5-1983

Cluster Analysis for Acid Rain Data in Norway

Ali Ghafourian
Utah State University

CLUSTER ANALYSIS FOR ACID RAIN DATA IN NORWAY

by

Ali Ghafourian

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Mathematics

UTAH STATE UNIVERSITY
Logan, Utah

1983

ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. Robert Gunderson, who through his patience and understanding helped me to accomplish and write this thesis.

I would also like to thank Dr. Robert Heal, who helped me through the years I spent at Utah State University. I also want to thank him for being a special friend to me.

A very special thanks to my wife Sherry, for her help and encouragement throughout my graduate program.

Ali Ghafourian

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Cluster Analysis for Acid Rain Data in Norway

by

Ali Ghafourian, Master of Science

Utah State University, 1983

Major Professor: Dr. Robert Gunderson
Department: Mathematics

This paper gives a description of three well known clustering methods, and discusses the advantages and disadvantages of each. Then, the results of these three clustering methods are compared through examining them on a specific set of data.

(64 pages)

1. INTRODUCTION

Clustering techniques are mathematical tools used for detecting the similarities between different groups of data.

This paper will begin by describing what cluster analysis is and then survey clustering methods in general. This discussion is followed by a more detailed discussion of three particular and important examples of clustering algorithms; single linkage, Ward's method and fuzzy c-means. These three examples of cluster analysis techniques will then be used on a set of data showing the chemical analysis of water samples taken from 80 lakes in Norway. The results of the application will be compared and the advantages and disadvantages of these algorithms will be discussed. The discussion includes finding a "best" number of clusters for this specific data set. Finally, all of the achieved results, such as the "best" cluster number, and some basic chemistry knowledge, will be used in order to identify the lakes in Norway with various degrees of acid rain pollution.

2. CLUSTER ANALYSIS

Classification is one of the fundamental processes in science, in the sense that one of the most primitive and common activities of man consists of sorting like things into categories. Cluster analysis offers many different techniques for discovering the relationships and similarities between a group of data. For example, one can have a body of data units (subjects, persons, cases) that are each described by scores on selected variables (attributes, characteristics, measurements). The objective is to classify these data units into different clusters so that the data that belongs to a cluster has a high degree of "natural association" and at the same time different clusters are "relatively distinct" from each other. It will be observed that there are many different ways to cluster a group of data, so the approach to a problem and the results achieved depend on certain choices made by the person who is doing the clustering. The phrases "natural association" and "relatively distinct," that have been used above, determine the method of clustering that the investigator is trying to use. The following are some practical examples which show where the need for cluster analysis arises in a natural way in several fields of study.

1. Medicine: the principal classification problem in medicine is the classification of disease.

2. Life Sciences: classification is important in the fields of biology, botany, zoology, ecology, and paleontology.

3. Behavioral and Social Sciences: classification is important in psychology, sociology, criminology, anthropology, linguistics, and archaeology.

4. Engineering Sciences: clustering is used in pattern recognition, artificial intelligence, and systems science.

5. Earth Sciences: classification is important in geology, geography, regional studies, soil sciences, and remote sensing.

3. TYPES OF CLUSTERING METHODS

This section will be concerned with a general discussion of the three main categories of clustering methods, as identified by Duda and Hart [1]. In the next section we select three particular methods for further examination and comparison.

3.1 Hierarchical clustering method

This is a sequence of classifications in which larger clusters are obtained through a merger of smaller ones in a nested, or hierarchical method. Because of their simplicity, hierarchical methods are very conceptual, well known, and have a high demand in different fields of science; especially in biological taxonomy where they have a classical application. A detailed discussion of these methods follows below.

Suppose there are n samples and the goal is to partition these n samples into c clusters. The procedure starts by assuming that we have n clusters; that is, we are assuming that each sample by itself makes an individual cluster. The next task is to try to decrease the number of clusters to n-1 and next to n-2 and next to n-3 and so on. It is very easy to see that at the k^{th} stage, the number of clusters will be $c = n-1 + k$. This procedure

will continue until the desired number of clusters is achieved. And, if there is no intention of getting any specific number of clusters, this procedure can go on until c = 1 is achieved; that is, every sample will belong to a single cluster. Throughout this procedure, if any two elements of the sample, say x_1 and x_2, are grouped together at any level of similarity, then they will remain in the same group at all higher levels. Thus, if x_1 and x_2 belong to the k^{th} cluster at some level, then they will remain in that same cluster at any higher level. For this reason, the method of procedure is called the hierarchical clustering method.

As was mentioned earlier, the way two elements are grouped together is based on some sort of similarity measure between data elements, and the investigator is the one who decides what the definition of "similarity" should be.

In order to make this procedure somewhat clearer, suppose that there are seven samples to which the hierarchical clustering method will be applied. For every hierarchical clustering there is a corresponding tree, called a dendogram, that shows how the samples are grouped. Figure 1 shows the tree for our example. The procedure starts at level 1 with seven different clusters $(x_1, x_2, x_3, \ldots, x_7)$, and each cluster contains only one sample. Then at level 2 the samples x_1 and x_7 join together and form a cluster. At level 4 the samples x_3

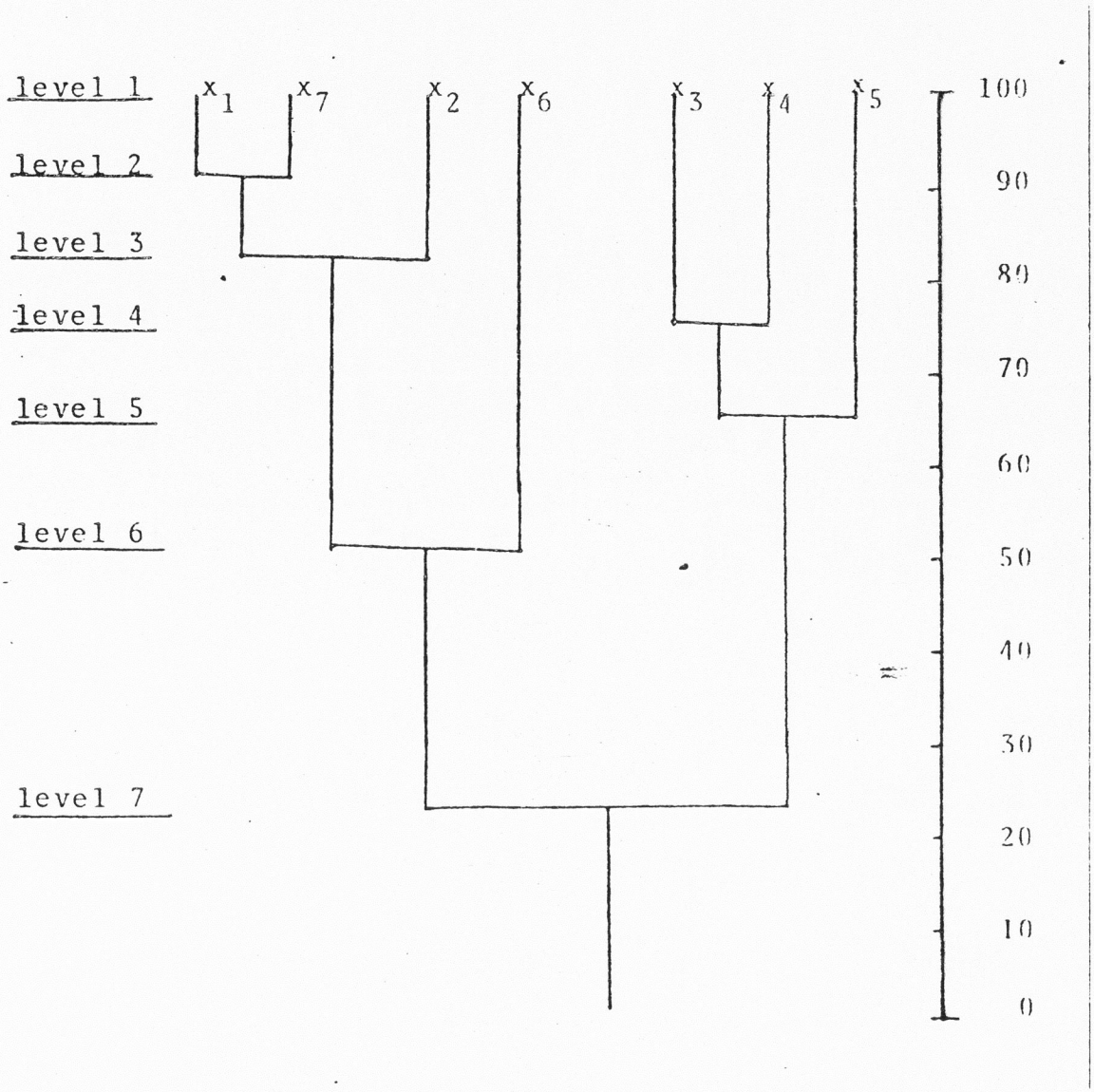


Fig. 1. A dendogram for a hierarchical clustering method.

and x_4 have been grouped to form a cluster. In this method it is usually possible for one to measure the similarity between clusters that are grouped. For example, in Fig. 1 the similarity at level 5 is 73, and at level 7--the last level--the similarity is 20. These similarity values (i.e., 73 and 20) usually come from the mathematical definition of similarity that the investigator uses, and depend on the nature of his or her investigation. The significance of the similarity measure is that it measures the relative distance between clusters and may provide useful insight into the choice of a "best" number of clusters for a given data set.

As was mentioned earlier, the investigator is the one who decides which similarity measures to use, according to the type of problem or research that he is doing. Listed here are some common similarity measures that are used [1].

$$d_{min}(X_i,X_j) = \min_{x\varepsilon X_i,\ x'\varepsilon X_j} || x-x' ||$$

$$d_{max}(X_i,X_j) = \max_{x\varepsilon X_i,\ x'\varepsilon X_j} || x-x' ||$$

$$d_{avg}(X_i,X_j) = \frac{1}{n_i n_j} \sum_{x\varepsilon X_i,\ x'\varepsilon X_j} \sum || x-x' ||$$

$$d_{mean}(X_i,X_j) = || m_i-m_j || \text{ where}$$

$$X_i = \{x_i\}\ ,\ i = 1, \ldots, n$$

Hierarchical clustering itself divides into distinct methods, with some of the most commonly used of the methods being the single linkage method where d_{min} is used, and the complete linkage where d_{max} is used, and the Ward's method [1], which is, in fact, a hybrid hierarchical-objective function method. The single linkage and Ward's methods are discussed in more detail below. Later on these two methods will be applied to some specific data, and the results compared, in order to discuss the advantages and disadvantages of each.

3.2 Graph theoretical

Notice that in Fig. 2 there are eight samples which are connected together by straight lines. The graph theoretical methods regard the nodes as the set of samples. Edge weights between pairs of nodes can be based on a measure of similarity between pairs of nodes. That is, a threshold distance d_0 is selected and two points are said to be in the same cluster if the distance between them is less than d_0. This procedure can easily be generalized to apply to arbitrary similarity measures. So, one can talk about a clustering strategy in this method as the "connectivity" between the nodes.

The graph theoretical method is highly adaptable to a data with "chains." However, if there exists a mixed data structure there will usually be a lot of trouble

because of the chaining tendencies of this method, such as is shown in Fig. 3a and 3b, which are badly distorted [2].

Since graph theoretical methods are not particularly popular, we will not discuss them in detail in this report. For further information refer to [3].

3.3 Objective function clustering

As was mentioned earlier, cluster analysis is one of the basic tools for identifying structure in data. For a given set of data, which consists of n samples $x_1,\ldots,x_n$, we want to partition these samples into c disjoint clusters such that the samples in the same clusters are more similar than the samples in different clusters. Objective function methods measure the clustering quality of any partition of the data, so that one is trying to find the cluster by minimizing some objective function.

