

# GENERAL APPLICATIONS OF HIERARCHICAL GROUPING USING THE HIER-GRP COMPUTER PROGRAM

## I. INTRODUCTION

HIER-GRP is a computer program which hierarchically groups a set of regression equations so as to minimize the overall loss of predictive efficiency at each stage of clustering. The HIER-GRP program is described in Gott (1978). The mathematical procedure upon which the HIER-GRP program is based is described in Bottenberg and Christal (1961). The HIER-GRP program has been used extensively for grouping regression equations in which the predictor covariance matrix and predictor means are the same for all equations being considered for grouping. This condition is called the proportionality assumption. When the unit vector is considered as a predictor, an equivalent statement is that the ratios of the corresponding elements of the sums-of-squares and cross-products matrices for any two equations to be clustered are equal to the ratios of the corresponding numbers of the cases within each equation. A problem in which this condition is met is referred to as a proportional regression clustering problem. This was the case in numerous Air Force applications of Judgment Analysis (JAN) (Christal, 1968), such as the development of enlistment promotion systems (Black, 1973; Gott, 1974; Koplyay, 1969). The purpose of this paper is to propose an expanded application of HIER-GRP, without program modification, for grouping analyses in more general situations. The proposed model which produces the input to the clustering is different in general from that used in past HIER-GRP applications, but is equivalent to it in the case of proportional regression clustering problems.

Some sample applications are discussed in section II of this paper, the new method is presented and contrasted with the standard HIER-GRP analysis in section III, and the HIER-GRP input procedure is discussed in section IV. In section V, relationships between the dissimilarity matrix, sum of squared errors,  $D^2$  (Ward & Hook, 1963), and interaction sum of squares are derived. In Section VI, the new method is compared with the  $D^2$  clustering method. Example analyses are then presented for clustering columns of a matrix, both to minimize the sum of squared errors within groups, in section VII, and to minimize the sum of squares due to interaction between columns within a group, in section VIII. Finally, a nonproportional regression clustering problem is presented and solved in section IX, and overall results are summarized in section X.

## II. EXAMPLE APPLICATIONS FOR GROUPING ANALYSES

### Grouping Prediction Systems With Different Predictor-Covariance Matrices

Frequently it is desired to cluster prediction systems in which each equation is derived from a different set of predictor data (in general, data with different predictor covariances and means). This situation usually occurs when observed criterion scores are not available for all sets of

predictor observations as, for example, when final technical school grades are predicted from data gathered on personnel prior to their entry into the Air Force.

### Grouping Jobs (or Technical Schools) to Minimize Loss of Differential Classification Effectiveness

In the practical problem of classifying personnel into different career areas, it is desired to assign persons to jobs (Person Job Match) so that the overall measure of predicted performance (the objective function) is maximized. Jobs can be clustered into groups in order to simplify this assignment problem. The clustering procedure should group jobs together such that persons considered for a classification within a job cluster can be interchanged among all jobs in the cluster with very small effect on the overall measure of classification effectiveness. Jobs should be clustered for which the differences between predicted job success among any pair of jobs within the cluster is nearly constant for all persons. This can be described as little interaction between people and jobs within the cluster (Ward, 1979, 1983).

### Grouping Raters to Minimize Interaction

Occasionally, it is of interest to group raters of tasks such that raters are "similar" according to the extent to which their ratings differ by a constant amount across all tasks. As in the Person Job Match situation, raters can be grouped to minimize the interaction between tasks and raters within the cluster.

### Grouping Any Values

Often it is desired to cluster data that have been observed, judged, or generated without reference to a set of prediction equations. For example, this can occur when different judges each give ratings of the learning difficulty associated with a list of job tasks. It might be of interest to group tasks based on the similarity of ratings provided by the judges. Also, it may be appropriate to group the judges based on the similarity of ratings given across all tasks (or anything else rated by judges). Similarity is frequently measured in terms of sum of squared differences,  $D^2$  (Ward, 1963; Ward & Hook, 1963). The connection between HIER-GRP groupings and  $D^2$  groupings is explained in section VI of this paper.

