Conference on Applied Statistics in Agriculture 2003 - 15th Annual Conference Proceedings

CLUSTERING ENVIRONMENTS BASED ON CROSSOVER INTERACTIONS AND USING GRAPHICAL APPROACHES TO VISUALIZE CLUSTERS

Ken Russell

Kent Eskridge

Daryl Travnicek

CLUSTERING ENVIRONMENTS BASED ON CROSSOVER INTERACTIONS AND USING GRAPHICAL APPROACHES TO VISUALIZE CLUSTERS

Ken Russell, Kent Eskridge, and Daryl Travnicek

Abstract

Crossover interactions occur in evaluation trails when ranks of cultivars change across environments. Determining groups of environments within which crossover interactions are minimized may facilitate making cultivar recommendations. Model-based approaches to finding such clusters have been previously described. Our goal was to describe a new, non-model based approach of defining these clusters and then apply this method to a 59 environment x eight maize (*Zea mays L.*) cultivar data set. Hierarchical clustering of a 59 x 59 distance matrix defined two environmental clusters within which the total crossover interaction was reduced by approximately one-third and four clusters within which the crossover interaction was reduced by one-half. Four graphical approaches to visualizing the environmental clusters in this data set also were considered. Multi-dimensional scaling (MDS) allowed visualization of clusters when the dimensionality of the crossover space was reduced by considering only some of the crossover interactions between pairs of cultivars. Another benefit of MDS may be identification of specific environmental variables associated with crossover interactions.

1. Introduction

Whenever a series of crop cultivars is evaluated over a series of environments, a statistically significant genotype x environment interaction often is observed. If the goal only is to identify those cultivars with the highest mean performance, then this interaction is of no real consequence unless the rankings of the cultivars change across environments. Interaction attributable to changes in rank is known as crossover interaction.

If crossover interaction is repeatable over time, then a breeder can use this interaction to his/her advantage by grouping the environments such that the crossover interaction within environmental groups is minimized. Crossa et al. (1995) and Crossa and Cornelius (1997) developed model-based procedures for identifying these environmental clusters. Russell et al. (2003) proposed a non-model based approach to defining distances between environments based on the use of a statistic developed by Gail and Simon (1985) for testing the significance of crossover interaction between two treatments over a series of environments. Our goals are i) to discuss the effectiveness of using this distance measure to determine clusters of environments within which crossover interaction is minimized and ii) to consider the value of several different graphical approaches in visualizing these clusters.

2. Our measure of distance between environments based on crossover interaction

In Figure 1, the two treatments are cultivars *A* and *B*. Each cultivar ranks the environments the same, but the environments do not rank the cultivars the same. Gail and Simon (1985) defined Q^+ as the difference between *A* and *B* squared, divided by the variance of a cultivar difference, summed across those environments in which *A* is superior to *B*. Q^- was defined similarly, but for those environments in which *B* is superior. The test statistic, which in Figure 1 is designated as Q_{AB}, is the minimum of Q^+ and Q^- . Gail and Simon (1985) determined

critical values, c, such that the probability of Q_{AB} being greater than c is no greater than specified Type 1 error levels. As noted by Baker (1988), this approach to quantifying crossover interactions seems particularly appropriate to the analysis of changes in ranks between cultivars in multi-environmental tests. The reason is that most plant breeders would consider the crossover interaction between two cultivars in terms of the entire environmental space of interest.

In Figure 2, each column in the matrix corresponds to a unique pair of cultivars and each row to a unique pair of environments. The absolute value of each element in the AB column is Q_{AB}, which is the Gail-Simon Q statistic. A particular element is assigned a positive value if in both environments of the corresponding environmental pair the ranking of A and B is the same. Otherwise, Q_{AB} has a negative value. Then for any pair of environments, XY, a Q_{XY} value is obtained by summing across all cultivar pairs. That is, Q_{XY} is the sum of all Q_{ij}, where ij denotes any cultivar pair, with the sign considered. In contrast, Q_{SUM} is the sum of all Q_{ij}, sign ignored. Then, we define the distance between environments X and Y as 1 – the ratio of Q_{XY} to Q_{SUM}. If all cultivar pairs are ranked the same way by environments X and Y, then Q_{XY} equals Q_{SUM} and the distance between X and Y is 0. The maximum distance between any two environments with this distance parameter is 2.

Russell et al. (2003) demonstrated the effectiveness of this distance measure in identifying clusters of environments within which crossover distance is minimized using both simulated data sets with known crossover interactions and an actual data set obtained from the evaluation of eight maize cultivars in 59 environments that had been previously analyzed by Crossa and Cornelius (1997). The total crossover interaction among the undivided 59 environments, as measured by Q_{SUM}, was 756. Using Proc Cluster, method complete linkage (SAS Institute Inc., 1989) and an input 59 x 59 distance matrix computed with our crossover distance measure, Q_{SUM} was reduced by approximately one-third by dividing the environments into two groups of 24 and 35 environments (Table 1) . At the four-cluster stage, the value of Q_{SUM} was only slightly less than one-half of its initial value. This indicated that the pattern of crossover interaction in this data set was quite complex.

Even though the same method of hierarchical clustering was used as Crossa and Cornelius (1997) had employed to develop their dendogram, the environmental clusters identified with each approach were quite different. For example, the 10 environments in their smallest cluster defined at the four-cluster stage were spread across all four clusters identified by our approach. The percentage reduction in Q_{SUM} achieved by each approach was approximately the same at each stage of clustering, although our approach was slightly superior in that regard. We believe the result of several different clusterings of environments producing approximately equal reductions in the amount of crossover is a reflection of the complexity of this type of interaction in this data set.

3. Graphical presentations of crossover interaction

Plant breeders and biometricians have long been interested in using graphical approaches to visualize and to help understand genotype x environmental interactions. Therefore, it is only natural we should investigate graphical approaches to visualizing crossover interactions. In each graphical approach, the 59 x 8 maize data set was used as an example.

