Direct Clustering of a Data Matrix

J. A. Hartigan

Journal of the American Statistical Association, Vol. 67, No. 337. (Mar., 1972), pp. 123-129.

© Journal of the American Statistical Association
March 1972, Volume 67, Number 337
Theory and Methods Section

Direct Clustering of a Data Matrix

J. A. HARTIGAN*

Clustering algorithms are now in widespread use for sorting heterogeneous data into homogeneous blocks. If the data consist of a number of variables taking values over a number of cases, these algorithms may be used either to construct clusters of variables (using, say, correlation as a measure of distance between variables) or clusters of cases. This article presents a model, and a technique, for clustering cases and variables simultaneously. The principal advantage in this approach is the direct interpretation of the clusters on the data.

1. INTRODUCTION

Consider the voting data, Table 1, consisting of the percentage Republican vote for President of the United States, in the southern states, over the years 1900–1968. It is desired to detect clusters of states, i.e., states which vote similarly, and also clusters of years, i.e., years for which the votes are similar.

The definition of similarity is crucial, and it has been customary to express similarity through a measure of distance between pairs of objects to be clustered. See, for example, [1,2,4,5,6,7,8,9,11,12,14,15,16,17,19,22].

It is customary to assume this distance matrix is given, and to develop algorithms to construct clusters using the distance matrix. For example, a cluster may be any subset of objects such that for any two objects x and y inside the cluster, and any z outside the cluster, x and y are closer to each other than to z. The family of clusters satisfying the previous property forms a tree, i.e., any two clusters in the family are disjoint, or one includes the other. Another common cluster family is a partition, a set of disjoint subsets whose union is the whole set. Partitions are special cases of trees, and trees will be the principal type of clustering considered here.

For case by variable data, there are several serious difficulties with the distance method which make alternative approaches desirable. Some of these are discussed in [9]. A list of the difficulties, in increasing order of importance follows.

1. *Expensive computations.* The computing and storing of $n(n-1)/2$ distances makes it expensive to cluster even moderate numbers of objects, say 500. Some algorithms avoid the $n(n-1)/2$ calculations by computing only a subset of the distances during the course of the algorithm. This seems plausible since principally the small distances are involved in constructing clusters. So this computational difficulty is not really an overriding one.

2. *Weighting decisions.* In choosing the distance function, decisions must be made about the relative weight to be given to each variable. This is not only a problem of the distance approach. In clustering insects, various variables describing mouth parts might be used, and a certain set of clusters obtained. Or variables describing genitalia might be used, and a different set of clusters obtained. A family of different clusterings will be obtained according to the relative weight given genitalia and mouthparts. See, for example, [3, p. 150].

A taxonomist chooses carefully the variables to be used in clustering, rejects many as irrelevant or too variable, and gives more or less importance to others. These decisions are often subjective ones, disagreed on among the experts, subject to later revision. The weighting decisions for variables are made interactively with the establishing of clusters. Thus a variable which does not distinguish well between established clusters will be reduced in weight.

In the distance approach, these decisions must be made in advance of any knowledge of clustering, in advance of any hints about which variables are good variables and which are bad for clustering. Real variables have usually been weighted using the sample covariance matrix, perhaps requiring all variables to have variance one, perhaps doing principal component analysis and ignoring all but the first few principal components corresponding to the largest eigenvalues, or perhaps even computing the Mahalanobis distance which gives all principal compo-

1. REPUBLICAN VOTE FOR PRESIDENT[a]

State	Year																	
	00	04	08	12	16	20	24	28	32	36	40	44	48	52	56	60	64	68
Alabama (AA)	35	21	24	8	22	31	27	48	14	13	14	18	19	35	39	42	70	14
Arkansas (AS)	35	40	37	20	28	39	29	39	13	18	21	30	21	44	46	43	44	31
Delaware (DE)	54	54	52	33	50	56	58	65	51	43	45	45	50	52	55	49	39	45
Florida (FA)	19	21	22	8	18	31	28	57	25	24	26	30	34	55	57	52	48	41
Georgia (GA)	29	18	31	4	7	29	18	43	8	13	15	18	18	30	33	37	54	30
Kentucky (KY)	49	47	48	25	47	49	49	59	40	40	42	45	41	50	54	54	36	44
Louisiana (LA)	21	10	12	5	7	31	20	24	7	11	14	19	17	47	53	29	57	23
Maryland (MD)	52	49	49	24	45	55	45	57	36	37	41	48	49	55	60	46	35	42
Mississippi (MI)	10	5	7	2	5	14	8	18	4	3	4	6	3	40	24	25	87	14
Missouri (MO)	46	50	49	30	47	55	50	56	35	38	48	48	42	51	50	50	36	45
North Car. (NC)	45	40	46	12	42	43	55	29	29	27	26	33	33	46	49	48	44	40
South Car. (SC)	7	5	6	1	2	4	2	9	2	1	4	4	4	49	25	49	59	39
Tennessee (TE)	45	43	46	24	43	51	44	54	32	31	33	39	37	50	49	53	44	38
Texas (TS)	31	22	22	9	17	24	20	52	11	12	19	17	25	53	55	49	37	40
Virginia (VA)	44	37	38	17	32	38	33	54	30	29	32	37	41	56	55	52	46	43
West Virginia (WV)	54	55	53	21	49	55	49	58	44	39	43	45	42	48	47	54	32	40

[a] Southern states by 20th century years, in percentages.

