J. Rubb. Res. Inst. Malaysia, 31(3), 175–187

Investigating Cluster Analysis as a Technique for Grouping Laboratories in SMR Round-Robin Interlaboratory Crosscheck Programme

LEONG YIT SAN

In most rounds, three compact and natural clusters are found by Ward's method and mode analysis. Ward's method, the relocation procedure and the mode analysis all generated topologically similar structures. When the eight clusters formed by Ward's method are superimposed on Youden's diagram, clusters of laboratories behaving similarly and lying close to one another in the Youden's diagram are produced indicating that Ward's method with the Euclidean measure of similarity tends to cluster together anomalous or acceptable laboratories. The existence of a cluster of size one out of three or four clusters of a population of about twenty-five laboratories may be used to signal the occurrence of an outlier. Such an outlier usually lies furthest away from the centre of the ellipse in either the first or third quadrant.

Youden's method[1] of analysing interlaboratory crosscheck data leads to the identification of anomalous laboratories which tend to overestimate or underestimate initial Wallace Plasticity (P_o). Such analysis results in a form of classifying or grouping of the laboratories, usually into three possible groups. The three clusters or groups can be tentatively visualised as a grouping of laboratories which tends to overestimate, underestimate or have acceptable results. Cluster analysis is usually employed in the formation of groups based on some similarity measure using a fusion or relocation procedure. It remains to be seen whether cluster analysis will lead to a classification similar to that of Youden's method where anomalous laboratories are grouped together.

The descriptive summarisation of large quantities of multi-variate data by clusters, undefined *apriori*, is increasingly practised by taxonomists[2–5]. Cluster analysis is a procedure in which one objectively groups together entities on the basis of their similarities and differences. Some of the purposes of performing cluster analysis[6,7] are as follows:

- To group the entities in a convenient manner for mental clarification and communication in the same sort of way as the values of a single variable are grouped in a frequency distribution.
- To search for specific sorts of organisational structure or grouping of the data
- To discover new fields of research
- To be used as a check list to explain previous findings or observed groupings.

COMMUNICATION 726

Most authors emphasise the first and third purposes of using cluster analysis to arrive at a useful description of the entities sampled for administrative purpose and to discover unsuspected clusterings which may prove to be important in reducing a sample scatter into a number of component clumps in order to search for regions of continuous density. This paper is mainly concerned with the first, second and fourth purposes of using cluster analysis *i.e.* to check whether cluster analysis will lead to a natural formation of clusters of anomalous laboratories and acceptable laboratories and perhaps provide some new information to explain the existing deficiency of laboratories not producing complete agreement among laboratories. The principles and implications of the results of applying the techniques to SMR round robin data which may provide a further insight into the mechanism of the clustering process and the cluster structure in summarising the interrelationships among the laboratories are discussed.

Measures of Similarity

Basic to any clustering process is the notion of similarity and differences (dissimilarity). Two laboratories P and Q are said to belong to the same cluster if the distance between their points is sufficiently small and to different clusters if the distance is sufficiently large. Some of the distance functions commonly used are:

- Euclidean distance
- City block metric
- Minkowsky distance
- Angular separation
- Correlation
- Profile similarity index
- Coefficient of nearness
- Canberra metric
- Mahalanobis distance
- Dispersion
- Jaccard coefficient
- Shape difference
- Size difference

The Euclidean measure of similarity (or error sum of squares) is discussed here primarily because it can be related to Youden's method. It measures the extent of the scatter about cluster centres and the results are characterised by the close clumping of points into spherical clusters of similar size. It is suitable for finding tight clusters which have the property that each cluster centre represents the constituent laboratories at a high level of similarity with respect to all the underlying variables.

MATERIALS AND METHODS

Clustering Techniques

Clustering techniques basically fall into four groups as follows:

- Hierarchic fusion methods
- Iterative relocation methods
- Monothetic division methods (for binary data)
- Other methods such as mode analysis and density method

Since the present data contain continuous variables, monothetic division methods are not examined.

Hierarchic Fusion Methods

Eight hierarchic fusion methods[8,9] described in detail in *Appendix A* were studied. They are:

- Single linkage
- Complete average
- Group average
- Centroid
- Median
- Ward

- Ward (standardised)
- Lance Williams
- McQuitty

Iterative Relocation Methods

Iterative relocation[10-13] with four different initial seed points as described in *Appendix B* was examined. The classifications produced by these four methods were used as initial seed points in the clustering process of the iterative relocation methods. They are:

- Ward
- Random
- Size
- Shape

Mode Analysis

Mode analysis as described in *Appendix C* was also examined.

