2200176

H. Niewodniczański
INSTITUTE OF NUCLEAR PHYSICS,
Kraków, Poland.

INP--1624/PL

INP - REPORT NO. 1624/PL

Investigation of Clustering in Sets of Analytical Data

JÓZEF KAJFOSZ

Kraków — April 1993

WYDANO NAKŁADEM
INSTYTUTU FIZYKI JĄDROWEJ
IM. HENRYKA NIEWODNICZAŃSKIEGO
KRAKÓW, UL. RADZIKOWSKIEGO 152

Kopię kserograficzną, druk i oprawę wykonano w IFJ Kraków

Wydanie I zam. 49/93 Nakład 75 egz.

Investigation of Clustering in Sets of Analytical Data

JÓZEF KAJFOSZ

H. Niewodniczański Institute of Nuclear Physics, 31-342 Kraków, Poland

Abstract: Foundations of the statistical method of cluster analysis are briefly presented and its usefulness for the examination and evaluation of analytical data obtained from series of samples investigated by PIXE, PIGE or other methods is discussed. A simple program for fast examination of dissimilarities between samples within an investigated series is described. Useful information on clustering for several hundreds of samples can be obtained with minimal time and storage requirements.

Streszczenie: Przedstawiono pokrótce podstawy statystycznej metody analizy danych, służącej do uzyskiwania informacji o grupowaniu się obiektów w obrębie pewnego zbioru, tzw. „cluster analysis". Przydatność tej metody dyskutowana jest na przykładzie jej zastosowań do danych z analizy pierwiastkowej serii próbek, badanych metodami PIXE i PIGE. Podano opis prostego programu komputerowego, umożliwiającego szybkie dokonywanie tego typu analiz na zbiorach, zawierających do kilkuset próbek, przy stosunkowo skromnych wymaganiach sprzętowych.

1. Introduction

In the application of all analytical methods in general, and of methods for elemental analysis of the composition of matter in particular, a considerable amount of data characterising the investigated samples is produced. Optimal utilization of the information contained in these data becomes increasingly important. Among many different statistical methods used in their evaluation a specific and still seldom used one is the method of cluster analysis (For compact information and numerous references the book in ref. [1] seems worth recommendation). Its objective is to uncover statistically significant grouping of items (samples, probes, objects) in a given series on the basis of their similarity in a predetermined respect and to expose dissimilar groups (called clusters), if any, and find their number and populations, in order to provide an aid for interpretation and drawing conclusions of merit from the experimental results.

Although the methods of cluster analysis are not new and by now are relatively well developed, and although elaborate software for their application exists, their usage is still limited. First, because many workers in various fields of science are still not sufficiently familiar with them and their features. Second, because the existing commercial software is not everywhere available. Third, because the introduction of cluster analysis either by the development of new programs or by the implementation of the existing ones is a time-consuming task. Fourth, let us mention that with increaseing of number of items in the series analysed both the computing time and the storage required in most programs grow rapidly, making them inapplicable to smaller computers.

The above led to the development of the program presented in this report. Its simplicity and low requirements for memory space, computing time and implementation efforts make it suitable for small personal computers and for laboratories with limited resources. The program may also be considered a first step in the introduction of cluster analysis thus facilitating further contacts with more sophisticated systems (e. g. [2] and [3]).

2. Essentials of the method

a) Terminology

Cluster analysis is not a method suddenly invented by somebody. Rather its individual elements have developed gradually at various places as specialists in many branches of scientific activity tried to draw conclusions from their experimental material. Therefore different terms are used for the description of essentially the same method. Some call it "multivariate analysis", others "hierarchical grouping". In psychology this method is often called "pattern recognition". In botany and zoology the terms "taxonomy" or "taxometry" are used. In this report we will use the term "cluster analysis", being aware, however, that some readers may recognise it as a method well known to them under another name.

b) Items and their properties

Generally speaking, the subject of cluster analysis is always a certain number (which we will call a set or a series) of items of some kind, which, depending on their character, may be e. g. objects, cases, individuals, localities, samples etc. Each item of the set in question is characterised by a certain number of properties, expressed quantitatively by some numeric values. These properties in turn can be called attributes, characteristics, patterns, etc., depending on their character.

This can be made clear by the following examples. In psychological studies items are individual persons and their properties may be their behavioural patterns, intellectual and/or somatic characteristics, scores attained in some tests, etc. In botany or zoology items are individual plants or animals, or their species, and their properties are e. g. their genetic and/or environmental characteristics. In environmental studies items could be e. g. cities and their propereties may be their environmental, geographical, industrial or social characteristics. In material studies items would be some samples and their properties, e.g. their mechanical, physical chemical or crystallographical characteristics.

Generally speaking, the set consists of N items (objects, samples) and each item is characterized by M properties (patterns, attributes). This gives a total set of $N \times M$ data values $C(i,j)$, where $i = 1,2,...,N; j = 1,2,...,M$. Hence the first index i denotes the item number while the second index j denotes the number of the property.

In our application of cluster analysis items are samples for elemental analysis and their properties under study are their elemental concentrations for a number of elements, determined in PIXE or PIGE analysis. Further on we will speak only about samples and their concentrations, one has to be aware, however, that cluster analysis can easily be similarly applied to a very broad spectrum of other objects and their properties.

c) Dissimilarity of samples

An important concept in cluster analysis is the similarity or dissimilarity of samples. Two samples i,j are similar, if the differences of their correspondent elemental concentrations are small, i. e. if

$$C(i,k) - C(j,k) = d(i,j,k) \tag{1}$$

are small numbers for all or at least most of the values of k. Of course the term "small" is relative and can be meaningful only in comparison to other similarities within the series.

In order to make similarities comparable and the concept useful, the definition of a dissimilarity variable is necessary. The simplest and most straightforward choice can be derived from the analogy with space coordinates. As the "dissimilarity" of points in geometry can be defined by their distance in an Euclidean space, so the dissimilarity of samples can be defined by their "distance" in a "concentration space". Each sample can be represented as a point in a M-dimensional Euclidean space, where the coordinate of the k-th dimension is simply the concentration $C(i,k)$ of the element k in the i-th sample. Then the "distance" $d(i,j)$ of two samples i and j can be expressed as

$$d(i,j) = \sqrt{\sum_k [C(i,k) - C(j,k)]^2}. \tag{2}$$

In such a concentration space the points representing similar samples will be close to one another, while large distance $d(i,j)$ will indicate a great dissimilarity of the correspondent samples.

