GEA-R: suite of R programs for Genotype × Environment Analyses

Mateo Vargas Hernández *
Universidad Autónoma Chapingo, Texcoco, México – vargas_mateo@hotmail.com
Centro Internacional de Mejoramiento de Maíz y Trigo, Texcoco, México

Rosa Angela Pacheco Gil
Centro Internacional de Mejoramiento de Maíz y Trigo, Texcoco, México – R.A.Pacheco@cgiar.org

Abstract

Appropriate use of statistical methods for selecting a genotype's best performance under different environmental conditions is essential for breeding programs. Sets of genotypes evaluated in single or multiple environments (locations, years, etc.) under different management conditions such as water stress or low nitrogen are known as multi-environment trials (METs). These trials can detect and explain whether there are repeatability or interaction mechanisms between genotypes and environments. The GEA-R is a suite of R codes for analyzing and interpreting the genotype × environment interaction (GEI) from METs. The models calculated by the GEA-R are: The well-known Additive Main Effects and Multiplicative Interaction (AMMI or GE model), Sites Regression (SREG or GGE model), and when external environmental or genotypic covariates (molecular markers) is available, it is possible to use Partial Least Squares (PLS) regression and/or Factorial Regression (FR) models. Also are available several parametric and non-parametric methods for Stability Analyses. GEA-R can adjust the models for sets of genotypes evaluated in different environments using complete and/or incomplete blocks designs, or using only the adjusted means. To make GEA-R more accessible and easy to use, it runs through a graphic interface created in JAVA software. This interface generates graphics statistics (biplots), as well as numerical summaries of results in comma-delimited files. Additionally, other important contribution of this suite of R programs resides in the fact that is based on free software.

Keywords: GE and GGE models; Partial Least Squares regression; Factorial Regression; Free Software.

1. Introduction

When assessing grain yield of a set of cultivars in multi-environment trials, changes are commonly observed in the relative yield performance of cultivars with respect to each other across sites. This differential yield response of cultivars from one environment to another is called genotype × environment interaction (GEI) and can be studied, described, and interpreted by statistical models. A commonly used procedure for modeling interaction is a simple regression of the cultivar performance on the site mean. This model can be depicted in a set of straight lines with different slopes, one for each cultivar, and the heterogeneity of slopes accounts for the interaction. A generalization of the regression on the site mean model is the multiplicative model also called Additive Main Effect and Multiplicative Interaction (AMMI) model. The AMMI model provides more opportunity for modeling and interpreting GEI than the simple regression on the site mean model because it allows modeling the GEI in more than one dimension.

When additional information is available on environments, cultivars, or both, GEI can be modeled directly by the factorial regression model (van Eeuwijk *et al*., 1996). However, this approach has the problem to be sensitive to multicollinearity and noise and is non-parsimonious. To overcome some of these problems, Aastveit and Martens (1986) proposed the partial least squares (PLS) regression method as a more direct and parsimonious linear-bilinear model. This method consists of relating **X** and **Y** matrices in one single estimation procedure. The **Y** matrix contains site × cultivar

grain yield data as dependent variables and the **X** matrix has the external environmental variables (or external genotypic variables like molecular markers) as the explanatory variables.

2. Models

2.1. The basic two-way model

The basic two-way, fixed-effect, linear model considers that the empirical response, y_{ij}, of the ith level of cultivar (i=1,2,…,I), and the jth level of site (j=1,2,…,J) with (r=1,2,..,n) replications in each of the $I \times J$ cells is expressed as

$$y_{ijr} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + \varepsilon_{ijr} \qquad (1)$$

where μ is the grand mean over all observations of both factors, τ_i is the additive effect of the i^{th} cultivar, δ_j is the additive effect of the jth site, $(\tau\delta)_{ij}$ is the non-additive interaction of the i^{th} cultivar in the j^{th} environment, and ε_{ijr} is the error associated with τ_i, δ_j, and assumed to be normally and independently distributed (NID) $(0, \sigma^2)$.