A Combined Method

Since there is an abundance of clustering techniques detailed studies of the relative merits of hierarchic fusion, iterative relocation methods or mode analysis will involve lengthy reports. This study does not compare hierarchial representations[14,4] by distortion measures such as cophenetic correlations. It is suggested by Wishart[5] that the following steps be performed for populations of size less than 150:

1. Use Ward's method selecting an output from eight clusters down to one.
2. Run the relocation method to select eight cluster groupings from *Step 1* as the starting classification using the error sum of squares as the similarity coefficient.
3. Re-run the relocation procedure from a random generation of the initial classification of eight clusters. Compare the configuration with the final configuration obtained in *Step 2*.
4. Execute a mode analysis.

RESULTS AND DISCUSSION

Table 1 shows the merge list when Ward's method of hierarchic fusion is applied to *Round Robin 37-1*. The list shows the successive levels at which the fusion process occurs and the coefficient in the table denotes the total within cluster sum of squares. Visual interpretation of the list is provided by the dendrogram of *Figure 1*. Using the suggestion that the fusion process could be stopped when a significant drop or discontinuity in the fusion coefficient value was observed, it was clear that three compact clusters were detected at the coefficient level of 2.6. This result was confirmed by executing a mode analysis as well as a density procedure. Eighteen dense points were found with an enclosure ratio of 0.76. Detailed results of the analyses as well as other hierarchic procedures and cluster diagnostic statistic are not presented here due to the lengthy output obtained. However, a concise summary *(Tables 2* and *3)* of all the analyses is given here. It was evident that Ward's method, the relocation procedure with Ward's classification as the initial seed points, the relocation procedure with random initial seed points and the mode analysis all generated topologically similar structures for three clusters. Further, the results of Ward's method and mode analysis obtained by normalising the variables as recommended by most workers were similar to those using unnormalised variables. On the other hand, the structures produced by the single linkage, complete linkage, the group average linkage, the centroid, the median, Lance-Williams and McQuitty hierarchic fusion methods appeared to

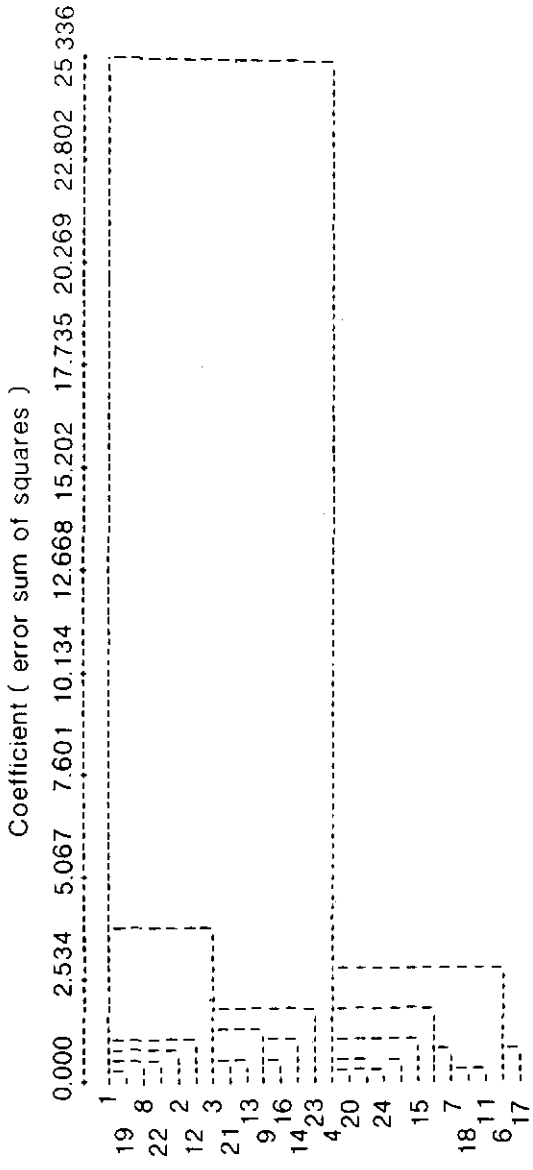

Figure 1. Dendogram of Round Robin 37-1 produced by Ward's method.

differ from the structure of Ward's method. *Laboratory 17* appeared to be an outlier or anomalous when the former methods were used; a result which was also obtained by the Youden's method[1,15]. Similar results were also obtained in the case of eight clusters, the only difference being the hierarchic structure obtained from the complete linkage fusion method did not differ substantially from the structure obtained by Ward's method. This result confirmed the finding by Khir[14] who reported that Ward's method seemed to be the most efficient.