Before going into more detail about objective function clustering, we shall first discuss the differences between "hard" and "fuzzy" clustering. Suppose that from the set of all people in the United States we want to locate the cluster of tall people. Suppose further that one chooses a method of hard clustering to do the task. In this instance, assume that everyone who is over 6 feet tall belongs to the cluster of tall people. One either belongs to the cluster or one does not. The disadvantage of this

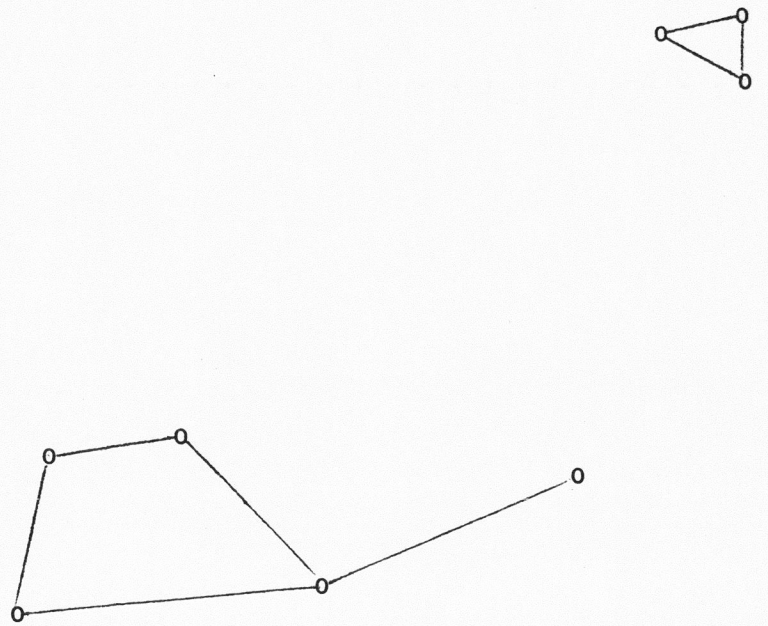
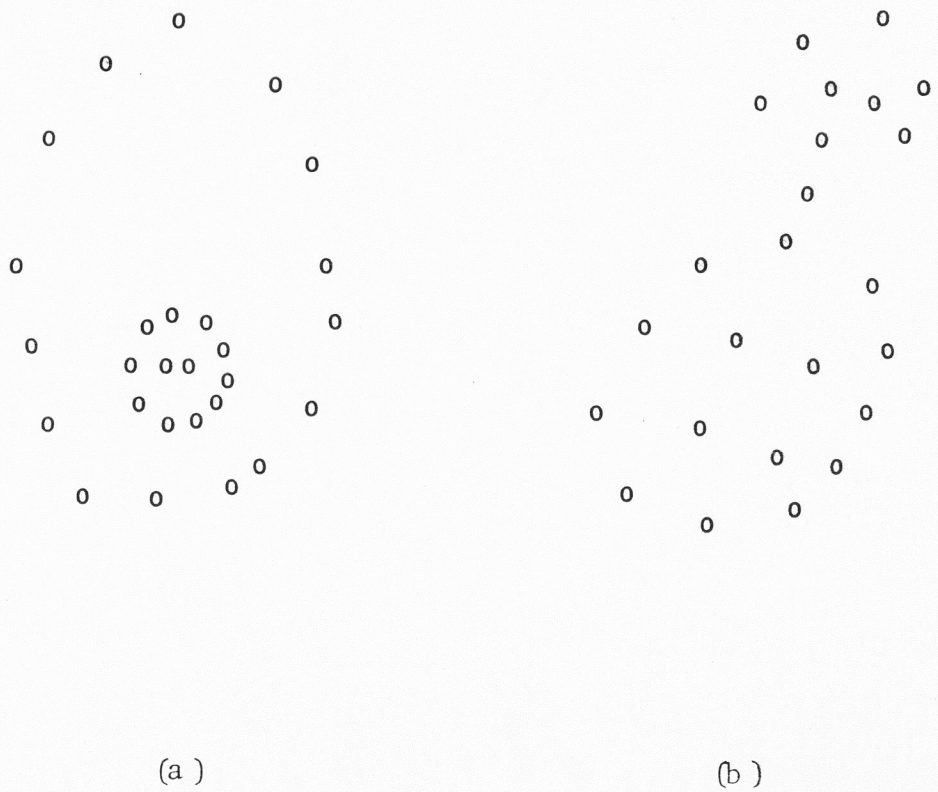


Fig. 2. Sample data for the graph theoretical method.



(a) (b)

Fig. 3. Examples of mixed data structure.

method is that someone who measures 5 feet 11-9/10 inches would be left out. The fuzzy clusterer reasons that a person who is 6 feet tall is certainly more likely to be considered as "tall" than one who is 5 feet tall. However, a person who is 5 feet 11-9/10 inches tall is also a very good candidate for the cluster of tall people. Therefore, rather than classifying people as being tall or not, everyone is given a membership value that describes how close he or she is to the cluster center. For instance, someone who is 6 feet tall may get a membership value of 0.9, while the person who is only 5 feet tall may get a membership of only 0.1.

To be more precise, let $X = (x_1,\ldots,x_n)$ be a finite set of n objects where each object is described by some number of features, f. Thus, X is a set of n vectors in R^f, which represents our data. Let c denote the number of clusters ($2 \leq c \leq n$). If we want to partition $X = (x_1,\ldots,x_n)$ into c disjoint subsets, call it $S_1,\ldots,S_c$, the procedure is called hard clustering. This process can be formalized by defining the "hard" membership functions

$$U_i: \quad X \rightarrow \{ 0,1 \} \qquad i = 1, \ldots, c$$

by $U_i(x_k) = 1$ if x_k is an element of the i^{th} cluster

$U_i(x_k) = 0$ otherwise

for all $c = 1, 2,\ldots,c$.

On the other hand, in fuzzy clustering procedures we assign <u>any</u> value <u>between</u> "0" and "1" to a sample of the data set to generate new membership functions $U_1,\ldots,U_c$ according to the definition

$$U_i: \quad X \rightarrow [\ 0,1\ ] \qquad i = 1,2,\ldots,c$$

where

$$\sum_{i=1}^{c} U_i(x_k) = 1$$

and

$$x_k \in X.$$

For example, if $U_3(x_k) = .8$, then the sample belongs more to cluster number 3 than to any other cluster. It is easy to see that an object may belong to several clusters at the same time. That is, it is possible to have many joint clusters.

Now that we are more familiar with the notion of hard and fuzzy clustering, we return to objective function clustering. Recall that the problem is one of finding the partition that minimizes an objective function. In this section we introduce the characteristics of several basically similar objective functions.

<u>The sum-of-squared-error function</u>. Perhaps the sum-of-squared-error function is the simplest and most widely used of all objective function clustering methods. Define m_i to be the "mean" of samples,

$$m_i = \sum_{k=1}^{n} U_{ik} x_k \Big/ \sum_{k=1}^{n} U_{ik}$$

where $U_{ik} = U_i(x_k)$, then we can define the sume of the squared error by

$$J_e = \sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik} \, || \, x_k - m_i \, ||^2$$

Thus, J_e measures the total squared error incurred in representing the n samples $x_1,\ldots,x_n$ by the c cluster centers $m_1,\ldots,m_c$ [1].

If, in the objective function J_e, we assign the value of either "1" or "0" to U_{ik}, then the result is the method of minimum variance discussed by Duda and Hart [1]. However, if we assign any value between "0" and "1" to U_{ik}, then the result is the fuzzy c-means clustering method developed by Bezdek [2]. This method is discussed in more detail in the next section.

Related minimum variance objective function. It can be easily shown that

$$J_e = 1/2 \sum_{i=1}^{c} n_i \bar{s}_i$$

$$n_i = \sum_{k=1}^{n} U_{ik}$$

where

$$\bar{s}_i = \frac{1}{n_i^2} \sum_{x \varepsilon X_i} \sum_{x' \varepsilon X_j} || \, x - x' \, ||^2$$

therefore, $\bar{S}_i$ is just the average squared distance between the points in the i^{th} cluster. One can replace $\bar{S}_i$ by the average or the median, and thereby generate additional functions to work with. For instance $\bar{S}_i$ can be replaced by

$$\bar{S}_i = \frac{1}{n_i^2} \sum_{x \in X_i} \sum_{x' \in X_j} S(x,x')$$

or

$$\bar{S}_i = \min_{x,\ x \in X_i} S(x,x')$$

where S_i is some similarity function.

Scattering objective functions. The scattering objective functions divide into several different classes such as the trace objective function, the determinant objective function and invariant objective functions. More detail is given about each of these methods in the following paragraph.

One usually uses scatter matrices when doing multiple discriminant analysis. The following definitions are found in Duda and Hart [1].

$$m_i = \frac{1}{n_i} \sum_{x \in X_i} x \quad \text{(mean vector for the } i^{th} \text{ cluster)}$$

$$m = \frac{1}{n} \sum_{X} x = \frac{1}{n} \sum_{i=1}^{c} n_i m_i \quad \text{(total mean vector)}$$

$$S_i = \sum_{x \in X_i} (x-m_i)(x-m_i)^t \quad \text{(scatter matrix for } i^{th} \text{ cluster)}$$

$$S_w = \sum_{i=1}^{c} S_i \quad \text{(within-cluster scatter matrix)}$$

$$S_B = \sum_{i=1}^{c} n_i (m_i-m)(m_i-m)^t \quad \text{(between-cluster scatter matrix)}$$

$$S_T = \sum_{x \in X} (x-m)(x-m)^t \quad \text{(total scatter matrix)}$$

and

$$S_T = S_W + S_B$$

The investigator usually tries to minimize S_W or maximize S_B [1].

For the trace objective function, the trace of S_W is the one to be minimized and is defined as follows [1].

$$t_r(S_W) = \sum_{i=1}^{c} t_r S_i = \sum_{i=1}^{c} \sum_{x \in X_i} ||x-m_i||^2 = J_e$$

For the determinant objective function we minimize

$$J_d = |S_W| = \left| \sum_{i=1}^{c} S_i \right|$$

as the objective function [1].

One might elect to maximize the objective function

$$t_r(S_W)^{-1} S_B = \sum_{i=1}^{d} \lambda_i$$

or the invariant objective function where λ's are the eigen-values of $S_W^{-1} S_B$ [1].

4. TWO HIERARCHICAL CLUSTERING ALGORITHMS

4.1 Single linkage

Single linkage is a very well known method in hierarchical clustering. As was mentioned earlier, a device is needed to measure the similarity between the objects. Consider the behavior when d_{min} is used as our device for a similarity measurement.

Note that in Fig. 4 data points are used as nodes of a graph and straight lines are used to form a path between the nodes. This path will be called subset X_i. At this stage we need to find the nearest subset, and our device to do so is d_{min} which measures the distance between the subsets. Now, by adding an edge between the nearest pairs of nodes in X_i and X_j, the merging between the subsets X_i and X_j is determined. Looking at Fig. 4, notice that there are no closed loops or circuits. The reason is edges linking clusters always go between distinct clusters. What we have in Fig. 4, as a result of this procedure, is called a tree. If we were to continue this procedure indefinitely, we would get a spanning tree, which means that all the subsets would link together.