## III. THE HIER-GRP GROUPING CRITERION AND THE INDUCED MODEL

### The Dissimilarity Matrix A and the Update Equations

The HIER-GRP program groups proportional regression equations so as to minimize the decrease in the overall squared multiple correlation coefficient ( $R^2$ ) at each stage of clustering. To this end, HIER-GRP computes a dissimilarity matrix A, such that  $A_{ij}$  (for  $i \neq j$ ) is the loss in overall  $R^2$  associated with clustering equations (or groups)  $i$  and  $j$

together at a given stage. The aim is to extend the use of HIER-GRP to more general problems, but this must be done by reformulating the more general problem as a proportional regression clustering problem, since HIER-GRP solves only this kind of problem. The goal in the next few paragraphs is to examine what HIER-GRP does with a proportional regression clustering problem. The equations referred to are those of the final proportional regression clustering problem, not those of some nonproportional problem before reformulation.

Assume the  $i^{\text{th}}$  equation has  $n_i$  observations, criterion standard deviation  $s_i$ , and criterion mean  $m_i$ . Assume also that  $M$  is the covariance matrix of the predicted standardized scores (note that  $M_{ii}$  is the squared multiple correlation coefficient of equation  $i$ ) and that  $SS_c$  is the total criterion sum of squares corrected for the grand mean. Then the corrected sum of squares is calculated by

$$SS_c = \sum_i [n_i (s_i^2 + m_i^2)] - (\sum_i [n_i m_i])^2 / (\sum_i n_i)$$

and the dissimilarity matrix  $A$  is computed by

$$A_{ij} = [(s_i^2 M_{ii} + s_j^2 M_{jj} - s_i s_j (M_{ij} + M_{ji}) + (m_i - m_j)^2) / SS_c] [n_i n_j / (n_i + n_j)]$$

for  $i$  not equal to  $j$  (Gott, 1978, pp. 17-18, 49-50).

This equation is used to determine the dissimilarity matrix before any groupings, but is never applied at later stages after groupings have been selected. The resulting dissimilarity matrix is scanned to find its smallest off-diagonal element, and the corresponding groups  $i$  and  $j$  are clustered together. Then the  $A$  matrix and the observation count vector are updated as follows:

$$n_g = n_i + n_j$$

$$A_{gq} = A_{qg} = [(n_i + n_q)A_{iq} + (n_j + n_q)A_{jq} - n_q A_{ij}] / (n_i + n_j + n_q)$$

where the index  $g$  is used for the new group formed by clustering  $i$  and  $j$ , and  $q$  is any group index other than  $i$  or  $j$  (Gott, 1978, p. 53).

Assuming (without loss of generality) that  $i$  is less than  $j$ , the program labels the new cluster with the index  $i$ . Index  $j$  is then deleted from another vector which lists the groups under consideration, and the observation count and dissimilarity elements for group  $g$  are stored in the space formerly used for group  $i$ . Additional calculations are accomplished in order to output information at each clustering stage. Once the dissimilarity matrix and observation counts have been updated to reflect the clustering,  $A$  is scanned again. This process continues until the number of groups is reduced to one.

## The Matrix Product B'V

Bottenberg and Christal (1961, p. 11) describe the use of the matrix product B'V. B is the matrix of standardized regression coefficients, with each column containing the coefficients for one particular equation of the k equations to be grouped. An element  $b_{ij}$ , therefore, contains the standardized least-squares weight for predictor i in equation j ( $i = 1, \dots, p$  and  $j = 1, \dots, k$ ). Matrix V is the p by k matrix of validity coefficients--the correlations of the criteria with the predictors. Under the proportionality assumption,  $M = B'V$  is a useful way to compute the covariance matrix M of the predicted standardized scores (predicted scores based on the standardized beta weights applied to all standardized predictor observations). HIER-GRP uses this formula to compute M before initially computing the dissimilarity matrix. This, together with the above formula for the dissimilarity matrix, explains why HIER-GRP requires the inputs described by Gott (1978).

