AN OUTLINE TO HOMALS – 1

August 1978

Jan van Rijckevorsel

Jan de Leeuw

Department of Datatheory
University of Leiden
Wassenaarseweg 80
Leiden
The Netherlands

DEPARTMENT OF DATATHEORY/FACULTY OF SOCIAL SCIENCES
UNIVERSITY OF LEYDEN - THE NETHERLANDS

An outline to Homals-1

Introduction [1]

Homals-1 is the name of a computer program for the multidimensional analysis of categorical data. Extensive instructions for the use and implementation of this program are given.

Also described are the data and the model. Great emphasis is given to the intuitive appealing, geometrical approach of model and results. Several examples of applications in political science,biology,sociology, psychology and archeology are shown.

The theoretically non-interested reader can skip par.3,4, and 5.

The Homals-1 model is a form of principal components analysis on binary indicator matrices (Guttman(1941),Burt(1950),de Leeuw(1973)).

The computer program is written and optimized by the Department W.T. of the computing centre (C.R.I.) of the University of Leiden. Especially we want to thank Tom Visser, who did all the programming.

[1] This study was partly financed by the Nederlandse Organisatie voor Zuiver-Wetenschappelijk Onderzoek, ZWO, Grant nr. 56-97

1.0 Several approaches and other work

Several approaches to Homals-1 are possible. In the first place Homals-1 tries to solve a particular multidimensional scaling problem, which also could be solved (less efficiently) by general multidimensional scaling programs as Smacof (Heiser,1977). In the second place Homals-1 is closely related to canonical analysis, in particular to discriminant-analysis. And finally Homals-1 can be represented as a nonlinear generalization of principal component analysis.

The multidimensional scaling approach has a strong geometrical orientation. It uses the intuitive concept of distance; the basic idea is that elements belonging to the same subsets should be represented by points that are close together, and subsets sharing the same elements should also be close together. Homals-1 essentially represents elements and subsets in a joint space. Thus it is a form of nonmetric unfolding of categorical data. It is well known by now that nonmetric unfolding very often gives degenerate and uninteresting representations. By strenghtening the nonmetric requirements Homals-1 very often finds interesting representations. The interpretation of Homals-1 as an unfolding algorithm seems to be new.

Another interpretation closely related to multidemsional scaling ideas has been used by Benzécri and his school. They present a technique which is equivalent to Homals-1 under the name 'analyse des correspondences'. The basic reference is Benzéçri et al. (1973). In analyse des correspondence we formulate a metric multidimensional scaling problem, which is solved by the classical technique of metric scaling. It turns out that the solution is the same as the solution to the nonmetric scaling problem mentioned earlier.

The discriminant or canonical approach to Homals-1 is the classical one. The key papers are Fisher (1940) and Guttman (1941). This approach can also be formulated in geometric terms, using distances, which makes it very similar to the multidimensional scaling problems discussed earlier. De Leeuw (1973) also derives the Homals-1 equations in this way. This approach is used in this paper in par. 3.

The principal component interpretation of Homals-1 is also quite old. It started with Hirshfeld (1935), was extended by Fisher, Maung, and Lancaster and further extended by Naouri (1971), and Pousse and Dauxois (1976). The work of Lancaster, Maung, Hirshfeld, Fisher, Sarmanov is reviewed in Lancaster (1968). Nonlinear principal component analysis is not very geometrical, but either algebraic or probabilistic or both. There are interesting connections with Pearson's theory of contingency, and thus with chi square.

2.0 A nontechnical description

2.1 The Data

The data consist of rectangular matrices which contain categorical information on which no a priori measurement characteristics are defined. In general the rows are elements and the columns partitions defined on these elements. Every partition has some mutually exhaustive disjoint subsets and every element has a profile of scores on these partitions. It is a question of terminology whether one talks about observations in stead of elements, about variables in stead of partitions and about categories in stead of subsets. To reach a general audience we prefer elements,partitions and subsets. Examples of data structures which can be represented in this way are given in Guttman (1941), Burt(1950), Lingoes(1968) and de Leeuw(1973). This kind of data are characterized by the fact that the partitions are able to classify the elements into discrete subsets. The partitions are therefor sometimes called classificatory variables. Sets of such classificatory categorical partitions are questionnaires, tests for mental ability, rating scales, social economic classifications, dental compositions in mammals, kill ratio's of sera on cells, etc.. The naming of rows and collumns is arbitrary so that in zoology they are called animals and phylogenetic characteristics; in archeology graves and gravegifts; in political science bills and political parties; in psychology observations and items; in medicine cells and sera etc.. All columns define partitions of elements (rows) into mutually exhaustive disjoint subsets.

2.2 The Model

We want to metricize subsets and elements. We do not metricize the partitions self, only their subsets. Metricizing means that we want attribute numerical values, i.e. coordinates in an euclidean space, to subsets and elements on such a way that we get a simple and attractive picture or plot.

We call a solution , i.e. plot, simple when it is of low dimensionality and we call it attractive when it has nice geometrical properties. Nice geometrical properties are:

) Subsets and elements are in a joint space.

) Elements that share most subsets with other elements are representative and therefore central in space.

) Elements that share the least subsets with all other elements are unique and therefore excentrical in space.

) Elements that share an unique group of subsets are homogeneous and therefore contiguous in space.

) Unique groups of subsets are heterogeneous and therefore separated in space.

We want to obtain such such simple and attractive plots without any assumption about the distribution of elements over subsets within one partition or between partitions. The only demand we make is that subsets within a partition are disjoint and mutually exhaustive. If one can make additional assumptions one should use stronger models. But if stronger models hold for the data this will easily be recognized as such in the Homals-1 solution.

The model tries to fit an underlying partition whose subsets are as disjoint as possible and as exhaustive as possible. This means geometrically that subsets have to be as homogeneous as possible and therefore contiguous in space, and that the set of subsets must be as heterogeneous as possible and therefore separated in space. This underlying partition has not to coincide with any of the dimensions of the solution space, neither does it have to coincide with a conceptual idea, although that would be quite convenient. In the case of only one manifest partition subsets are perfectly disjoint and exhaustive, so there will be perfect homogeneity within subsets and perfect heterogeneity between subsets. Hence it is easy to fulfill our requirements for a simple and attractive plot.

2.2.1 Stress

If we could replace all partitions in the dataset by just one, it would again be easy to make our simple and attractive plot. The more difficult it is to replace all partitions by just one other one, the more difficult it will be to create a simple and attractive plot. A measure for the degree of difficulty is stress. The lower the stress the better the model succeeded with representing all partitions by one other and finding 'our' plot.It of course hardly ever happens that a set of partitions can perfectly be represented by just one. It would be rather trivial to do any analysis on such data. Generally one finds a representative partition that is the best fitting average partition for all partitions in the dataset. It depends on the data wether this 'average' partition is very representative for the whole set. If so stress will be low; if not stress will be high. In the trivial case there will be perfect homogeneity and stress will be zero. and vice versa in the case of perfect heterogeneity, also trivial, stress will be equal to the number of dimensions. Stress per dimension is equal to one minus the eigenvalue belonging to that dimension.

2.2.2 Normalization and minimization

As explained stress is a measure of fit. The overall stress is the sum of the stresses per dimension . Depending on the normalization chosen the eigenvalues are computed from the metricized elements, the first normalization, or from the metricized subsets, the second normalization. This overall stress is the real loss minimized in the algorithm.

How things are minimized and what the differences are between the two normalizations is dealt with in par. 3. The different normalizations exist only for computational reasons. If there are more elements than subsets the second normalization is more efficient and if there are more subsets than elements the first one is more efficient.

For understanding the model the difference in normalization is only misleading.

2.2.3. Missing data or incomplete datasets

Although missing data are a nuisance in most models, they are not in Homals-1. Every subset defines a geometrical restriction on the elements in the final solution, but only for the elements in that subset. Combinations of restrictions locate the points of elements and the points of subsets in the plot. If an element is missing in a certain partition, which means that it cannot be attributed to any of the subsets of that particular partition, then its point in the final plot will not be restricted by that very partition but only by the other ones. Elements- and subsetpoints are only defined by the non-missing entries of the dataset.

2.2.4 Dimensionality,generalization and fit

The dimensionality of the solution is decided upon by the user. As soon as the plot satisfies him or her in a conceptual sense there is no need for more dimensions. In most cases something like a real underlying dimensionality does not exist so we don't have to look for it. One chooses a two- or threedimensional solution.because it generates simple plots. There is nothing wrong with looking upon Homals-1 as a technique to create such plots. The stress per dimension reflects the relative importance of a dimension.

Inductive generalization from sample to population is generally not possible without extra empirical information or strong additional assumptions.

Because the partitions are not fitted by straight lines or other specified functions any functional relationship between subsets can be represented.

2.2.5 Partitioning

If there exists one dominant a priori partitioning of the data, one can use its subsets as elements in a Homals-1 analysis without modifying the original data. This means that all elements within an a priori subset are made equal. The Homals-1 analysis will be done on these a priori subsets in stead of the original data. This is a rather rough procedure which ignores differences between elements within an a priori subset. We doubt wether this option is very useful. An example of an application is given in the analysis of the salmon data. (par.9)

3.0 A technical description

3.1 The data

Suppose $I=\{1,2,\ldots\ldots,n\}$ is a finite index set. The data are partitions of I into disjoint subsets. Partition Π_r is written as $\Pi_r = \{\Pi_1^r,\Pi_2^r,\ldots,\Pi_{m(r)}^r\}$, i.e. the number of subsets in Π_r is m(r), and all these subsets are supposed to be nonempty and disjoint. The number of subsets is n and apart from missing data, the subsets are supposed to be exhaustive as well. We also define the binary indicator matrix $\Delta^r = \{\delta_{ij}^r\}$ of the partition Π_r as a n x m(r) matrix with $\delta_{ij}^r=1$ if and only if $i \in \Pi_j^r$; $\delta_{ij}=0$ otherwise. And we define a n x n binary diagonal matrix with $M^r = \{m_i^r\}$ with $m_i^r=1$ if and only if $i \in \Pi^r$ and $m_i^r=0$ otherwise; and $M^* = \sum_{r=1}^{s} M^r$.

3.2 The quantifications

A (p-dimensional) quantification of I is a mapping of I into R^p. Each quantified element of I is thus a p-element vector x_i. For a given quantification $x_1, x_2, \ldots, x_n$ and a given partition Π with indicator matrix Δ we can define the within-subset means

$$x_j^* = \frac{\sum_{i=1}^{n} \delta_{ij} x_i}{\sum_{i=1}^{n} \delta_{ij}} \tag{1}$$

and the overall mean..

$$x^* = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

Each x_i can be written as

$$x_i - x^* = \sum_{j=1}^{m} \delta_{ij}(x_j^* - x^*) + \sum_{j=1}^{m} \delta_{ij}(x_i - x_j^*) \tag{3}$$

The first term of the right is the deviation of the subset mean from the total mean for the subset containing x_i, the second term is the deviation of x_i from its subset mean. A quantification is <u>homogeneous</u> if the last term is relatively small for each i. Geometrically a quantification (of a given partitioning) is homogeneous if the points in the same subset are close together (in R^p), and the subsets are far apart from each other.

3.3 Measurement of homogeneity

To give a precise quantitative definition of homogeneity we define the symmetric matrices.

$$T = \sum_{i=1}^{n} (x_i - x^*)(x_i - x^*)^T \tag{4a}$$

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{l=1}^{m} \delta_{ij}\delta_{il}(x_i - x_j^*)(x_i - x_l^*)^T \tag{4b}$$

and

$$B = \sum_{j=1}^{m} n_j^*(x_j^* - x^*)(x_j^* - x^*)^T \tag{4c}$$

with

$$n_j^* = \sum_{i=1}^{n} \delta_{ij} \tag{5}$$

some easy algebra gives the important identity

$$T = B + W \tag{6}$$

Moreover $B = 0$ if and only if all subsets have the same vector of means, and $W = 0$ if and only if all elements in the same subset coincide. Thus the case $B = 0$ (or $W = T$) corresponds with perfect non-homogeneity (or heterogeneity), and the case $W = 0$ (or $B = T$) corresponds with perfect homogeneity (or non-heterogeneity). A quantification is homogeneous if W is small relative to T (or if B is large relative to T). A convenient measure of homogeneity is the ratio of determinants

$$\lambda = \frac{|B|}{|T|}, \tag{7a}$$

and in the same way a convenient measure of heterogeneity is

$$\mu = \frac{|W|}{|T|} \tag{7b}$$

Both measures lie between zero and one, and assume these bounds only in the case of perfect homogeneity of heterogeneity. Another very important property of both λ and μ is that they are invariant under nonsingular linear transformations of the quantification X, i.e. if X is an n x p matrix containing the quantifications and S is a nonsingular p x p matrix then $\tilde{X} = XS$ defines a different quantification with exactly the same value of λ and μ. There is similar invariance under translations of the space, i.e. we can add a constant vector to all rows of X.

3.4 Finding the optimal quantifications

We can now study the problem of finding X in such a way that λ is maximized (or μ is minimized). The problem is trivial if there is only a single partition (s=1). because we can always make $\lambda=1$ and $\mu=0$ by setting all x_i within a subset equal to x_j^* which is completely arbitrary. If there are more partitions. however, a perfect trivial solution with respect to all partitions is no longer possible in general. With obvious changes of notation we define

$$B_r = \sum_{j=1}^{m(r)} n_{jr}^*(x_{jr}^* - x^*)(x_{jr}^* - x^*)^T \tag{8a}$$

$$W_r = T - B_r. \tag{8b}$$

$$B = \sum_{r=1}^{s} B_r \tag{9a}$$

$$W = \sum_{r=1}^{s} W_r \tag{9b}$$

With these new definitions the basic matrix equation (6) is still true except for a constant s, but the problem of maximizing λ or minimizing μ is no longer trivial. Observe that the invariance results and the bounds for λ and μ remain true with this new definition of B and W.

The problem of maximizing λ is studied first. We can restrict ourselves (because of the invariance of λ under t translations) to quantifications $x_1, x_2, \ldots, x_n$ with $X^*=0$. In this case T reduces to

$$T = \sum_{i=1}^{n} M_i^* x_i x_i^T \qquad (10a)$$

and B_r reduces to

$$B_r = \sum_{j=1}^{m(r)} n_{jr}^* x_{jr}^* x_{jr}^{*T} \qquad (10b)$$

A more convenient matrix expression for B_r can be obtained by using the n x m(r) indicating matrix Δ_r, the diagonal matrix $\Psi_r = \Delta_r^T M_r \Delta_r$ with elements n_{jr}^*, and the n x p matrix X with the X_i then,.

$$B_r = X^T M_r \Delta_r \Psi_r^{-1} \Delta_r^T M_r X \qquad (11)$$

of course

$$T = X^T M^* X \qquad (12)$$

and

$$W_r = T - B^r = X^T\ M^* - M_r \Delta_r \Psi_r^{-1} \Delta_r^T M_r \quad X \qquad (13)$$

letting

$$A_r = \quad M_r \Delta_r \Psi_r^{-1} \Delta_r\ r \qquad (14a)$$

$$A = \sum_{r=1}^{s} A_r \qquad (14b)$$

we find

$$B = X^T A X \qquad (15)$$

The matrix A is a symmetric matrix of order n. A_r can be computed very easely by observing that element (i,1) is zero if i and 1 are not in the same subset of partition Δ_r, and element (i,1) is equal to $(n^{+}_{jr})^{-1}$ if i and 1 are both in I_{jr}.

It follows from (12) and (15) that

$$\lambda(X) = \frac{X^T A X}{X^T M^* X} \qquad (16)$$

Because λ is invariant under nonsingular transformations we can restrict our attention to quantifications X that satisfy $T = X^T M^* X$ (and also, of course, $eM^*X = 0$).
Thus the problem of maximizing λ (X) is equivalent to maximizing $X^T AX$ over all X that satisfy $X^T M^* X = I$ and $eM^*X = 0$. By using the fact that all rows and columns of A add up to one, and by using a familiar theorem of Ky Fan (of Beckenbach and Bellman, 1961, section 32) we find that

$$\max_{\substack{X^T M^* X = I \\ eM^+ X = 0}} \lambda(X) = \prod_{k=1}^{p} \Theta_k \qquad (17)$$

if

$$1 = \Theta_0 \geq \Theta_1 \geq \ldots \geq \Theta_{n-1} \qquad (18)$$

are the ordered eigenvalues of A. Moreover the maximum is attained by taking X equal to the eigenvector corresponding with Θ_1, Θ_2,..., Θ_p (if some of the eigenvalues are equal additional qualifications are needed, but they are not essential). It is clear that after the eigenvector solution is computed we can apply translations and non-singular linear transformations of the space, if that seems useful for some reason. It is also clear by the same reasoning that , at the same time, the solution that minimizes μ is found and that the two solutions are identical.

Given the solution for X we can compute the within subset means (collected in the m(r) x p matrices X_r^*) by the rule

$$X_r^* = \Psi_r^{-1}\Delta_r^T M^r X \tag{19}$$

If we use the eigenvector solution for X this implies that

$$\sum_{r=1}^{s} (X_r^*)^T \Psi_r (X_r^*) = X^T A X = \Theta \tag{20}$$

with Θ the diagonal matrix of eigenvalues.

4.0 The algorithm and its rationale

The computatinal procedure suggested by our analysis in section 4 is very straightforward. We construct the n x n matrix A and find its p + 1 largest eigenvalues with corresponding eigenvectors (the dominant eigenvector corresponds with a trivial solution, we have $\Theta_0 = 1$,, and all elements of the vector are equal to $n^{-\frac{1}{2}}$]. It is, however, often true that n is very large, in which case an eigen-analysis of A may become prohibitive. Observe that usually p <<n, and that consequently we do not want to use techniques that compute all values and vectors anyway.

Define the matrices

$$H_r = M_r \Delta_r \Psi_r^{-\frac{1}{2}} \tag{21}$$

and collect them in the supermatrix

$$H = \{ H_1, H_2, \ldots , H_s \} \tag{22}$$

Clearly

$$A = HH^T \tag{23}$$

Now suppose

$$H = X\Theta^{\frac{1}{2}}Y^T \tag{24}$$

is the singular value decomposition of H. Then X and Θ can be found by computing eigenvectors and eigenvalues of A, but also by first computing Y and Θ as eigenvectors and eigenvalues of

$$C = H^TH = Y\Theta Y^T, \tag{25}$$

and then X by

$$X = HY\Theta^{-\frac{1}{2}} \tag{26}$$

Observe that from (19)

$$X^* = \Psi^{-\frac{1}{2}}H^TX \tag{27}$$

With Ψ diagonal matrices containing Ψ_r. If we combine this with (26) we find

$$X^* = \Psi^{-\frac{1}{2}}H^T\ HY\Theta^{-\frac{1}{2}} = \Psi^{-\frac{1}{2}}Y\Theta^{\frac{1}{2}}. \tag{28}$$

Thus X^* is a simple row and columnwise rescaling of Y, and singular value decomposition (truncated at rank p + 1) gives both X^* and X.

Operating on C may be more efficient than operating on A, because C is often of a smaller order than A. Moreover computing C usually gives more interesting information, because the submatrices C_{rt} defined by

$$C_{rt} = H_r^T H_t = \Psi_r^{-\frac{1}{2}} \Delta_r^T M_r M_t \Delta_t \Psi_t^{-\frac{1}{2}} \tag{29}$$

contain the (rescaled) bivariate mariginals of the s-dimensionnal discrete distribution defined by the classifications. Moreover the elements of C have interesting relations with the chi-square values that can be computed for these bivariate tables (Guttman 1941, Burt 1950, De Leeuw 1973, section 3.7). But again the order of C may be too large to make an eigen-analysis practical. We may consequently have to use methods that do not construct either A or C, and that use the sparseness of H in an efficient way. Thus we define the new loss function:

$$\Phi(X;Y_1,Y_2,\ldots,Y_s) \equiv \sum_{r=1}^{s} \text{tr } (X-\Delta_r Y_r)^T M_r (X-\Delta_r Y_r) \tag{30}$$

and we also define $\Phi_Y(X)$ as the minimum of $\Phi(X;Y_1,Y_2,\ldots,Y_s)$ over $Y_1,Y_2,\ldots,Y_s$ and $\Phi_X(Y)$ as the minimum of $\Phi(X;Y_1,Y_2,\ldots,Y_s)$ over X. Straightforward computation gives

$$\Phi_Y(X) = \sum_{r=1}^{s} \text{tr } X^T(M^* - M_r \Delta_r \Psi^{-1} \Delta_r^T M_r)X = \text{tr } X^T W X \tag{31a}$$

and

$$\Phi_X(Y) = \text{tr } Y^T C Y \tag{31b}$$

with W defined by (9b) and C defined by (29).

So maximizing λ gives the same solution as minimizing $\Phi_Y(X)$ over all X that satisfy $X^TM^*X = I$ and as minimzing $\Phi_X(Y)$ over all Y that satisfy $Y^T\Psi Y = I$. Moreover minimizing $\Phi(X;Y_1,\ldots,Y_s)$ over X satisfying X^TM^*X as well over unrestricted $Y_1,\ldots,Y_s$ or minimizing $\Phi(X,Y_1,\ldots,Y_s)$ over Y satisfying $Y^T\Psi Y = I$ as well over unrestricted X. The normalization on X we call more norm 1 and the normalization over Y norm 2. Both problems can be solved conveniently by an alternating least squares (or block relaxation) technique. The idea is to minimize Φ alternatingly over Y with X fixed at its current value, and over X with Y at its current value. In the algorithm of norm 1 we start with $X^{(0)}$ and set the iteration counter γ equal to zero, we then alternate the two following steps:

$$Y_r^{(\gamma)} = \Psi_r^{-1}\Delta_r^TM_rX^{(\gamma)} \tag{32a}$$

$$X^{(\gamma+1)} = M^{*-\frac{1}{2}}\tilde{X}^{(\gamma)}\left[\tilde{X}^{(\gamma)T}\tilde{X}^{(\gamma)}\right]^{-\frac{1}{2}} \tag{32b}$$

with

$$\tilde{X}^{(\gamma)} = M^{*-\frac{1}{2}}\sum_{r=1}^{s}\Delta_rY_r^{(\gamma)} \tag{32c}$$

In norm 2 we start with $Y^{(0)}$;we alternate the following two steps:

$$X^{(\gamma)} = M^{*-1}\sum_{r=1}^{s}\Delta_rY_r^{(\gamma)} \tag{33a}$$

$$Y^{(\gamma+1)} = \Psi^{-\frac{1}{2}}\tilde{Y}^{(\gamma)}\left[\tilde{Y}^{(\gamma)T}\tilde{Y}^{(\gamma)}\right]^{-\frac{1}{2}} \tag{33b}$$

with

$$Y^{(\gamma)} = \Psi^{-\frac{1}{2}}\sum_{r=1}^{s}X^{(\gamma)}M_r\Delta_r \tag{33c}$$

The method has the enormous advantage that the matrix products in (32a) and (32c) or (33a) and (33b) can be computed as simple additions, using only information on groups or category membership of the individuals.Thus we do not have to store the wasteful indicator matrices Δ_r, if p is much smaller than both n and m then the required storage space is of order n x m(r).If we compare (19) and (32a) we see that the X_r^* are computed simultaneously. Of course (32b) or (33b) is the most expensive step. If n or sm $(= \sum_{r=1}^{s} \sum_{k=1}^{m} \Pi_{k(r)}^{r}) \gg p$ it requires approximately $\frac{3}{2}np^2$ (or $\frac{3}{2}(sm)p^2$) multiplications to carry out this step. Computation of the inverse symmetric square root requires less than 10 p^3 multiplications, especially in the final iterations the product matrix will be already close to diagonal...