2.2. Fixed-effect linear-bilinear models

Linear-bilinear models have been used extensively for studying and modeling the genotype × environment (GE) interaction. Detailed descriptions of these family of models can be found in several articles Gauch (1988), Crossa and Cornelius (1997), Vargas *et. al*., (1999, 2001). From Eq. 1, the general formulation of the fixed-effect linear-bilinear model is

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij} \qquad (2)$$

where the constant λ_k is the singular value of the kth multiplicative component that is ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$; the α_{ik} are elements of the kth left singular vector of the true interaction and represent genotypic sensitivity to hypothetical environmental factors represented by the kth right singular vector with elements γ_{jk}. The α_{ik} and γ_{jk} satisfy the constraints $\sum_i \alpha_{ik} \alpha_{ik'} = \sum_j \gamma_{jk} \gamma_{jk'} = 0$ for k≠k′ and $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$.

2.3. The site regression model (SREG)

The SREG model is a variant of the AMMI model (Crossa and Cornelius, 1997), in which the bilinear term models the main effects of genotypes [G] plus the GE interaction, and the composition of the two-way $I \times J$ matrix that is subjected to singular value decomposition is different than the one used in the AMMI model. The SREG model is useful for agricultural and plant breeders researchers for analyzing multi-environment trials to study the response pattern of genotypes and environments.

2.4. The Factorial Regression (FR) model

When external climatic variables and genotypic covariables such as diseases, molecular markers, etc. are available, can be used for examining the causes of interaction. These external variables are introduced as covariables in a regression analysis for explaining the interaction.

Factorial regression model (van Eeuwijk et al., 1996; Vargas et al., 1999; 2001) has been used for studying the effects of both genetic and environmental covariables and for explaining the causes of GEI. The FR models are ordinary linear models that approximate the interaction effects in Eq. 1 by the products of one or more: (1) τ_i covariables (observed) × δ_j effects (estimated); and (2) τ_i effects (estimated) × δ_j covariables (observed).

For m=1,…,G covariables of τ_i represented by $x_{i1},...,x_{iG}$, Eq. 1 becomes $y_{ijk} = \mu + \tau_i + \delta_j + \sum_{g=1}^{G} x_{ig}\xi_{jg} + \varepsilon_{ijk}$, $G \leq I$-1, where ξ_{jg} represents the regression coefficient of factor δ_j with respect to the covariable x_{ig} of τ_i. For h=1,…,H covariables of δ_j represented by $z_{j1},...,z_{jH}$, Eq. 1 is $y_{ijk} = \mu + \tau_i + \delta_j + \sum_{h=1}^{H} \varsigma_{ih} z_{jh} + \varepsilon_{ijk}$, $H \leq J$-1, where ς_{ih} represents the regression coefficient of τ_i with respect to the covariable z_{jh} of δ_j.

2.5. Partial least squares (PLS) regression

Vargas et al. (1999, 2001) described PLS in the context of agronomic and GE interactions and its relationship with the FR. When genotype responses over environments (**Y**) are modeled using environmental covariables, the J×H matrix **Z** of H (h=1,2,…,H) environmental covariables can be written as

$$\mathbf{Z}=\mathbf{t}_1\mathbf{p}'_1+\mathbf{t}_2\mathbf{p}'_2+\ldots+\mathbf{t}_M\mathbf{p}'_M+\mathbf{E}_M=\mathbf{TP}'+\mathbf{E} \quad (3)$$

where the matrix **T** contains $\mathbf{t}_J$, J×1 vectors called latent environmental covariables or Z-scores (indexed by environments), the matrix **P** contains the $\mathbf{p}_1 \ldots \mathbf{p}_H$, H×1 vectors called Z-loadings (indexed by environmental variables), and **E** has the residuals. The matrix **Y** is

$$\mathbf{Y}=\mathbf{t}_1\mathbf{q}'_1+\mathbf{t}_2\mathbf{q}'_2+\ldots+\mathbf{t}_M\mathbf{q}'_M+\mathbf{F}_M=\mathbf{TQ}'+\mathbf{F} \quad (4)$$

where the matrix **T** is same as in the previous equation, and the matrix **Q** contains the $\mathbf{q}_1 \ldots \mathbf{q}_I$ I×1 vectors called Y-loadings (indexed by genotypes) and **F** has the residuals. Principal component analysis of **Z** and of **Y** allows reducing the number of variables in each system. A reduced number of PLS latent variables give a low rank representation of the least squares estimates of the FR with environmental covariables.

