Theor Appl Genet (1983) 66:131–133

HASS cluster analysis: a new method of grouping genotypes or environments in plant breeding

T. B. Ramey and A. A. Rosielle
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Received March 3, 1983
Communicated by P. M. A. Tigerstedt

Summary. A new method of cluster analysis, termed the hierarchical agglomerative sums of squares method (HASS), is proposed to facilitate clustering of genotypes or environments where genotype × environment interaction exists. The method is a modification of that proposed by Lin (1982), who used the equivalent of the genotype × environment interaction mean squares for clustering. Lin fused on the minimum genotype × environment interaction mean square within potential new clusters at each fusion cycle. HASS clustering, however, uses the pooled genotype × environment interaction sums of squares within all clusters (new and old) at each fusion cycle. HASS clustering is shown to more nearly achieve the objective of minimizing the pooled interaction mean square within clusters and maximizing the interaction mean square among clusters when compared with Lin's (1982) method and the widely used average linkage method.

Key words: Genotype × environment interaction – HASS clustering – Cluster analysis – Grouping genotypes – Grouping environments

Introduction

Genotype × environment interaction is a problem in most plant breeding programs concerned with improvement of quantitative traits such as grain yield. One method breeders have used to reduce the impact of these interactions is to stratify genotypes or environments into groups so that interactions within groups are minimized. Several methods have been proposed to achieve this objective. Horner and Frey (1957) grouped locations by empirically examining variety × location interaction mean squares for various combinations of environments and choosing groups which had low interactions. The most widely used technique, however, for classifying environments or genotypes into groups has been cluster analysis.

Cluster analysis operates on a matrix of interclass dissimilarities (or similarities) where the class may be either genotypes or environments, depending on which is being clustered. Dissimilarity indexes used have been distances, squared distances, and correlation coefficients (Abou-El-Fittouh et al. 1969; Mungomery et al. 1974; Campbell and Lafever 1977). In addition, various attributes have been clustered including raw data, standardized data, and interaction effects (Abou-El-Fittouh et al. 1969; Fox and Rosielle 1982).

A fundamental aim of all methods of cluster analysis is the grouping of genotypes or environments in a way which minimizes the genotype × environment interaction mean square within clusters. Lin (1982) showed that dissimilarity indexes based on genotype × environment interaction mean squares within clusters are equivalent to those based on averaged squared distances between class attributes adjusted for the main effects of the class. Lin's (1982) fusion method involved calculating at each fusion cycle the genotype × environment interaction mean square within potential new clusters and fusing on the minimum. However, the objective of grouping genotypes or environments should be to minimize the genotype × enviroment interaction mean square within all clusters, not merely that within potential new clusters.

In this paper, a new procedure is described, called Hierarchical Agglomerative Sums of Squares Clustering (HASS), which achieves the aim of minimizing overall genotype × environment interaction mean square within clusters at each fusion cycle. HASS clustering is shown to be superior not only to Lin's (1982) method, but also to the commonly used agglomerative method of minimizing the average distance (linkage) between members of 2 separate clusters which are to be fused (Dickson et al. 1981). The method is illustrated by example.

Methods

Let Y_{ij} be the observed value of the i^{th} genotype ($i = 1, 2 \ldots r$) in the j^{th} environment ($j = 1, 2, \ldots, n$). The interaction effect in any cell of the two-way table of genotypes and environments is given by:

$$I_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

where

$$\bar{Y}_{i.} = \sum_{j=1}^{n} Y_{ij}/n$$

$$\bar{Y}_{.j} = \sum_{i=1}^{r} Y_{ij}/r,$$

and

$$\bar{Y}_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n} Y_{ij}/n\,r.$$

Lin (1982) defined a dissimilarity index between two genotypes as:

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

$$= 1/[2(n-1)] \sum_{j=1}^{n} [I_{ij} - I'_{ij}]^2. \quad (2)$$

The dissimilarity index for a subset r_k of the r genotypes was defined by Lin (1982) as:

$$d(1, 2, \ldots, r_k) = 2/[r_k(r_k - 1)] \sum_{1 \leq i_k < i'_k}^{r_k} d(i_k, i'_k). \quad (3)$$

And, he showed:

$$d(1, 2, \ldots, r_k) = 1/[(r_k - 1)(n-1)] \sum_{i=1}^{r_k} \sum_{j=1}^{n} I_{ij}^2. \quad (4)$$

Thus, Lin's dissimilarity index at any level of clustering is equivalent to the genotype × environment interaction mean square within a newly fused cluster.