* J. A. Hartigan is associate professor, Department of Statistics, Yale University, New Haven, Conn. 06520.

2. MARGINAL TREES USING DISTANCES

State	Year						Distance between years	Distance between states
	32	36	40	60	64	68		
SC	2	1	4	49	59	39	32	SC
MI	4	3	4	25	87	14	3 | 40	12 | IA
LA	7	11	14	29	57	23	6 | 5 | 36	18 | 14 | MI
KY	40	42	40	54	36	44	25 23 24 | 60	23 26 36 | KY
MD	36	41	37	46	35	42	18 18 18 | 8 | 68	28 21 35 | 4 | MD
MO	35	48	38	50	36	45	50 45 47 30 34 | 64	28 25 36 | 4 | 4 | MO

nents (including the many junk ones with small eigenvalues) equal weight. But the sample covariance matrix does not reveal clustering of cases, and it is on this clustering that variables must be evaluated. Admittedly there is a circularity here in that variables are used to construct the clusters which are used to evaluate the variables.

3. *Remoteness from data.* If distances are used, the results of the clustering must be interpreted as—such and such objects are close in this distance. This information is not useless if the distance is well chosen. An example can be seen in Table 2 selected from the voting data. Euclidean distances are used between years, giving all states equal weight, and a tree of clusters is obtained for years. Similarly a tree is obtained from states. The year clusters are interpreted by saying 32, 40, 36 are close, 60, 68 are close and 64 is far from all years. And SC, LA, MI are close, and KY, MD, MO even closer. But this information does not reveal the main happening in the data, i.e., the unusual interaction between SC, LA, MI and 60, 68, 64.

This article introduces a clustering technique in which the model for a single cluster relates a cluster of variables to a cluster of cases. Variables and cases are thus clustered simultaneously, and the results of the clustering are interpreted directly on the data matrix.

Tryon [20] and Tryon and Bailey [21] cluster both variables and cases, first clustering variables using the correlation matrix and then using a distance measure across the clusters of variables to cluster cases. Their technique differs from the present one in not relating specific case clusters to specific variable clusters. In a survey paper, Good [10] sketches a technique for simultaneous clustering of cases and variables.

2. DIRECT CLUSTERING MODELS

The data are assumed to be in data matrix form with rows of the matrix designated as cases and columns of the matrix designated as variables. There is a response for each variable, for each case. (Some responses may be missing.) There may be various degrees of comparability among responses in the matrix. Usually, responses to various cases within the same variable are comparable. More specially, if the variables are all on the same scale (perhaps after some previous standardization), responses may be comparable between variables as well.

A cluster is a submatrix of the data matrix. The corresponding set of cases will be called a marginal case (or row) cluster; the corresponding set of variables is a marginal variable (or column) cluster. There are thus three clusters present: the cluster of response values, the marginal cluster of cases, and the marginal cluster of variables. A model for a cluster specifies the form of the data within a cluster. For example, if responses are comparable across variables, as in the voting data, the model specifies that all responses within the cluster are equal. If responses are not comparable across variables, the model specifies that each variable in the marginal cluster is constant over the cases in the marginal case cluster. Other models will be appropriate for different forms of data, e.g., a two-way analysis of variance model, or the requirement that the submatrix be of low rank. See Table 3.

If a number of clusters are present, it is desirable for presentation purposes to restrict the way they relate to each other. It makes an ugly picture if all clusters cannot be presented as contiguous blocks after permuting rows and columns. The three-tree cluster structure requires that all three families of clusters, on the responses, the cases, and the variables, are trees. (A family of clusters is a tree if no two clusters "overlap"; either they are disjoint, or one includes the other.) This structure implies contiguity of clusters after permutation of rows and columns. (The converse is false.) A slight adjustment must be made in the implications of a model in the presence of

3. DIRECT CLUSTERING MODELS[a]

		Case	Variable				
			1	2	3	4	5
A.	Responses comparable over variables	1	4	4	4	4	7
		2	4	4	4	4	2
		3	4	4	4	4	1
		4	6	9	3	2	1
		5	5	4	1	4	6
		6	7	8	1	6	2
B.	Responses not comparable between variables	1	4	7	A	K	1
		2	4	7	A	K	2
		3	4	7	A	K	3
		4	7	1	B	L	4
		5	8	2	C	L	7
		6	9	6	E	F	1
C.	ANOVA model	1	4	5	2	8	1
		2	7	8	5	11	6
		3	11	12	9	15	1
		4	2	4	9	8	1
		5	1	7	2	6	7
		6	3	1	4	3	4

[a] Case cluster (1, 2, 3); variable cluster (1, 2, 3, 4).

a number of clusters. The model implies that the response values in a cluster C be of a certain form, *regarding the values in other clusters included in C as missing.*

Two interesting specializations of the three tree cluster structure are:

1. The response clusters form a partition. This case is considered in detail for the equal-response-values model on the voting data.
2. All three sets of clusters are partitions. The three-tree structure, and the above specializations, are outlined in Figure A.