The final clusterings obtained by the iterative relocation method using random initial seed points did not differ substantially from the same relocation procedure using Ward's classification as the initial seed point. Further, in most cycles, a single iteration was sufficient to obtain a stable cluster. Using the suggestion by Wishart[5] to find a global optimum, it was found that the final clusterings obtained using the initial seed points of the classification by the size and shape coefficients were similar to the iterative relocation procedure using random initial seed points indicating a global optimum had been obtained for three clusters.

Bearing in mind that the existence of a cluster containing a single isolated laboratory might be used to signal the occurrence of an outlier, it was decided to investigate how effective this signal was. *Table 4* shows the cluster of size one obtained from Ward's method for eight clusters in *Round Robins 37-1* to *34-2*. It was found that in *Round Robin 35-2*, if three clusters were formed, one of them would be of size one containing *Laboratory 22*. The total within error sum of squares dropped from 22.29 to 10.02 when three clusters were formed, indicating that *Laboratory 22* was anomalous. This finding was in agreement with the finding of Youden's method.

It must be pointed out that there is no statistical test to detect outliers in this manner but clearly such a test depends on the change in the total error sum of squares, the number of clusters in which a cluster of size one or perhaps two is

TABLE 1 MERGE LIST FOR WARD'S METHOD IN ROUND ROBIN 37-1
OUTPUT CLASSIFICATIONS FOR 1 TO 8 CLUSTERS

CYCLE	1	NOW FUSE POINTS	1	19	AT COEFFICIENT[a]	0.200 –	24	CLUSTERS AND NEW CLUSTER CODE IS	1
CYCLE	2	NOW FUSE POINTS	7	18	AT COEFFICIENT	0.238 –	23	CLUSTERS AND NEW CLUSTER CODE IS	7
CYCLE	3	NOW FUSE POINTS	4	20	AT COEFFICIENT	0.263 –	22	CLUSTERS AND NEW CLUSTER CODE IS	4
CYCLE	4	NOW FUSE POINTS	4	10	AT COEFFICIENT	0.338 –	21	CLUSTERS AND NEW CLUSTER CODE IS	4
CYCLE	5	NOW FUSE POINTS	4	24	AT COEFFICIENT	0.369 –	20	CLUSTERS AND NEW CLUSTER CODE IS	4
CYCLE	6	NOW FUSE POINTS	7	11	AT COEFFICIENT	0.380 –	19	CLUSTERS AND NEW CLUSTER CODE IS	7
CYCLE	7	NOW FUSE POINTS	1	8	AT COEFFICIENT	0.384 –	18	CLUSTERS AND NEW CLUSTER CODE IS	1
CYCLE	8	NOW FUSE POINTS	9	16	AT COEFFICIENT	0.388 –	17	CLUSTERS AND NEW CLUSTER CODE IS	9
CYCLE	9	NOW FUSE POINTS	3	21	AT COEFFICIENT	0.438 –	16	CLUSTERS AND NEW CLUSTER CODE IS	3
CYCLE	10	NOW FUSE POINTS	4	25	AT COEFFICIENT	0.492 –	15	CLUSTERS AND NEW CLUSTER CODE IS	4
CYCLE	11	NOW FUSE POINTS	3	13	AT COEFFICIENT	0.580 –	14	CLUSTERS AND NEW CLUSTER CODE IS	3
CYCLE	12	NOW FUSE POINTS	1	22	AT COEFFICIENT	0.592 –	13	CLUSTERS AND NEW CLUSTER CODE IS	1
CYCLE	13	NOW FUSE POINTS	1	2	AT COEFFICIENT	0.645 –	12	CLUSTERS AND NEW CLUSTER CODE IS	1
CYCLE	14	NOW FUSE POINTS	5	7	AT COEFFICIENT	0.653 –	11	CLUSTERS AND NEW CLUSTER CODE IS	5
CYCLE	15	NOW FUSE POINTS	6	17	AT COEFFICIENT	0.838 –	10	CLUSTERS AND NEW CLUSTER CODE IS	6
CYCLE	16	NOW FUSE POINTS	9	14	AT COEFFICIENT	0.963 –	9	CLUSTERS AND NEW CLUSTER CODE IS	9