Equation (4) shows that the sum of squares for genotype × environment interaction within the k^{th} cluster is:

$$SS(k) = [(r_k - 1)(n-1)]\, d(1, 2, \ldots, r_k).$$

Substituting for $d(1, 2, \ldots, r_k)$ from equation (3) gives:

$$SS(k) = 2(n-1)/r_k \sum_{1 \leq i_k < i'_k}^{r_k} d(i_k, i'_k). \quad (5)$$

Finally, if there are c clusters after f fusion cycles the total sum of squares for genotype × environment interaction within clusters is:

$$2(n-1) \sum_{k=1}^{c} (1/r_k) \sum_{1 \leq i_k < i'_k}^{r_k} d(i_k, i'_k). \quad (6)$$

The method we propose, HASS clustering, uses equation (6) at each fusion cycle to minimize the total sum of squares for genotype × environment interaction within clusters.

The HASS method was used to group environments in the data set of Table 1: a two-way table of mean yields of 15 wheat cultivars grown at 9 locations in Western Australia in 1975. The HASS method was compared to Lin's (1982) method (equation 3) and the widely used average linkage method (Dickson et al. 1981). The average linkage method is based on the equation:

$$d(k, k') = 1/(r_k\, r_{k'}) \sum_{i_k=1}^{r_k} \sum_{i_{k'}=1}^{r_{k'}} d(i_k, i_{k'}) \quad (7)$$

where k and k′ refer to a pair of clusters present at the previous fusion cycle, r_k and $r_{k'}$ are the respective number of items within each cluster, and $d(i_k, i_{k'})$ is the distance between item i_k in cluster k and $i_{k'}$ in cluster k′, as defined by equation 1.

Results and discussion

Table 2 shows that the HASS method of cluster analysis was generally superior to the methods of both Lin (1982) and the average linkage method when minimum mean square for cultivar × location interaction within clusters is used as the criterion. All three methods give identical results at the first and last fusion. In the remaining six fusion cycles, the HASS method was superior in 3, identical in 2, and inferior in 1 fusion

Table 1. Mean yields (kg/ha) of 15 wheat cultivars grown at 9 locations in Western Australia in 1975

	Location									
Cultivar	1	2	3	4	5	6	7	8	9	Mean
1	2831	3159	1359	3802	2793	2539	1383	1828	1336	2337
2	2273	5737	1476	4218	2343	2500	1481	2016	1327	2597
3	2709	3768	1523	3463	2754	2422	1570	1744	1230	2354
4	2545	4443	1664	3671	2773	2773	1926	1828	973	2511
5	2807	5071	1500	4348	2343	2558	1261	1744	1008	2516
6	1425	3083	961	2682	1504	1543	1008	703	187	1455
7	1547	2231	1289	3359	1699	3007	1903	1931	1101	2007
8	2690	3609	1500	3073	2968	2715	1715	1898	1265	2381
9	2695	4265	1828	3854	2383	2246	1542	1870	1265	2439
10	2868	2578	1383	3567	2617	2597	1476	2034	1277	2266
11	2709	4518	1898	3177	2812	2597	1251	1706	926	2399
12	2358	4321	1687	3619	2597	2578	1228	1575	762	2303
13	2461	4434	1758	3567	2422	2461	1205	1622	621	2283
14	2986	4228	1781	4036	2363	2558	1495	1603	1183	2470
15	2634	3703	1804	3828	2539	2558	1458	1894	1219	2404
Mean	2503	3943	1561	3618	2461	2510	1460	1733	1045	2315

Table 2. Degrees of freedom, location groupings, and mean squares for cultivar × location interaction within clusters using three methods of clustering

