A Novel Approach to Plant Genotypic Classification in Multi-site Evaluation

Mathura Thillainathan1 and George C.J. Fernandez2
Department of Applied Economics and Statistics (MS 204), University of Nevada, Reno, NV 89557

Abstract. A user-friendly SAS statistical and graphical application to classify genotypes evaluated under multiple sites is presented. First, the test sites are classified into three environments, LOW (\( I_Y < Q_1 \)), MEDIUM (\( Q_1 < I_Y < Q_3 \)), and HIGH (\( I_Y > Q_3 \)) yielding environments, using the first (Q1) and third (Q3) quartile of the site mean yield (\( I_Y \)) as the cutoff value. Then, in each environment, the genotypes are classified as low (L: \( I_Y < (Q_1) \)), medium (M: \( (Q_1) < I_Y < (Q_3) \)), and high (H: \( I_Y > (Q_3) \)) yielding under each of the three environments, by comparing each genotype mean (\( I_Y \)) with the overall genotype mean (\( I_Y \)) based on \( \text{LS}_{\text{gen}} \) statistic computed from a separate two-way ANOVA models for LOW, MEDIUM, and HIGH yielding environments. Using the user-friendly SAS MACRO, EXPLORGE horticulturists can effectively and quickly perform genotype classification under multi-site evaluation. The steps involved in downloading the necessary MACRO-CALL file from the author’s home page [http://www.ag.unr.edu/gf] and the instructions for running the SAS MACRO are presented. The features of this graphical method and the graphics produced by the EXPLORGE MACRO are demonstrated and validated by published data.

Yield trials conducted with many genotypes grown in multiple sites and years form the basis for comparative genotypic evaluation. Genotypes that perform well over a wide range of environmental conditions are preferred in general. But some genotypes are not suitable for diverse environments due to the presence of genotype \times environment interaction (GEI) effects. Such interactions may confound the genotypic performance with environmental effects. Hence, selection of superior genotypes can be severely limited in the presence of a significant GEI (Finlay and Wilkinson, 1963; Shafii and Price, 1998; Zobel, 1990).

The concept of GEI and its stability statistics are being analyzed in different stability analysis methods for the selection of stable genotypes. Consistent performances across different sites and/or years are referred to as yield stability (Fernandez, 1991; Tai, 1971). Plaisted and Peterson’s (1959) mean variance component for pair wise GE interaction, Plaisted’s (1960) variance component for GE interaction, Wricke’s (1962) evolavence stability measure, Finlay and Wilkinson’s (1963) and Perkins and Jinks’ (1968) regression coefficient for the deviation from regression parameter, Tai’s (1971) stability statistics and Shukla’s (1972) stability variance are some of the stability statistics used in past studies. Tai’s, Shukla’s, and Finlay’s stability statistics were widely used under the regression approach of GEI analysis (Lin et al., 1986).

TAI’s stability plot (Tai, 1971) and Additive Main Effects and Multiplicative Interaction (AMMI) biplot analysis (Shafii and Price, 1998) are popular graphical methods that are used to identify the stable and environmentally sensitive genotypes. In TAI’s method, the GE \((g_{ij})\) interaction term is partitioned into two components: the linear response to environmental effects \((\alpha_i)\), and the deviation from the linear response \((\lambda_i)\). TAI’s stability estimates for perfectly stable and averagely stable genotypes are \((\alpha_i, \lambda_i) = (-1,1)\) and \((\alpha_i, \lambda_i) = (0,1)\), respectively. TAI’s analysis provides a method of obtaining the prediction interval for \(\alpha_i = 0\) and a confidence interval for \(\lambda_i\) values, so that the genotypes could be distributed graphically in different stability regions of the TAI’s plot and could be separated based on stability regions. Although TAI’s stability analysis is classifying the genotypes based on \(\alpha_i\) and \(\lambda_i\) estimates, the information on genotypic mean yield is not included in their graphical plot. TAI’s method may also identify a genotype that gives an average yield in a low yielding environment but a higher yield in a high yielding environment as an unstable genotype.