WARDS METHOD GROUP 17 FUSE POINTS							1	12	AT COEF	1.080		8 CLUSTERS							
1	1	3	4	5	6	5	1	9	4	5	1	3	9	15	9	6	5	1	4
3	1	23	4	4															
WARDS METHOD GROUP 18 FUSE POINTS							4	15	AT COEF	1.095		7 CLUSTERS							
1	1	3	4	5	6	5	1	9	4	5	1	3	9	4	9	6	5	1	4
3	1	23	4	4															
WARDS METHOD GROUP 19 FUSE POINTS							3	9	AT COEF	1.180		6 CLUSTERS							
1	1	3	4	5	6	5	1	3	4	5	1	3	3	4	3	6	5	1	4
3	1	23	4	4															
WARDS METHOD GROUP 20 FUSE POINTS							3	23	AT COEF	1.683		5 CLUSTERS							
1	1	3	4	5	6	5	1	3	4	5	1	3	3	4	3	6	5	1	4
3	1	3	4	4															
WARDS METHOD GROUP 21 FUSE POINTS							4	5	AT COEF	1.828		4 CLUSTERS							
1	1	3	4	4	6	4	1	3	4	4	1	3	3	4	3	6	4	1	4
3	1	3	4	4															
WARDS METHOD GROUP 22 FUSE POINTS							4	6	AT COEF	2.790		3 CLUSTERS							
1	1	3	4	4	4	4	1	3	4	4	1	3	3	4	3	4	4	1	4
3	1	3	4	4															
WARDS METHOD GROUP 23 FUSE POINTS							1	3	AT COEF	3.887		2 CLUSTERS							
1	1	1	4	4	4	4	1	1	4	4	1	1	1	4	1	4	4	1	4
1	1	1	4	4															
WARDS METHOD GROUP 24 FUSE POINTS							1	4	AT COEF	25.336		1 CLUSTER							
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1															

[a]Error sum of squares

TABLE 2. CLUSTER CODES OF SIXTEEN CLUSTER ANALYSES FOR THREE CLUSTERS IN ROUND ROBIN 37-1

Laboratory	Hierarchic fusion methods									Relocation method				Mode analysis		
	1	2	3	4	5	6	6a	7	8	1	2	3	4	1	2	3
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	3	3	3
3	1	1	1	1	1	2	1	2	2	1	1	1	1	1	1	1
4	1	2	2	2	2	3	3	3	1	3	3	3	3	3	3	2
5	1	2	2	2	2	3	3	1	1	3	3	3	3	3	3	3
6	1	3	2	2	2	3	3	3	1	3	3	3	3	3	3	2
7	1	2	2	2	2	3	3	1	1	1	1	1	1	3	3	3
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1
10	1	2	2	2	2	3	3	3	1	3	3	3	3	3	3	2
11	1	2	2	2	2	3	3	1	1	3	3	3	3	3	3	3
12	1	2	1	1	1	1	3	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	2	2	2	2	2	2	2	2	1	1	1
14	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1
15	1	2	2	2	3	3	3	3	1	3	3	3	3	3	3	2
16	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1
17	2	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3
18	1	2	2	2	2	3	3	1	1	3	3	3	3	3	3	3
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	2	2	2	2	3	3	3	1	3	3	3	3	3	3	2
21	1	1	1	1	1	2	2	2	2	2	2	2	2	1	2	1
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	3	1	1	1	1	2	2	2	2	2	2	2	2	1	2	1
24	1	2	2	2	2	3	3	3	1	3	3	3	3	3	3	3
25	1	2	2	2	2	3	3	3	1	3	3	3	3	3	3	2

Hierarchic fusion methods

1 = Single linkage
2 = Complete linkage
3 = Group average
4 = Centroid
5 = Median
6 = Ward
6a = Ward (standardised)
7 = Lance Williams
8 = McQuitty