To see why  $M = B'V$ , let  $X_i$  ( $i = 1, \dots, k$ ) be the  $n_i$  by p matrix of standardized predictor observations for equation i. Let  $Y_i$  be the  $n_i$  by 1 vector of standardized criterion observations for equation i. Then the standardized beta weight vectors (p by 1) are given by

$$B_i = (X_i'X_i)^{-1}X_i'Y_i.$$

Let n be the sum of the  $n_i$  and let X be the n by p predictor matrix which combines the predictor data for all the equations. That is,

$$X' = [X_1', \dots, X_k'].$$

Let the beta matrix B be the p by k matrix whose  $i^{\text{th}}$  column is  $B_i$ . That is,

$$B = [B_1, \dots, B_k].$$

The predicted standardized scores are contained in the n by k matrix Z defined by

$$Z = XB.$$

The covariance matrix M of the predicted standardized scores is

$$\begin{aligned} M &= Z'Z/n \\ &= B'X'XB/n. \end{aligned}$$

But the validity matrix  $V$  (the correlation matrix of the criteria with the predictors) is the  $p$  by  $k$  matrix given by

$$\begin{aligned}V &= X'Y/n \\ &= [X'X (X'X)^{-1}] X'Y/n \\ &= X'XB/n.\end{aligned}$$

Therefore,  $M = B'V$  as claimed earlier.

### The Induced Model

Although  $B$  and  $V$  are normally thought of as matrices of standardized regression coefficients and validities, the definition and use of the  $A$  matrix makes it clear that the only relevant items for grouping with HIER-GRP are the initial  $A$  matrix and the observation counts. That is, if two grouping problems with different beta matrices, different validity matrices, different criterion means, and/or different criterion standard deviations result in the same initial dissimilarity matrix (or if one is a positive scalar multiple of the other) and the same observation counts, then the two corresponding HIER-GRP runs will result in the same groupings. Because of this, it may be useful to avoid using  $B$  and  $V$  with their usual interpretations. For instance, if  $M$  is known, then  $M$  and an identity matrix may be input in place of  $B$  and  $V$  (or vice-versa).

The method proposed here depends on using the following simple model, which we have named the "induced" model. Given a set of  $n$  observations on  $k$  variables expressed as a matrix  $P$  of dimension  $n$  by  $k$  with elements  $P_{ij}$ , consider the  $k$  regression equations derived by using all of the variables as predictors in each equation and using each variable as the criterion in turn. HIER-GRP is used in the usual way to cluster these regression equations, using observation counts that reflect the amount of data one has about each criterion. The result is that an identity matrix is input for  $B$ , and the correlation matrix of the variables is input for  $V$ . This model is a proportional regression clustering problem, whatever the original data; hence HIER-GRP applies as usual to this induced model. This is an interesting example of a case in which the dissimilarity of the regression weights reveals nothing about the similarity of the criteria (columns of  $P$ ), since the beta weight vectors for two criteria are orthogonal, no matter how closely the two criteria may be related. Notice that the regression equations of the induced model, before grouping, all have  $R^2$  values of 1. Therefore, the initial overall system  $R^2$  value computed by HIER-GRP (a weighted average of the individual equation  $R^2$  values) will also be 1.

### Application of the Induced Model to a Regression Clustering Problem

In case the original problem was a regression clustering problem, compute the raw regression weights for each individual criterion (based on the data for that criterion) and compute a predicted criterion score on each criterion for each observation of predictor scores. Use of the above

induced model on these predicted scores (regressing the predicted scores against themselves, and using observation counts from the original problem) yields a proportional regression clustering problem, whether the original problem was proportional or not.

If the original problem was a proportional regression clustering problem, the induced model will lead to the same groupings as the usual HIER-GRP analysis, because the initial observation counts are the same and the initial dissimilarity matrix in one HIER-GRP run is a positive scalar multiple of that in the other run. This can be seen as follows. Assume that each equation  $i$  has criterion standard deviation  $s_i$  and criterion mean  $m_i$ , and  $M$  is the covariance matrix of the predicted standardized scores. The covariance matrix  $Q$  of the predicted scores can be written as

$$Q_{ij} = s_i M_{ij} s_j.$$

Let  $s_i^*$  be the standard deviation of the predicted scores for equation  $i$  and let  $C$  be the correlation matrix of the predicted scores. It follows that the covariance matrix  $Q$  of the predicted scores can be written as

$$Q_{ij} = s_i^* C_{ij} s_j^*.$$

From these two ways of writing the matrix  $Q$ , it is established that

$$s_i^* C_{ij} s_j^* = s_i M_{ij} s_j.$$

In the proportional case, the means of the predicted scores ( $m_i^*$ ) are the same as the means of the observed scores ( $m_i$ ). The total corrected sum of squares for the predicted scores, divided by the total corrected sum of squares for the observations, is the  $R^2$  value for the original system of equations. Using the last equation and these two observations, the formula for the dissimilarity matrix  $A$  shows that the dissimilarity matrix computed for the induced model is simply the dissimilarity matrix computed for the standard analysis, divided by that  $R^2$  value. That is, one of the two dissimilarity matrices is a positive scalar multiple of the other. It follows that the two HIER-GRP analyses will result in exactly the same groupings, as claimed earlier. The decision values (decreases in overall  $R^2$ ) computed in the two HIER-GRP runs will be in this same ratio to one another. The  $R^2$  values at each stage of grouping are also in this same ratio.