Figure 4a represents two obvious separate clusters,

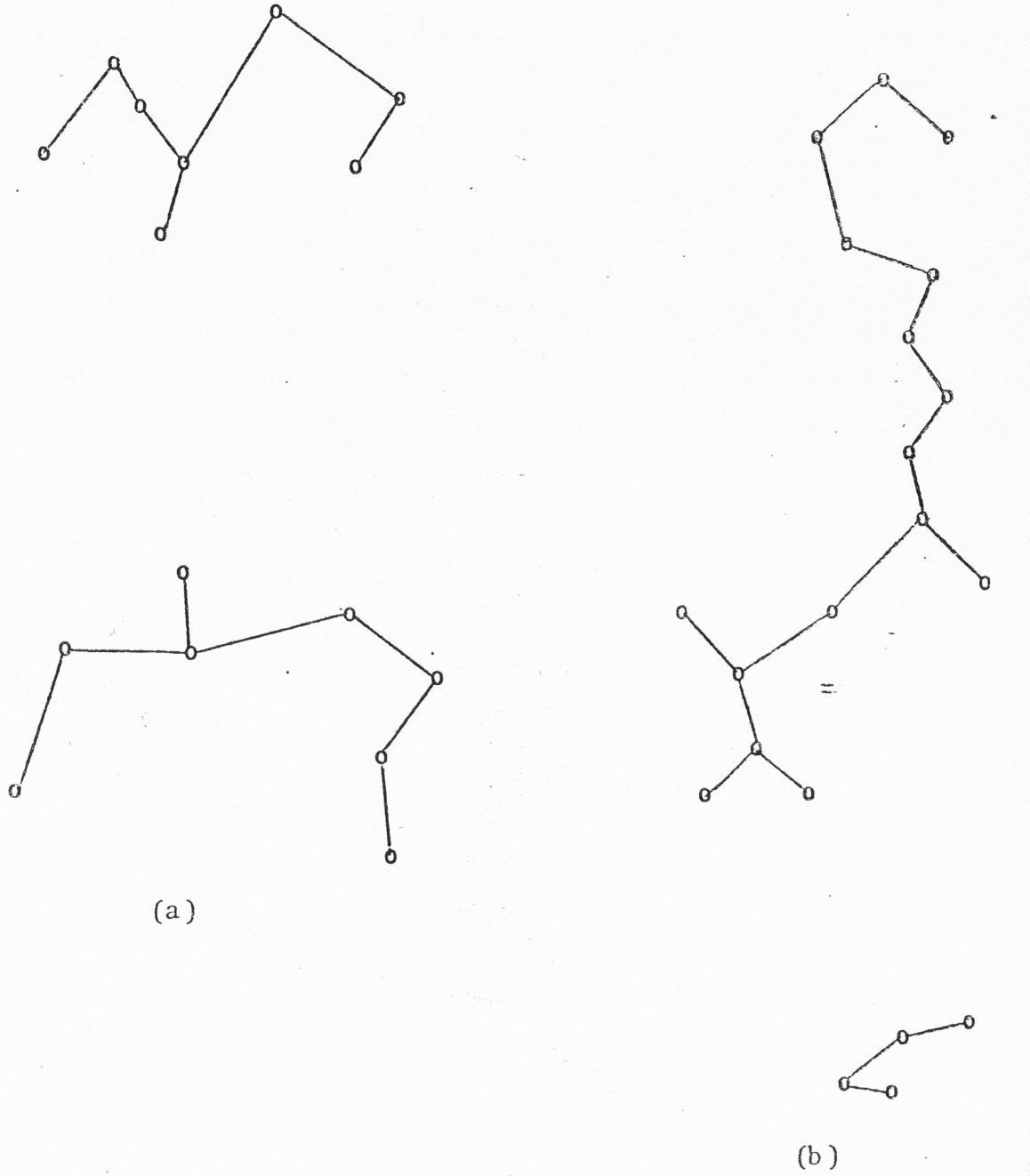


Fig. 4. Results of the single linkage algorithm using d_{min} on artificial two dimensional data.

where the data are very well separated and the computed clusters similar in size. However, in Fig. 4b, the extra data present cause a very significant difference in the clustering structure. As we can see in this figure, there is one small and one large cluster. We used c = 2 as the number of clusters wanted in both a and b.

The preceding procedure is known as the nearest-neighbor or minimum algorithm, with single linkage being the specific method used. In this instance we stopped the process for an arbitrary threshold, if the distance between the nearest clusters exceeds the threshold value then the process stops.

4.2 Ward's method

As was mentioned earlier Ward's method is a general and widely used method of hierarchical clustering. While it can be called a hierarchical method, it shows the major features of the objective function methods in that this method chooses the points of merger at each stage so that an objective function is maximized according to the purpose of the investigator in a particular problem.

Define the following quantities:

X_{ijk} = score on the i^{th} of n variables for the j^{th} of mk data units in the k^{th} of h clusters.

$\bar{X}_{ik} = \sum_{j=1}^{mk} X_{ijk}/mk$ be the mean on the i^{th} variable for data units in the k^{th} cluster.

$$E_k = \sum_{i=1}^{n} \sum_{j=1}^{mk} (X_{ijk} - \bar{X}_{ik})^2 = \sum_{i=1}^{n} \sum_{j=1}^{mk} X_{ijk}^2 - mk \sum_{i=1}^{n} \bar{X}_{ik}^2$$

Thus, E_k is the error sum of squares for cluster k; so what we really have is the sum of the euclidian distance from each data point in cluster k to the mean of the very same cluster [4]. Since for every k we have an E_k, denote the total error sum of squares for the collection of clusters by

$$E = \sum_{k=1}^{h} E_k$$

At this point we want to find two clusters such that when they merge we have the minimum increase in the error sum of the squares.

The increase in E for two cluster p and q, if we denote it by E_{pq} is:

$$E_{pq} = E_t - E_p - E_q$$

where t is the resulting cluster after clusters p and q have merged.

It can be verified that

$$E_{pq} = \frac{m_p m_q}{m_p + m_q} \sum_{i=1}^{n} (\bar{X}_{ip} - \bar{X}_{iq})^2 \text{ [1].}$$

So the above equation gives the increase in the error sum of the square due to the merger of cluster p and q.

A computational consideration that is very important for an investigator involved in large problems is the accumulation of round-off error. To put the equation into a form less sensitive to round-off error, we need to define some additional quantities:

$$T_{ik} = \sum_{j=1}^{mk} X_{ijk} = m_k \bar{X}_{ik}$$

be the total of scores on the j^{th} variable for data units in the k^{th} cluster.

$$S_k = \sum_{i=1}^{n} \sum_{j=1}^{mk} X_{ijk}^2$$

be the sum of squared scores on all variables for all data units in the k^{th} cluster. Then, E_k can be written as

$$E_k = S_k - \sum_{i=1}^{n} T_{ik}^2/mk$$

Also it was shown that

$$E_{pq} = E_t - E_p - E_q$$

and it can be verified that

$$E_{pq} = S_p+S_q - \sum_{i=1}^{n} (T_{ip}+T_{iq})^2/(m_p+m_q) - E_p-E_q$$

where

$$m_t = m_p+m_q, \quad S_t = S_p+S_q, \quad \text{and } T_{it} = T_{ip}+T_{iq}$$

thus, E_{pq} can be written in terms of S_p, S_q, T_{ip}, and T_{iq}.

The above expression can easily be reduced further, or put into many other forms; however, accomulating S_k, E_k, M_k and $\{T_{ik}: \ i=1,\ldots,n\}$ for each cluster primarily involves simple addition and avoids the repeated multiplication and division required when using cluster means.

5. AN OBJECTIVE FUNCTION METHOD

5.1 Fuzzy c-means

Among the various types of objective function clustering algorithms, perhaps fuzzy c-means is the most widely used one of all. This method chooses the sum-of-squared-error as the objective function to minimize.

Recall from the previous sections that U_i was denoted as a membership function, and we assigned a value between $\underline{0}$ and $\underline{1}$ to each function. That is:

$$U_i: \quad X \to [0,1] \quad \text{for all } i=1,\ldots,c$$

such that

$$\sum_{i=1}^{c} U_i(x_k) = 1 \qquad \text{for all } k=1,\ldots,n$$

Denote $v_i \in R^f$ for $i=1,\ldots,c$ as the mean of the data vectors in S_i. We use the fuzzy c-means algorithm to produce a fuzzy clustering of the data set. The vectors $v_1,\ldots,v_c$ are also called the center of the clusters. Define U_{ik} to be the value of the i^{th} membership function on the k^{th} data point x_k. We would like to measure the similarity between the objects by the distance between data vectors in such a way that, if the cluster centers and membership functions are chosen so that if we have a datapoint close to the corresponding cluster center,

has a high membership value. The fuzzy c-means algorithm produces c fuzzy clusters so that for any real number $m > 1$, it finds a membership matrix $U = [U_{ik}]$ and cluster centers $V = (v_1,\ldots,v_c)$ to minimize the objective function

$$J(U,V) = \sum_{i,k} (U_{ik})^m \, ||x_k - v_i||^2 \qquad (1)$$

Therefore, we minimize the distance between the k^{th} data point to its corresponding cluster center.

Using LaGrange multipliers on J with the constraint that $\sum_{i,k} U_{ik} = 1$, we can easily obtain the necessary conditions for a local minimum as follows:

$$v_i = \left[\sum_k (U_{ik})^m \, x_k \right] / \sum_k (U_{ik})^m \qquad (2)$$

$$U_{ik} = \sum_i \left(1/|x_k - v_i|^2\right)^{1/(m-1)} / \sum_j \left(1/|x_k - v_j|^2\right)^{1/(m-1)} \qquad (3)$$

As was mentioned, m is any real number greater than 1 (requirement for LaGrange multiplier method), and is called the exponent weight. Using values of $m \gg 1$ in the algorithm results in minimizing the effect of those data points whose membership values are uniformly low. In other words, those data points do not play as significant a role in determining the cluster centers and membership functions.

Here are the necessary steps that one should take when using the fuzzy c-means algorithm. First, choose

a value for c and m. Next, guess the initial membership matrix U, which is a c by n matrix, and then compute the cluster centers using the membership and equation (2). By using equation (3), recompute memberships and cluster centers. Last, compare the successive membership matrices. The procedure can be stopped at any point, depending of course on some prescribed value; the value by which the cluster centers in successive iteration differ.

6. ADVANTAGES AND DISADVANTAGES OF THE ALGORITHMS

6.1 Single linkage

Advantages. The first advantage is that this method can easily be followed by persons who may not be very familiar with mathematics. This algorithm does not involve very sophisticated mathematical equations.

The second advantage is that this clustering method is relatively inexpensive. The computer algorithm is simple, when compared to many other algorithms.

Thirdly, this algorithm, as was mentioned before, produces a tree where one can actually see where two objects link or join together at a certain level of similarity.

Lastly, often it happens that the data set may have a few data points that are distant from the majority of the data (Fig. 5).

