FUZZY CLUSTERING AND ITS APPLICATIONS

Noboru Ohsumi

The Institute of Statistical Mathematics

1979

Contents

Summary

This report will be mainly concerned with the techniques of cluster analysis, which are practically useful tools for the analysis of multidimensional data. Especially, during about the recent eighteen years, a large number of techniques of cluster analysis are proposed and studied by many researchers. Indeed the needs for these techniques arise in many fields of applied science.

Particularly hierarchical techniques are perhaps the most commonly used category of clustering methods. Because of this, the methods of hierarchical cluster analysis, especially agglomerative hierarchical clustering (AHC) methods, are used widely, since most of these methods are suitable for various kind of data and may be simply carried out. Nevertheless, though the investigations for evaluation or comparison of the properties between these methods are only little discussed systematically in works, it is quite importance and necessary to discuss them in order to enable us to formulate in terms of a satisfactory validity of clustering process.

Generally, in the AHC methods, relationships among the objects being classified are represented by a dissimilarity or similarity matrix. Therefore it is quite natural and meaningful to describe the matrix by taking as a representation of a relation or a graph. On such a case, especially, the concept of fuzzy relations proposed by Zadeh is more relevant and useful to examine reasonably the clustering models.

Thus we shall firstly attempt to summarize several properties among the AHC methods, especially, single linkage, complete linkage, and the synonymous techniques, based on the fuzzy theory. And we can clearly crarify the following features: a) the solution (i.e. dendrogram) obtained from single or complete linkage is equivalent to the fuzzy equivalence relation, b) especially, the solution of single linkage is identical to the transitive closure formed from min-max (or max-min) composition of the original dissimilarity (or similarity) matrix, namely an arbitrary reflexive and symmetric relation, and c) a minimal spanning tree may be generated by using the result of single linkage.

Furthermore, we shall propose the fuzzy degree of fitness which is a new index of evaluating and comparing relationship between two relations, the original similarity (or dissimilarity) matrix R and the matrix R^* derived from R by excuting AHC methods. And this index may be generated by using the fuzzy symmetric difference between two relations. Moreover we shall propose a modified clustering procedure, say modified linkage method, which approximates a given relation R in the sense of minimizing this index.

Successively we consider the comparison between the set of partitions formed by clustering process, since the evaluating problems of partitions is very important in the practical use of cluster analysis. Here we examine the following problems: a) comparison between two dendrograms (i.e. two equivalence relations), b) evaluation between partitions produced from two or many dendrograms, and c) estimation of the number of clusters, in other words, determining the level of cut on a dendrogram.

Finally several practical examples are given to illustrate and explain our consideration.

1. *Introduction*

Many researchers have argued that classification is fundamental process in all fields of science. Especially, there are a large number of automatical classification techniques which are known and distributed under the generic name *cluster analysis*. In the present paper, we shall discuss several considerations for the cluster analysis.

More concretely speaking, if we are given a data set of n objects or individuals and each of them is observed on each m characteristics or variables, then we are faced with the problem how to think out a procedure for grouping the objects into k groups. And we shall attempt to investigate the procedure for exploring the intrinsical tendency of data and to evaluate the characteristic of the groups.

Usually, the most well known terms for techniques which classify data into several groups are *stratification* and *discrimination*. But cluster analysis is essentially different from these concepts in the point that they are used to describe for assigning objects to group having *a priori* given labels. For example, in the social survey data, the technique which classifies data into groups by using the demographic factor, such as sex and occupation, is the stratification. And the discrimination is, for example, the term to describe the process for classifying the patients into several categories, such as the smoking and the no-smoking. On the other hand, the term cluster analysis is used for techniques which group objects by the use of proximity or similarity between objects. And in most cases, we *cannot determine* groups for assigning each object *a priori*. In this respect, cluster analysis is different from other methods of multivariate analysis. Already many comprehensive reviews on clustering techniques and their applications appeared and their detailed explanations were given by Cormack(1977), Everitt(1974 , 1977), Bock(1974), Sneath and Sokal(1973), Blashfield(1977) and by many researchers in many fields. But at the same time there were a number of case techniques misunderstood and misused.

During about the past eighteen years, especially, since 1960, there has been a growing interest in clustering methods for forming the meaningful classification. Though there are actually a large number of different methods of cluster analysis, most of them can be arranged under the two categories, namely, *hierarchical techniques* and *non-hierarchical techniques*. Hierarchical techniques may be also subdivided into *agglomerative methods* which perform the cluster by a successive fusions of the given objects on data set into several groups, and *divisive methods* which partition successively the data set into groups. And yet, non-hierarchical techniques are the iterative partitioning or optimization-partitioning, and there are many other methods included in this category, for example, mode-seeking methods, mixture problems, clumping techniques, and so on.

The term cluster analysis was firstly appeared and discussed in the social science and the psychology during the 1940's. At that time cluster analysis was considered as one which be comparable to factor analysis and principal component anaysis [see Tryon(1970)]. And it did not attract significant attention until about the early 1960's. But it must be emphasized that the so-called *numerical taxonomy* developed by Sokal and Sneath gave the main stimulus for biological taxonomy and attracted a great deal of attention in many fields. At the same time, another reason of the growth of interest to cluster analysis depends upon spread and existence of large high-speed computers made a possible to use practically many methods during the 1960's.

In this report, we shall discuss mainly the properties of the *agglomerative hierarchical clustering (AHC) methods*. Most of AHC methods start the clustering process by forming a matrix which represents the pairwise similarities or dissimilarities of all objects being groups. In general, the solution of AHC method can be represented by a *hierarchical structure*, that is, *a hierarchical tree* or *a dendrogram*. But it is rarely that the hierarchical structure or dendrogram is constructed explicitly by fusing of the objects. Therefore, the AHC methods can be interpreted as a result of successive approximations to form a hierarchical structure from the original similarity or dissimilarity matrix which represents a kind of relationship between the objects. There are a large number of AHC methods, for example, as well-known methods, single linkage, complete linkage, centroid method, group average method, Ward's method, and so on.

And yet, since the applied fields of technique such as cluster analysis are more interdisciplinary, there are many similar concepts. Really, the various methods of cluster analysis play an important role in such fields as psychology, sociology, biology, pattern recognition, systematic zoology, ecology, and so on. Therefore, there are many synonyms for cluster analysis, for example, Q-mode analysis in factor analysis, typology, grouping, clumping, numerical taxonomy or classification, and unsupervised pattern recognition.

Thus, our present aim is systematically to arrange several techniques of cluster analysis to avoid the confusion caused by the above mentioned. Furthermore, we shall extend to more generalized situation. In addition, examining the techniques proposed by many researchers in the distinct fields, for instance, biology and psychology we can observe that most of those are the same or almost same methods. For example, the researchers refferring to Johnson's paper in psychology use the terms "maximum method" and "minimum method", but these two methods are known as "complete linkage" and "single linkage" in the biological field. And the terms "complete linkage", "furthest neighbor", "rank order typal analysis", "diameter analysis" are synonyms. The terms "single linkage", "nearest neighbor", "minimum method", "elementary linkage analysis", "connected method" and a kind of minimal sapnning tree are synonyms.

We shall turn our attention to the facts that different terms have been used to describe the same thing and that they may be explained with a common conception that is said the *fuzzy set theory* proposed by Zadeh. Thus, it is natural that we attempt to introduce the fuzzy set theory and the fuzzy relation into the systematical consideration of cluster analysis.

In the section 2, firstly, we attempt to define and characterize several terminologies and properties of AHC techniques. For example, we discuss hierarchical structure, hierarchical partitioning set, dendrogram, ultrametric property, and so on. Moreover, solving our problems appears to require some adaptive tools. In such a case, fortunately, it seemed to us that the concept of *fuzzy ralation* is more useful to examine a clustering model. Thus, section 2 will serve to examine the several characteristics of the AHC methods and the relationship between them and fuzzy relations.

Secondly, in the section 3, we discuss the examinations for evaluation and comparison of the typical agglomerative hierarchical clustering techniques, especially, the *complete linkage* and the *single linkage*, by using the fuzzy relation. As the clustering process depends on the structure of data, we cannot directly evaluate the ability of cluster analysis by the comparison of various algorithm. Therefore, the examination by the fuzzy relation is more available and useful.

Successively, in the section 4, we discuss the problems of evaluating the number of clusters and of comparing between the different hierarchical partitions based on the same data set. Generally, the solution of a hierarchical clustering technique is represented by a dendrogram, but it is not always clear beforehand many clusters we can expect. Under the some assumption concerning a cluster, we propose the several criteria for evaluating the number of clusters and propose a procedure for investigating the clustering process.