#### IV. INPUT PROCEDURE FOR HIER-GRP

For all problems it is assumed that there is given a matrix  $P$  of dimension  $n$  by  $k$  with elements  $P_{ij}$ . These values can be generated in many different ways, depending on the application. For example, the columns could be the predicted criterion values for  $n$  persons in  $k$  schools (or

career fields). Or the columns could be the observed judgments for each of  $k$  judges about the criticality of a set of  $n$  tasks to be performed. Also, the element  $P_{ij}$  could be the expected payoff (or utility) of assigning person  $i$  to job  $j$ .

Assume that it is desired to cluster the  $k$  columns of  $P$ . (Note: If row clustering is desired then interchange rows and columns; i.e., replace  $P$  by its transpose  $P'$ .) Perform the following steps, with input formats as described by Gott (1978, pp. 9-10):

1. Compute the means, the standard deviations and the correlation matrix  $C$  of the columns of  $P$ .

2. Enter the standard deviations computed in step 1 as the criterion and predictor standard deviations.

3. Input the criterion and predictor means according to the following options:

- a. If it is desired to cluster columns so as to minimize the sum of squares of error within groups (SSE), then input the  $k$  means obtained from step 1 as the criterion and predictor means.

- b. If it is desired to cluster columns so as to minimize the sum of squares due to interaction (S) between columns within each cluster, then input 0's (or any constant value) for the criterion means and the predictor means.

4. Enter  $k$  as the number of criteria and the number of predictors.

5. Enter a  $k$  by  $k$  identity matrix for the beta weight matrix  $B$ .

6. Enter  $C$  for the validity matrix  $V$ .

7. Enter the observation counts from the original problem.

Note that all inputs and outputs are related to the trivial regression clustering problem mentioned in Section III. As a result, the degrees of freedom and significance figures calculated by HIER-GRP may be inappropriate for the original problem. In fact, as shown in the examples in Appendices A and B, the degrees of freedom calculated by HIER-GRP may be negative. This in no way affects the validity of the groupings.

#### V. RELATIONSHIPS BETWEEN THE DISSIMILARITY MATRIX, SUM OF SQUARED ERRORS, $D^2$ , AND INTERACTION SUM OF SQUARES

Option 3a above will cluster the columns of  $P$  to minimize the sum of squared errors within groups (SSE) as described by Ward and Hook (1963). Option 3b will cluster the columns of  $P$  to minimize the interaction sum of squares (S) between the columns within groups. These options can be seen to

achieve the desired results by observing relationships between the elements of the dissimilarity matrix ( $A_{ij}$ ), the sum of squared errors ( $SSE_{ij}$ ), the squared distance ( $D_{ij}^2$ ), and the interaction sum of squares ( $S_{ij}$ ).

For the proportional case, as noted in section III,

$$s_i M_{ij} s_j = s_i^* C_{ij} s_j^*$$

and

$$m_i = m_i^*$$

Thus,  $A_{ij}$  can be written as

$$A_{ij} = [(s_i^{*2} + s_j^{*2} - 2s_i^* s_j^* C_{ij} + (m_i^* - m_j^*)^2) / SS_c] [n_i n_j / (n_i + n_j)].$$

Now let

$$SSE_{ij} = \sum_q [(P_{qi} - (P_{qi} + P_{qj})/2)^2 + (P_{qj} - (P_{qi} + P_{qj})/2)^2],$$

$$D_{ij}^2 = \sum_q [(P_{qi} - P_{qj})^2] \quad \text{and}$$

$$\begin{aligned} S_{ij} = & \sum_q [P_{qi}^2 + P_{qj}^2 - (P_{qi} + P_{qj})^2 / 2] \\ & - ((\sum_q [P_{qi}])^2 + (\sum_q [P_{qj}])^2) / n \\ & + (\sum_q [P_{qi} + P_{qj}])^2 / (2n) \end{aligned}$$

where  $n$  is the number of rows in  $P$  (Ward, 1983).