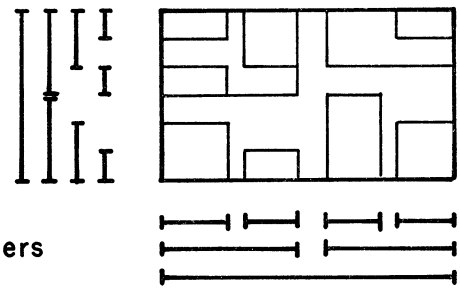
An example of the three-tree cluster structure is given in Table 4, with the model that variables in a cluster be constant over cases in the cluster. The data are UN votes in 1969–1970. It is natural to expect clusters of countries (similar interests or political systems) and clusters of propositions (series of propositions about the same underlying issues). The eye-and-hand algorithm attempts to maximize the number of equal values implied by the model and, given this maximum, to minimize the number of blocks. The clustering expresses the voting in summary form—a few propositions were agreed on by all countries, but otherwise the Assembly splits into an Eastern and Western bloc who disagree on most issues. On the China question, Senegal and Dahomey defect from the Eastern bloc and France, Sweden and Norway from the Western bloc. The Soviet bloc defects on the South African expulsion issue, as do Mexico and Venezuela, etc. The large blocs are more stable than the small ones.

A. 3T CLUSTER STRUCTURES

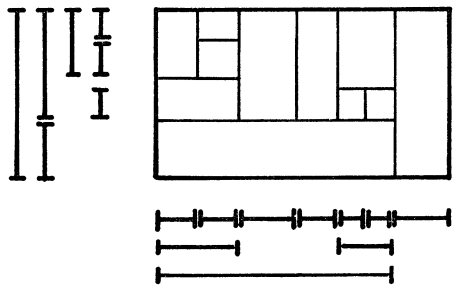
1. Three tree structure

Case clusters

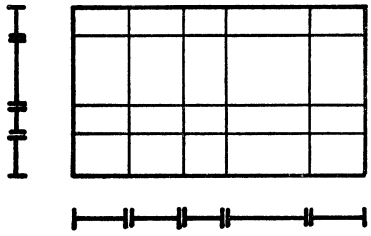
Variable clusters



2. Partitioned responses



3. Three partitions



4. UN VOTES IN 1969–1970[a]

State	EASE				HUNG				CHINA				KOREA				SO AF				PAPUA	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14								
USR	1	1	1	1	3	1	2	3	1	3	2	2	1	3								
BGA	1	1	1	1	3	1	1	3	1	3	2	2	1	3								
YUG	1	3	3	3	3	1	1	3	1	2	3	1	1	2								
SYR	1	2	2	2	3	1	1	3	1	2	3	1	1	3								
UAR	1	3	3	3	3	1	1	3	2	2	3	1	1	3								
KEN	1	3	3	3	3	1	1	3	2	5	3	1	1	3								
TAN	1	2	2	2	3	1	1	3	2	5	3	1	1	3								
SEN	1	3	3	3	1	2	2	2	2	1	3	1	1	2								
DAH	1	3	3	3	1	3	1	3	5	1	3	1	2	2								
USA	1	3	3	3	1	3	3	1	3	1	1	3	3	1								
UNK	1	3	3	3	1	1	3	2	3	1	1	3	3	1								
FRA	1	3	3	3	3	1	2	3	3	1	1	3	2	2								
SWE	1	3	3	3	3	1	2	3	3	1	1	3	3	1								
NOR	1	3	3	3	3	1	3	2	3	1	1	3	3	1								
AIA	1	3	3	3	1	3	1	3	3	1	1	3	3	1								
NZ	1	3	3	3	1	3	1	1	3	1	1	3	3	1								
MEX	1	2	2	2	1	3	3	1	3	1	1	1	2	1								
VEN	1	2	2	2	1	3	3	1	3	1	2	1	1	1								
BRA	1	2	2	2	1	3	3	1	3	1	1	3	1	1								

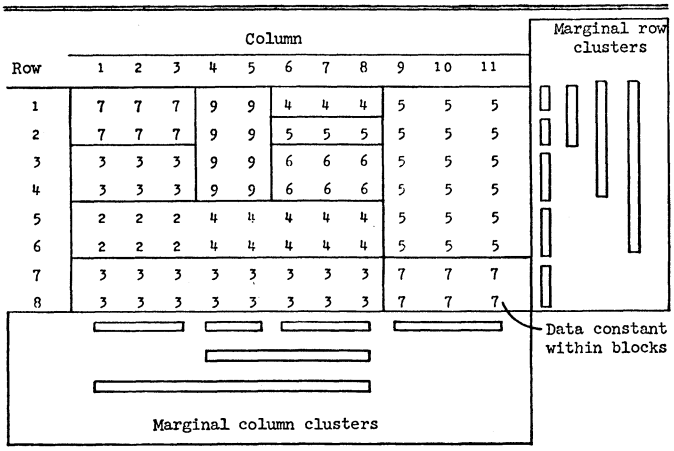
[a] 1 = Yes, 2 = Abstain, 3 = No, 5 = Absent.

