Published March, 1969

# Classification of Environments to Control Genotype by Environment Interactions with an Application to Cotton[1]

H. A. Abou-El-Fittouh, J. O. Rawlings, and P. A. Miller[2]

**ABSTRACT**

**Cluster analysis as a tool for classifying locations in order to minimize the within-cluster genotype by location interactions was discussed and applied to the data for lint yield per hectare in upland cotton (*Gossypium hirsutum* L.) obtained from the Regional Cotton Variety Tests for years 1960-1964. The distance coefficient, which was more efficient as a measure of similarity than the product-moment correlation coefficient in preliminary analyses, was used to study the zoning of the cotton belt. Some modifications in the currently recognized zones of adaptation for cotton were suggested.**

***Additional index words.*** **Distance coefficient, Product-moment correlation coefficient, Upland cotton, *Gossypium hirsutum* L., Cluster analyses.**

THE genotypic value, breeding value, and commercial worth of breeding lines or varieties are all defined as averages over some reference set of environments. Entry by environment interaction is treated quite appropriately as part of the experimental error in the evaluation of the entry differences. The existence of this interaction has been long recognized as evidenced by the standard requirement of testing breeding material over many environments and the delineation of "testing regions" within which part of the environmental conditions are more or less standardized.

The commonly used categorization of environments into years and locations identifies a portion of the interaction effects, the entry by location interaction, which may in part be controlled by the delineation of breeding regions. The principal basis for defining these regions appears to have been a subjective evaluation of the mean difference in certain environmental factors thought to be important with respect to level of performance, *per se,* with only secondary attention being given to effects of these factors on genotype by environment interaction. Insofar as varietal testing is concerned, the interaction effects of environmental factors are more important than their main effects. Furthermore, the size of the direct effect of an environmental factor may not be at all related to the size of the interaction attributable to that factor.

Developing varietal testing zones by use of a reduction in genotype by location interaction has been reported by Horner and Frey (1957). They estimated the genotype by location interaction component from nine locations for 5 years in oats. This component was reduced by 11%, 21%, 30%, and 40% from that of all locations when the test area was divided into two, three, four and five subareas, respectively.

As a first step in controlling genotype by environment interactions, without requiring any knowledge of the environmental factors responsible, locations can be classified according to the similarity of their interactions with a set of entries. Such a procedure would result in dividing a set of locations into not necessarily contiguous regions such that the interactions of entries and locations within regions would be small. The precision to be gained by such zoning is dependent on the proportion of the entry by environment interaction that is due to locations and on the efficacy of the classification method to form discrete classes of what must be a continuous variable. Other ways of control will undoubtedly be developed as the environmental factors causing the interactions are identified.

The purpose of this study is to present a method for studying the similarity of environments insofar as genotype by environment interactions is concerned and to evaluate and possibly to modify the system of zoning used for the Cotton Belt of the United States.

## METHODS OF ANALYSES

Two methods of measuring similarity of locations were studied, both utilizing a set of data from v varieties grown in each of $l$ locations for y years. Each location was represented by a vector ($l$ × v) of estimated variety by location *interaction effects,*

$$l_1 = [\widehat{(vl)}_{11}, \ \widehat{(vl)}_{12}, \ \ldots, \ \widehat{(vl)}_{1v}]$$

Then the two measures of similarity between locations i and i′ were defined as follows:

(1) Distance coefficient:

$$d_{ii'} = \{ \sum_{j=1}^{v} [ \widehat{(v\ell)}_{ij} - \widehat{(v\ell)}_{i'j} ]^2 / v \}^{1/2}$$

(2) Product-moment correlation coefficient:

$$r_{ii'} = \sum_{j=1}^{v} \widehat{(v\ell)}_{ij} \widehat{(v\ell)}_{i'j} \Big/ [ \sum_{j=1}^{v} \widehat{(v\ell)}^2_{ij} \sum_{j=1}^{v} \widehat{(v\ell)}^2_{i'j} ]^{1/2}$$

Distance coefficients were based on a geometrical model of v dimensions, for the v varieties, in which each location was represented by a point whose coordinates were the estimated interaction effects. The use of distance to measure resemblance was discussed by Sokal (1961) where he used average squared distances. Rohlf and Sokal (1965), however, preferred the use of linear distances. Correlation coefficients were first introduced into classification problems by Michener and Sokal (1957) and Sokal and Michener (1958).