3.1 Plot of raw data

A plot of the raw data from only four of the eight cultivars (Figure 3) clearly demonstrated that considerable crossover interaction existed in this data set, as each cultivar had both the best performance in some environments and the worst performance in others. But this graph is not of much help in seeing any patterns of the crossover interaction. The situation would only become less interpretable if information from the other cultivars was added to the plot.

3.2 Cultivar regression on environmental index

Figure 4 shows a regression line for each cultivar that was obtained by regressing the cultivar's performance at each environment on an environmental index, which is the mean yield of all eight cultivars. This graphical approach has been widely used to investigate cultivar responses to environments for nearly 40 years (Finlay and Wilkinson, 1963; Eberhart and Russell, 1966). The R^2-value of 0.94 is the percentage of the variation within cultivars that was accounted for by the fitted regression lines. Among these lines, there are 11 crossover points. Six of these crossover points occur close to the average yield of the lowest yielding environment or to the average yield of the highest yielding environment and thus do not seem that important. The dashed lines show the other five crossover points. Based on these crossover points, there appears to be three key groups of environments: a low-yielding group, a high-yielding group, and a middle group. However, this division of the environments only removed 13% of the total crossover interaction in this data set, based on Q_{SUM}. As expected, there was considerable crossover interaction remaining in the middle group, but there also was considerable crossover interaction in the high group. Based on the Gail-Simon test statistic, the greatest crossover interaction occurred between cultivar pairs 1 and 8 and 4 and 8. Neither of these cultivar pairs accounted for any of the crossover points in Figure 4. Why did this discrepancy occur?

Even though the genotype x environmental interaction was highly significant and significant crossover interaction existed in this data set, this variation is much smaller than the environmental variation. This is almost always true in data sets from multi-environmental evaluations of cultivars. Therefore, much of the high R^2 was attributable to the good fit of the regression lines to the environmental component of the within cultivar variation. The slopes of the lines do not give a good representation of the actual genotype x environmental or crossover interaction unless that interaction is a large percentage of the total variation in the data set.

3.3 Genotype and genotype x environmental (GGE) biplot

Biplot graphs have become a commonly used technique to display multi-environmental data. Figure 5 is an environment-centered genotype, genotype x environment (GGE) biplot, for which the model is

$$\hat{Y}_{ij} - \mu - B_j = \sum_{p=1}^{2} \lambda_p \xi_{ip} \eta_{jp} + \varepsilon_{ij}, \text{ where} \qquad \text{(Eq. 3.3.1)}$$

$\hat{Y}_{ij}$ = observed value of the i^{th} genotype in the j^{th} environment,

μ = the overall mean,

β_j = the effect of the j^{th} environment,

λ_p = the singular value for the p^{th} principal component,
ξ_{ip} = the eigenvector for the i^{th} genotype and the p^{th} principal component,
η_{pj} = the eigenvector for j^{th} environment and the p^{th} principal componet, and
ε_{ij} = the residual error (Yan and Kang, 2002).

This figure was drawn using version 3.4.50 of GGE Biplot Pattern Explorer (copyright W. Yan, 2001). The first two principal components accounted for only slightly more than 50% of the variation among cultivars and the genotype x environmental interaction. The connection of the outerlying-most cultivars by lines forms a polygon, which Yan and Kang (2002) refer to as the "Which Won Where" view of the biplot. By drawing lines from the origin that bisect the sides of the polygon at right angles, the graph is divided into six sectors. For example, the area bounded by the lines bisecting lines *GA* and *AD* is referred to as sector *A*. The only environment in this sector is environment *6*. Likewise, the area bounded by the bisecting lines of lines *AD* and *DF* is sector *D*, within which there are nine environments. According to Yan and Kang (2002), this polygon view not only shows the best cultivar for each test environment but that these sectors also divide the environments into what they call mega-environmental groups. Thus, the GGE biplot suggested that cultivar *A* was the highest yielding cultivar in environment *6*, cultivar *D* was the highest yielding culitvar in each of the nine environments in sector *D*, and likewise for the other sectors. It then follows that division of the environments into the groups defined by the sectors should substantially reduce crossover interaction. In environment *6*, however, cultivar *A* was not the highest yielding cultivar. In fact, it was only the fifth highest yielding out of eight cultivars in this environment. In the nine environments of sector *D*, cultivar *D* was the highest yielding in only three of these nine environments and the second highest yielding in two others. The total value of Q_{SUM} within the six environmental groups defined by the sectors was 439. In comparison, using our distance measure and clustering procedure, a substantially greater reduction in Q_{SUM} was realized with only three clusters (756 to 388). Thus, we conclude that in a large data set with complex patterns of genotype x environmental interaction, a GGE biplot is of little use in defining or visualizing the environmental clusters that give the greatest reduction in crossover interaction.

3.4 Multi-dimensional scaling

Multi-dimensional scaling is a graphical procedure that has been used by social scientists, taxonomists, and ecologists, but to our knowledge has not been used to illustrate interactions between genotypes or environments in multi-environmental trial data sets. To understand the concept of MDS, consider a simple map showing the location of a few cities (Figure 6). Using a ruler, the distance between each pair of cities is easily determined. The opposite problem is determining the best spatial orientation of the cities when only the distance between each pair of cities is known. This is what MDS accomplishes. That is, given distances or some other measures of relationship between objects, which in this case are environments, MDS determines spatial coordinates for the objects such that the difference between the actual distances (d_{ij}) and the spatial distances (the right-hand term in the equation shown in Figure 6) is minimized.

Multiple criteria may be used to define the relationship among the cities, and likewise, the number of coordinates (*i.e.*, dimensions) may be greater than two.