This suggests that we use the procrustus type orthogonalization and it also suggests that we stop iterating when the Procrustus subroutine tells us that the matrix we feed into it is already diagonal.

The method (32a and 33b) is closely related to Bauer's generalization of the power method (Bauer 1957), and to the modifications of this method proposed by Rutishauser (1967,1970).

The main difference is that we do not use an eigenvalue method to compute $X(X^TX)^{-\frac{1}{2}}$, but a class of iterative methods proposed by Leipnik (1971). It is often suggested that methods of this type should not be used for the general symmetric eigen problem when n/p < 4, but in our case the alternative methods based on (tri) diagonalization are often not feasible or very wasteful, because A and C can be extremely large, and computing A and/or C destroys the sparseness of Δ_r.

5.0 Partitioning

The set I of par.3.0 itself can be a partitioning of another set Q, $Q = \{1,\ldots\ldots,t\}$, $t>n$ into n subsets which are nonempty, disjoint and exhaustive. Let P be this partitioning then there exists a binary indicator matrix $\Delta^P = \{\delta^P_{ij}\}$ with $\delta^P_{ij}=1$ if and only if $i\epsilon P_j$, and $\delta^P_{ij}=0$ otherwise.

If Π^Q_r is a partitioning of Q , then there exists a binary indicator matrix $\Delta^Q_r = \{\delta^{Qr}_{ij}\}$ with $\delta^{Qr}_{ij}=1$ if and only if $i\epsilon \Pi^{Qr}_j$ and $\delta^{Qr}_{ij}=0$ otherwise. There exists also a t x t binary diagonal matrix M^{Qr}, $M^{Qr} = \{m^{Qr}_i\}$ with $m^{Qr}_i=1$ if and only if $i\epsilon\Pi^{Qr}$ and $m^{Qr}_i=0$ otherwise.

Define $$\tilde{\Delta}_r = \Delta^{P^T}\Delta^Q_r$$

and $$\tilde{M}_r = \Delta^{P^T}M^Q_r\Delta^P$$

$\tilde{\Delta}_r$ is not a binary indicator matrix anymore

If we replace Δ_r by $\tilde{\Delta}_r$ and M_r by $\tilde{M}_r$ and apply the results of par.3 and par.4 we do a kind of discriminant analysis on categorical data where there is no within group variation.

It is not clear at the moment how useful this a priori partitioning is.

6.0 The Program

6.1 Some general remarks

Homals-1 is an ANS Fortran IV computer program which works satisfactorily on IBM,CDC and Univac installations and which can easily be implemented on other computers.

The IBM version has dynamical storage allocation by means of an Assembler routine. For other installations this routine is replaced by a static Fortran array allocation routine.

There are two version of the program:

Homals-1 version 3.02A	The data are read from any specified unit formatted. The whole raw data matrix is kept in core.
Homals-1 version 3.02B	The data are read unformatted from any rewindable unit. Only one row-vector of the data matrix is kept in core at the time. This B version is especially made for handling large datasets on smaller computers and CDC installations.

6.2 The structure

Homals-1 is written according to the principles of structured programming because is has to be read and understood by other programmers and it has to be portable.

The program is a member of a set of Homals programs which share several subprograms. This means that also the set of Homals programs is structured and that,given one complete Homals program, one can easily build another one out of that with some extra routines.

The structure of the program has several levels. The main level is dominated by the two normalizations, which are divided into four combinations of missing data (yes-no) and a priori partitionings (yes-no) each. In these eight (4+4) subprograms always exist four phases:

The initialization
The quantification
The orthogonalization
The I/O phase

The iterative process takes place during quantification and orthogonalization. The other two phases are only done once.

6.3 Optimization

Variable dimensioning is used for all arrays except for one, dynamically or statically allocated, superarray, which shares storage with all arrays. Starting pointers of these arrays are computed according to the problem parameters.

We tried to reach more efficiency in execution time by timing and optimizing the program. Optimization is done, using the IBM Fortran X compiler, by rewriting comparatively slow parts of the program. Timing is done during analysis of large datasets. Performance on other type installations is roughly just as good (Univac) or better (CDC). The utmost attention is given to obtain optimal efficiency within the iterative process.

6.4 Performance

Except for a certain overhead of program instructions and I/O the execution time of Homals-1 is approximately linear with ISUC or NOBS, depending on the normalization, all other parameters fixed. The execution time is quadratic with the numbers of dimensions all other parameters fixed. To give some idea of the performance we analyzed a small set of randomly generated datasets.

IT Number of iterations	NVAR Number of partitions	NDIM Number of dimensions	ISUC Number of subsets	NOBS Number of elements	Execution-time in secs. CPU Norm 1	Norm 2
75	80	2	228	57	12.14	13.79
75	80	2	228	114	21.72	22.80
75	80	2	228	228	40.55	39.91
75	80	2	228	456	77.37	72.97
75	80	2	228	912	153.64	143.14
49	80	2	50	400	9.46	9.19
75	80	2	100	400	23.00	19.60
75	80	2	200	400	39.39	35.61
75	80	2	400	400	70.80	70.01

Table 1 Execution times in secs. CPU on IBM 370/158

The difference in execution times between the two normalizations becomes substantial when ISUC is at least ten times as big or ten times as small as NOBS.

Randomly generated data are the most heterogeneous as possible and therefore there is no convergence within 75 iterations. Dealing with real data one can expect 30 iterations as average. This is based on our limited experience of ± 200 runs. The execution times in Table 1 should be divided by two to find a rough estimate of the execution time of analysis of a dataset of corresponding size.

The B version is necessarily slower, caused by the heavy I/O within the iterative process.

6.5 Specific remarks for the user

elements=observations
partitions=variables
subsets=categories

6.5.1 The data

)The data are supposed to be on unit INPU, the first parameter of the I/O card; if the data are on card, they must follow the parametercards.

)In version A rows have to correspond with observations and columns with variables. Data will be read formatted according to the user's format.

In version B the input data matrix has to be transposed: rows correspond with variables and columns with observations. The data are read unformatted from any rewindable unit.

data

) The must be positive integers starting with the number one. The number on the category card is regarded as the highest meaningful category number and thus the total number of categories of that particular variable.
If the missing data parameter IMIS, the second parameter of the analysis card, is set to one, any integer number outside the range 1 - the total number of categories of a variable, is considered as a missing value
)The first NVR2 variables are the variables analyzed. When there are more variables in the datamatrix according to format, they must be used either as labels for the plot of observation points, or as an a priori analysis partition.

6.5.2 The plots

There are two kinds of plots available: labeled- and unlabeled plots.
Unlabeled plots : The points coinciding in one cell are counted. The symbols printed show how many points fall in a particular cell; the symbols 1-9 are selfevident,the symbols A-Z correspond with the numbers 10-35. When there are more than 35 points the rest will be indicated with the '+' symbol. The exact number of points for each '+' is specified after the plot.
Labeled plots : The printed symbols correspond with the category numbers of the labeling variable. The symbolyzing is analogue the unlabeled plot. The only difference is that when there are more than 35 labels the symbol attribution starts again with number 1. In this case a '+' is printed when two points coincide; the labels for those points are specified after the plot.

When the number of dimensions is greater than two, only the first two dimensions will be plotted.

6.5.3 The normalizations

As far as output is concerned the two normalizations are the same with the exception that the plot of all category scores is only available for the second normalization.
The default option is that the most efficient normalization is chosen. The NORM parameter can override this default.

NORM 1	NORM 2
more efficient when: NOBS >> ISUC	more efficient when : ISUC >> NOBS
no categories scores plot of all categories	categories scores plot of all categories
category scores are on a smaller scale because they are the centroids of the observation scores.	observation scores are on a smaller scale because they are the centroids of the category scores.

6.6 Notes for implementation

) The first two assigment statements of the subroutine HOMAIN have the following meaning :
'ICAR'=a ; a is the logical unit number for the card reader.
'IPRI'=b ; b is the logical unit number for the printer.

) The subroutine DECLAR allocates a superarray of a certain specified length. It has to be large enough to contain all arrays used in the program. If the array is not large enough the program will return from DECLAR with an error message and the correct size of the superarray. In stead of the DECLAR subroutine any storage allocator can be used, for instance a dynamic storage allocator.

) The labeled common block 'VARBLS' contains parameters and array sizes, all stored as integers. This common block is defined in the HOMAIN subroutine.

) Array-types and their wordlengths

Type	Lenght	Name
Integer	32 bits	datain,datapv,icatgo, marfre,mafrpv,prodat iparti
Real	32 bits	auxili
Double Precision	64 bits	marfin,mfpvin,accumu obscor,catsco,stress

) The random number generator in the subroutine RANDMA requires a necessary available integer arithmetic of 125 * 348525375 (< 2**29). If this number is to large another random number generator, generating real numbers in double precision between -1 and +1, must be implemented.

)If the plots are not square, or if they are to large or to small for your printer, you can adapt some statements in PLOTTO, the plotting subroutine, according to the notes in the source of the subroutine. These notes are indicated with the symbols :
c?>>>

)In subroutine IMTQL2 is machep the machine dependent parameter specifying the relative precision of floating point aritmetic. The actual value is 2**-20

7.0 Deck set up HOMALS-1, version 3.01 A/B.

PARAMETER CARDS:

column	format	name	meaning
1 - 5	I5	NJOB	number of jobs

Then for every job:

Title card:

1 - 80	20A4	NAME	title of the job

Problem card:

1 - 5	I5	NOBS	number of observations
6 - 10	I5	NVR1	number of variables in the datamatrix
11 - 15	I5	NVR2	number of analysis variables (≤ NVR1)
16 - 20	I5	NDIM	number of dimensions
21 - 25	I5	MAXC	number of categories of the variable with the maximum amount of categories (analysis-partitioning-variable, if present, included)
26 - 30	I5	ISUC	total number of categories of the analysis variables
31 - 35	I5	NUPA	=0: no analysis-partition =I: (0 < I ≤ NVR1) variable I is partitioning

Analysis card:

1 - 5	I5	NORM	normalization method (1 or 2) (0: method internally chosen)

column	format	name	meaning
Analysis card (cont.):			
6 - 10	I5	IMIS	=0: the datamatrix does not contain missing data =1: the datamatrix possibly contains missing data
11 - 15	I5	MAXI	maximum number of iterations (default 75)
16 - 20	E10.8	EPSI	convergence criterium (default .5E-6 = 0.0000005)
I/O card:			
1 - 5	I5	INPU	unit-number of the datamatrix (default: card unit-number)
6 - 10	I5	IDAT	- only for version A - printing of the datamatrix =0: no =1: yes
11 - 15	I5	IXCH	- only for version A - printing cross-tabs, chi-squares and degrees of freedom =0: no =1: chi-sqares and degrees of freedom =2: all
16 - 20	I5	IWRI	printing of observation- and category scores =0: no =1: observation scores only =2: yes =3: category scores only
21 - 25	I5	IPLO	plotparameter =0: no plots =1: plot of observation scores, unlabeled, or (if NUPA ≠ 0) labeled by obs.number; if NORM = 2, plot of category scores, labeled by their variable number =2: same as 1 + plots, according to the contents of the vector "IPARTI", that is in this case read from the card(s), preceding the format-cards

column	format	name	meaning
I/O card (cont.):			
26 - 30	I5	IOUT(1)	unit-number for output other than print of the observation scores, preceeded by their identification number; the format used is (I5,8F9.7) (=0: no such output)
31 - 35	I5	IOUT(2)	unit-number for output other than print of the category scores, preceeded by their variable and category number; the format used is (2I4,8F9.7) (=0: no such output)
Category card(s):			
1 - 80	16I5	ICATGO	highest category per variable (for all variables of the datamatrix)

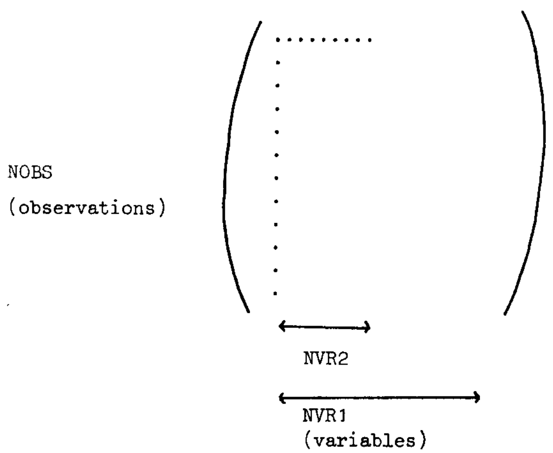
If IPLO = 2 only:

column	format	name	meaning
Plotcard(s):			
1 - 80	80I1	IPARTI	when a '1' is punched in column j: the observation scores will be plotted, labeled by variable j; when a '2' is punched in column j: same as '1' + the category scores of variable j will be plotted, labeled by their category number; when a '3' is punched in column j: same as '2' minus '1'; note: when NUPA ≠ 0 no labeling of the observation scores plot is possible, also such a request is ignored
3 Format cards:			
1 - 80	20A4	FMT1	- only for version 3.01 A - I-format to read the datamatrix
1 - 80	20A4		
1 - 80	20A4		

Additional remarks Deck setup HOMALS 1 version 3.01 $^{A}/_{B}$

About the data matrix:

The data matrix has to have the following form:



also NVR1 $\geq$ NVR2

The program considers the first NVR2 variables from the data matrix as the analysis variables....

Two possibilities:

a. no partition in the analysis (in terms of parameters: NUPA=0)
   when observation score plots are requested(i.e. when IPLO=1 or IPLO=2) the program uses, if, NVR1 > NVR2, the last (NVR1-NVR2) variable(s) each once as a partitioning variable(s) in the plot(s).

b. partition in the analysis (i.e. NUPA > 0)
   1. if the analysis partitioning variable is also one of the analysis variables (i.e. 0 < NUPA $\leq$ NVR2), NVR1 must be equal to NVR2.
   2. if the analysis partitoning variable is not one of the analysis variables, (i.e. NUPA > NVR2), the program considers the variable, immediately following the analysis variable, as the analysis partitioning variable, and NVR1 must be equal to NVR2+1.

8.0 References

Benzécri, J. P. (1974) L'analyse des données. Vols I, II Paris:Dunod

Burt, C. (1950) The factorial analysis of qualitative data. British J. Statist. Psychol.,3, 166-185

Fisher, R. A. (1940) The precision of discriminant functions. Ann. Eugen. London, 10, 422-429

Guttman, L. (1941) The quantification of a class of attributes: a theory and method of scale construction. In: The prediction of Personal Adjustement (P. Horst, ed.), pp. 251-364. New York: Social Science Research Council.

Heiser, W. (1977) How to use SMACOF. Dep. of Datatheory, Univ. of Leiden.

Lancaster, H. O. (1969) The Chi-squared Distribution. New York: Wiley.

de Leeuw, J. (1969) Some contribution to the analysis of categorical data. Dep. of Datatheory, Univ. of Leiden nr RN 004-69

- (1973) Canonical Analysis of Categorical Data. Psychol. Inst., Univ. of Leiden.

Leipnik, R. B. (1971) Rapidly Convergent Recursive Solution of Quadratic Operator Equations. Numer. Math. ,17,1-17

Lingoes, J. (1968) The multivariate analysis of qualitative data. Multivariate Behavioural Research ,3,61-94

Rutishauser, H. (1969) Computational Aspects of F.L. Bauer's simultaneous Iteration Method. Numer. Math. ,13, 4-13

- (1970) Simultaneous Iteration Method for Symmetric Matrices. Numer. Math. ,16, 205-223

9.0 The examples

9.1	Salmo in north western America	Systematic Zoology
9.2	Roll call data	Political Science
9.3	Japanese Religion	Sociology
9.4	Dentition of Mammals	Biology
9.5	Functional Learning	Psychology
9.6	Seriation	Archeology

The examples 9.1 - 9.5 are analysed with an earlier version of the program, version 3.01, This means that the solutions are not oriented towards the principal components. The total stress is still the same. There are only differences in rotation of the final solution and in the stress per dimension. Examples 9.1-9.4 would be somewhat rotated and the stresses in example 9.5 would be very low for the first axis in stead of the second axis.

9.1 SALMO IN NORTH WESTERN AMERICA

Source: Legendre, P., Schreck, C. B., Behnke, R. I. (1972), Taximetric Analysis of Selected Groups of Western North American Salmo with respect to Phylogenetic Divergences. Syst. Zool., 21, 292-307.

This example deals with the classification of North American Salmo along 8 character states (see table 1). The total material consists of 849 specima of the genus Salmo which are all described along these 8 character states. Together they represent 5 different species in addition to some unnamed forms of Northern American trouts.

Table 1 Definition of Character States

1: Pyloric cacca number

state	number
1	17-26
2	26.1-35.9
3	36-44
4	45-60

2: Number of vertebrae

state	number
1	56-60
2	60.1-62.6
3	62.7-64

3: Number of scales in lateral series

state	number
1	121-137
2	138-148
3	148.1-159
4	159.1-200

4: Rows of scales above lateral line

state	number
1	20-24
2	(void state)
3	25-32.9
4	33-37.9
5	38-48

5: Rays in pelvic fin

state	number
1	8.0-9.2
2	9.3-10.2

6: Color of the top of dorsal fin

state	color
1	bright orange or cream
2	tip not of a bright color

7: Color of the cutthroat marks

state	color
1	red
2	yellow
3	(no marks)

8: Size and location of the spotting

state	color
1	large, posterior
2	medium, posterior and more anterior
3	fine and profuse, generalized

The character state score patterns, collection sites and the names of species and forms are shown in table 2

Table 2

Species or form	Source	Sample size	Character states
Rainbow	Whiskey L. Outlet, Alaska	21	43132233
	Wood R., Alaska	30	
	Brooks L., Alaska	16	
	Tebay L., Alaska	22	
	Tikchik L., Alaska	11	
	Alagnak R. (or Branch R.), Alaska	29	
	Coquihaila R., Alaska	30	
	N. Fork Salmonberry R., Alaska	28	
	Lwr. Kathleen L., Yukon	17	
	Mulberly C., Alberta, Canada	32	
	Ruby Valley Fish Hatchery, Nevada	30	
Rainbow	San Pablo C., W. fork, California	30	43132233
Gila	Diamond C., New Mexico	1	21342123
		3	21232123
		2	22332123
		2	22232123
		2	22242123
		2	21241123
		1	21332123
Gila	Diamond C., New Mexico	1	22341123
Mexican Golden	Rio Verde (Fuerte), Mexico	1	11211112
		4	11111112
	Rio Sinaloa, Mexico	5	11111112
	Rio Culiacan, Mexico	4	11111112
Mexican Golden		1	11211112
Rio Truchas	Rio Truchas, Mexico	5	32232233
		1	32112233
		2	31332233
		3	32332233
		1	33232233
		1	31432233
		1	31312233
		1	31432233
Rio Truchas	Rio Truchas, Mexico	1	32231233
Apache	Ord C., Arizona	1	21451122
		1	21441122
		1	12242122
		1	11242122
		1	11252122
		1	11442122
		2	21342122
Apache	Ord C., Arizona	1	21242122

Table 2 continued

Species or form	Source	Sample size	Character states
Cutthroat	Yellowstone L., Wyoming	30	32451211
	Reservoir Canyon near Pine Valley, Utah	13	32451211
	Indian C. and Rio Seco, Colorado	33	32451211
	Headwaters Big Thompson R., Colorado	20	22451211
	L. Victor, N. Fk. Boulder C., Wyoming	15	32451211
	R. Arriba, Canones C., Trib. R. Chama, N. Mex.	6	32451211
	Pacific C., Wyoming	8	22451211
	Cottonwood C., Wyoming	21	32451211
	Home C., Wyoming	8	32452211
Cutthroat	Big Sandstone C., Douglas C., Wyoming	41	32451211
Red-banded	Sheephaven C., California	1	21341122
		1	32441122
		1	32442122
		2	22442122
		1	42442122
		1	42441122
		3	32452122
		1	32451122
		1	42452122
Red-banded	Sheephaven C., California	1	42451122
Calif. Golden	Alpine L., Wyoming	10	11451111
	Cottonwood L. & C.,California	9	21451111
	Golden Trout C., California-types	24	
	Golden Trout C., California	13	
	Whitney C., California	10	
	Golden Trout C., California	31	
	Cottonwood C., California	39	
	S. Fk. Kern R., california	34	
Calif. Golden	Salley Keyes L., California	22	21451111
Kern River	Kern R., California-types	7	32342122
	(Soda Spg., S. Fk. Kaweah, Little Kern, Coyote C., California	8	
	Rifle C., California	22	
	Coyote C., California	33	32342122
	Soda Springs C., California	12	22352122
Kern River	Wet Meadows C., California	17	32342122

Table 3

Rainbow Trout	Salmo gairdneri
Gila Trout	Salmo gilae
Mexican Golden Trout	Salmo chrysogaster
Rio Truchas Trout	unnamed form
Apache Trout	unnamed form
Cutthroat Trout	Salmo clarki
Red-banded Trout	unnamed form
Californian Golden Trout	Salmo aguabonita
Kern River Trout	Salmo aguabonita gilberti

The published material in Legendre et al. (1972) included a total of 104 different objects collected from a large number of different sites, all characterized by their " profile " on the eight character states shown in table 1. These 104 objects represent combinations of geographical collection sites and character state patterns in the sense that there may be two or more objects with the same characteristic in the set, but only if they were caught at different locations.