Using the measures of similarity, the locations with similar patterns of interaction were clustered into groups at several successive stages. The clustering procedure used was an adaptation of "variable group clustering" based on "average linkage" originally used by Sokal and Michener (1958) and described by Sokal and Sneath (1963), section 7.3.2.4. The method was originally applied to correlation coefficient matrices but is applicable to both measures of similarity used in this study. The procedure as used herein will be outlined and then the differences from the method as described by Sokal and Sneath will be described.

The measures of similarity were computed for all possible pairs of locations giving an $l$ × $l$ matrix of similarity coeffi-

[1] Paper number 2561 of the Journal Series of the North Carolina State University Agricultural Experiment Station, Raleigh, N. C. 27607. This investigation was supported in part by Public Health Service Research Grant No. GM 11546 from the Division of General Medical Sciences and by the National Institutes of Health Grant No. FR-00011. Part of a thesis submitted by the senior author in partial fulfillment of the requirements for the Ph.D. degree. Received for publication February 9, 1968.

[2] Former graduate student, Department of Experimental Statistics (now Statistician, Ministry of Agriculture, Giza, UAR); Professor of Experimental Statistics; and Professor of Crop Science, North Carolina State University, respectively.

cients. Then, letting $d_{min(1)}$ and $r_{max(1)}$ be the minimum distance coefficient and the maximum correlation coefficient, respectively, in the set of all possible pairwise measures in Stage 1, a subset of the $l$ points was grouped to form a cluster if

*i*) all possible pairwise measures within the subset were less than $d_{min(1)} + \alpha$, for the distance coefficient method, or greater than $r_{max(1)} - \beta$, for the correlation coefficient method and *ii*) if each point in the subset was more similar, on the average, to the other points in the same subset than to any point outside the subset.

In our application $\alpha = 22.23$ and $\beta = .05$ were used, the choice of which will be explained later. After grouping was completed in the first stage, each grouped subset was represented by its center point in space whose coordinates were the average of coordinates for the points in the subset and a new $l' \times l'$ matrix ($l' < l$) of similarity coefficients was computed. The procedure was repeated until a single cluster was formed out of the $l$ locations.

This procedure differs from that described by Sokal and Sneath in the following respects. First, the basic variable is defined differently in that similarities are measured on a vector of estimated interaction effects rather than on a vector of primary measurements on a set of traits as is generally the case in Numerical Taxonomy. This, however, does not alter in any way the application of the clustering techniques. Secondly, in the original description a new point was admitted to a cluster if the admission of the point did not lower the average similarity, $\bar{S}_n$, of the cluster by more than 0.03. The value 0.03 referred to correlation coefficients as the measure of similarity and was arrived at empirically. As they suggested, the value would need to be adjusted for different studies or different similarity coefficients. If the value is set too small, the clustering process may terminate before it is complete simply because no point can be found which would not lower the $\bar{S}_n$ of the existing clusters by more than the prechosen amount. Conversely, setting the value too high would mean the clustering would proceed at too rapid a pace with the consequent obscuring of intermediate steps. The choice of $\alpha$ and $\beta$ in our application plays the same role except that the maximum similarity coefficient in a particular stage, $r_{max(m)}$ or $d_{min(m)}$, is used to provide a sliding scale as a point of reference to insure that some clustering occurs at every stage, so that the process does not terminate prematurely. As $\alpha$ or $\beta$ approaches zero, the clustering approaches the so-called 'pair-group' method where only the two most similar points are combined at any one stage.

The only disadvantage of the pair-group method is the greater amount of computing required since ($l$ — 1) stages are required to complete the clustering of $l$ points. It has the advantage of being "devoid of an arbitrary criterion for group formation" and is probably to be preferred if computing facilities permit although, as pointed out by Sokal and Sneath, "it is well known that the two alternatives produce very similar results." In our case, preliminary analyses indicated that the resultant clustering was relatively insensitive to reasonable changes in $\alpha$ or $\beta$. (For a more complete discussion, see Sokal and Sneath (1963) section 7.3.2.6 and a recent paper by Gower (1967) in which three methods of cluster analysis are compared.) Thirdly, our procedure differs in that a new matrix of similarity coefficients was computed after every stage of clustering based on the average coordinates of the points in each cluster rather than using Spearman's sums of variables method as suggested by Sokal and Sneath (1963) for correlation. The method employed for taking averages was analogous to the unweighted average of similarities described by Gower (1967), equation (2).

