Theor. Appl. Genet. 62, 277–280 (1982)

Grouping Genotypes by a Cluster Method Directly Related to Genotype-environment Interaction Mean Square*

C.-S. Lin

Statistical Research Section, Engineering and Statistical Research Institute, Agriculture Canada, Ottawa (Canada)

Summary. To facilitate the interpretation of data from a genotype by environment (GE) experiment when the GE interaction is large, a cluster method is proposed to group genotypics according to their response to the environments. The dissimilarity index between a pair of genotypes is defined in terms of distance adjusted for the average effects of genotypes, and Sokal and Michener's (1958) unweighted pair-group method is used in the clustering algorithm. The new index, constructed in each cluster cycle for any group, is shown to be equivalent to within group GE interaction mean square under 2-way ANOVA. Thus, if the F-value is used as an empirical stopping criterion for clustering, there will be no significant GE interaction within groups and the genotypes within the groups can be compared by their average effects. The method of analysis is illustrated by an example.

Key words: Grouping genotypes – Genotype-environment interaction – Cluster analysis

Introduction

In a 2-way classification, two aspects of the data structure are of special importance; the "level"aspect represented by the marginal means (or average effects), and the "shape" aspect represented by the differential responses of individual to one factor at different levels of the other factor. When response "shapes" are similar, straightforward comparison by "levels" is possible. However if the "shapes" are different, which implies the presence of interaction, comparisons by "levels" may be misleading. An analytical way of investigating the data under these circumstances is to stratify the data by "shapes" and then to compare the "levels" within each stratum. In the context of a genotype by environment (GE) experimentation, this means that genotypes (or environments) should be grouped so that there will be no GE interaction within groups.

Grouping can be done by cluster analysis. In the literature, there are two major approaches for grouping genotypes based on their GE interaction (or GE interaction and genotypic effect jointly). One approach is to consider genotype as a vector of n-attributes represented by n environments and to use the distance coefficient (Abou-El-Fittouh et al. 1969; Hanson 1970) or the squared distance (Mungomery et al. 1974; Johnson 1977) as a similarity index for clustering. The other approach is to impose a linear model for GE interaction based on the environmental index (Finlay and Wilkinson 1963) and to use the deviation mean square from a joint regression as dissimilarity index for clustering (Lin and Thompson 1975). The former approach is less restrictive, compared to the latter, in terms of assumptions. However, it does not provide a natural stopping criterion for clustering. Thus if groups are to be constrained in such a way that the mean square of the GE interaction should not be significantly different from the estimated error, then the stopping point has to be determined by the 2-way ANOVA on a "trial and error" basis (see, e.g. Ghaderi et al. 1980). This is too laborious if the number of genotypes is large. In contrast, the regressions approach of Lin and Thompson's (1975) method provides a natural and well-defined stopping criterion which would obtain groups with the same intercept and the same slope (the former represents average effect and the latter the GE interaction). However this approach requires that residual MS's from regressions are homogeneous with respect to genotypes since its dissimilarity index is based on the test statistic for a joint regression.

From the point of view of stratifying the data, imposing a structural model for GE interaction is not necessary but having a well-defined stopping criterion is important. These requirements can be most conveniently fulfilled if the index has some functional relationship with the GE interaction mean square in the 2-way ANOVA. Williams (personal communication, see also De Pauw et al. 1981) thus proposed an algorithm by defining the dissimilarity index as the GE interaction mean square and the new indices constructed in each cluster cycle are calculated from the data of grouped genotypes.

In this paper, it will be shown that same result can be obtained based on the method of Abou-El-Fittouh

* Contribution no. I-348 from the Engineering and Statistical Research Institute

0040-5752/82/0062/0277/$ 1.00

et al. (1969), with a slight adjustment to their distance coeffcient. Thus, a direct link between a cluster method and the GE interaction mean square is obtained. The data from the paper of Yates and Cochran (1938) are used as an example to illustrate the method of analysis.

Method

Let y_{ij} be the observed value of the i-th ($i = 1, \ldots, m$) genotype in the j-th ($j = 1, .., n$) environment and let the dissimilarity index between two genotypes i and i′ be defined as

$$d(i, i') = 1/[2(n-1)] \sum_{j=1}^{n} [(y_{ij} - \bar{y}_{i.}) - (y_{i'j} - \bar{y}_{i'.})]^2, \quad (1)$$

where

$$\bar{y}_{i.} = \sum_{j=1}^{n} y_{ij}/n \quad \text{and} \quad y_{i'.} = \sum_{j=1}^{n} y_{i'j}/n.$$

(Note $d(i, i') = n/[2(n-1)]\, d_{ii'}^2$;

where $d_{ii'}$ is Abou-El-Fittouh et al. (1969) distance coefficient.)