From this data matrix, as used originally, two more matrices may be constructed. The first one consists of all the 45 different character states patterns in the original matrix, regardless of collection-sites. The second of the constructed data-matrices includes all the specima in the material with all replications within or across collection sites; this defines a 849 by 8 matrix.The contents of the three mentioned matrices are identical in respect to the character states represented, they differ only in respect to the weight given to each of the patterns.Those weights are based on the marginal frequency of each pattern.

In the analysis of these three different matrices the objective was to give an indication of the phenotypic similarities that are assumed to exist between certain members of the genus Salmo,especially in respect to the relict populations of limited distribution,which cannot be readily assigned to either S. gairdneri or S. clarki.The various types that are included in this study are listed in table 3.

In order to find the phylogenetic similarities between species and unnamed forms we partitioned each dataset into 9 groups. Members of a species or unnamed form are hence treated as belonging to a a-priori group which is to be quantified This partitioning is done within the analysis by means of a n added partitioning variable.To see wether species and unnamed forms really were homogeneous groups at all a non partitioned analysis on the same three datasets was done as well.

All six analyses show the same homogeneous (i.e. non-overlapping and separated) series:

1) The Rainbow Trout Series	-Rainbow trout -Rio Truchas trout
2) The Golden Trout Series	-Cutthroat trout -Californian Golden trout -Mexican Golden trout
3) The miscellaneous Series	-Gila trout -Apache trout -Red-banded trout -Kern River trout

The results of the 6 analyses are in fig. 1 to 6.

As the number of character states patterns in the data sets increases from 45 via 104 to 849, there is a tendency for the clearly distinct series to remain distinct as far as the rainbow, cutthroat and californian golden are concerned. However, the other forms and species, save the kern river trout, become less distinct. Separation of the series is still possible but the contiguity is decreased. This effect is caused by the higher marginal frequencies of the distinct species in the bigger data sets.

Fig. 1, 50 unique character patterns (unpartitioned)

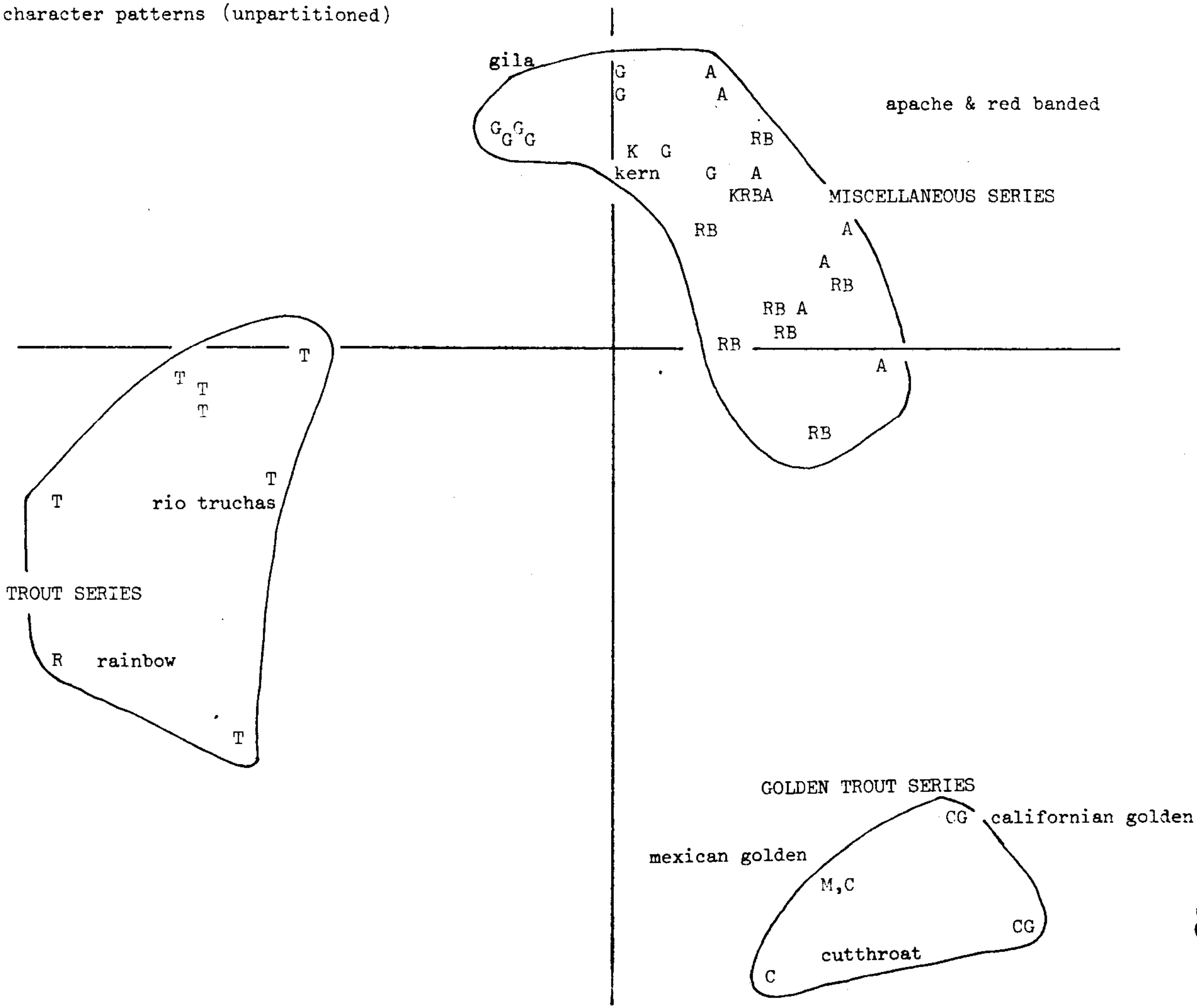


Fig. 2, 50 unique character patterns (partitioned)

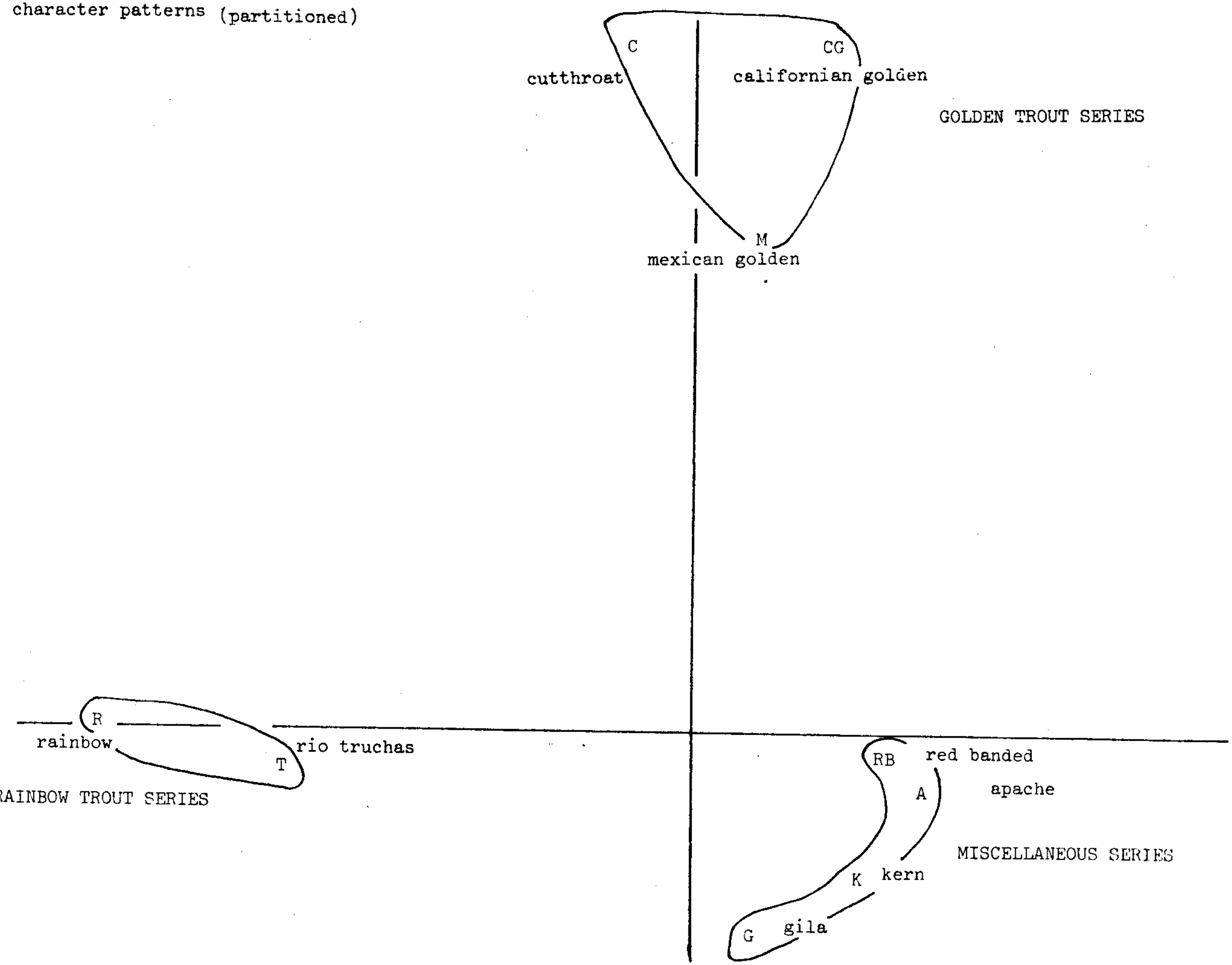


Fig. 3, 849 specima (unpartitioned)

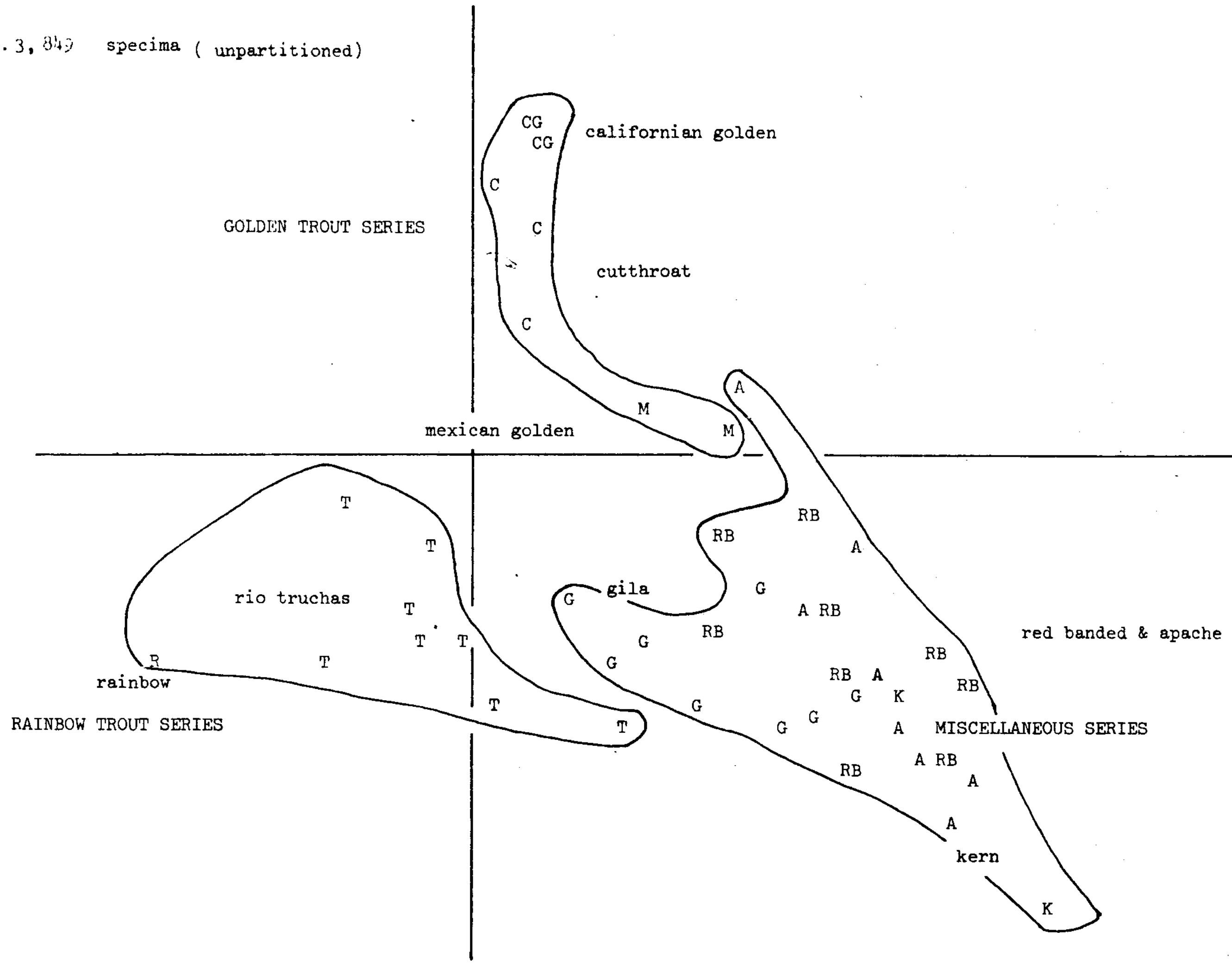


Fig. 4. 849 specima (partitioned)

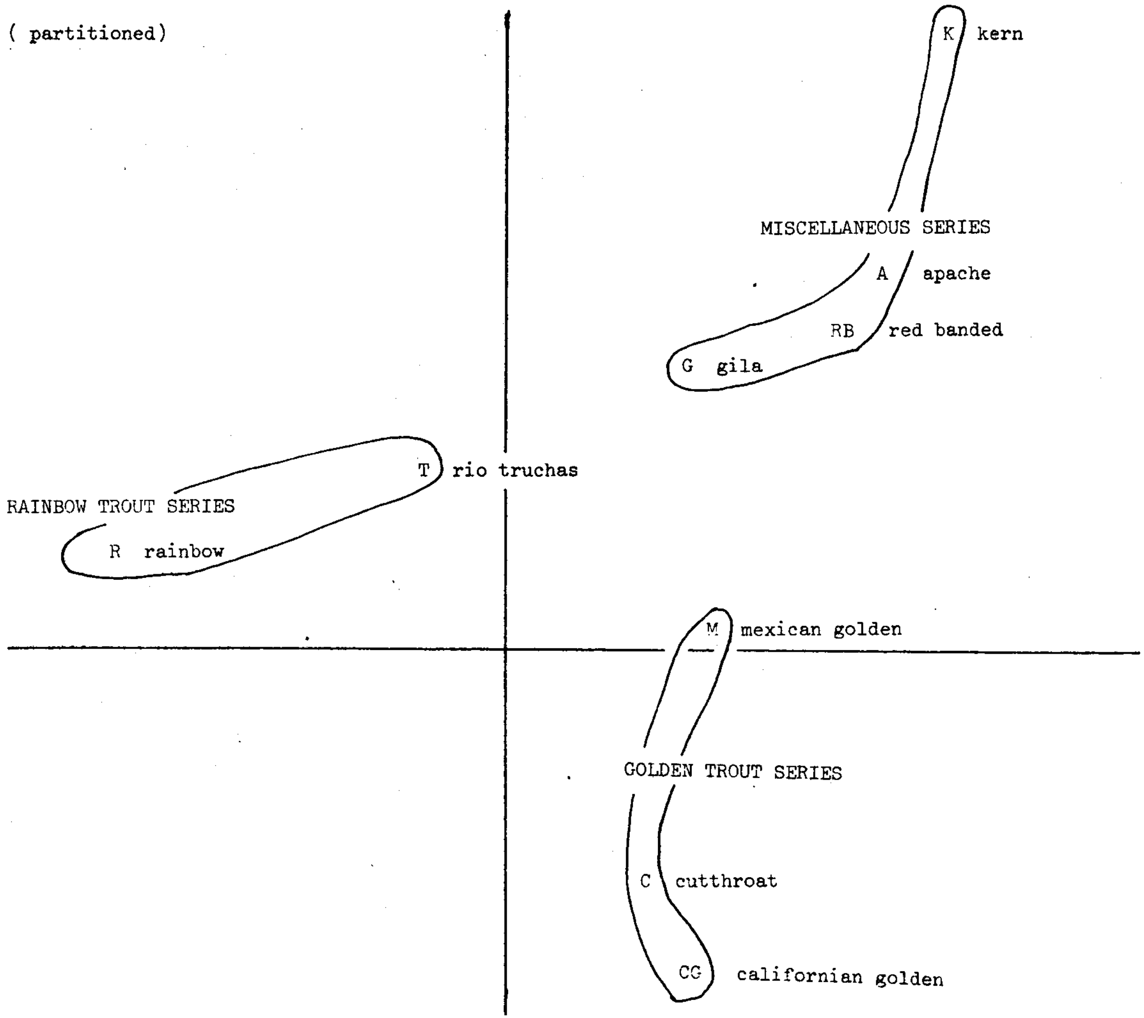


Fig. 5. 104 Objects (unpartitioned)

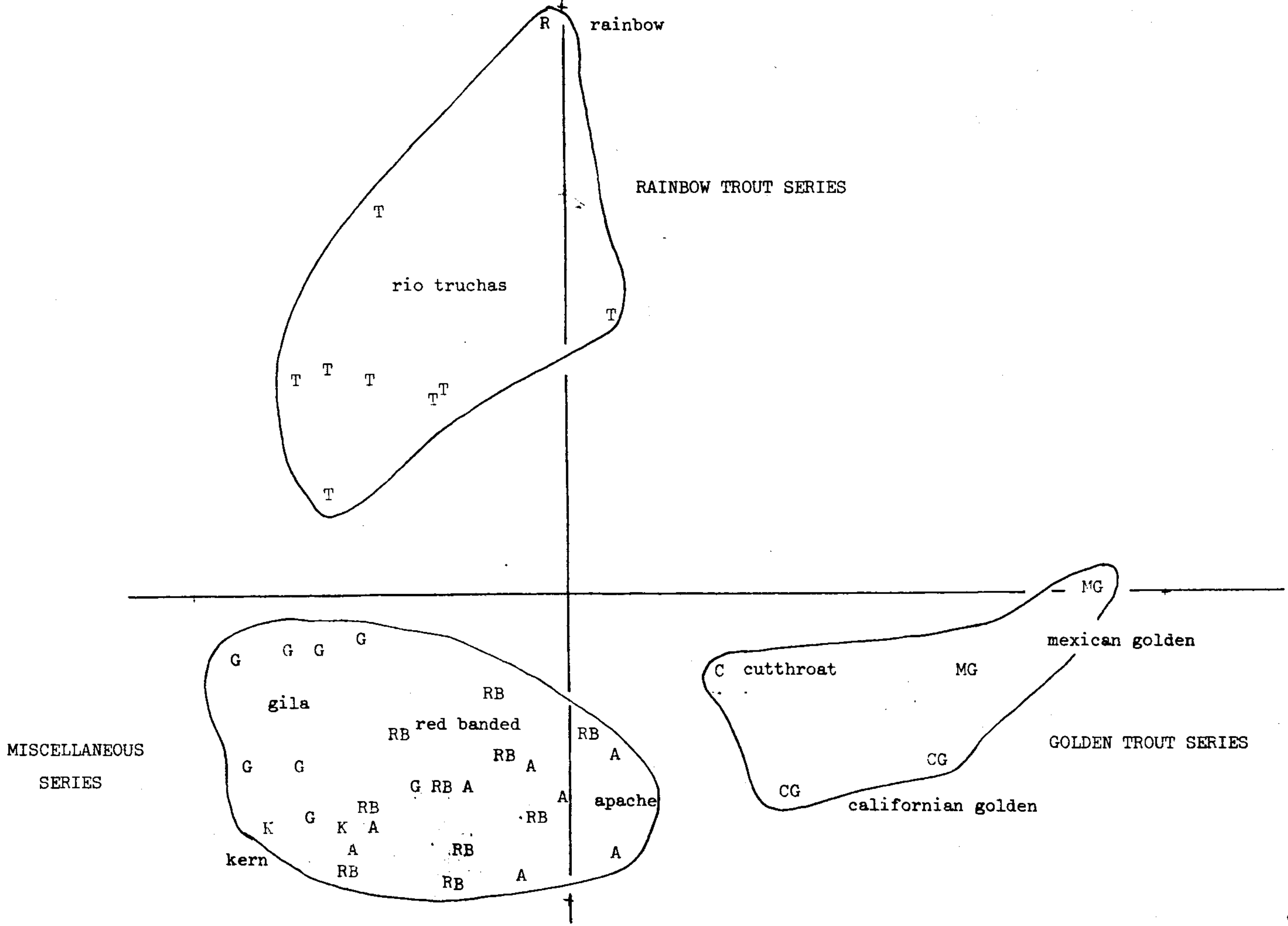
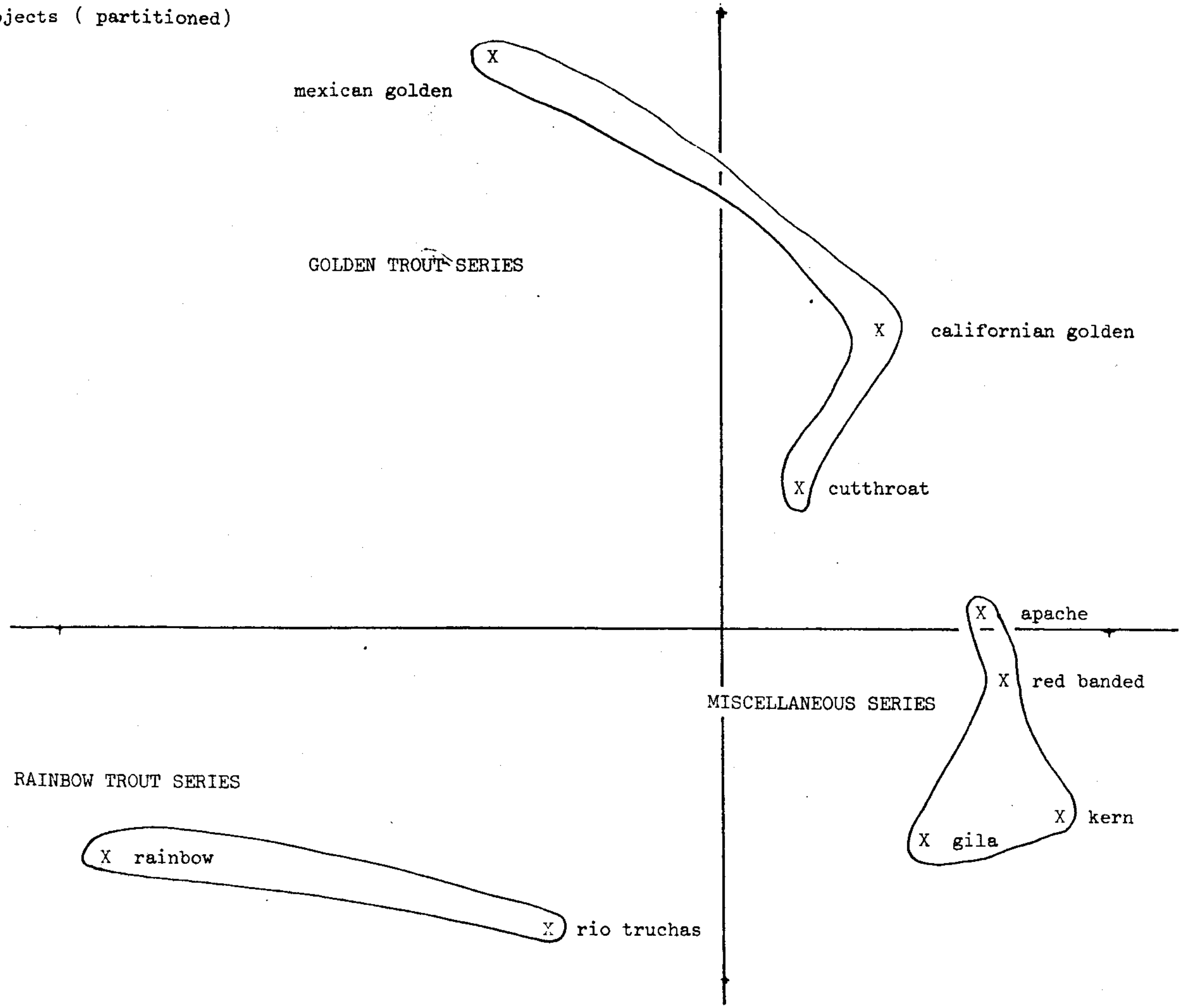


Fig. 6 . 104 objects (partitioned)



9.2 ROLL CALL DATA

The data consist of aye, nay votes on 58 bills that passed the first chamber of the dutch parliament. The votes are classified according to the 11 political parties and to the government's point of view. These data are collected by Menno Wolters, Institute of Datatheory, Leiden.