As a complement to our discussion, in the following, we shall examine the behavior of clustering techniques by the observing effects of a little change in the data set disturbed by adding noise. By this procedure, that is called the *sensitivity analysis*, we can objectively *handle the problems of evaluating the number of clusters and investi*gating the validity of clustering process. Thus we can find a clue to the number of clusters and obtain a useful tool which is suitable for examining the structure included in data.

In the last section, to examine our consideration, we shall illustrate several practical examples and briefly summarize our argument.

2. *Notations and preliminary definitions*

2.1 *Conception of hierarchical structure*

For simplicity of our discussion, firstly, we shall prepare the several notations and terminologies.

We define the set of n objects

$$E = \{ O_1, O_2, O_3, \cdots, O_n \}$$

or for abbreviation,

$$E = \{ 1, 2, 3, \cdots, i, \cdots, n \}$$

and denote the raw data consisted of a $n \times m$ matrix,

$$X = (x_{il}) \quad (i = 1, 2, \cdots, n \ ; \ l = 1, 2, \cdots, m)$$

where $\underline{x}_i = (x_{i1}, x_{i2}, \cdots, x_{im})'$ is the observed vector for the i th object. Then, the AHC methods begin with the computation of a similarity matrix $S = (s_{ij})$ or a dissimilarity (i.e. distance-like measure) matrix $D = (d_{ij})$ between the objects formed from X. The distance d_{ij} represents the degree of difference between i th object and j th object and the similarity s_{ij} represents the degree of proximity between the objects. The dissimilarity d_{ij} or the similarity s_{ij} is said to be a *metric* for E if it satisfies the following three conditions:

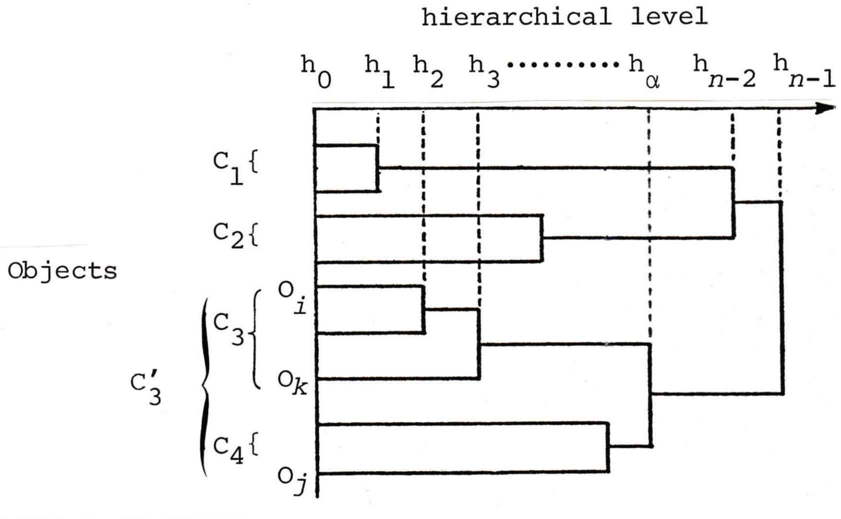
i) reflexivity or anti-reflexivity,

ii) symmetry, (1)

iii) transitivity, that is, triangular inequality.

If d_{ij} or s_{ij} satisfies only the condition i) and ii), it is said to be a *non-metric*. Yet most of AHC methods can be commonly suitable for the use of various kinds of dissimilarity or similarity.

The basic clustering procedure with AHC methods is very similar and surprisingly simple. In the AHC methods, in short, the goal of a clustering process can be represented as a *dendrogram*. In other words, the input is a matrix D or S, the end of a clustering process is a dendrogram which is a graphical representation of *hierarchical structure*. Namely, the hierarchical structure or the dendrogram may be presented in the form of a tree diagram as shown in Figure 1, which is a two dimensional diagram configurating the fusions between objects which have been constructed at each successive level. As shown in Figure 1, when the order of fusion level is monotonically changing, it is said that the hierarchical structure possesses a property of *monotone transformation*.



This example consists of the nine objects. And there exists the set of cluster $C^6 = \{C_1, C_2, C'_3\}$ at the hierarchical level $h_\alpha (=h_6)$. Also, $O_i, O_k \varepsilon C_3$ and $O_i \varepsilon C_4$. Obviously, this tree has the ultrametric property.

Figure 1. A dendrogram with monotonic invariant property.

These conceptions more precisely are defined as follows.

Definition 1.

Let us define a hierarchical structure H on the set E as follows. We assume now all possible subsets on E being non-empty A, B, C, D, $\cdots$. If any one of the next three conditions is satisfied for any two sets A, B then a partition $H = \{A, B, C, D, \ldots\}$ is said to be *a hierarchical.*

$$\begin{aligned} &\text{i)} \quad A \cap B = \phi \quad \text{(empty)} \\ &\text{ii)} \quad A \subset B \qquad\qquad (2) \\ &\text{iii)} \quad A \supset B \end{aligned}$$

Definition 2.

Let us define a non-negative function $h(A)$ for $A \varepsilon H$ as an index characterizing H. Then,

$$h(B) < h(A) \qquad \text{for all } A, B \varepsilon H \text{ and } B \subset A \qquad (3)$$

Especially, if $h(A)=0$ then A indicates the set of each object. And if $h(A) \leq h(B)$ then it is said to be weak. Such function h is called the *index of hierarchy* and $h(A)$ indicates a *level* or *step* of cluster $A \varepsilon H$. If $h(B) > h(A)$ for $B \subset A$, then we call it the *inversion.*

Definition 3.

A *dendrogram* is considered as a hierarchical structure H specified by the index $h(\cdot)$. We shall write such dendrogram by $<H, h>$.

For example, Figure 2 shows a dendrogram with seven objects. And there exists the following hierarchical structure H.

$$H = \Big\{ \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{1,2\}, \{4,5\}, \{1,2,3\}, \{4,5,6\}, \{4,5,6,7\}, \{1,2,3,4,5,6,7\} \Big\}$$

Furthermore, we can observe $h(A_1) > h(A_2) = h(A_3) = 0$, $h(A) > h(B)$ and $h(A) > h(C) > h(B)$ in Figure 2.

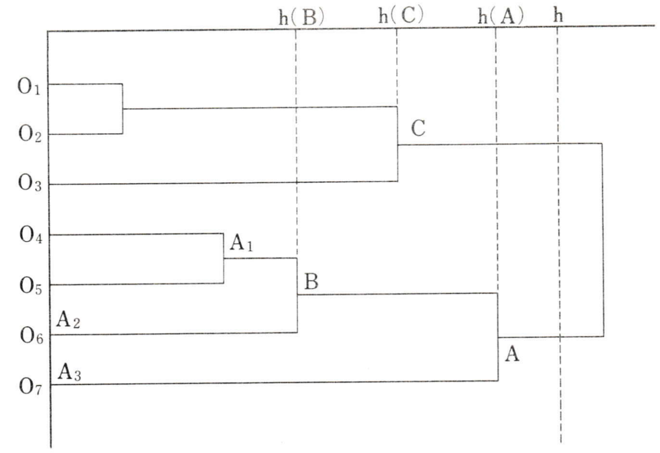


Figure 2. Relationship between hierarchical structure and index of hierarchy.

Definition 4.

Let A(h) be a set of mutually disjunctive partitions at the level h. For example, in Figure 2, $A(h) = \{A, C\} = \{\{1,2,3\}, \{4,5,6,7\}\}$. Then we can define a *hierarchical partitioning set* H* as follows. Let now $h_0 = 0 < h_1 < h_2 < \cdots < h_\alpha < \cdots < h_{n-1}$ be a monotonically increasing sequence of the index $h(\cdot)$.(*) Then we shall consider the following partitions.

$$\begin{aligned} C^0 &= A(h_0) = \{\{1\}, \{2\}, \cdots, \{n\}\} \\ &\vdots \\ C^\alpha &= A(h_\alpha) = \{C_1^\alpha, C_2^\alpha, C_3^\alpha, \cdots, C_{m_\alpha}^\alpha\} \\ &\vdots \\ C^{n-1} &= A(h_{n-1}) = \{\{1, 2, 3, \cdots, n\}\} \end{aligned} \tag{4}$$

(*) If we consider the relationship between objects as the similarity, then the sequence $\{h_\alpha\}$ is a monotonically decreasing. In the following discussion we shall use these description according to situations.

where $m_\alpha = n - \alpha$ indicates *the number of clusters* at the level h_α, obviously

$$n = m_0 > m_1 > m_2 > \cdots > m_{n-1} = 1.$$

In addition, C^α $(0 \leq \alpha \leq n-1)$ is a partition at the level h_α. Thus H^* is represented by the following,

$$H^* = \{ C^0, C^1, C^2, \cdots, C^\alpha, \cdots, C^{n-1} \} \tag{5}$$

Obviously, the partition C^α at the level h_α may be generated from the set of clusters (i.e. partitions) at the level $h_{\alpha-1}$. The dendrogram with the above described properties is called that is a *monotonic invariant.*

Definition 5.