Buchhorn (1980) has shown that

$$S_{ij} = (n/2) (s_i^{*2} + s_j^{*2} - 2s_i^* s_j^* C_{ij})$$

and it can be shown that

$$D_{ij}^2 = n(s_i^{*2} + s_j^{*2} - 2s_i^* s_j^* C_{ij} + (m_i^* - m_j^*)^2)$$

and

$$D_{ij}^2 = 2(\text{SSE}_{ij}).$$

Then by substitution it can be seen that

$$A_{ij} = [(2/n) \text{SSE}_{ij} / \text{SS}_c] [n_i n_j / (n_i + n_j)].$$

Observing that

$$\text{SSE}_{ij} = S_{ij} + (n/2)(m_i - m_j)^2,$$

it follows that

$$A_{ij} = [(2/n) (S_{ij} + (n/2)(m_i - m_j)^2) / \text{SS}_c] [n_i n_j / (n_i + n_j)].$$

Therefore  $\text{SSE}_{ij}$  is composed of two parts, the interaction sum of squares and the sum of squares due to differences between means. When  $m_i = m_j$  (as in option 3b), then the error sum of squares is due only to interaction.

## VI. $D^2$ VERSUS $R^2$ CLUSTERING

Assume that we wish to apply the induced model to an  $n$  by  $k$  observation matrix  $P$ , with columns  $P_j$  ( $j = 1, \dots, k$ ) and observation counts  $n_j$ . The criterion observation vector  $Y_j$  for equation  $j$  is given by

$$Y_j = [P_j', \dots, P_j']'$$

where  $P_j'$  is repeated  $n_j$  times. That is,  $Y_j$  is a vector ( $n_j n$  by 1) containing  $n_j$  copies of column  $P_j$ . The predictor matrix for equation  $j$  is similarly given by

$$X_j = [P', \dots, P']'$$

where the matrix  $P'$  is repeated  $n_j$  times. That is,  $X_j$  is an  $n_j n$  by  $k$  matrix consisting of  $n_j$  copies of the  $P$  matrix.

The observation count for equation  $j$  of this model actually is  $n_j n$  and the observation count for the system is  $n^2$ , while the earlier discussion of the induced model used observation counts  $n_j$  for equation  $j$  and  $n$  for the system. Yet the dissimilarity matrix is exactly the same with the larger observation counts, since

$$(n_j n)(n_q n) / (n_j n + n_q n) = n n_j n_q / (n_j + n_q),$$

and the factor  $n$  is exactly compensated for by the fact that the corrected sum of squares with the larger observation counts is also  $n$  times that with the smaller counts. The predictor matrices  $X_j$  clearly satisfy the proportionality condition since  $X_j'X_j/n_j$  does not depend on  $j$ .

The between-groups sum of squares matrix BSS as defined by Myer (1969) is computed by

$$BSS_{jq} = D_{jq}^2 n_j n_q / (n_j + n_q).$$

The  $BSS_{jq}$  entry is the increase in sum of squared deviations that would result from the clustering of groups  $j$  and  $q$  at this stage. The BSS matrix is scanned for its smallest off-diagonal term to determine the grouping which results in the smallest possible increase in the sum of squared deviations. Once a grouping is chosen, the BSS matrix is updated exactly the way HIER-GRP updates the dissimilarity matrix  $A$ .

When the induced model is applied to the groups of  $P$ , the variable represented by column  $P_j$  is regressed against all  $k$  variables corresponding to the columns of  $P$ . Since the criteria are perfectly predicted in this case (before any grouping), the regression sum of squares is the total corrected sum of squares. If two groups  $j$  and  $q$  are clustered, the total corrected sum of squares does not change. but the regression sum of squares changes to reflect the fact that  $n_j$  copies of column  $P_j$  and  $n_q$  copies of column  $P_q$  are replaced with  $n_g$  copies of column  $P_g$ , where  $g$  is the index for the new group and:

$$n_g = n_j + n_q$$

and 
$$n_g P_g = n_j P_j + n_q P_q.$$

The regression sum of squares is reduced by an amount equal to the sum of squares due to columns  $j$  and  $q$  minus that for column  $g$ . This difference is given by:

$$\begin{aligned} & n_j P_j' P_j + n_q P_q' P_q - n_g P_g' P_g \\ &= [n_j - n_j^2 / (n_j + n_q)] P_j' P_j + [n_q - n_q^2 / (n_j + n_q)] P_q' P_q \\ &\quad - [2n_j n_q / (n_j + n_q)] P_j' P_q \\ &= [n_j n_q / (n_j + n_q)] [P_j' P_j + P_q' P_q - 2P_j' P_q] \\ &= [n_j n_q / (n_j + n_q)] (P_j - P_q)' (P_j - P_q) \\ &= D_{jq}^2 n_j n_q / (n_j + n_q) \\ &= BSS_{jq}. \end{aligned}$$

The above derivation in fact motivates the name of the BSS matrix. The decrease in  $R^2$  is precisely equal to the decrease in regression sum of squares divided by the total corrected sum of squares. Therefore the dissimilarity matrix entry for clustering groups  $j$  and  $q$  would be equal to

$$A_{jq} = \text{BSS}_{jq} / \text{SS}_c^*$$

where (summing  $i=1, \dots, k$ )  $\text{SS}_c^* = \sum_i (n_i P_i' P_i) - n m^2 = n \text{SS}_c$ .

These equations show that the BSS matrix used in the  $D^2$  grouping of the columns of  $P$  is a positive scalar multiple of the dissimilarity matrix  $A$  used by HIER-GRP to group the regression equations of the induced model. Grouping by the  $D^2$  criterion involves using the BSS matrix in exactly the same way that HIER-GRP uses the dissimilarity matrix  $A$ , and the update equations used are the same as those used to update  $A$  in HIER-GRP. The result is that regression clustering of the induced model using HIER-GRP is equivalent to  $D^2$  grouping on the columns of  $P$ .

#### VII. GROUPING COLUMNS BASED ON THE NO-DIFFERENCES HYPOTHESIS

HIER-GRP's decision value at each clustering stage reflects the error sum of squares introduced by clustering the columns. If the criterion means are not set to the same value, then the HIER-GRP clustering attempts to put those columns together that have the most similar corresponding elements. Consider the following example of matrix  $P$ .

$$P = \begin{bmatrix} 1 & 11 & 14 & 0 \\ 2 & 12 & 13 & 0 \\ 3 & 13 & 12 & 0 \\ 4 & 14 & 11 & 0 \\ \text{mean} & 2.5 & 12.5 & 12.5 & 0 \\ \text{variance} & 1.25 & 1.25 & 1.25 & 0 \end{bmatrix}$$

Observe the differences between the corresponding elements of the adjacent columns. (Differences should be considered for all possible pairs of columns, but for the sake of simplicity only the adjacent columns are compared here.)

|       | Large<br>Differences |       | Small<br>Differences |      | Large<br>Differences |       |   |
|-------|----------------------|-------|----------------------|------|----------------------|-------|---|
| $P =$ | 1                    | (-10) | 11                   | (-3) | 14                   | (+14) | 0 |
|       | 2                    | (-10) | 12                   | (-1) | 13                   | (+13) | 0 |
|       | 3                    | (-10) | 13                   | (+1) | 12                   | (+12) | 0 |
|       | 4                    | (-10) | 14                   | (+3) | 11                   | (+11) | 0 |



Table 2. Summary of Constant-Differences Grouping Example

| Number of Groups | Overall $R^2$ | Decision Value | Column Number                          |
|------------------|---------------|----------------|--|
| 4                | 1.0000        | -              | 1      2                      3      4 |
| 3                | 1.0000        | .0000          |  |
| 2                | .8333         | .1667          | └───┬───┬───┬───┘                      |
| 1                | .0833         | .7500          | └──────────────────┘                   |

Columns 1 and 2 were grouped first, with a decision value of 0, reflecting the fact that there is no interaction between rows and columns within those two columns. This means that the differences are constant (-10) for each of the four pairs of elements. The next stage grouped columns 3 and 4, with a decision value of 0.1667. Finally, all columns were combined, with a loss of 0.7500. The regression sum of squares for interaction, and the increase in the error sum of squares for interaction, can be obtained by multiplying the  $R^2$  and decision values by the total interaction sum of squares (15.0). These results have been inserted on the computer printout in Appendix B. Note that degrees of freedom and significance figures should be disregarded since they are inappropriate for this problem. Changes in the program to suppress such printing are being considered.