There is no assurance that any of the clustering methods provide the "best" solution since the optimum clustering, or zoning in this case, can be determined only if all possible groupings are studied.

## DESCRIPTION OF DATA

The data analyzed were the 1960-1964 results of the Regional Cotton Variety Tests conducted by the U. S. Department of Agriculture in cooperation with the Agricultural Experiment Stations in the Cotton Belt[3]. In this testing program the U. S. Cotton Belt is divided into five regions (Fig. 1, current system). Table 1 lists the code numbers for the locations involved. The data belonged to two cycles of varietal testing, the first for the years 1960-1962 and the second for 1963-1964, and were collected at 39 different testing sites.

**Table 1. Code numbers and current regions of the classified locations.**

Plains region

1. Brownfield, Texas
2. Lubbock, Texas
3. Halfway, Texas
4. Spur, Texas
5. Chillicothe, Texas
6. Altus, Okla.
7. Chickasha, Okla.
8. Mangum, Okla.
9. Chickasha, Okla.

Eastern region

10. Jackson, Tenn.
11. St. College, Miss.
12. Crossville, Ala.
13. Experiment, Ga.
14. Tifton, Ga.
15. Florence, S.C.
16. Rocky Mount, N.C.

Central region

17. Stillwater, Okla.
18. College Sta., Texas
19. Beeville, Texas
20. Weslaco, Texas
21. Hope, Ark.
22. Bossier City, La

Delta region

23. St. Joseph, La.
24. Stoneville, Miss.
25. Tunica, Miss.
26. Clarkedale, Ark.
27. Sikeston, Mo.
28. Ft. Pillow, Tenn.

Western region

29. Shafter, Calif.
30. Brawley, Calif.
31. Yuma, Ariz.
32. Tempe, Ariz.
33. Marana, Ariz.
34. Artesia, N. Mex.
35. Univ. Park, N. Mex.
36. Ysleta, Texas
37. Pecos, Texas
38. Logandale, Nev.
39. Pecos, Texas

The data were unbalanced since some locations were added to or dropped from the testing program after it was initiated. A two-way table of varietal mean lint yield/hectare at each location over years, within each testing cycle, was constructed. Grand mean and appropriate variety and location effects were subtracted from each cell yielding a ($l \times$ v) table of the estimates of variety by location interaction effects. Resemblances between pairs of locations were computed from the resulting two-way table using the rows as the $l_i$ vectors.

The grouping stages were then represented by a "tree" diagram of relationships. Analyses of variance over years were computed for each stage. The first meaningful stage of zoning was taken as the last stage of grouping before the variety by location within region (group) mean square exceeded the variety by year by location within region mean square. This decision rule was chosen arbitrarily to define a point beyond which branching became excessively sensitive to yearly fluctuations.

## COMPARISON OF CLASSIFICATION METHODS

The distance coefficient was proposed as one measure of similarity because, in addition to its rather common usage for cluster analysis, it is closely related, as used herein, to the variety by location interaction sum of squares. It is easily shown that $\frac{v}{2}\, d^2_{ii'}$ is the sum of squares due to variety by locations interaction within the region formed by combining locations i and i'. More generally, the sum of squares due to variety by location interactions within any size cluster of locations is v times the sum of squared distances of each point from the midpoint of the cluster.

The correlation coefficient was included initially because, in addition to being commonly used in cluster analysis, it has previously been used in a similar context to measure the similarity of 'behavior' of two locations. The relationship between the correlation coefficient and the distance coefficient can be seen by noting that

$$r_{ii'} = 1 - \frac{v}{2}\,(d^*_{ii'})^2$$

where $d^*_{ii'}$ is the distance measure after each vector, $l_i$, has been standardized to unit length. In 2-dimensions with similar interpretations for n-dimensions this is equivalent to moving each point along the line of its vector to the circumference of a circle with radius 1. Thus, only the information contained in the angle between two vectors is retained; the information in the lengths of the vectors is lost.

These considerations suggest that the distance coefficient would be more efficient for the purpose of defining breeding regions when the criterion is the minimization of within-region genotype by location interactions. To obtain some empirical check on their relative efficiencies, both measures were utilized to classify the 39 locations in the Cotton Belt into groups on the basis of the differential lint yield performance of the four varieties which were common to all locations in 1960-1962 ('Acala 4-42,' 'Coker 100A,' 'Deltapine 15,' and 'Lankart 57').

