Genotype by Environment analysis using RAP

Bart-Jan van Rossum

2019-03-05

Genotype by Environment Analysis using the RAP package

The RAP package is developed as an easy to use package for analyzing data of plant breeding experiments with many options for plotting and reporting the results of the analyses.

The package has three main components:

- Modeling trial data for single trials and extracting results
- Genotype by Environment (GxE) analysis
- QTL analysis

This vignette deals with the GxE part of the package and describes in detail how to perform the different types of analysis that are available in the package.

The following types of analysis can be done using RAP:

- Best variance-covariance model
- AMMI Analysis
- GGE Analysis
- Finlay-Wilkinson Analysis
- Computation of mega environments
- Computation of stability measures

Data preparation

Just as for the analysis of single field trials, the input for GxE analysis in the RAP package is an object of class TD. For a detailed description on how to construct such an object see the vignette Modeling field trials using RAP. The TD object created in the final step of this vignette, `TDGxE`, will be used for the GxE analyses in the current vignette.

Best variance-covariance model

The function gxeVarComp fits models with different variance-covariance structures to the GxE data and determines the best model for the data. With the default settings for the function lme4 is used for fitting the models and only a compound symmetry model is fitted. When changing the modeling engine to asreml eight models with different variance-covariance structures are fitted:

- identity
- compound symmetry
- diagonal
- heterogeneous compound symmetry
- outside
- factor analytic with one factor
- factor analytic with two factors

- unstructured

The best model for the data is selected based on either the Akaike Information Criterion (AIC) or the Baysian Information Criterion (BIC). Which criterion is used is determined by the parameter `criterion` in the function `gxeVarComp`.

Using the `TDGxE` TD object created in the vignette Modeling field trials using RAP the function can be used as follows:

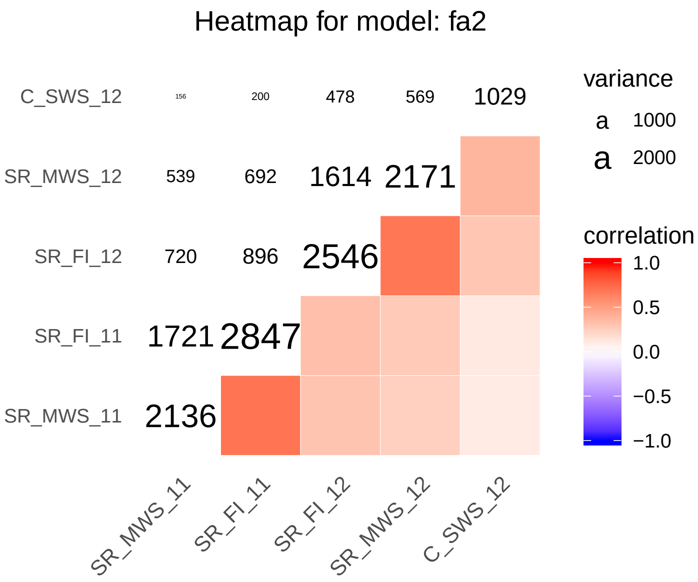
```
## Use lme4 for fitting the models - only compound symmetry.
geVC <- gxeVarComp(TD = TDGxE, trait = "BLUEs_GY")
summary(geVC)
#> Best model: cs, based on BIC.
#>          AIC      BIC Deviance NParameters
#> cs 31188.53 31199.65 31184.53           2
```

```
## Use asreml for fitting the models - eight models fitted.
## Use AIC as criterion for determining the best model.
if (requireNamespace("asreml", quietly = TRUE)) {
  geVC2 <- gxeVarComp(TD = TDGxE, trait = "BLUEs_GY", engine = "asreml",
                      criterion = "AIC")
  summary(geVC2)
}
#> Best model: fa2, based on AIC.
#>                   AIC      BIC Deviance NParameters
#> fa2          27310.09 27384.80 27282.09          14
#> unstructured 27311.58 27391.63 27281.58          15
#> fa           27518.37 27571.74 27498.37          10
#> outside      27548.83 27580.86 27536.83           6
#> hcs          27639.91 27671.93 27627.91           6
#> cs           27672.67 27683.34 27668.67           2
#> diagonal     27904.72 27931.40 27894.72           5
#> identity     28001.40 28006.74 27999.40           1
```

As becomes clear from the summary, the best model based on AIC is the model with a factor analytic variance-covariance structure with two factors. Note that for the both factor analytic models to be fitted a minimum of five environments are needed. If the data contains less environments, those two models are skipped.