Figure 7 is a plot of the two-dimensional MDS coordinates of the 59 environments obtained by using SAS Proc MDS. The input data set was a 59 x 59 distance matrix obtained from the crossover interactions between 28 pairs of maize cultivars. The lines connect the five pairs of environments for which the actual crossover distance was a minimum or a maximum. Although these lines indicate that MDS and actual distance are correlated, the environmental points do not clearly show any clusters.

Figure 8a is a plot of the predicted distance based on two-dimensional MDS coordinates against the actual distances. As expected based on results shown in Figure 7, a positive correlation obviously exists. But, there also exists a substantial amount of scattering of the points. In the SAS MDS procedure, the degree-of-fit in MDS is measured by a variable called "badness-of-fit". In the two-dimensional fit, the badness-of-fit was 0.28. In the four-dimensional fit (Figure 8b), the badness-of-fit was reduced to 0.12, and in the six-dimensional fit (Figure 8c) to 0.08. Clearly, the dimensionality of the crossover space in this data set was more than two dimensions. This conclusion is underscored by Figure 9, which is the same as Figure 7 except now the four groups of environments that were defined by hierarchical clustering of the 59 x 59 crossover distance matrix are identified. Although there is some aggregation of the environments based on their cluster membership, the definition of the clusters is not apparent in this plot. Thus, for this particular data set, it appears that MDS is not of much greater value than any of the other graphical techniques in visualizing environmental clusters within which crossover interaction is minimized.

Nonetheless, we believe that analysis by MDS may be of some value in interpreting crossover interactions. First, the badness-of-fit provides an indication of the dimensionality of the crossover space, which is valuable information. In this data set, the first decrease of the badness-of-fit below 10% occurred at five dimensions. Therefore, the five-dimensional coordinates could be used to define environmental clusters using a non-hierarchical rather than a hierarchical clustering approach. If the only initial data that are available are distances, then only a hierarchical clustering is possible prior to obtaining the MDS coordinates. In some data sets, non-hierarchical clustering may result in better clusters than hierarchical clustering.

Secondly, a key objective in evaluation of crossover interactions should be identification of specific environmental variables that cause these interactions. One approach to finding such variables is to use canonical correlation, where the MDS coordinates are one set of variables and the environmental variables suspected of being associated with crossover interactions are the other set of variables. Causation of environmental variables with high multiple correlations with the MDS coordinates could then be verified in controlled tests. We could not perform this analysis with the 59 x 8 maize data set because information on specific environmental variables was not available. However, we will be testing this approach with new data sets containing this type of information.

Finally, MDS may be of more value in visualizing environmental clusters when the crossover space is less complex. Often, plant breeders are not particularly interested in crossover interactions that occur among lower performing entries. In this data set, there were only four crossover interactions among the 28 cultivar pairs that based on the Gail-Simon test were significant at the 0.20 level and in which both cultivars had above average yield. Figure 10a is

the plot of the two-dimensional MDS fit of the 59 x 59 distance matrix in which distances were calculated from only the crossover interaction between four pairs of cultivars. All 59 environments plotted to only 11 unique points, and within all the environments at each point no crossover interaction existed. It appears that the environments fall into a left group of four points, which represents 24 environments, and a right group of seven points of 35 environments. However, in the two clusters that give the greatest reduction in Q_{SUM}, the environments represented by the top point in the right-hand group actually belong in the cluster represented by the left points. This occurred because the dimensionality of the crossover space still is greater than two (badness-of-fit = 0.18).

Figure 10b is a three-dimensional MDS plot of the same data set. Not only does the definition of the two environmental clusters become much more conspicuous, but this plot actually suggests the existence of four environmental clusters. These clusters are identical to those defined by hierarchical clustering of this distance matrix. Within these four clusters of 7, 11, 18, and 23 environments, the total Q_{SUM} was reduced from 156 to 33.

4. Summary

Crossover interactions complicate the interpretation of multi-environmental evaluations of cultivars. If the crossover interactions are repeatable over time, then identifying clusters of environments within which these interactions are minimized is appropriate. Graphical procedures that do not focus directly on crossover interaction data will likely be of limited value in visualizing these clusters. Multi-dimensional scaling of crossover interactions should be helpful in determining the specific environmental variables that are associated with these interactions. Also, if the crossover patterns are not too complex, then the plots of two- or three-dimensional MDS coordinates may graphically reveal the spatial orientation of these clusters.

5. References

Baker, R.J. 1988. Tests for crossover genotype-environmental interactions. Can. J. Plant Sci. 68: 405-410.

Crossa J., P.L. Cornelius, K. Sayre, and J.I. Ortiz-Monasterio R.A. 1995. Shifted multiplicative model funsion method for grouping environments without cultivar rank change. Crop Sci. 35:54-62.

Crossa J. and P.L. Cornelius. 1997. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. Crop Sci. 37:406-415.

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

Finlay, K.W. and G.N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Austr. J. Agric. Res. 14:742-754.

Gail, M. and R. Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41: 361-372.

Russell, W.K., K.M. Eskridge, D.A. Travnicek, and F.R. Guillen-Portal. 2003. Clustering environments to minimize change in rank of cultivars. Crop Sci. 43:858-864.

SAS Institute Inc. 1989. SAS/STAT user's guide (version 6, 4th ed.). SAS Institute Inc., Cary, NC.

Wan, Y. and M.S. Kang. 2002. GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists. CRC Press, Boca Raton

Figure 1. Crossover interaction between two cultivars across eight environments.