[3] Results of the Regional Cotton Variety Tests are published annually by the Cotton and Cordage Fibers Research Branch, USDA, Beltsville, Md.

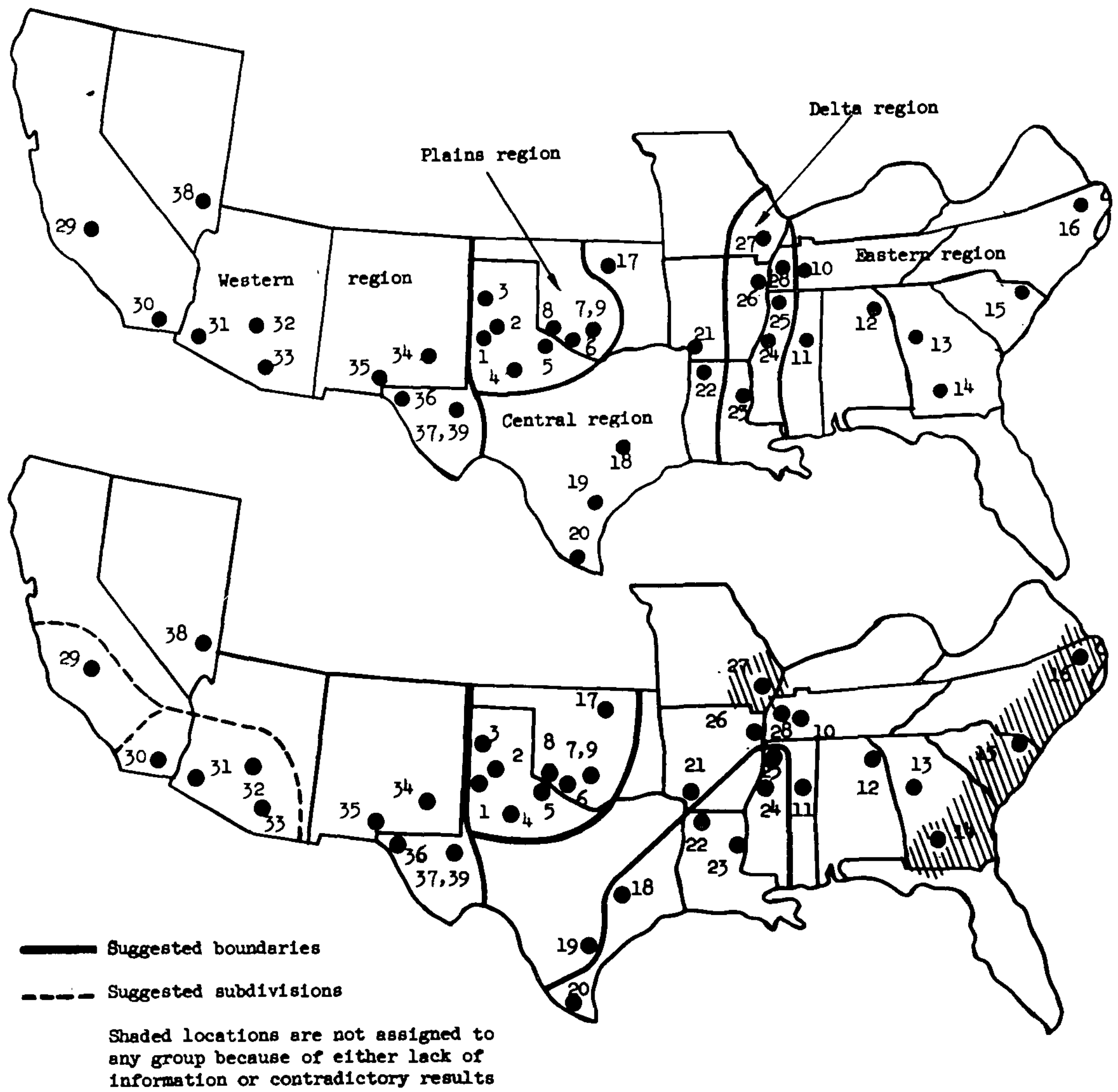


Fig. 1. Current (above) and suggested (below) zoning systems of the Cotton Belt.

Table 2. Number of groups formed by the methods of distances and correlation coefficients and the within groups interaction sum of squares at each stage of grouping.