We are not interested in the classification of bills but in the classification of political parties by means of their aye votes only and by means of their aye and nay votes. Absences and abstentions are treated as missing data. The raw data and their marginal frequencies are shown in table 1.

Table 1

	political parties											
Nr.	1	2	3	4	5	6	7	8	9	10	11	12
Bills Nr. 1	1	1	1	1	1	1	2	2	3	1	1	1
2	1	1	1	2	1	2	1	1	3	2	1	1
3	1	1	1	2	1	2	1	1	3	2	1	1
4	1	1	1	1	1	1	2	2	3	2	1	1
5	1	1	1	2	1	1	1	1	3	2	1	1
6	1	2	1	2	2	2	2	2	3	2	1	1
7	1	1	1	2	1	2	2	2	3	2	1	1
8	1	1	2	1	1	1	2	2	3	1	2	2
9	1	1	1	1	1	1	2	2	3	1	1	1
10	1	1	1	2	1	1	1	1	3	2	1	1
11	3	1	1	1	1	1	1	1	2	2	1	1
12	1	1	1	2	1	2	1	1	2	2	1	1
13	1	1	1	2	1	1	1	1	2	2	1	1
14	1	1	1	2	1	2	1	3	2	2	1	1
15	1	1	1	2	1	3	1	3	2	2	1	1
16	1	1	1	2	1	3	1	3	2	2	1	1
17	2	2	2	1	2	1	2	3	1	1	2	2
18	1	3	1	2	1	2	1	3	2	2	1	1
19	1	1	1	1	1	1	2	3	1	1	1	2
20	3	1	2	1	1	1	2	3	1	1	2	2
21	3	2	1	3	2	2	1	3	2	2	1	1
22	1	1	1	2	1	1	2	3	2	2	1	1
23	1	3	1	2	3	2	2	3	2	2	1	1
24	1	1	1	2	1	1	1	3	3	2	1	1
25	1	1	1	2	1	1	2	3	2	2	1	1
26	2	2	2	1	2	2	2	3	1	1	2	2
27	3	2	2	2	2	2	1	3	2	2	2	1
28	2	2	2	1	2	2	2	3	1	1	2	2
29	1	1	1	2	1	1	1	3	2	2	1	1
30	1	1	1	1	1	1	2	3	1	1	1	1

codes

1 = aye

2 = nay

3 = abstained

Table 1 continued

		political parties											
	Nr.	1	2	3	4	5	6	7	8	9	10	11	12
Bills	Nr. 31	1	1	1	2	1	1	2	1	3	2	1	1
	32	3	1	1	1	2	2	1	2	3	2	1	1
	33	1	2	1	1	1	1	2	2	3	2	2	2
	34	1	2	1	2	1	2	1	3	3	2	1	1
	35	2	2	2	1	2	1	1	2	3	1	2	2
	36	1	1	1	2	1	2	2	1	3	2	1	1
	37	2	2	2	1	2	1	1	1	3	2	1	2
	38	1	1	1	2	1	3	2	1	3	1	1	1
	39	1	2	1	2	1	2	2	1	3	2	1	1
	40	1	1	1	2	1	1	1	2	3	1	1	1
	41	1	1	1	2	2	1	2	3	2	3	1	1
	42	1	2	1	1	2	1	1	3	2	2	1	1
	43	1	1	1	2	1	1	2	3	1	1	1	1
	44	1	1	1	2	1	1	2	3	1	1	1	2
	45	1	1	1	2	1	1	1	3	1	2	1	1
	46	1	1	1	2	1	1	2	3	1	1	1	2
	47	3	2	1	1	2	2	1	3	2	2	1	1
	48	3	2	1	1	1	2	1	3	2	2	1	1
	49	1	1	1	2	1	1	1	3	1	1	2	1
	50	1	1	1	2	1	1	2	3	1	1	1	1
	51	1	1	1	2	1	1	2	3	1	1	1	1
	52	1	1	1	2	1	2	1	3	2	1	1	1
	53	2	1	2	1	1	1	1	3	1	1	1	2
	54	2	1	2	1	2	1	2	3	1	1	2	2
	55	1	2	1	2	2	2	1	3	2	2	2	1
	56	1	1	1	2	1	1	2	3	1	1	1	1
	57	1	1	1	2	1	1	2	3	1	1	1	1
	58	1	1	1	1	1	1	1	3	1	1	2	2

Table 2

			Nr.
In charge is a left-center coalition of the following parties:	PVDA	social democrat	3
	D'66	progressive liberal	11
	ARP	protestant	5
	KVP	catholic	2
	PPR	radicals	12
In opposition were:	VVD	conservative liberal	4
	CPN	communist	7
	PSP	pacifistic socialist	8
	BP	right wing protest	9
	SGP	orthodox calvinistic	10
	CHU	protestant	6

To see wether homogeneity between political parties is affected by aye or nay voting their classification was once based on the aye votes only and once based on aye and nay votes.These classifications were of course done by means of two Homals 1 analyses. The aye votes structure is shown in fig. 1. Coalition and opposition are separated by dotted lines. The analysis of the aye and nay votes produced a party structure of striking similarity, when only the aye votes points are plotted. See fig. 2. This means that the aye votes structure is rather stable and not very influenced by the nay votes. Only the PSP moved a little bit in the direction of the coalition in fig. 2. Both figures show the homogeneity in aye voting of the coalition and the heterogeneity in aye voting of the opposition.

The plot of the nay votes points of political parties is not complementary with figure 1 or figure 2. See fig. 3. Apparently the reasoning for aye voting is not the counterpart of the reasoning for nay voting. Also can be concluded that as far as nay voting is concerned the opposition is more homogeneous and the coalition more heterogeneous.

political parties , aye votes based on aye votes

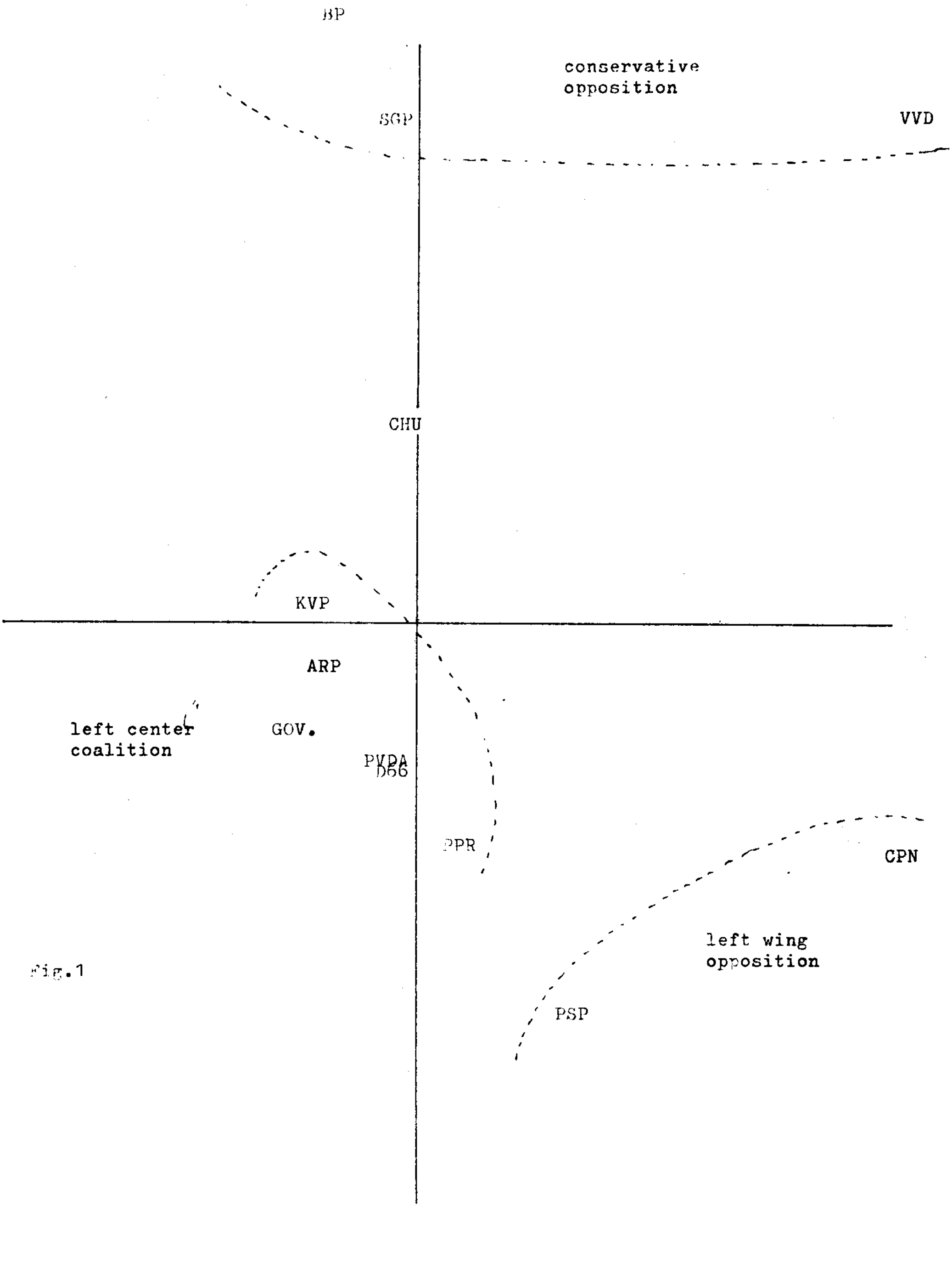


fig.1

political parties, aye votes based on aye and nay votes

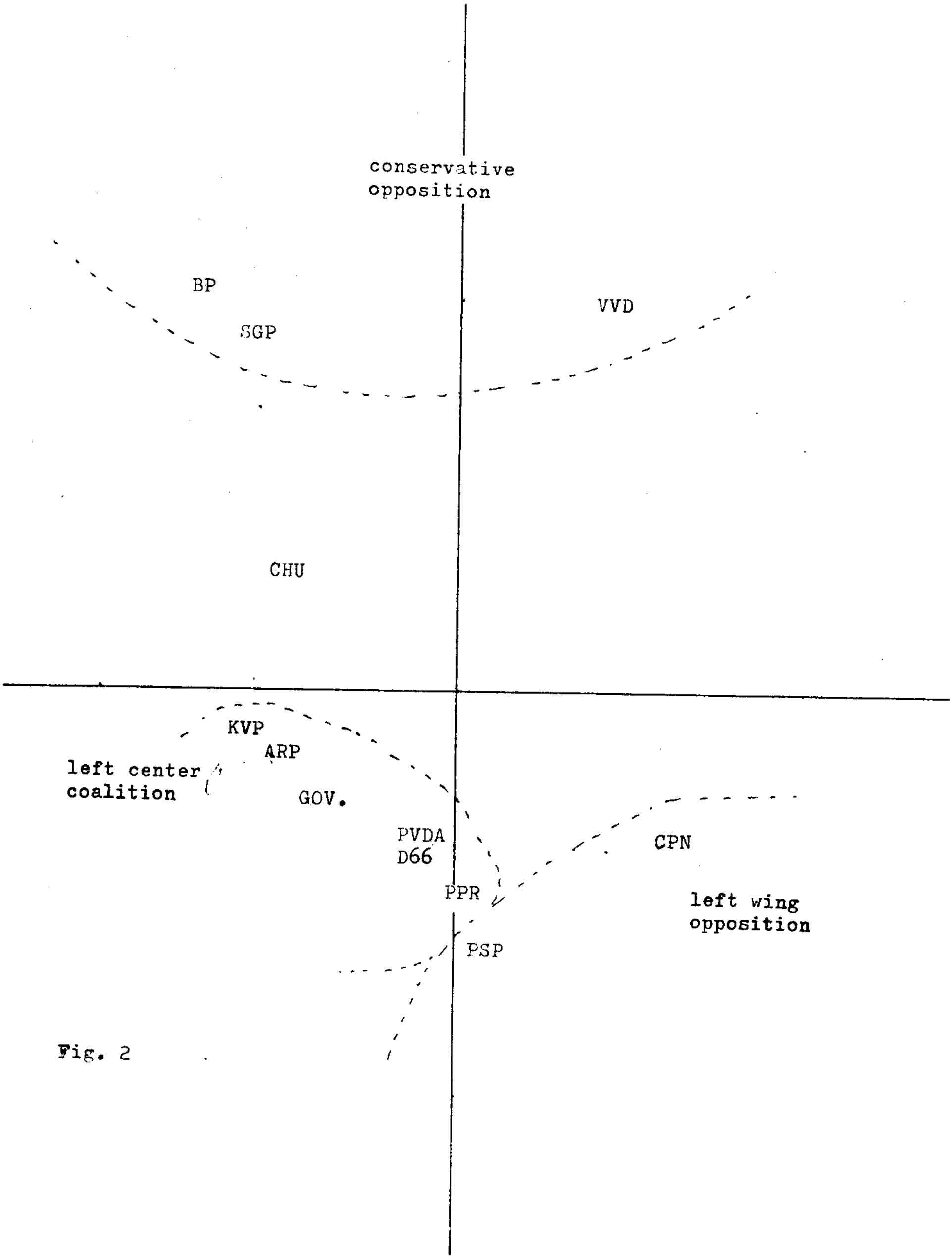


Fig. 2

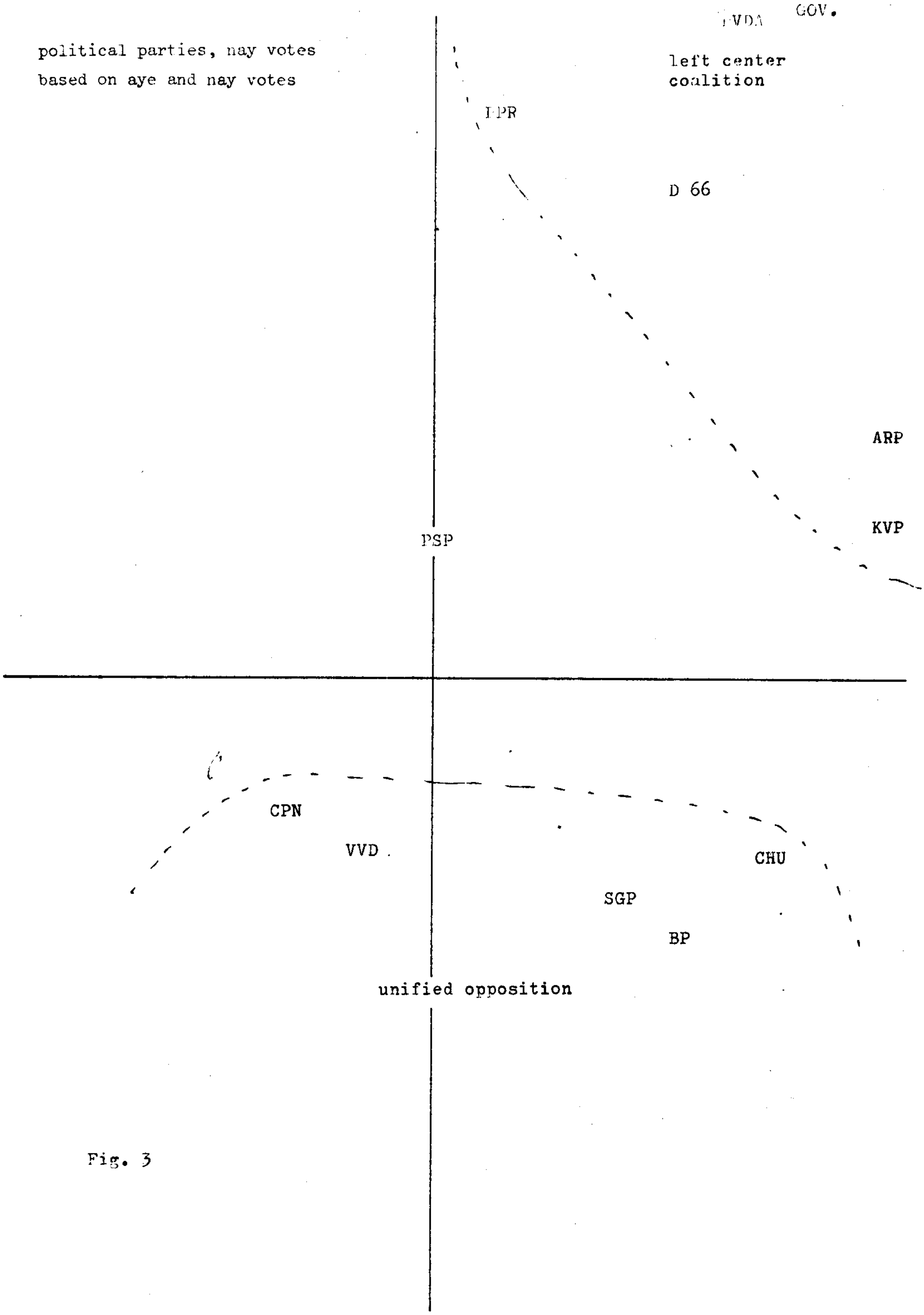

political parties, nay votes
based on aye and nay votes
PVDA
GOV.
left center
coalition
PPR
D 66
ARP
KVP
PSP
CPN
VVD
CHU
SGP
BP
unified opposition

Fig. 3

9.3 Japanese Religion

Source:Meiko Sugiyama (1974) Religious Behaviour of the Japanese.
US-Japan seminar on multidimensional scaling San Diego.

The data consist of 4243 responses on 6 binary questions about religious behaviour.The possible answers are yes or no.

List of questions:

1) Do you make it a rule to practice religious conduct,such as attending religious services,religious worship and missionary works and do you occasionally offer prayers or chant sutras?

2) Do you visit a grave once or twice a year?

3) Do you occasionally read religious books such as the Bible or the Buddhist Scriptures?

4) Do you visit shrines and temples to pray for business prosperity,succes in an entrance examination and so forth?

5) Do you keep a talisman,such as an amulet,charm or mascot near you?

6) Did you draw a fortune,consult a diviner or had you your fortune told within the last years?

