sciendo

DOI: 10.2478/bile-2018-0008

Biometrical Letters
Vol. 55 (2018), No. 2, 97-121

AMMI and GGE Biplot for genotype × environment interaction: a medoid–based hierarchical cluster analysis approach for high–dimensional data

Anderson Cristiano Neisse[1], Jhessica Letícia Kirch[2], Kuang Hongyu[3]

[1] Master's Student in Statistics, Federal University of Viçosa, Brazil, e-mail: a.neisse@gmail.com
[2] Master's Student in Statistics, Luiz de Queiroz College of Agriculture, University of São Paulo, Brazil
[3] Department of Statistics, Federal University of Mato Grosso, Brazil

Summary

The presence of genotype-environment interaction (GEI) influences production making the selection of cultivars in a complex process. The two most used methods to analyze GEI and evaluate genotypes are AMMI and GGE Biplot, being used for the analysis of multi environment trials data (MET). Despite their different approaches, both models complement each other in order to strengthen decision making. However, both models are based on biplots, consequently, biplot-based interpretation doesn't scale well beyond two-dimensional plots, which happens whenever the first two components don't capture enough variation. This paper proposes an approach to such cases based on cluster analysis combined with the concept of medoids. It also applies AMMI and GGE Biplot to the adjusted data in order to compare both models. The data is provided by the International Maize and Wheat Improvement Center (CIMMYT) and comes from the 14th Semi-Arid Wheat Yield Trial (SAWYT), an experiment concerning 50 genotypes of spring bread wheat (Triticum aestivum) germplasm adapted to low rainfall. It was performed in 36 environments across 14 countries. The analysis provided 25 genotypes clusters and 6 environments clusters. Both models were equivalent for the data's evaluation, permitting increased reliability in the selection of superior cultivars and test environments.

Key words: genotype × environment interaction, adaptability and stability, additive main effects and multiplicative interaction model, multi-environment trials, cluster analysis, medoids

1. Introduction

Environmental conditions strongly influence agricultural production, leading to considerable variations in yield. Such influence is discriminated when yield experiments are performed in various locations and in different years (Pacheco et al., 2005; Akbarpour el al., 2014). Such influence is termed genotype-environment interaction (GEI). In the case of multi-environment trial (MET) data, GEI is frequently present. Due to the nature of this kind of data, it is often represented in two-way tables containing genotype means across all of the environments in the study (Rodrigues et. al., 2014; Hongyu et al., 2014).

MET studies are essential since the presence of GEI causes the relative performance ranking to change across environments, which complicates the evaluation of genotypes. Were it not for GEI, one single genotype would prevail in any environment, and it would take a single experiment to correctly choose the best genotypes (Gauch and Zobel, 1996; Hongyu et al., 2015). The key to significantly increasing agricultural production is to increase productivity per hectare and per dollar, which includes understanding and exploiting GEI as well as possible (Kang, 2002). In plant breeding and crop improvement, the main objectives of MET are: (i) to study GEI; (ii) to evaluate genotypic adaptability and stability; (iii) to establish relations between genotypes, environments and environment-genotypes simultaneously; and (iv) to predict the production of certain genotypes, enabling the precise selection of environments for subsequent cropping cycles (Gauch, 2013). Inefficient methods in genotype-environment interaction analysis can also represent a problem for breeders, who aim to select genotypes with superior performance in different environments (Hongyu et al., 2014).

Among various statistical techniques used for evaluating GEI, the two most frequently used are AMMI (Additive Main-effects and Multiplicative Interaction) and GGE Biplot. Several researchers have used the AMMI model as an effective method for analyses of GEI (Crossa, 1990; Annicchiarico, 1997; Gauch, 2006). Proposed by Gauch (1992), the AMMI model uses analysis of variance and principal component analysis to achieve a better understanding of GEI, its causes and consequences. Yan et al. (2000) proposed the GGE Biplot analysis, which considers both genotype main effects and GEI effects as important for the analysis (Miranda et al., 2009). The only difference between these models is in the initial steps of the analysis, where GGE analyzes G plus GE (or GEI) while AMMI separates G from GE; and at the final steps where the biplots for the interpretation

are built. Despite the possibility of their complementing each other due to their equivalent features, there has been discussion among authors about the effectiveness of AMMI and GGE in depicting the adaptive responses of genotypes over environments (Yan and Tinker, 2005; Gauch, 2006; Yan and Tinker, 2006; Yan et. al., 2007; Gauch et al., 2008). However, such differences do not imply the superiority of either of the methods. AMMI Biplot's graphic analysis provides relatively simple analysis for breeding researchers. Based on the data, it allows conclusions to be drawn concerning phenotypic stability, genotype behavior, genetic divergence between genotypes, and environments with optimal performance. As for GGE Biplot, it complements AMMI Biplot's environmental stratification, making it possible to delineate mega-environments and genotypes with optimal performance in such groups (Miranda et al., 2009).

Since both the AMMI and GGE approaches depend on principal component analysis (PCA), high-dimensional data may eventually become difficult to interpret visually in biplot analysis. In cases where too many components are needed to capture considerable proportions of the original variance, the researcher has to plot multiple biplots in order to be able to interpret enough of the original GEI variability. One approach that may facilitate interpretation is to apply clustering analysis on GEI in order to group genotypes and/or environments with similar genotype-environment interactions. With groups at hand, one option for a cluster representative which is free from possible outlier influence is the medoid, that is, the genotype that is most similar to each other genotype in the group on average (Xu and Wunsch, 2008).

This study's purpose was to apply AMMI and GGE analyses in order to depict GEI from 50 wheat genotypes in 36 environments across 13 countries. The measured variable was grain yield (t/ha). The data used comes from the 14^{th} Semi-Arid Wheat Trials, a two-replicate experimental study performed by the International Maize and Wheat Improvement Center (CIMMYT). To deal with the high dimensionality of the data, cluster analysis was applied based on GEI in order to obtain representative genotypes and environments.

2. Material and methods

The data used in the analysis has been provided by the International Maize and Wheat Improvement Center (CIMMYT) since 2015 in their online Research Data Repository. It concerns the 14^{th} Semi-Arid Wheat Yield Trial

(SAWYT), which is a two-duplicate multi-environmental trial performed during the 2006 cycle. The trials concerned spring bread wheat (*Triticum aestivum*) germplasm adapted to low rainfall. Various traits were measured from 50 different genotypes in 36 environments across 13 countries. The environments are drought-prone and typically receive less than 500 mm of rainfall during the cropping cycle. Figure 1 presents further information on the number of countries, their geographic distribution, and the number of locations in each.