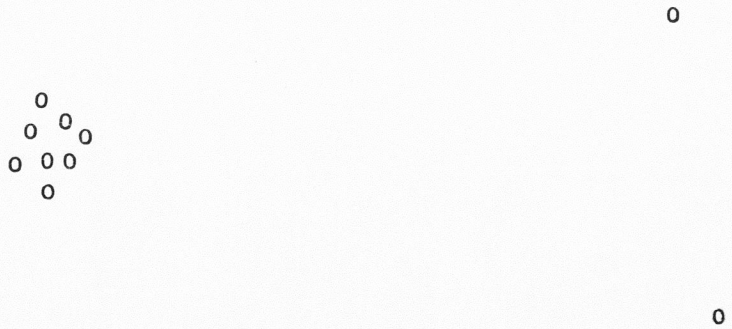


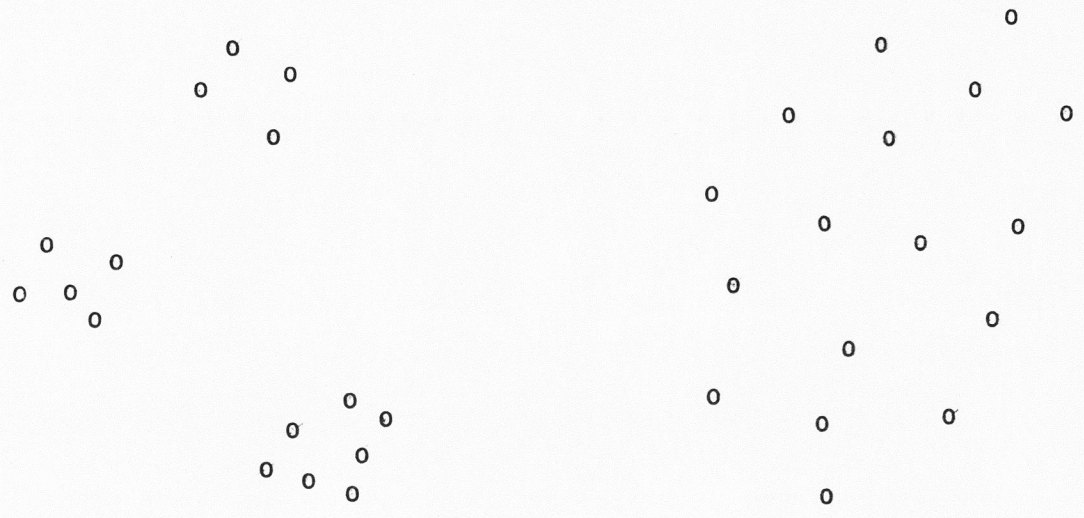
Fig. 5. Illustrated example of "outliers."

If it is possible to detect these points, known as "outliers," they might be omitted from the data set, making it easier to find the "best" number of clusters for the data. The advantage of the single linkage method is that it is relatively easy to locate such points by looking at the dendogram.

Disadvantages. This clustering method is not a good method for sets of data that are fuzzy, as opposed to data that are well-separated (Fig. 6). If well-separated data are used with the single linkage method, the result is a good clustering. Otherwise, the result may be misleading. Unfortunately, one usually does not know ahead of time whether the data is well-separated or not.

When this clustering method is used, there is no precise device (mathematical formula) to determine the best number for c. In other words, we don't know what a "best" number of clusterings for a given set of data would be.

Another disadvantage of this method is the effect of "linking" data points. In order to demonstrate this effect, consider the following figure (Fig. 7). The data structure in Fig. 7 is referred to as hybrid points, and the solid line around the data points show the clustering obtained when the single linkage algorithm is applied. In this case however, the results may not be the best clustering for the data.



(a) well-separated (b) fuzzy

Fig. 6. Fuzzy data versus well-separated data.

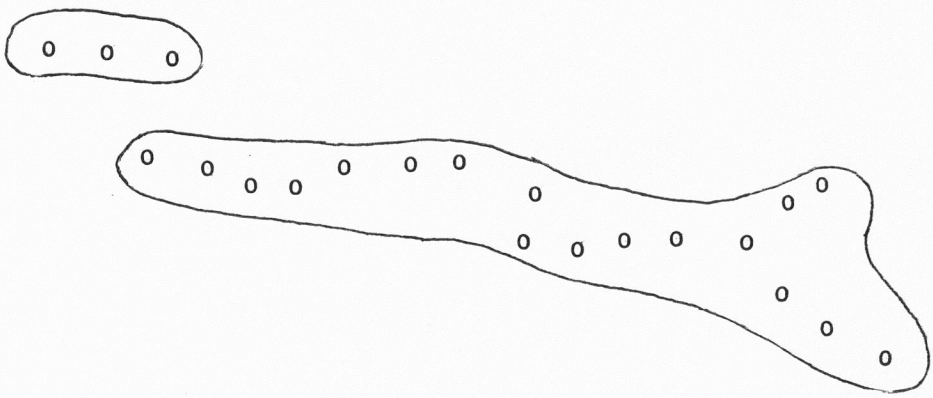
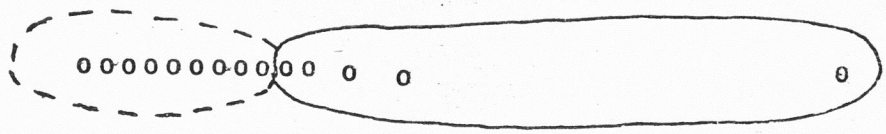


Fig. 7. Results of the single linkage algorithm applied to "linking" data points.

6.2 Ward's method

As was mentioned, Ward's method is another type of hierarchical clustering method. The advantages and disadvantages are somewhat similar to those of the single linkage method. However, because Ward's method has some of the same features of objective function clustering, it does not have the advantage of spotting "outliers." Figure 8 demonstrates this point by illustrating how the Ward's method may cluster a data set containing "outliers." Where the single linkage method does not give a good clustering when applied to "linking" data (refer to Fig. 7), if the Ward's method is used, the results are much better.



1st cluster 2nd cluster

Fig. 8. Clustering obtained when Ward's method is applied to data with "outliers."

6.3 Fuzzy c-means

Advantages. The main advantage is that this method takes into account the effect of "fuzzy" data. Hybrid points are recognized as such and have a corresponding effect on the final cluster configuration. In addition, the final membership values are of great practical value in interpreting the significance and meaning of the final cluster configuration.

Suppose there is a set of data that looks like the data in Fig. 9. If the fuzzy c-means algorithm is applied to these data, the point x in between will not belong to either cluster. For instance, if one chooses c = 2 as the number of clusters, then the point x, if exactly located between the two groups of data, will have the membership value of 0.5 for each cluster. However, if the single linkage algorithm is applied, x only belongs to one or the other of the clusters. Therefore, it appears that the point x has no similarity whatsoever with one of the clusters, and this is not a very accurate result.

Another advantage of this method of clustering is that the investigator has some control over the number of clusters. Before even starting this procedure, one must choose what to use as a cluster number (c = n where n is greater than 2). Then it is possible to see what objects belong to what clusters for different numbers of

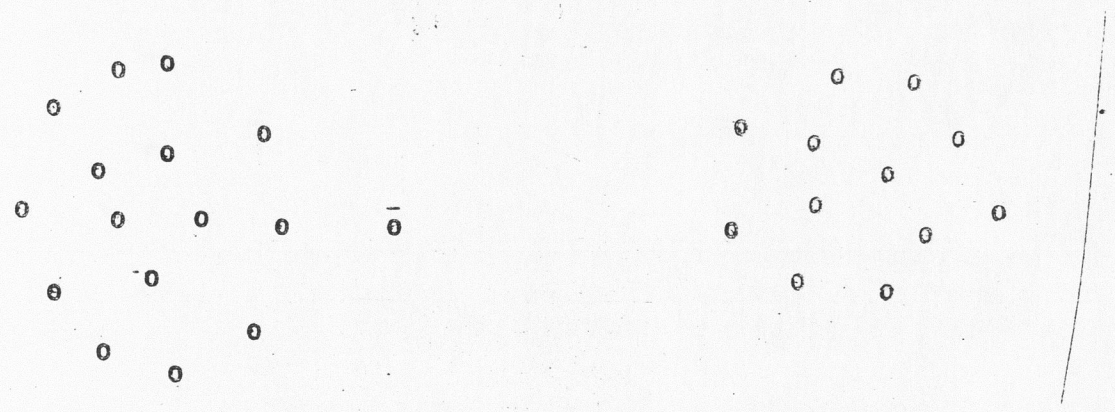


Fig. 9. Shared membership example.

c, and also what objects belong to which clusters based on the degree of membership value. Thus, by trying the various numbers for c, one may eventually find a good choice for the number of clusters for a specific problem.

It should be emphasized that the relative "fuzziness" of the final cluster configuration provides a possible measure of the "goodness" of that configuration. Thus, fuzzy clustering methods provide an opportunity for a mathematical solution to the cluster validity problem. While this topic is beyond the scope of this report, the reader is referred to the book by Bezdek [2] and the papers by Windham [5] for more details.

Consider the following two dimensional data picture in Fig. 10. If the fuzzy c-means algorithm is applied to this data using c = 3, we will probably get a very

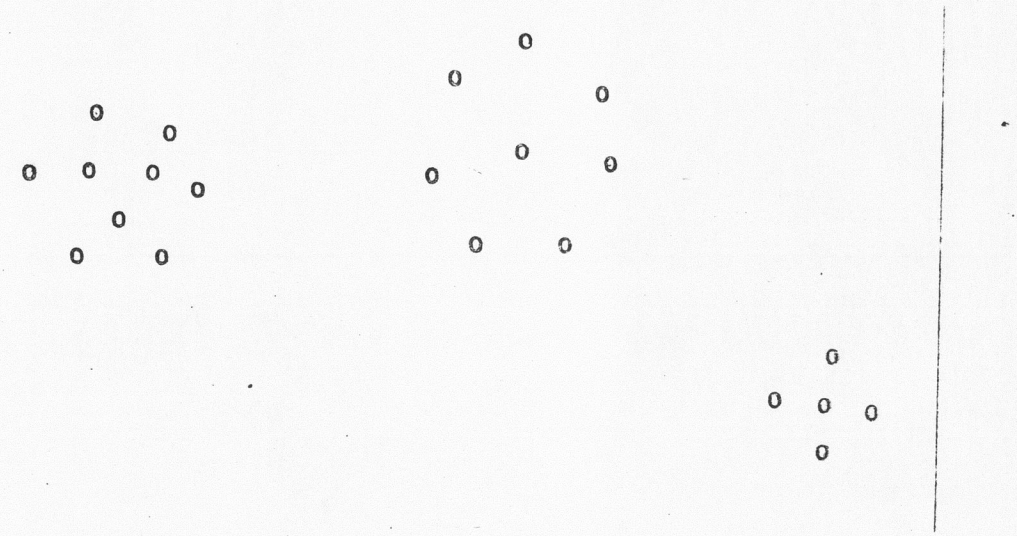


Fig. 10. Round-shaped data.

result. However, even if we don't start with $c = 3$, through the effectiveness of cluster validity, $c = 3$, which is the best number for clustering the data, can easily be achieved.

Disadvantages. The fuzzy c-means algorithm can be considered an expensive method of clustering relative to the cost of other types of clustering, such as the single linkage or Ward's methods.

The worst disadvantage of this method is that it is only good when there are round shaped data, although it is possible to modify it to detect other shapes [2]. It is not possible to modify the other methods of this report, which share this common disadvantage. Recall that

the fuzzy c-means algorithm chooses subsets $S_1,\ldots,S_c$ of our data set X, which minimizes

$$J_e = \sum_i \sum_{x \varepsilon s:} ||x - v_i||^2$$

where $v_i \varepsilon R^f$ is the mean of the data vectors in S_i. In other words, it can be said that this algorithm chooses the clusters to minimize the distance of the points in the clusters to the centers of the clusters. If there are several points in the clusters whose distances are close to the center of the corresponding cluster, then these points are naturally close together, which means that they are strongly belonging to that cluster.

Next, consider another two dimensional data set which is pictured in Fig. 11. It is easy to see that the obvious number for clustering the data in Fig. 11 is two. We can achieve this result using the single linkage method. However, if the fuzzy c-means algorithm is applied to this set of data with c = 2, the results will not be as good. The fuzzy c-means algorithm will cluster the data into two clusters as is illustrated in Fig. 11. In conclusion, the results obtained from the fuzzy c-means algorithm when using non-round-shaped data are not good, unless the clusters are fairly well-separated.