NOTE:
1. Call for eased tensions in Korea.
2. Add Hungarian preamble to South Africa expulsion from UNCTAD.
3. Replace last paragraph of preamble by Hungarian amendment.
4. Hungarian amendment of paragraph 1 and 2 of SA expulsion.
5. Declare the China admission question important.
6. Recognize mainland China and expel Formosa.
7. To make study commission on China admission important.
8. To form study commission on China admission.
9. To adopt USSR proposal to delete item on Korea unification.
10. Reaffirm the UN mission in Korea.
11. Declare SA expulsion from UNCTAD important.
12. Adopt SA expulsion.
13. Right of Papua and New Guinea to independence in principle.
14. Call for powers to turn government over to Papua and New Guinea.

3. A DIRECT CLUSTERING ALGORITHM

In the Republican voting data, the entries in the data matrix are comparable across both states and years. An appropriate model for a single cluster is that all responses within the cluster are equal. The cluster structure will assume that the clusters partition the responses, but that the marginal row clusters form a tree, as do the marginal

5. PARTITIONED RESPONSE MODEL



column clusters. The names "rows" and "columns" will be used rather than cases and variables, since the rows and columns are treated the same in the equal-response model. See Table 5 for a data matrix satisfying the above model and cluster structure.

The deviation of an observed data matrix A from the ideal model on a particular partition $B_1, \cdots, B_k$ of responses is measured by the sum of squares

$$\text{SSQ} = \sum_{i,j} (A_{ij} - A^*_{ij})^2,$$

where A^*_{ij} is the ideal data matrix closest to A_{ij}. Of course A^*_{ij} is constant within each cluster of the partition, so it is defined by

$$\sum_{i,j \in B_p} (A^*_{ij} - A_{ij}) = 0, \qquad p = 1, 2, \cdots, k.$$

Equivalently,

$$\text{SSQ} = \sum_p \sum_{i,j \in B_p} (A_{ij} - b_p)^2,$$

where b_p is the average value of A_{ij} in the cluster B_p. Every partition $B_1, \cdots, B_k$ can be evaluated by computing its SSQ and for very small data matrices it might be possible to look at all partitions. Of course, if k is the size of the matrix, SSQ will be zero, and generally comparisons of SSQ should only be made for k fixed.

For real data, some quick method of reaching a reasonable partition must be devised. The following splitting algorithm is of a familiar type used in one-way clustering, with some complications due to the requirement that marginal clusters form trees (see Figure B). At the kth step of the algorithm, there will be a partition into k clusters $B_1, \cdots, B_k$. The cluster B_p has the set of rows R_p and columns C_p and will sometimes be denoted (R_p, C_p). The number of rows in R_p is r_p, and the number of columns in C_p is c_p. A "split" of B_p is a division of B_p into two clusters B'_p and B''_p either by rows or columns. For a row split, $B'_p = (R'_p, C_p)$ and $(B''_p = R''_p, C_p)$, where R'_p, R''_p is a partition of R_p.

Beginning with the partition consisting of a single set (the set of all indices of the matrix), the algorithm proceeds by splitting selected clusters in the partition. Thus at the kth step, the partition changes from $B_1, B_2, \cdots, B_p, \cdots, B_k$ to $B_1, B_2, \cdots, B_{p-1}, B'_p, B''_p, \cdots, B_k$. The reduction in SSQ due to splitting $B_p = (R_p, C_p)$ into $B'_p = (R'_p, C_p)$ and $B''_p = (R''_p, C_p)$ is

$$\text{SSQR} = c_p r'_p (A(B'_p) - A(B_p))^2 + c_p r''_p (A(B''_p) - A(B_p))^2,$$

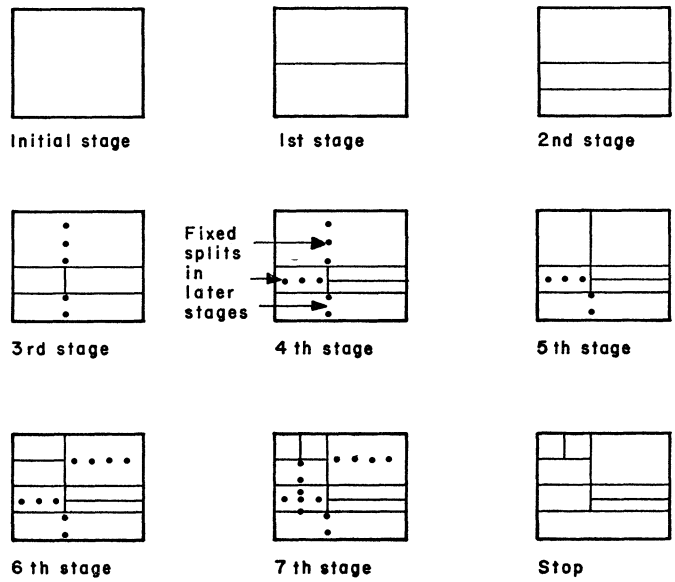
where $A(B)$ denotes the average of A over the block B. It is evident that only the row means are necessary to determine this reduction, whatever R'_p and R''_p, and that the maximum reduction over all possible 2^p splits will occur in order of the row means, i.e., all row means in R'_p will be less than all row means in R''_p. It is therefore sufficient to test $(r_p - 1)$ divisions of the ordered means to find the optimal row split of B_p.