Fusion cycle	Degrees of freedom for cultivar × location interaction within clusters	Location groupings			Mean squares for cultivar × location interaction within clusters		
		HASS	Lin	Average linkage	HASS	Lin	Average linkage
1	14	(8, 9)	(8, 9)	(8, 9)	12,763	12,763	12,763
2	28	(6, 7); (8, 9)	(6, 8, 9)	(6, 7); (8, 9)	20,162	24,296	20,162
3	42	(1, 5); (6, 7); (8, 9)	(6, 7, 8, 9)	(6, 7, 8, 9)	25,141	27,912	27,912
4	56	(1, 5); (6, 7, 8, 9)	(1, 5); (6, 7, 8, 9)	(1, 5); (6, 7, 8, 9)	29,708	29,708	29,708
5	70	(1, 3, 5); (6, 7, 8, 9)	(1, 5); (3, 6, 7, 8, 9)	(1, 5); (3, 6, 7, 8, 9)	37,303	37,838	37,838
6	84	(1, 3, 5); (4, 6, 7, 8, 9)	(1, 5); (3, 4, 6, 7, 8, 9)	(1, 3, 5, 6, 7, 8, 9)	52,160	51,796	57,804
7	98	(1, 3, 4, 5, 6, 7, 8, 9)	(1, 3, 4, 5, 6, 7, 8, 9)	(1, 3, 4, 5, 6, 7, 8, 9)	67,606	67,606	67,606
8	112	(1, 2, 3, 4, 5, 6, 7, 8, 9)	(1, 2, 3, 4, 5, 6, 7, 8, 9)	(1, 2, 3, 4, 5, 6, 7, 8, 9)	145,845	145,845	145,845

cycle to Lin's (1982) method. Compared to the average linkage method, the HASS method was superior in 3 and equivalent in 3 of the 6 fusion cycles.

The fact that the HASS method was inferior in 1 fusion cycle to Lin's (1982) method shows that the method will not always result in optimum grouping because of restrictions that are imposed by previous fusion cycles. Such restrictions result from the hierarchical nature of the method, which, at fusion cycle f, restricts the choice of new groupings to a fusion of 2 clusters already obtained at fusion cycle f−1. Thus, HASS clustering was inferior to Lin's (1982) method at the sixth fusion cycle because of the restraint imposed by the inclusion of location 3 in a different cluster at the fifth fusion cycle (Table 2).

Minimization of the total sum of squares for interaction within clusters by equation (6) is equivalent to maximization of the total sum of squares for interaction among clusters, since the sum of these components is equal to the total interaction sum of squares. At fusion cycle f, the degrees of freedom within clusters is $f(n-1)$ and the degrees of freedom among clusters is $(r-f-1)(n-1)$ where r is the total number of items to be clustered and n is the number of measurements on each item. The degrees of freedom could also be expressed in terms of c, the total number of clusters at a fusion cycle (including single item clusters), because $f = r - c$. Thus, the degrees of freedom within clusters is $(r-c)(n-1)$, and among clusters is $(c-1)(n-1)$. Because the among and within cluster degrees of freedom are constant for any particular fusion cycle, minimization of the total sum of squares for interaction within clusters results in minimization of the pooled interaction mean square within clusters and maximization of the interaction mean square among clusters. Therefore, HASS clustering will produce the best partitioning of the within and among cluster interaction mean squares that a hierarchical agglomerative method can produce. As noted before, depending on restrictions imposed by previous fusion cycles, other methods may give superior results at particular fusion cycles. However, if the results from the previous fusion cycle are equivalent using different methods, the HASS method cannot produce an inferior result in the next fusion cycle.

The key difference between HASS clustering and Lin's (1982) method is that Lin determined mean squares for new clusters only at each fusion cycle, while HASS clustering determines the pooled within cluster interaction sum of squares over all clusters, old and new, present at each fusion cycle. HASS clustering could also operate on mean squares, but would require that the overall within cluster mean square be calculated from the pooled sum of squares.

References

Abou-El-Fittouh HA, Rawlings JO, Miller PA (1969) Classification of environments to control genotype by environment interactions with an application to cotton. Crop Sci 9: 135–140

Campbell LG, Lafever HN (1977) Cultivar × environment interactions in soft red winter wheat yield tests. Crop Sci 17: 604–608

Dickson WJ, Brown MB, Engelman L, Frane JW, Hill MA, Jennrich RI, Toporek JD (1981) BMDP statistical software 1981. University California Press, Berkeley

Fox PN, Rosielle AA (1982) Reducing the influence of environmental main-effects on pattern analysis of plant breeding environments. Euphytica 31:645–656

Horner TW, Frey KJ (1957) Methods for determining natural areas for oat varietal recommendations. Agron J 49: 313–315

Lin CS (1982) Grouping genotypes by a cluster method directly related to genotype-environment interaction mean square. Theor Appl Genet 62: 277–280

Mungomery VE, Shorter R, Byth DE (1974) Genotype × environment interactions and environmental adaptation. 1. Pattern analysis – application to soya bean populations. Aust J Agric Res 25: 59–72