In this case, the HIER-GRP decision values reflect the error sum of squares associated with the no-interaction hypothesis. The question is "How constant are the differences between corresponding row elements?" This approach is appropriate for clustering jobs (or people) in the Person Job Match situation. If there is no interaction among a group of jobs, then the jobs can be combined, because all one-to-one assignment choices within the cluster are equally good. This clustering technique might be used to reduce the number of source rows (or demand columns) in any standard transportation problem. Once the reduced problem is solved, so that demands have been assigned to each cluster of sources, the original problem is solved by making specific assignments of supplies within each cluster to demands already assigned to that cluster. If the interaction between source rows within each cluster is zero, an arbitrary assignment within each cluster gives an optimal solution to the original problem. If the interaction is nonzero, no assignment within clusters can be guaranteed to be optimal for the original problem, but choosing optimal assignments within each cluster will give a near-optimal solution if the interactions are small. More details are discussed by Ward (1979, 1983).

#### IX. A REGRESSION CLUSTERING PROBLEM

The following problem is a nonproportional modification of the proportional regression clustering problem cited by Bottenberg and Christal (1961). The problem consists of four regression equations with two

predictors, derived from four data sets having four, five, six, and seven observations, respectively. Table 3 displays the data involved, with  $X_{1j}$  and  $Z_{1j}$  being predictor values for observation 1 in data set  $j$  while  $Y_{1j}$  is the criterion score for observation 1 in data set  $j$ . Table 4 displays the raw regression weights for the four equations, for the model

$$Y_{1j} = b_{0j} + b_{1j}X_{1j} + b_{2j}Z_{1j} + e_{1j}.$$

Table 5 displays the predicted scores given by applying the weights in Table 4 to the  $X$  and  $Z$  values in Table 3. That is, the predicted score on criterion  $j$  for observation  $u$  from data set  $v$  is determined by

$$P_{tj} = b_{0j} + b_{1j}X_{uv} + b_{2j}Z_{uv}$$

where  $j=1, \dots, 4$ ;  $v=1, \dots, 4$ ;  $u=1, \dots, n_v$ ; and  $t = u + (n_1 + \dots + n_{v-1})$ . Note that predicted scores are used even for the cases in which observations are available. This and some other details of the process are dictated by an arbitrary decision that the induced model groupings should agree with the standard HIER-GRP groupings in the proportional case.

Table 5 contains the means and standard deviations of the columns of  $P$ , while the validity matrix in section II of Appendix C contains the correlation matrix for the columns of  $P$ . These are input to HIER-GRP along with the observation counts to get the output shown in Appendix C. The inputs are shown in sections I and II of the output. As shown in that output, criteria 1 and 2 are clustered at the third (three-cluster) stage, for a drop in overall system  $R^2$  value from 1.0 to 0.9579. At the second (two-cluster) stage, criterion 4 is added to the (1,2) cluster for an overall system  $R^2$  value of 0.7338. Finally, criterion 3 is added to the (1,2,4) cluster for an overall  $R^2$  value of 0.1260. The regression sum of squares and the increase in the error sum of squares can be obtained by multiplying the  $R^2$  and decision values by the total corrected sum of squares (401.93). These results have been inserted on the computer printout in Appendix C.