If B_p is split by rows, it may happen that the marginal row cluster R_p contains some other row clusters R'_q, R''_q due to a previous splitting of some other cluster B_q. In this case, the only split of B_p conforming to the marginal tree requirement will have $R'_p = R'_q$, $R''_p = R''_q$. Such a split is called a *fixed* split; if there are no prior constraints, the split is called a *free* split. Since a free split selects among all possible divisions of the row cluster, a larger sum of square reduction will be expected just due to chance. It will be shown in Section 4 that the SSQ reduction due to free splits should be multiplied by $\pi/2r_p$, $r_p > 2$.

The algorithm thus proceeds by choosing that split which maximizes SSQ reduction at the kth step, using adjusted SSQ for free splits. The splitting stops when the reduction in SSQ due to further splitting is less than that expected by chance. Once again some complications enter due to the presence of free and fixed splits. Let $SS1$ be the sum of squares reduction due to all free splits, and let N_1 be the total number of columns and rows involved in free splits. Let $SS2$ be the sum of squares reduction due to fixed splits and let N_2 be the total number of such fixed

B. SCHEME OF SPLITTING ALGORITHM

SEQUENCE OF STEPS OF ALGORITHM



splits. Let $SS3$ be the sum of squared deviations within clusters, and let N_3 be the total number of responses less the number of clusters. The algorithm stops when $SS1$ and $SS2$ are not large compared with their expected values based on $SS3$—more precisely, when $SS3/N3 > (SS1/2 + SS2)/(N_1/\pi + N_2)$. The unusual weighting of $SS1$ and $SS2$ will be justified in Section 4.

4. FIXED AND FREE SPLITS

During the algorithm, the rows of clusters are split in two ways: (1) by a free split which maximizes the SSQ reduction over all divisions into two disjoint sets of rows, or (2) by a fixed split which divides the rows into two predetermined disjoint sets. Let $((A_{ij},\ j=1,\ \cdots,\ n),\ i=1,\ \cdots,\ m)$ denote the values in the cluster, and suppose that $A_{ij}=\mu+\sigma\xi_{ij}$ where ξ_{ij} are unit normal, and independent. It is desired to know the distribution of SSQ reduction due to free or fixed splits, so that the value of splitting can be assessed. If the actual SSQ reduction is not high compared to the expected reduction under the null model, then the splitting will not be executed.

The fixed split reduction has a well known distribution. Suppose the given split is at row k, and let $\overline{X}_1$ denote the mean of all observations $\{A_{ij},\ i \le k\}$ and $\overline{X}_2$ denote the mean of all observations $\{A_{ij},\ i > k\}$. Then SSQ reduction $= (\overline{X}_1 - \overline{X}_2)^2 nk(m-k)/m$, and this is distributed as $\sigma^2\chi_1^2$.

Because it is chosen to be the maximum reduction over all divisions into two sets of rows, the distribution of the free split reduction is complicated. If $X_{(1)}, X_{(2)}, \cdots, X_{(m)}$ denote the ordered row means, then

$$\text{SSQ reduction} = \max_k \left(\frac{X_{(1)} + \cdots X_{(k)}}{k} - \frac{X_{(k+1)} + \cdots X_{(m)}}{m-k} \right)^2 \cdot nk(m-k)/m.$$

Under the null hypothesis, $X_{(1)}, X_{(2)}, \cdots, X_{(m)}$ are order statistics from a normal distribution $N(\mu, \sigma^2/n)$. My colleague L. J. Savage has shown that, for m large, the optimal k is within $0(\sqrt{m})$ of $\frac{1}{2}m$, and that the SSQ reductions for all k within $0(\sqrt{m})$ of $\frac{1}{2}m$ differ by $0(1)$. For large m, then, the distribution of SSQ reduction may be determined by splitting at any convenient place near the median of $X_{(1)}, \cdots, X_{(m)}$. Using k, $X_{(k)} \le \mu \le X_{(k+1)}$, as the splitting point, it follows that conditionally on k,

$$\text{SSQ reduction} = (\overline{X}_1 + \overline{X}_2)^2 nk(m-k)/m + 0(1),$$

where $\overline{X}_1$ is the mean of k half normals and, independently, $\overline{X}_2$ is the mean of $(m-k)$ half normals. The mean and variance of the leading term is available conditionally on k, and k is binomially distributed with expectation $\frac{1}{2}m$ and variance $\frac{1}{4}m$. It then follows that

$$\text{E (SSQ reduction)} = 2\sigma^2 m/\pi + 0(1).$$
$$\text{Var (SSQ reduction)} = 8\sigma^4 m/\pi + 0(1).$$

Asymptotically then

$$\text{SSQ reduction} \sim 2\sigma^2 \chi^2_{m/\pi} + 0(1).$$

Using this distribution theory, the algorithm may be modified to treat fixed and free splits differently. For free splits of more than two rows,

$$\text{MSQ} = (\text{SSQ reduction})\ \pi/2m,$$

where m is the number of rows. For other splits

$$\text{MSQ} = \text{SSQ reduction}.$$

The algorithm proceeds at each stage by executing that split with smallest MSQ.