The distribution of all possible response patterns:

binary code number	questions 1 2 3 4 5 6	marginal frequency
00	1 1 1 1 1 1	042
01	1 1 1 1 1 0	033
02	1 1 1 1 0 1	006
03	1 1 1 1 0 0	017
04	1 1 1 0 1 1	012
05	1 1 1 0 1 0	029
06	1 1 1 0 0 1	008
07	1 1 1 0 0 0	082
08	1 1 0 1 1 1	051
09	1 1 0 1 1 0	069
10	1 1 0 1 0 1	020
11	1 1 0 1 0 0	054
12	1 1 0 0 1 1	034
13	1 1 0 0 1 0	124
14	1 1 0 0 0 1	027
15	1 1 0 0 0 0	317
16	1 0 1 1 1 1	001
17	1 0 1 1 1 0	002

binary code number	questions 1 2 3 4 5 6	marginal frequencies
18	1 0 1 1 0 1	000
19	1 0 1 1 0 0	009
20	1 0 1 0 1 1	001
21	1 0 1 0 1 0	011
22	1 0 1 0 0 1	007
23	1 0 1 0 0 0	059
24	1 0 0 1 1 1	008
25	1 0 0 1 1 0	023
26	1 0 0 1 0 1	007
27	1 0 0 1 0 0	035
28	1 0 0 0 1 1	010
29	1 0 0 0 1 0	055
30	1 0 0 0 0 1	013
31	1 0 0 0 0 0	194
32	0 1 1 1 1 1	011
33	0 1 1 1 1 0	007
34	0 1 1 1 0 1	002
35	0 1 1 1 0 0	005
36	0 1 1 0 1 1	004
37	0 1 1 0 1 0	008
38	0 1 1 0 0 1	004
39	0 1 1 0 0 0	044
40	0 1 0 1 1 1	072
41	0 1 0 1 1 0	126
42	0 1 0 1 0 1	045
43	0 1 0 1 0 0	142
44	0 1 0 0 1 1	080
45	0 1 0 0 1 0	258
46	0 1 0 0 0 1	137
47	0 1 0 0 0 0	760
48	0 0 1 1 1 1	000
49	0 0 1 1 1 0	002
50	0 0 1 1 0 1	000
51	0 0 1 1 0 0	004
52	0 0 1 0 1 1	004
53	0 0 1 0 1 0	003
54	0 0 1 0 0 1	006
55	0 0 1 0 0 0	030
56	0 0 0 1 1 1	033
57	0 0 0 1 1 0	048
58	0 0 0 1 0 1	038
59	0 0 0 1 0 0	064
60	0 0 0 0 1 1	042
61	0 0 0 0 1 0	096
62	0 0 0 0 0 1	090
63	0 0 0 0 0 0	718

We did a Homals-1 analysis on the 4243 x 6 datamatrix to find some homogeneous sets of response patterns and equivalently some homogeneous sets of response categories.We talk about response patterns in stead of observations because observations with the same pattern have the same quantification and nearly all possible patterns are in the dataset.

By inspecting the resulting response pattern plot we can place a, generally not rectangular, grid on the configuration in such a way that we can order the patterns according to two cumulative scales. One scale is for utilitarian faith,questions 4,5,6,and the other scale is for religious faith,questions 1,3. See fig. 1 .

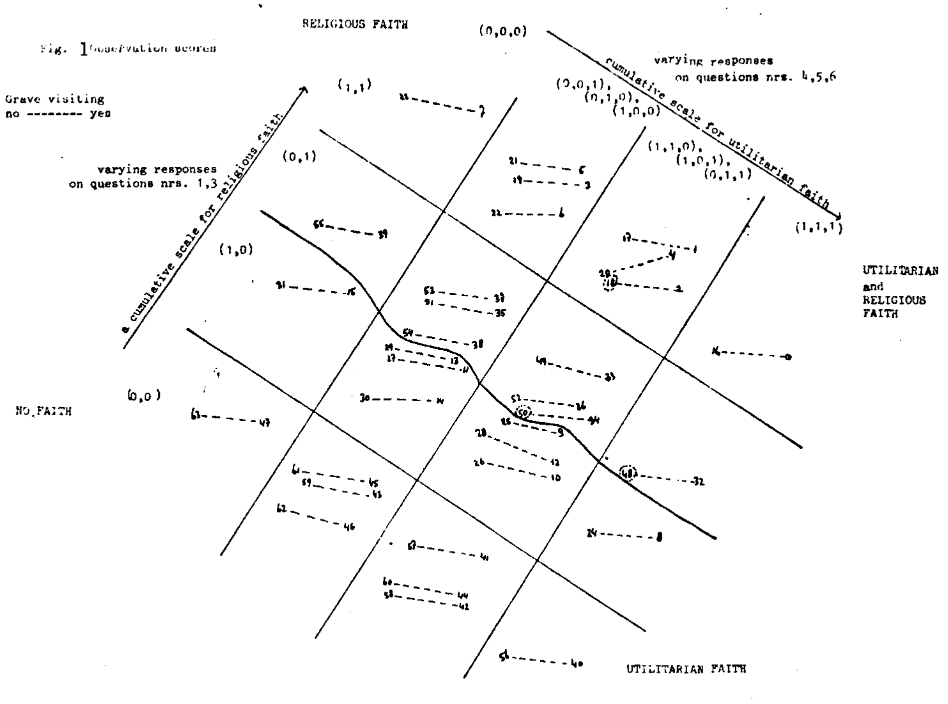
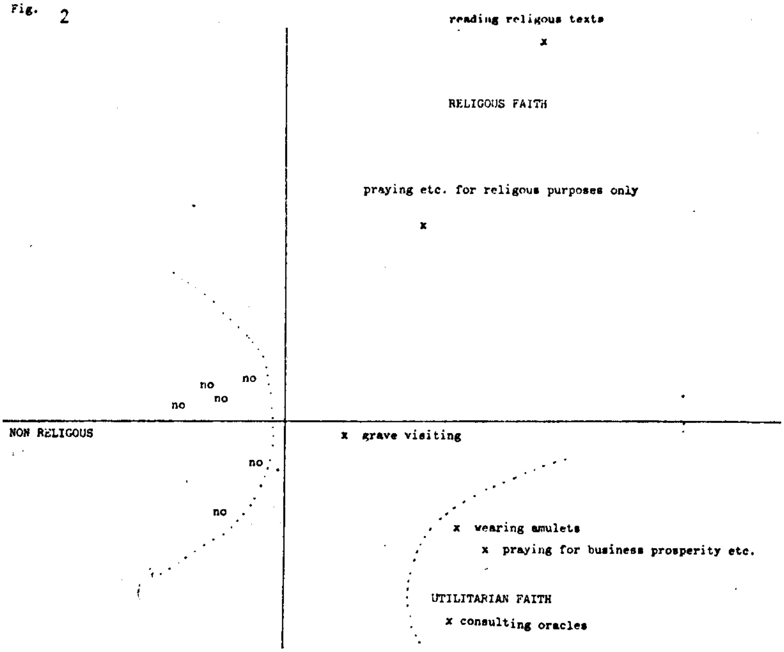


Fig. 1 observation scores

The 64 response patterns are divided into two groups of 32 each by the variable grave-visiting.This division is remarkably independent in respect to the cumulative scales.i.e. the difference in grave-visiting exists for every partition in the grid.Patterns within a partition who only differ in grave-visiting are connected by dotted lines.

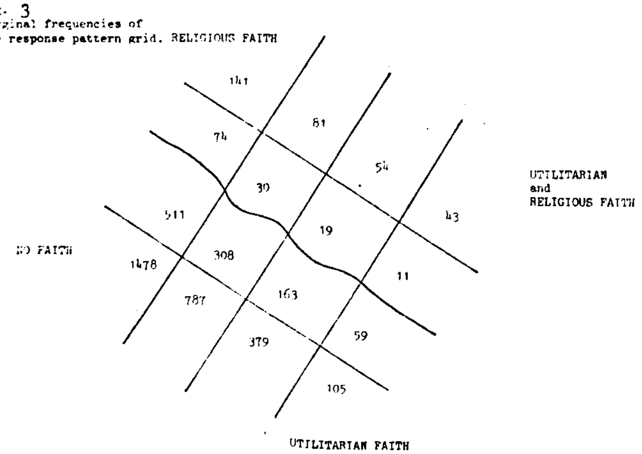
The two scales,being more or less orthogonal,imply a classification into four types of faith:religious ,utilitarian,religious and utilitarian,and no faith.Three of them are also recovered in the category scores plot.See fig. 2 .

Fig. 2



The fourth type,utilitarian and religious faith,is the weakest classification because of its low marginal frequencies.This is the reason that it is not in the category scores plot. See fig. 3 .

Fig. 3
Marginal frequencies of
the response pattern grid.



Additional remarks:

a Only because nearly all possible response patterns are represented ted by the data and because the three homogeneous groups are that homogeneous it is possible to apply the grid.

b The quantification of the three missing response patterns,nrs.18,48 and 50,is obtained by adding per dimension the category quantifications of a pattern.This means that new observations can always be scaled into an already existing Homals solution.

c The execution times for normalizations one and two is respectively 70 and 36 seconds.

9.4 Dentition of Mammals

Source: Palmer, E.L. (1957) Fieldbook of Mammals. Dutton, N.Y., cited in: Hartigan J.A. (1975)Clustering in algorithms, Whiley, N.Y.

The idea is to classify mammals by means of their dentition. There are four groups of teeth, incisors, canines, premolars, and molars; all four divided over the upperjaw (top) and lower jaw (bottom).

List of teeth and their possible frequencies for one mammal:

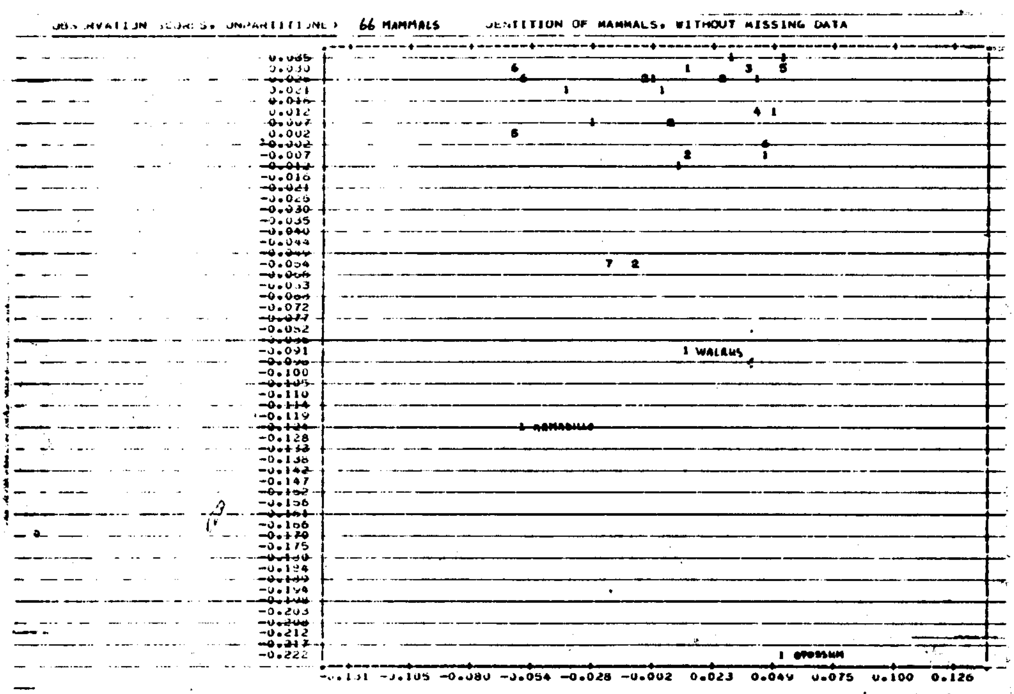
1	top incisors	0 1 2 3 5
2	bottom incisors	0 1 2 3 4
3	top canines	0 1
4	bottom canines	0 1
5	top premolars	0 1 2 3 4
6	bottom premolars	0 1 2 3 4
7	top molars	0 1 2 3 4 8
8	bottom molars	0 1 2 3 4 8

List of mammals and their set of teeth:

opossum	54113344
hairy tale mole	33114433
common mole	32103333
star nose mole	33114433
brown bat	23113333
silver hair bat	23112333
pigmy bat	23112233
house bat	23111233
red bat	13112233
hoary bat	13112233
lump nose bat	23112333
armadillo	00000088
pika	21002233
snowshoe rabbit	21003233
beaver	11002133
marmot	11002133
groundhog	11002133
prairy dog	11002133
ground squirrel	11002133
chipmunk	11002133
gray squirrel	11001133
fox squirrel	11001133

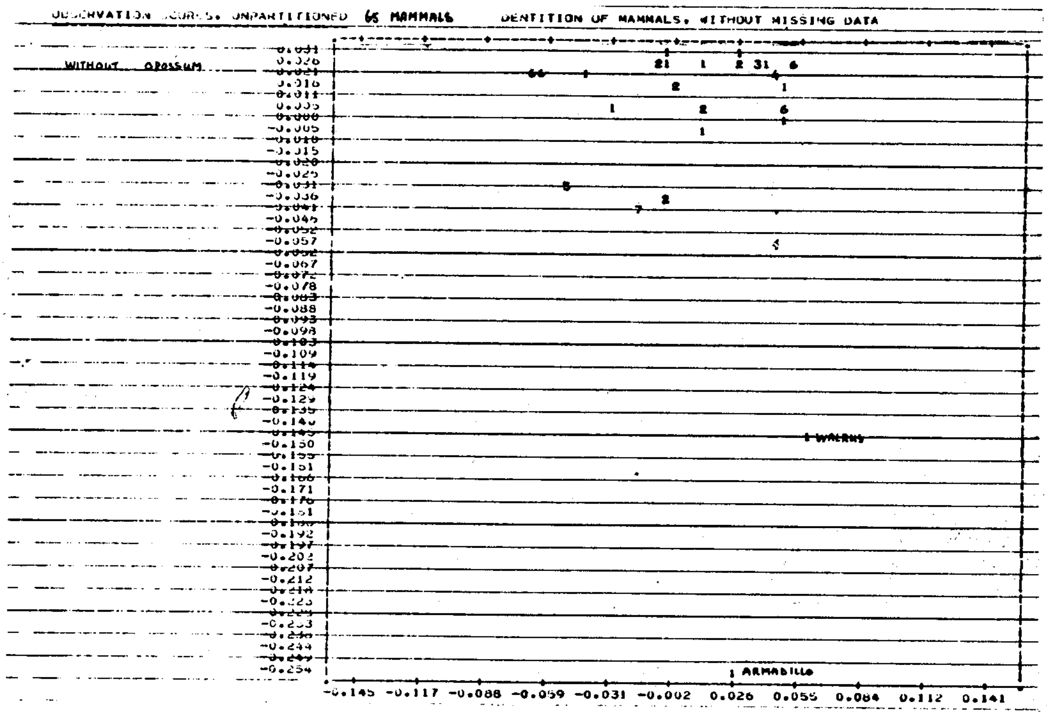
pocket gopher	11001133
kangeroo rat	11001133
pack rat	11000033
field mouse	11000033
muskrat	11000033
black rat	11000033
house mouse	11000033
porcupine	11001133
guinea pig	11001133
coyote	13114433
wolf	33114423
fox	33114423
bear	33114423
civet cat	33114422
raccoon	33114432
marten	33114412
fisher	33114412
weasel	33113312
mink	33113312
ferrer	33113312
wolverine	33114412
badger	33113312
skunk	33113312
river otter	33114312
sea otter	32113312
jaguar	33113211
ocelot	33113211
cougar	33113211
lynx	33113211
fur seal	32114411
sea lion	32114411
walrus	10113300
grey seal	32113322
elephant seal	21114411
peccary	23113333
elk	04103333
deer	04003333
moose	04003333
reindeer	04103333
antelope	04003333
bison	04003333
mountain goat	04003333
muskox	04003333
mountain sheep	04003333

In the first HOMALS run of 66 mammals we see one extreme outlier in the observations plot: the opposum. This is the only mammal in the set with four topmolars, four bottom-molars and five top incisors. This combination is unique in this set. See plot 1.



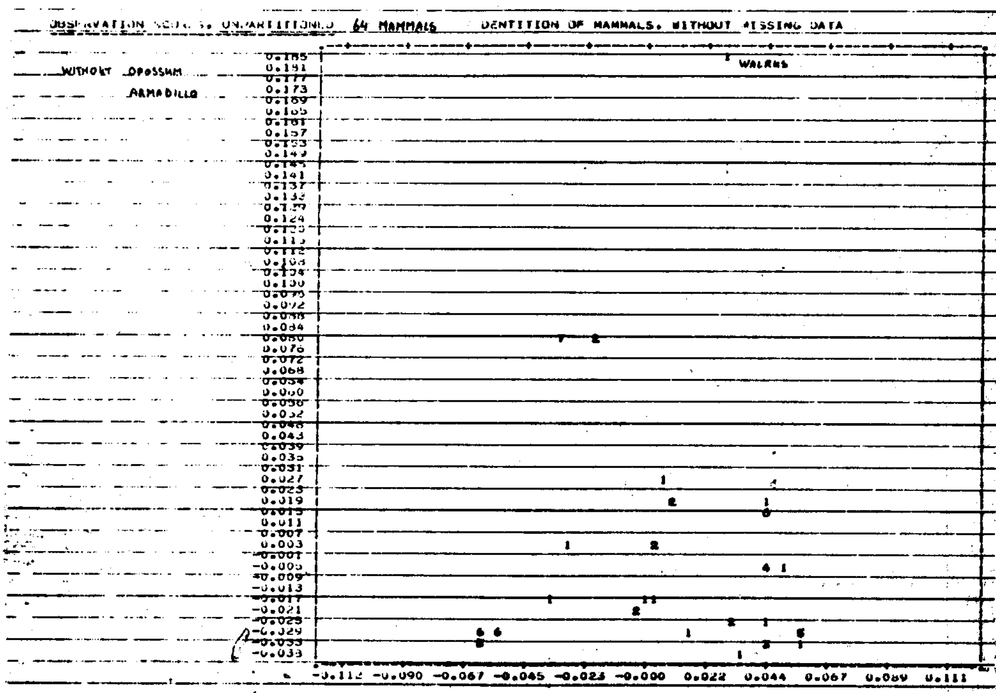
plot 1

In the second HOMALS run we removed the opossum and analysed the remaining 65 mammals, because the opossum possibly could have dominated the solution. As we see in plot 2 the whole configuration didn't change much and we found a new outlier, the armadillo, an animal with a most peculiar set of teeth; it has no incisors, no canines nor premolars, but eight molars in each jaw.



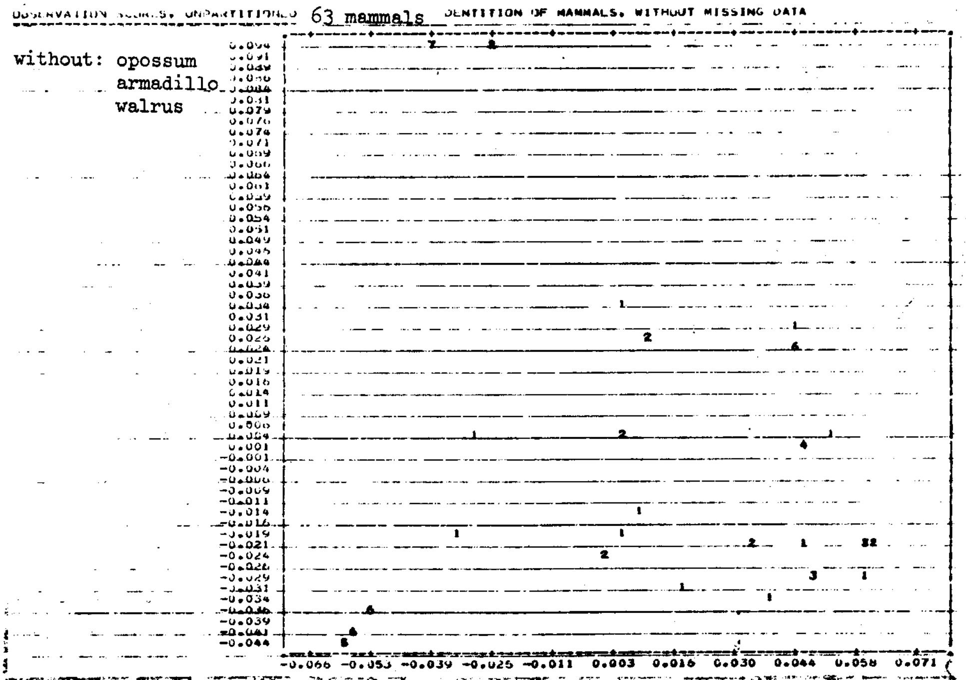
plot 2

Again we removed the outlier, the armadillo, from the dataset and analysed the 64 remaining mammals and found a new outlier, the walrus. See plot 3.



plot 3

The configuration of the other 63 mammals remained stable but the whole picture is rotated along the horizontal X-axis, a freedom of rotation that exists for all HOMALS solutions.
Now we come to the last run without the three outliers. The conclusion is that outliers don't play an important role in finding a stable configuration for this dataset, because the configuration of the 63 mammals does not differ much wether there are outliers in the dataset or not. If one wants to dig out the difference within the 63 mammals one can enlarge the scale of the plot by leaving out outliers. We stopped with removing outliers from the dataset when they became sets of outliers and hence no outliers anymore. But one can go on removing as long as there is a conceptual reason for doing so.
Plot 4 shows the results of the last analysis.



plot 4

Labeling the mammals in plot 4 according to their number of teeth of a certain group shows us some nicely separated regions wherein mammals with the same dentition are situated.
Plots 5 and 6 illustrate this labeling for top canines (plot 5) and for bottom premolars (plot 6).
The centroids of mammals with the same number of bottompremolars are the computed category scores for bottom premolars. Those category scores are indicated with an asterix in the plots 5 and 6 and separately plotted in the plots 7 and 8.
It is clear that classification of mammals by means of their dentition is rather straightforward if one uses these kinds of plots.