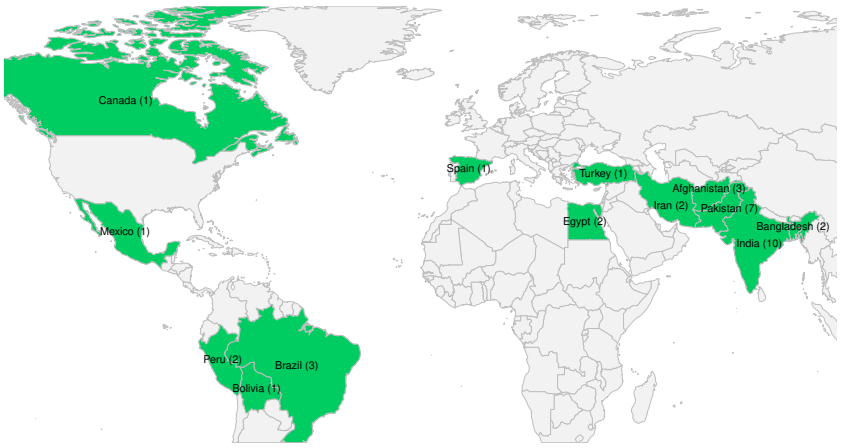

Figure 1. World map showing the geographical distribution of environments in the countries included in the 14^{th} Semi-Arid Wheat Trial

All of the analyses in this study concern only the yield weight, which was the only measured trait with replicates. Replication is a key characteristic for the AMMI model, making it possible to perform analysis of variance (ANOVA). The yield weight (t/ha) was primarily analyzed with simple ANOVA and conjoint analysis to assess the genotypic and environmental main effects as well as the GEI effects. Once GEI was evaluated as a significant effect present in the data, adaptability and phenotypic stability analyses were performed using the AMMI and GGE models. All of the analyses presented in this study were performed using R statistical software, version 3.4.1 (R DEVELOPMENT CORE TEAM, 2017).

2.1. Principal component analysis and biplot

Also called singular values decomposition (SVD), principal component analysis (PCA) was proposed by Pearson (1901) as a method to visualize a data

matrix. Each row can be geometrically visualized as a point in a space with as many dimensions (or axes) as there are columns (Gauch, 2006). PCA reduces the data dimensionality by deriving new axes, the principal components, which retain as much as possible from the original variation in a monotonic decreasing pattern, where the first component retains the most variation (Neisse and Hongyu, 2016). Any two-way data matrix $\boldsymbol{Z}$, with elements z_{ij} where $i = 1, ..., g$ represent the rows (or genotypes) and $j = 1, ..., e$ represent the columns (or environments), can be decomposed by SVD into p principal components (PC):

$$z_{ij} = \sum_{k=1}^{p} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij} \tag{1}$$

with $p \leqslant min(e, g-1)$. Every PC is composed by the genotype scores matrix α_{ik}, the environment scores matrix γ_{jk}, the singular value λ_k, and the residual ε_{ij} which is not explained by the model. The model restrictions are: (i) $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$, and (ii) α_{ik} scores are orthonormal, i.e. $\sum_{k=1}^{g} \alpha_{ik}\alpha_{ik'} = 1$ (if $k = k'$) and $\sum_{k=1}^{g} \alpha_{ik}\alpha_{ik'} = 1$ (if $k \neq k'$) with similar restrictions for γ_{jk} (Yan, 2011; Hongyu et al., 2015). Successive principal components are denoted by PC1, PC2, PC3 and so on.

In cases where the first few components capture a considerable proportion of the original variation, PCA provides a useful, low-dimensional data representation (Johnson and Wicher, ; Gauch, 2006). In such cases, they may be analyzed in graphical representations using biplots. This graphical analysis method was introduced by Gabriel (1971), and is useful for PCA because it represents simultaneously the rows and columns of a data matrix. Biplot graphical analysis allows the detection of groups in the observations, while also showing the dispersion and correlations between variables or columns (Hongyu et al., 2014; Gauch, 2006). Any two-way data matrix with rank r that can be approximated by a rank 2 matrix, i.e. the first two components are the ones which explain the most variation ($p = 2$ in equation 1), can be graphically interpreted in a two-dimensional biplot with a proper singular value partitioning:

$$z_{ij} = (\lambda_1^f \alpha_{i1})(\lambda_1^{1-f} \gamma_{j1}) + (\lambda_1^f \alpha_{i2})(\lambda_2^{1-f} \gamma_{j2}) + \varepsilon_{ij} \tag{2}$$

where f is the singular value partition factor (SVP). In the biplot, the abscissa and ordinate for the genotypes are $\lambda_1^f \alpha_{i1}$ and $\lambda_1^f \alpha_{i2}$ respectively, while for the environments the abscissa is $\lambda_1^{1-f} \gamma_{j1}$ and the ordinate is $\lambda_2^{1-f} \gamma_{j2}$

(Yan, 2011). The purpose of f is to re-dimension the scores for better visual interpretation of the biplot. In the MET data context, $f = 1$ results in the allocation of the singular values entirely in the genotype scores (genotype-centered or $SVP = 1$), $f = 1$ allocates them in the environment scores (environment-centered or $SVP = 2$), and $f = 0.5$ will allocate the singular values' square roots for both genotype and environment scores (symmetric or $SVP = 3$) (Hongyu et al., 2015). In the case of GGE Biplot analysis, the genotype-centered and the environment-centered SVP are used for the evaluation of genotypes and environments respectively (Yan, 2011).

2.2. AMMI model

Introduced by Gauch (1992), the additive main effects (G and E) and multiplicative interaction (GE) model, or AMMI model, combines ANOVA and PCA in a single model. In the case of AMMI analysis, principal component analysis is applied to the GEI effects only after some preliminary verifications are made based on ANOVA analysis. These verifications are based on three numbers: the sums of squares (SS) for genotypes (G), GEI effect signal (GEI_S) and GEI noise (GEI_N) (Hongyu et al., 2014). The sum of squares from G and GEI can be easily obtained from the ANOVA analysis, while GEI_S and GEI_S are obtained from GEI. The GEI_S SS is obtained by simply multiplying the mean square error (MSE) by the degrees of freedom for GEI, then GEI_N is obtained by subtracting GEIs from GEI (Gauch, 1992; Gauch, 2013). Occasionally GEI is buried in noise, then the SS for GEI_N will be approximately equal to that of GEI; in such cases it is not appropriate to apply AMMI analysis to the data (Gauch, 2013). The statistical model equation for the i^{th} genotype in the j^{th} environment in r blocks or replications is:

$$Y_{ijr} = \mu + g_i + e_j + b_r(e_j) + \sum_{k=1}^{n} \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij} + \varepsilon_{ij} \tag{3}$$

where Y_{ijr} is the phenotypic trait (e.g. yield) of genotype i in environment j for replicate r, μ is the grand mean, g_i are the genotype main effects as deviations from μ, e_j are the environment main effects as deviations from μ, λ_k is the eigenvalue for the interaction Principal Component (PC) axis k, α_{ik} and γ_{jk} are the genotype and environment PC scores (i.e. the left and right eigenvectors) for axis k, $b_r(e_j)$ is the effect of the replication r within the environment j, r is the number of replications, ρ_{ij} is the residual containing

all multiplicative terms not included in the model (1), n is the number of axes or principal components (PC) retained by the model, and ε_{ijr} are the experimental errors, assumed independent with identical distribution, $\varepsilon_{ij} \sim N(0, \frac{\sigma^2}{r})$ (Gauch, 1992; Hongyu et al., 2014).