To stop, consider all free splits of more than two rows or columns, with total sum of squares reductions

$$SS1 \sim 2\sigma^2 \chi^2_{N_1/\pi},$$

where N_1 is the total number of rows or columns freely split. And all other splits, with total sum of squares reduction

$$\text{SS2} \sim \sigma^2 \chi^2_{N_2},$$

where N_2 is the number of such splits. Finally consider the sum of squares within blocks,

$$SS3 \sim \sigma^2 \chi^2_{N_3},$$

where N_3 is the total number of data points, less the number of blocks.

Stop if $SS3/N_3 > (\frac{1}{2}SS1 + SS2)/(N_1/\pi + N_2)$. Thus the algorithm stops when further splits, on average, do not reduce prediction error.

An example of the algorithm appears in Table 6. In that table the quantity $SS3/N_3$ is called mean square within blocks; $(\frac{1}{2}SS1 + SS2)/(N_1/\pi + N_2)$ is mean square due to new splits. These are recorded at each step in the algorithm.

5. DIRECT CLUSTERING VERSUS DISTANCE CLUSTERING

The average distance algorithm is a joining algorithm which begins with a set of clusters consisting of single objects, and forms new clusters by coalescing that pair of clusters between which the average (over pairs of objects in the two clusters) distance is smallest. This algorithm was applied to the voting data for both years and states. The resulting clusters are compared with the marginal clusters of the two-way clustering algorithm, in Figure C.

There are disappointingly few contradictions in the two trees. In the states, Texas is placed by the average distance algorithm with (FA, AS) rather than (GA, AA, LA). Examination of Table 6 shows that Texas was rather similar to (GA, AA, LA) up to 1948, but has since become more similar to (FA, AS). This fact is signalled in the joint clustering by the splitting off of Texas in later years. In the years, 1904 is placed by the average distance algorithm with 1908 rather than with 1944, 1948, 1916. In Table 6, there seems no clear cut advantage in either assignment. Year 1968 is not grouped with (1908, 1924, 1900, 1920) in the average algorithm, which

6a. TWO-WAY CLUSTERING OF REPUBLICAN VOTE

State	Year																		
	12	36	32	40	44	48	16	04	68	08	24	00	20	28	56	60	52	64
SC	1	1	2	4	4	4	2	5	39	6	2	7	4	9	25	49	49	59
MI	2	3	4	4	6	3	5	5	14	7	8	10	14	18	24	25	40	87
GA	4	13	8	15	18	18	7	18	30	31	18	29	29	45	33	37	30	54
LA	5	11	7	14	19	17	7	10	23	12	20	21	31	24	53	29	47	57
AA	8	13	14	14	18	19	22	21	14	24	27	35	31	48	39	42	35	70
TS	9	12	11	19	17	25	17	22	40	22	20	31	24	52	55	49	53	37
FA	8	24	25	26	30	34	18	21	41	22	28	19	31	57	57	52	55	48
AS	20	18	13	21	30	21	28	40	31	37	29	35	35	39	46	43	44	44
VA	17	29	30	32	37	41	32	37	43	38	33	44	38	54	55	52	56	46
NC	12	27	29	26	33	33	42	40	40	46	40	45	43	55	49	48	46	44
TE	24	31	32	33	39	37	43	43	38	46	44	45	51	54	49	53	50	44
KY	25	40	40	42	45	41	47	47	44	48	49	49	49	59	54	54	50	36
MD	24	37	36	41	48	49	45	49	42	49	45	52	55	57	60	46	55	35
MO	30	38	35	48	48	42	47	50	45	49	50	46	55	56	50	50	51	36
WV	21	39	44	43	45	42	49	55	40	53	49	54	55	58	54	47	48	32
DE	33	43	51	45	45	50	50	54	45	52	58	54	56	65	55	49	52	39

from the table is again not decisively right or wrong. The average algorithm groups 1956 with 1952, 1960 rather than with 1928. The 1956, 1928 grouping is based on the large Republican votes in those two years in the border states. But the grouping (1956, 1952, 1960) is also unobjectionable.

Another sort of difference between the two trees is that there are a larger number of clusters in the average distance method, the number of objects less one. For example there are four clusters within the (KY, MO, WV, MD, DE) cluster, versus one for the two-way clustering. All of these states are very similar to each other, and the four clusters are not all all reliable when sampling errors in the distances are considered. The two-way clustering seems more correct in suppressing tiny, irrelevant details. Next, looking at large clusters, the average algorithm has two groups (SC, MISS, GA, AA, LA) and (rest), whereas the two-way method has four groups, (SC, MISS), (GA, AA, LA, TS), (FA, AS), and (rest). From the table, there does seem to be a good case for this set of four clusters; (SC, MISS) seem distinctively different from (GA, AA, LA) in particular. The grouping (SC, MISS, GA, AA, LA) is also very unstable under sampling error on the distances. The absence of splitting in the small clusters is due to the stopping rule. The suppression of clusters at the larger level is more interesting in that it can only happen when early free splits are later verified by further fixed splits. Thus 64 is first split off in the cluster 28–64, SC–AS. Then this split is verified independently for the states VA–DE.