Relocation methods with initial seed points

1 = Ward
2 = Random
3 = Size
4 = Shape

Mode analysis

1 = Mode analysis
2 = Mode (standardised)
3 = Density

TABLE 3. CLUSTER CODES OF SIXTEEN CLUSTER ANALYSES FOR EIGHT CLUSTERS IN ROUND ROBIN 37-1

Laboratory	Hierarchic fusion methods									Relocation method				Mode analysis		
	1	2	3	4	5	6	6a	7	8	1	2	3	4	1	2	3
1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–
2	1	1	1	2	2	1	1	1	1	1	1	2	2	–	–	–
3	1	2	2	1	1	2	1	2	2	2	2	3	3	–	–	–
4	1	3	3	3	3	3	3	3	3	3	3	4	4	–	–	–
5	1	4	3	3	3	4	4	4	3	4	4	5	5	–	–	–
6	2	5	3	3	3	5	5	3	3	5	5	6	6	–	–	–
7	1	4	3	3	3	4	4	4	3	4	4	2	2	–	–	–
8	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–
9	3	2	2	1	1	6	6	2	2	6	6	7	7	–	–	–
10	1	3	3	3	3	3	3	3	3	3	3	4	4	–	–	–
11	1	4	3	3	3	4	4	4	3	4	4	5	5	–	–	–
12	4	6	4	4	4	1	4	4	4	1	1	1	1	–	–	–
13	1	2	2	1	1	2	6	2	2	2	2	3	3	–	–	–
14	5	7	5	5	5	6	2	5	5	6	6	7	7	–	–	–
15	6	3	6	6	6	7	7	6	6	7	7	4	4	–	–	–
16	3	2	2	1	1	6	6	2	2	6	6	7	7	–	–	–
17	7	5	7	7	7	5	5	7	7	5	5	6	6	–	–	–
18	1	4	3	3	3	4	4	4	3	4	4	5	5	–	–	–
19	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–
20	1	3	3	3	3	3	3	3	3	3	3	4	4	–	–	–
21	1	2	2	1	1	2	6	2	2	2	2	3	3	–	–	–
22	1	1	1	1	2	1	1	1	1	1	1	1	1	–	–	–
23	8	8	8	8	8	8	8	8	8	8	8	8	8	–	–	–
24	1	3	3	3	3	3	3	3	3	3	3	5	5	–	–	–
25	1	3	3	3	3	3	3	3	3	3	3	4	4	–	–	–

Hierarchic fusion methods

1 = Single linkage
2 = Complete linkage
3 = Group average
4 = *Centroid*
5 = Median
6 = Ward
6a = Ward (standardised)
7 = Lance Williams
8 = McQuitty

Relocation methods with initial seed points

1 = Ward
2 = Random
3 = Size
4 = *Shape*

Mode analysis

1 = Mode analysis
2 = Mode (standardised)
3 = Density

TABLE 4. CLUSTER OF SIZE ONE IN WARD'S METHOD FOR EIGHT CLUSTERS

Round Robin	Laboratory code			Total number of clusters of size one
37-1	23 (6)	15 (8)	–	2
37-2	20 (6)	24 (7)	–	2
36-1	12 (6)	15 (7)	–	2
36-2	21 (5)	23 (8)	–	2
35-1	14 (5)	15 (6)	11 (8)	3
35-2	22 (3)	–	–	1
34-1	16 (3)	17 (4)	8 (8)	3
34-2	–	–	–	0

Figures in brackets denote the number of clusters in which the cluster of size one was found to be first formed.

detected first and the total number of laboratories. This problem is posed as a challenge to statisticians wishing to develop further the theoretical or statistical basis of this type of test. Our main concern regarding this problem is not to develop such complicated test statistics of the existence of outliers but rather in the implications of such results.

Figure 2 shows the eight clusters formed by Ward's method superimposed on Youden's two-sample diagram[15] in *Round 37-1*. It was clear that Ward's method produced a cluster of laboratories behaving similarly and lying close to one another in the Youden's diagram, *i.e.* laboratories such as *9*, *16* and *14* which overestimated P_o to a similar degree were grouped together whereas *Laboratories 17* and *6* which underestimated P_o to a similar degree formed one of the eight clusters. This means that cluster analysis using Ward's method tends to cluster together anomalous laboratories since *Laboratories 17, 14* and *23* (a cluster by itself) were found to be anomalous by the Youden's method. Further, the Ward's method did not group together laboratories which behaved differently since none of the eight clusters contained laboratories in the first and third quadrant and far away from the centre of the ellipse. The cluster with *Laboratory 23* as its single member was located farthest away at the upper first quadrant. Similar behaviour was observed in *Round Robins 37-2* to *34-2*. It was also clear that eight clusters were necessary to produce a meaningful detailed diagnosis of the structure of round robin data of twenty to twenty-five laboratories instead of three natural clusters as indicated by the mode analysis and the density procedure.

The Euclidean distance used in Ward's method had been found to be a meaningful measure of similarity. It has the property of giving extra weight to outlying values of a single variate[4].

The effect of this property can be seen from *Table 4*. In *Round 37-1, Laboratory 15* was picked out as a cluster of size one. This laboratory had unusually large material differences[15].