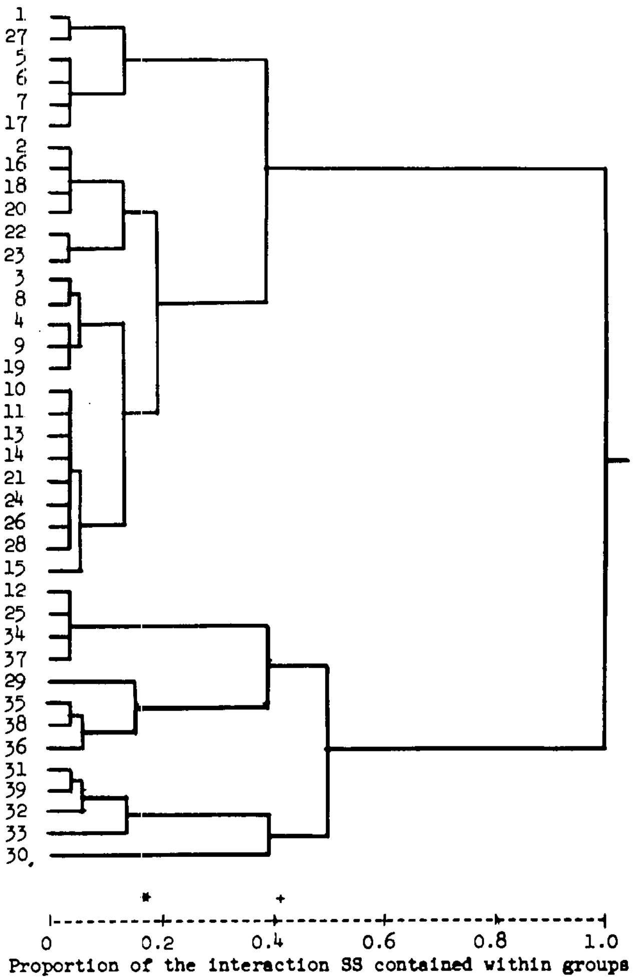
| Grouping stage | Method of distances | | Method of correlation coefficients | |
|---|---|---|---|---|
| | No. of groups | S.S. | No. of groups | S.S. |
| 1 | 16 | 37,591 | 23 | 59,491 |
| 2 | 12 | 54,855 | 19 | 80,278 |
| 3 | 7 | 117,289 | 14 | 107,974 |
| 4 | 6 | 170,570 | 9 | 141,574 |
| 5 | 3 | 339,912 | 7 | 173,395 |
| 6 | 2 | 439,318 | 6 | 212,224 |
| 7 | 1 | 862,777 | 4 | 310,455 |
| 8 | - | - | 2 | 453,036 |
| 9 | - | - | 1 | 862,777 |

This was the largest balanced set of data available involving all test locations. Table 2 shows the number of groups formed by each method and the within groups interaction sum of squares at each zoning stage. In the entire clustering sequence, the distances method consistently resulted in the smaller within groups interaction sum of squares for any given number of groups formed. In addition, the clustering formed using the correlation coefficient showed a definite arrangement into rays oriented toward the origin with no cluster crossing the origin. This reflects the fact that the similarity of two points using the correlation coefficient is determined only by the angle between the two vectors. The clusters formed using the distance method were, in contrast, much more compact.

On the basis of the above results and the theoretical considerations, distance coefficients were used to study similarities between the interaction effects for lint yield/hectare at the different locations using larger sets of varieties. This is not to say that for other objectives or other criteria of success the correlation would not be the more desirable measure of similarity.

## ZONING RESULTS

Distance coefficients were used to measure the resemblance between pairs of locations on the basis of the estimated interaction effects for yield. For illustrative purposes, the diagram of relationships resulting from the method of distances is presented in Fig. 2. The relationships in this diagram are based on the differential yield performance of the four varieties tested at all 39 locations during 1960-1962. In the figure, the first meaningful stage of zoning, designated by an asterisk, yielded seven groups with variety by location within group interaction sum of squares equal to .18 of the total variety by location interaction sum of squares. The vertical lines connecting similar subgroups are positioned in the figure to reflect the pro-



* Latest stage with MSVL(R)/MSVYL(R) < 1

+ Proportion of interaction SS contained within regions for the current zoning system, number of regions = 5

**Fig. 2. Diagram of relationships for the method of distances.**

portion of the total variety by location interaction sum of squares contained within groups after each stage of grouping. The dagger designates the proportion of the interaction sum of squares found within regions for the current system of zoning when the same set of data was analyzed.