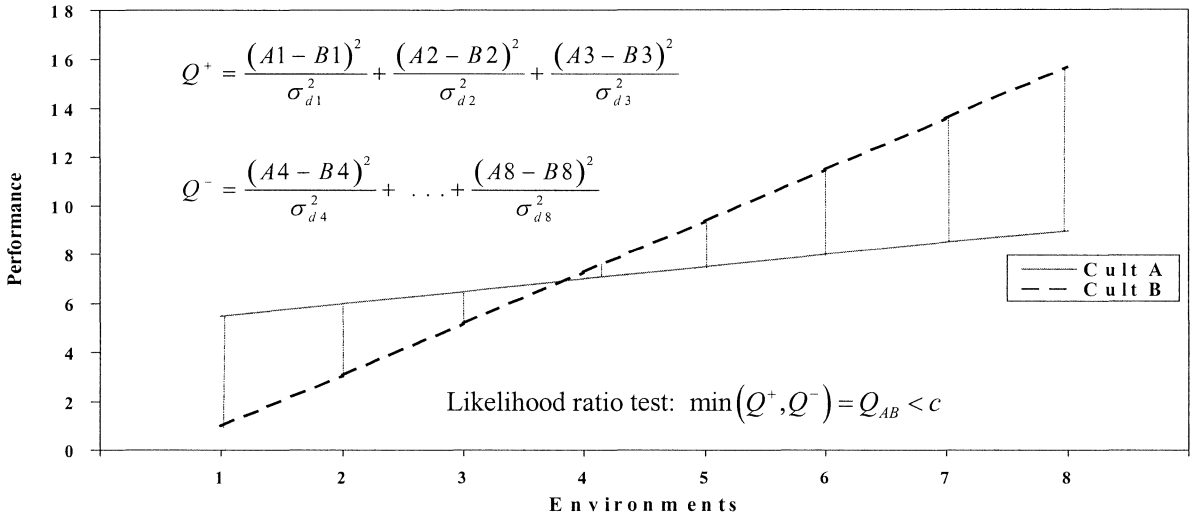
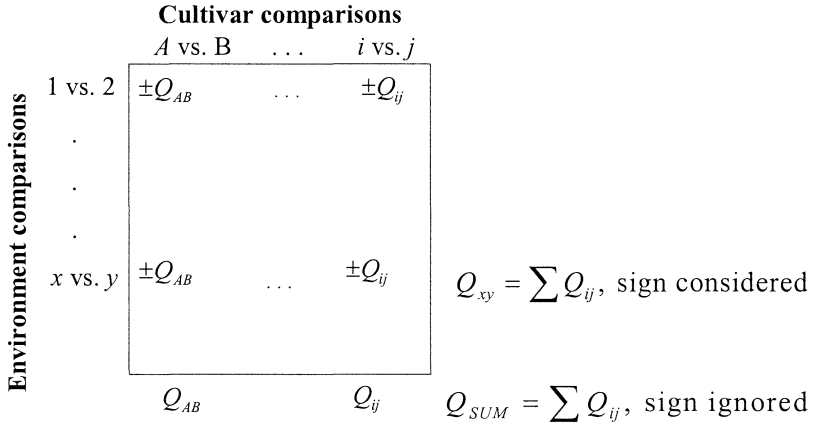


Figure 2. Calculation of a crossover distance between two environments based on Q statistic of Gail and Simon (1985).



$$\text{Distance}(xy) = 1 - \frac{Q_{xy}}{Q_{SUM}}$$

Table 1. Reduction in crossover interaction achieved by hierarchical clustering of a 59 x 59 crossover distance matrix generated from grain yield data of eight maize cultivars in 59 environments.

Number of clusters	Number of environments in cluster	Q_{SUM}†
1	59	756
2	a) 24	158
	b) 35	325
	Combined	483
3	a) 24	158
	b) 12	104
	c) 23	126
	Combined	388
4	a) 12	51
	b) 12	75
	c) 12	104
	d) 23	126
	Combined	356

† Q_{SUM} is a measure of the amount of crossover interaction that is based on the Gail and Simon (1985) Q statistic.

Figure 3. Grain yields of four maize cultivars across 59 environments.

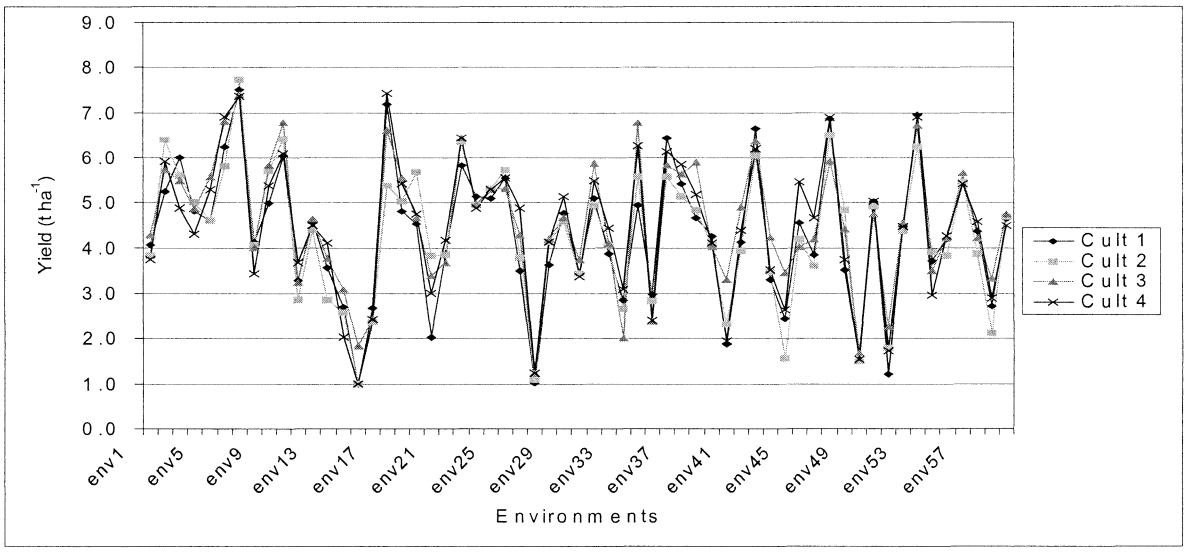


Figure 4. Regression lines of grain yields of eight maize cultivars on environmental index across 59 environments. Dashed lines denote crossover points between any pair of cultivars that have moderate environmental index values. The index value of an environment is the mean yield of all eight cultivars in that environment.