Let us define an *ultrametric* produced from a hierarchical structure H or a dendrogram $< H , h >$ as follows.

$$\delta_{ij} = \min \{ h(A) | A \in H, \quad \text{for any } i, j \in A \} \tag{6}$$

It is clear that δ_{ij} satisfies symmetry and reflexivity. The above (6) is rewritten by the following.

$$\delta_{ij} \leq \max \{ \delta_{ik}, \delta_{kj} \} \quad \text{for any} \quad i, j, k \in E \tag{7}$$

The distance which satisfies the above (7) as the condition of transitivity is said that is an *ultrametric.* By the duality, if a similarity δ' satisfies the next expression, then it is said to be a *inframetric.*

$$\delta'_{ij} \geq \min \{ \delta'_{ik}, \delta'_{kj} \} \quad \text{for any} \quad i, j, k \in E \tag{7}'$$

These inequalities may be always derived from the dendrogram. In fact, we can observe easily the existence of the ultrametric property on the dendrograms shown in Figure 1 and 2. And if $\delta_{ij} \geq \delta_{ik} \geq \delta_{kj}$ then $\delta_{ij} \leq \max \{ \delta_{ik}, \delta_{kj} \} = \delta_{kj}$ (for any $i, j, k \in E$), that is, any i, j, k construct an *isosceles triangle.*

Thus in the following discussion, we treat only the methods such that the result of clustering may be represented by the monotonic hierarchical structure. A clustering method which transforms a D or S into a hierarchical structure (i.e. a dendrogram) may be regarded as a procedure which imposes the *ultrametric property* of a dissimilarity or similarity, whether the original one is metric or non-metric.

2.2 *A brief description of AHC techniques*

In general, the hierarchical methods is used as a strategy to reproduce strictly hierarchical structure in the data. The first stage in many AHC techniques is the conversion of the matrix X into an $n \times n$ matrix of inter-object similarities or dissimilarities, with the exception of some procedures such as Ward's method. AHC techniques first form an initial set on n clusters (that is, each object is a cluster) and then, in a stagewise way reduce the number of clusters one at a time until all n objects form one cluster. Difference between methods of this kind arises from different ways of defining dissimilarity or similarity between objects or between two clusters of objects. Thus we obtain the following definition.

Definition 6.

The AHC method is a procedure which forms the $<H , h>$ with a monotonic hierarchical structure. Namely, by an AHC method, we can obtain a hierarchical partition

$$C^{\alpha} = \{ C^{\alpha}_{1}, C^{\alpha}_{2}, \cdots , C^{\alpha}_{m_{\alpha}} \}$$

or in abbreviation

$$C^{\alpha} = \{ C_{1}, C_{2}, \cdots , C_{n-\alpha} \}$$

at the level h_{α}, and which is derived from $C^{\alpha-1}$ at the level $h_{\alpha-1}$.

We can illustrate the typical algorithm of AHC methods as follows.

[*Basic algorithm of AHC method*]

[Step 1] We consider each object as one cluster, that is, put label i to each object and $C^0=\{C_1, C_2, \cdots, C_n\}=\{\{1\},\{2\},\cdots,\{n\}\}$.

[Step 2] Calculate a dissimilarity matrix $D=(d_{ij})$ where d_{ij} is the dissimilarity between i th and j th objects, and find a pair of clusters (C_p, C_q) for which the distance between C_p and C_q, δ_α is the smallest, where $C_p, C_q \varepsilon C^{\alpha-1}$. Namely,

$$\delta_\alpha = d^*_{pq} = \min\{ d^*_{ij} \mid O_i \varepsilon C_p, O_j \varepsilon C_q, p \neq q \} \qquad (8)$$

And merge clusters C_p and C_q and form $C_t(=C_p \cup C_q)$. Then recalculate the distances between cluster C_t and all other clusters except C_t. The above δ_α is the distance obtained by this merge.

[Step 3] Repeat [Step 2] $(n-2)$ times, (namely, $\alpha=1, 2, 3, \cdots, n-1$) or a suitable number of times preassigned. In each stage, record the information about the pair of clusters merged and the distance between them.

[Step 4] Lastly, draw a dendrogram.

In case of using similarity matrix S in the above algorithm, set S, s_{ij}, "similarity", and "max" instead of D, d_{ij}, "dissimilarity", and "min". Since the purpose of this paper is firstly to discuss the problem concerned with the relationship between the AHC methods and the fuzzy relation, we state mainly about the well-known two methods single linkage and complete linkage. The both procedure define by replacing the right hand side of (8) in the above algorithm by the following formulas.

Definition 7.

Single linkage method

$$\min \{ d_{ij} | O_i \varepsilon C_p , O_j \varepsilon C_q \}$$

$$\triangleq \min_{l,m} [\min_{r,s} \{ d_{rs} | O_r \varepsilon C_l , O_s \varepsilon C_m \}] \qquad (1 \leq l , m \leq n-\alpha , l \neq m , p \neq q)$$

Complete linkage method

$$\max \{ d_{ij} | O_i \varepsilon C_p , O_j \varepsilon C_q \}$$

$$\triangleq \min_{l,m} [\max_{r,s} \{ d_{rs} | O_r \varepsilon C_l , O_s \varepsilon C_m \}] \qquad (1 \leq l , m \leq n-\alpha , l \neq m , p \neq q)$$

where the symbol "$\triangleq$" indicates the meaning of definition.

The most essential difference between these two methods is that complete linkage takes a *maximum* operation and single linkage takes a *minimum* operation. In other words, complete linkage is exactly the opposite of the single linkage. Moreover, it is clear that the both methods possess the following properties. Firstly, a sequence of distances, say $\{ \delta_\alpha \}$ ($\alpha = 0, 1, 2, \cdots$), generated by the algorithm has a property of monotonically increasing, namely,

$$\delta_0 = 0 < \delta_1 < \delta_2 < \cdots < \delta_\alpha < \cdots < \delta_{n-1} \tag{9}$$

Let us now denote by $\Delta = (\tilde{\delta}_{ij})$ $(i,j = 1, 2, \cdots ,n)$ a distance matrix derived from a dendrogram formed by single linkage. Similarly, let $\nabla = (\underset{\sim}{\delta}_{ij})$ denote a distance matrix produced by complete linkage. Then the following relationships are always satisfied.

$$\tilde{\delta}_{ij} \leq d_{ij} \tag{10}$$

or

$$\underset{\sim}{\delta}_{ij} \geq d_{ij} \tag{11}$$

Finally it is shown that $\tilde{\delta}_{ij}$ or $\underset{\sim}{\delta}_{ij}$ generates the hierarchical structure, namely the dendrogram.

2.3 *Fundamental concept of fuzzy relations*

Turning our attention to the fact that single linkage and complete linkage are characterized only by the *maximum* or *minimum* operation, it is more reasonable that we try to introduce the concept of fuzzy set theory, especially *fuzzy relation* or *fuzzy graph* into the generalized extension of AHC methods. And it is the next aim to examine the relationship between fuzzy relation and the AHC methods.

We shall now define the subset A of E to which $\mu(i|A)$ or μ_i represents the degree of belongingness. Under the consideration of the ordinary set theory, we can regard that if any $i \in A$ then $\mu_i = 1$, and if any $i \notin A$ then $\mu_i = 0$, say, μ_i is a *characteristic function.* But if the value of μ_i takes in the interval $[0, 1]$, μ_i is called a *membership function.* A subset A of this kind is said to be a *fuzzy subset.* We assume two fuzzy subsets A, B and define as follows:

$$A \subseteq B \quad \text{iff} \quad \mu(i|A) \leq \mu(i|B) \quad \text{for any } i \in E \tag{12}$$

$$A \wedge B = \{(i, \min\{\mu(i|A), \mu(i|B)\}) | i \in E\} \tag{13}$$

$$A \vee B = \{(i, \max\{\mu(i|A), \mu(i|B)\}) | i \in E\} \tag{14}$$

Therefore, the operators $\vee$ and $\wedge$ stand for union and intersection in the sense of fuzzy set theory, that is, $\vee$ and $\wedge$ indicate the maximum and minimum, respectively. And we can define also a fuzzy relation in $E_1 \times E_2$ as follows:

$$R = [\{(i,j), \mu(i,j|R)\} \mid i \in E_1, j \in E_2] \tag{15}$$

Especially, if $E_1 = E_2 = E$, we have the following *fuzzy (binary) relation.*

$$R = [\{(i,j), \mu(i,j|R)\} \mid i,j \in E] \tag{16}$$

where $\mu(i,j|R)$ is a membership function which represents the degree of belongingness of pair (i,j) to the subset $E^2 = E \times E$.