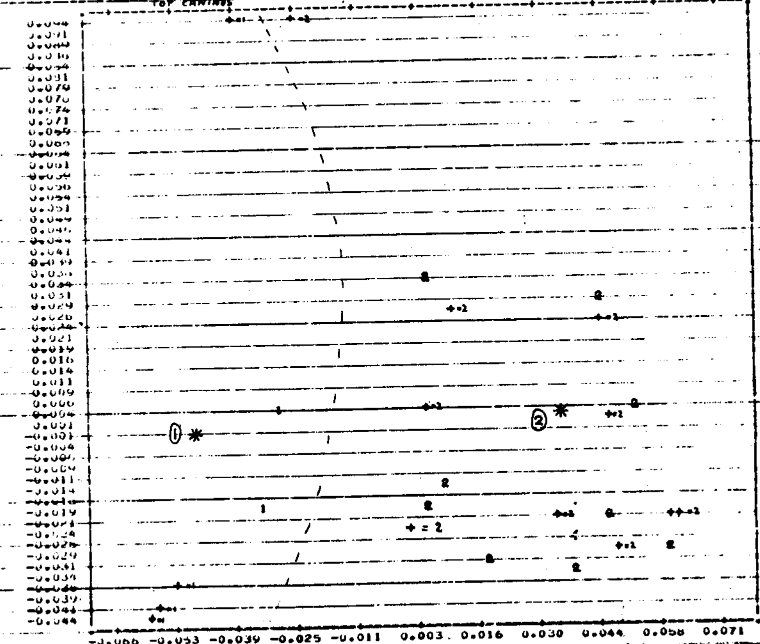
OBSERVATION SCORES PARTITIONED BY VAR. 3 DENTITION OF MAMMALS, WITHOUT MISSING DATA

TOP CANINES

* CATEGORY SCORES

CODEBOOK

REAL NRS	CODES
0	1
1	2

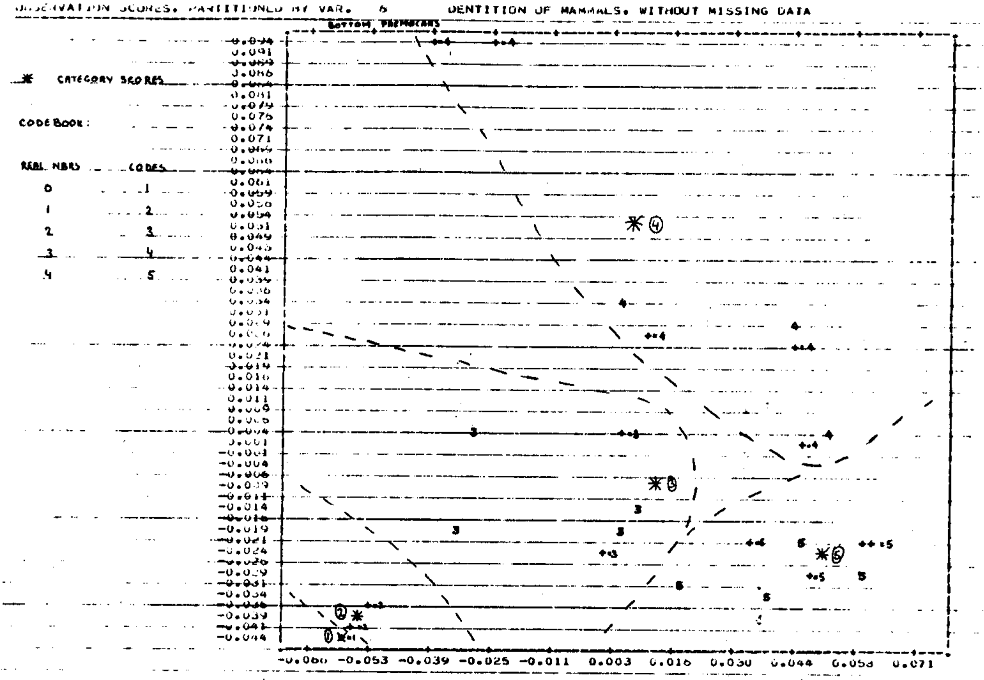


SUMMARY OF ALL CELLS (X,Y), MARKED : + IN THE PLOT, CONTAINING MORE THAN 1 POINT IDENTIFICATION

Y	X	NUMBER OF POINTS	POINT-IDENTIFICATION
[illegible]	[illegible]	7	1111111
[illegible]	-0.02[illegible]	2	22
[illegible]	0.013	2	22
[illegible]	0.044	6	222222
[illegible]	0.005	2	22
[illegible]	0.040	4	2222
-0.0[illegible]	0.054	2	22
-0.0[illegible]	0.060	3	222
-0.021	0.052	2	22
-0.024	[illegible]	2	22
-0.029	0.043	3	222
-0.036	-0.053	[illegible]	111111
-0.041	-0.057	6	111111
-0.044	-0.0[illegible]	5	11111

Plot 5

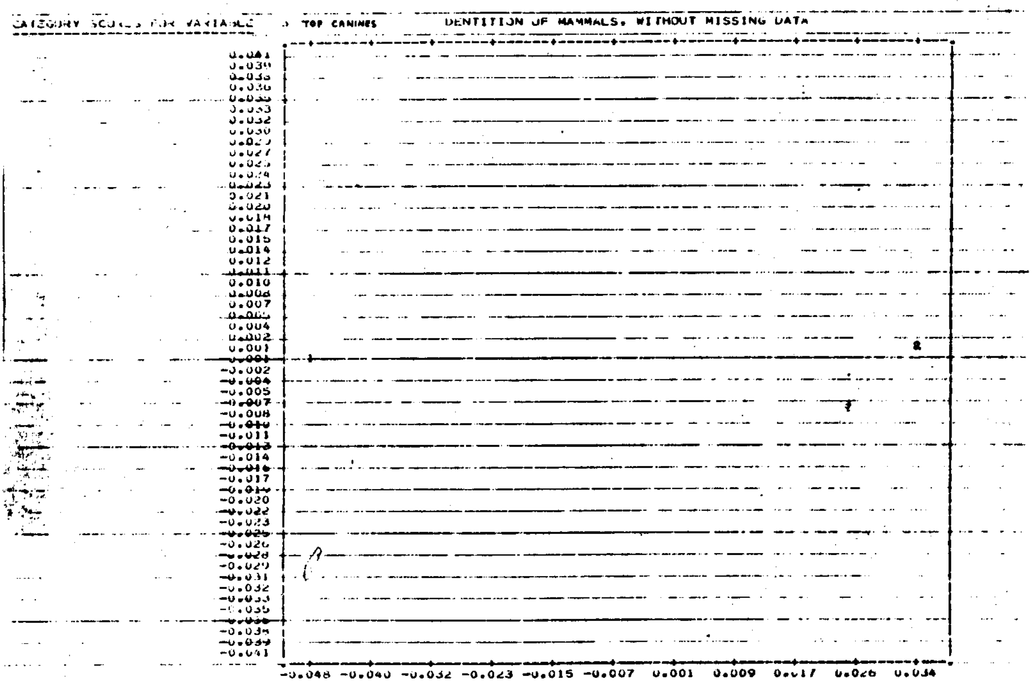
UBSERVATION SCORES, PARTITIONED BY VAR. 6 DENTITION OF MAMMALS, WITHOUT MISSING DATA



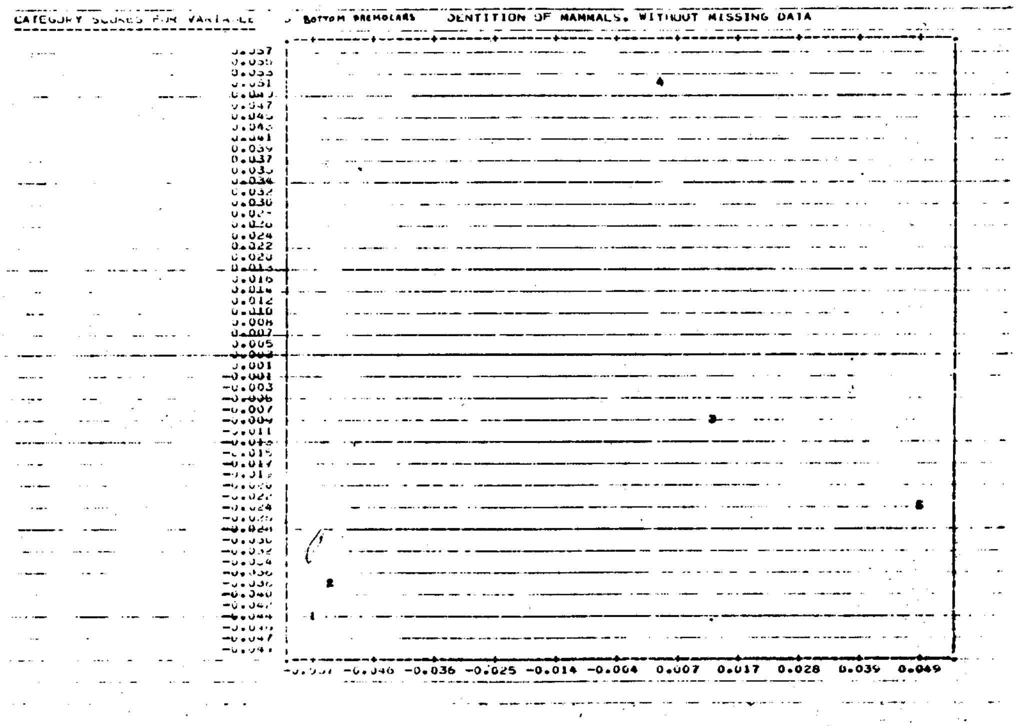
SUMMARY OF ALL CELLS (X,Y), MARKED : + IN THE PLOT, CONTAINING MORE THAN 1 POINT IDENTIFICATION

Y	X	NUMBER OF POINTS	POINT-IDENTIFICATION
0.094	-0.039	7	4444444
0.094	-0.025	2	44
0.026	0.010	2	44
0.024	0.044	5	44444
0.004	0.053	2	44
0.001	0.046	4	3333
-0.021	0.034	2	55
-0.021	0.060	3	555
-0.021	0.062	2	55
-0.024	0.001	2	33
-0.029	0.048	3	555
-0.036	-0.053	6	222222
-0.041	-0.057	6	222222
-0.044	-0.059	5	11111

plot 6


CATEGORY SCORES FOR VARIABLE 3 TOP CANINES
DENTITION OF MAMMALS, WITHOUT MISSING DATA
-0.048 -0.040 -0.032 -0.023 -0.015 -0.007 0.001 0.009 0.017 0.026 0.034

plot 7


CATEGORY SCORES FOR VARIABLE
BOTTOM PREMOLARS
DENTITION OF MAMMALS, WITHOUT MISSING DATA

plot 8

THE OBSERVATION SCORES ARE: without opossum
armadillo
walrus

DIMENSIONS

OBSERVATIONS	1	2
~~1~~	~~0.0362~~	~~-0.0240~~
2	0.0056	0.0347
~~3~~	~~0.0352~~	~~-0.0210~~
4	0.0110	0.0264
~~5~~	~~0.0662~~	~~-0.0647~~
6	0.0087	-0.0152
~~7~~	~~[illegible]~~	~~[illegible]~~
8	0.0010	-0.0249
~~9~~	~~0.0010~~	~~-0.0242~~
10	0.0052	0.0047
11	-0.0328	~~-0.0177~~
12	-0.0230	0.0040
~~13~~	~~-0.0516~~	~~-0.0366~~
14	-0.0515	-0.0366
~~15~~	~~-0.0515~~	~~-0.0366~~
16	-0.0515	-0.0366
~~17~~	~~-0.0515~~	~~-0.0366~~
18	-0.0515	-0.0366
~~19~~	~~-0.0546~~	~~-0.0404~~
20	-0.0546	-0.0404
~~21~~	~~-0.0546~~	~~-0.0404~~
22	-0.0546	-0.0404
~~23~~	~~-0.0570~~	~~-0.0440~~
24	-0.0570	-0.0440
~~25~~	~~-0.0570~~	~~-0.0440~~
26	-0.0570	-0.0440
~~27~~	~~-0.0570~~	~~-0.0440~~
28	-0.0546	-0.0404
~~29~~	~~-0.0546~~	~~-0.0404~~
30	0.0201	-0.0309
~~31~~	~~0.0436~~	~~-0.0281~~
32	0.0436	-0.0281
~~33~~	~~0.0436~~	~~-0.0281~~
34	0.0610	-0.0275
~~35~~	~~0.0476~~	~~-0.0204~~
36	0.0606	-0.0216
~~37~~	~~0.0606~~	~~-0.0216~~
38	0.0439	0.0250
~~39~~	~~0.0439~~	~~0.0250~~
40	0.0439	0.0250
~~41~~	~~0.0606~~	~~-0.0216~~
42	0.0439	0.0250
~~43~~	~~0.0439~~	~~0.0250~~
44	0.0520	0.0027
~~45~~	~~0.0464~~	~~-0.0292~~
46	0.0466	0.0011
~~47~~	~~0.0466~~	~~0.0011~~
48	0.0466	0.0011
~~49~~	~~0.0466~~	~~0.0011~~
50	0.0621	-0.0224
~~51~~	~~0.0621~~	~~-0.0224~~
52	0.0455	0.0234
53	0.0397	-0.0344
~~54~~	~~0.0110~~	~~0.0264~~
55	-0.0241	0.0936
~~56~~	~~-0.0379~~	~~0.0939~~
57	-0.0379	0.0939
~~58~~	~~-0.0241~~	~~0.0936~~
59	-0.0379	0.0939
~~60~~	~~-0.0379~~	~~0.0939~~
61	-0.0379	0.0939
~~62~~	~~-0.0379~~	~~0.0939~~
63	-0.0379	0.0939

```
H O M A L S - 1                    VERSION 3.01 (T.B.M.)              TOM VISSER
DECEMBER 1977                                                         DEPARTMENT W.I.
                                                                      CENTRAAL REKENINSTITUUT

                                                                      JAN VAN RIJCKEVORSEL      JAN DE LEEUW
                                                                      DEPARTMENT DATATHEORY
                                                                      UNIVERSITY OF LEIDEN

                                                                      WASSENAARSEWEG 80
                                                                      LEIDEN
                                                                      HOLLAND
=========
(JOB NR. 1)
=========

MEMORY REQUIRED:  128K

INPUT-DATA SPECIFICATIONS:
--------------------------

NAME - DENTITION OF MAMMALS, WITHOUT MISSING DATA

                       = PROBLEM PARAMETERS =

NOBS - NUMBER OF OBSERVATIONS                                         63
NVAT - TOTAL NUMBER OF VARIABLES IN THE DATAMATRIX                     8
NVAR - NUMBER OF ANALYSIS-VARIABLES                                    8
NDIM - NUMBER OF DIMENSIONS                                            2
MAXC - MAXIMUM NUMBER OF CATEGORIES OVER ALL VARIABLES                 5
ISUC - TOTAL NUMBER OF CATEGORIES OVER THE                            28
       ANALYSIS-VARIABLES
NDPA - PARAMETER FOR THE ANALYSIS-PARTITION                            0
       =0  : NO ANALYSIS-PARTITION
       =I>0: VARIABLE I IS PARTITIONING

                       = ANALYSIS PARAMETERS =

NORM - METHOD OF ANALYSIS                                              1
       (IF NOBS >> ISUC YOU SHOULD USE NORM=1; IF NOBS << ISUC
       NORM=2; NORM=0: CHOISE IS MADE INTERNALLY)
IMIS - PARAMETER FOR MISSING DATA                                      1
       = 0 : THE DATAMATRIX POSSIBLY CONTAINS MISSING DATA
       = 1 : THE DATAMATRIX DOES NOT CONTAIN MISSING DATA
MAXI - MAXIMUM NUMBER OF ITERATIONS                                  200
EPSI - CONVERGENCE CRITERIUM                                    0.60E-06

                       = INPUT/OUTPUT PARAMETERS =

INPD - UNIT NUMBER OF THE DATAMATRIX                                   8
IDAT - PARAMETER INDICATING WHETHER OR NOT (1 AND 0 RESP.)             1
       TO PRINT THE DATAMATRIX
IXCH - PARAMETER INDICATING WHETHER OR NOT TO PRINT CROSS-             1
       TABS, CHI-SQUARES AND DEGREES OF FREEDOM
          =0: NO PRINT
          =1: PRINT OF CHI-SQUARES AND DEGREES OF FREEDOM
          =2: PRINT ALL
IWRI - PARAMETER INDICATING WHETHER OR NOT TO PRINT                    2
       OBSERVATION- AND CATEGORY SCORES
          =0: NO PRINT
          =1: OBSERVATION SCORES ONLY
          =2: BOTH
          =3: CATEGORY SCORES ONLY
IPLO - PARAMETER INDICATING WHETHER OR NOT AND HOW TO                  2
       PLOT OBSERVATION AND CATEGORY SCORES
          =0: NO PLOT
          =1: OBSERVATION SCORES, UNPARTITIONED AND PARTITIONED BY
              THE NON-ANALYSIS VARIABLES ONLY
          =2: PLOT PARTITIONED ACCORDING TO THE USERS REQUEST
IOUT - UNIT NUMBER FOR OUTPUT OF THE OBSERVATION SCORES TO             0
       CARD, TAPE OR DISK (=0: NO SUCH OUTPUT REQUIRED)

ICAT(J) - NUMBER OF CATEGORIES PER VARIABLE:

            NUMBER OF  |             NUMBER OF  |             NUMBER OF  |             NUMBER OF
 VARIABLE   CATEGORIES |  VARIABLE   CATEGORIES |  VARIABLE   CATEGORIES |  VARIABLE   CATEGORIES
------------------------------------------------------------------------------------------------
     1          4      |      3          2      |      5          5      |      7          3
     2          4      |      4          2      |      6          5      |      8          3

THE CATEGORY SCORES OF THE FOLLOWING ANALYSIS-VARIABLES ARE PLOTTED, AND, IF LABELED WITH A STAR,
THEY ACT AS PARTITIONING VARIABLES IN THE OBSERVATION SCORES PLOT:

        1 *
        2 *
        3 *
        4 *
        5 *
        6 *
        7 *
        8 *

FMT1 - FORMAT TO READ THE DATAMATRIX: (5X,8I2)
```

9.5 FUNCTIONAL LEARNING

Source: De Klerk, L. W. F., De Leeuw, J., Oppe, S. (1970), Functional Learning, investigations with real valued functions. Report E 024-70, Psychological Institute, University of Leiden, The Netherlands.

To create this dataset a list of S-R pairs was presented to a subject by the experimenters; one pair at the time. After the presentation of one S-R pair only one stimulus item was presented for a short interval of time. When the subject responded to the presentation according to an association rule based on the preceding S-R pairs, his response R' was recorded and a new pair was presented. All stimulus items were presented once.

Let S stand for one continuum and R for the second. Let s_i be an array of points on the S continuum and r_i be the set of points into the set R. Then the items were construcued in such a way that there is a mathematical function that defines the mapping of the points s_i into the set R. See fig. 1.

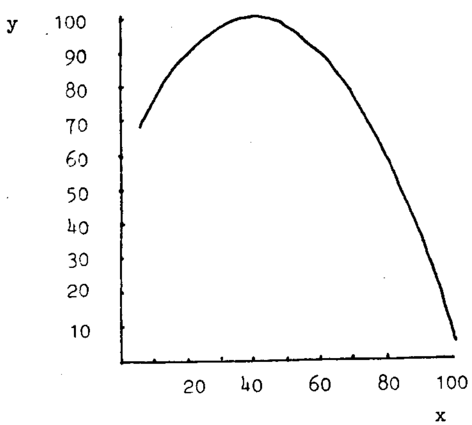


Fig. 1 The graphical representation of the parabola: $y = -\frac{19}{720}x^2 + \frac{19}{9}x + \frac{520}{9}$

Due to the associative character of this type of items the subjects are supposed to learn the specific assignment of R to S. This assumption is tested by inserting new stimuli. When the subjects have learned the function R = f (S), then the responses, 'R', to the interpolated (or extrapolated) stimuli are completely predictable.

The set S is divided into 20 equally spaced intervals represented by lines varying from 5 to 100 mm.. The subjects responses, also expressed as lines, are also recorded in mm..

To analyse the data the responses are discretized into sets with 5,10 and 20 equally spaced intervals for every subject. There are 57 subjects in the experiment and 10 runs of each 20 learn and reproduction trials. See table 2. Only the first and the last run are analysed, because in the first run the subjects are supposed to reproduce the function R = f (S) worse than in the last run. And this difference we want to mark. The idea is to recover the individual reference curves and the optimal reference curve for all subjects for each run. A reference curve describes the functional relation between the stimuli and the subject's responses.

Per run and for every discretization we did a Homals-1 analysis on the transposed datamatrix (i.e. a 20 by 57 matrix). In the Homals terminology there are observation scores for stimuli and category scores for subjects, one for every interval the subject scored in.

The second axis has the lowest stress in all cases, which is equivalent to the dominant eigenvalue. This is the reason only this second axis is used.

Table 1 The stresses for both axes

Discretization level	5		10		20	
Axis number	1	2	1	2	1	2
Run 1	.67	.51	.46	.38	.31	.26
Run 10	.54	.29	.27	.15	.16	.09

The quantifications on the second axis of all three discretizations are plotted against the stimuli lenghts. This resulted in three optimal reference curves per run. See fig. 2 and 3.

.The stresses of run nr.10 are much lower than those of run nr.1,which means that the subjects responded more uniformly in the last run but not necessarily better,although the curve of run nr.10 follows f (S) better than the first run's curve. See again fig. 2 and 3.

A quantification of an interval of the optimal reference curve is the centroid of quantifications of the corresponding interval of all individual reference curves.