A heat map of the correlation between the environments based on the best fitted model can be plotted.

```
if (requireNamespace("asreml", quietly = TRUE)) {
  plot(geVC2)
}
```



The upper left of the plot displays the variance between environments. Larger values are displayed in a bigger font. In the lower right of the plot correlations between environments are shown. A dark red color indicates a strong positive correlation between environments, a dark blue a strong negative correlation. Environments are clustered by their correlations and ordered according to the results of the clustering.

A pdf report containing the most important results of the analysis can be made using the `report` function.

```
report(geVC2, outfile = "./myReports/varCompReport.pdf")
```

AMMI Analysis

The Additive Main Effects and Multiplicative Interaction (AMMI) model fits a model which involves the Additive Main effects (i.e. genotype and trial) along with the Multiplicative Interaction effects. Then a principal component analysis is done on the residuals (multiplicative interaction). This results in an interaction characterized by Interaction Principal Components (IPCA) enabling simultaneous plotting of genotypes and trials.

The AMMI analysis can be performed with the RAP package using the function gxeAmmi.

```
## Run gxeAmmi with default settings.
geAm <- gxeAmmi(TD = TDGxE, trait = "BLUEs_GY")
summary(geAm)
#> Principal components
#> ====================
#>                              PC1       PC2
#> Standard deviation     1030.58353 715.42427
#> Proportion of Variance    0.49975   0.24083
#> Cumulative Proportion     0.49975   0.74058
#>
#> Anova
#> =====
#>                        Df     Sum Sq    Mean Sq  F value    Pr(>F)
#> Genotype              383  760625194    1985967   3.7378 < 2.2e-16 ***
#> Environment             4 8310680494 2077670123 3910.4240 < 2.2e-16 ***
#> Interactions         1532  813975825     531316
#> PC1                   386  406785224    1053848    3.8030 < 2.2e-16 ***
#> PC2                   384  196031610     510499    1.8422  6.41e-13 ***
#> Residuals             762  211158990     277112
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Environment scores
#> ==================
#>                  PC1         PC2
#> C_SWS_12   0.2278314 -0.81879355
#> SR_FI_11  -0.5946811  0.16138788
#> SR_FI_12   0.4069946  0.53407849
#> SR_MWS_11 -0.4826721 -0.01143373
#> SR_MWS_12  0.4425272  0.13476091
```

With the default settings in the principal components analysis a maximum of two principal components are used. This can be changed using the `nPC` parameter in the function. The number of principal components can never be larger than the number of environments and the number of genotypes in the data. By specifying `nPC = NULL` the algorithm will determine the number of principal components by a method of forward selection.

```
## Run gxeAmmi. Algorithm determines number of principal components.
geAm2 <- gxeAmmi(TD = TDGxE, trait = "BLUEs_GY", nPC = NULL)
summary(geAm2)
#> Principal components
#> ====================
#>                              PC1       PC2
#> Standard deviation     1030.58353 715.42427
#> Proportion of Variance    0.49975   0.24083
#> Cumulative Proportion     0.49975   0.74058
#>
#> Anova
#> =====
#>                        Df     Sum Sq    Mean Sq   F value    Pr(>F)
#> Genotype              383  760625194    1985967    3.7378 < 2.2e-16 ***
#> Environment             4 8310680494 2077670123 3910.4240 < 2.2e-16 ***
#> Interactions         1532  813975825     531316
#> PC1                   386  406785224    1053848    3.8030 < 2.2e-16 ***
#> PC2                   384  196031610     510499    1.8422  6.41e-13 ***
#> Residuals             762  211158990     277112
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Environment scores
#> ==================
#>                  PC1         PC2
#> C_SWS_12   0.2278314 -0.81879355
#> SR_FI_11  -0.5946811  0.16138788
#> SR_FI_12   0.4069946  0.53407849
#> SR_MWS_11 -0.4826721 -0.01143373
#> SR_MWS_12  0.4425272  0.13476091
```

It is possible to exclude certain genotypes, e.g. outliers, from the analysis using the parameter `excludeGeno`.