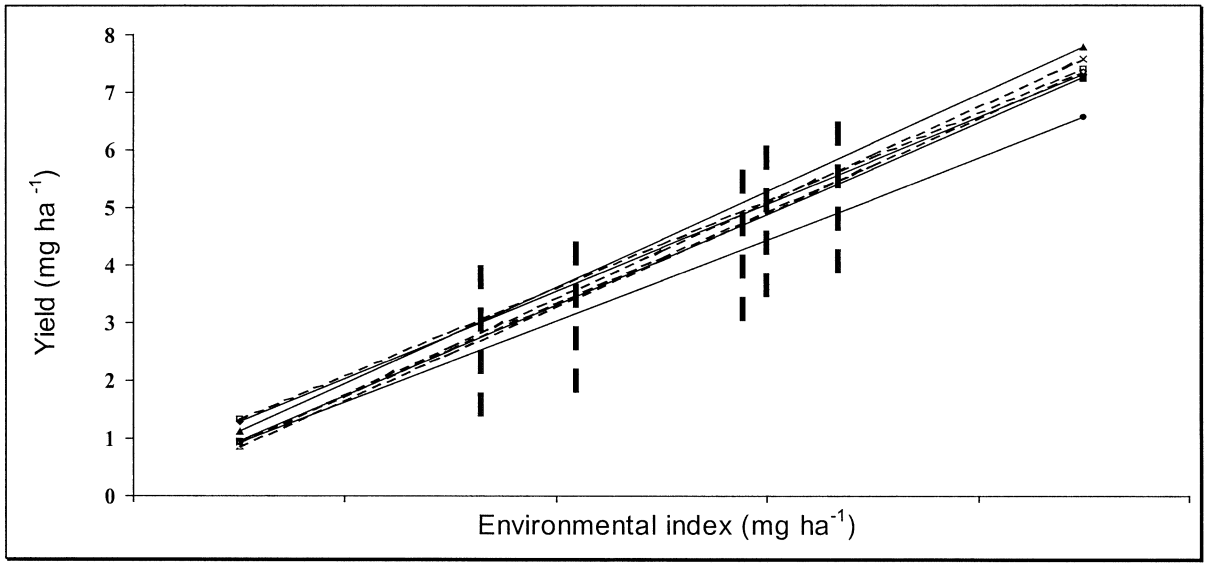


Figure 5. Genotype, genotype x environment (GGE) biplot of grain yields of eight maize cultivars (designated by letters) evaluated across 59 environments (designated by numerals).

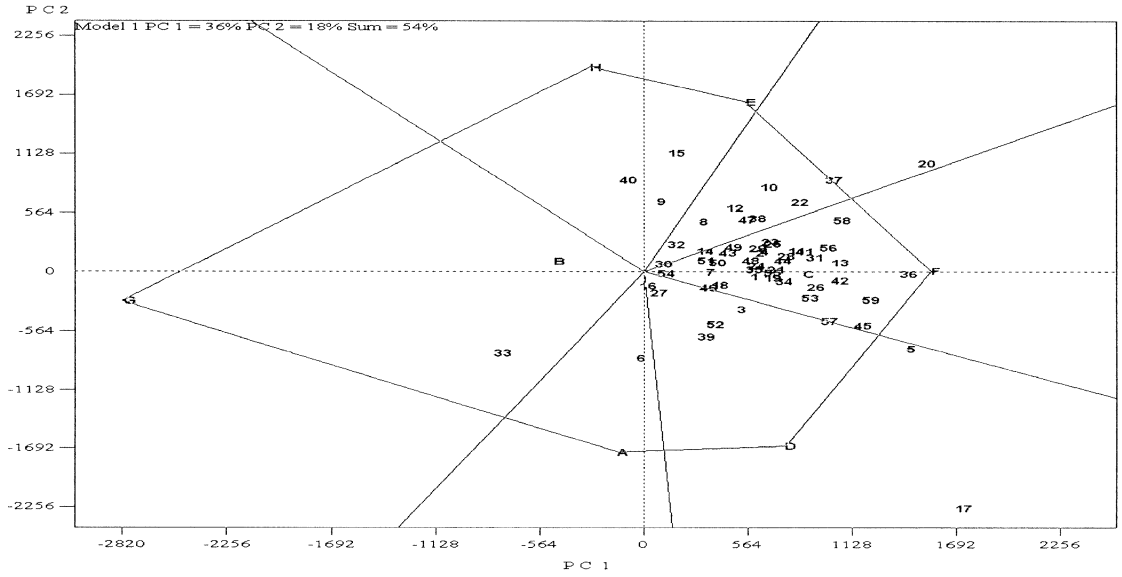


Figure 6. Multiple-dimensional scaling is a procedure that determines coordinates of objects (cities, environments, etc.) given the distances between these objects.