We suppose that the value of $\mu(i,j|R)$, in abbreviation $\mu(i,j)$ or μ_{ij}, takes only in the interval $[0,1]$. Then there are many fuzzy relations with the several conditions. We shall define the condition of some fuzzy relations as follows:

(a) $\mu(i,i)=1$ for any $i \in E$ (reflexivity)

(a)′ $\mu(i,i)=0$ for any $i \in E$ (anti-reflexivity)

(b) $\mu(i,j)=\mu(j,i)$ for any $i,j \in E$ (symmetry)

(c) $\mu(i,j) \geq \max_{k} [\min\{\mu(i,k), \mu(k,j)\}]$ (17)
for any $i,j,k \in E$ (max-min transitivity)

(d) $\mu(i,j) \leq \min_{k} [\max\{\mu(i,k), \mu(k,j)\}]$
for any $i,j,k \in E$ (min-max transitivity)

Table 1. Summary of some fuzzy relations

condition / relation	(a)	(a)'	(b)	(c)	(d)
similitude	X		X	X	
dissimilitude		X	X		X
resemblance	X		X		
dissemblance		X	X		

Besides, in the above condition (c), the operation of the righthand side indicates the following meaning.

$$\max_{k} [\min\{\mu(i,k), \mu(k,j)\}]$$
$$= \max [\min\{\mu(i,1), \mu(1,j)\}, \min\{\mu(i,2), \mu(2,j)\}, \cdots,$$
$$\min\{\mu(i,k), \mu(k,j)\}, \cdots, \min\{\mu(i,n), \mu(n,j)\}] .$$

That is, the composition of fuzzy relations is a kind of matrix calculation and an extension of ordinary matrix calculations by product-sum operation. It is the same in the expression (d). As shown in Table 1, we can consider the several fuzzy relations by the suitable combination of each condition. For example, the relation that satisfies the conditions (a), (b), (c) is a similitude relation.

Thus, we can easily find that the non-metric dissimilarity is identical to the *fuzzy dissemblance relation* and that the non-metric similarity is identical the *fuzzy resemblance relation.* And we can recognize that there is an important connection between fuzzy relations and clustering property. Moreover, for simplicity, let $R \circ R \subseteq R$ or $R^2 \subseteq R$ denote (c), and $R * R \supseteq R$ or $R^2 \supseteq R$ denote (d), where the symbols " $\circ$ " and " $*$ " denote max-min and min-max operations, respectively. We call the *max-min* and *min-max* (*two-fold*) *composition*(*) the relation R^2 and R^{2*} , respectively. Furthermore, we provide the following definitions.

Definition 8.

Let $\hat{R}$ denote the *max-min transitive closure* of a symmetric and reflexive relation.

$$\begin{aligned}\hat{R} &= R \vee (R \circ R) \vee (R \circ R \circ R) \vee \cdots \vee (\underbrace{R \circ R \circ \cdots \circ R}_{k\text{-fold max-min composition}}) \vee \cdots \\ &= R \vee R^2 \vee R^3 \vee \cdots \vee R^k \vee \cdots \qquad (18)\end{aligned}$$

Similarly, let $\check{R}$ denote the *min-max transitive closure* of a relation.

$$\begin{aligned}\check{R} &= R \wedge (R * R) \wedge (R * R * R) \wedge \cdots \wedge (\underbrace{R * R * \cdots * R}_{k\text{-fold min-max composition}}) \wedge \cdots \\ &= R \wedge R^{2*} \wedge R^{3*} \wedge \cdots \wedge R^{k*} \wedge \cdots \qquad (19)\end{aligned}$$

(*) An arbitrary fuzzy relation satisfies the properties of associativity and commutativity with respect to the operation "$\circ$" and "$*$". But the distributivity is not always satisfied. For example, let R, S and Q be three relations, $R \circ (Q \circ S) = (R \circ Q) \circ S$, $R \circ Q = Q \circ R$. However, in the distributivity, $R \circ (Q \vee S) = (R \circ Q) \vee (R \circ S)$, but $R \circ (Q \wedge S) \neq (R \circ Q) \wedge (R \circ S)$ [see Kaufmann(1973)].

In addition, it is well known that the fuzzy relations possess several important properties as follows, and these properties play an important role in the agglomerative type clustering.

$$\hat{R} = R \vee R^2 \vee R^3 \vee R^4 \vee \cdots \rightarrow \hat{R}^2 \subset \hat{R} \qquad (20)$$

$R^2 \subset R \leftrightarrow R = \hat{R} \leftrightarrow R$ is transitive

$R^2 = R \rightarrow R = \hat{R} \leftrightarrow R$ is transitive (21)

If $R^{k+1} = R^k$ for any positive integer k, that is, idempotent,

$$\hat{R} = R \vee R^2 \vee R^3 \vee \cdots \vee R^k \qquad (22)$$

and k is sometimes called a *number of reachability*.

$$\hat{R} = R \vee R^2 \vee R^3 \vee \cdots \vee R^n \qquad (23)$$

Let $\check{R}$ denote the min-max transitive closure. Then

i) $\overline{\hat{R}} = \check{\overline{R}}$

ii) $\overline{R \circ R} = \overline{R} * \overline{R}$ (24)

where ii) is the relationship between max-min and min-max, and $\overline{R}$ is a complement of R. In the above notations, without the loss of generality we can put $\check{R}$, $\wedge$, $\supset$, R^{k*} instead of $\hat{R}$, $\vee$, $\subset$, R^k, respectively.

We can find more the relationships between a relation and a *graph*. Namely we can identify an ordinary relation with a graph. By the similar consideration, a fuzzy (binary) relation may be considered as a *fuzzy graph*. Accordingly, let G denote a graph , the nodes of G is the set of objects and each weight of G corresponds with each value of membership function on E^2.

3. *On the evaluation of the AHC methods by fuzzy relations*

3.1. *Relationship between the AHC methods and fuzzy relations*

Since the investigations for evaluation or comparison of AHC methods are only little discussed in formal works up to the present, it is really important and necessary to discuss them in order to study cluster analysis. Therefore, in the following, we shall turn our attention to these problems. By the aid of the results described in the previous section, firstly, we can easily find the following property.

Property 1.

An ultrametric inequality for the distance is identical to a min-max transitivity, that is, *an ultrametric is a fuzzy dissimilitude relation.* Similarly, *an inframetric is a fuzzy similitude relation.* Occassionally, the former is called the dissimilarity relation, the latter is called the similarity relation. And either construct the *equivalence relation.*

We can easy prove this property by comparing (7) with (d) of (17). Let now δ_{ij} be an ultrametric distance. Then,

$$\delta_{ij} \leq \min_k [\max \{ \delta_{ik}, \delta_{kj} \}]$$
$$\leq \max \{ \delta_{ik}, \delta_{kj} \} \qquad (\text{for any } k \in E)$$

Therefore it is obvious that (7) implies (d) of (17), and is implied by (d) of (17). And we can find the analogous relationship about inframetric.

We next consider a *path* or *chain* from O_i to O_j (i.e. an ordered r-tuple with or without duplication) in the finite graph $G \subset E \times E$ (i.e. in the relation R especially the *dissemblance relation*),

$$w_{ij} = (i = i_0, i_1, i_2, \cdots, i_{r-1} = j) \tag{25}$$
$$\text{where } i_t \in E, \quad t = 0, 1, 2, \cdots, r-1 .$$

With each path $(i_0, i_1, i_2, \cdots, i_{r-1})$ we shall set a value defined by the following,

$$\ell (i_0, i_1, i_2, \cdots, i_{r-1})$$

$$= \max\{ \mu(i_0, i_1), \mu(i_1, i_2), \cdots, \mu(i_{r-2}, i_{r-1})\} \qquad (26)$$

for abbreviation, rewritten the above

$$\ell (i,j) = \max_{1 \le t < r-1} \{ \mu(i_{t-1}, i_t) \} \qquad (27)$$

Let us now consider all possible paths existing between O_i and O_j $(i,j \in E)$ and let W_{ij} be the set of all such paths. Namely,

$$W_{ij} = \{ w_{ij} \mid w_{ij} = (i_0, i_1, \cdots, i_{r-1})\} \qquad (28)$$

Moreover, we shall define the minimal path W^*_{ij} from O_i to O_j by

$$\ell^*(i,j) = \min_{W^*_{ij}} \{ \ell(i,j)\}$$

$$= \min_{W^*_{ij}} [\max\{ \mu(i_0, i_1), \mu(i_1, i_2), \cdots, \mu(i_{r-2}, i_{r-1})\}] \qquad (29)$$

Then we can obtain several relationships from the discussion by Kaufmann (1973).

