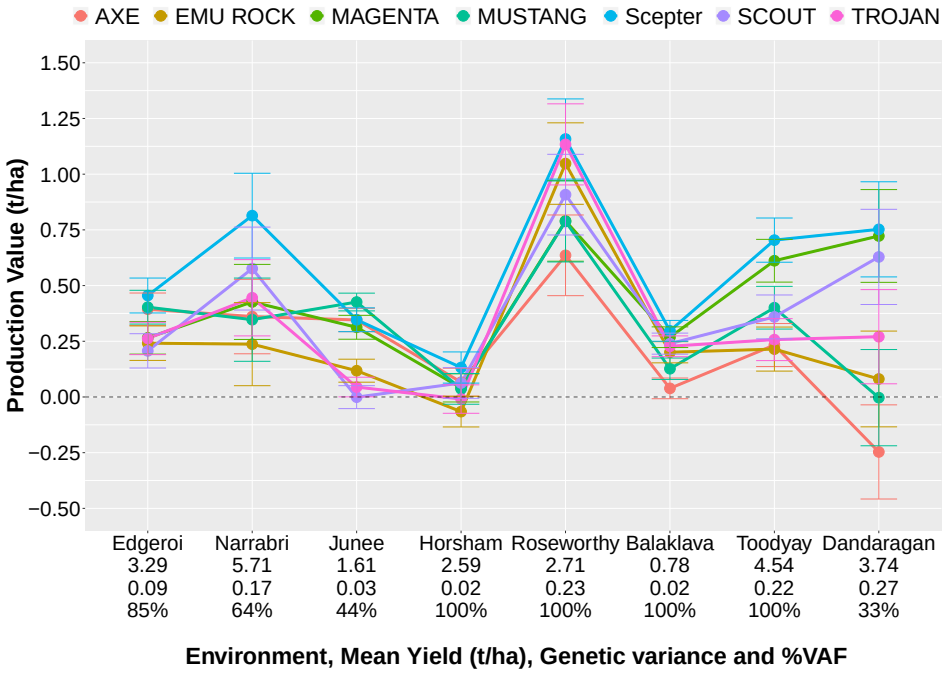
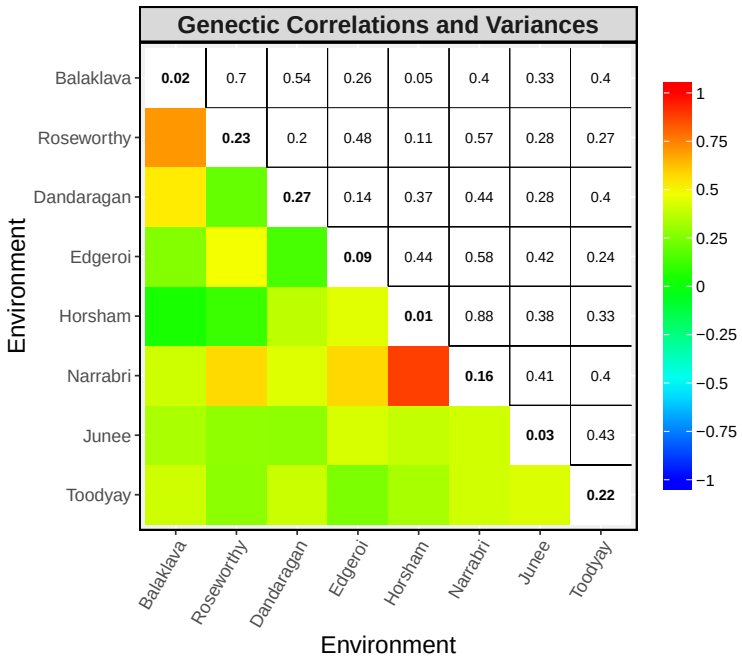


Figure 4: Overall performance (OP) vs stability measure (RMSD) for yield displaying all 315 varieties in CAIGE Bread Wheat 2018.



(a) PV-PLUS graphical representation of the MET results for Bread Wheat for selected varieties



(b) Heatmap representation of the genetic correlation matrix for Bread Wheat.

Figure 5: PV-PLUS graphic and heatmap of the residual maximum likelihood (REML) estimate of the total genetic correlation matrix (with genetic variances on the diagonal) for environments in CAIGE Bread Wheat 2018.

3 Glossary

Accuracy: the model-based correlation between the true values and the EBLUPs of the CVE effects.

Between environment genetic variance matrix, $\boldsymbol{G_e}$: the $t \times t$ matrix (where t is the number of environments) that reflects variance and covariance assumptions for VE effects across environments. The diagonal elements of the matrix are the variances of the VE effects for individual environments; the off-diagonal elements are the covariances between VE effects in different environments. This is often converted to a correlation matrix, which will be called the Between environment genetic correlation matrix.

Breeder trial: a comparative variety trial managed by a public breeding program. Each trial comprises a single randomisation of varieties to a set of field plots.

Common Variety by Environment (CVE) effect: the component of the VE effect that corresponds to the linear combinations of the common factors in the FA model. By definition the CVE effects for one environment are correlated with CVE effects in at least one other environment. Note that VE effects are obtained as the sum of CVE and SVE effects.