Fernandez et al. (1989) applied a segmented regression analysis model based on the study of Verma et al. (1978). In their study, the genotypes were separated into low yielding and high yielding groups and separate regression coefficients for low yielding \((\beta_i)\) and high yielding \((\beta_i)\) were computed. Then, based on the magnitude of regression coefficients estimate, the genotypes were categorized into seven different yield stability groups.

However, a graphical separation of genotypic groupings was not presented in this method.

AMMI biplot analysis enables clustering of genotypes based on similarity of response and the degree of stability were performed across diverse environments (Gauch, 1988; Shafii and Price, 1998). Here, the biplot displays of the Principal Components Analysis (PCA) scores where PCA1 vs. PCA2 or PCA1 vs. mean genotypic yield are plotted against each other to provide visual inspection and interpretation of the GEI components (Shafii and Price, 1998).

Thillainathan and Fernandez (2001) developed user-friendly SAS applications to perform enhanced TAI’s graphical Stability analysis and AMMI biplots and provided instructions for downloading and using these methods. However, the interpretation of the TAI’s stability statistics, AMMI/PCA scores, and the features in the biplot displays are usually considered difficult to understand by horticulturists. Therefore, the objective of this study is to develop a user-friendly graphical analysis of genotype selection that analyzes the GEI interaction graphically, thus being able to visually study how the different genotypes perform under different environments by grouping and naming them.

Materials and Methods

Data and analysis requirements. At least three genotypes and six sites, ranging from low yielding to high yielding, should be included in a replicated yield trial. If multi-year data are available, genotypic classification under multi-sites could be performed for each year separately, since the year effect is usually treated as random and significant genotype \times year interaction effects usually exist. In the case of horticultural crops with multiple harvests, the total yield per site per year can be used.

1) Grouping the sites into LOW, MEDIUM, and HIGH yielding environments

Assuming the mean yield of the \(i^{th}\) genotype for the \(j^{th}\) site is \(\bar{Y}_{ij}\).

Step 1. Estimate the site mean \(\bar{Y}_j\) for each \(j^{th}\) site.

Step 2. Rank the site means \(\bar{Y}_j\) and estimate the first quartile \((Q_1 = 25\text{-}\%\text{ percentile})\) and the third quartile \((Q_3 = 75\text{-}\%\text{ percentile})\) values using the SAS PROC UNIVARIATE procedure.

Step 3. Group the sites into three separate groups as LOW, MEDIUM, and HIGH yielding environments according to the following conditions: a) if the site mean \(\bar{Y}_j\) is less than \(Q_1\), the sites are classified as a low yielding environment (LOW); b) if the site mean \(\bar{Y}_j\) falls between \(Q_1\) and \(Q_3\), the sites are classified as a medium yielding environment (MED); c) if the site mean \(\bar{Y}_j\) is more than \(Q_3\), the sites are classified as a high yielding environment (HIGH).
2) Fitting separate two-way ANOVA models for LOW, MEDIUM, and HIGH yielding environments

Step 1. Because the genotype and sites are usually treated as fixed, a fixed ANOVA model can be fitted using the SAS GLM (SAS) procedure. The Type III SS in SAS GLM procedure can handle unbalanced data in a fixed effects model. Fit three, separate, two-way ANOVA models, as shown below, with a heteroscedastic error structure for each environment using the SAS PROC GLM procedure:

\[ y_{ijkl} = \mu + g_{i(l)} + s_{j(l)} + b_{k(l)} + e_{ijkl} \] \hspace{1cm} [1]

where:

- \( l \) = 1 for low yielding environment (LOW) \( l \) = 2 for moderate yielding environment (MEDIUM) \( l \) = 3 for high yielding environment (HIGH)

- \( g_{i(l)} \) is the response variable of the \( i \)th genotype
- \( j \)th site of \( k \)th replicate in \( l \)th environment
- \( \mu \) is the grand mean
- \( s_{j(l)} \) is the effect of site in \( l \)th environment, \( b_{k(l)} \) is the effect of \( k \)th replicate or block in \( l \)th environment, and \( e_{ijkl} \) is the random error term. Since similar sites are grouped within each environment, interaction between genotype and sites within each environment was assumed biologically nonsignificant.