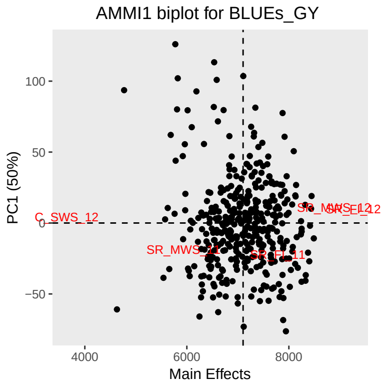
```
## Run gxeAmmi with three principal components.
## Exclude genotypes G278 and G279.
```

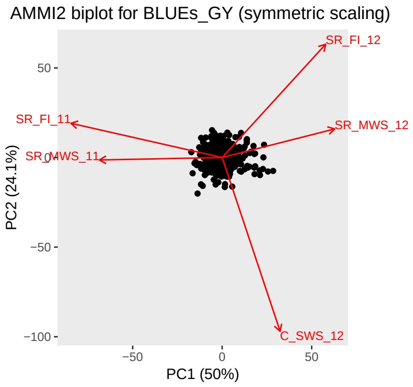
```
geAm3 <- gxeAmmi(TD = TDGxE, trait = "BLUEs_GY", nPC = 3,
                 excludeGeno = c("G278", "G279"))
summary(geAm3)
#> Principal components
#> ====================
#>                               PC1       PC2       PC3
#> Standard deviation     1031.82428 715.68094 536.71735
#> Proportion of Variance    0.49982   0.24046   0.13524
#> Cumulative Proportion     0.49982   0.74029   0.87552
#>
#> Anova
#> =====
#>                     Df     Sum Sq    Mean Sq   F value    Pr(>F)
#> Genotype           381  754528274    1980389    3.7189 < 2.2e-16 ***
#> Environment          4 8264149230 2066037308 3879.7532 < 2.2e-16 ***
#> Interactions      1524  811556993     532518
#> PC1                384  405635969    1056344    3.9527 < 2.2e-16 ***
#> PC2                382  195147899     510858    1.9115  2.01e-10 ***
#> PC3                380  109752961     288824    1.0807    0.2251
#> Residuals          378  101020163     267249
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Environment scores
#> ==================
#>                    PC1        PC2        PC3
#> C_SWS_12    0.2297733 -0.8196600  0.0736193
#> SR_FI_11   -0.5951463  0.1602307  0.6434218
#> SR_FI_12    0.4048303  0.5315277 -0.1391278
#> SR_MWS_11  -0.4825132 -0.0128036 -0.7329387
#> SR_MWS_12   0.4430560  0.1407052  0.1550254
```

If the data contains a column year, it is possible to perform the AMMI analysis per year in the data. This can be done by specifying the parameter `byYear = TRUE`.

The results of an AMMI analysis can be displayed in a biplot. Two types of biplot are available. "AMMI1" plots the main effects against the first principal component. "AMMI2" plots the first against the second principal component.

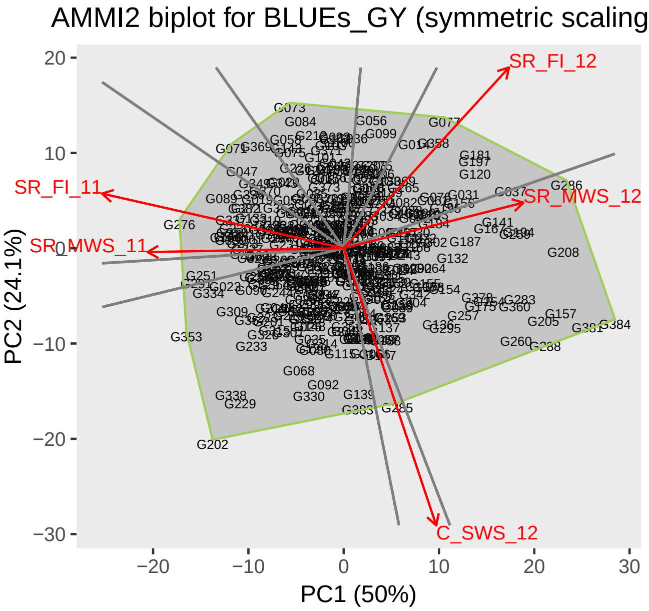
```
## Create an AMMI1 and AMMI2 biplot.
plot(geAm, scale = 0.5, plotType = "AMMI1")
plot(geAm, scale = 0.5, plotType = "AMMI2")
```