2.6. Experimental data

Several data sets are included in GEA-R as illustration of the different experimental designs that can be analyzed in GEA-R. The data must include sets of genotypes evaluated in different environments (locations, years, management conditions, a combination of them, etc.), the experimental design could be a RCBD, a Lattice or simply the adjusted means. Also can include environmental variables (maximum and minimum temperatures, sun radiation, relative humidity, etc. in the vegetative, flowering and grain filling stages), or genotypic covariables (molecular markers).

3. Suite of R programs

GEA-R was created in a Windows environment and consists of a set of R codes. This suite was linked to a JAVA GUI file and optimized for use with R 3.1.1, but it is compatible with earlier and higher versions of R. The data are loaded as a first step and then it is highly advisable to do an exploratory analysis through the descriptive statistics, which can be calculated easily with this suite. The next step involves choosing the type of experimental design to be analyzed, and the response variables. Finally, you can select the type of analysis to be performed: AMMI and/or SREG models, PLS regression, Factorial Regression, or Stability Analysis. JAVA GUI facilitates doing the analyses

in GEA-R because it works with a graphic interface screen that tells the user which steps to follow until the analysis is successfully completed.

3.1. Running GEA-R

To run GEA-R, the user must open the interface GEA-R.jar. When the program is used for the first time, an R-version has to be loaded by clicking on the 'Load R' button. GEA-R will display a list of R-versions detected as installed in the computer (Fig. 1a). If user has not installed the R software, he/she needs to download it from the internet or run it directly from a file included in GEA-R. The user must select the R-version to use. It is possible to switch to another R-version by clicking on the 'Change R' button. Once you have selected the R version to use, the program ask if user is interested in update the R packages, if this option is selected, the updates will be done automatically but requires you have an internet connection available.

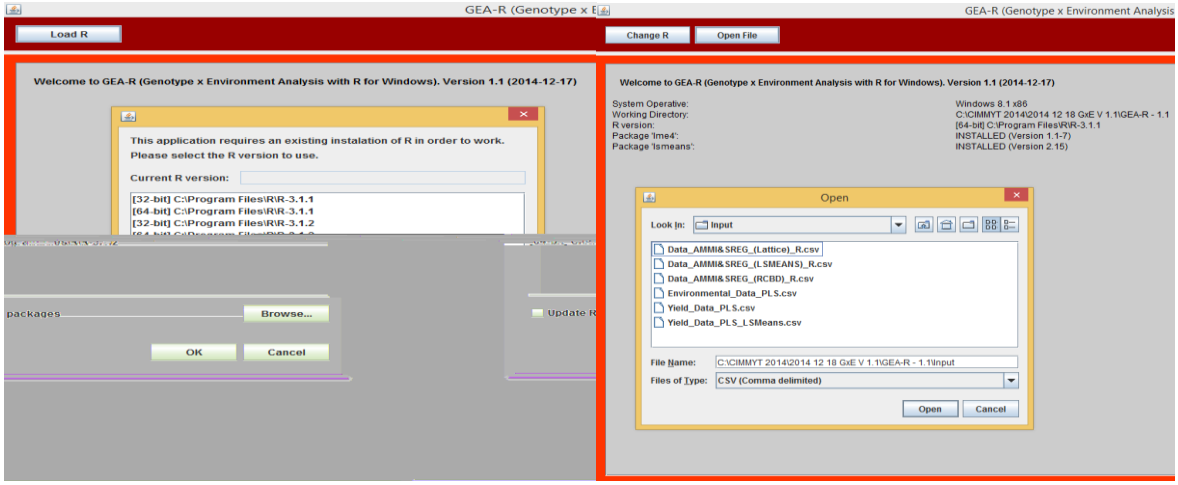


Figure 1. First Menu shown by GEA-R for to (a) left side, select the appropriate R version, (b) right side, select the data to be used in the analyses

The next step is to load the data to be analyzed by clicking on 'Open File' (Fig. 1b**)**. Data must be in a comma-separated file (.csv) with the names of the factors and traits in the first row. After loading the data, the user must specify the details of the analysis which he/she wishes to run (Fig. 2a**)**: the experimental design used, factor names, and the name of the output folder where the results will be saved. In addition, he/she will be asked if he/she wants descriptive statistics: boxplots, histograms or a basic statistics summary. The basic statistics summary will display the minimum, maximum, quantiles, means, and standard deviation for each trait-location combination. The boxplots and histograms may be saved in PDF files while the summary statistics may be saved in CSV files.