Step 2. Three separate error structures for LOW, MEDIUM, and HIGH yielding environments are estimated in two-way ANOVA models by environments to reduce effects of the unequal error variance on mean comparison. Therefore, three separate mean square errors are computed for each of the LOW, MEDIUM, and HIGH environments.

3) Grouping the genotypes into low yielding (L), moderate yielding (M), and high yielding (H) under each of the three environments

Step 1. Estimate the LSD\(_{0.01}\) critical value for comparing each genotypic mean with the grand mean in each environment.

Based on a two-way ANOVA model, assuming that genotype is a fixed factor and the genotype \( \times \) site interaction within each environment is negligible, three separate LSD\(_{0.01}\) values are computed. The LSD\(_{0.01}\) for comparing each genotypic mean \( \bar{y}_{i(l)} \) with the grand genotypic mean, \( \bar{y}_{..(l)} \), for \( l \)th environment can be computed as follows,

\[ \text{LSD}_{0.01} = t_{0.01/2, a} \cdot \sqrt{\frac{MSE}{r_{l}}} \quad [2] \]

where \( t_{0.01/2, a} \) is the \( t \)-statistic from Student’s \( t \)-table, and \( (r_{l}) \) is the MSE (mean squared error) for the \( l \)th environment. The number of replications used to compute the \( i \)th genotypic mean and the grand mean within each environment differs substantially. Therefore, a weighted denominator, \( r = 0.5 \left[ 1/r_{L} + 1/r_{H} \right] \), where \( r_{l} \) is the number of observations used to compute the \( i \)th genotypic mean yield \( (\bar{y}_{i(l)}) \), and \( r \) is the number of observations used to compute the \( l \)th environmental grand genotypic mean \( (\bar{y}_{..(l)}) \), is used. Problems with the missing values and unbalanced data could be handled correctly since each LSD value estimation was adjusted for unbalanced replication.

Step 2. The genotypes are also classified subsequently as High (H), Medium (M), and Low (L) yielding, under each of the three environments, based on the following criteria:

- High (H) yielding in \( l \)th environment: yield of \( i \)th genotype \( (\bar{y}_{i(l)}) \) in \( l \)th environment \( \geq \bar{y}_{..(l)} + \text{LSD}_{0.01} \)
- Medium (M) yielding in \( l \)th environment: yield of \( i \)th genotype \( (\bar{y}_{i(l)}) \) in \( l \)th environment \( < \bar{y}_{..(l)} - \text{LSD}_{0.01} \)
- Low (L) yielding in \( l \)th environment: yield of \( i \)th genotype \( (\bar{y}_{i(l)}) \) in \( l \)th environment \( < \bar{y}_{..(l)} - \text{LSD}_{0.01} \)

Naming the genotypes based on three-letter codes

A genotype with a three-letter code of “LMH” can be interpreted as low yielding (L) under LOW environment, average yielding (M) under MEDIUM environment, and high yielding (H) under HIGH environment. This genotype is considered highly environmentally sensitive and shows a below average stability rating based on traditional stability analysis. A genotypic code of “MMM” can be interpreted as the genotype performing average in all three LOW, MEDIUM, and HIGH environments. Therefore, this genotype is considered similar to an averagely stable genotype based on traditional stability methods.