The definition of the sample distance as the dissimilarity variable introduced in equation (2) is neither the only one possible nor the only one used. In the next section several other choices of dissimilarity variables will be shown.

d) Clustering of samples

The representation of samples as points in a "concentration space" offers a possibility to observe their mutual relations. If subsets of similar samples exist within the entire set, their correspondent points will be apparent as clusters of close points, more or less separated from the rest of the points. This could lead to conclusions of merit about the samples, e. g. their origin, history, previous handling etc., or could help to uncover some unknown analogies and affinities among them.

There are a number of reasons for such affinities and thus for grouping items into clusters. In the course of PIXE analyses we found, e. g. the following examples:

— Ancient clay pottery: origin, methods of manufacturing, conditions of preservation;

— Plants from polluted areas: pollution sources, accumulation mechanisms, time of sampling;

— Human body fluids or tissues: nutrition, environment, habits, health status;

— Geological samples: mineral constituents, time periods of formation, conditions of erosion.

Usually there exists some natural and known grouping of samples, e. g. according to their sampling place, sampling time or other criteria. The simplest way of drawing conclusions from experimental data is through comparison of mean values of concentrations of such groups. In many applications, however, it proves to be useful to investigate the grouping of the samples on the basis of the analytical data alone, apart from the known or anticipated natural grouping within the set. Knowledge of the clustering of samples, i. e. of their distribution into several groups, with the samples within a group being similar to one another, while samples from different groups differ in some respect, can be of importance for several reasons:

— It can prove or disprove the correctness of the natural grouping. E. g. if the data set consists of elemental concentrations in specimens taken from sick and healthy humans, but the data do not indicate dissimilarities between those two groups larger than the dissimilarities between members of the same group, then the analysis can obviously serve neither for the diagnostics of the disease nor for drawing valid conclusions about correlations between the disease and the elemental contents of the investigated specimens;

— The grouping on the basis of analytical data alone can reveal facts of merit not anticipated and thus lead to some new findings.

— It can be used to test the suitability of analytical methods for solving individual problems. By including into a set several independent analyses of each sample of a series, one can check the relation between the accuracy of the method (visible in dissimilarities between analyses of the same sample) and the effects under study (visible in dissimilarities between analyses of different samples). For a method of high accuracy analyses for each sample should form a separate cluster, thus indicating the possibility to determine safely the differences between samples.

— It also renders the possibility to search for samples of predetermined properties. If one includes a "model sample" into a set, the clustering will give information about the similarity of all samples to this model.

The aim of the cluster analysis is to evaluate the clustering of samples within a set and to present it in a form suitable for easy inspection by the user. The two subsequent sections describe the way of doing it.

3. Evaluation of the clustering

Prior to the actual calculation of the clustering in a sample set several preparatory steps are necessary. The options taken in those steps are of great importance since the result of the clustering will generally depend on them: different sets of options will lead to different clusters with different membership.

The preparatory steps are described shortly in the following four subsections. The fifth one describes the main calculation process.

a) Initial modification (standardization) of data

In many cases it is reasonable not to work with the original experimental values (like elemental concentrations) but to modify them in some way. The aim may be for instance to diminish the influence of outsiders, to moderate the influence of experimental errors, to make the contribution of all elements approximately equal or proportional to their relevance or to the accuracy of their concentration values, etc.

Several of the modifications usually adopted are mentioned below. In all of them the initial data $C(i,j)$ are replaced by modified data $C'(i,j)$. The formulae indicate how the modified data are obtained.

— Equalization of mean values:

$$C'(i,j) = C(i,j)/a(j) \quad \text{where} \quad a(j) = \sum_i C(i,j)/N \tag{3}$$

— Equalization of ranges:

$$C'(i,j) = C(i,j)/b(j) \quad \text{where} \quad b(j) = \max_i C(i,j) - \min_i C(i,j) \tag{4}$$

— Equalization of maxima:

$$C'(i,j) = C(i,j)/c(j) \quad \text{where} \quad c(j) = \max_i C(i,j) \tag{5}$$

— Equalization of standard deviations:

$$C'(i,j) = C(i,j)/d(j) \quad \text{where} \quad d(j) = \sqrt{\sum_i [C(i,j) - a(j)]^2/(N-1)} \tag{6}$$

— Logarithmisation:

$$C'(i,j) = \log C(i,j) \tag{7}$$

— Various combinations of the above ones;

— Others.

The necessity of modifying the data is obvious e. g. in situations where concentrations of some elements are much larger (even by several orders of magnitude) than those of others. In clustering with unmodified data the former elements would dominate in the clustering eliminating the contributions from the latter ones.

b) Choice of dissimilarity variable

The results of the clustering process may be further modified and made more legible by a proper definition of the dissimilarity between two samples. Earlier in eq. (2) we gave the formula for the simplest and most natural one: the Euclidean distance (for unmodified data). Below several other possible choices are shown below:

— Euclidean distance:

$$d(i,j) = \sqrt{\sum_k [C'(i,k) - C'(j,k)]^2} \tag{8}$$

— Squared Euclidean distance:

$$d(i,j) = \sum_k [C'(i,k) - C'(j,k)]^2 \tag{9}$$

— Cosine of vectors of values (insensitive to dilutions):

$$d(i,j) = \frac{\sum_k [C'(i,k)C'(j,k)]}{\sqrt{\sum_k [C'(i,k)^2] \sum_k [C'(j,k)^2]}} \tag{10}$$

— City block (Manhattan) distance:

$$d(i,j) = \sum_k |C'(i,k) - C'(j,k)| \tag{11}$$

— Chebyshev distance:

$$d(i,j) = \max_k |C'(i,k) - C'(j,k)| \tag{12}$$

— Minkowski distance (generalised Euclidean):

$$d_p(i,j) = \left(\sum_k |C'(i,k) - C'(j,k)|^p\right)^{1/p} \tag{13}$$

— Distance in absolute power metric:

$$d_{p,r}(i,j) = \left(\sum_k |C'(i,k) - C'(j,k)|^p\right)^{1/r} \tag{14}$$

— Others.