Fig. 2 The optimal reference curves for run nr. 10

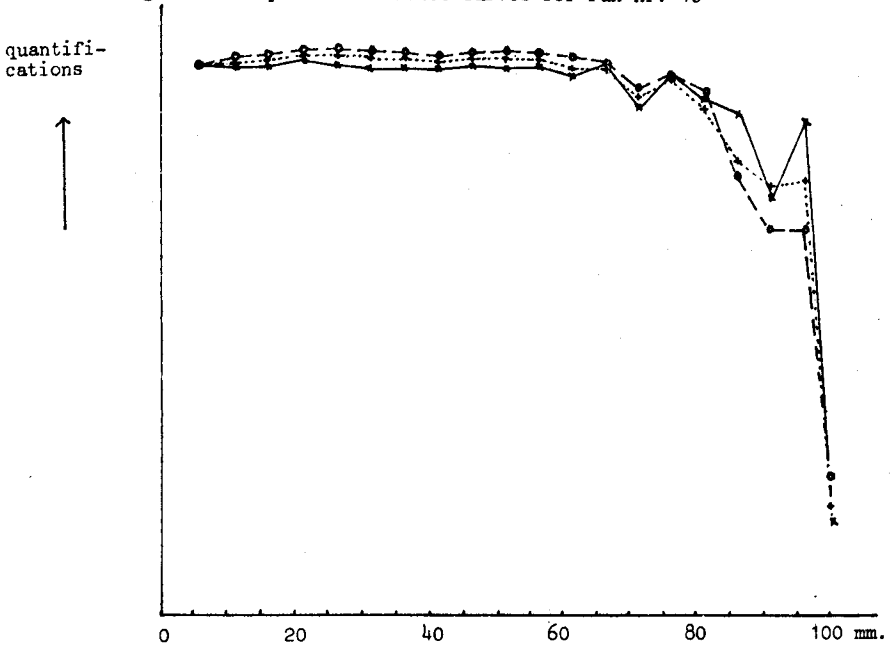
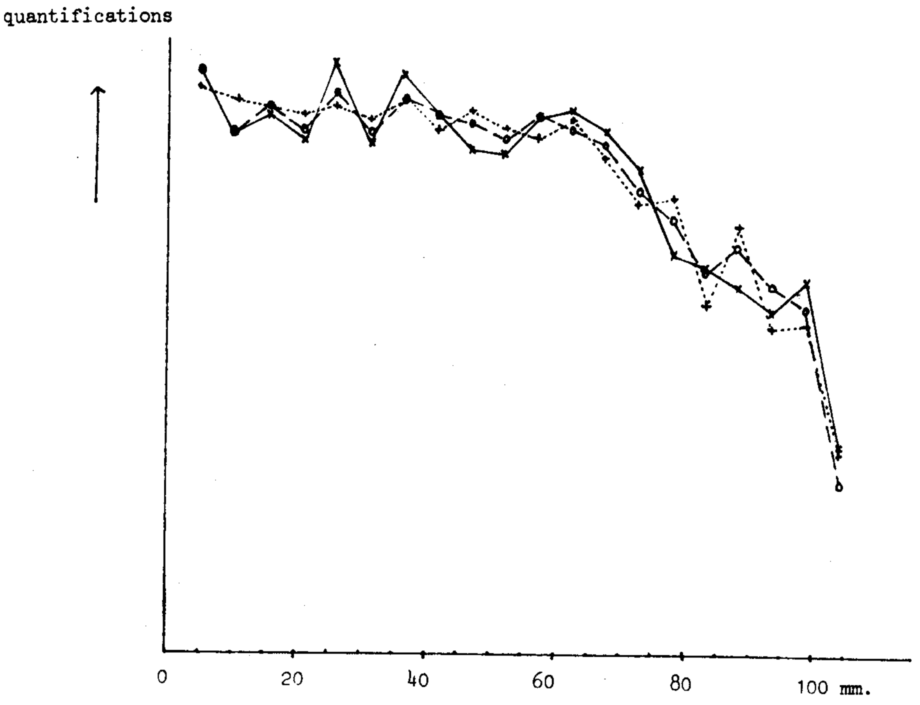


Fig. 3 The optimal reference curves for run nr. 1



Compared with a classical principal component analysis on the rank orders (De Klerk a.o. 1970) the Homals solutions fit all kinds of functions used by the subjects better because the quantification of the used intervals is done for every subject according to the particular function a subject has in mind,while in pca it is done for all subjects simultaneously. For instance subjects who use a straight line are not in discordance with parabola users in a Homals context, though they are in a pca solution.

Plots of quantifications of the used intervals of a subject against the ordered stimuli show the individual reference curves. See fig. 4 and 5. Figure 4 shows the curve of subject nr. 23, who has an extreme good fit ($D^2 = 32$) in terms of the squared average deviation ($=D^2$) between his own curve and the parabola of figure 1. The other curve,in figure 5,belongs to subject nr. 24, who has a bad fit ($D^2 = 1188$).

The different lines within each plot depict the different discretizations. Both the individual and general curves are much alike in respect to the three discretizations, although the stresses differ widely. The finest discretization of 20 intervals has the lowest stress in all cases within one run. But even the widest intervals s still reflect the reference curve adequately enough.In table 3 are the meanings of the three different lines per plot.

Table 3

——x——x——x——	the 20 intervals line
--o-----o-----o---	the 10 intervals line
··+·····+·····+···	the 5 intervals line

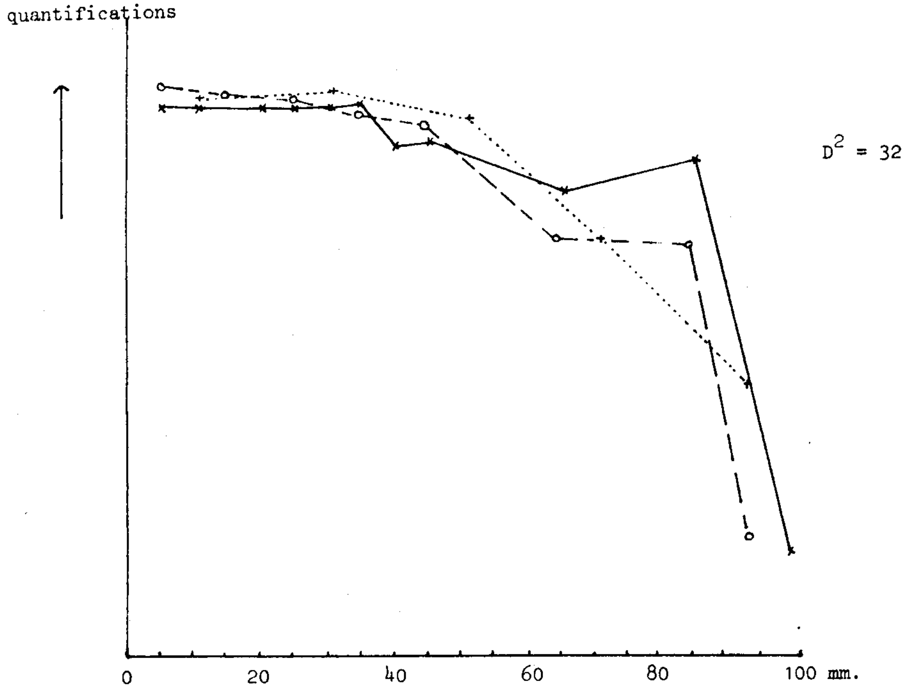


Fig. 4 interval quantifications vs. stimuli. subject nr. 23 . run 10

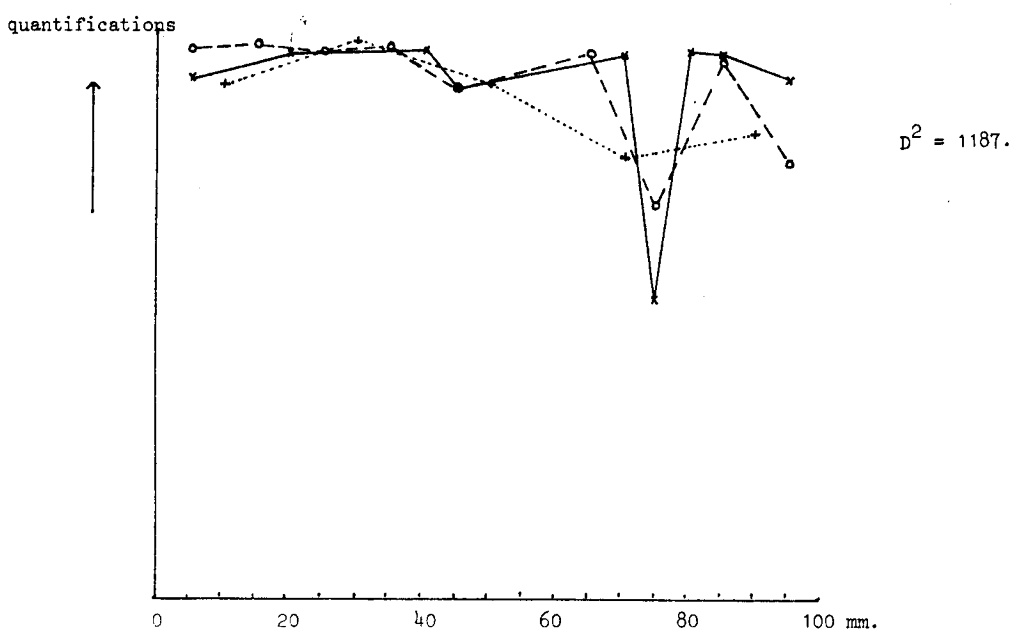


Fig. 5 interval quantifications vs. stimuli. subject nr 24. run 10

Table 1 The responses of 57 subjects (row-entries) to 20 stimuli (collumn-entries) The stimuli are lines varying from 5 to 100 mm. The responses are also lines varying from 0 to 100 mm.

RUN 1

5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
95	84	98	86	74	85	48	92	67	68	80	50	52	40	43	70	8	22	85	88
84	93	73	64	25	71	88	83	58	76	55	42	41	34	60	7	97	29	63	7
55	69	39	18	100	84	54	55	25	74	19	34	70	66	44	50	80	11	86	6
85	80	85	77	57	70	65	48	80	90	66	91	78	41	60	35	45	29	21	18
70	83	87	76	75	62	77	84	72	67	68	76	73	54	87	9	55	29	64	13
77	74	70	58	74	79	64	42	43	66	76	76	57	48	83	88	13	18	99	85
100	94	92	77	70	79	68	63	70	59	72	78	79	52	53	43	30	29	9	18
63	95	89	100	91	91	55	35	93	87	93	75	87	20	78	45	32	31	32	18
89	86	73	87	56	63	74	55	55	87	65	76	75	80	61	18	50	74	68	14
59	76	84	90	91	72	61	27	70	33	72	41	86	17	84	23	40	67	27	8
88	93	88	88	90	84	95	85	80	90	81	85	90	90	20	15	16	9	18	10
79	83	73	80	91	82	90	94	45	100	82	89	29	20	19	30	28	28	20	27
51	74	58	81	55	63	79	77	71	39	25	54	78	34	45	37	60	35	33	33
92	86	40	42	68	76	43	52	51	75	64	72	56	6	27	60	78	9	34	5
91	63	58	82	76	71	.79	75	59	85	55	43	16	15	26	62	33	33	51	6
93	94	60	71	78	75	42	98	24	97	63	87	54	94	97	11	34	52	19	71
4	63	61	69	48	66	32	75	54	74	45	47	71	70	21	87	60	80	63	65
81	87	73	7	15	59	58	64	25	81	75	26	19	43	63	53	46	47	51	65
93	87	87	80	80	84	80	81	71	91	59	51	55	60	82	64	95	60	22	29
77	65	68	52	10	52	97	62	85	26	84	53	100	60	38	87	24	21	30	40
100	100	87	75	100	100	100	100	89	100	100	84	32	100	22	56	45	38	40	25
91	76	76	86	57	85	70	90	88	55	65	55	76	53	37	11	85	42	17	31
100	85	89	70	74	63	65	51	91	62	65	77	72	55	31	36	19	21	8	6
20	98	73	35	58	51	72	21	95	90	44	42	100	40	49	18	44	7	32	7
70	68	62	72	90	88	90	58	71	89	92	96	88	39	57	47	68	44	62	41
97	94	91	85	72	75	84	69	90	79	68	68	70	67	77	.45	38	39	17	12
100	88	75	50	71	98	60	83	32	80	87	79	100	95	90	21	42	78	62	14
71	66	72	93	39	84	68	52	72	44	41	51	58	63	53	67	70	75	71	88.
91	79	32	44	41	57	65	59	24	69	73	64	67	65	27	28	21	41	7	15
93	76	66	66	85	78	67	85	56	63	.36	73	28	67	64	30	24	24	41	5
88	79	94	70	69	80	83	73	65	71	65	49	60	25	27	10	26	13	13	13
82	58	77	63	69	73	80	86	76	94	72	66	68	9	71	57	31	25	23	4
80	50	63	61	67	83	55	60	42	88	78	47	82	57	40	63	61	9	42	28
96	69	73	93	75	69	53	69	64	86	73	84	75	58	56	57	31	62	15	10
100	75	50	47	60	45	60	50	60	67	65	65	85	45	55	90	87	70	70	72
69	74	74	54	69	74	90	91	35	80	73	82	79	62	52	34	57	38	66	22
67	80	70	84	80	74	83	70	52	58	44	51	71	11	33	73	15	55	52	49
100	71	67	88	36	67	85	69	39	59	60	55	60	27	61	57	78	24	28	8
59	72	78	91	70	44	78	49	49	26	76	76	100	54	26	94	54	99	20	16
62	91	48	83	84	10	97	69	58	57	60	67	60	58	35	32	13	41	19	7
91	88	93	66	61	64	65	73	60	70	31	78	39	10	31	35	14	40	28	34
60	64	55	71	73	72	80	72	70	66	80	20	84	48	37	75	40	56	40	42
99	96	91	98	65	55	62	61	61	97	63	75	31	81	88	43	22	42	22	21
60	70	53	75	53	53	65	63	41	61	77	26	26	66	63	25	20	29	20	24
96	85	95	77	55	56	94	35	82	84	64	35	61	80	27	63	44	87	43	12
71	70	70	66	41	76	78	32	54	50	79	36	45	33	31	55	55	8	33	20
98	91	92	84	45	80	47	90	85	35	67	30	81	89	47	28	50	.15	20	95
80	71	67	60	43	50	55	73	57	80	62	59	72	61	73	78	68	84	90	79
84	63	61	64	63	52	66	60	74	71	88	53	56	55	20	55	50	45	65	75
86	82	80	85	75	73	83	28	75	79	85	71	75	76	68	29	12	7	95	37
84	82	83	74	40	54	67	22	82	11	68	44	59	25	48	42	35	62	74	12
84	93	73	64	25	71	48	63	56	76	55	42	41	34	60	7	97	29	63	7
95	84	98	86	74	85	48	92	57	68	80	50	52	40	43	70	8	22	85	88
55	69	39	18	100	84	54	55	25	70	19	34	70	66	44	50	80	11	86	6
98	80	85	100	90	60	85	84	45	76	80	17	69	78	50	37	78	27	32	20
100	93	100	91	95	80	84	68	68	84	75	80	82	73	82	75	17	42	16	3
62	58	73	72	7	55	64	47	52	81	73	85	85	62	58	64	17	72	85	52

Table 1 continued

RUN 10

5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
70	81	79	87	85	79	93	73	51	60	66	76	75	64	31	50	30	19	24	7
80	77	75	68	90	75	66	97	73	70	45	70	48	34	70	55	10	40	45	32
95	89	89	79	59	76	81	100	63	100	60	78	63	45	64	36	19	15	9	6
52	80	75	81	83	75	93	63	75	90	82	90	78	60	65	65	60	55	5	15
62	68	79	79	82	67	91	84	92	96	82	33	93	58	71	42	44	40	45	5
100	90	100	77	100	90	73	95	95	84	94	92	54	92	100	71	99	81	79	7
59	99	69	83	91	67	74	49	54	82	68	88	74	57	68	92	30	19	32	5
73	78	77	73	65	74	71	64	72	78	82	95	17	17	22	32	40	19	33	18
42	85	88	81	51	59	68	52	74	66	92	86	88	54	51	23	51	13	61	10
59	72	65	73	86	100	100	92	85	64	74	79	59	63	69	42	38	29	31	6
58	75	93	93	88	90	95	92	94	95	94	92	62	74	66	73	20	24	55	8
55	91	90	65	96	95	96	100	100	69	99	95	85	87	65	67	22	27	80	5
44	55	58	61	64	57	67	74	59	73	65	68	72	69	77	68	47	46	44	60
99	79	90	97	99	90	100	95	72	85	91	89	73	60	67	28	14	35	20	5
63	85	90	90	95	100	100	55	55	58	60	66	58	76	21	20	10	65	32	5
62	85	82	83	84	82	87	45	100	84	85	72	82	20	16	18	23	13	18	7
56	62	66	58	76	70	73	36	79	70	85	70	77	65	64	45	66	37	20	9
84	87	85	74	82	76	70	77	91	74	63	71	53	55	69	56	66	9	64	14
62	54	64	73	71	64	70	54	57	42	43	49	51	55	32	55	24	22	19	11
58	74	61	88	65	98	100	93	93	81	85	93	75	62	33	50	43	47	45	7
100	73	90	91	91	40	92	84	68	84	81	91	80	85	86	94	63	75	66	4
56	94	91	81	100	90	90	73	63	91	90	77	57	59	66	62	34	39	35	30
66	78	85	94	93	97	100	98	100	100	95	77	71	64	60	58	37	36	18	6
70	60	65	65	66	58	66	73	23	58	85	99	32	30	20	50	100	58	6	28
64	78	57	93	77	78	81	75	63	91	95	78	72	43	68	39	68	47	43	3
63	60	66	68	99	94	99	98	100	88	68	77	74	76	72	24	17	15	20	5
61	78	80	78	85	83	77	74	79	79	73	84	72	78	72	69	41	51	6	18
56	69	58	63	54	61	65	68	86	85	94	89	62	66	53	52	78	54	90	6
63	100	65	83	66	66	64	73	92	69	79	43	21	73	70	40	45	34	82	5
100	81	85	87	95	96	100	97	98	98	92	85	86	76	31	30	27	25	34	5
100	82	85	85	93	75	99	83	78	70	76	65	57	80	85	20	12	19	5	9
92	76	76	61	68	62	63	91	75	72	69	73	85	58	65	40	53	41	20	6
81	79	95	91	100	99	99	100	69	97	100	72	71	46	84	31	100	34	100	37
63	86	62	95	95	95	100	93	75	93	70	74	77	63	66	44	32	39	23	6
65	80	65	82	100	80	100	100	100	100	100	90	40	70	65	50	100	45	37	60
63	68	80	78	75	69	73	81	81	68	84	80	86	82	61	60	31	33	41	7
75	76	70	61	81	48	51	22	63	61	70	6	77	21	82	90	34	17	24	7
65	77	81	73	99	100	100	87	100	99	100	78	57	53	83	53	23	37	88	5
46	60	70	97	91	85	85	81	82	52	100	81	75	60	51	57	60	80	17	6
94	92	82	73	78	72	83	45	65	95	64	55	63	58	51	47	18	36	15	6
71	60	57	74	77	71	67	82	100	78	100	80	82	36	65	71	19	19	28	5
75	81	86	89	79	76	79	75	73	77	67	80	77	79	58	74	21	9	24	18
83	72	75	71	69	75	74	84	70	75	94	86	80	32	92	37	20	15	24	9
70	87	69	80	83	81	73	77	72	77	75	74	70	66	23	31	33	44	20	7
62	86	65	65	71	94	95	97	86	96	75	82	78	77	70	34	74	30	42	5
76	73	71	75	78	99	100	28	72	67	66	31	52	47	65	27	59	23	18	5
78	73	78	80	84	85	100	75	35	75	70	90	75	69	55	42	44	23	70	9
76	75	70	69	90	96	82	100	95	77	97	86	89	73	70	74	66	81	30	6
75	75	76	71	78	75	89	76	83	88	45	76	71	80	69	60	7	31	78	79
60	71	70	81	90	76	84	99	96	100	92	75	75	62	67	68	43	35	20	7
57	68	77	68	81	91	69	44	100	59	70	69	63	65	47	48	45	27	48	6
70	81	79	87	85	79	93	73	61	60	66	76	75	64	31	60	30	19	24	7
80	77	75	68	90	75	55	97	73	70	45	70	48	34	70	55	10	40	45	32
95	89	89	79	59	76	81	100	63	100	60	78	63	45	64	36	19	15	9	6
68	70	80	85	78	90	85	94	92	57	92	82	83	90	92	64	54	22	78	16
62	74	89	67	82	80	74	79	84	89	92	76	30	64	57	53	20	45	17	3
82	23	64	42	41	59	48	49	52	52	60	54	33	41	37	56	44	29	25	9

9.6 Seriation

Source: Hodson, F. R. (1968) The La Tène Cemetery at Münsingen Rain, Catalogue and relative chronology. Bern: Stämpfli.

Kendall, D. G. (1971) Seriation. In: Mathematics in the archeological and historical sciences (F.R. Hodson, ed.). Edinburgh: University of Edinburgh Press.

The data are 59 'closed find' graves and 70 variaties of grave-gifts from the early and middle La Tène period, situated at the La Tène cemetery at Münsingen Rain, Switzerland.

The aim of the analysis is to find the relative chronology of the graves based on the gifts found in them. In archeology this is called sequencing,seriation or the application of Petrie's concentration principle (Kendall,1971).

The matrix analyzed is an incidence matrix; that is to say a matrix where each row represents a grave and each column a variety (of jewellery etc.).The cells contain codes for presence (=1) or absence (=2) of a certain gift in a certain grave. Referring to Hodson's plate 123 (Hodson,1968) we analysed 59 graves, corresponding to the first 59 rows and 70 varieties, corresponding to the first 70 columns. see Table 1

We disturbed the original order on several not very systematic ways, more like shuffling cards, but in whatever order the data were, when analysed, the same minimum and the same final solution was reached every time, which was to be expected. The analysis discussed here was on the data in the order of Hodson's table 123.