Connectivity: Number of unique varieties in common between pairs of environments and represented by a $t \times t$ matrix (where t is the number of environments).

Dendrogram order: describes the ordering of environments in regards to a clustering algorithm on the Between environment genetic correlation matrix. We use the agglomerative (nested) hierarchical algorithm implemented in the `agnes` function from the R library `mclust`.

Diagonal model: a model that can be used in the MET analysis to provide a structure for the Between environment genetic variance matrix. The VE effects are assumed to be independent so that all covariances (off-diagonal elements of $\boldsymbol{G_e}$) are zero, with the result that $\boldsymbol{G_e}$ has a diagonal form. For t environments the diagonal model for $\boldsymbol{G_e}$ contains t (unknown) variance parameters.

Empirical Best Linear Unbiased Estimate (EBLUE): all fixed effects in the linear mixed model analysis are estimated using the method of best linear unbiased estimation, but with variance parameter estimates replaced with their REML estimates. Thence all fixed effect estimates are termed Empirical Best Linear Unbiased Estimates (EBLUEs).

Empirical Best Linear Unbiased Prediction (EBLUP): all random effects in the linear mixed model analysis are predicted using the method of best linear unbiased prediction, but with variance parameter estimates replaced with their REML estimates. Thence all random effect predictions are termed Empirical Best Linear Unbiased Predictions (EBLUPs).

3 Glossary

EMY: the mean yield of all field plots at an environment.

Environment loading: the linear combinations in the FA model comprise products of environment loadings and variety scores. For each common factor there is a loading for each environment and in the MET analysis these are (unknown) variance parameters. They may be interpreted as (unknown) environmental covariates.

Factor Analytic (FA) model: a model that can be used in the MET analysis to provide a structure for the Between environment genetic variance matrix. The aim is to account for the covariances of the V×E effects between environments (the off-diagonal elements of $\boldsymbol{G_e}$) in terms of a small number, k, of (unknown) common factors. The model is postulated in terms of the V×E effects as linear combinations of the common factors, plus an error term. For t environments and k factors the FA model for $\boldsymbol{G_e}$ contains $t(k+1) - k(k-1)/2$ (unknown) variance parameters.

Highway 1 order: describes the ordering of environments with respect to geographic location, commencing in QLD and roughly following Highway 1 in a clockwise direction around the Australian grain belt through NSW, Vic, SA and WA.

Multi-Environment Trial (MET): a series of comparative variety trials grown in different environments (typically indexed by year and geographic location). In general there is a single CAIGE trial in each environment, but in some cases there may be multiple trials.

CAIGE trial: a comparative variety trial managed by CAIGE. Each trial comprises a single randomisation of varieties to a set of field plots.

One-stage MET analysis: a linear mixed model analysis of a MET dataset that comprises individual plot yield data combined across all environments. The analysis requires the modelling of both the genetic effects (in particular the Variety by Environment effects), and non genetic effects for each trial.

Percentage variance accounted for (%VAF): the percentage of V×E variance for an environment that is accounted for by the common factors in an FA model. Note that 100-%VAF represents the percentage contribution of the specific variance for the environment.

Production Value Plus System (PV-PLUS): a system for the provision of grower information from the MET analysis. The information comprises production values (PVs) which are EBLUPs of CVEs. The PVs, together with a measure of accuracy, are displayed graphically for nominated sets of varieties and environments. For each environment they are expressed as deviations (t/ha) from the EBLUE of the mean parameter for the environment. **The PV-PLUS system will be made available through an online APP, please contact Dr. Ky Mathews to gain access.**

3 Glossary

Residual Maximum Likelihood (REML): the method of estimation for variance parameters in the linear mixed model analysis. Thence all variance parameter estimates are termed REML estimates.

Specific variance: the error term in the FA model is assumed to comprise independent effects that may have a different variance, called a specific variance, for each environment.

Specific Variety by Environment (SVE) effect: the component of the VE effect that corresponds to the error term in the FA model. By definition the SVE effects for one environment are uncorrelated with SVE effects in every other environment. Note that VE effects are obtained as the sum of CVE and SVE effects.

Unstructured model: a model that can be used in the MET analysis to provide a structure for the Between environment genetic variance matrix. This is a completely general form so that for t environments the unstructured model for $\boldsymbol{G_e}$ contains $t(t+1)/2$ (unknown) variance parameters.

Variety: an entry in a CAIGE trial.

Variety by Environment (VE) effect: (unknown) random genetic effect of a variety in an environment. It is expressed as a deviation (t/ha) from the (unknown) fixed mean parameter for the environment.

Variety score: the linear combinations in the FA model comprise products of environment loadings and variety scores. For each common factor there is a score for each variety and in the MET analysis these are (unknown) random effects. They may be interpreted as varietal responses (sensitivities) to the environment loadings.

References

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2017). ASReml-R Reference Manual Version 4.

Gogel, B. J., Smith, A. B., & Cullis, B. R. (2018). Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica* **214**, 44–64.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smith, A. B. & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* **214**, 143–161.

Smith, A. B., Cullis, B. R., & Gilmour, A. R. (2001). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics* **43**, 129–145.

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.