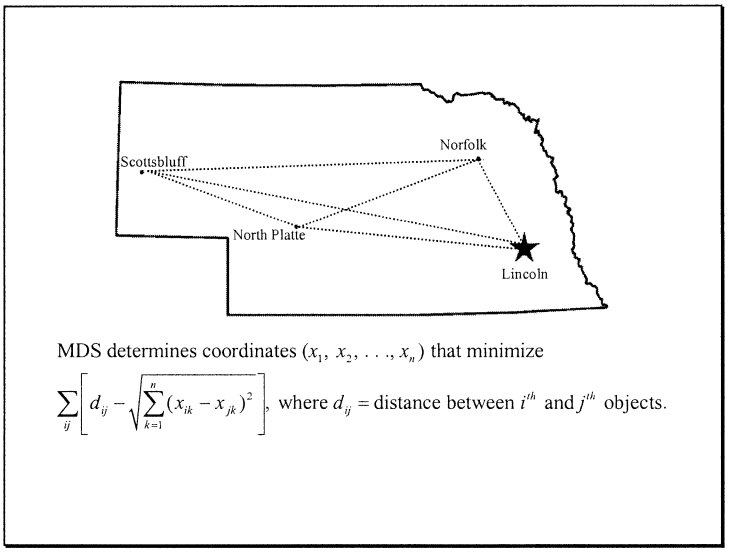


Figure 7. Two- dimensional plot of multiple-dimensional scaling coordinates from a 59 x 59 crossover distance matrix generated from the grain yields of eight maize cultivars evaluated across 59 environments. Solid lines connect the pairs of environments with the smallest crossover distances; dotted lines connect the pairs of environments with the largest crossover distances.

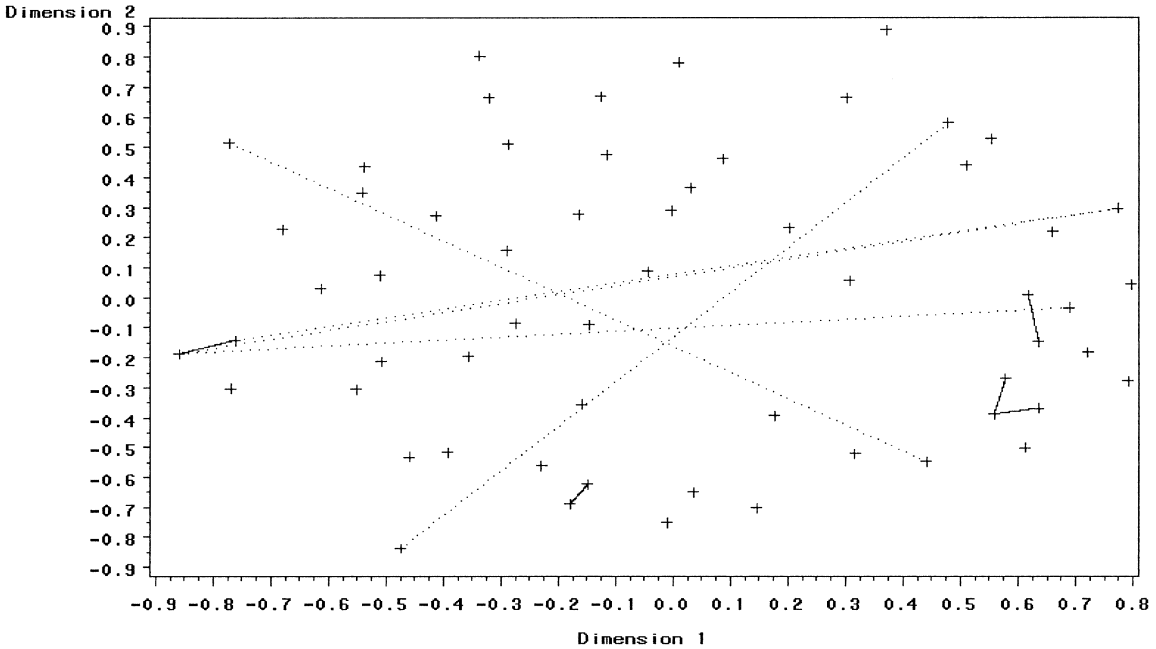
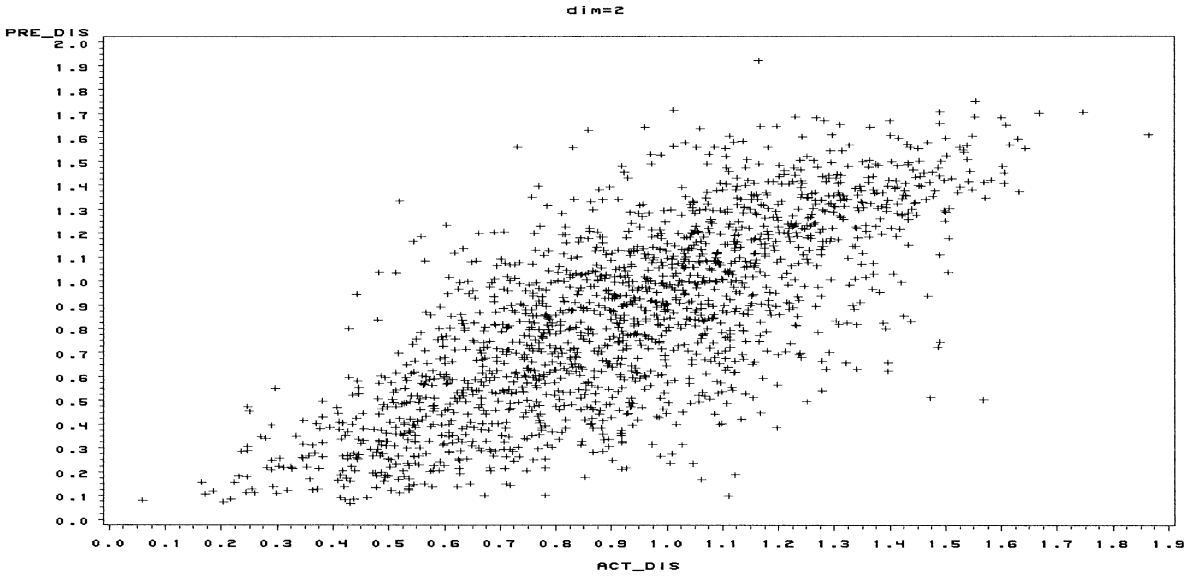
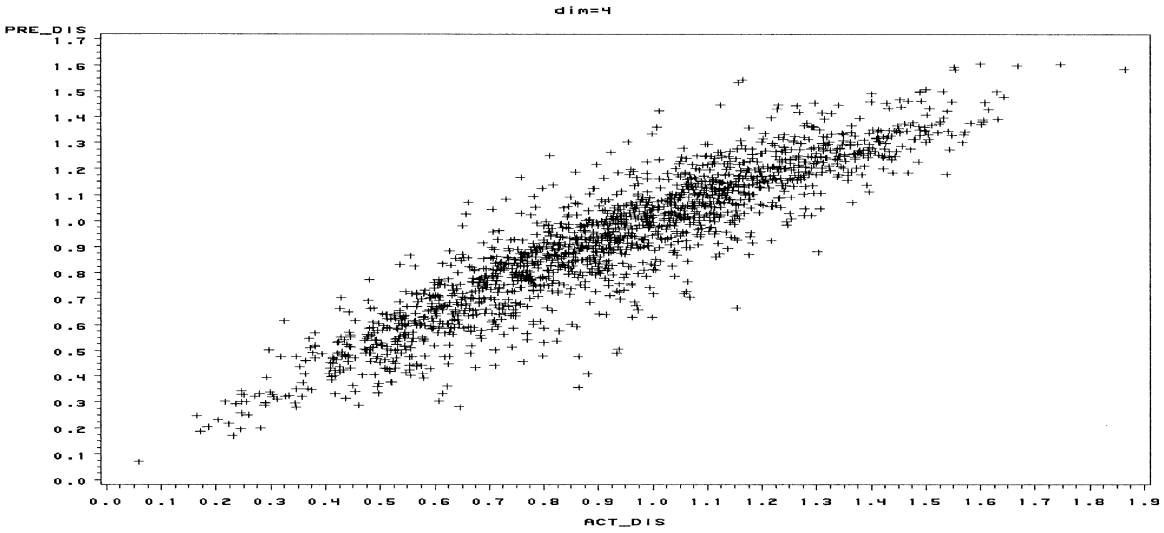