First the means matrix $\boldsymbol{Y}_{g\times e}$ is generated for the g genotypes and e environments across all of the replicates, then a traditional ANOVA is fitted to it. The residual from the fitted ANOVA is the GEI effect, which represents the multiplicative part of the AMMI model. Based on the GEI, PCA is then applied to the interaction matrix $\boldsymbol{GE}_{g\times e} = [(ge)_{ij}]$, which is obtained by the equation:

$$(ge)_{ij} = Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{\cdot\cdot} \tag{4}$$

where Y_{ij} is the mean across replications for the genotype $i = 1, ..., g$ in the environment $j = 1, ..., e$; $\overline{Y}_{i\cdot}$ is the mean of genotype i across all environments and replications; $\overline{Y}_{\cdot j}$ is the mean of environment j across all genotypes and replications; and $\overline{Y}_{\cdot\cdot}$ is the global mean. Since $\overline{Y}_{\cdot\cdot}$ is subtracted twice among $\overline{Y}_{i\cdot}$ and $\overline{Y}_{\cdot j}$, it needs to be added back once (Gauch, 1992; Pacheco et al., 2005; Hongyu et al., 2014).

An efficient technique to assign degrees of freedom to the AMMI model has similar results to traditional ANOVA and unfolds degrees of freedom corresponding to the GEI. It determines the degrees of freedom for each part of the interaction sum of squares SS_{GEI} (or λ_k^2). Subsequently, the F test is applied to each AMMI component to assess its significance relative to $MS_{meanerror}$. Therefore, the number of axes to correctly explain the interaction's behavior can be determined based on the proportion of SS_{GEI} accumulated up to the n^{th} axis ($\sum_{k=1}^{n} \lambda_k^2 / SS_{GEI}$) (Hongyu et al., 2014). The number k of components to consider in the analysis is based on the F significance test for each of the successive interaction terms; Gollob's F (Gollob, 1968) is one of the most common methods used in the AMMI model. This method uses the expression $DF_{PC_k} = g + e - 1 - 2k$, where $k = 1, ..., p$, $p = min(g - 1, e - 1)$ and PC_k is the model's k^{th} axis or PC. Since AMMI is rather a family than a single model, the model's name is denoted by the number of components it contains, for example, AMMI0 for zero components and AMMIF for all components (Gauch, 1992; Akbarpour el al., 2014).

It is common in the literature analyses for PC1 and PC2 to capture a significant proportion of the GEI total variation; consequently AMMI1 and AMMI2 are the most frequently used models to interpret genotype×envi-

ronment interaction. The AMMI1 model is the most well-known AMMI model; its abscissa represents the main effects (G and E), and its ordinate represents the PC1 scores. Such a biplot allows the researcher to evaluate genotypes and environments in terms of their mean production, and also to obtain a first look at their stability in terms of GEI with PC1. As for AMMI2, its abscissa represents the PC1 scores and its ordinate the PC2 scores. In this way the researcher may evaluate genotypes in terms of their stability and specific adaptability to environments, and vice versa. AMMI2 is also useful for the delineation of mega-environments, that is, groups of environments that have the same genotype as most productive (Hongyu et al., 2014).

2.3. GGE model

The GGE Biplot model (Yan et. al., 2000) was introduced based on biplots, which are an effective tool for visualizing two-way data, and are frequently used for the analysis of MET data. A GGE biplot is able to simultaneously display genotype main effects (G) and genotype × environment effects (GE) from a two-way data table (Yan et. al., 2000). Its first component, when highly correlated with genotype main effect (G), represents the proportion of production solely attributed to the genotype. The second represents the proportion explained by GEI. For the GGE Biplot to be generated, the mean matrix must be environment-centered and then decomposed into principal components by SVD; the first two PCs are then used to generate the graphic (Yan and Tinker, 2005). The GGE biplot is based on the model:

$$Y_{ij} - \overline{Y}_{\cdot j} = \lambda_1 \xi_{i1} \eta_{j1} + \lambda_2 \xi_{i2} \eta_{j2} + \varepsilon_{ij} \tag{5}$$

where Y_{ij} is the mean across replications for genotype i ($i = 1, ..., g$) in environment j ($j = 1, ..., e$); $\overline{Y}_{\cdot j}$ is the mean of environment j across all genotypes and replications; $\lambda_1 \xi_{i1} \eta_{j1}$ and $\lambda_2 \xi_{i2} \eta_{j2}$ are PC1 and PC2 respectively; λ_1 and λ_2 are the eigenvalues associated with each PC; ξ_{i1} and ξ_{i2} are the PC's scores in the i^{th} genotype; η_{j1} and η_{j2} are the scores for each PC in the j^{th} environment, and ε_{ij} is the error associated with the model (Yan et. al., 2000; Miranda et al., 2009).

A method to evaluate the quality of the fit of the GGE biplot to the data was proposed by Yan and Tinker (2006), called the "information relation" (IR). Imagine a two-way data matrix with g genotypes and e environments; the maximum number of PCs required to completely represent such a table is $k = min(e, g - 1)$. If environments are uncorrelated, each

PC should be independent and explain a proportion of exactly $1/k$ of the original variation. If environments are correlated, however, the proportion of the variation explained by the first components should be greater than $1/k$, while the proportion explained by the last components should be less than $1/k$. The calculation of IR is simple: it is obtained by multiplying k by the proportion of the variation explained by each PC. Any PC with IR $\geqslant 1$ expresses correlation between environments. Thus, if the first two components from a PCA have IR $\geqslant 1$, then the two-dimensional biplot represents the data properly.

Given the GGE Biplot definition, its main difference compared with the AMMI model becomes clear: while the AMMI model applies SVD only to GE effects, GGE Biplot considers G and GE together, additively. This difference is the main topic of discussion among authors concerned with the effectiveness of the two models in depicting adaptive responses of genotypes over environments (Yan and Tinker, 2005; Gauch, 2006; Yan and Tinker, 2006; Yan et. al., 2007; Gauch et al., 2008). The differences are not such as to make either of the methods superior. Nevertheless, GGE Biplot presents some features based on the presence of G in the analysis, which naturally the AMMI model does not offer.

2.4. Mega-environment delineation and agricultural recommendations

Whenever different genotypes are adapted to different environment groups and the variation between groups is greater than the variation within groups, those environment groups are called mega-environments (Hongyu et al., 2015). When there are crossovers between winning genotypes, the subdivision of a region into two or more mega-environments is necessary so that the researcher can exploit the narrow adaptation, gaining substantial opportunities to increase yield. However, there are three considerations to be made (Gauch, 2013): (i) In order to select the best model to represent the data, a proper model diagnosis must be performed, because as the order of the selected model grows the number of mega-environments tends to grow also; (ii) It is important for mega-environments to have predictive potential for locations and years, and this role is greatly enhanced if the mega-environments have an evidential and environmental interpretation, beyond the delineation of winning genotypes; (iii) With several mega-environments, the process is costly for breeders, unless a practical portion of GE becomes available for

exploring narrow adaptations to increase yield, so it is necessary to select low-order models in order to delineate a small and manageable number of mega-environments. Fortunately, just two or three mega-environments are often sufficient to allow GE to capture a sizeable portion of the interaction signal. (Gauch, 2013; Hongyu et al., 2015).

It is a major purpose of MET data studies on yield to select the best genotypes for use or recommendation for certain regions. This task is remarkably difficult though – true winners are often obscured by noise and generate improper complexity, reducing genetic gains. Pursuing both high yield and stability is a frequent approach for genotype selection. However, this method has five considerable problems (Gauch, 1992): (i) any particular choice is difficult to justify since there are many stability parameters (Annicchiarico, 2002; Gauch, 2013); (ii) many ways to integrate high yield and stability fail to optimize known and agriculturally significant outcomes; (iii) stability is a meaningful objective only within an individual mega-environment, a requirement frequently ignored by the literature (Gauch, 2013); (iv) it is a requirement to have at least 8 trials in each mega-environment for reasonably reliable estimates of stability (Annicchiarico, 2002; Gauch, 2013); and finally (v) it is a defective paradigm to consider stability solely as a problem to be minimized, since instability in fact presents breeders with both problems and opportunities.