Another disadvantage of this method of clustering is that for someone who is not familiar with mathematical concepts, the following and understanding of this method

is relatively harder than in any of the other methods; for example single linkage. Since, in general, the concept of clustering is a widely used technique in all fields, people who are not familiar with mathematics may need to use it, but may not understand the complex mathematical techniques involved with this method.

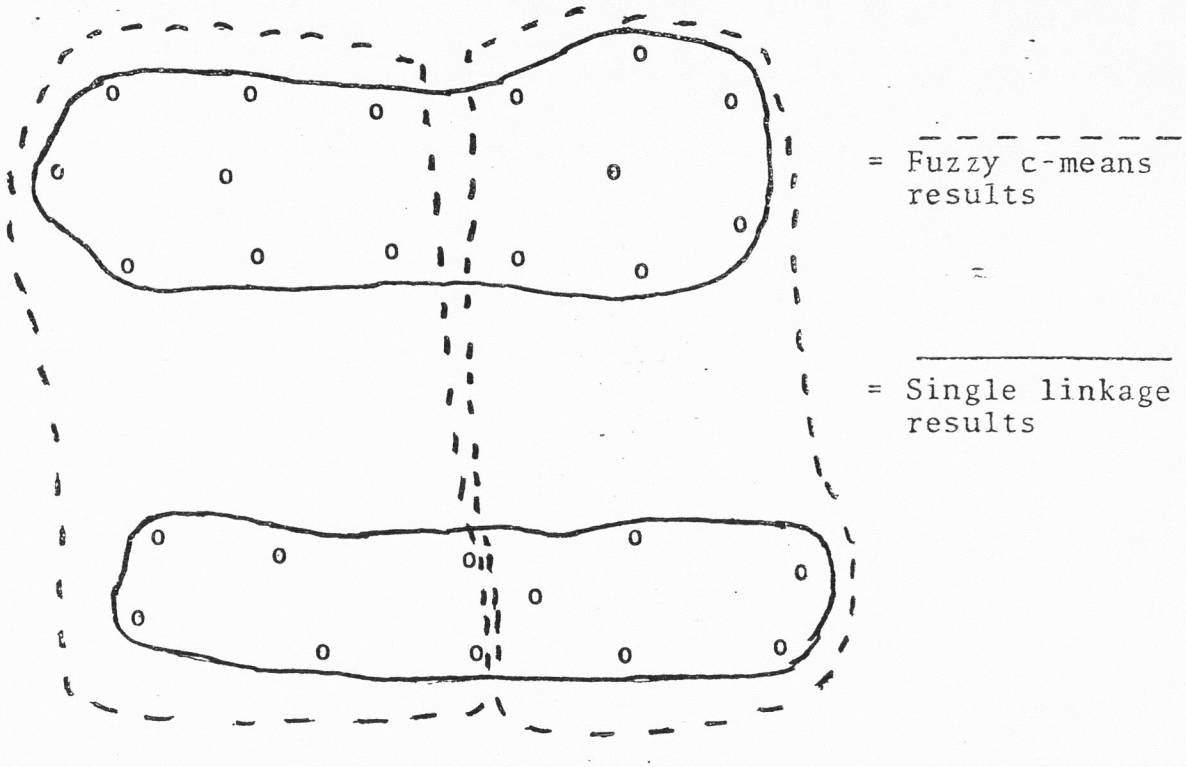


Fig. 11. Single linkage versus fuzzy c-means on artificial data.

7. APPLICATION TO ACID RAIN DATA IN NORWAY

7.1 The data

Disappearance of valuable fish population in the lakes of the southern part of Norway brought to attention the possible pollution in the water of the lakes. This turned out to be a result of a high amount of acid in the lake waters. In Norway, acid precipitation was, at that time, seen as a possible cause of the increasing acidity of the water sources in the southern part of the country. There was an assumption that the acid rain was originating in the industrialized part of Europe. There was also the possibility that over a long period of time, the penetration of the acid rain could cause changes in the soil and therefore a reduction in forest growth.

In all, about 150 lakes were sampled during October-November, 1974. The southern part of Norway was divided into square blocks. Preferably the lakes and watershed should be pristine with no major disturbances by agriculture, siviculture, or lake level regulation, and the lake should be situated at the head of the drainage basin. Water samples were collected at 0.5 meter depth and 2 meters above the bottom. Analysis was carried out on rain ions (H^+, Na, K, Ca, Mg, Al, SO_4, Cl, NO_3), and the

survey was repeated in the years 1975-1978 using only the water chemistry. Control samples proved that the data was representative for the area (refer to Table 2).

7.2 Analysis of clustering results and comparison of the methods

The purpose of this section is to discuss the results of the computer's output when different types of algorithms are applied to the acid rain data, and to try and find the "best" number of clusters for this data.

The results obtained when the single linkage method was applied will be discussed first. Since this method does not work well when there are fuzzy data, as was discussed in previous sections, naturally the computer's output does not say very much about the number of clusters. It is very difficult to find a reasonable number of clusters using this algorithm. In order for the reader to become more familiar with the output of this algorithm, a copy of the dendogram (spanning tree) is submitted with this report. By looking at this the reader can see that the output does not specify what the cluster number for this algorithm should be.

During the investigation, the data were normalized according to

$$x \rightarrow \log(x+1)$$

However, the results were unchanged, and the output

was almost the same. Finding a good cluster number was still almost impossible.

Next, Ward's method was applied to this data. The interesting thing about the output of this algorithm is that it can be easily studied, and finding a good cluster number is simple. Through looking at the dendogram, the number "4" was chosen as the cluster number for the output. Then the results were compared with those of the fuzzy c-means algorithm using c=4, c=5, c=6 applied to the same data.

The best way of comparing results is to plot each different algorithm output on the five different maps of Norway. Different numbers on the maps indicate different clusters (all areas marked "1" indicate one cluster, and so on).

An interesting fact about the five outputs is that all of them show that the lakes in the southern part of Norway are all clustered together (Figs. 12, 13, 14, 15, and 16). This indicates that there is uniform pollution in the southern part of Norway. The output is as follows:

Ward's method (1st cluster)

Lake number: 84, 17, 85, 10, 5, 81, 1, 80, 2, 77

Fuzzy c-means (1st cluster) c=4

Lake number: 84, 17, 85, 10, 5, 81, 1, 80, 2, 77

Fuzzy c-means (1st cluster) c=5

Lake number: 84, 17, 85, 10, 5, 1, 80, 2, 77

Fuzzy c-means (1st cluster) c=6

Lake number: 84, 17, 85, 10, 5, 81, 1, 80, 2, 77, 4, 6, 82, 14, 92

Ward's method--normalized data (1st cluster)

Lake number: 84, 17, 85, 10, 5, 81, 1, 80, 2, 77

The lakes included in the first cluster for the different algorithms are designated as "1' on the different maps of Norway. Figure 17 shows a map of Norway with the lakes numbered for reference purposes.

Notice that the results of the first cluster for the different algorithms is very much the same. Another common thing about the different output is that the lakes in the western part of Norway are also clustered together.

7.3 Conclusions with respect to methods and data

At this point let us forget about clustering and consider the overall pollution in Norway using the data on hand. By referring to the data, one can see that there are several ionized substances in the lake water. For example, the amount of H^+ in lake number one is 19.5, and the amount of NO_3 in lake number seventeen is 3.6. However, there are positive and negative ions, for instance "H" is positive chile "Cl" is negative. According to Utah State University's Chemistry Department, in order to have a really polluted lake, the sum of all the positive

ions should be equal to the sum of all the negative ions. The more the amount of positive ions equals the amount of negative ions, the higher the chances that the lake is polluted. If the amounts of negative and positive ions differ greatly, then the lake is either not polluted seriously or the measurements are not accurate. As an example, look at the measures of the sums of the positive and negative ions for two different lakes, lake number 22 and lake number 95.

Lake #22

positive ions = 0.1 + 47.0 + 7.4 + 2(421.7) + 2(87.2) = 1072.3

negative ions = 6.4 + 2(156.1) + 36.7 + 364.1 = 719.4

Lake #95

positive ions = 7.8 + 63.1 + 3.6 + 2(17.5) + 2(21.4) = 152.3

negative ions = 6.4 + 2(37.5) + 70.5 + 1.0 = 152.9

Notice that the results for the two lakes are different. In lake number 22 the sum of positive ions is much larger than the sum of negative ions. However, in lake number 95 the sums are very close. Table 1 shows the sum of positive ions and negative ions for each individual lake.

At this point, one needs to establish a way to

compare the values in Table 1 in order to indicate the seriousness of each value as compared to the pollution levels. To do this, all lakes whose sums of positive and negative ions are in the high ten percent level will be in one group, while those whose values are in the top twenty percent level will form another group, and those in the thirty percent level will form yet another group. The remaining lakes whose sums are not included in the aforementioned groups will form their own group. Figure 18 shows a map of Norway which has these groups plotted out for further reference.

Also, there are other signs on this map that should be explained. Some of the lake measurements show respectively higher values of substances than others. For example, lake number 34 has Σ positive ions = 70.0 while lake number 22 has Σ positive ions = 1072.3. Therefore, in order to make some distinction between such lakes, they are separated into three different categories as designated in the map key.

The interesting fact about all of these calculations is that the grouping of the lakes according to chemical analysis is very similar to the grouping that was done through the fuzzy c-means and Ward's method. There is a region in the southern part of Norway where all of the lakes are in the first ten percent group, and it is interresting to note that most of these lakes have relatively

high amounts of substances in them (Fig. 18). Recall that the fuzzy c-means (c=4, c=5, c=6) and Ward's method show nearly the same conclusions. Also, note that the western part of Norway has some pollution, although it is not as heavy as that found in the southern part of Norway. Again, the fuzzy c-means and Ward's method showed this same grouping. According to the map there is almost no problem in the northwest of Norway, as well as in the eastern and central parts, with the exceptions of locally polluted areas.

The above conclusions bring up the idea that maybe the data set can be put into four different groups, or in other words, there are four clusters that best describe the data. This idea correlates with the four cluster obtained through using the fuzzy c-means and Ward's methods, because as was mentioned before, the results obtained using the fuzzy c-means algorithm with c=4, c=5, and c=6 and both Ward's methods were not much different from a four cluster method using the fuzzy c-means with c=4 only and the Ward's method with normalized data (refer to page 36). It is now reasonable to conclude, that through the use of the fuzzy c-means with four clusters and the Ward's method with four clusters and the map of Norway plotted according to the chemical analysis of the lakes, that the best number for clustering the data is four.

Now that we have established those lakes in southern and western parts of Norway are polluted, it is beneficial to know which chemical substances, existing in Norway's lakes, are actually harmful to the environment. In looking at the data sheet, and without going through a lot of chemistry, it is easy to note that the most harmful chemical compounds existing in the polluted areas are:

$$HNO_3 \rightarrow H^+ + NO_3$$

and

$$HCl \rightarrow H^+ + Cl^-$$

These two acids are capable of killing much of the life that thrive in the lakes. For instance, most fish cannot survive in water containing the above mentioned acids. The next most abundant chemical compounds found in the lakes are those substances that make hard water, for example $CaSO_4$. These are not that harmful, and therefore they are not considered polluting materials. There are also many other compounds existing in the lakes, such as sodium chloride found mostly in the southern part of Norway, but this report is not concerned with them.

The gas known as NO_2, produced by heavy industrial companies, exists in the atmosphere above Norway. When it rains, this gas mixes with the water, thus producing the "acid rain." The interesting question is, where are

the harmful gases coming from? Since Norway does not have heavy industrial companies, and neither do any of Norway's neighbors, the closest candidates for producing these gases are West Germany and Great Britain. In order to know whether the gases are indeed coming from the above mentioned countries, the wind patterns traveling from them to Norway must be studied. Since this is a totally new aspect, it cannot be dealt with in this report. So, it is left to others for further study.