Figure 8. Predicted multi-dimensional scaling (MDS) distances for 59 environments plotted against actual crossover distances. Crossover distances are calculated from grain yields of eight maize cultivars in those environments.

a) Two MDS dimensions



b) Four MDS dimensions



c) Six MDS dimensions

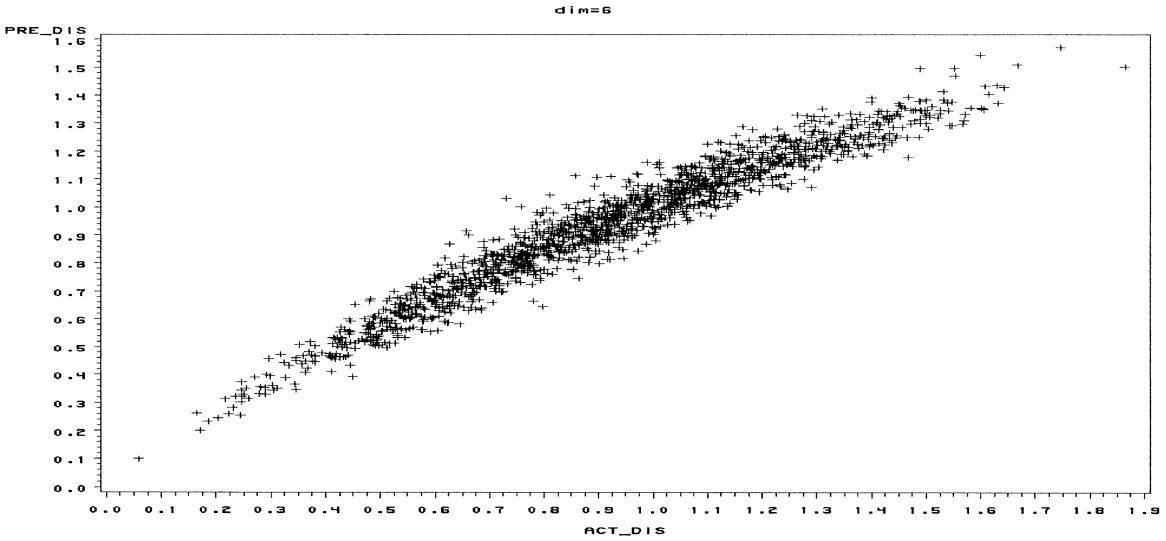


Figure 9. Identification in a multi-dimensional scaling plot (two dimensions) of the four environmental groups defined by hierarchical clustering of a 59 x 59 crossover distance matrix that give the greatest reduction in crossover interaction. Crossover distances are calculated from grain yields of eight maize cultivars in 59 environments.