(1) Let $R \subset E \times E$, which is a dissemblance relation, then we have

$$\ell^*_k(i,j) = \mu(i,j \mid R^{k*}) \text{ for any } i,j \in E \times E$$

where $\ell^*_k(i,j)$ indicates *the strongest path* or *the largest link distance* existing from O_i to O_j of *length* k. This is proved by induction as follows :

a) if $k = 1$, then $\ell^*_1(i,j) = \mu(i,j \mid R)$.

b) if $k = 2$, then $\ell^*_2(i,j) = \min\{ \ell(i,j)\}$

$$= \min_{i_1} [\max\{ \mu(i,i_1 \mid R), \mu(i_1,j \mid R)\}]$$

$$= \mu(i, j \mid R*R)$$

because, in general,

$$\mu(i,j\,|\,R*R) = \min_{k}\,[\,\max\{\mu(i,k\,|\,R),\ \mu(k,j\,|\,R)\}]\qquad (i,j,k \in E).$$

c) if $k = 3$, then

$$\begin{aligned}\ell_3^*(i,j) &= \min_{i_1,i_2}\,[\,\max\{\mu(i,i_1\,|\,R),\ \mu(i_1,i_2\,|\,R),\ \mu(i_2,j\,|\,R)\}]\\ &= \min_{i_2}\,[\,\max\{\ell_2^*(i,i_2),\ \mu(i_2,j\,|\,R)\}]\\ &= \min_{i_2}\,[\,\max\{\mu(i,i_2\,|\,R*R),\ \mu(i_2,j\,|\,R)\}]\\ &= \mu(i,j\,|\,R*R*R),\end{aligned}$$

Therefore we can obtain

$$\ell_k^*(i,j) = \mu(i,j\,|\,\underbrace{R*R*\cdots*R}) = \mu(i,j\,|\,R^{k*}) \qquad (30)$$

(k-fold min-max composition)

(2) We can find the following relation.

$$\ell^*(i,j) = \mu(i,j\,|\,\check{R}) \qquad (31)$$

where $\check{R}$ is the transitive closure of a dissemblance relation R.

(3) Furthermore, we may explain

$$\ell_k^*(i,j) = \ell_{j \le n}(i,j)$$

where $\ell^*_{j \le n}(i,j)$ indicates the value of the strongest path of length less than or equal to n from O_i to O_j. After removing the closed-loops or circuits in G, there remains a chain which has at most length n. Thus, the above (3) is built strictly.

Property 2.

By the above (2) and Definition 5, we can see that $\ell^*(i,j)$ has the ultrametric property. That is,

$$\ell^*(i,j) \le \max\{\ell^*(i,k),\ \ell^*(k,j)\}\qquad (i,j,k \in E) \qquad (32)$$

Of course, $\ell^*(i,i) = 0$, $\ell^*(i,j) = \ell^*(j,i)$.

Next, we shall verify that there exists a maximal one in the family of ultrametric δ_{ij} which is dominated by the dissimilarity d_{ij}, namely

$$\delta_{ij} \le d_{ij} \ .$$

We now denote such δ by

$$\delta^*_{ij} = \sup\{\delta_{ij}\}\ , \quad i, j \in E \tag{33}$$

obviously,

$$\begin{aligned} \delta^*_{ij} &= \sup\{\delta_{ij}\} \le \sup\,[\,\max\{\delta_{ik}\,,\,\delta_{kj}\}\,] \\ &= \max\,[\,\sup\{\delta_{ik}\}\,,\,\sup\{\delta_{kj}\}\,] \\ &= \max\{\delta^*_{ik}\,,\,\delta^*_{kj}\} \end{aligned} \tag{34}$$

Thus, δ^*_{ij} is an ultrametic.

And such δ^* is called a *maximal dominant ultrametric*. This is identical to one which is a *subdominant ultrametric* called by M. Roux.

We can now see the next interesting property based on the above results.

We assume the ultrametric δ_{ij} being dominated by d_{ij}, namely $\delta_{ij} \le d_{ij}$.

Noting that d_{ij} is the same one as $\mu(i\ ,j)$, we can find that

$$\delta^*_{ij} \le \max_{1 \le t \le r} \{\delta_{i_{t-1}, i_t}\} \le \max_{1 \le t \le r} \{d_{i_{t-1}, i_t}\} = \ell(i\ ,j)\ .$$

Furthermore since a minimum of $\ell(i,j)$ is $\ell^*(i,j)$,

$$\ell(i,j) \geq \ell^*(i,j) \ge \delta^*_{ij} \tag{35}$$

where δ^*_{ij} is (33).

Moreover let us consider a path whose length is one, that is, $\ell(i,j) = \mu(i,j|R)$, then $\ell(i,j) = d_{ij}$ and always $\ell^*(i,j) \le \ell(i,j)$ by (29).

Thus,

$$d_{ij} \ge \ell^*(i,j) \tag{36}$$

By (35) and (36), we can obtain

$$\delta^*_{ij} \le \ell^*(i,j) \le d_{ij} \quad .$$

However, δ^*_{ij} is a maximal of δ_{ij}, therefore $\ell^*(i,j)$ must be identical to $\delta^*(i,j)$. Namely $\ell^*(i,j) = \delta^*_{ij}$. Thus it is easy to see the following.

Property 3.

$\ell^*(i,j)$ is a maximal dominant ultrametric. In other words, $\mu(i,j|\check{R})$ or $\check{R}$ generates an ultrametric which is maximal dominant.

We examine, in the following, the connection the fuzzy relation and the result of single linkage. We shall denote a sequence of the hierarchical index, which is derived from $\Delta = (\tilde{\delta}_{ij})$ by $\{\tilde{h}_\alpha\}$ $(\alpha=0, 1, \cdots, n-1)$. Then the dendrogram may be represented by $< H , \tilde{h} >$. And the present problem is to verify that $< H , \tilde{h} >$ produced by single linkage is identical to the transitive closure of D. Before the following statement, we show that these relationships are immediately clarified by the next simple illustration.

Example 1.

We set $E = \{1, 2, 3, 4\}$ and a dissimilarity matrix $D = (d_{ij})$ (i.e. dissemblance relation).

$$D = \begin{array}{cccc} 1 & 2 & 3 & 4 \\ \end{array}$$
$$D = \left[\begin{array}{cccc} 0.0 & 0.3 & 0.4 & 0.2 \\ & 0.0 & 0.1 & 0.4 \\ & & 0.0 & 0.5 \\ & & & 0.0 \end{array}\right] .$$

Clustering D by algorithm of single linkage, we can represent the result by a dendrogram in Figure 3. In this example, at stage one of the clustering process object 2 and 3 are fused to form a group, because $d_{23} = 0.1$ is the smallest value in D. Next we calculate the distance between this group and the remaining two objects, 1 and 4 as follows.

$$d_{(23).1} = \min\{d_{21}, d_{31}\} = 0.3 = d_{21},$$
$$d_{(23).4} = \min\{d_{24}, d_{34}\} = 0.4 = d_{24}$$

and form a new matrix D′.

$$D' = \begin{array}{c} \begin{array}{ccc} 1 & \{2,3\} & 4 \end{array} \\ \begin{bmatrix} 0.0 & 0.3 & 0.2 \\ & 0.0 & 0.4 \\ & & 0.0 \end{bmatrix} \end{array}$$

The smallest value in D′ is d_{41}=0.2, and so objects 4 and 1 are fused and become a second group. Recalculate the distances and we obtain the matrix D′′.

$$D'' = \begin{array}{c} \begin{array}{cc} \{1,4\} & \{2,3\} \end{array} \\ \begin{bmatrix} 0.0 & 0.3 \\ & 0.0 \end{bmatrix} \end{array}$$

Lastly fusion of the two groups take place to form a single group.

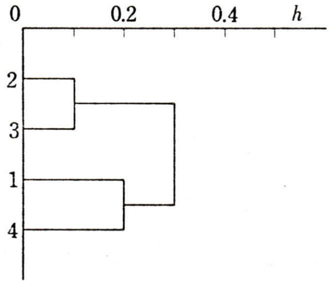


Figure 3. Dendrogram produced by applying single linkage to a dissimilarity matrix.