If Sokal and Michener's (1958) unweighted pair-group method is used for the clustering algorithm, the new dissimilarity index for a subset of r genotypes can be written

$$d(1, 2, \ldots, r) = 2/[r(r-1)] \sum_{1 \leq i < i'}^{r} d(i, i'). \quad (2)$$

Lemma

$$d(1, 2, \ldots, r) \quad (3)$$
$$= 1/[(r-1)(n-1)] \sum_{i=1}^{r} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2,$$

where

$$\bar{y}_{.j} = \sum_{i=1}^{r} y_{ij}/r \quad \text{and} \quad \bar{y}_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n} y_{ij}/r\,n.$$

Proof

By definition

$$2(n-1)\, d(i, i') = \sum_{j=1}^{n} [(y_{ij} - \bar{y}_{i.}) - (y_{i'j} - \bar{y}_{i'.})]^2$$
$$= \sum_{j=1}^{n} [(y_{ij} - y_{i'j})^2 - n(\bar{y}_{i.} - \bar{y}_{i'.})^2.$$

It was shown by Gaylor (1956) that

$$\sum_{1 \leq i < i'}^{r} \left[\sum_{j=1}^{n} (y_{ij} - y_{i'j})^2 - n(\bar{y}_{i.} - \bar{y}_{i'.})^2 \right]$$
$$= r \sum_{i=1}^{r} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

Therefore

$$2(n-1)/r \sum_{1 \leq i < i'}^{r} d(i, i') = \sum_{i=1}^{r} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2,$$

and

$$d(1, 2, \ldots, r)$$
$$= 1/[(r-1)(n-1)] \sum_{i=1}^{r} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

QED

Equation (3) indicates that if the dissimilarity index for a pair of genotypes is defined by the squared distance adjusted for the average effects of genotypes (Eq. 1) and if the unweighted pair-group method is used in the clustering algorithm (Eq. 2), then the new dissimilarity indices constructed in each cluster cycle will be equivalent to the GE interaction mean square for corresponding genotypes under the 2-way ANOVA. Thus if the F-value for the GE interaction mean square, as tested against the esimated error, is used as the empirical criterion, and the cluster process is stopped when the smallest dissimilarity index in the circle exceeds the critical F-value, then the resulting group will have no significant GE interaction.

Table 1. Two-way table of variety by location (based on the total of three replications and two years' data), cited from the paper by Yates and Cochran (1938)

Variety	Coded	Location						Mean
		1	2	3	4	5	6	
'Manchuria'	(1)	161.7	247.0	185.4	218.7	165.3	154.6	188.8
'Svansota'	(2)	187.7	257.5	182.4	183.3	138.9	143.8	182.3
'Velvet'	(3)	200.1	262.9	194.9	220.2	165.8	146.3	198.4
'Trebi'	(4)	196.9	339.2	271.2	266.3	151.2	193.6	236.4
'Peatland'	(5)	182.5	253.8	219.2	200.5	184.4	190.1	205.1
Mean		185.8	272.1	210.6	217.8	161.1	165.7	202.2

Example

The example is taken from Yates and Cochran's (1938) paper: a 2-way table of barley yield (totalled for 2 years and 3 replications) for five varieties at six locations is shown in Table 1.

The dissimilarity indices based on Eq. (1) for all pairs of genotypes are calculated. The m × m matrix for the first clustering cycle see Table 2.

Because d(2, 3) is the smallest index in the cycle, varieties 2 and 3 are clustered. The dissimilarity matrix for the second clustering cycle based on the unweighted pair-group method is shown in Table 3; where, for example, d(1, 2, 3) is calculated in the following way,

$$d(1, 2, 3) = 1/3[d(1, 2) + d(1, 3) + d(2, 3)] = 160.7.$$

As a check, ANOVA for varieties 1, 2, 3, gives the following results (Table 4):

i.e. $d(1, 2, 3) = MS(V \times L) = 160.7$.

By proceeding in the same way up to (m − 1) cycle, the smallest index in each clustering cycle can be summarised as follows (Table 5).