Data

To validate the genotypic groupings and to compare the effects of previous stability statistics, we used the published winter rapsaeed variety yield data (Shafii and Price 1998) downloaded from the web site at http://www.uidaho.edu/ag/statprog/ammi/. This winter rapsaeed variety yield data (Mg ha\(^{-1}\)) consists of yield data for six genotypes grown in 14 sites for 3 years with four replicates per trial.

Results and Discussion

The yield characteristics of the sites classified as LOW (1052 Mg ha\(^{-1}\) ± 38), MEDIUM (1967 Mg ha\(^{-1}\) ± 50), and HIGH (3709 Mg ha\(^{-1}\) ± 102) yielding based on the mean site yield over genotypes and year are presented in Table 1. Out of the 14 sites included in this study, three, six, and five sites were classified as a LOW, MEDIUM, and HIGH yielding environment, respectively (Table 1). There was about a 2-fold increase in the mean yield between LOW to MEDIUM and MEDIUM to HIGH yielding environments.

The genotypic yield variation in each environment for 3 years are presented in comparative box-plots in Figure 1. In all 3 years, the degree of yield variation in the high yielding environment was at least 2-fold more than the genotypic variation in the LOW yielding environment. These findings clearly justify the need for the separate error model in the ANOVA and the LSD computation.

The genotypic groupings based on graphical GEI analysis by running SAS MACRO EXPLORGE for three different years—1987, 1988, and 1989—are presented in Figure 2. In these EXPLORGE plots, the Y axis represents the seed yield mean (Mg ha\(^{-1}\)) and the X axis represents the three classified environments, LOW, MEDIUM, and HIGH. The genotypes and their genotypic groupings (L, M, H), based on their relative performance over the mean genotypic yield in three environments, are presented in the figure legends. All the users can visually compare the mean yield performance of the genotypes in three environments directly from these plots. The graphical plots of EXPLORGE appear to be simpler and easier to understand by breeders, horticulturists, and agronomists, than interpreting TAI’s plot or AMMI biplots. But, the presence of low yielding sites (>15) in the data set might cloud the interpretation and separation of genotypes in the EXPLORGE graphical plots. However, the inclusion of 3-letter codes for each genotype in the graphics helps to overcome the problem of crowding caused by too many genotypes in the data set.

In the original analysis of Shafii et al. (1998), 14 sites and 3 years with different combinations were treated as 27 fixed environments. But in this study, the year effect was separated from the site effect and EXPLORGE analyses were performed separately by year. In the 1987–89 EXPLORGE analysis, one out of the six genotypes was classified as ‘MMM’ (Table 2), indicating that the mean yield potentials of the ‘Cascade’ genotype was average in all three LOW, MEDIUM, and HIGH yielding environments for all three years. The genotypes ‘Glacier’ and ‘Bridge’ performed average or below in all three environments. The genotype ‘Jet’ was sensitive to the low yielding environments since, in 1987, the yield

| Attributes | LOW-yielding sites | MEDIUM-yielding sites | HIGH-yielding sites |
|------------|-------------------|-----------------------|--------------------|
| Mean yield (Mg ha\(^{-1}\)) | 1052 (±38) | 1967 (±50) | 3709 (±102) |

The sites included in the above table: Georgia (GGA), TGA), Idaho (ID), Kansas (KS), Mississippi (MS), Montana (MT), New York (NY), North Carolina (NC), Oregon (OR), South Carolina (SC), Tennessee (TN), Texas (TX), Virginia (VA), and Washington (WA).
Fig. 1. Comparative box-plots of yield variation (Mg·ha⁻¹) of winter rapeseed in three (LOW, MEDIUM, and HIGH) environments for 3 years.
Fig. 2. Mean genotypic yield (Mg·ha⁻¹) of winter rapeseed cultivar and genotypic groupings based on EXPLORGE analysis under 'LOW', 'MEDIUM,' and 'HIGH' yielding environments for year (a) 1987, (b) 1988, and (c) 1989.
Table 2. Genotypes used in the winter rapeseed cultivar trial (Shafii and Price, 1998) and their respective performance codes, named according to their performance in all three environments, LOW, MED, and HIGH, for 3 years.