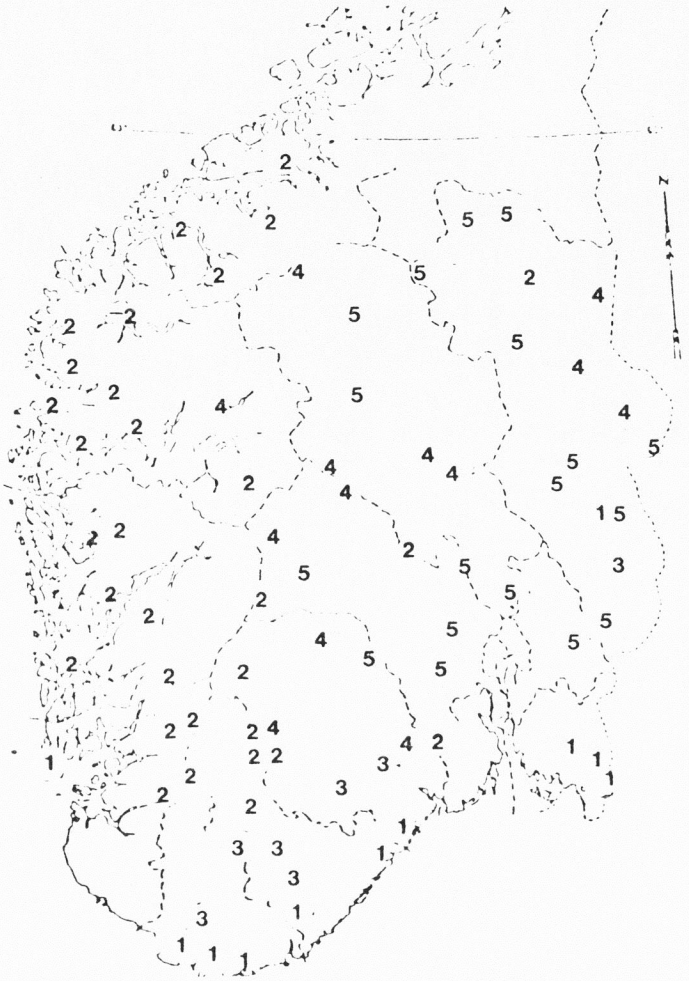


Fig. 12. Ward's method with five clusters.

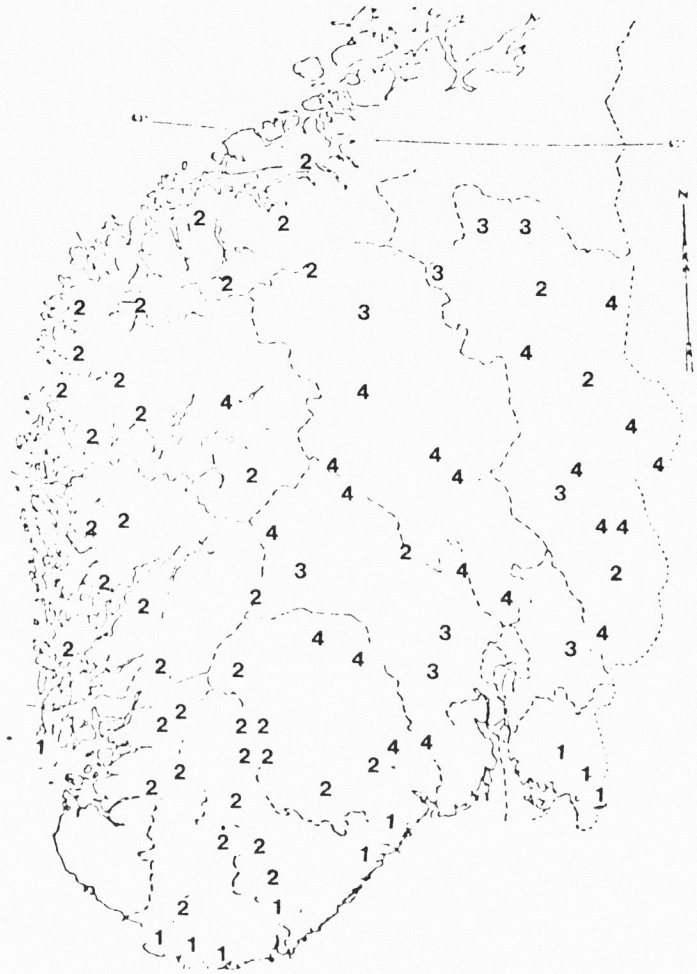


Fig. 13. Fuzzy c-means with c=4.

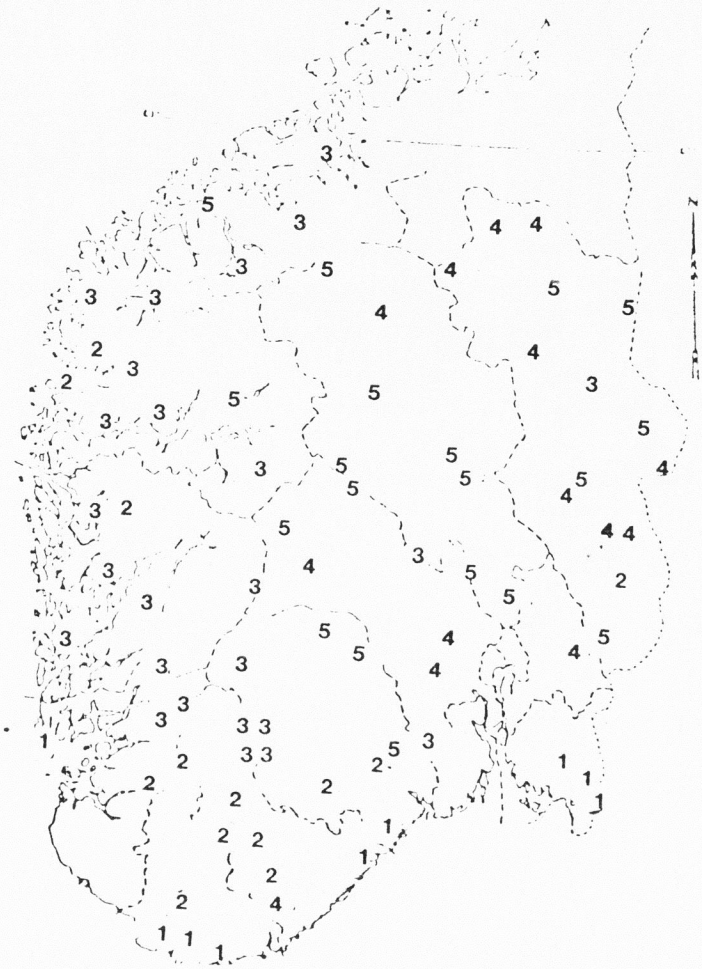


Fig. 14. Fuzzy c-means with c=5.

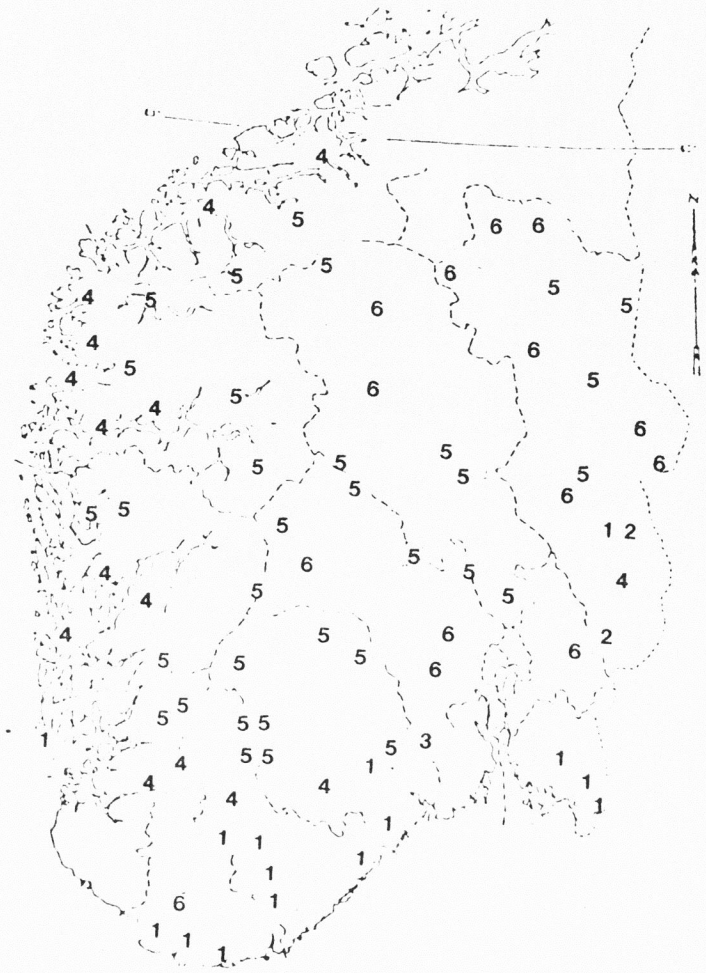


Fig. 15. Fuzzy c-means with c=6.