The mean square for error converted from the original paper is $23.28 \times 2 \times 3 = 139.7$ with 216 df. Using this value for the F test, V × L is significant at the fourth cycle suggesting two groups; variety 4 and the four other varieties. It appears that the significant GE interaction shown in this set of data is attributable to variety 4. For the other four varieties 1, 2, 3, 5 the interaction is not significant and thus they can be compared by their average effects. Newman-Keuls' multiple test of means (see, e.g. Steel and Torrie 1960) gives the following results, see Table 6.

Table 2.

Coded varieties	1	2	3	4
2	260.7			
3	133.8	88.5		
4	748.3	658.3	755.0	
5	198.3	278.1	336.5	976.4

Table 3.

Coded varieties	1	2, 3	4
2, 3	160.7		
4	748.3	500.3	
5	198.3	234.0	976.4

Source	df.	MS
Variety (V)	2	393.5
Location (L)	5	4524.2
V × L	10	160.7

Table 5.

Cluster cycle	Varieties grouped	Smallest index	Calculated F-value	Tabular F-value ($\alpha = 5\%$)
1	2, 3	87.5	0.63	3.84
2	(2, 3), 1	160.7	1.15	3.00
3	(1, 2, 3), 5	215.8	1.54	2.60
4	(1, 2, 3, 5), 4	443.3	3.17	2.37

Table 6.

Variety	5	3	1	2
Mean	205.1	198.4	188.8	182.3

The underlining indicates that the differences between the respective varieties are not significant at the 5% level.

Conclusion

Interpretation of 2-way classification data when interaction is present is often difficult. This is particularly true when the numbers of categories in each classification are large. The present cluster method was prepared as an analytical tool for investigating such data. Stratification of the data by the similarity of response "shape" provides a logical base to compare the individuals within strata (groups) by their average effect, and also makes it easier to identify the interaction structure.

The criterion for the same response "shape" defined in the present method (and in the context of a genotype-environment experiment) is the size of the GE interaction mean square. Genotypes are considered to respond in the same way to environments if their GE interaction mean square is not significantly different from the estimated error. Because all the indices, defined or constructed, are GE interaction mean squares for the corresponding genotypes, grouping based on the critical F-value as a stopping point should satisfy this criterion and the resulting response "shapes" within group should be approximately homogeneous.

Acknowledgement

I wish to thank Dr. C. J. Williams for useful discussion.

Literature

Abou-El-Fittouh, H.A.; Rawling, J.O.; Miller, P.A. (1969): Classification of environments to control genotype by environment interactions with an application to cotton. Crop Sci. **9,** 135–140

De Pauw, R.M.; Faris, D.G.; Williams, C.J. (1981): Genotype-environment interaction of yield in cereal crops in northwestern Canada. Can. J. Plant Sci. **61,** 255–263

Finlay, K.W.; Wilkinson, G.M. (1963): The analysis of adaptation in plant-breeding program. Aust. J. Agric. Res. **14,** 742–754

Gaylor, D.W. (1956): Equivalence of two estimates of product variance. J. Am. Stat. Assoc. **51,** 451–453

Ghaderi, A.; Everson, E.H.; Cress, C.E. (1980): Classification of environments and genotypes in wheat. Crop. Sci. **20,** 707–710

Hanson, W.D. (1970): Genotypic similarity. Theor. Appl. Genet. **40,** 226–231

Johnson, G.R. (1977): Analysis of genotypic similarity in terms of mean yield and stability of environmental response in a set of maize hybrids. Crop Sci. **17,** 837–842

Lin, C.S.; Thompson, B. (1975): An empirical method of grouping genotypes based on a linear function of the genotype-environment interaction. Heredity **34,** 255–263

Mungomery, V.E.; Shorter, R.; Byth, D.E. (1974): Genotype × environment interactions and environmental adaptation. I Pattern analysis – application to soya bean populations. Aust. J. Agric. Res. **25,** 59–72

Sokal, R.R.; Michener, C.D. (1958): A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38,** 1409–1438

Steel, R.G.D.; Torrie, J.H. (1960): Principles and Procedures of Statistics New York: McGraw-Hill

Yates, F.; Cochran, W.G. (1938): The analysis of groups of experiments. J. Agric. Sci. **28,** 556–580

Received February 23, 1982
Communicated by R. W. Allard

Dr. C.-S. Lin
Statistical Research Station
Engineering and Statistical Institute
Agriculture Canada
Ottawa K1A OC6 (Canada)