The purpose of mega-environment analysis is to divide the studied region into significant subregions in a way that makes it possible to explore GEI. When a two-dimensional GGE biplot is considered as a significant representation for the data, it is also a tool for the delineation of mega-environments, also called the "which won where" plot (Yan, 2011; Hongyu et al., 2015). In the GGE Biplot analysis, when delineating mega-environments, the mean presented in the graphic is related to the mega-environment mean, not to the grand mean, which helps to identify genotypes with broad or specific adaptations to some environments or groups of environments (Yan and Kang, 2003). The "which won where" plot point of view is built by an irregular polygon and as many lines as there are sides in the polygon, with these lines starting at the biplot's origin and intercepting the polygon perpendicularly. The polygon's vertices mark genotypes that are further from the origin in all directions; thus all genotypes are inside the polygon (Yan, 2011). A hypothetical environment is represented by a line that perpendicularly crosses a side of the polygon, if both genotypes that formed that side have good levels of productivity, the genotype's relative rank would be inverted

in environments at the line's opposite extreme (crossed GE). Thus, the lines radiating from the origin divide the biplot into sections, and there is a vertex (genotype) for each section which had the best yield performance in environments contained in that section, which is called a mega-environment.

2.5. Hierarchical cluster analysis

When analyzing MET data, when ANOVA analysis shows the existence of significant GEI to be studied, there may be some complications when PCA is applied. When the data is high-dimensional not only by genotypes but also by environments, the researcher might need more than three principal components to explain a significant amount of the variation. In the case of biplot analysis, complications arise due to the visual limitation. For instance, the data presented in this study has 50 genotypes and 36 environments, so it would take 12 principal components to explain 82.9% of the original GEI variability; consequently a proper analysis would require at least six two-dimensional biplots. This study proposes a method to reduce the data dimensionality by forming groups of genotypes and environments that are similar in terms of GEI. Based on the obtained groups, representative genotypes and environments will be selected based on medoids to take the place of their groups in the analysis.
Clustering methods are generally classified as partitional clustering or hierarchical clustering, depending on the properties of the generated clusters. Partitional clustering starts from a pre-specified number of clusters. Hierarchical clustering starts with every observation as its own cluster and performs a sequence of nested clusterings (agglomerative hierarchical clustering) or else starts with a unique cluster containing all of the observations (divisive hierarchical clustering). Hierarchical clustering generates a binary tree or a dendrogram which depicts all of the nested clustering steps (Xu and Wunsch, 2008). In this way the number of clusters is defined by cutting the dendrogram at some level or proximity height. There are methods that may provide a good guess at the right number of groups, for example, evaluating the sum of squares within groups (SS_w) at each step and picking the number of groups where SS_w stabilizes. However, the researcher may make a decision based on empirical information or numerical criteria, or just for the sake of simplicity. The researcher may also have practical reasons to establish a certain number of clusters based on their intended use. A frequent approach is to pick the number of clusters based on the

researcher's previous knowledge concerning the similarity relation between the data being studied (Barroso, 2003).

Both partitional and hierarchical clustering methods perform clustering based on a distance measure matrix. There are many distance measures for all kinds of clustering problems; examples include the Mahalanobis distance (Mahalanobis P.C., 1939), the Euclidean distance and the Manhattan distance. It is also a common approach to use the correlation distance in the case of pattern recognition in microarray gene expression data (Datta and Datta, 2003). In this study the correlation distance between two genotypes (or environments) x and y was obtained by $d(x, y) = 1 - corr(x, y)$ where $corr(x, y)$ is the statistical correlation between the mean production of the genotype in each environment, or vice-versa. Thus genotypes (or environments) that are positively correlated will be considered similar, having lower distances than uncorrelated or negatively correlated objects. Since the aim of this study is to analyze GEI, the distance metric based on the correlation was applied in the interaction matrix 4, which does not consider genotypic and environmental main effects.

With groups of genotypes and environments at hand, a method for picking representatives is needed. There are two most common methods: the centroid method, which is used by the k-means clustering methodology (Hartigan and Wong, 1979); and the medoid method used in the PAM (partitioning around medoids) algorithm (Kaufman and Rousseeuw, 1990). The method for obtaining a cluster's centroid consists in taking the mean production of all genotypes (or environments) in the cluster. However, this method has two important disadvantages that should be accounted for: (i) it is susceptible to outliers, which distort the analysis; and (ii) it is a prototype representation, which means that it is not a real member picked from the cluster. The medoid method overcomes such disadvantages, as the medoid is in fact a member of the cluster and consequently cannot be affected by outliers, nor can it be an outlier itself (Xu and Wunsch, 2008). Considering $\boldsymbol{D}$ as a distance matrix, c as a specific cluster and $\boldsymbol{D}^c$ as the cluster's distance matrix, then $medoid^c = min(\boldsymbol{S}^c)$ with $S_j^c = sum_{i=1}^n D_{ij}^c$ is the medoid for the cluster c. Imagining the cluster as a cloud of points in the space, the medoid would be the point closest to the center with the lowest mean distance to all of the other points.

3. Results and discussion

3.1. Cluster analysis

The original data provided by the 14^{th} Semi-Arid Wheat Trials comes from a MET with 50 genotypes adapted to low rainfall environments tested in 14 different countries, totaling 36 environments. A first analysis with the AMMI model revealed significant main effects and interaction effects. However, when PCA was performed, using both AMMI and GGE models, the outcome consisted of 35 principal components, of which the first 12 were significant, explaining 82.9% of the original variability. The first two components retained only 30% of the original variability, insufficient by far for a reliable interpretation. Considering only one Biplot model, it would take at least six graphs conjointly analyzed to explain a significant enough amount of variability to allow reliable decision-making.

This being the case, a cluster analysis was applied to reduce the genotypes and environments into groups. The agglomerative hierarchical clustering method was applied, considering the correlation distance matrix obtained from the genotype-environment interaction matrix 4. The cluster analysis was applied in genotypes and environments independently; that is, both steps considered the original data. With the groups at hand, one medoid was chosen for each group to be the cluster's representative in the analysis. Figures 2 and 3 present two dendrograms for genotypes and environments respectively, showing the clustering steps for different distances.

Figure 2 shows blue boxes representing the 25 genotype clusters formed, with blue marks for each cluster's medoid genotype. The number of clusters for the genotypes is chosen only for the sake of allowing visual interpretation. The number of 25 clusters was chosen so that the average number of genotypes by clusters was 2 – that is, on average, pairs of the most similar genotypes would be grouped in order to maintain as many genotypes as possible while gaining advantages in terms of visual analysis.

The red boxes in Figure 3 delineate the six chosen environmental clusters, while the medoid environments are marked with a red label. It can be seen from the dendrogram that a number of 5, 6 or 7 clusters would be a reasonable choice given the clustering heights. However, since the maximum number of principal components is $min(e, g-1)$ and the AMMI model requires a greater number of genotypes than environments, the criterion was to choose as many clusters as possible while maintaining more than 70% variability in the first two principal components. The whole of the AMMI

and GGE analysis considers the group representatives obtained by cluster analysis.