To go on to the next step, click on 'Continue'. In this step, the user has to select which analysis to perform (AMMI, SREG, PLS, FR, or Stability) and the traits to be analyzed (Fig. 2b). Finally, when GEA-R finished the selected methodology will show the results generated which will be automatically saved in the output folder, you can see them directly from the links shown in the lower right side of the display

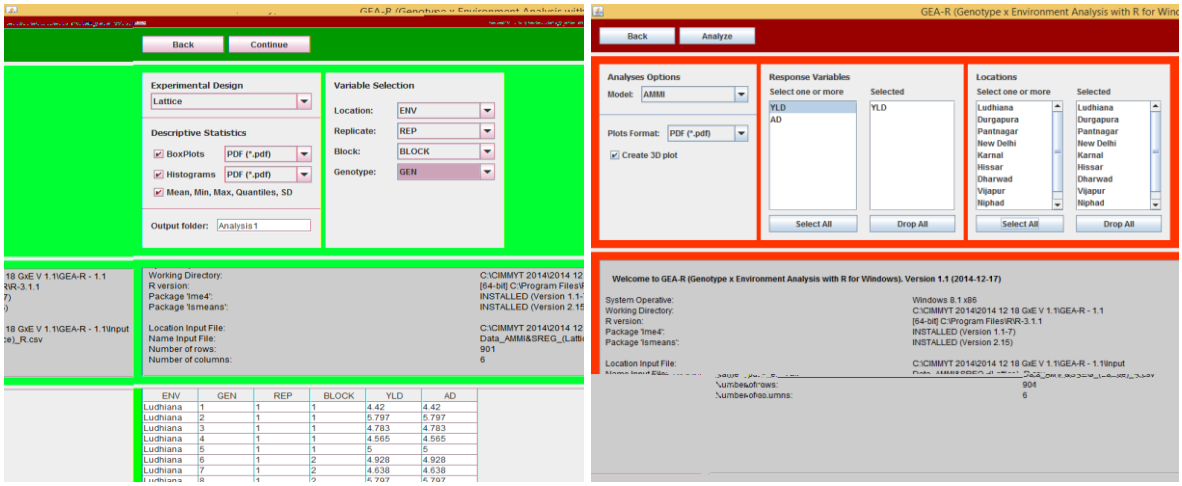


Figure 2. Submenus for to (a) left side, select the experimental design, descriptive statistics, folder name where to save the results, and names of the factors, (b) right side, selection of model (AMMI, in this example), variable and environments to include in the analysis

4. Results generated by GEA-R

Some of the graphics (biplots) as well as summaries of statistical results are shown without explanation because obvious limitations of space. The purpose is only to show the versatility for using data coming from different experimental designs, managing both location and/or genotypes as numerical or character variables, including or not external information as environmental or genotypic covariables, etc., and how the GEA-R can analyze data using complementary methodologies, which is the best and highly recommended strategy when interpreting the GEI. The interpretation of the biplots and a comparison of the results obtained from the different methodologies can be consulted in several articles (van Eeuwijk et al., 1996; Vargas, et al., 1999, 2001; Yan and Kang, 2002).