The conclusion is that the clusters resulting from the two methods are not very different, but that the average algorithm has more clusters than the two-way clusters and that these extra clusters can be suppressed without loss because they are not reliable. The average algorithm is conceptually simpler than the two-way method, though somewhat more expensive in computation. The real ad-

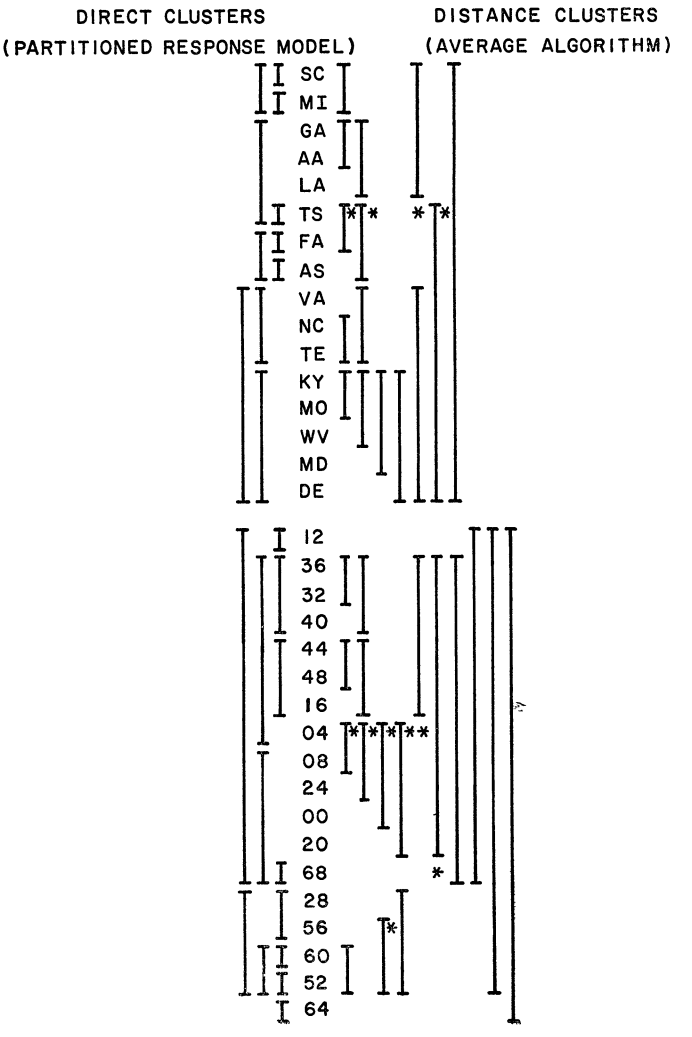
6b. MEAN SQUARE ANALYSIS[a]

	Boundary	Best split	All new splits	Within blocks		Boundary	Best split	All new splits	Within blocks
1	AS, 12-64	2645	2045	277	19	12, FA-AS	338	91	46
2	20, SC-AS	1765	940	184	20	AA, 28-52	228	89	45
3	20, VA-DE	1924	507	113	21	56, VA-DE	185	73	43
4	MI, 12-20	840	363	107	22	04, GA-TS	165	66	42
5	52, SC-AS	545	269	92	23	04, VA-TE	518	72	38
6	MI, 28-52	1240	315	87	24	04, KY-DE	345	57	36
7	52, VA-DE	1227	279	83	25	04, FA-AS	229	49	35
8	MI, 64-64	672	219	79	26	TE, 64-64	154	46	34
9	TE, 12-20	394	205	76	27	28, SC-MI	121	41	34
10	SC, 64-64	392	179	69	28	60, MI-MI	113	39	33
11	56, SC-MI	372	173	68	29	FA, 68-20	90	37	33
12	TS, 12-20	288	166	65	30	TS, 64-64	85	37	33
13	TS, 28-52	285	135	61	31	AA, 64-64	408	43	33
14	FA, 28-52	300	134	61	32	68, KY-DE	83	36	31
15	MI, 60-52	272	126	60	33	68, FA-FA	205	34	30
16	12, KY-DE	229	121	59	34	40, VA-TE	78	30	29
17	12, VA-TE	1120	143	52	35	40, KY-DE	302	34	28
18	12, LA-TS	628	102	48	36	40, FA-AS	149	27	27

[a] See Sections 3, 4.

NOTE: The first split runs from years 1912 to 1964, beneath state AS. The second split runs from states SC to AS, after year 20.

C. REPUBLICAN VOTING DATA



vantage of the two-way method does not lie in the marginal clusters, but in the interpretation directly on the data matrix expressing the interaction between the two marginal clusters. It may be worthwhile to work out the two marginal trees in advance, using one of the distance algorithms. Then construct the joint clustering by considering the variance in all blocks formed from pairs of marginal clusters, retaining only those blocks whose variance is sufficiently small.

6. ERROR ANALYSIS

A general technique for assessing statistical error using subsamples is described in [13]. Applied here, a randomly selected subsample (each observation lies in the subsample with probability 0.5) of the data is used, with the remaining observations treated as missing. This process is repeated many times, with the two-way clusterings obtained regarded as a random sample from the "true" two-way clustering for an infinite amount of data. Two such clusterings are displayed in Table 7.

The principal divisions are maintained in the subsamples (with some fringe adjustments)—for states (MI, SC), (LA, AA, GA, TS, FA, AS), (NC, VA, TE), (MO, WV, KY, DE) and for years (1964), (1952, 1956, 1960, 1928), (rest). The finer divisions do not appear. (It was in these finer divisions that contradictions appeared with the average distance algorithm.)