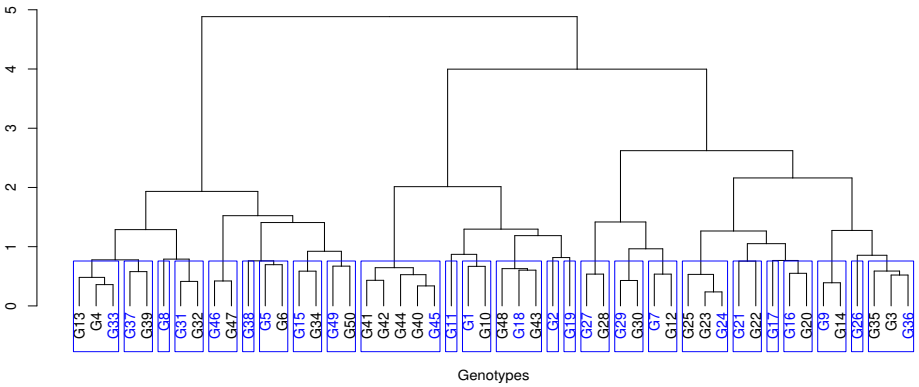

Figure 2. Agglomerative hierarchical clusters and respective medoids for 50 genotypes

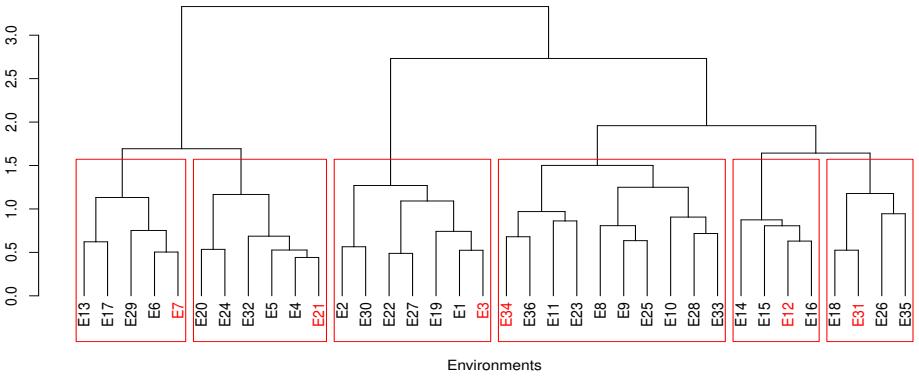

Figure 3. Agglomerative hierarchical clusters and respective medoids for 36 environments

3.2. AMMI model analysis

The results provided by the AMMI conjoint analysis of variance for the yield (t/ha) in the 14^{th} Semi-Arid Wheat Yield Trials showed the presence of significant genotypic and environmental main effects as well as interaction effects ($p<0.001$). The coefficient of variation for the data was low (14.3645%), indicating correct experimental design and performance. It is also important to note that the levels of significance and experimental precision match those from the ANOVA fitted to the original data. Such results

allow further investigation based on the genotype-environment interaction (GEI) effects for the 14^{th} SAWYT data.

The interaction sum of squares ($SS_{GE} = 81.91$) obtained by ANOVA in the AMMI analysis (Table 1) corresponds to the sum of all eigenvalues ($\sum_{k=1}^{n} \lambda_k^2$). However, SS_{GE} may be inflated because of the possible presence of considerable noise. To remove noise, five principal components (PC) were fitted to the interaction matrix $\boldsymbol{GE}$. The first principal component (PC1) turned out to be significant with p<0.001 according to Gollob's F test, as did the second (PC2). While PC1 explained 51.4% of the variability, the proportion attributed to PC2 was 22.1%. PC1 and PC2 together explain 73.4% of the variability, which is sufficient, since 70% is considered the minimum amount of variability for the model to be relatively reliable. Despite the fact that PC3 was also significant (p<0.05), explaining 14.0% of the variability and bringing the cumulative total up to 87.4%, it was omitted from further analysis so that the simplicity of two-dimensional analysis would be maintained. Figure 4 presents the AMMI1 model, with yield on the abscissa and PC1 scores for genotypes and environments on the ordinate. In the figures, genotypes are denoted by G1 to G50 and environments by E1 to E36, bearing in mind that only the medoid representatives for each cluster are shown.

Table 1. Conjoint analysis of variance of the wheat trial productivity (t/ha) and GEI sum of squares decomposition

Source	DF	SS	MS	F	p-value
Environment (E)	5	552.22	110.4440	270.50	<0.001***
Replicate/Environment	6	2.45	0.4083	1.49	0.1860^{NS}
Genotype (G)	24	16.46	0.6858	2.50	<0.001***
Interaction (GE)	120	81.91	0.6826	2.49	<0.001***
PC1	28	42.07	1.5025	5.48	<0.001***
PC2	26	18.07	0.6950	2.53	<0.001***
PC3	24	11.46	0.4775	1.74	0.0250*
PC4	22	8.24	0.3745	1.36	0.1443^{NS}
PC5	20	2.07	0.1035	0.38	0.9927^{NS}
Residual	144	39.51	0.2744	-	-
Total	299	692.55	-	-	-
General Mean	3.6467				
CV (%)	14.3645				

It is an objective of SAWYT to study genotypes adapted to low rainfall in environments receiving typically less than 500 mm of rain during the cropping cycle; this may explain the agglomeration of genotypes around

the mean in Figure 4. The genotypes with the highest average yield were G27 > G8 > G29 > G18 > G21 > G7 > G45 > G31 > G1, in decreasing order (Figure 4) and from this group, those with the most stability were G8, G45, G21, and G7, considering both axes' proportion of explained variability (Figure 5). The groups of genotypes with yield fairly close to the mean, i.e., with a maximum distance relative to the mean of 3%, were G11 > G2 > G9 > G36 > G37 > G38 > G15 > G26 > G19, while the groups of genotypes with yield below the mean were G49 > G46 > G24 > G16 > G5 > G17 > G33, both in decreasing order. In the average yield genotype group, those with the most stability were G19, G15, G26, G2, G36 and G11. In the low production group, the only genotype with stability was G49.

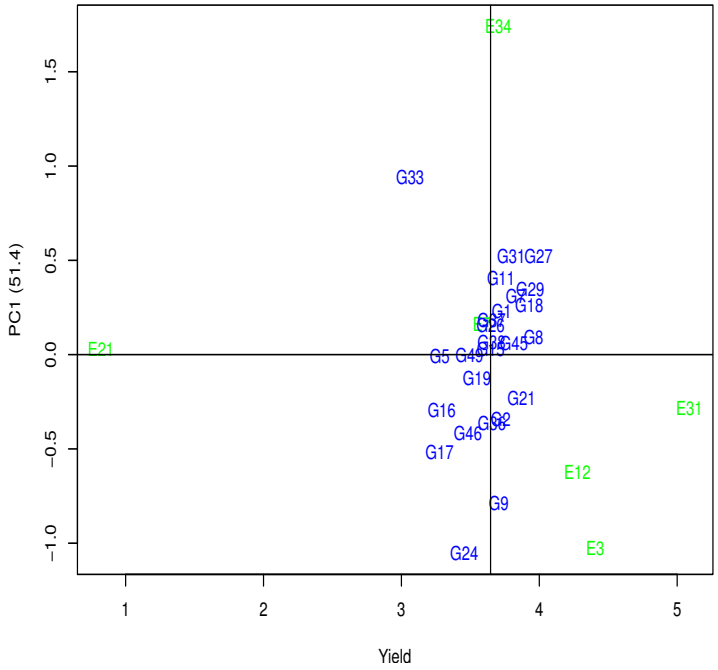

Figure 4. Biplot AMMI1 (Mean Yield vs PC1) for the wheat trials data with 25 genotypes (G) and 6 environments (E)

It is important to note that despite being separated into three groups in terms of their mean yield, all of the genotypes fall relatively close to the mean. For instance, the two genotypes with the most stability were placed in two different groups despite being very similar in terms of mean yield (G49=3.5475 and G19=3.4939) and falling relatively close to the grand

mean (3.6467). Considering the fact that PC1 explains most of the variability (51.4%), the genotypes that exhibited both stability and high production on average were G8, G45, G21 and G7 (Figure 4).