It is difficult to give any general recommendations. The best choice can be made after some tests and gaining experience in the specific application.

c) Choice of clustering method

The term "clustering method" here means the prescription for calculating the dissimilarity between a sample and a group of samples (a cluster) as well as between two groups of samples (two clusters).

Let us have a cluster P which consists of m samples numbered $p_1, p_2, ..., p_m$; and a cluster R of n samples denoted $r_1, r_2, ..., r_n$. There are generally $m \times n$ values $d(p_i, r_j)$ of dissimilarity between their members. The following methods of clustering differ by the definition of the inter-cluster dissimilarity in terms of these values $d(p_i, r_j)$.

— The simplest and probably most frequently used is the single linkage, also called the nearest-neighbor method. The dissimilarity of clusters is defined as the dissimilarity of their nearest members, i. e. as the minimal of all the $d(p_i, r_j)$ values.

— In average linkage the average value of all $d(p_i, r_j)$ is taken as the dissimilarity between clusters.

— In complete linkage, also called the furthest-neigbor method, the maximal $d(p_i, r_j)$ value is used.

— In centroid clustering and median clustering methods first the coordinates of the centroids or medians for the two clusters are determined and then the inter-cluster dissimilarity is calculated from them as the dissimilarity of two samples.

— In some cases even more sophisticated clustering methods are used.

d) Treatment of missing values

A small but important problem which requires solution before the clustering can be carried out is how to deal with missing values in the data set. The omission of items (samples) or variables (elements) where the data are incomplete, e. g. where some values are missing, may be undesirable, while assigning them zero values may distort the results (by unduly increasing

some of the dissimilarities). The safest solutions is probably to give them mean values for the element in question.

e) Clustering procedures

The process of clustering consists in joining samples into clusters starting from the lowest values of the dissimilarity variable $d(i,j)$ and proceeding with increasing d until all samples are joined together. The starting point thus is N clusters, each consisting of one sample, and the final point is one cluster which consists of all the samples. Since every joining step diminishes the number of clusters by one, there are totally $N-1$ joining steps. Each of them is characterised by a certain value of the dissimilarity variable and they can be sorted by increasing d.

Most significant for the investigation are those of the $N-1$ clusters created in the process, which are formed relatively "early", i. e. at low values of d (thus indicating high similarity of the samples within them) and are sustained for a "long time" i. e. merge with other clusters at high values of d (thus indicating large dissimilarity to samples from outside the cluster).

Essentially two ways of carrying out the clustering can be chosen. The first one is more straightforward but less practical. It consists in the following steps:

— Calculate $d(i,j)$ for all pairs i,j of samples. (The number of $d(i,j)$ values is $N(N-1)/2$ and thus increases rapidly with increasing N);

— Reorder the $d(i,j)$ triangular matrix by ascending values of d;

— Beginning with the lowest d find the $N-1$ clustering steps.

This method is inconvenient because of high computer memory and computing time requirements. The memory space is proportional to N^2 while the time needed is proportional to N^4. Because of that a more economic approach was proposed in ref. [4]. Its use leads to substantial reduction in memory size (proportional to N) and time (proportional to N^2). The steps of the modified method are the following:

— Carry out the clustering for k samples (starting with $k=2$);

— Add the next $(k+1)$-th sample, calculate the k dissimilarity values $d(i,k+1), i=1,2,...,k$ and update the clustering record;

— Increase k by 1 and return to the previous step. Repeat the two steps until all N samples are included.

This method makes possible cluster analysis for hundreds of samples even on small personal computers. Its drawback is that it is limited to the "single-link" or "nearest-neighbor" method, because only for this method the algorithm for the stepwise updating of the clustering record was developed.

4. Presentation of results of clustering

This section describes shortly the ways or tools commonly used for displaying the results of the clustering process.

a) The dissimilarity spectrum

A simple feature for visual examination of the clustering within an investigated set, which seems not to have been used before in any of the existing descriptions of the method of cluster analysis, is the dissimilarity spectrum. Each of the $N(N-1)/2$ dissimilarity values $d(i,j)$ is represented in it by one count in an appropriate channel proportional to the d value. We used 100 channels for the whole range between d_{min} and d_{max}. In a set with no grouping (i. e. with all the samples forming one big cluster) the dissimilarity spectrum would be a broad structure with only one maximum in the region of most frequently occurring dissimilarity values (apart from statistical fluctuations in the channel contents). A typical spectrum of this kind is shown

in Fig. 1. It represents a set of data from 127 hair samples from cancer patients [5]. A smooth single-peak structure indicates no clustering within this set.

In the case of some clustering the shape of the dissimilarity spectrum is substantially different. Let us suppose that the samples of the set form two clusters consisting of s and t samples, respectively. Then there would be $s(s-1)/2 + t(t-1)/2$ intra-cluster dissimilarities and $s \times t$ inter-cluster dissimilarities. Since the values of the former are smaller than those of the latter (otherwise the clusters would not be separated), they will form two distinct peaks in the spectrum, at lower and at higher dissimilarities.

This was demonstrated on a data set of 125 "samples" randomly created and separated into two distinct groups (Fig. 2). That this idea works also for real analytical data sets follows from Fig. 3, where the dissimilarity spectrum for a set of 76 sedimentary rock samples is shown. The distribution of these data points in a 2-dimensional concentration space is shown in Fig. 4.

We have found, moreover, that even some information on the number of clusters can be obtained from the dissimilarity spectrum. A simple comparison of the number of intra-cluster and inter-cluster dissimilarities indicates that for two clusters the area of the second peak cannot be larger than the area of the first one. For more than two clusters, however, this relation can be changed, as seen in Fig. 5. This dissimilarity spectrum thus indicates the presence of more than two clusters.

The dissimilarity spectrum has been found to be a convenient and easily obtainable tool for a fast examination of clustering in a data set.

b) The dendrogram

The oldest and most familiar tool for presentation of the clustering results is the dendrogram (Fig 6 a). Its name is derived from the Greek word "dendron", which means "tree", because when watched upside down it has a tree-like shape.