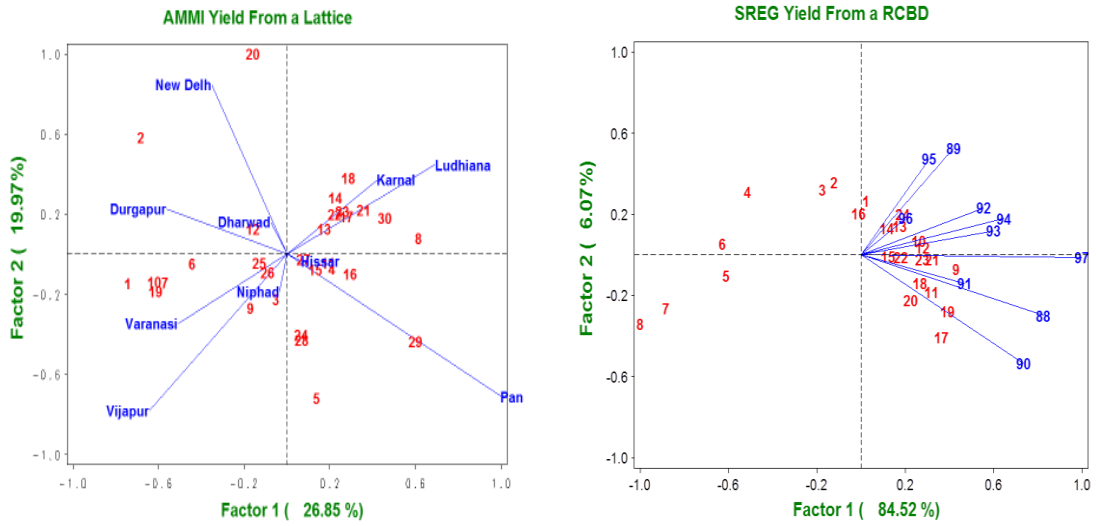
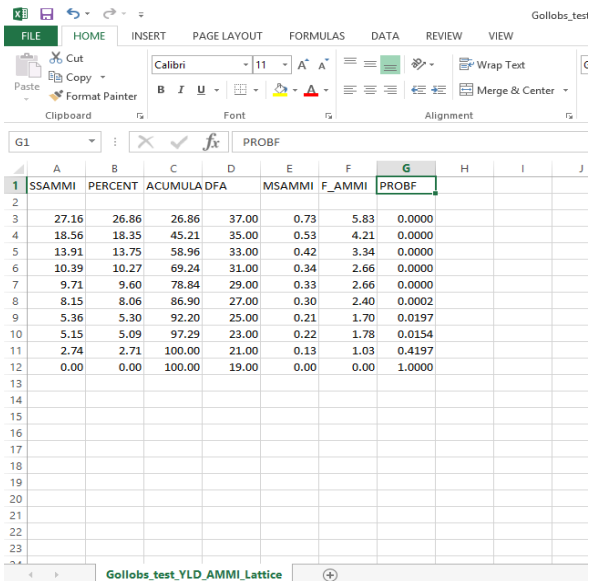


Figure 3. (a) Left side, biplot for AMMI model, data coming from a Lattice Design, genotypes (in red) as numerical and environments (locations, in blue) as character variables. (b) Right side, biplot for the SREG (GGE) model for data coming from a Lattice design.



SSAMMI	PERCENT	ACUMULA	DFA	MSAMMI	F_AMMI	PROBF
27.16	26.86	26.86	37.00	0.73	5.83	0.0000
18.56	18.35	45.21	35.00	0.53	4.21	0.0000
13.91	13.75	58.96	33.00	0.42	3.34	0.0000
10.39	10.27	69.24	31.00	0.34	2.66	0.0000
9.71	9.60	78.84	29.00	0.33	2.66	0.0000
8.15	8.06	86.90	27.00	0.30	2.40	0.0002
5.36	5.30	92.20	25.00	0.21	1.70	0.0197
5.15	5.09	97.29	23.00	0.22	1.78	0.0154
2.74	2.71	100.00	21.00	0.13	1.03	0.4197
0.00	0.00	100.00	19.00	0.00	0.00	1.0000