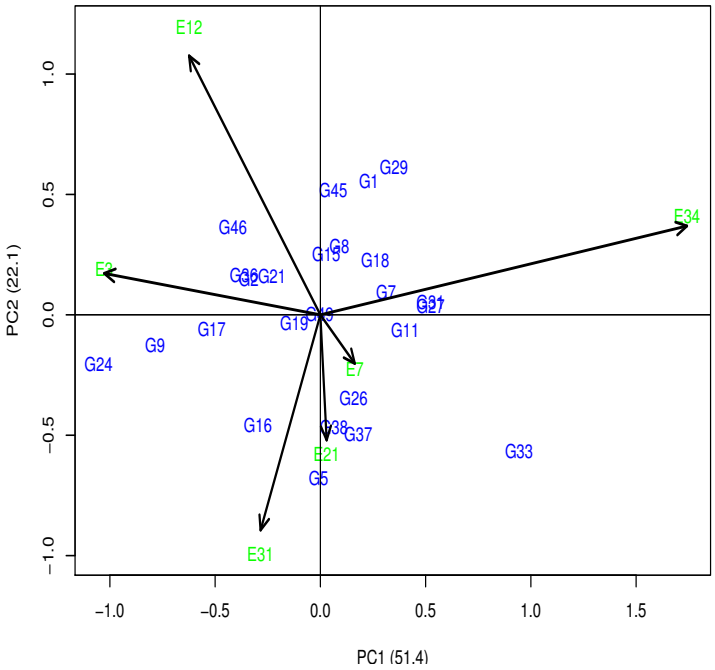

Figure 5. AMMI2 biplot (PC1 vs PC2) for the wheat trials yield data (t/ha) with 25 genotypes (G) and 6 environments (E)

The environments with the highest productivity on average were E31 (Shesham Bagh, Afghanistan) followed by E3 (Wheat Research Institute, Pakistan) and E12 (Dwr-Karnal, India). An environment with mean yield was E34 (Dinajpur, Bangladesh) followed closely by E7 (Londrina, Brazil), while an environment positioned way below average was E21 (Quetta Ari Sariab, Pakistan). Environment E7 was also shown to be the most stable. There were no groups formed according to Figures 4 and Figure 5, which is a result of the clustering analysis applied to delineate groups of environments.

As for specific adaptations, Figure 5 showed G33 with high yield performance in environments E34, E21 and E7 while performing poorly in E12 and E3. Environment 21 had genotypes G5, G38 and G37 with high performance due to specific adaptation; G5 also adapted well in E31, as did G16. The environment E3 had G24 as the genotype with the highest per-

formance due to specific adaptation, followed by G9, but these genotypes had low performance in E34. The genotypes with the highest yield performance for E34 were G31 and G27. Genotype G46 had better performance in environments E3 and E12 while performing poorly in other environments. Genotype G29 had the best performance in E12 and E34 in comparison with other environments, followed by G1 and G45; these three genotypes had a low yield in environments E7, E21 and E31.

In terms of model fitting, the work of Hongyu et al. (2014) led to similar results: two principal components were able to retain 71.2% of the original variability from 9 genotypes in 20 different environments with the AMMI2 model. According to Gauch et al. (2008), the fundamental reason AMMI is appropriate for agriculture is that the ANOVA step of the analysis can effectively separate G and E main effects from GE interaction, which presents researchers with many problems and opportunities.

3.3. GGE Biplot analysis

From the six principal components fitted by the GGE Biplot analysis, according to the information relation (IR), only PC1 (IR=2.6394) and PC2 (IR=1.6926) retained patterns in the data. This shows that the first two principal components were adequate for visually representing the data. The biplot graph is shown in Figure 6, where the abscissa represents PC1 and the ordinate represents values of PC2. Similarly to the AMMI plots, genotypes are denoted by G1 to G50 and environments by E1 to E36, bearing in mind that only the medoid representatives for each group are shown.

Table 2. Explained proportion and information relation (IR) for six GGE principal components

PC	Explained Variability (%)	IR
1	43.99	2.6394
2	28.21	1.6926
3	12.43	0.7458
4	8.85	0.5310
5	4.62	0.2772
6	1.9	0.1140

The GGE biplot in Figure 6 was built with environment-centered data (centering = 2), not scaled (scale = 0), and with column singular value partitioning (SVP = 2). The "which won where" GGE biplot in Figure 6

allowed the visual grouping of environments based on crossed GEI between the highest yield genotypes. The polygon's vertices comprised genotypes G33, G16, G9, G45, G29, G27. The seven environments were grouped into 5 mega-environments formed by: (i) E21, (ii) E31 and E3, (iii) E12, (iv) E9, (v) E34 and E7. The genotype G33 was the vertex in the mega-environment formed by E21, which means that this genotype had the highest yield in E21. Similarly, G27 was the best genotype in terms of yield in the mega-environment formed by E34 and E7 (Figure 6).

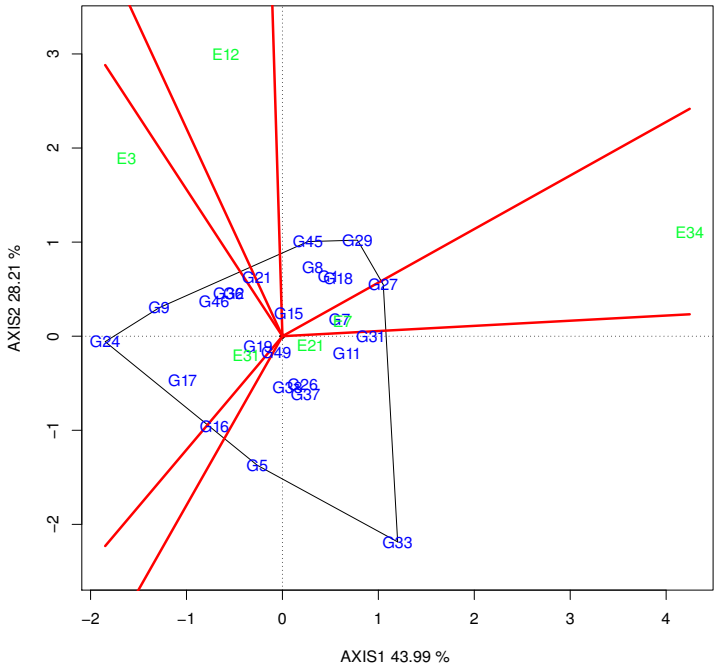

Figure 6. "Which won where" GGE biplot for the wheat trials yield data (t/ha) for the delineation of mega-environments

There were two cases of mega-environments with two vertices; however, there was still a better genotype in each environment. In the mega-environment formed by E9 the two genotypes with the best performance were G29 and G45, in decreasing order. In the E31 and E3 mega-environment the best genotypes were G24 and G9. There were also genotypes in regions with no environment at all, which means that such genotypes had a poor performance in all environments, as was the case with G16, and even G21 to a certain degree (Figure 6).