The AMMI plot function has many options to customize the plot. It is possible to plot different principal components on the axis using `primAxis` and `secAxis`. Genotypes can be grouped and colored by a variable in the data using `groupBy` and `colorBy`. A convex hull can be plotted around the genotypes in an AMMI2 biplot with lines from the origin perpendicular to the edges of the hull. This is usefull for identifying mega environments. Genotypes can be left out of the plot completely by setting `plotGeno = FALSE` and similarly `plotEnv = FALSE` assures no environments are plotted. For displaying genotypes by their names instead of points, use `sizeGeno` with a size larger than zero. `envFactor` can be used to blow up the environmental scores in the plot. A value for `envFactor` between zero and one effectively blows up the genotypic scores. Some more options are available for sizing and coloring. Run `help(plot.AMMI)` for full details.

```
## Create an AMMI2 biplot with convex hull around the genotypes and genotype names
## displayed. Blow up genotypic scores by using envFactor = 0.3
plot(geAm, scale = 0.5, plotType = "AMMI2", sizeGeno = 2, plotConvHull = TRUE,
     envFactor = 0.3)
```



For the AMMI analysis a report can be made using the `report` function.

```
report(geAm, outfile = "./myReports/AMMIReport.pdf")
```

GGE Analysis

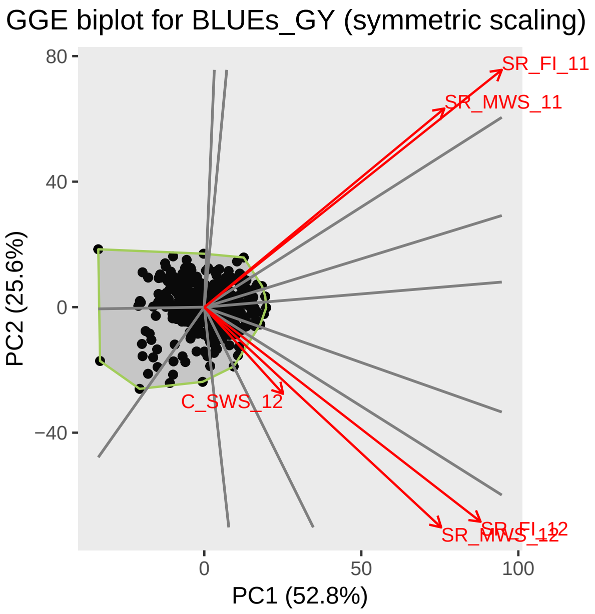
A Genotype plus Genotype by Environment analysis is very similar to an AMMI analysis. The difference is in the first step where, instead of genotype and environment, only environment is fitted as a main effect in the model. The rest of the analysis is identical to an AMMI analysis and is done in RAP by running the function `gxeAmmi` with parameter `GGE = TRUE`.

```
## Run gxeAmmi with default settings.
geGGE <- gxeAmmi(TD = TDGxE, trait = "BLUEs_GY", GGE = TRUE)
summary(geGGE)
#> Principal components
#> ====================
#>                               PC1        PC2
#> Standard deviation     1473.83381 1025.25811
#> Proportion of Variance    0.52835    0.25568
#> Cumulative Proportion     0.52835    0.78403
```

```
#> 
#> Environment scores
#> ==================
#>                  PC1        PC2
#> C_SWS_12  0.1476429 -0.1947368
#> SR_FI_11  0.5574381  0.5335308
#> SR_FI_12  0.5177392 -0.4824069
#> SR_MWS_11 0.4498476  0.4466397
#> SR_MWS_12 0.4439005 -0.4951959
```