Using this dendrogram, we can make $\Delta = (\tilde{\delta}_{ij})$ and index $\{ \tilde{h}_\alpha \}$.

$$\tilde{\delta}_{ii} = 0 = \tilde{h}_0 \qquad (i = 1, 2, 3, 4)$$

$$\tilde{\delta}_{23} = 0.1 = \tilde{h}_1$$

$$\tilde{\delta}_{24} = 0.2 = \tilde{h}_2$$

$$\tilde{\delta}_{12} = \tilde{\delta}_{13} = \tilde{\delta}_{24} = \tilde{\delta}_{34} = 0.3 = \tilde{h}_3$$

Thus a dendrogram $\langle H , \tilde{h} \rangle$ for matrix D was constructed.

Here we now consider the case that single linkage produces C^{α} from $C^{\alpha-1}$ according to Definition 6, then we can write as follows,

$$d_{pq} = \min\{ d_{ij} \mid i \in C_p ,\ j \in C_q ;\ C_p ,\ C_q \in C^{\alpha-1} \} \qquad (37)$$

Accordingly,

$$\tilde{h}_{\alpha} = \min\{ d_{pq} \mid \text{for any } C_p ,\ C_q \in C^{\alpha-1}\} \qquad (38)$$

and $\tilde{h}_{\alpha} > \tilde{h}_{\alpha-1}$ $(\alpha = 0, 1, \cdots, n-1)$. Under the above relations (37) and (38), let us consider a path defined on $<H, \tilde{h}>$

$$\ell(i,j) = \max_{1 \leq t \leq r-1} \{d_{i_{t-1}, i_t}\}$$

then $\ell(i,j)$ satisfies all the properties of path described previously.

Next, let C_A, C_B denote two clusters in $C^{\alpha-1}$, then it is easy to find that:

i) if $C_A = C_B$, then $\ell(i,j) \leq \tilde{h}_{\alpha}$ for any $i, j \in C_A (= C_B)$

ii) if $C_A \neq C_B$, then $\ell(i,j) > \tilde{h}_{\alpha}$ for any $i \in C_A$, $j \in C_B$.

Moreover, considering $\ell^*(i,j) = \min\{\ell(i,j)\}$ on $<H, \tilde{h}>$, we can observe that the next i)′, ii)′ are correspondent with the above i), ii).

i)′ $\ell^*(i,j) \leq \tilde{h}_{\alpha}$ $i, j \in C_A (= C_B)$

ii)′ $\ell^*(i,j) > \tilde{h}_{\alpha}$ $i \in C_A$, $j \in C_B$

Especially we now specify as $C_A = C_p$, $C_B = C_q$ $(C_p, C_q \in C^{\alpha-1})$, that is, consider two clusters which are merged at the next fusion-level $\tilde{h}_{\alpha}$, then it is clear to be always $\ell^*(i,j) \leq \tilde{h}_{\alpha}$. In the other hand, $\ell^*(i,j)$ is a maximal dominant ultrametric by Property 3. Therefore, $\ell^*(i,j)$ consists with $\tilde{h}_{\alpha}$, accordingly, $\tilde{h}_{\alpha}$ is a maximal dominant ultrametric and is derived from the transitive closure of a dissemblance relation, namely a dissimilarity matrix.

Finally, summarizing the previous discussion, we can obtain several important properties.

Property 4.

(1) A distance matrix $\Delta = (\tilde{\delta}_{ij})$ derived from a dendrogram formed by single linkage possesses the *maximal dominant ultrametric property* and generates a *fuzzy dissimilitude relation*, namely, which is reflexive, symmetric and min-max transitive.

(2) *The distance matrix* Δ *is identical to a fuzzy min-max transitive closure* derived from the original dissimilarity matrix $D = (d_{ij})$.

(3) These results include fairly well-known several methods which are proposed by many researchers.

For example, an algorithm to construct a kind of hierarchy proposed by M. Roux is surely identical to single linkage. Furthermore, a hierarchical r-clique grouping procedure by E. Peay is really equivalent to the operation of transitive closure. And, of course, minimum method (Johnson), the nearest neighbor (Lance and Williams), elementary linkage analysis (McQuitty) are the same or almost same methods as the solution of transitive closure, since these methods perform the clustering process by the use of minimum and maximum operation.

In the above discussion, it is needless to say that we can set "similarity", "similitude", "max-min" and "S" instead of "dissimilarity", "dissimilitude", "min-max" and "D" without the loss of generality.

In addition, we can see another interesting characteristic.

Property 5.

Let us set $\tilde{\delta}_{ij} = d_{ij}$, then we have a *minimal spanning tree (MST).* In other words, we shall now denote by T the relation or graph,

$$T = (t_{ij}),$$

where
$$t_{ij} = \begin{cases} d_{ij}, & \text{if } \tilde{\delta}_{ij} = d_{ij} \\ 0 \quad , & \text{if } \tilde{\delta}_{ij} < d_{ij} \end{cases}$$

According to Definition 7 of single linkage, the fusion distance at the level $\tilde{h}_\alpha$ is exactly the smallest one between objects or clusters. Accordingly, let us consider two clusters C_p, C_q fusing successively at the level $\tilde{h}_\alpha$, it is obvious that there exists exactly one distance which satisfies $\tilde{\delta}_{ij} = d_{ij}$ (except for the case of tie).

Thus, we can observe easily that T derives a MST. A MST is a tree whose weight is minimum among all spanning trees of graph (i.e. relation). A *tree* is a connected graph without circuits and a *spanning tree* of connected graph G is a tree in G which contains all nodes (i.e. objects) of G. And we define the weight of a tree to be the sum of the weights of edges (i.e. each element of relation or dissimilarity matrix) constituting the MST. Moreover if we consider complete linkage method, we can find the analogous property as follows.

Property 6.

A distance matrix $\nabla = (\underset{\sim}{\delta}_{ij})$ derived from a dendrogram formed by complete linkage method satisfies the ultrametric property and constructs a *fuzzy dissimilitude relation.* However, in general, the above ∇ is not always coincident with the transitive closure. But we can form a kind of spanning tree by the similar procedure as the above described method. Since always $\underset{\sim}{\delta}_{ij} \geq d_{ij}$, there exists a kind of spanning tree which derives from $\underset{\sim}{\delta}_{ij} = d_{ij}$.

Now it is easy to verify each characteristic described above.

Example 2.

Citing the result in Example 1, we can obtain

$$\Delta = (\tilde{\delta}_{ij}) = \begin{bmatrix} 0.0 & 0.3 & 0.3 & 0.2 \\ & 0.0 & 0.1 & 0.3 \\ & & 0.0 & 0.3 \\ & & & 0.0 \end{bmatrix}$$

On the other hand, let $\check{D}$ be the min-max transitive closure of D,

$$D * D = D^{2*} = \begin{bmatrix} 0.0 & 0.3 & 0.3 & 0.2 \\ & 0.0 & 0.1 & 0.3 \\ & & 0.0 & 0.3 \\ & & & 0.0 \end{bmatrix}$$

Similarly,

$$D^{3*} = D * D^{2*} = \begin{bmatrix} 0.0 & 0.3 & 0.3 & 0.2 \\ & 0.0 & 0.1 & 0.3 \\ & & 0.0 & 0.3 \\ & & & 0.0 \end{bmatrix} = D^{2*}$$

Therefore, $D^{2*} = D^{3*} = D^{4*}$, and

$$\check{D} = D \wedge D^{2*} = \begin{bmatrix} 0.0 & 0.3 & 0.3 & 0.2 \\ & 0.0 & 0.1 & 0.3 \\ & & 0.0 & 0.3 \\ & & & 0.0 \end{bmatrix}$$

Thus we can see $\check{D} = \Delta$.

Furthermore, comparing $\check{D}$ with Δ or D,

$$\tilde{\delta}_{12} = d_{12} = 0.3$$

$$\tilde{\delta}_{14} = d_{14} = 0.2$$

$$\tilde{\delta}_{23} = d_{23} = 0.1$$

Then,

$$T = \begin{bmatrix} 0.0 & 0.3 & 0.0 & 0.2 \\ & 0.0 & 0.1 & 0.0 \\ & & 0.0 & 0.0 \\ & & & 0.0 \end{bmatrix}$$

Surely, this matrix generates a minimal spanning tree.

3.2 *A measure of difference between two similarity matrices*

"How to get some good measure of evaluations and comparisons of clustering techniques" is a common problem encountered in works. On these characteristics of hierarchical structures, many workers have discussed from various points of view. For example, Hartigan (1967) has introduced a measure of distance between two similarity matrices from a statistical point of view, and it is similar to a measure of stress proposed by Kruskal (1964), Farris (1969), Rohlf and Sokal (1962) have investigated the so-called CPCC (Cophenetic Correlation Coefficient) and it has been also much used practically. On the other hand, Jardine and Sibson (1971), Lerman (1970) have examined a method of evaluation based on an ordinary relation. Clearly, our situation is the extension of the latter.