CONCLUSION

Sixteen cluster analyses have been applied to SMR round robin data. In most rounds, three compact and natural clusters are found by Ward's method and mode analysis. Ward's method, the relocation procedure with Ward's classification as the initial seed points and random initial seed points as well as the mode analysis all generated topologically similar structures. Further, the results of Ward's method and mode analysis obtained by normalising the variables as recommended by most workers are similar to those using unnormalised variables. On the other hand, the structures produced by the other hierarchic fusion methods except the complete linkage method appear to differ markedly from the structure obtained by Ward's method. The tentative global iterative relocation procedure is

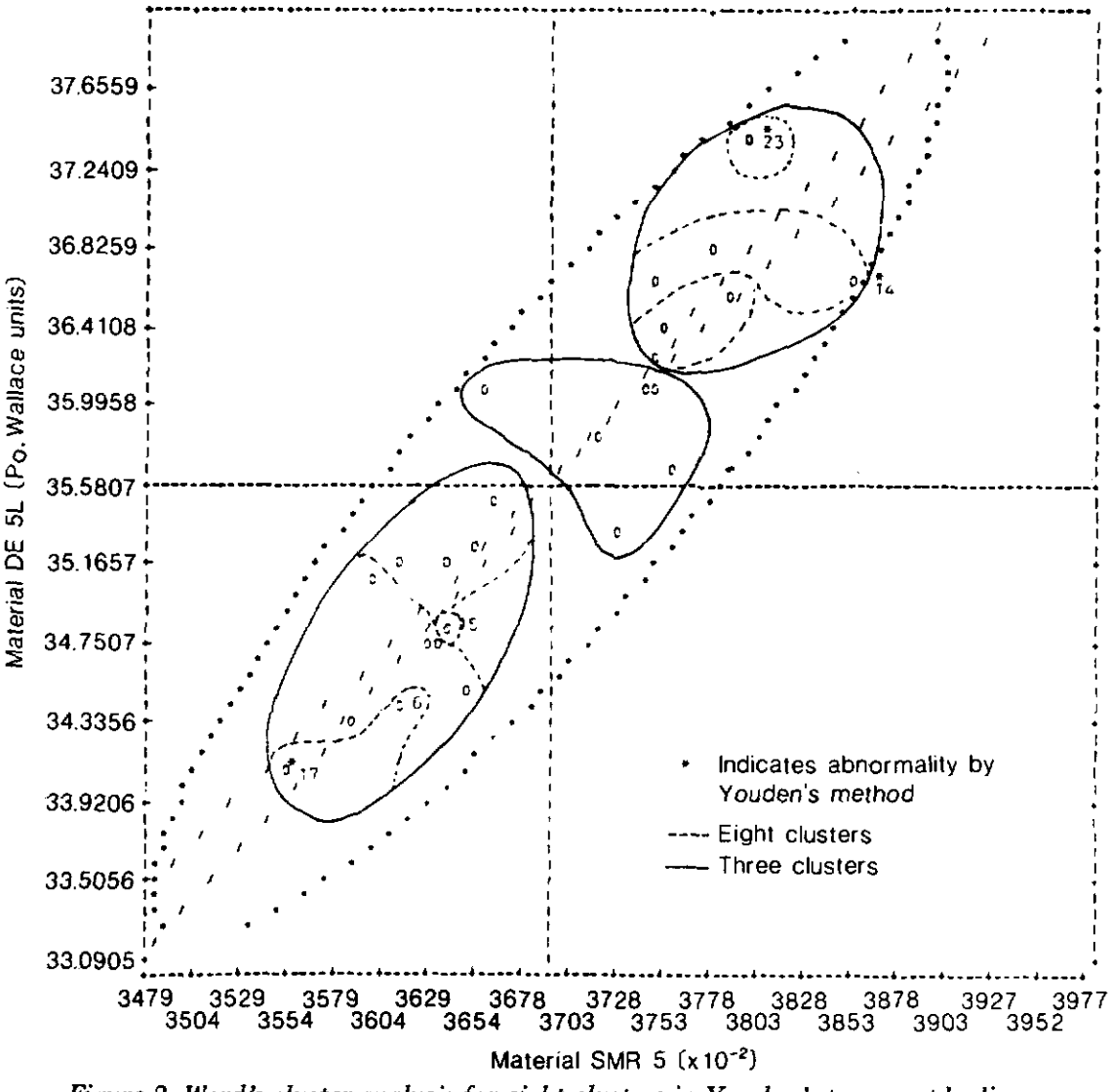

Figure 2. Ward's cluster analysis for eight clusters in Youden's two-sample diagram.

found to produce a structure similar to Ward's method.