Options for plotting results of a GGE analysis are identical to those for an AMMI analysis. `plotType` "GGE1" and "GGE2" may be used as substitutes for "AMMI1" and "AMMI2", but the latter are valid options as well.

```
## Create an GGE1 and GGE2 biplot.
plot(geGGE, scale = 0.5, plotType = "GGE2", plotConvHull = TRUE)
```



Finlay-Wilkinson Analysis

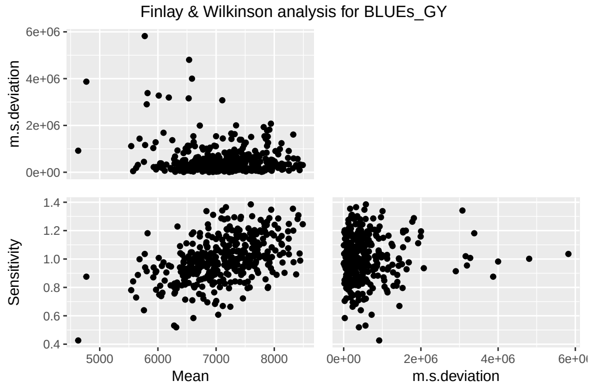
With the Finlay-Wilkinson Analysis (Finlay and Wilkinson 1963) a modified joint regression analysis is used to rank genotypes based on phenotypic stability for each individual trait.

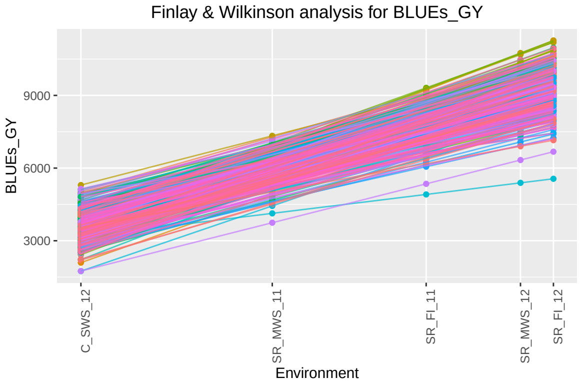
In the RAP package this analysis can be done using the `gxeFW` function. By default all trials in the `TD` object are used in the analysis, but this can be restricted using the parameter `trials`. The genotypes included in the analysis can be restricted using `restrictGeno`.

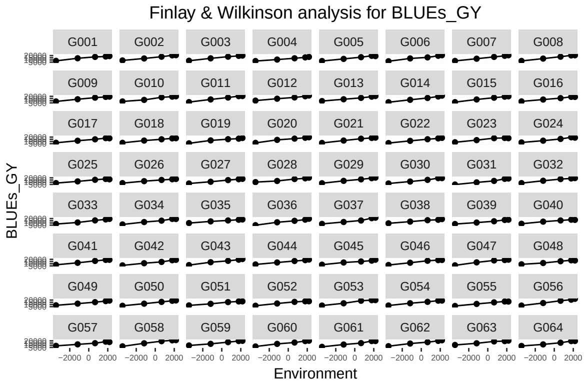
```
## Perform a Finlay-Wilkinson analysis for all trials.
geFW <- gxeFw(TD = TDGxE, trait = "BLUEs_GY")
summary(geFW)
#> Environmental effects
#> ====================
#>       trial     envEff se_envEff  envMean rank
#> 1  C_SWS_12 -3450.6949  56.48279 3651.430    5
#> 2  SR_FI_11   661.1322  56.58694 7762.838    3
#> 3  SR_FI_12  2176.7587  56.48279 9279.343    1
#> 4 SR_MWS_11 -1171.7041  56.58694 5930.812    4
#> 5 SR_MWS_12  1784.5081  56.48279 8887.059    2
#>
#> Anova
#> =====
#>                        Df     Sum Sq    Mean Sq   F value Pr(>F)
#> genotype              383  760927259    1986755    3.7129 <2e-16 ***
#> trial                   4 8308722428 2077180607 3881.8460 <2e-16 ***
#> Sensitivities         383  200189459     522688    0.9768 0.6051
#> Residual             1147  613761116     535101
#> Total                1917 9883600263    5155764
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Most sensitive genotypes
#> =========
#>  genotype      sens  se_sens  genMean se_genMean MSdeviation rank
#>      G202 0.4261900 0.157044 4630.041   326.7125   920155.42    1
#>      G229 0.5190070 0.157044 6314.679   326.7125   394503.53    2
#>      G338 0.5311154 0.157044 6278.680   326.7125   564566.01    3
#>      G330 0.5840103 0.157044 6609.668   326.7125    29086.74    4
#>      G285 0.6073710 0.157044 7036.768   326.7125   724189.99    5
```