| Genotype   | 1987 | 1988 | 1989 |
|------------|------|------|------|
| Bienvenu   | L/M/H| MMH  | MHL  |
| Bridge     | MMM  | MMM  | MMM  |
| Cascade    | MMM  | MMM  | MMM  |
| Dwarf      | LLM  | LLM  | MLM  |
| Glacier    | MMM  | MMM  | MMM  |
| Jet        | HMM  | LMM  | LMM  |

1. Indicates whether the genotype is low yielding (L), medium yielding (M) or high yielding (H) in LOW environment.
2. Indicates whether the genotype is low yielding (L), medium yielding (M) or high yielding (H) in medium (MED) environment.
3. Indicates whether the genotype is low yielding (L), medium yielding (M) or high yielding (H) in HIGH environment.

Conclusions

The proposed method of genotype classification evaluated under multiple sites for multiple years has the following advantages:

- GE interaction effect and yield potentials of each genotype in LOW, MED, and HIGH yielding environments can be viewed directly in the same graphics.
- Easy to follow genotype groupings are made for the purpose of effective genotype selection under wide environmental conditions using a standard two-way ANOVA model.
- Problems caused by missing or unbalanced replications are treated by computing individual LSMEAN values for each genotype mean with the overall genotype mean in a given environment.

A user-friendly SAS MACRO ‘EXPLORGE’ is currently available for immediate use by horticulturists. They can effectively and quickly perform the genotype selection under multi-environments, and this MACRO can help them to analyze their data immediately following the trial, spend more time in data exploring, interpretation of graphics, and output rather than spending time on writing SAS program codes or depending on SAS programmers to analyze their data.

Literature Cited

Eberhart, S.A. and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36–40.

Fernandez, G.C.J. 1991. Analysis of genotype × environment interaction by stability estimates. HortScience 26:947–950.

Fernandez, G.C.J., H.K. Chen, and J.C. Miller, Jr. 1989. Adaptation and environmental sensitivity of mungbean genotypes evaluated in the international mungbean nursery, Euphytica 41:253–261.

Finlay, K.W. and G.N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding program. Aust. J. Agr. Res. 14:742–754.

Gauch, H.G. 1988. Model selection validation for yield trials with interaction. Biometrics 44:705–715.

Lin, C.S., M.R. Binnis, and L.P. Lefkovitch. 1986. Stability analysis: Where do we stand. Crop Sci. 26:894–900.

Perkins, J.M. and J.L. Jinks. 1968. Environmental and genotype-environment components of variability. Ill. Multiple lines and crosses. Heredity 23:339–356.

Plaisted, R.L. 1960. A shorter method for evaluating the ability of selections to yield consistently over locations. Amer. Potato J. 37:166–172.

Plaisted, R.L. and L.C. Peterson. 1959. A technique for evaluating the ability of selections to yield consistently in different locations or seasons. Amer. Potato J. 36:381–385.

Shafii, B. and W.J. Price. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. J. Agr. Biol. Environ. Stat. 3:335–345. [http://www.ovidado.com/ag/statprog/ammi/]

Shukla, G.K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237–245.

Tai, G.C.C. 1971. Genotypic stability analysis and its application to potato regional trials. CropScience 11:184–190.

Thullinathan, M. and G.C.J. Fernandez. 2001. SAS applications for Tai’s stability analysis and AMMI model in Genotype × Environment Interaction (GEI) effects. J. Hered. 93(4):367–371.

Verma, M.M., G.S. Chahal, and B.R. Murty, 1978. Limitations of conventional regression analysis: A proposed modification. Theor. Appl. Genet. 53:83–91.

Wricke, G. 1962. Uber eine Methode zur Erfassung der ökologischen Streubreite in Feldversuchen Z. Pflanzenzuecht 47:92–96.