Similarity relationships were first established within the currently defined regions and then among adjacent sets of regions utilizing at each step information on all common varieties. Similarities or dissimilarities based on intraregional analyses were favored over those based on interregional analyses if they were contradictory, since the former were obtained from more information than the latter.

## Intraregional Similarities

Relationships established among the locations in the Eastern region of the current zoning system were based on the interaction pattern in the yields of nine varieties in each of the two testing cycles. These relationships indicated that location 15 had an interaction pattern similar to that of location 16 in the first testing cycle and of location 14 in the second. All three locations occupy the extreme eastern section of the Cotton Belt. Furthermore, the remaining locations of this region were considerably similar in the two cycles.

Interaction patterns in the yields of 14 varieties in the first testing cycle and 10 varieties in the second were utilized to establish resemblances among the locations in the current Delta zone. Locations 23, 24, and 25 had a considerable degree of resemblance in both cycles. These three locations occupy the southern part of the Delta region. Location 26 fits rather well with location 27 in the first cycle of testing and with location 28 in the second.

Variety by location interaction component in the Central region did not differ significantly from zero and was considerably smaller than the variety by year by location interaction component in the second cycle. Data from the first cycle suggested that location 17, the northernmost site in the region, did not show any degree of similarity with the remaining locations insofar as interaction pattern was concerned. This was based on data from 12 varieties. In spite of being geographically located between locations 18 and 20, location 19 was dissimilar to either one. Rather, it joined location 21 in the first stage of clustering, in which stage locations 18, 20, and 22 were joined together.

Relationships calculated within the current Plains region were based on the interaction pattern in the yields of 12 varieties in the first cycle and 11 varieties in the second. Clustering results were quite inconsistent between the two testing cycles and this was considered an indication against any attempt to subdivide this region.

In the Western region of the Cotton Belt, the variety by site interaction was not significantly different from zero and unimportant when compared with the sizable component due to the three-factor interaction in the second testing cycle. However, relationships based on eight varieties in 1961 and 1962 suggested that locations 30, 31, 32, and 33; and locations 35, 36, 37, and 38 formed two reasonably homogeneous clusters. Location 29 was dissimilar to all locations in the region.

## Interregional Similarities

Figure 2, based on the interaction pattern in the yields of four varieties in the first testing cycle, suggested that with exception of locations 12 and 25, the Western region had a relatively different pattern of interaction from that of all other four regions in this cycle.

Data from the current Plains and Central regions were combined in the first cycle. Of interest was the indication that location 17 resembled some of the locations in the current Plains region. In the second cycle, the pooled interaction effects from the two regions were found nonsignificant and relatively small, thus no attempt was made to control this interaction by studying the pattern of location clustering.

Distance coefficients were computed between the locations of the Eastern, Delta, and Central regions of the present zoning regime. These distances were based on the interaction pattern in the yields of seven varieties which were common to all locations in the first cycle of testing. Two major clusters were observed. The first cluster consisted of locations 10, 11, 13, 14, 15, 19, 21, 22, 26, and 28. The inclusion of locations 15 and 22 in this group is contradictory to the within regional results and therefore the two locations can be considered as questionable members. The second cluster contained locations 16, 18, 20, and 24. Within region analyses suggest that locations 22, 23, and 25 should be included in this group. Location 17 was quite dissimilar to the other sites studied.

In the second cycle, and on the basis of four varieties which were common to the three eastern regions, three major groups were found. The first group consisted of locations 14 and 15. The second consisted of the remaining sites in the current Eastern region as well as locations 17, 19, 21, 26, and 28. On the basis of the intraregional analyses, location 17 would be excluded from that group, and location 16 would be considered a questionable member. The last group included locations 18, 22, 23, 24, and 25. Regional information put location 20 with this cluster of sites.

When all test locations in the Cotton Belt were combined in the second cycle, the three common varieties did not show a substantial degree of interaction with the set of environments and, in addition, the Western region had an outstanding contribution to the three factor interaction used as reference in determining the first meaningful stage of zoning. For these reasons, distance coefficients were not estimated for this environmental set.