Three types of plots can be made to investigate the output of the analysis. `plotType = "scatter"` creates three scatter plots where genotypic mean, mean squared deviation and sensitivity are plotted against each other. `plotType = "line"` creates a plot with fitted lines for all genotypes in the analysis. `plotType = "trellis"` creates a trellis plot with individual slopes per genotype. At most 64 genotypes are plotted. It is possible to select genotypes using the parameter `genotypes`.

```
plot(geFW, plotType = "scatter")
plot(geFW, plotType = "line")
plot(geFW, plotType = "trellis")
```







A report can be made as well containing a summary of the analysis.

```
report(geFW, outfile = "./myReports/FWReport.pdf")
```

Computation of mega environments

For the computation of mega environments, an AMMI model is fitted and then, using the fitted values from this model, the environments are clustered. Mega environments can be created by two clustering methods. The first method groups environments based on their best performing genotype. Environments that share the same best genotype belong to the same mega environment, regardless whether environments correspond to years or locations.
In the second method, genotypes that are above a certain quantile are used to classify locations into mega environments that are consistent across years. In this method, genotypes are scored according to whether they are above the `cutOff` threshold for the genotypic ranking within each location (1 if a genotype is above the threshold and 0 otherwise). This gives a genotype by location matrix with 1's and 0's that is used to calculate the correlation between locations. Then correlations across years are combined using the method by Charter and Alexander (Charter and Alexander 1993). The combined correlations are used to calculate Euclidean distances for hierarchical clustering. The number of mega environments obtained with the hierarchical clustering procedure is chosen to maximize the correlated response to selection within mega environments, as proposed in Atlin et al (Atlin et al. 2000).
Since the test data doesn't contain information about the year, only the first method is available for this data. This is the default setting for the `gxeMegaEnv` function.

```
geMegaEnv <- gxeMegaEnv(TD = TDGxE, trait = "BLUEs_GY")
#>  Mega factor     Trial Winning genotype AMMI estimates
#>            1 SR_MWS_12             G037      10863.969
#>            2  SR_FI_12             G056      11528.407
#>            3  C_SWS_12             G188       5423.140
#>            4  SR_FI_11             G276      10137.297
#>            4 SR_MWS_11             G276       7953.709
```

As can be seen in the column Mega Factor in the output, four mega environments have been created. In the environments SR_FI_11 and SR_MWS_11 G276 is the best genotype, so these two environments are clustered together. The other three environments have different winning genotypes and therefore form their own mega environment.

The values for the BLUPs and associated standard errors for the genotypes based on the calculated mega environments, can be computed using the function `gxeTable`. This can be done using either “asreml” or “lme4” as an engine for fitting.

```
if (requireNamespace(package = "asreml", quietly = TRUE)) {
  geMegaEnvPred <- gxeTable(TD = geMegaEnv, trait = "BLUEs_GY", engine = "asreml")
  head(geMegaEnvPred$predictedValue)
}
#>             1        2        3        4
#> G001 7080.357 7152.093 7107.182 8096.040
#> G002 8237.376 8317.016 7691.374 7785.526
#> G003 7665.637 7746.164 7398.335 8045.609
#> G004 6578.156 6522.595 6851.599 6448.847
#> G005 7507.483 7494.956 7297.933 7396.632
#> G006 7864.135 7943.569 7466.156 7534.489
```

Computation of stability measures

Different measures of stability can be calculated using the RAP package, the cultivar-superiority measure of Lin & Binns (LIN and BINNS 1988), Shukla's (Shukla 1972) stability variance and Wricke's (Wricke 1962) ecovalence.