On the horizontal axis are the numbers of samples of the set arranged in a certain sequence while on the vertical axis is the dissimilarity. The horizontal bars indicate the dissimilarity levels, at which individual samples or previously merged groups of samples (clusters) merge into one cluster. Thus at the bottom is the initial stage of N separate one-sample clusters while at the top is the final stage of one N-sample cluster. The $N-1$ horizontal bars at different dissimilarity levels indicate the $N-1$ joining steps of the clustering.

Notice that the dissimilarity values listed below are not attributes of the samples but of the joining steps, showing the level at which the sample (or the cluster) at the left side of the value merges with the sample (or the cluster) at its right side. Consequently, the dendrogram indicates two well-resolved clusters within the set, one consisting of samples no. 6, 4 and 1, and the other containing samples 7, 2, 5 and 3.

The dendrogram proves to be a proper tool for smaller sets of samples. With increasing N, it very soon (at about $N = 50$) becomes too complex and difficult to follow. For larger sets, therefore, another way of presentation is more suitable.

c) The icicle plot

This form of presentation of the clustering was probably discovered independently several times. It is essentially a modification of the dendrogram. The idea is apparent from Fig. 6 b), where the way of obtaining an icicle plot from a dendrogram is shown. The three bars (two vertical ones and one horizontal) of a joining mode in a dendrogram are replaced in the icicle plot by a single vertical bar, as shown in Fig. 6 c). This shape justifies the name "icicle plot", because when watched upside-down the bars resemble icicles.

This plot is both easy to produce and easy to inspect even for large numbers of samples. One only has to keep in mind that the vertical "borderline" divides the samples at its left from those at its right side (at the dissimilarity level determined by its height) or, in other words,

joins the left cluster with the right one at the level determined by the top point of the bar. The two clusters (which may also be single samples) extend to both sides until a higher "borderline" is encountered. This determines unambiguously the membership of the joining clusters for all of the $N-1$ joining steps.

The state of the clustering at an arbitrary value of the dissimilarity variable can be easily inspected by means of a rule placed in the icicle plot horizontally at the chosen level. Each vertical bar which intersects this horizontal line (i. e. is higher than the chosen level) separates two clusters. All the samples between two nearest such vertical bars form one cluster (at the chosen level). It is also to be noticed that the dissimilarity between two arbitrary samples (in the sense of the chosen clustering method) is determined by the highest bar which separates them in the icicle plot. (This dissimilarity, however, may not be identical with $d(i,j)$ of those two samples because it may be an inter-cluster dissimilarity, calculated in a different way).

Because of its convenience and suitability for sets with large N we use only this last form of presentation of clustering results, although the name "dendrogram" is still in use in the program.

Fig. 7 shows the icicle plot for the set of hair samples the dissimilarity spectrum of which was shown in Fig. 1. Indeed, one big cluster is formed at low dissimilarity levels and for higher d only individual samples (the outsiders) are added. Fig. 8 shows the icicle plot for the randomly created data set (Fig. 2) while in Fig. 9 the clustering of the sedimentary rock series (Fig. 3) is shown. Both Fig. 8 and 9 show a pronounced two-cluster structure of the data. The icicle plot for the series illustrated by the dissimilarity spectrum of Fig. 5 confirms the finding that more than two clusters were involved. Fig. 10 shows clearly the presence of three clusters.

5. The program

A simple computer program in FORTRAN was developed for the application of cluster analysis of data from PIXE and PIGE elemental analyses. Its main characteristics and a short instruction for its use are given below.

a) Characteristics and options

In the present version of the program its properties and possibilities are the following:

— Clustering method: single-link (nearest-neighbour);

— Method of calculation: stepwise updating of the clustering record (ref. [4]);

— Possible definitions of dissimilarity variable: Euclidean distance or squared Euclidean distance;

— Possible modifications of input data: Logarithmisation and/or equalisation of mean values;

— Treatment of missing values: replacement by mean values;

— Maximal number of samples: 200 with the present dimensions; may be easily extended;

— Maximal number of variables (concentrations): 20 with the present dimensions; may be easily extended;

— Way of presentation of results: Tables of data for the dissimilarity spectrum and for dendrogram construction; simple icicle plot produced within the ASCII output file;

— Further possibilities: Creation of tables for dissimilarity spectra after the rejection of preset fractions of outsiders.

b) How to use the program

The FORTRAN computer code of the program is available in the file DIST.FOR and its executable file is DIST.EXE. Two files: DIST.DAT and SAMPLE.DAT contain the input data.

The file DIST.DAT contains the settings and options, and SAMPLE.DAT the set of data for the clustering. The program produces an ASCII output file DIST.OUT with the results of the clustering ready for printing.

Both input files are read by the program according to the rules for list-directed input (with no format specification (*)) in which the separators between individual items are spaces or commas.

The file DIST.DAT contains the following data:

N — Number of samples in the set (up to 200);

M — Number of elements in the data (up to 20);

K — Number of elements included in the clustering (up to M);

$JE(J), J = 1, ..., K$ — Sequential numbers of the K elements included;

IL — The logarithmization option. For $IL = 0$ concentrations (raw data), for $IL = 1$ logarithms of concentrations (logarithmed data) will be used;

IQ — The choice of dissimilarity variable. For $IQ = 0$ Euclidean distances, for $IQ = 1$ squares of Euclidean distances will be used;

IM — The equalization of mean values option. For $IM = 0$ not equalized mean values, for $IM = 1$ equalized mean values will be used;

KO — Requested number of dissimilarity spectra with outsiders rejected (up to 10);

$OUT(J), J = 1, ..., KO$ — The fraction of the number of all dissimilarities which are to be regarded as outsiders and rejected. If $KO = 0$ one dummy value should be supplied.

The file SAMPLE.DAT contains the following data:

— First line: General description of the data set (up to 60 characters); will be printed in the output heading;

— Second line: Names of elements present in the data (up to 60 characters); will be printed in the output heading;

— Third and subsequent lines: The sequence $(I, (C(i,j), J = 1, ..., M), I = 1, ..., N)$, i. e. the analytical data for N samples of the set. The data for each sample begin with the sample number I and continue with M values of concentration $C(i,j)$. Each sample must start at the beginning of a new record.