[*Received January 1971. Revised August 1971.*]

REFERENCES

[1] Ball, G. H., "Data Analysis in the Social Sciences," American Federation of Information Processing Societies Conference Proceedings, *Fall Joint Computer Conference* 27, 1, Washington: Spartan Books, 1965.

[2] Beale, E. M. L., "Euclidean Cluster Analysis," *Bulletin of the ISI*, 43, 2 (1969), 92–4.

[3] Blackwelder, R. A., *Taxonomy*, New York: John Wiley & Sons, Inc., 1966.

[4] Bonner, R. F., "On Some Clustering Techniques, *IBM Journal*, 22 (1964), 22–32.

[5] Cheetham, A. H. and Hazel, J. E., "Binary (Presence-Absence) Similarity Coefficients," *Journal of Palaentology*, 43 (1969), 1130–6.

[6] Edwards, A. W. F. and Cavalli-Sforza, L. L., "A Method for Cluster Analysis," *Biometrics*, 21 (1965), 372–5.

[7] Fisher, W. D., *Clustering and Aggregation in Economics*, Baltimore, Md.: Johns Hopkins, 1969.

[8] Fortier, J. J. and Solomon, H., "Cluster Procedures," in Krishnaiah, ed., *Multivariate Analysis*, New York: Academic Press, 1966.

[9] Friedman, H. P. and Rubin, J., "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association*, 62 (December 1967), 1159–78.

[10] Good, I. J., *Categorization of Classification Mathematics and Computer Science in Biology and Medicine*, London: Her Majesty's Stationery Office, 1965.

7. TWO SUBSAMPLE CLUSTERINGS OF VOTING DATA[a]

State	Year																
	12	32	48	36	44	40	24	16	00	08	68	04	56	20	28	52	60	64
SC	1	2	4				2		7	6					9			59
MI	2		3		6				10		14	5	24		18	40		87
LA	5	7			19	14			21		23			31	24		29	
GA		8	18	13	18		18	7		31		18				30	37	
AS	20	13		18		21	29	28			31	40		39	39		43	
TS			25		17	19	20	17				22		24	52			37
AA	8	14	19	13	18	14			35		14	21			48	35		70
FA				24	30	26				22	41	21	57		57	55		
NC	12		33		33					46	40					46	48	44
VA		30					33	32			43	37					52	46
TE	24	32				33	44			46							53	44
MD	24					41	45					49					46	
KY		40			45						44					50	54	
WV	21				45			49	54			55			58		47	32
MO			38	38	48		50				45	50	50					36
DE		51					58	50		52		54	55	56	65			39

	44	12	36	32	48	16	40	68	24	04	00	20	08	64	60	52	56	28
MI					3			14	8			14	7	87				
SC	4		1	2			4				7	4				49		
LA			11	7	17	7				10				57		47		
GA	18	4		8		7	15	18	18	29			31	54	37		33	43
AA		8	13			22	14	14		21		31		70	42		39	48
TS	17					17	19			22				37			55	52
FA	30	8	24		34	18	26			21		31		49				57
AS	30				21	28	21					39	37	44	43	44	46	
VA		17		30	41		32	43			44	38	38			56		
NC			27	29	33				40	40		43	46		48		49	
TE	39	24					33	38		43	45	51		44	53	50		54
MO		30	38		42		48						49			51	50	56
KY		25	40	40	41	47			49				48	36		50	54	59
WV				44	42			40	49	55				32		48	54	58
MD			37			45		42			52	55	49	35				57
DE					50		45			54		56	52	39	49		55	65

[a] Each data point is set missing with probability 0.5.

[11] Gower, J. C. and Ross, G. J. S., "Minimum Spanning Trees and Single Linkage Cluster Analysis," *Applied Statistics*, 18 (1969), 54–64.

[12] Hartigan, J. A., "Representation of Similarity Matrices by Trees," *Journal of the American Statistical Association*, 62 (December 1967), 1140–58.

[13] ———, "Using Subsample Values as Typical Values," *Journal of the American Statistical Association*, 64 (December 1969), 1303–17.

[14] Jardine, N. and Sibson, R., "A Model for Taxonomy," *Mathematical Biosciences* 2 (1968), 465–82.

[15] Johnson, S. C., "Hierarchical Clustering Schemes," *Psychometrika*, 32 (1967), 241–54.

[16] King, B. F., "Market and Industry Factors in Stock Price Behaviour," *Journal of Business*, 39 (1966), 139–90.

[17] McQuitty, L. L., "Expansion of Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data," *Educational and Psychological Measurement*, 27 (1967), 253–5.

[18] Peterson, S., *A Statistical History of the American Presidential Elections*, New York: Frederick Ungar, 1963.

[19] Sokal, R. R. and Sneath, P. H. A., *Principles of Numerical Taxonomy*, San Francisco: W. H. Freeman and Co., 1963.

[20] Tryon, R. C., *Cluster Analysis: (Correlation Profile and Orthometric Factor Analysis for the Isolation of Unities in Mind and Personality)*, Ann Arbor, Mich.: Edwards Brothers, 1939.

[21] ——— and Bailey, D. E., *Cluster Analysis*, New York: McGraw-Hill Book Co., 1970.

[22] Ward, J. H., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58 (March 1963) 236–44.