## DISCUSSION

Several steps toward improving the current system of zoning were suggested by the results. In the current Western region, interactions were sizable and their pattern changed appreciably over the years. The data indicated that the variety by location interaction can be reduced by combining locations 30, 31, 32, and 33 in one subzone; and locations 34, 35, 36, 37, and 38 in another subzone. Location 29 was dissimilar to all other sites studied within the region. The results did not favor subdividing the current Plains region. Rather, this region was expanded to include location 17.

Among the remaining locations, two major groups were observed. The first group included locations 18, 20, 22, 23, 24, and 25. This group combined parts of the current Delta and Central regions. The second group was comprised of locations 10, 11, 12, 13, 19, 21, 26, and 28. This group combined parts of the current Eastern, Delta, and Central regions. It was questionable, however, as to whether or not locations 14, 15, and 16 should be combined together in a separate region or combined with the one above. These locations are on the boundary of the Cotton Belt. The information concerning location 27 was inadequate for the purpose of assigning it to some subset of the population of environments and there was some inconsistency regarding location 28. Fig. 1 shows the modified zoning system suggested by the results. The boundaries drawn in the figure show only the grouping of the test locations involved and are not intended to represent exact geographical boundaries.

The suggested zoning system is expected to reduce the average estimate of the variety by location within region interaction component to about one half of that estimated under the current system. This speculation is based on several individual analyses of variance performed on the same set of data used to develop the proposed system and consequently the size of this reduction may be an overestimate.

Subdivision of the current Western region agrees with the modification introduced recently to the analyses of the data obtained in this region[4]. Inclusion of location 17 into the current Plains region seems to be geographically and climatically justified. The southern Mississippi Valley and the southeast portion of Texas are represented by "river bottom" locations which tended to have similar patterns of varietal performance.

The data used in this study constituted a sample of five yearly environments for most locations and a similarity of locations over all years was required before being combined. This should comprise an adequate sample of yearly differences. Most comparisons of locations within the current regions involved a reasonable number of varieties and should be fairly reliable. However, many of the decisions regarding interregional similarities were based on a minimum of data. Consequently, it is recommended that this system of zoning be studied further, before being adopted, using an independent set of data on the performance of a large number of varieties grown for several years at each of a sample of sites representing the cotton growing area in the United States.

It should be pointed out that the classification of environments obtained by a method such as used herein is a function of the varieties used to measure the interaction effects. Compared to random selections, the varieties used in this particular case might tend to give a clustering biased in favor of the current zoning insofar as the admission of varieties to the regional testing program is conditioned on consistency of performance throughout one or more of the regions. There is no way from the data available to measure the magnitude of this bias or even to verify its presence. It should not be presumed, however, that such bias is necessarily undesirable since the varieties used in this analysis may be more representative of the future entries in the testing program than random selections would be.

Similarity between locations was based on lint yield data. Although yield is probably the most important character one could expand the vector of estimated interaction effects to include all traits of importance. In this case the distance coefficient between locations i and i′ will be defined as:

$$d_{ii'} = [\sum_{j=1}^{c} w_j d_{ii'\,j}^{2}]^{1/2}$$

where $d_{ii'j}$ and $w_j$ are the distance coefficient and the relative weight for trait j and c is the number of traits included. The relative weights are presumably a function of the relative economic importance as well as the relative importance of interaction in the trait. For cotton, however, additional traits may not be needed since yield has been previously reported by Miller et al (1962) to be the only character showing appreciable variety by environment interaction.

[4] These analyses are performed by the Biometrical Services, USDA, Beltsville, Md.

## LITERATURE CITED

Gower, J. C. 1967. A comparison of some methods of cluster analysis. Biometrics 23:623-637.

Horner, T. W., and K. J. Frey. 1957. Methods of determining natural areas for oat varietal recommendations. Agron. J. 49:313-315.

Michener, C. D., and R. R. Sokal. 1957. A quantitative approach to a problem in classification. Evolution 11:130-162.

Miller, P. A., H. F. Robinson, and O. A. Pope. 1962. Cotton variety testing: Additional information on variety × environment interactions. Crop Sci. 2:349-352.

Rohlf, F. J., and R. R. Sokal. 1965. Coefficients of correlation and distance in numerical taxonomy. Kansas Univ. Sci. Bull. 45:3-27.

Sokal, R. R. 1961. Distance as a measure of taxonomic similarity. Systematic Zool. 10:70-79.

————, and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Kansas Univ. Sci. Bull. 38:1409-1438.

————, and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Company.