— Last line: An empty record as an indication of the end of data.

Missing values are always replaced by the mean value for the element in question. Each value $C(i,j) \leq 0.1$ is treated as a missing value.

The output file DIST.OUT contains the headings, information on the settings and options and the results of the clustering in the form of tables and of an icicle plot.

The results start with a table of the dissimilarity spectrum without rejection of outsiders. Then the dendrogram data follow. They contain for each sample the dissimilarity value at which the sample joins and the sequential number of the sample it joins. The transposed dendrogram data contain the level of the joining with the sample at the left side and the sequence number of the sample in the original data. In the icicle plots of Figs. 7 – 10 these transposed dendrogram data are displayed. In the transformed (reordered) dendrogram data the same information is contained but the sequence of the samples is changed. At the left side are the two samples which join first (at the lowest dissimilarity value). This reordering makes the sequence of samples in the icicle plot independent on the sequence in which the samples appear in the input data. In the present version of the program the icicle plot displays the transformed (reordered) dendrogram

data. Finally, if $KO > 0$, the tables of the dissimilarity spectra with the preset fractions of outsiders rejected are displayed.

6. Conclusions

We have found cluster analysis to be a valuable tool in the evaluation of sets of analytical data. Its application is possible without great effort. The computer program developed, despite its simplicity and limitations has proved to be a helpful aid in a number of studies based on large quantities of analytical data. This method seems to be worth recommendation for every laboratory where such data are produced and analysed.

References

[1] B. Everitt: *Cluster Analysis*, Heinemann Educational Books, London 1977

[2] M. J. Norušis: *SPSSx — Introductory Statistics Guide*, McGraw-Hill — SPSS Inc., New York 1983

[3] E. V. Sayre: *Brookhaven Procedures for Statistical Analyses of Multivariate Archaeometric Data*, Rep. No. BNL-21693, Upton, N. Y., 1984

[4] R. Sibson: "SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method", *Computer J.* **16** (1973) 30

[5] W. M. Kwiatek, M. Cholewa, J. Kajfosz, K. W. Jones, R. E. Shore, A. L. Redrick: "Correlation of Trace Elements in Hair of Patients With Colon Cancer", *Nuclear Instruments and Methods in Physics Research* **B22** (1987) 166

Figure Captions

Fig. 1 Dissimilarity spectrum of data from a hair sample series. Elements included: Ca, Fe, Cu, Zn and Br. Means equalized, data logarithmed. Euclidean distances used.

Fig. 2 Dissimilarity spectrum for a set of randomly created data with a two-cluster structure.

Fig. 3 Dissimilarity spectrum for data from sedimentary rock sample series. Ca and Fe included. Means equalized. Euclidean distances used.

Fig. 4 Two-dimensional scattergram of the rock sample series.

Fig. 5 Dissimilarity spectrum for a randomly created three-cluster data set.

Fig. 6 Tools for visual presentation of the results of clustering: A dendrogram (a) and an icicle plot (c). In (b) the equivalence and interrelation of both these methods is shown.

Fig. 7 Icicle plot of data from the hair sample series.

Fig. 8 Icicle plot of the two-cluster random data set.

Fig. 9 Icicle plot of data from the rock sample series.

Fig. 10 Icicle plot of the three-cluster data set.

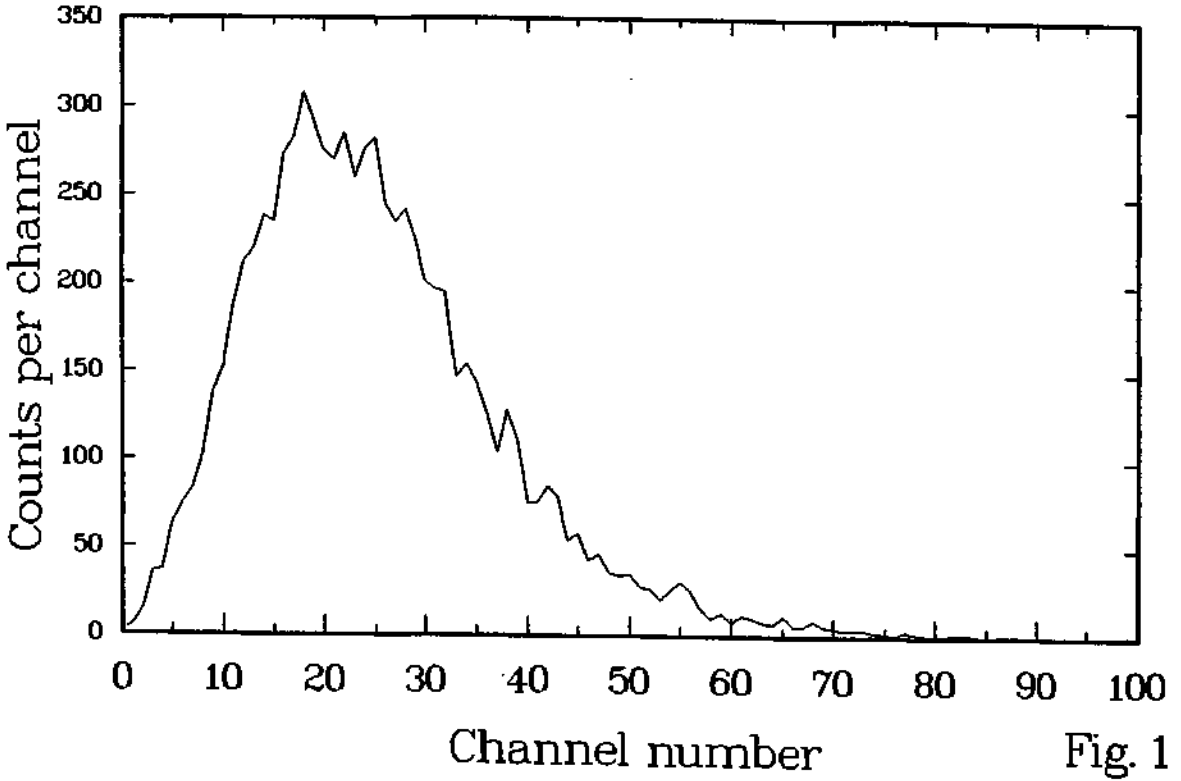
Counts per channel
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Channel number