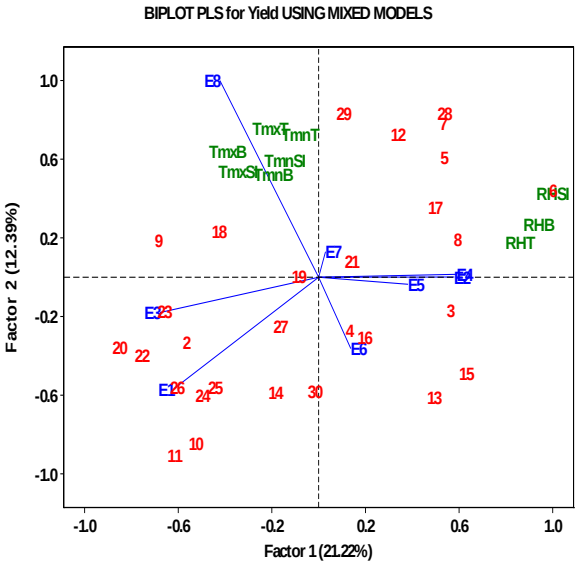


Figure 4. (a) Left side, summary of the Gollob's Test for determining the number of significant factors in the AMMI model, including also the sum of squares, individual and accumulated proportion of the total GEI sum of squares explained, degrees of freedom, mean square, F value and probability associated, for each factor; (b) Right side, biplot from Partial Least Squares (PLS) regression, including simultaneously information related to the genotypes (in red), environments (in blue), and environment covariables as external information (in green) for explaining the GEI. Tmx: maximum temperature, Tmn: minimum temperature, RH: relative humidity, evaluated in three different stages of the crop development.

Table 1. Stepwise selection summary of results obtained from the factorial regression (FR) using 600 molecular markers (M) as genotypic covariables for explaining the genotype × environment (year) interaction. The best model was selected using the AIC criterion and a significance level of 5%.

Effect Entered	DF	Sum of Squares	Mean Square	AIC	F Value	Pr > F	% GEI explained
Model	**206**	**174003.87**	**844.67**		**5.65**	**<.0001**	
Year	3	5034.46	1678.15		27.41	<.0001	
Gen	175	102727.10	587.01		3.75	<.0001	
Year × Gen	546	100776.75	184.57			<.0001	
Year×M321	3	7668.60	2556.20	4673.71	11.03	<.0001	7.61
Year×M392	3	3048.69	1016.23	4645.65	8.57	<.0001	3.03
Year×M246	3	3218.69	1072.89	4626.97	6.14	0.0004	3.19
Year×M454	3	3516.44	1172.14	4610.22	5.62	0.0008	3.49
Year×M82	3	3338.80	1112.93	4592.29	5.88	0.0006	3.31
Year×M579	3	2398.46	799.48	4579.73	4.52	0.0038	2.38
Year×M566	3	1747.75	582.58	4569.59	3.90	0.0090	1.73
Error	**525**	**78440.60**	**149.41**				
Total	**731**	**252444.48**					

5. Discussion

For the description of the mean response of genotypes over environments and for studying and interpreting GEI in agricultural experiments, three classes of models are commonly used: (1) linear models; (2) bilinear models, and (3) linear-bilinear models. One class of linear models, namely factorial regression (FR) models, and one class of bilinear models, namely partial least square (PLS) regression, allow incorporation of external environmental and genotypic covariables directly into the model. In general, the AMMI biplot and the PLS biplot offer similar interpretations of the results. Interpretation of biplots is useful for researchers because it helps to identify major environmental (or cultivar) variables that cause positive or negative interactions between subsets of cultivars with subsets of environments. One advantage of the PLS approach is that a large number of environmental (or cultivar) covariables can be used. Furthermore, PLS is insensitive to multicollinearity. The main advantage of the FR is that parameters are estimated and hypotheses are tested in relation to the available external covariables. When environmental and cultivar covariables are considered simultaneously, multiple FR with a stepwise variable selection procedure provides a useful tool for selecting the most relevant covariables, and their cross products, for explaining GEI.

6. Conclusions, contribution of GEA-R

The AMMI, SREG, PLS and FR analyses complement each other and offer an aid to researchers not only for determining the importance of individual environmental and cultivar covariables in explaining GEI, but also for finding sub sets of covariables that adequately describe GEI in terms of understandable covariables.

GEA-R allows plant breeders to calculate and to interpret the GEI for comparing and selecting the best performing genotypes in specific environments using different strategies. This means that we can identify which genotypes within locations have more similarities or differences. Based on these methodologies, locations which, singly or in conjunction, provide effective screening for genotypes can be identified. The flexibility, power, and ease of use of this suite of programs make it a valuable instrument in the breeders' toolbox. GEA-R and its user's manual may be downloaded from CIMMYT's web page at no cost.

References

Aastveit, H., and H.Martens. 1986. ANOVA interactions interpreted by partial least squares regression. Biometrics 42:829–844.