As we mentioned already, the AHC methods that proposed up to the present are considered as an exact method to form the fuzzy equivalence relation itself by a kind of successive approximation. We are interested, in the following, to evaluate the difference between two distance or similarity matrices. We need an index which examines a difference between the original similarity matrix and the matrix derived from a dendrogram. Since the distance or similarity formed by single and complete linkage is regarded as a fuzzy dissimilitude or similitude relation, as the extention of ordinary symmetric difference, it is natural and valid that we consider the *fuzzy symmetric difference as a measure of evaluating and comparing the result of clustering process.*

Definition 9.

Let now $S = (s_{ij})$ and $S^* = (s^*_{ij})$ denote the original and derived similarity matrix, respectively. Then, the fuzzy symmetric difference is defined as follows:

$$\rho(S, S^*) = (S \wedge \bar{S}^*) \vee (\bar{S} \wedge S^*) \tag{39}$$

where $\bar{S}$ and $\bar{S}^*$ represent the complement of S and S^*, respectively. Though we will mainly describe here about the case of similarity, our consideration can be easily extended to the case with distance measures. And let ρ_{ij} denote an element of matrix $\rho(S,S^*)$, then we have the following relationships,

$$\begin{aligned}\rho_{ij} &= (s_{ij} \wedge \bar{s}^*_{ij}) \vee (\bar{s}_{ij} \wedge s^*_{ij}) \\ &= \frac{1}{2}\{1 - |s_{ij} + s^*_{ij} - 1| + |s_{ij} - s^*_{ij}|\}\end{aligned}$$

accordingly,

$$\text{if } s_{ij} + s^*_{ij} \leq 1, \quad \text{then } \rho_{ij} = \frac{1}{2}\{s_{ij} + s^*_{ij} + |s_{ij} - s^*_{ij}|\}$$

$$\text{if } s_{ij} + s^*_{ij} > 1, \quad \text{then } \rho_{ij} = 1 - \frac{1}{2}\{s_{ij} + s^*_{ij} - |s_{ij} - s^*_{ij}|\}$$

where $\bar{s}_{ij} = 1 - s_{ij}$, $\bar{s}^*_{ij} = 1 - s^*_{ij}$.

Thus,

$$\rho_{ij} = \begin{cases} \max(s_{ij}, s^*_{ij}) & \text{if } s_{ij} + s^*_{ij} \leq 1 \\ 1 - \min(s_{ij}, s^*_{ij}) & \text{if } s_{ij} + s^*_{ij} > 1, \end{cases} \tag{40}$$

where the range of ρ_{ij} is in the interval $[0, 1]$.

We shall call ρ_{ij} *a fuzzy distance*, since it is a distance taking fuzzy informations into consideration. In many cases, it may be convenient to employ a scalar rather than a matrix for comparison between relations. Therefore, we investigate a degree of goodness of fit between two relations by a measure of $\rho(S,S^*)$, $r = \sum\sum_{i<j} \rho_{ij}$.

In the sense of fuzzy set theory, a difference between S in itself is not always zero,namely let us write it by $\rho_0(S,S)$, $r_0 = \|\rho_0(S,S)\| = \|S \wedge \bar{S}\|$ is not always zero. Moreover, the maximum difference is given by $r' = \|S \vee \bar{S}\|$. Thus, we can obtain the expression $r_0 \le r \le r'$. Using these results, we propose an index.

Definition 10.

We shall denote by r^* the index which indicates a *fuzzy degree of fitness* between two relations.

$$r^* = (r - r_0)/(r' - r_0) \tag{41}$$

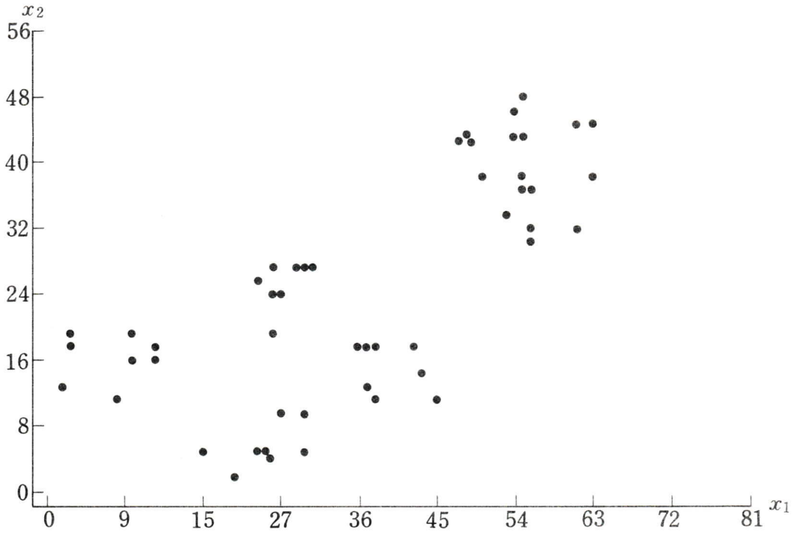
Obviously, this expression satisfies the inequality $0 \le r^* \le 1$. Therefore, we can examine a degree of the goodness of fit by the r^*'s . Such consideration based on the fuzzy symmetric difference is considered as the generalized extention of the absolute deviation or the rank order statistic in the traditionally statistical method.

Example 3.

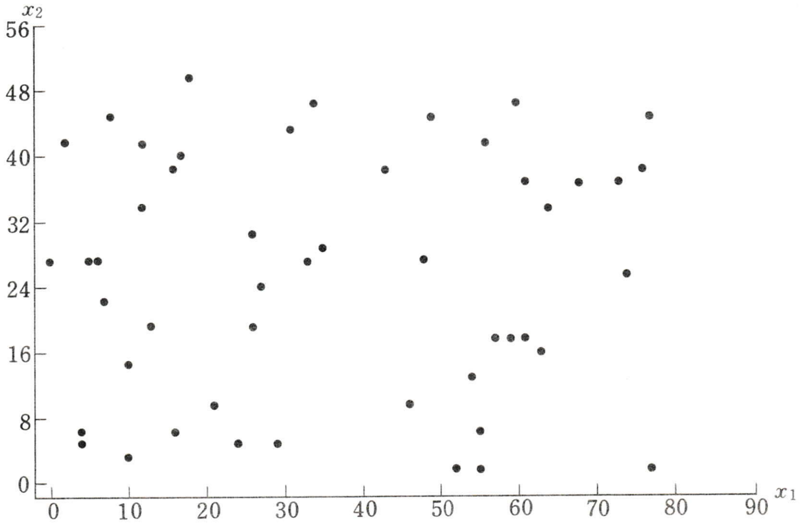
Let us suppose we wish to cluster the two sets of data in which two measurements are observed for each of fifty objects, respectively. These sets of data are shown as scatter diagrams in Figure 4-(A) and (B). Data (B) seems to consist of a single group apparently and data (A) consists of several well separated and compact groups. The clustering methods to be used are single linkage and complete linkage, and the results are shown in the following table.

Table 2.

Method	data set (A)	 (B)
complete linkage	0.229	0.318
single linkage	0.235	0.286
# of reach-ability (k)	6	12
# of size (n)	50	50



(A) In this figure, it seems that there exist several groups well separable in shape.



(B) In this figure, the configuration of data is considerably fuzzy.

Figure 4. The sets of data constructed by generating random variables for experiments.

The results shown in the above table suggest that our forcast is almost surely valid. Examination of the results obtained from data (B) shows that single linkage is very likely to be the situation rather than complete linkage. However, the two values is not almost different and both values are larger than the values produced from the data (A). Investigating the results for data (A), we can observe that complete linkage is little superior to single linkage but almost same. The above results illustrate clearly that data (A) is more closer to the hierarchical structure than data (B). Finally, observing the behavior of index r^*, we can evaluate more quantitatively the validity of clustering process which have judged by empirical and subjective interpretability as usual.