Fig. 1

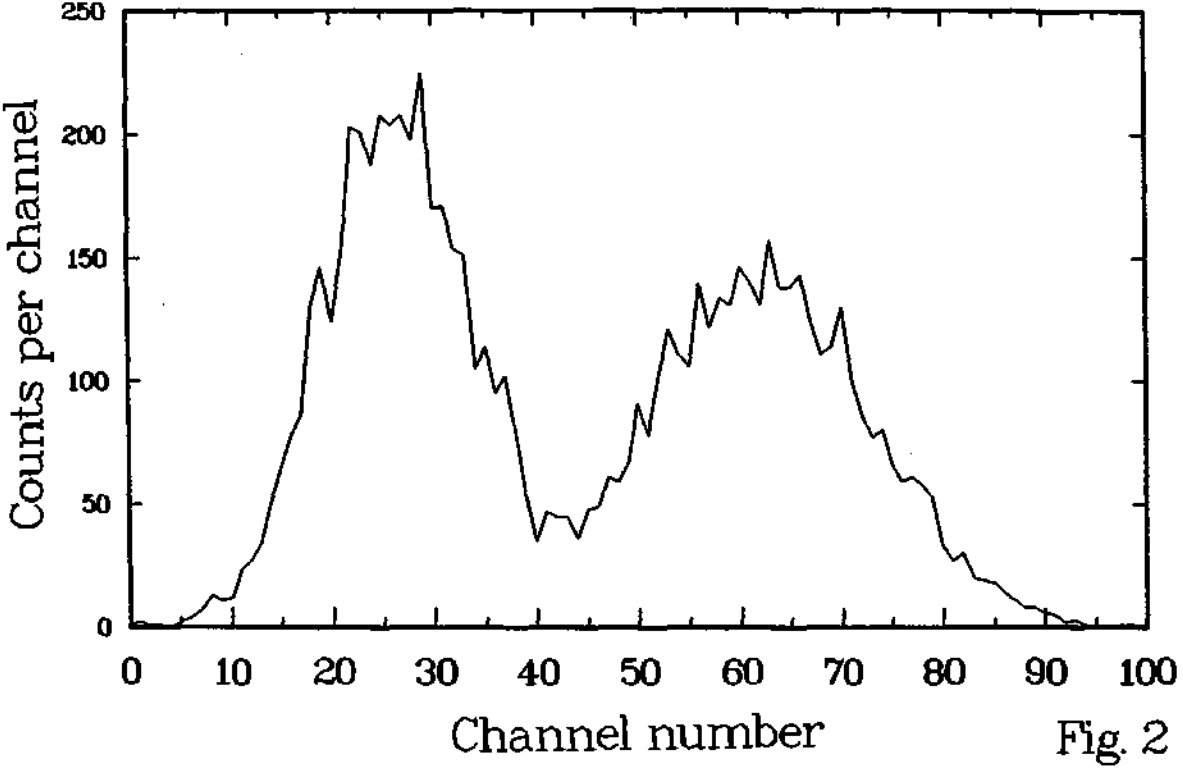
Counts per channel
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Channel number

Fig. 2

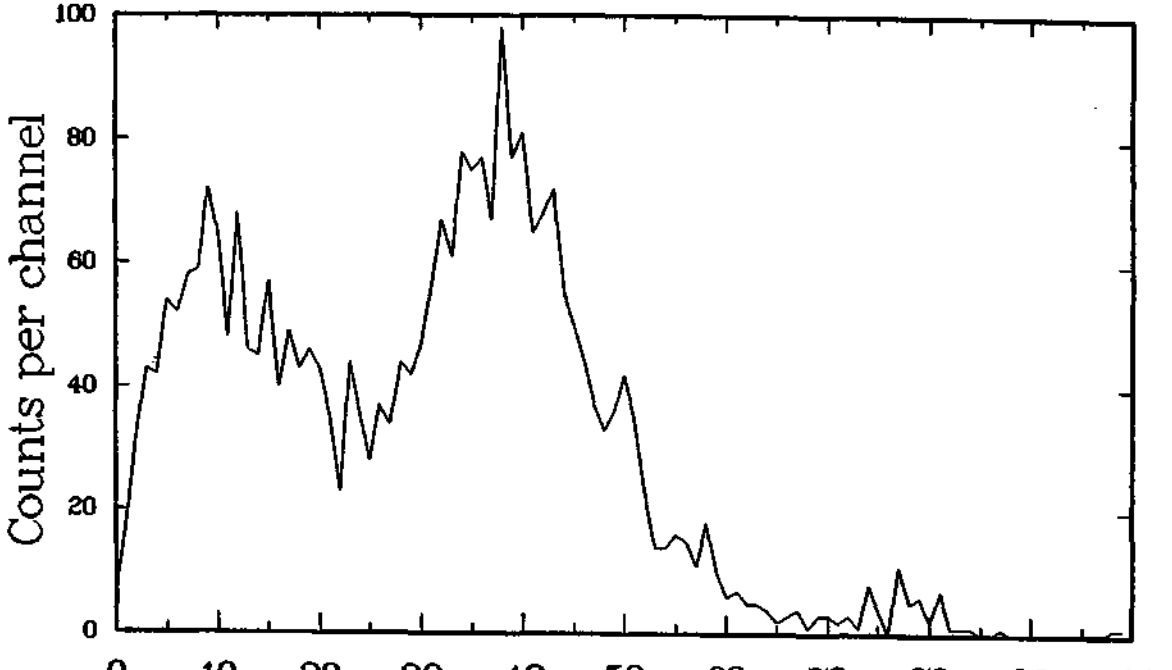
Counts per channel
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Channel number