Zobel, R.W. 1990. A powerful statistical model for understanding Genotype-by-Environment Interaction. Proc. Genotype-by-Environment Interaction and Plant Breeding. Louisiana State Univ., Baton Rouge.

Appendix

Running the user-friendly SAS MACRO ‘EXPLORGE’ to select genotypes

SAS version 6.12 for PC Win/NT was used to develop SAS MACRO EXPLORGE and later modified to include the advanced features available in SAS Version 8.2. The requirements for using this SAS MACRO are:

- A valid licence to run the SAS software on your PC.
- SAS modules such as SAS/BASE, SAS/STAT, and SAS/GRAPH should be installed in your computer.
- A working internet connection to access the EXPLORGE macro from the CABNAR server at the Univ. of Nevada–Reno while executing the downloaded macro-call file.

The steps for running the EXPLORGE MACRO are given below:

Step 1. Create a SAS data set containing multi-site, multi-year and replicated yield data. The users can also download the sample data set (Shafii and Price, 1998) from Fernandez’s home page [http://www.ag.unr.edu/agf] when downloading the SAS MACRO call file. This data set should contain the following variables:

- Genotype (GEN), a categorical variable
- Site (ENV), a categorical variable
- Block or Replications (BLK), the replications will be treated as blocks
- Response (YIELD), a continues numeric variable
- Optional Year (YEAR) variable for multi-year data

Step 2. Visit Fernandez’s (2001) home page [http://www.ag.unr.edu/agf], click the running dog icon, and follow the instructions given to go to the download page. Download the EXPLORGE.SAS MACRO CALL file and/or sample data file by clicking download link, save the file to a disk, and open it in the SAS PROGRAM EDITOR WINDOW. Click the ‘RUN’ icon to open the EXPLORGE MACRO-CALL window (Fig. 3).
Step 3. Input the required values by following the instructions provided in the SAS MACRO–CALL window for EXPLORGE (Fig. 3). Input the required SAS data set name, site, response, genotype, and block or replication variable names. If data is available for only one particular year, then leave the year variable field (field number 6, Fig. 3) blank and, if multiple year data is available, input the year variable name.

Options for saving the SAS output and SAS graphics files:

Users can select the folders to save the SAS output and the graphic files by inputting the folder names in the MACRO–CALL window. The users can also select one of the following output formats when saving the graphics and the output produced by the SAS MACRO:

DISPLAY: Files are not saved, but displayed in the SAS output and graphics window.

WORD: In version SAS 6.12, the graphics’ files are saved as CGM files and the output files are saved as TXT files suitable for including in Microsoft office products. But, in version SAS 8.0 and above, both SAS graphics and the output file are saved together as an RTF file suitable for opening in MS WORD.

WEB: In version SAS 6.12, the graphics’ files are saved as GIF files and the output files are saved as TXT files suitable for including in Microsoft office products. But, in version SAS 8.0 and above, both SAS graphics and the output file are saved together as an HTML file suitable for opening in any INTERNET browser.

PDF: In version SAS 6.12, the graphics’ files are saved as CGM files and the output files are saved as TXT files suitable for including in Microsoft office products. But, in version SAS 8.0 and above, both SAS graphics and the output file are saved together as a PDF file suitable for opening in ACROBAT Reader.

TXT: In version SAS 6.12, the graphics’ files are saved as CGM files and the output files are saved as TXT files suitable for including in Microsoft office products. But, in version SAS 8.0 and above, the graphics are saved as EMF files and the output files are saved as TXT files suitable for including in Microsoft office products.

Step 4. Submit the SAS MACRO

After inputting all the required fields, move your cursor to the last MACRO field, 10 (Fig. 3) and hit the ENTER KEY to run the SAS MACRO. The MACRO–CALL window file automatically accesses the appropriate SAS MACRO from the Internet server, College of Agriculture, Univ. of Nevada, and provides the users with the required graphs, genotypic classification, and plots.