The cultivar-superiority measure is the sum of the squares of the difference between genotypic mean in each environment and the mean of the best genotype, divided by twice the number of environments. Genotypes with the smallest values of the superiority tend to be more stable, and closer to the best genotype in each environment.

Shukla's stability variance (static stability) is defined as the variance around the genotype's phenotypic mean across all environments. This provides a measure of the consistency of the genotype, without accounting for performance.

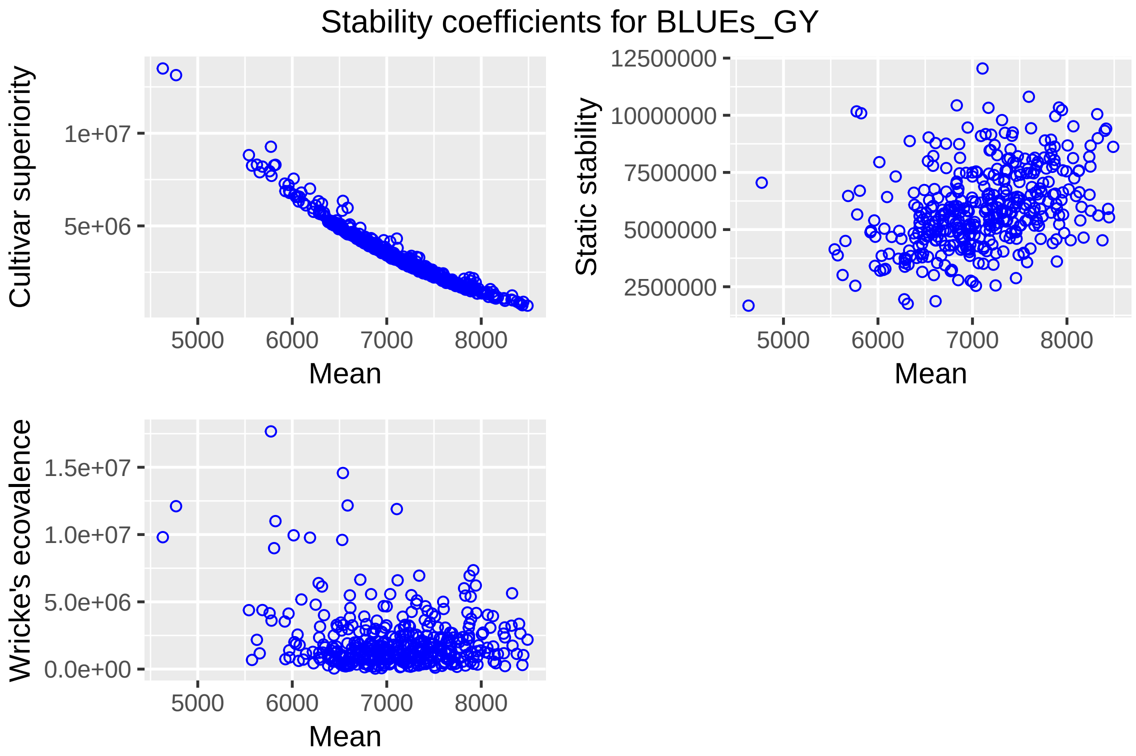
Wricke's Ecovalence Stability Coefficient is the contribution of each genotype to the GxE sum of squares, in an unweighted analysis of the GxE means. A low value indicates that the genotype responds in a consistent manner to changes in environment; i.e. is stable from a dynamic point of view. Like static stability, the Wricke's Ecovalence does not account for genotype performance.

```
geStab <- gxeStability(TD = TDGxE, trait = "BLUEs_GY")
summary(geStab, pctGeno = 2)
#> 
#> Cultivar-superiority measure (Top 2 % genotypes)
#>  genotype superiority     mean
#>      G202    13494769 4630.041
#>      G288    13135607 4770.062
#>      G384     9272209 5774.064
#>      G249     8821607 5541.442
#>      G316     8303387 5625.830
#>      G208     8296293 5822.169
#>      G260     8266599 5807.864
#>      G248     8251252 5573.973
#> 
#> Static stability (Top 2 % genotypes)
#>  genotype   static     mean
#>      G286 12046595 7106.561
#>      G077 10809302 7596.784
#>      G181 10433051 6833.935
#>      G197 10340757 7915.891
#>      G358 10325838 7169.372
#>      G073 10225251 7947.329
#>      G384 10169164 5774.064
#>      G208 10089354 5822.169
```

```
#>
#> Wricke's ecovalence (Top 2 % genotypes)
#>  genotype   wricke     mean
#>      G384 17665165 5774.064
#>      G381 14574450 6536.117
#>      G157 12162796 6585.634
#>      G288 12107556 4770.062
#>      G286 11896559 7106.561
#>      G208 10997293 5822.169
#>      G360  9942805 6014.688
#>      G202  9805416 4630.041
```