Fig. 3

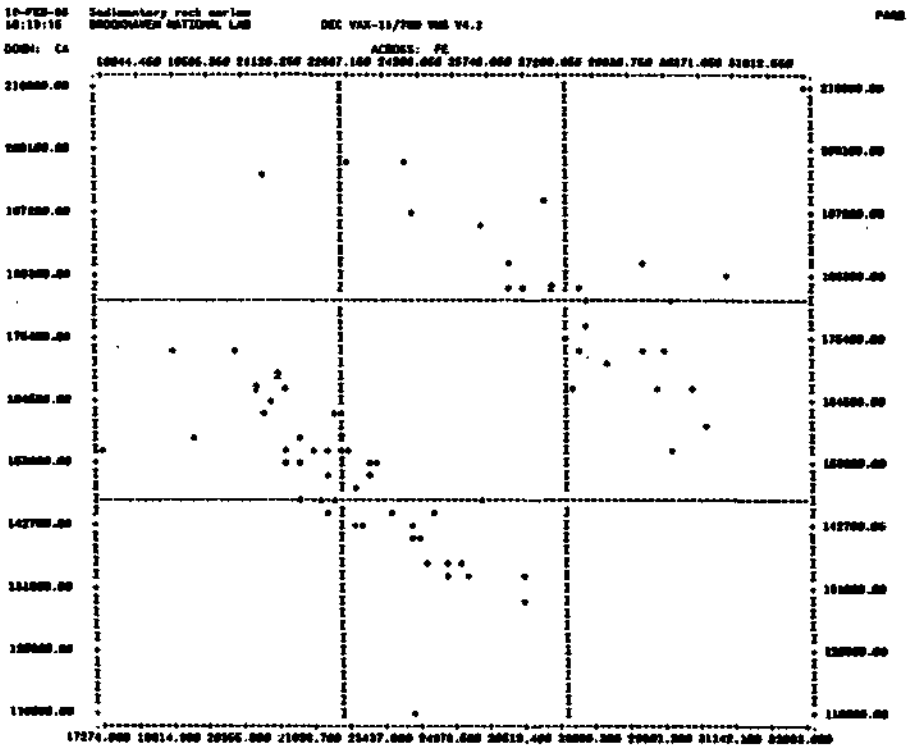

Fig. 4

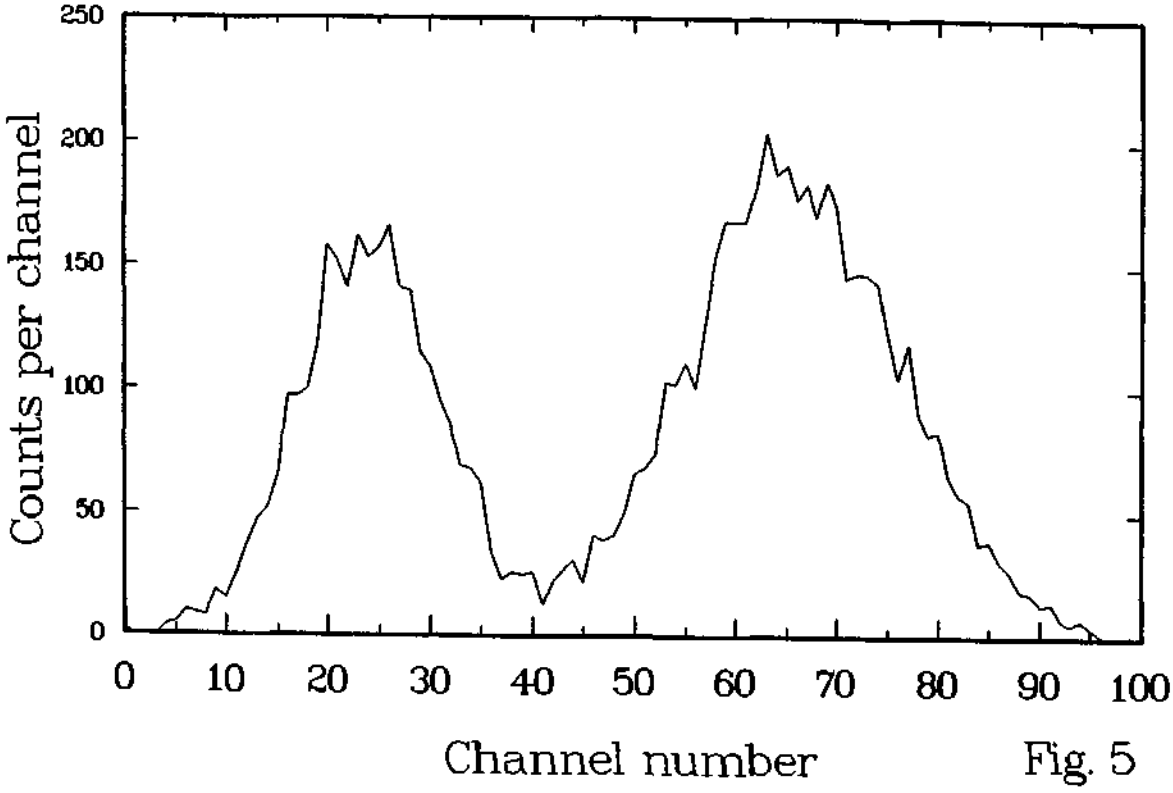
Counts per channel
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Channel number