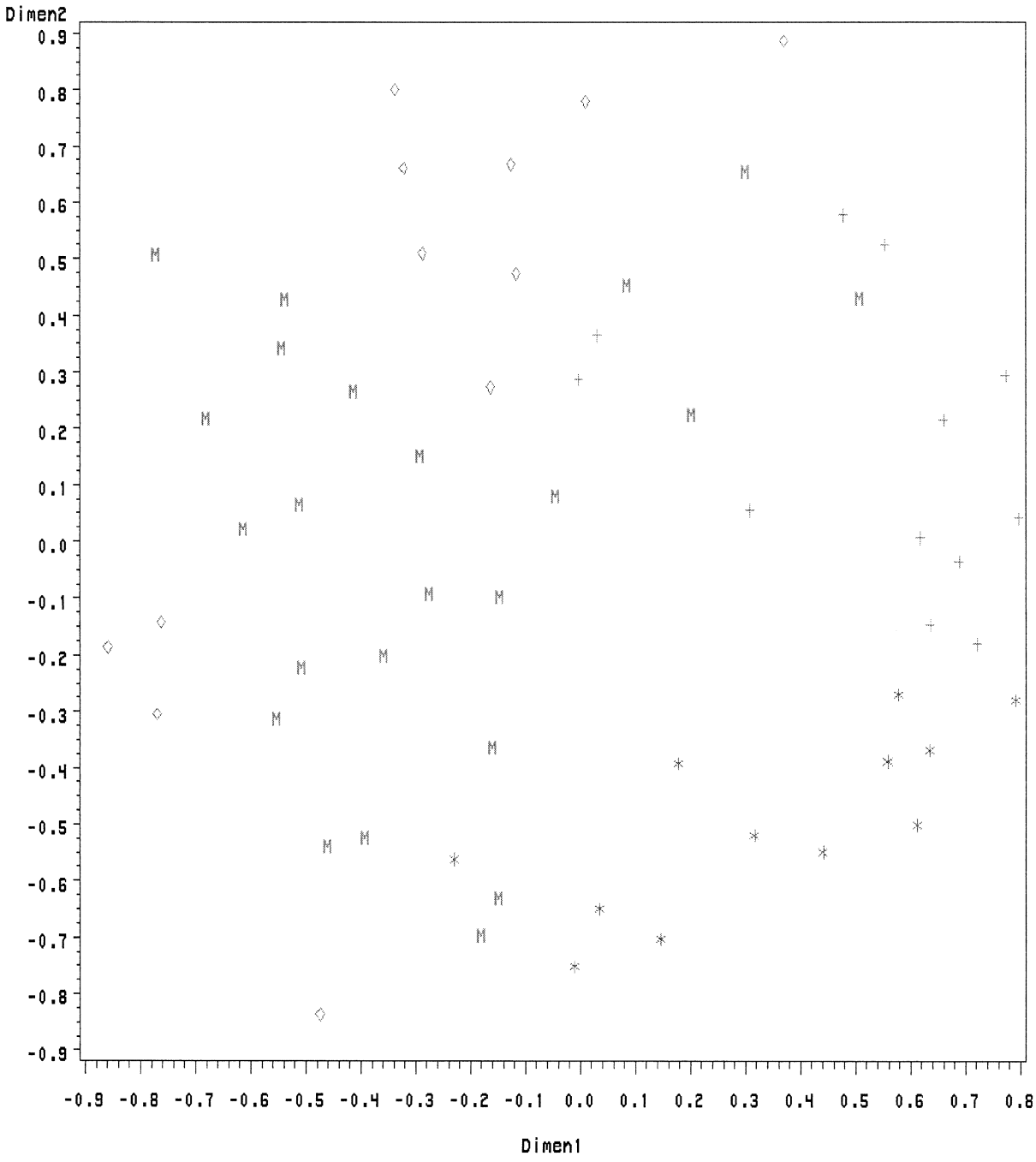
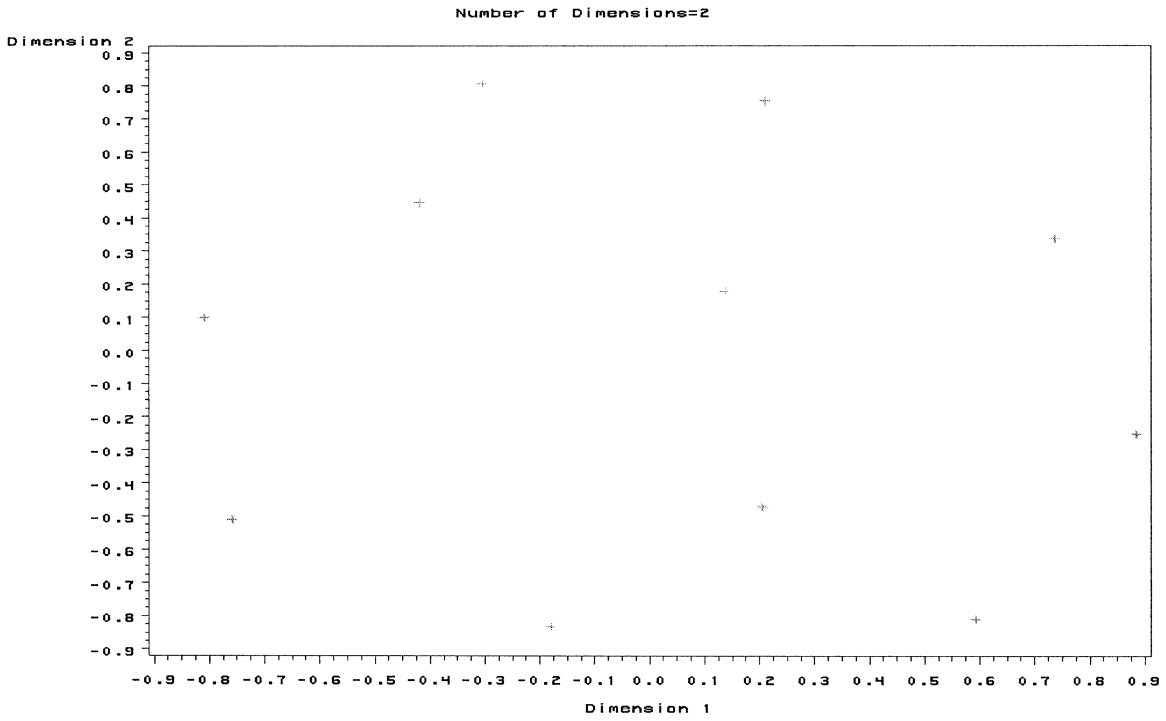


Figure 10. Environmental groups in a multi-dimensional scaling plot of 59 environments. Distances between environments are based on selected crossover interactions between eight maize cultivars for grain yield.

a) Two dimensions



b) Three dimensions