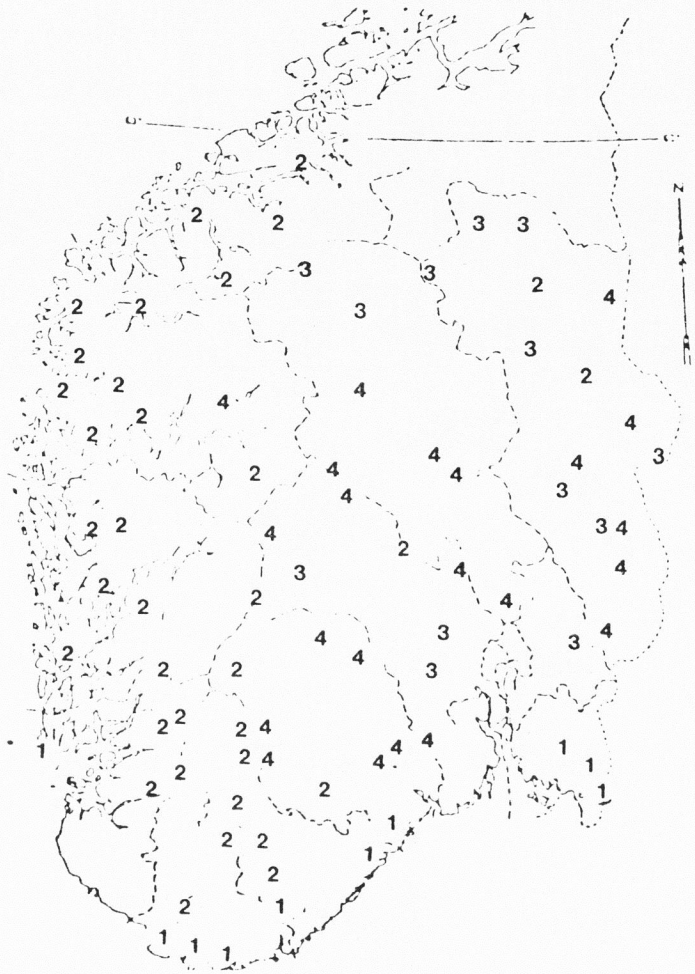
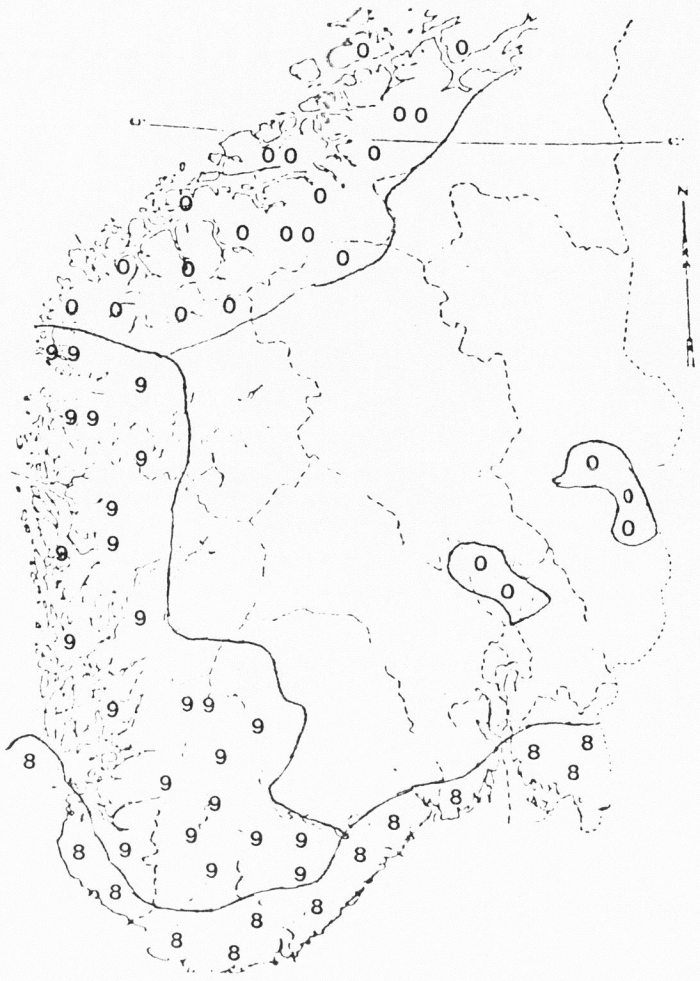
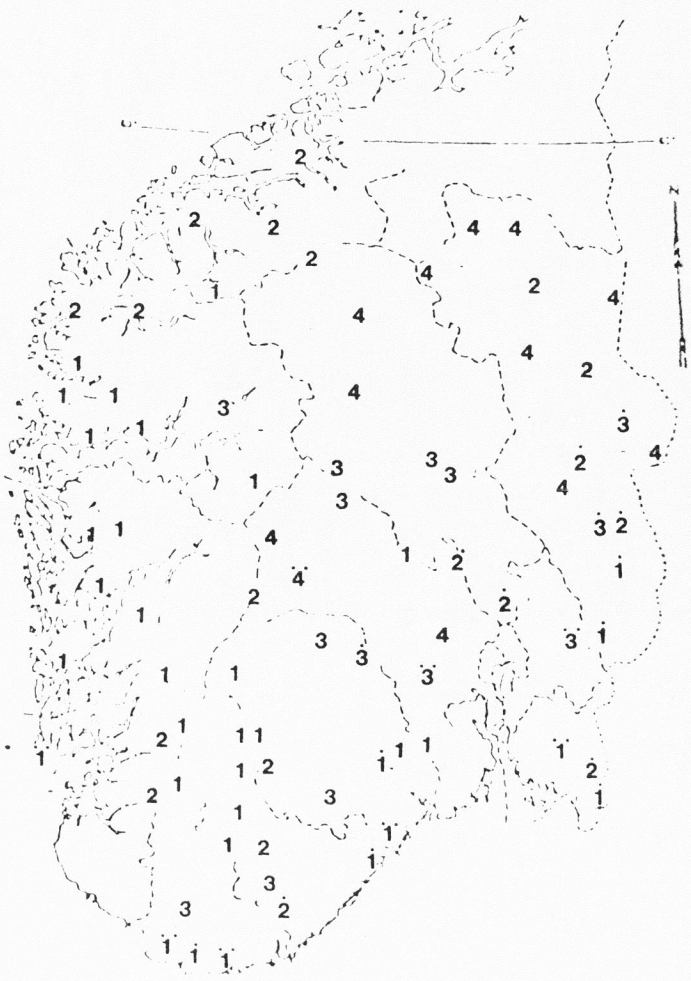


Fig. 16. Ward's method applied to normalized data with four clusters.



map key: 8 = highly polluted
9 = lightly polluted
0 = possible light pollution

Fig. 17. Lake pollution in Norway.



map key: 1= sum of negative and positive ions are in the 0 to 10% group

2= sum of negative and positive ions are in the 10 to 20% group

3= sum of negative and positive ions are in the 20 to 30% group

4= sum of negative and positive ions are in the 30 to 50% group

·= relatively high amounts of ions

··= extremely high amounts of ions

Fig. 18. Clustering based on chemical analysis.

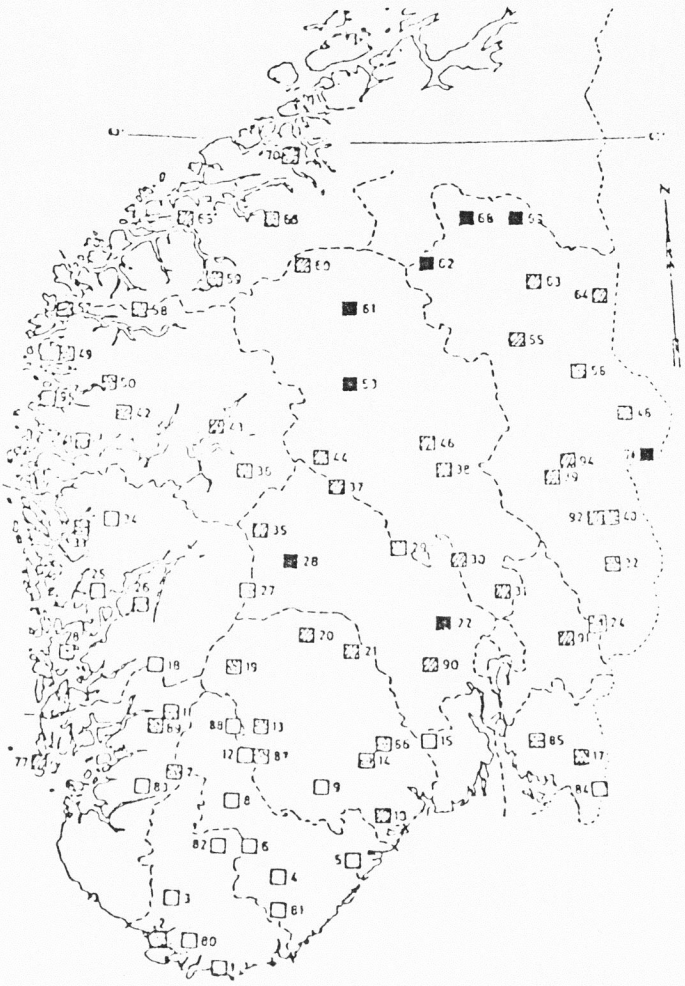


Fig. 19. Map of Norway with lakes numbered.

Table 1

List of the sums of negative and positive ions

Lake #	Sum of positive ions	Sum of negative ions
1	593.2	600.6
2	533.1	513.9
3	222.4	279.2
4	195.7	256.3
5	399.7	422.0
6	170.1	201.1
7	123.4	116.4
8	127.1	135.7
9	122.3	174.5
10	594.7	542.8
11	112.6	108.5
12	132.8	134.5
13	226.9	216.6
14	251.4	251.3
15	250.0	256.2
17	483.0	428.9
18	146.9	152.1
19	104.7	105.4
20	268.9	182.3
21	341.5	279.4
22	1072.3	719.4
24	354.7	320.6
25	143.9	138.9
26	94.2	90.3
27	111.7	97.6
28	527.0	357.0
29	117.5	121.5
30	321.9	271.8
31	345.1	300.8
32	256.3	263.5
33	199.4	192.0

Table 1. cont'd

Lake #	Sum of positive ions	Sum of negative ions
34	70.1	74.4
35	311.3	217.6
36	94.6	86.9
37	310.7	239.1
38	319.2	242.2
39	586.0	405.6
40	392.7	347.4
41	167.7	159.7
42	116.9	111.2
43	236.3	170.3
44	285.7	209.7
46	303.7	226.1
48	390.7	292.0
49	178.4	166.6
50	100.5	97.3
53	393.1	274.7
55	545.7	372.7
56	101.3	116.8
57	279.1	251.7
58	146.6	125.7
59	103.4	95.6
60	150.8	130.1
61	1918.1	1168.8
62	889.2	578.2
63	150.0	129.9
64	333.0	231.6
65	286.4	243.4
66	97.6	117.4
68	557.7	360.5
69	701.6	435.2
70	208.2	180.1
74	590.7	415.0

Table 1. cont'd

Lake #	Sum of positive ions	Sum of negative ions
77	866.9	786.1
78	260.9	246.8
80	398.5	409.7
81	337.9	375.9
82	148.1	162.6
83	101.7	122.7
84	453.3	493.3
85	587.9	587.1
86	244.4	226.7
87	215.4	189.4
88	95.2	101.5
89	187.8	160.2
90	658.0	501.3
91	672.9	527.5
92	581.0	487.1
94	377.0	330.2
95	152.3	152.9

Table 2

Data set used in this report

#	H^+	NO3	SULF	CL	NA	K	CA	MG	HCO3
1	19.5	10.0	145.7	279.3	287.1	10.2	69.9	68.3	19.9
2	9.3	8.9	114.5	239.8	248.0	8.4	70.4	63.3	36.2
3	25.1	40.7	70.8	95.9	89.2	3.1	24.5	28.0	1.0
4	33.1	11.4	89.5	64.9	59.2	3.8	25.9	23.9	1.0
5	22.9	6.4	145.7	124.1	121.8	8.4	62.4	60.9	0.1
6	27.5	2.1	75.0	48.0	46.1	3.1	24.5	22.0	1.0
7	9.8	6.4	20.8	31.0	55.2	2.8	15.5	12.3	37.4
8	11.0	2.9	47.9	31.0	35.2	2.3	24.5	14.8	6.0
9	18.2	7.9	68.7	28.2	26.1	3.8	21.5	15.6	1.0
10	0.9	10.7	177.0	138.2	122.7	13.3	132.7	96.2	39.9
11	4.6	2.1	33.3	28.2	31.3	2.3	24.0	13.2	11.6
12	6.0	7.9	50.0	25.4	27.4	2.6	35.4	14.0	1.2
13	3.3	3.6	81.2	31.0	37.0	3.6	64.4	27.1	19.6
14	17.8	2.9	95.8	33.9	37.4	4.1	64.9	31.3	22.9
15	3.9	8.6	97.9	50.8	47.4	7.7	65.9	29.6	1.0
17	5.9	3.6	114.5	132.6	134.0	5.9	102.8	65.8	63.7
18	4.8	9.3	54.1	33.9	37.4	6.9	32.4	16.5	0.7
19	1.6	2.9	37.5	19.7	25.7	3.8	22.0	14.8	7.8
20	0.2	1.4	33.3	14.1	23.5	5.6	105.8	14.0	100.2
21	1.0	2.1	87.4	28.2	36.1	5.6	119.8	29.6	74.3
22	0.1	6.4	156.1	36.7	47.0	7.4	421.7	87.2	364.1
24	3.5	1.4	110.3	50.8	55.7	6.9	93.3	51.0	47.8
25	2.6	7.9	35.4	59.2	54.8	3.3	23.5	18.1	1.0
26	5.4	5.0	20.8	33.9	36.1	3.3	11.5	13.2	9.8
27	0.5	1.4	29.1	19.7	22.2	2.8	34.9	8.2	18.3
28	0.1	0.7	77.0	14.1	26.5	6.4	216.6	30.4	188.2
29	2.2	0.7	47.9	14.1	24.4	3.1	32.4	11.5	10.9
30	0.5	1.4	89.5	25.4	34.8	7.4	108.3	31.3	66.0
31	1.0	1.4	102.0	71.0	44.8	6.9	113.3	32.9	64.4
32	11.2	3.6	104.1	36.7	46.5	4.6	55.9	41.1	15.0
33	2.8	2.9	39.6	93.1	97.0	5.4	20.0	27.1	16.8