The GGE biplots in Figures 7 and 8 were built with environment-centered data (centering = 2), not scaled (scale = 0), with column singular value partitioning (SVP = 1). An ideotype is the ideal genotype for a certain environment or cropping objective; the ideotype thus combines high mean yield and stability in a mega-environment. The "Mean vs Stability" GGE biplot (Figure 7) allows the efficient evaluation of genotypes by both characteristics. The small green circle in Figure 7 represents the "mean-environment", which is an environment built on the coordinate means for all environments in the analysis. The green line in Figure 7 with the arrow passing through the origin represents the "mean-environment axis" and the direction in which the arrows point represents higher mean yield for the genotypes. The second axis represents stability; genotypes that are closer to the origin are more stable (Yan, 2011; Hongyu et al., 2015).

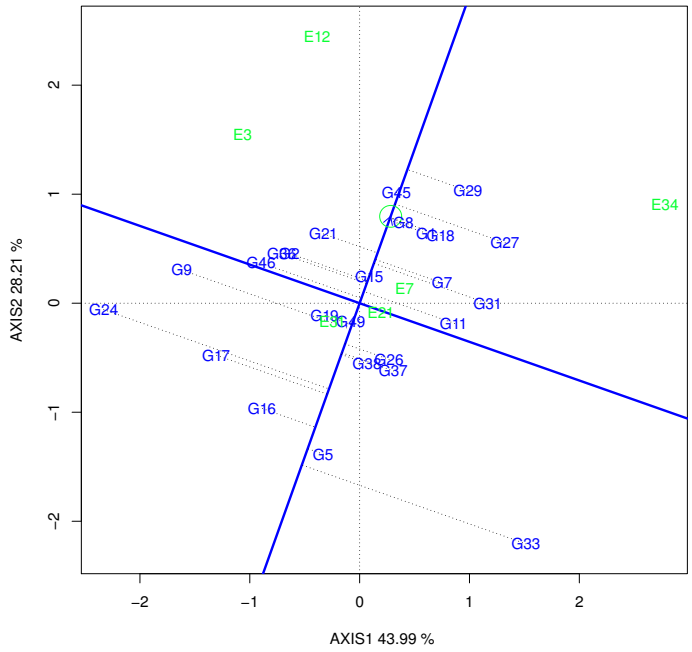

Figure 7."Mean vs Stability" GGE biplot for the wheat trials yield data (t/ha) with 25 genotypes (G) and 6 environments (E)

In terms of mean yield, the genotypes classification is G29 > G45 > G27 > G1 > G8 > G18 > ... > G11 > G15 > G46 > grand mean > G49 > G9 > G19 > ... > G16 > G5 > G33 (Figure 7). Among the unstable genotypes, G33 had the highest instability, due to its high performance in

E21 and low yield in other environments; other unstable genotypes exhibited similar behavior. Although not the most stable, G29 and G27 were among the most productive genotypes. There were genotypes with high stability and yield close to the grand mean, which was the case for G15, G49 and G19. It is important to note that the biplot represents only a fraction of the total variation; it is possible to wrongly evaluate a genotype as stable if its variability is not significantly retained by both principal components. According to Figure 7 the genotypes that are closest to the definition of ideotype for the analyzed data were G45, G8 and G1, being among the most productive (second, fifth and fourth places respectively) and highly stable.

The plot in Figure 8 enables evaluation of the test environments, to identify environments that may serve to select superior genotypes in an efficient way for a mega-environment. The selected test environment should have high genotype discrimitiveness and representativeness. Environments with shorter vectors have less discrimitiveness in relation to genotypes, i.e., all genotypes tend to perform equally and almost no information about genotypic differences can be revealed by such environments. A short vector could also mean that PC1 and PC2 do not represent that environment very well in cases where G + GE has not been retained properly. The environments E34, E12, and E3 presented long vectors, which means they have high discrimitiveness for the genotypes. It is also possible, by Figure 8, to identify environments with high representativeness: the smaller is the angle between an environment and the mean-environment axis (blue axis), the higher is its representativeness. An environment that has both characteristics more than the others is E12. Environments E3 and E34 have long vectors but greater angles, implying that they should not be recommended (Figure 8).

The GGE Biplot analysis used in this study included mega-environment analysis, genotype evaluation and test-environment evaluation, all of the three aspects that should be addressed according to Yan et al. (2007). As stated by Yan et al. (2000), the first principal component in the graphical analysis represents genotypic productivity, while the second represents genotypic stability. However, such properties tend to appear only in cases where the first principal component is highly correlated to the genotype effects. Also it has to be considered that, as mentioned by Yan et al. (2007), G and GE are first of all mathematical definitions. There is little evidence that G and GE are controlled by different genes and can thereby be separately selected.

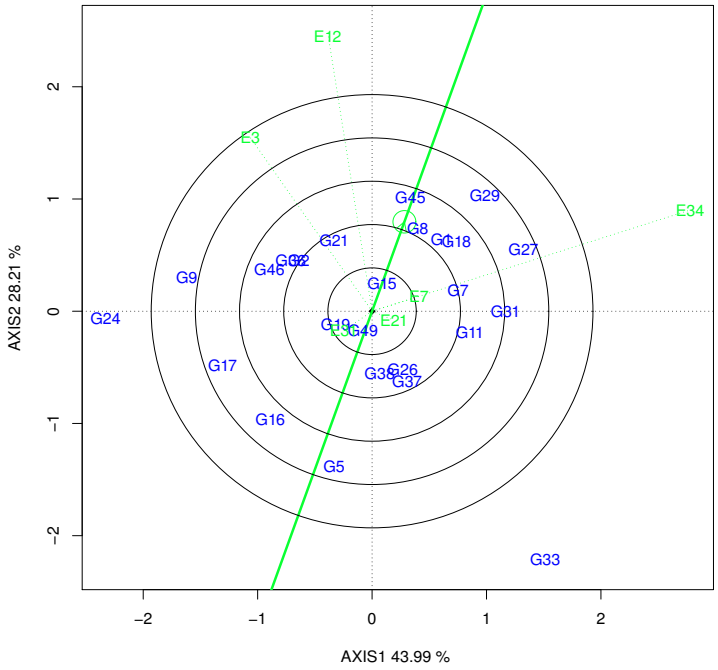

Figure 8."Discrimitiveness vs Representativeness" GGE biplot for the environments in the study

3.4. Comparison of AMMI and GGE Biplot analyses

The proportion of explained variability proved to be highly correlated when compared between AMMI and GGE Biplot analyses ($r = 0.9706$). According to Camargo-Buitrago et al. (2011), such high correlation indicates that both models performed equivalently in the MET data analysis. AMMI retained 51.40% of the total variation in PC1 while GGE retained 41.32%; the sum of the total variation retained in PC1 and PC2 was 73.50% for AMMI and 68.06% for GGE. Considering this, both models explained very similar amounts of variation; however, AMMI still retained a greater proportion. Nevertheless, GGE Biplot made it possible to analyze the MET data by a different approach and to confirm some interpretations, thus enhancing the reliability of AMMI.