Table 3. A Regression Clustering Problem

| $j,v$ | X   | Z    | Y    | $i,u$ | t  |
|-------|-----|------|------|-------|----|
| 1     | 2.0 | 4.0  | 1.0  | 1     | 1  |
| 1     | 2.0 | 8.0  | 5.0  | 2     | 2  |
| 1     | 4.0 | 4.0  | 9.0  | 3     | 3  |
| 1     | 8.0 | 8.0  | 5.0  | 4     | 4  |
| 2     | 0.0 | 8.0  | 3.0  | 1     | 5  |
| 2     | 5.0 | 4.0  | 5.0  | 2     | 6  |
| 2     | 5.0 | 6.0  | 7.0  | 3     | 7  |
| 2     | 5.0 | 10.0 | 9.0  | 4     | 8  |
| 2     | 0.0 | 2.0  | 1.0  | 5     | 9  |
| 3     | 0.0 | 6.0  | 3.0  | 1     | 10 |
| 3     | 2.0 | 2.0  | 7.0  | 2     | 11 |
| 3     | 2.0 | 8.0  | 7.0  | 3     | 12 |
| 3     | 3.0 | 6.0  | 19.0 | 4     | 13 |
| 3     | 3.0 | 6.0  | 15.0 | 5     | 14 |
| 3     | 8.0 | 8.0  | 15.0 | 6     | 15 |
| 4     | 6.0 | 9.0  | 1.0  | 1     | 16 |
| 4     | 1.0 | 1.0  | 2.0  | 2     | 17 |
| 4     | 1.0 | 1.0  | 3.0  | 3     | 18 |
| 4     | 3.0 | 7.0  | 5.0  | 4     | 19 |
| 4     | 3.0 | 7.0  | 8.0  | 5     | 20 |
| 4     | 7.0 | 12.0 | 8.0  | 6     | 21 |
| 4     | 0.0 | 5.0  | 8.0  | 7     | 22 |

Table 4. Regression Weights

| $j:$     | 1       | 2      | 3       | 4       |
|----------|---------|--------|---------|---------|
| $b_{0j}$ | 4.6000  | -.4545 | 7.0000  | 2.3857  |
| $b_{1j}$ | 0.4000  | 0.8364 | 1.4667  | -1.6714 |
| $b_{2j}$ | -0.2000 | 0.4909 | -0.0667 | 1.2714  |

Table 5. Predicted Scores

| P <sub>tj</sub>         | 1     | 2      | 3      | 4      |
|-------------------------|-------|--------|--------|--------|
| 1                       | 4.600 | 3.182  | 9.667  | 4.129  |
| 2                       | 3.800 | 5.145  | 9.400  | 9.214  |
| 3                       | 5.400 | 4.855  | 12.600 | 0.786  |
| 4                       | 6.200 | 10.164 | 18.200 | -0.814 |
| 5                       | 3.000 | 3.473  | 6.467  | 12.557 |
| 6                       | 5.800 | 5.691  | 14.067 | -0.886 |
| 7                       | 5.400 | 6.673  | 13.933 | 1.657  |
| 8                       | 4.600 | 8.636  | 13.667 | 6.743  |
| 9                       | 4.200 | 0.527  | 6.867  | 4.929  |
| 10                      | 3.400 | 2.491  | 6.600  | 10.014 |
| 11                      | 5.000 | 2.200  | 9.800  | 1.856  |
| 12                      | 3.800 | 5.145  | 9.400  | 9.214  |
| 13                      | 4.600 | 5.000  | 11.000 | 5.000  |
| 14                      | 4.600 | 5.000  | 11.000 | 5.000  |
| 15                      | 6.200 | 10.164 | 18.200 | -0.814 |
| 16                      | 5.200 | 8.982  | 15.200 | 3.800  |
| 17                      | 4.800 | 0.873  | 8.400  | 1.986  |
| 18                      | 4.800 | 0.873  | 8.400  | 1.986  |
| 19                      | 4.400 | 5.491  | 10.933 | 6.271  |
| 20                      | 4.400 | 5.491  | 10.933 | 6.271  |
| 21                      | 5.000 | 11.291 | 16.467 | 5.943  |
| 22                      | 3.600 | 2.000  | 6.667  | 8.743  |
| Means:                  | 4.673 | 5.152  | 11.267 | 4.696  |
| Standard<br>Deviations: | 0.826 | 3.076  | 3.535  | 3.694  |
| Observation<br>Counts:  | 4     | 5      | 6      | 7      |

#### X. SUMMARY

Procedures were described for applying a hierarchical grouping program to a wider range of problems than the program was originally designed to handle. Any clustering problem is converted, through an induced model, into a proportional regression clustering problem to which HIER-GRP applies without change. In the most general application, a matrix of any values obtained through estimation (predicted scores) or any other means can be evaluated for systematic differences. It was noted that, for proportional regression clustering problems, this method gives the same groupings as a standard HIER-GRP analysis. It was also shown that the method is closely related to grouping according to the  $D^2$  statistic. Empirical investigation is needed to evaluate the usefulness of this new clustering technique as a decision aid.