Table 2. cont'd

#	H^+	NO3	SULF	CL	NA	K	CA	MG	HCO3
34	9.1	5.0	18.7	25.4	30.0	2.0	5.5	9.0	6.6
35	0.4	3.6	43.7	22.6	28.7	7.2	117.8	19.7	104.0
36	0.6	1.4	29.1	8.5	14.8	4.6	27.9	9.9	18.8
37	0.2	2.1	68.7	14.1	26.5	3.6	104.8	35.4	85.5
38	0.3	0.7	66.6	11.3	26.1	5.6	117.3	26.3	97.0
39	0.2	2.1	93.7	28.2	32.2	5.6	232.0	42.0	187.9
40	4.7	2.9	120.8	50.8	50.0	5.6	105.3	60.9	52.1
41	4.7	3.6	35.4	70.5	69.6	6.6	19.5	23.9	14.8
42	5.1	2.1	22.9	50.8	51.8	2.8	13.0	15.6	12.5
43	0.4	1.4	35.4	28.2	26.1	7.2	84.8	16.5	69.9
44	0.3	3.6	52.1	11.3	24.4	4.6	83.8	44.4	90.6
46	0.2	2.1	58.3	14.1	27.0	4.9	101.3	34.5	93.3
48	0.6	1.4	77.0	31.0	36.1	2.8	123.8	51.8	105.6
49	7.9	7.1	31.2	81.8	80.9	3.8	19.0	23.9	15.3
50	2.6	4.3	25.0	36.7	37.4	4.1	15.0	13.2	6.3
53	0.1	2.1	60.4	8.5	25.7	9.7	116.3	62.5	143.3
55	0.3	2.9	75.0	19.7	40.9	8.7	172.2	75.7	200.1
56	4.8	0.7	50.0	11.3	20.9	6.6	23.0	11.5	4.8
57	1.8	0.7	50.0	104.4	118.3	3.8	43.9	33.7	46.6
58	0.8	0.7	25.0	53.6	50.9	3.1	29.4	16.5	21.4
59	0.9	2.1	31.2	16.9	21.7	2.8	32.4	6.6	14.2
60	0.4	0.7	35.4	16.9	32.6	5.6	45.4	10.7	41.7
61	0.0	0.0	174.9	25.4	69.2	0.5	718.6	205.6	793.6
62	0.1	2.9	108.3	19.7	38.3	12.0	374.2	45.2	339.0
63	0.3	1.4	35.4	11.3	33.1	5.6	39.9	15.6	46.4
64	0.2	0.7	41.6	25.4	41.8	5.4	100.8	42.0	122.3
65	0.3	1.4	39.6	101.6	115.3	5.6	46.4	36.2	61.2
66	1.0	4.3	20.8	28.2	30.5	3.3	21.5	9.9	12.8
68	0.1	0.0	58.3	25.4	32.2	14.1	229.5	26.3	218.5
69	0.1	0.7	58.3	28.2	37.0	15.3	269.5	55.1	289.7
70	0.9	0.7	27.1	104.4	94.4	4.1	24.0	30.4	21.6

Table 2. cont'd

#	H^+	NO3	SULF	CL	NA	K	CA	MG	HCO3
74	0.2	0.0	87.4	36.7	53.9	10.2	196.6	66.6	203.5
77	0.2	2.9	154.1	451.4	387.2	9.7	86.8	148.1	23.6
78	2.2	3.6	56.2	129.8	105.3	3.8	41.9	32.9	1.0
80	13.8	4.3	106.2	186.2	187.0	7.7	48.9	46.1	6.8
81	19.5	6.4	120.8	126.9	118.3	9.5	50.9	44.4	1.0
82	24.5	2.9	56.2	42.3	37.8	2.0	25.4	16.5	5.0
83	10.0	5.7	31.2	53.6	46.5	2.0	6.0	15.6	1.0
84	37.2	3.6	154.1	180.5	168.8	8.7	49.4	69.9	1.0
85	7.2	2.1	183.2	180.5	200.1	12.8	103.3	80.6	38.1
86	1.3	2.1	85.4	25.4	31.8	4.3	83.8	19.7	28.4
87	5.6	5.7	64.5	25.4	26.5	2.3	72.4	18.1	29.3
88	5.5	8.6	37.5	16.9	21.3	2.6	23.0	9.9	1.0
89	0.8	7.1	41.6	45.1	45.2	3.6	46.9	22.2	24.8
90	0.2	2.9	145.7	28.2	48.7	4.3	241.5	60.9	178.8
91	0.7	3.6	147.8	81.8	77.9	7.9	224.1	69.1	146.5
92	0.3	2.1	156.1	56.4	64.8	16.1	174.2	75.7	116.4
94	0.5	1.4	120.8	25.4	35.2	5.9	128.2	39.5	61.8
95	7.8	6.4	37.5	70.5	63.1	3.6	17.5	21.4	1.0

REFERENCES

[1] R. Duda and P. Hart, Pattern Classification and Scene Analysis (John Wiley and Sons, New York, 1973).

[2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm (Plenum, New York, 1980).

[3] C. T. Zahn, Graph-theoretical methods for detecting and describing Gestalt clusters, IEEE Trans.Comp. (1971) 67-79.

[4] M. R. Anderberg, Cluster Analysis for Application (Academic Press, New York, 1973).

[5] W. P. Windham, Cluster validity for the fuzzy c-means clustering algorithm, USU Mathematics Research Report, Utah State University, Logan, Utah, (1981).

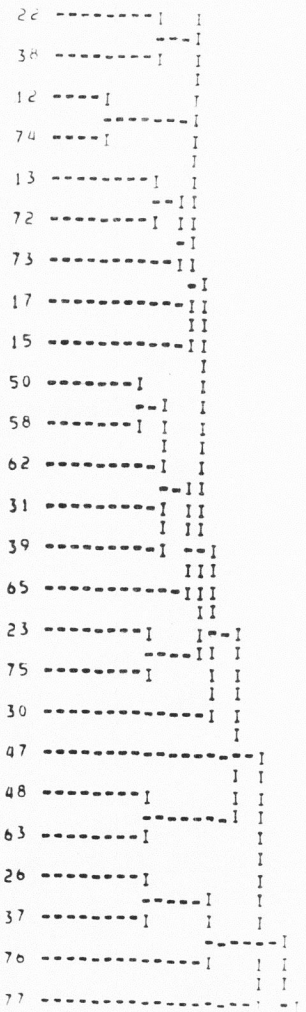
APPENDIX

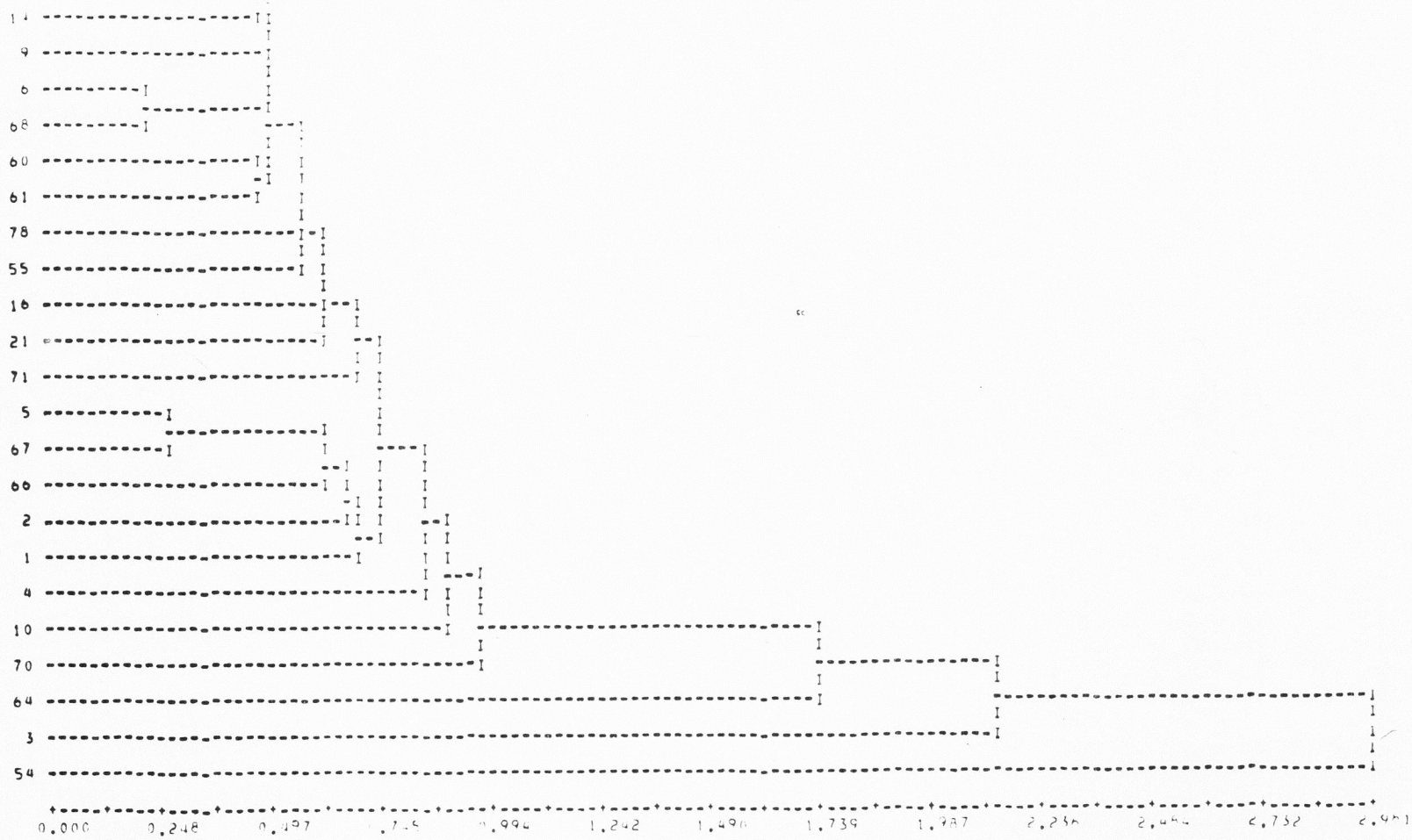
- - - - - - - TREE - - - - - - -

RESEMBLANCE MATRIX NAME : RES~
TREE NAME : TREE
NODE COUNT OPTION : 0
COPHENETIC OPTION : 0
CLUSTERING METHOD : SINGLE LINKAGE
MINIMUM VALUE ON TREE : 0.0000
MAXIMUM VALUE ON TREE : 2.9867

- - - - - - - TREE - - - - - - -

56
11
40
28
29
20
79
35
43
36
42
19
41
33
57
44
45
80
32
69
7
49
51
8




14
9
6
68
60
61
78
55
16
21
71
5
67
66
2
1
4
10
70
64
3
54
0.000
0.248
0.497
0.994
1.242
1.490
1.739
1.987
2.235
2.484
2.732
2.981