When the eight clusters formed by Ward's method is superimposed on Youden's diagram, it is found that the method produces clusters of laboratories behaving similarly and lying close to one another in the Youden's diagram. This means that cluster analysis using Ward's method and the Euclidean measure of similarity tends to cluster together anomalous laboratories or acceptable laboratories. Eight clusters are necessary to produce a meaningful detailed diagnosis of the structure of round robin data instead of three natural clusters as indicated by the mode analysis.

The existence of a cluster of size one out of three or four clusters of a population of about twenty-five laboratories

may be used to signal the occurrence of an outlier. Such an outlier usually lies farthest away from the centre of the ellipse in either the first or third quadrant.

The Euclidean distance which has the property of giving extra weight to outlying values of a single variate has been found to be a useful measure of similarity.

Rubber Research Institute of Malaysia
Kuala Lumpur *November 1982*

REFERENCES

1. LEONG, Y.S. AND LOKE, K.M. (1975) International and Local Round Robin Crosschecks for SMR Testing. *Proc. Int. Rubb. Conf. Kuala Lumpur 1975*, 5, 380.

2. SNEATH, P.H.A. AND SOKAL, R.R. (1970) *Numerical Taxonomy*. San Francisco: W.H. Freeman and Company.

3. ANDERBERG, M.R. (1973) *Cluster Analysis for Applications*. New York: Academic Press.

4. CORMACK, R.M. (1971) A Review of Classification. *Jl. R. Statist. Soc. Ser. A*, 134, 321.

5. WISHART, D. (1978) *CLUSTAN I C. User Manual*. Program Library Unit. Edinburgh University.

6. GOOD, I.J. (1965) Categorization of Classification. *Mathematics and Computer Science in Medicine and Biology*, H.M.S.O.

7. HILLS, M. (1971) Discussion on Dr Cormack's paper. *Jl. R. statist. Soc. Ser. A*, 134, 353.

8. LANCE, G.N. AND WILLIAMS, W.T. (1967) A General Theory of Classificatory Sorting Strategies. I. Hierarchial Systems. *Comput. J.*, 9, 373.

9. McQUITTY, L.L. (1966) Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Date. *Educ. psychol. Measur.*, 26, 825.

10. MACQUEEN, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proc. 5th Berkely Symp.*, 1, 281.

11. BEALE, E.M.L. (1969) *Cluster Analysis*. London Scientific Control Systems.

12. THORNDIKE, R.L. (1953) Who Belongs in the Family? *Psychometrika*, 18, 267.

13. JANCEY, R.C. (1966) Multi Dimensional Group Analysis. *Aust. J. Bot.*, 14, 127.

14. KHIR, M. (1979) Investigating Cluster Analysis as a Technique for Grouping and Selection of *Hevea* Clones. Doctor of Agricultural Science Thesis, State University of Ghent, Belgium.

15. LEONG, Y.S. (1981) Statistical Methods in International and Local Round Robin Interlaboratory Crosschecks for Standard Malaysian Rubber Testing. Doctor of Agricultural Science thesis. University of Ghent, Belgium.

APPENDIX A

Hierarchic Fusion Methods

Hierarchic fusion using eight combinatorial transformations of similarity matrix is performed. Hierarchy starts with N clusters, each being a single individual, which are numbered according to the input of the laboratories *(Table 1)*. In each of the *N-1* fusion cycles, the two clusters which are most similar are fused, the resulting cluster is labelled with the lesser of the two codes of its constituent clusters. It has been suggested that the process can be stopped when a significant drop or discontinuity in the fusion coefficient value is observed[5]. The transformation has been the subject of several publications[8,9,5] and is expressed as follows.

Let clusters P and Q be fused, then the similarity $S\ (R, P + Q)$ between any other cluster R and new fused cluster $(P + Q)$ is obtained from the transformation:

$$S(R, P+Q) = AP^* S(R, P) + AQ^* S(R, Q) + B^* S(P, Q) + G^* S(R, P) - S(R, Q)$$

where AP, AQ, B, G are assigned the following values according to the method

- Single linkage (nearest neighbour, minimum method)
  $AP = AQ = 0.5$, $B = 0$, $G = 0.5$ (similarity) or
  $G = -0.5$ (dissimilarity)
- Complete linkage (furthest neighbour, maximum method)
  AP = AQ = 0.5, B = 0, G = –0.5 (similarity) or
  $G = 0.5$ (dissimilarity)
- Group average linkage (unweighted pair group UPGMA method)
  $AP = NP/(NP + NQ)$, $AQ = NQ/(NP + NQ)$, $B = G = 0$
- Centroid (unweighted pair group centroid UPGMC method)
  $AP = NP/(NP + NQ)$, $AQ = NQ/(NP + NQ)$, $B = -AP^* AQ$, $G = 0$
- Median (unweighted pair group centroid WPGMC method)
  $AP = AQ = 0.5$, $B = -0.25$, $G = 0$
- Ward's method (error sum of squares or minimum variance)
  $AP = (NR + NP)/(NR + NP + NQ)$, $AQ = (NR + NQ/(NR + NP + NQ)$
  $B = -NR/(NR + NP + NQ)$, $G = 0$
- Lance-Williams flexible BETA method
  $AP = AQ = (1 - \text{BETA})/2$, $B = \text{BETA}$, $G = 0$
- McQuitty's similarity analysis (weighted average)
  $AP = AQ = 0.5$, $B = G = 0$
  NR, NP, NQ are cluster sizes and BETA is a variable input parameter.

APPENDIX B

Iterative Relocation Methods

The iterative relocation procedure of classifying N objects or laboratories into k groups can be considered as composed of three steps:

1. Create the initial configuration, that is, an initial partition into k clusters or k initial seed points.
2. During each relocation scan, each laboratory is reconsidered in turn and its similarities with all k clusters computed. Suppose that the similarity between laboratory X and its parent cluster is $S(P,X)$, then if $S(Q, X) > S(P, X)$, the method moves X from cluster P to cluster Q. The procedure is repeated until a local optimum is obtained.
3. Next, similarities between all pairs of clusters are computed and the two clusters which are most similar are fused, thereby reducing the classification to (k-1) clusters and *Step 2* is repeated until the number of terminal clusters specified is reached.

The initial seed points can be chosen randomly[10], regularly spaced[11], mutually farthest apart[12] or supplementary to the data[13].

It is often difficult to find a global optimum solution with a large population. However, the size and shape difference similarity coefficient comprise additive components of Euclidean distance and can be shown to yield orthogonal classifications. Thus, if the same stable error sum of squares solution is obtained with the relocation procedure from a random start, then a global optimum error sum of squares solution[5] is likely to have been achieved by repeatedly applying the relocation procedure using different similarity measures and initial seed points according to the following steps:

1. Use size difference as the similarity measure and random initial seed points.
2. Use shape difference as the similarity measure and random initial seed points.
3. Use error sum of squares as the similarity measure with results of *Step 1* as initial seed points.
4. Use error sum of squares as the similarity measure with results of *Step 2* as the initial seed points.
5. Use error sum of squares as the similarity measure with random seed points.

A convenient number of initial clusters to start the relocation procedure is 10.

APPENDIX C

Mode Analysis

Mode analysis is a method of deriving 'natural' clusters by estimating the disjoint density surfaces according to a probabilistic model.

For a density parameter k (taken to be one in our case), the average, $A(I)$, of the $2k$ smallest distance coefficients for each individual, I, is calculated. This value provides a measure of the density of the space in the immediate vicinity of each individual and small values are associated with points that lie in the region of high density.

Next, the individuals are ordered according to their $A(I)$ values. This ordering determines the sequence in which the individuals are introduced to the cluster nuclei (or become dense). At the start of the hierarchic clustering process, the individual with the least $A(I)$ value is introduced and initiates the first cluster nucleus. During each subsequent cycle, the coefficient threshold, R, is increased to the next smallest $A(I)$ value and the associated individual is said to become dense. Four actions are possible:

- The new point is separated from all other dense points by a distance which exceeds R. When this happens, the point initiates a new cluster nucleus and the number of clusters is increased by one.
- The new point is within distance R of one or more dense points which belong to only one cluster nucleus. In this case, the new point joins that cluster.
- The new point is within distance R of dense points belonging to two or more clusters. If this happens, the clusters concerned are fused.
- At each introduction cycle, the smallest distance D between dense points belonging to different clusters is found.

If at some cycle, the next smallest $A(I)$ threshold value exceeds D, then those two clusters separated by the distance D are combined. The cluster nuclei are defined as the groupings of dense points at coefficient R of any dense point. All other points which are not dense and separated from the cluster nuclei by a distance greater than R are deemed sufficiently remote to be unclassified (such individuals are coded 'o' in the classification array). For the purpose of classifying every individual on a best fit basis, the complete classifications are obtained by grouping each point which is not classified at the nuclei level with the cluster containing its nearest dense point.

The Density procedure in Clustan[5] can be used to estimate all the modes of a multivariate sample density.