Fig. 5

Fig. 6

Fig. 7

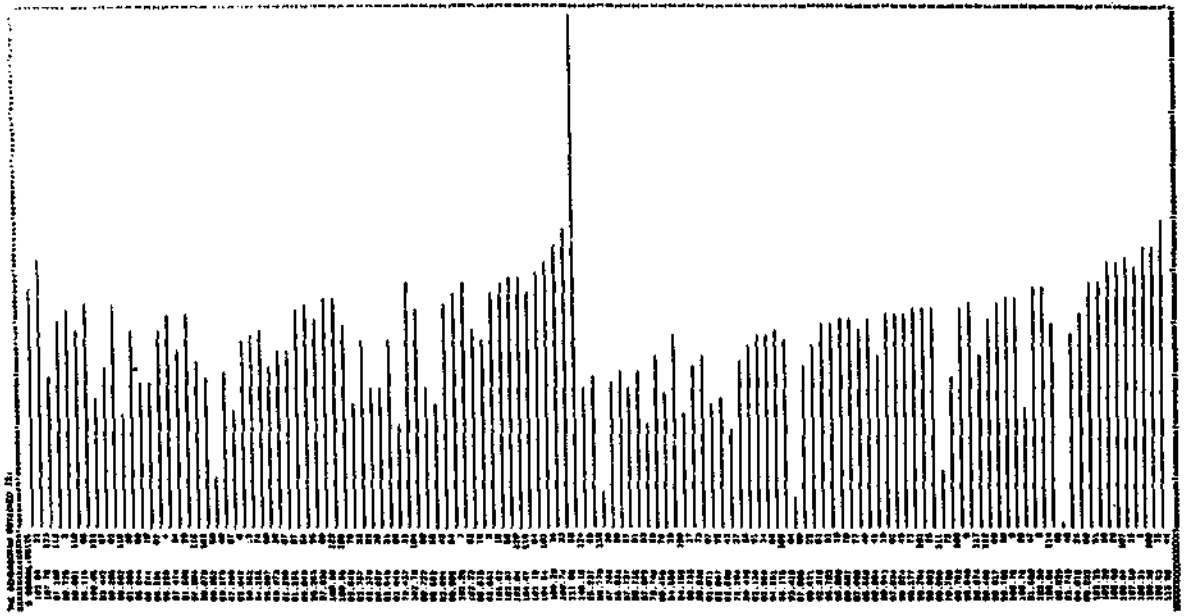

Fig. 8

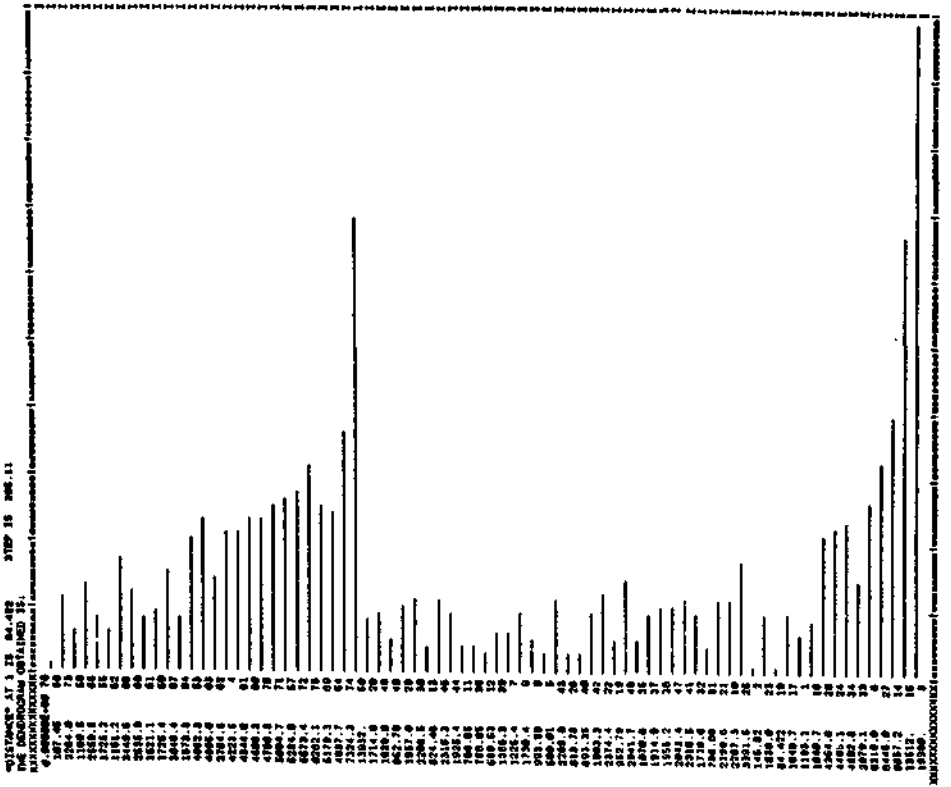

Fig. 9

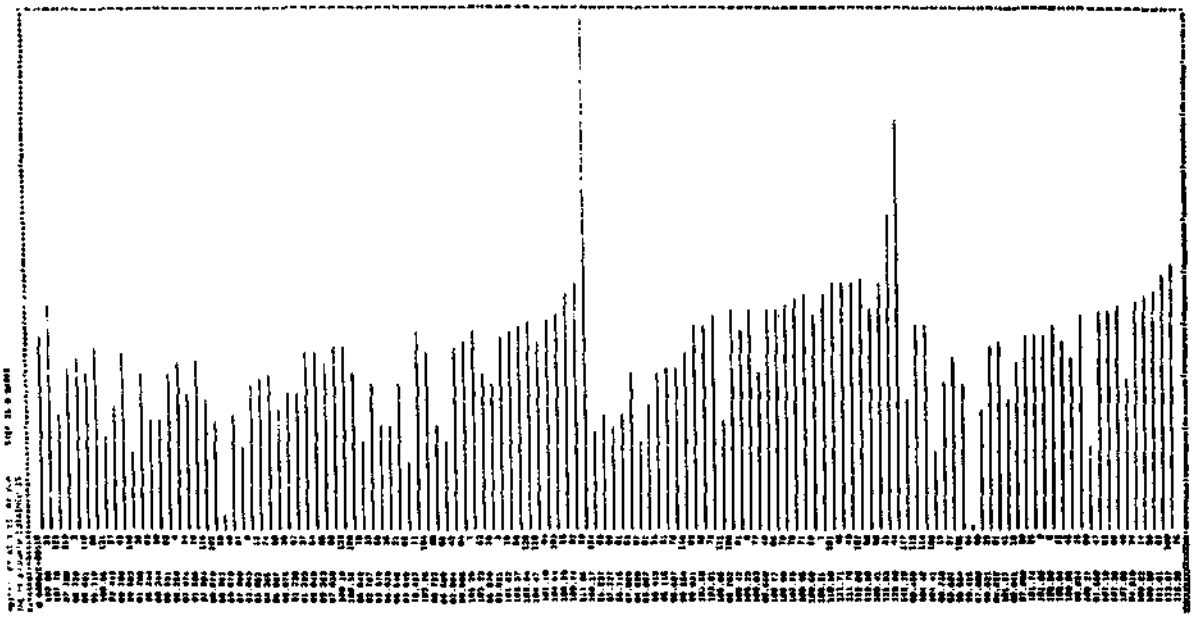

Fig. 10