There exist two typologies of the graves and they are drawn into the Homals-1 solutions. see also Table 2,3

Ia	Torcs,Marzabotto, roof-bow and certosa fibulae
Ib	Hollow armlets and anklets with continuous relief decoration
Ic	Bent rings
IIa	Middle La Tène fibulae

Table 2 Hodson's basic typology (Hodson, 1968 , pag. 28)

Table 3

Tomb identifier Hodson's plate 123	Ranknumber	Hodson's period Identification	Wiedmer's period Identification
13	1	Ia	A
32	2	Ia	A
7	3	Ia	A
9	4	Ia	A
16	5	Ia	A
23	6	Ia	A
44	7	Ia	A
12	8	Ia	A
8a	9	Ia	A
8b	10	Ia	A
6	11	Ia/b	B
31	12	Ia/b	C
51	13	Ia/b	C
40	14	Ia/b	D
48	15	Ib early	E
46	16	Ib early	E
62	17	Ib early	E
91	18	Ib early	E
49	19	Ib early	E
80	20	Ib early	E
107	21	Ib early	E
50	22	Ib late	F
68	23	Ib late	G
61	24	Ib late	G
152	25	Ib late	G
121	26	Ib late	H
90	27	Ib late	H
79	28	Ib late	H
84	29	Ib late	H
102	30	Ic early	I
136	31	Ic early	I
138	32	Ic early	I
94	33	Ic early	J
106	34	Ic early	J
135	35	Ic early	J
140	36	Ic early	K
134	37	Ic early	K
81	38	Ic late	L
130	39	Ic late	M
145	40	Ic late	N
132	41	Ic late	O
157	42	Ic late	O
158	43	Ic late	O
75	44	Ic late	P
149	45	Ic/II	Q
119	46	Ic/II	R
171	47	Ic/II	R
170	48	Ic/II	R
101	49	Ic/II	S
161	50	Ic/II	T
168	51	IIa	U
166	52	IIa	U
178	53	IIa	U
184	54	IIa	U
164	55	IIa	U
181	56	IIa	U
180	57	IIa	U
182	58	IIa	U
211	59	IIa	U

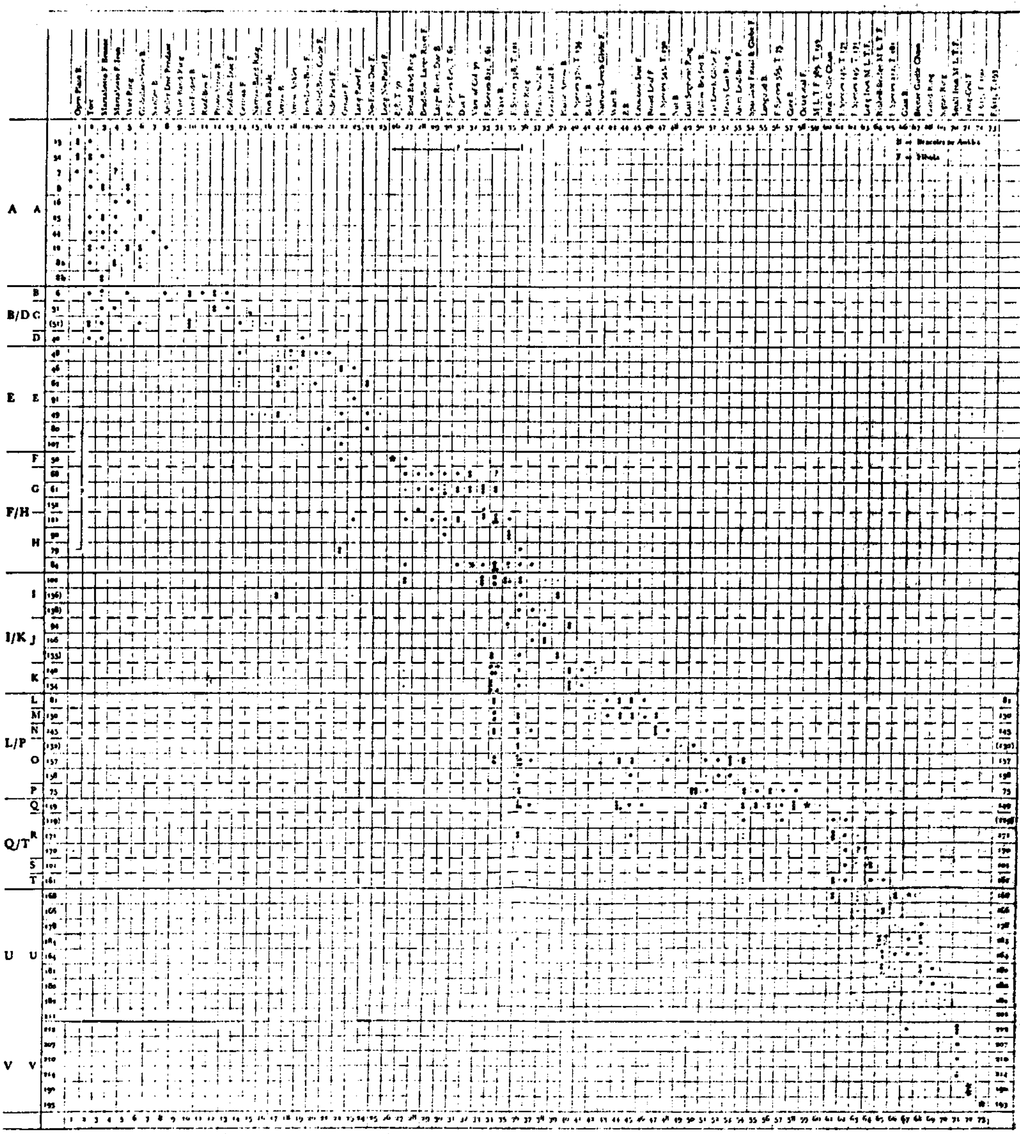


Table 1 Hodson's table 123

Apart from Hodson's seriation the graves were also seriated geographically by reason of the almost linear form of the cemetery. Both seriations are confirmed strongly by the one category solution and very weakly by the two category solution. This leads us to the two ways we analysed the data. First we analysed the data with two meaningful categories per gravegift: absent or present. See fig. 1,2,3
Figure 1 gives the grave-points labeled with their ranknumbers.
Figure 2 gives the same but labeled with Wiedmer's periods
Figure 3 gives the same but labeled with Hodson's periods

Secondly; the absence category can be caused by two reasons: the variety is not in the grave because it did not exist at that time or the variety was available in that time, but it is not in the grave for other reasons. These two causes may make the data heterogeneous. Therefore we analysed the data also with only one meaningful category: the presence of varieties. The absence is treated as missing value. That the absence structure is heterogeneous can be seen from the great difference in stress between the one and two category solution and also from the fact that the one category seriation is confirmed by the other seriations (see fig. 4,5,6); in fact the two category seriation can hardly be called a seriation at all.

	One category solution	Two category solution
First Dimension	.036	.891
Second dimension	.101	.915
Total stress max.=2	.138	1.806

Table 4 The stresses of the two Homals-1 solutions

Fig. 1 Münsingen Rain data
Two categories
Hodson's rank nrs.

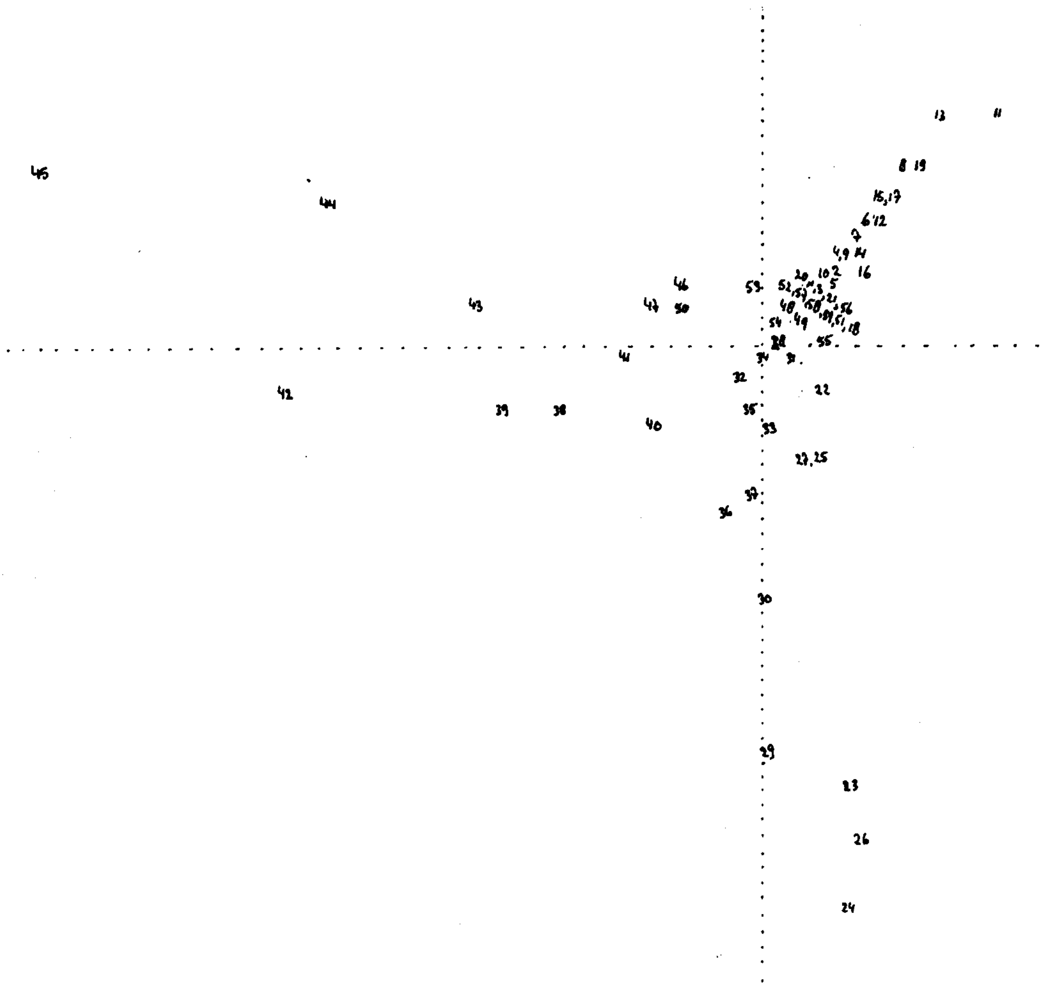


Fig. 2 Münsingen Rain data
Two categories
Wiedmer's periods

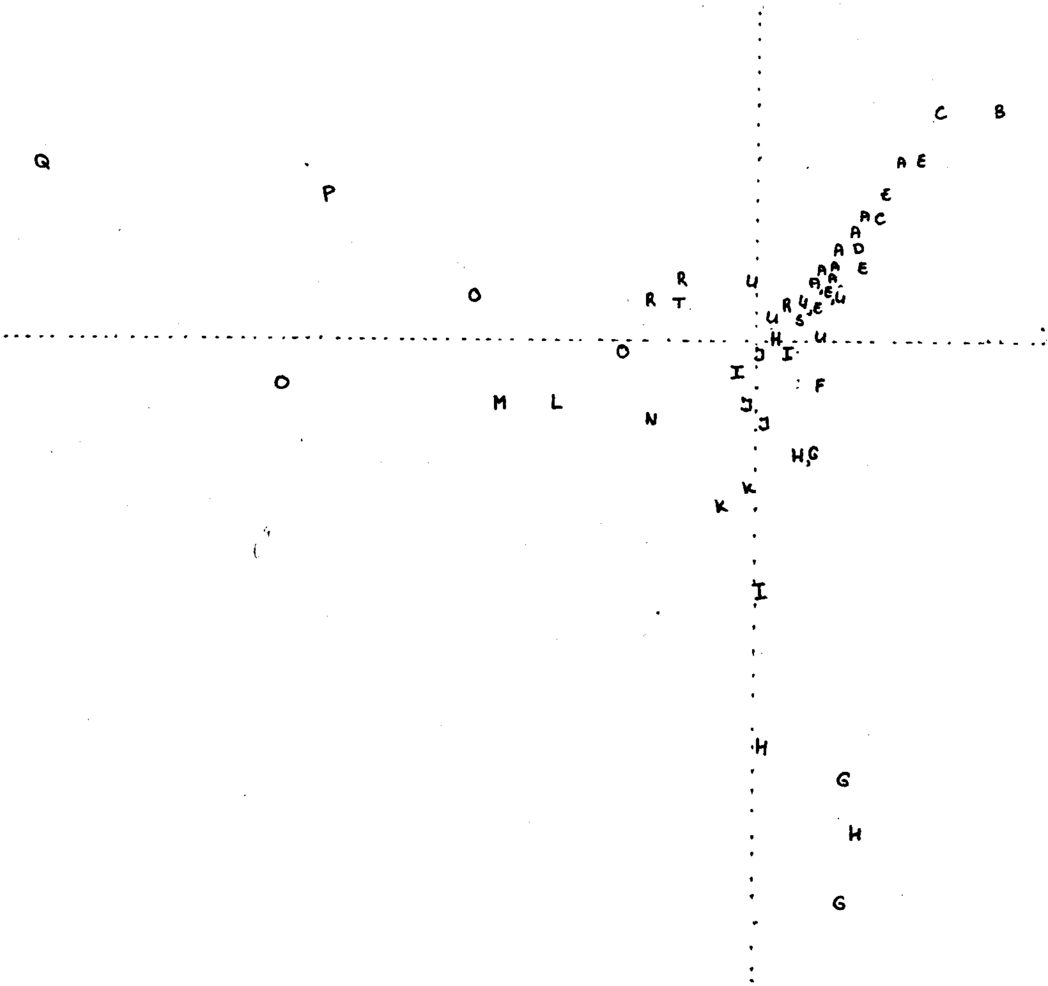


Fig.3 Münsingen Rain data
Two categories
Hodson's periods

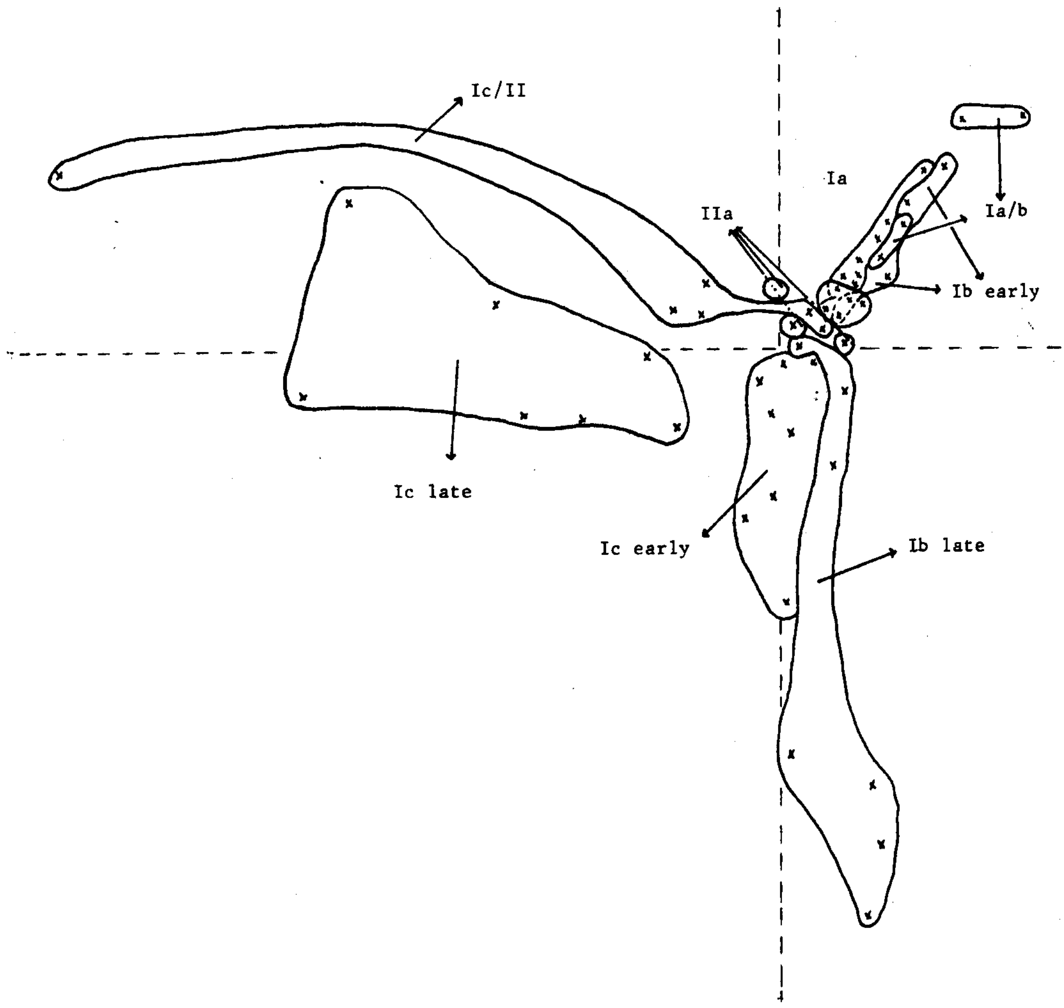


Fig.4 Münsingen Rain data
One category
Ranknumbers

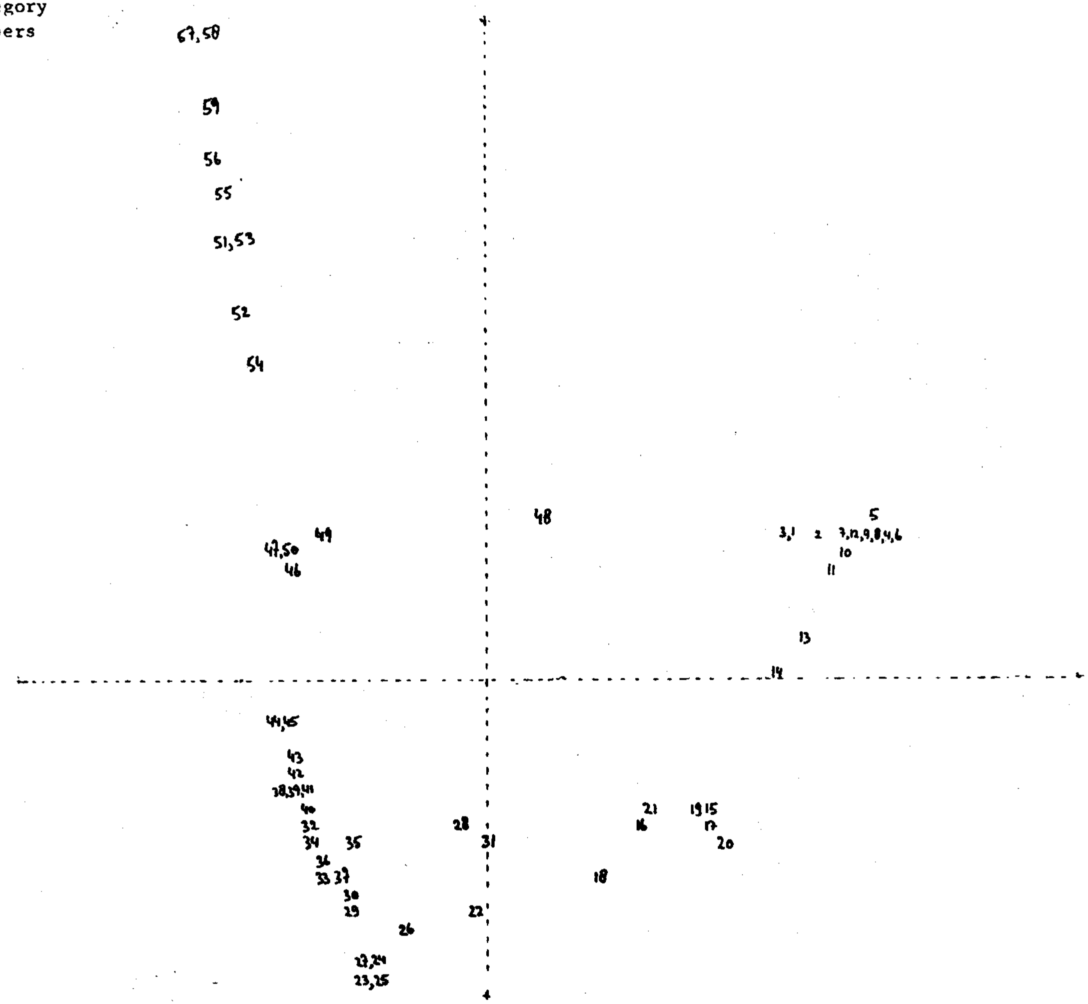


Fig.5 Münsingen Rain data
one category
Wiedmer's periods

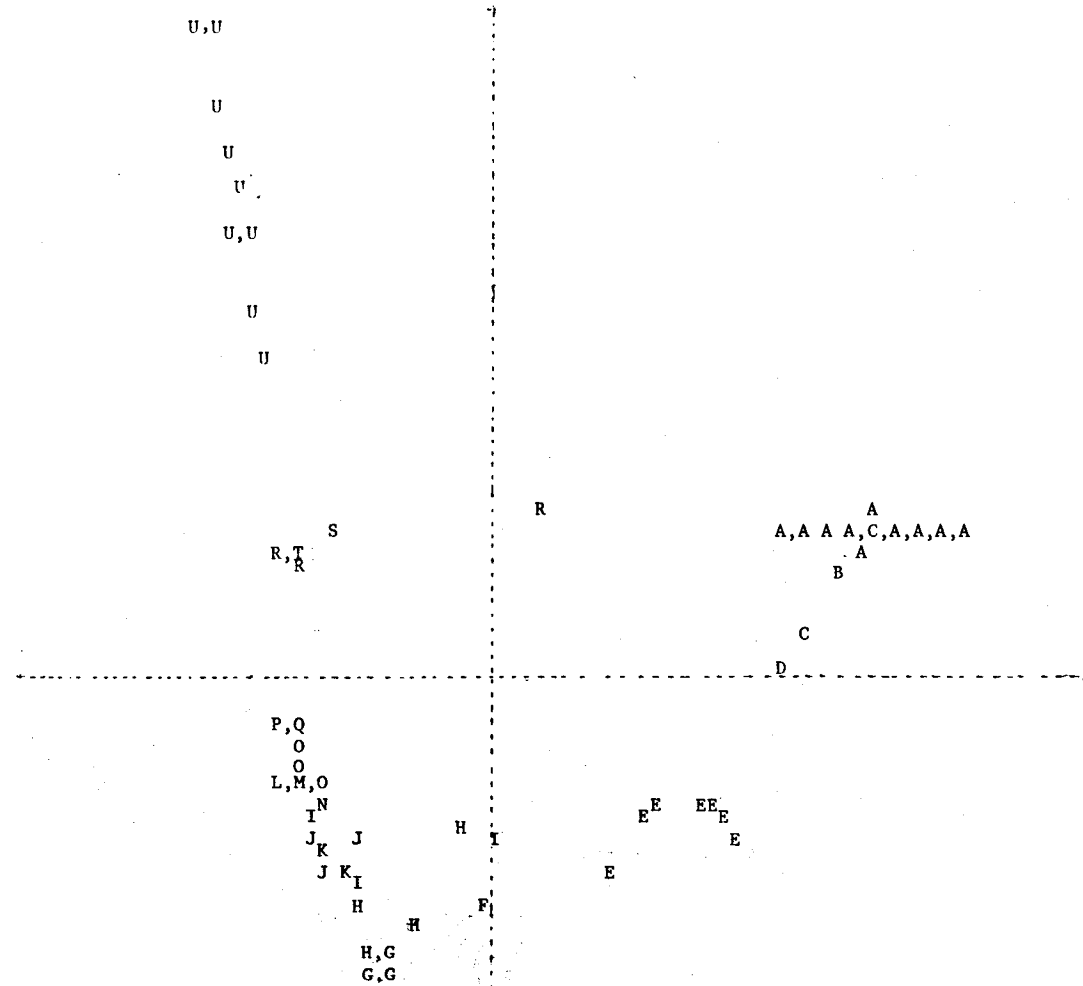


Fig.6 Münsingen Rain data
one category
Hodson's periods