Similar amounts of total variability explained by AMMI and GGE, despite AMMI explaining a slightly greater amount, were also reported by Hongyu et al. (2015). As was shown by Gauch et al. (2008), GGE Biplot alone would not be of interest to crop and soil scientists since it ignores envi-

ronmental main effects, which are considered in the AMMI model. However, Yan et al. (2007) stated that GGE Biplot serves an additional purpose not provided by AMMI analysis: the evaluation of test environments. The work of Hongyu et al. (2015) also compared AMMI and GGE analysis in a similar way to that presented in this study, showing the advantages of including both models in the analysis to exploit their strengths.

4. Conclusion

Cluster analysis has been shown to be efficient in delineating groups with similar genotypes and environments, as is implied by the fact that most environments formed a mega-environment on their own. Its effectiveness is also reflected in the fact that the levels of significance presented in the ANOVA of the reduced data remained the same as in the ANOVA of the original data. The interpretations for the genotypes and environments could thus be extended to their respective groups, always considering the degree of similarity based on their distance matrix.

According to the results, both AMMI and GGE were able to efficiently explore the variability present in MET data due to genotype-environment interaction. Also, by complementing each other, it was possible to gain reliability in the analysis, where there is a danger that some genotype may be wrongly identified as stable due to the lower than 100% proportion of variability explained by both models. With the 14^{th} Semi-Arid Wheat Yield Trials data, AMMI (73.5%) explained more variability than GGE (68.06%) in the first two components. However both models proved to be approximately equivalent, leading to substantially the same conclusions about the genotypes with the highest yield and stability. Further, the researcherâĂŹs attention is drawn to cases where certain genotypes and environments lead to different conclusions in each model. Both AMMI and GGE agreed that G8 and G45 were the genotypes closest to the definition of ideotype.

References

Annicchiarico P. (1997): Additive Main Effects and Multiplicative Interaction (AMMI) Analysis of Genotype-location Interaction in Variety Trials Repeated over Years. Teor. Appl. Genet. 94: 1072-1077.

Annicchiarico P. (2002): Genotype × environment interaction: Challenges and opportunities for plant breeding and cultivar recommendations. Food & Agriculture Org 174.

Akbarpour O., Dehghani H., Sorkhi B., Gauch Jr. H.G. (2014): Evaluation of Genotype × Environment Interaction in Barley (*Hordeum Vulgare L.*) Based on AMMI model Using Developed SAS Program. J. Agr. Sci. Tech. 16: 909-920.

Barroso L.P. (2003): Análise Multivariada. Lavras: UFLA, 151p.

Camargo-Buitrago I., Intire E.Q.M., Gorddón-Mendoza R., (2011): Identificación de mega-ambientes para potenciar el uso de genótipos superiores de arroz em Panamá. Pesquisa Agropecuária Brasileira 46(9): 1061-1069.

Crossa J. (1990): Statistical Analyses of Multilocation Trials. Adv. Agron. 44: 55-85.

Datta S., Datta S. (2003): Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19(4): 459-466.

Gabriel K.R. (1971): The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3): 453-467.

Gauch H.G. (1992): Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, Amsterdam.

Gauch H.G., Zobel R.W. (1996): AMMI analysis in yield trials. KANG, M. S., GAUCH, H. G. (Ed) Genotype by environment interaction. New York: CRC Press: 416-428.

Gauch H.G. (2006): Statistical analysis of yield trials by AMMI and GGE. Crop Science 46: 1488-1500.

Gauch H.G., Piepho H.P., Annicchiarico P., (2008): Statistical Analysis of Yield Trials by AMMI and GGE: Further Considerations. Crop Science 48: 866-889.

Gauch H.G. (2013): A Simple Protocol for AMMI Analysis of Yield Trials. Crop Science (in press).

Gollob H.F. (1968): A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika 33: 73-115.

Hartigan J.A., Wong M.A. (1979): K-means clustering algorithm. Journal of the Royal Statistical Society 28(1): 100-108.

Hongyu K., Penña M.G., Araújo L.B., Dias C.T.S. (2014): Statistical analysis of yield trials by AMMI analysis of genotype × environment interaction. Biometrical Letters 51(2): 89-102.

Hongyu K., Silva F.L., Oliveira A.C.S., Sarti D.A, Araújo L.C., Dias C.T.S. (2015): Comparação entre os modelos AMMI e GGE Biplot para os dados de ensaios multi-ambientais. Rev. Bras. Biom., São Paulo 33(2): 139-155.

Johnson R.A., Wichern D. (1998): Multivariate Analysis. Wiley StatsRef: Statistics Reference Online.

Kang M.S. (2002): Genotype-environment Interaction: Progress and Prospects. In: "Quantitative Genetics, Genomics and Plant Breeding". CAB International, Wallingford, England: 221-243.

Kaufman L., Rousseeuw P. (1990): Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis: 68-125.

Mahalanobis P.C. (1936): On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India: 49-55.

Miranda G.V., Souza L.V., Guimarães L.J.M., Namorato H., Oliveira L.R., Soares M.O. (2009): Multivariate analyses of genotype x environment interaction of popcorn. Pesq. agropec. bras., Brasília 44(1): 45-50.

Neisse A.C., Hongyu K. (2016): Application of Principal Components and Factor Analysis to Crime Data From 26 US States. Pesq. agropec. bras., Brasília 44(1): 45-50.

Pacheco R.M., Duarte J.B., Vencovsky R., Pinheiro J.B., Oliveira A.B. (2005): Use of supplementary genotypes in AMMI analysis. Theor Appl Genet 110: 812-818.

Pearson K. (1901): On lines and planes of closest fit to systems of points in space. Philos. Mag. 6(2): 559-572.

R DEVELOPMENT CORE TEAM (2017): R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Rodrigues P.C., Malosetti M., Gauch H. G., Van Eeuwijk F.A. (2014): A weighted AMMI algorithm to study genotype-by-environment interaction and QTLby-environment interaction. Crop Science 54(4) : 1555-1570.

Xu R., Wunsch D.C. (2008): Recent advances in cluster analysis. International Journal of Intelligent Computing and Cybernetics 1(4) : 484-508.

Yan W., Hunt L.A., Sheng Q., Szlavnics Z. (2000): Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Science 40(3) : 597-605.

Yan W., Kang M.S. (2003): G GE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists. CRC Press, Boca Raton, FL, USA, 271p.

Yan W., Tinker N.A. (2005): An Integrated Biplot Analysis System for Displaying, Interpreting, and Exploring Genotype × Environment Interaction. Crop Science 45 : 1004-1016.

Yan W., Tinker N.A. (2006): Biplot analysis of multi-environment trial data: Principles and applications. Canadian Journal of Plant Science 86(3) : 623-645.

Yan W., Kang M.S., Ma B., Woods S., Cornelius P.L. (2007): GGE biplot vs. AMMI analysis of genotype-by-environment data. Crop Sci. 47 : 643-655.

Yan W. (2011): GGE Biplot vs. AMMI Graphs for the Genotype-by-Environment Data Analysis. Journal of the Indian Society of Agricultural Statistics 65(2): 181-193.