Plotting the results yields a scatter plot for each stability measure, plotted against the genotypic mean.

```
plot(geStab)
```



For the computation of stability measures a summary report can be made.

```
report(geStab, outfile = "./myReports/stabReport.pdf")
```

It is possible to calculate the stability measures based on mega environments instead of regular environments. To do so the parameter `useMegaEnv` has to be set to `TRUE`.

```
## Compute stabilities measures based on mega environments computed in the
## previous paragraph.
geStabME <- gxeStability(TD = geMegaEnv, trait = "BLUEs_GY", useMegaEnv = TRUE)
summary(geStabME, pctGeno = 2)
#>
#> Cultivar-superiority measure (Top 2 % genotypes)
#>  genotype superiority     mean
#>      G202    16062254 4630.041
#>      G288    12951183 4770.062
#>      G249    10485808 5541.442
#>      G250     9531106 5655.763
#>      G248     9465304 5573.973
#>      G316     9317957 5625.830
#>      G383     9113378 5760.066
#>      G384     8779234 5774.064
#> 
#> Static stability (Top 2 % genotypes)
#>  genotype   static     mean
#>      G286 13999250 7106.561
#>      G077 13531672 7596.784
#>      G197 13367454 7915.891
#>      G073 12691302 7947.329
#>      G181 12527941 6833.935
#>      G084 12483187 8068.285
#>      G358 12390526 7169.372
#>      G037 12314466 7876.846
#> 
#> Wricke's ecovalence (Top 2 % genotypes)
#>  genotype   wricke     mean
#>      G384 12366447 5774.064
#>      G381 10236181 6536.117
#>      G202  9024066 4630.041
#>      G288  8961553 4770.062
#>      G286  8802312 7106.561
#>      G208  7854699 5822.169
#>      G157  7377751 6585.634
#>      G205  6494081 6187.658
```

References

Atlin, G.N., R.J. Baker, K.B. McRae, and X. Lu. 2000. "Selection Response in Subdivided Target Regions." *Crop Science* 40 (1): 7. https://doi.org/10.2135/cropsci2000.4017.

Charter, Richard A., and Ralph A. Alexander. 1993. "A Note on Combining Correlations." *Bulletin of the Psychonomic Society* 31 (2): 123–24. https://doi.org/10.3758/bf03334158.

Finlay, KW, and GN Wilkinson. 1963. "The Analysis of Adaptation in a Plant-Breeding Programme." *Australian Journal of Agricultural Research* 14 (6): 742. https://doi.org/10.1071/ar9630742.

LIN, C. S., and M. R. BINNS. 1988. "A SUPERIORITY MEASURE OF CULTIVAR PERFORMANCE FOR CULTIVAR LOCATION DATA." *Canadian Journal of Plant Science* 68 (1): 193–98. https://doi.org/10.4141/cjps88-018.

Shukla, G K. 1972. "Some Statistical Aspects of Partitioning Genotype-Environmental Components of Variability." *Heredity* 29 (2): 237–45. https://doi.org/10.1038/hdy.1972.87.

Wricke, G. 1962. "Evaluation Method for Recording Ecological Differences in Field Trials." *Z Pflanzenzücht* 47: 92–96.