3.3 *Extension of single linkage and complete linkage*

We shall now attempt to modify the algorithm of single and complete linkage, and to extend to more general case. Usually let us now define the dissimilarity (or similarity) measures between clusters used by AHC techniques is represented by the following recurrence formula.

$$d_{tr} = \frac{1}{2}(d_{pr} + d_{qr}) + (\gamma - \frac{1}{2})\left|d_{pr} - d_{qr}\right| \tag{42}$$

where d_{tr} is the distance between a cluster C_r and a cluster C_t formed by the fusion of cluster C_p and C_q, and d_{ij} is the distance between clusters C_i and C_j. And γ is a parameter and its value is given beforehand in the interval [0 , 1]. If $\gamma=0$, we can obtain single linkage and if $\gamma=1$, then complete linkage. Moreover, if $\gamma = \frac{1}{2}$, then the above relation shows the so-called *weighted pair group* (*WPG*) *method* proposed by Sokal. Surely,

if $\gamma=0$, then

$$d_{tr} = \frac{1}{2}(d_{pr}+d_{qr}) - \frac{1}{2}\left| d_{pr} - d_{qr} \right| = \min\{d_{pr}, d_{qr}\} \tag{43}$$

if $\gamma=1$, then

$$d_{tr} = \frac{1}{2}(d_{pr}+d_{qr}) + \frac{1}{2}\left| d_{pr} - d_{qr} \right| = \max\{d_{pr}, d_{qr}\} \tag{44}$$

and if $\gamma= 1/2$, then

$$d_{tr} = \frac{1}{2}(d_{pr}+d_{qr}) \text{ , namely average distance.} \tag{45}$$

Obviously, all of the results given by applying the above formula to the data, that is dendrograms, have the monotonic hierarchical structure. Therefore, by changing γ variously, clustering schemes with distinct characteristics can be obtained. Especially, if we attempt to adjust the value of γ keeping the value defined by the expression (41) relatively as small as possible, then we can investigate the solution which is more reasonably fitting to a given data.

Thus, it has been shown that our approach includes a natural generalization and extension for many AHC methods, especially which are similar to single linkage and complete linkage. And we shall call this method *modified linkage technique.*

Example 4.

We shall investigate several illustrations using artificial data.

(1) We consider the next relation as first example,

$$R = \begin{bmatrix} 1.00 & 0.05 & 0.20 & 0.50 \\ & 1.00 & 0.20 & 0.40 \\ & & 1.00 & 0.01 \\ & & & 1.00 \end{bmatrix}$$

R is a resemblance relation (i.e. similarity), so we compute a transitive closure by max-min composition and obtain the following relation,

$$\hat{R} = \begin{bmatrix} 1.0 & 0.4 & 0.2 & 0.5 \\ & 1.0 & 0.2 & 0.4 \\ & & 1.0 & 0.2 \\ & & & 1.0 \end{bmatrix}.$$

In this case, the number of reachability is two. And $\hat{R}$ coincides with the result of single linkage. Let us consider the complement of $\hat{R}$ and denote it by $\overline{\hat{R}}$

$$\overline{\hat{R}} = \Delta = (\tilde{\delta}_{ij}) = \begin{bmatrix} 0.0 & 0.6 & 0.8 & 0.5 \\ & 0.0 & 0.8 & 0.6 \\ & & 0.0 & 0.8 \\ & & & 0.0 \end{bmatrix}$$

and we can obtain a dendrogram $<H,\tilde{h}>$ characterized by the index $\{\tilde{h}_\alpha\} = \{\tilde{h}_0, \tilde{h}_1, \tilde{h}_2\} = \{0, 0.5, 0.8\}$ and the hierarchical structure $H = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,4\}, \{1,4,2\}, \{1,4,2,3\}\}$. Of course, it is clear that $\hat{R}$ generates a dendrogram in itself. Then we may consider each complement of them instead of "Δ", "$\tilde{\delta}_{ij}$", "$\tilde{h}_\alpha$", "H", namely "$\overline{\Delta}$", "$\overline{\tilde{\delta}}_{ij}$", "$\overline{\tilde{h}}_\alpha$", "$\overline{H}$", and define a dendrogram $<\overline{H},\overline{\tilde{h}}>$.

In the case of complete linkage, similarly, we can obtain

$$\overline{\nabla} = (\overline{\underset{\sim}{\delta}}_{ij}) = \begin{bmatrix} 1.00 & 0.01 & 0.01 & 0.50 \\ & 1.00 & 0.20 & 0.01 \\ & & 1.00 & 0.01 \\ & & & 1.00 \end{bmatrix}$$

where $\overline{\nabla}$, $\overline{\underset{\sim}{\delta}}_{ij}$ are complement for ∇, $\underset{\sim}{\delta}_{ij}$ respectively. And it is obvious that $\overline{\nabla}$ represents a dendrogram.

Moreover, in the formula (42),

i) if $\gamma = 0$, $r^* = 0.0$ (complete linkage)
ii) if $\gamma = 1/2$, $r^* = 0.097$ (WPG)
iii) if $\gamma = 1$, $r^* = 0.165$ (single linkage).

Therefore, in this case, complete linkage gives a best fitting solution.

(2) Next, we consider the following relation

$$R = \begin{bmatrix} 1.0 & 0.7 & 0.6 & 0.8 \\ & 1.0 & 0.9 & 0.6 \\ & & 1.0 & 0.5 \\ & & & 1.0 \end{bmatrix}$$

Then,

$$\hat{R} = R \circ R^2 \circ R^3 = \begin{bmatrix} 1.0 & 0.7 & 0.7 & 0.8 \\ & 1.0 & 0.9 & 0.7 \\ & & 1.0 & 0.7 \\ & & & 1.0 \end{bmatrix} .$$

This example is, in fact, the complement of D in Example 1 and the number of reachability is 3 in this case (Surely, we can verify the relation $\overline{\check{D}} = \hat{R}$).

Moreover, we can obtain

i) if $\gamma = 0$, $r^* = 0.182$

ii) if $\gamma = 1/2$, $r^* = 0.045$

iii) if $\gamma = 1$, $r^* = 0.0$

Thus, in this example, single linkage is the best solution and this result gives completely the opposite feature to the above (1).

Example 5.

Using the sets of data in Example 3, we shall examine the behavior of r^*'s with the change of γ in (42). Table 3 shows the result of such experiment. obseving this table, we can find that it is adequate to take γ's about 0.5 for each data. However, the values of r^* indicate that data (A) is closer to the hierarchical structure than data (B). This situation may be also observed by the size of reachability. Namely, the number of reachability obtained from (A) is six and that of (B) is twelve, thus the former is closer than the latter to the hierarchy.

Table 3. The behavior of r*'s with the change of γ.

	data set	
$\gamma(\times 10^{-1})$	(A)	(B)
0	0.229	0.318
1	0.201	0.291
2	0.173	0.262
3	0.150	0.260
4	0.136	0.237
5	0.136	0.222
6	0.149	0.214
7	0.163	0.271
8	0.179	0.279
9	0.200	0.286
10	0.235	0.286
# of reach-ability (k)	6	12
# of size (n)	50	50

Example 6.

Successively, let us investigate the several artificial data as in Figure 5. We shall calculate the Euclidean distance d_{ij} among the objects and form a distance matrix $D=(d^*_{ij})$, where d^*_{ij} is derived by dividing d_{ij} by the maximum element in it. Then, a matrix $S=(s_{ij})$ is produced by the transformation of d^*_{ij}, for example, which is represented by equation $s_{ij}=1-d^*_{ij}$. Thus, the matrix S is immediately considered as a similarity matrix, i.e., a fuzzy relation.

The above consideration is applied to our example of artificial data, and the results of computation are shown in Table 4. From these results, we can recognize the following characteristics in each figure of Figure 5. Firstly, we shall examine the three methods, that is, single linkage, complete linkage and WPG method. In figure (A), there are three clusters of the structure which is cloudly compact and apparently distiguishable, that is, the between-cluster distance is much larger than the within-cluster distance, and the gaps or moats between clusters can be apparently observed. In this case, complete linkage and WPG method are better and single linkage method give the poor results.

In figure (B), when there are two clusters which are internally homogeneous but connected by the so-called *bridge* or *chain* between them, it may be said that complete linkage and WPG method are again better than single linkage. Furthermore, let us consider the example which makes within-clusters more vague such as in figure (D). Then it is found that the values of r^* for complete linkage and WPG method become larger than the values of r^* of figure (B), and the values of single linkage are slightly larger. Thus, it may be reasonable if it is scarcely reasonable to apply complete linkage and WPG method to the clusters of such noisy structures as in figures (B) and (D).

And in figure (C), the two clusters are composition of several clusters connected by bridges. In this case, it is seen that the *gaps* among two clusters can be detected by any of the three methods, but complete linkage and WPG method and single linkage produce different results within two clusters. More precisely, we find a dendrogram which has *the chaining effect* by using single linkage while we find a dendrogram with reasonably compact clusters by complete linkage and WPG method. As shown in the above discussion, the similarity of complete linkage and WPG method which is empirically known is apparent according to the value of r^* shown in Table 4 .

Finally, we consider the situation in which the shapes of clusters are represented as figures (E), (F), and (G). These clusters possess the structure which can be characterized by *non-ellipsoidal* or *serpentine* *shapes*, but the border-lines of clusters are fuzzy. And it is immediately found that a difference between three methods can almost not be recognized. Under the consideration that complete linkage is related closely to single linkage by means of using max and min operations but the usage of those two operations is entirely different, we can find these results as a most interesting features. In these examples, it is suitable to utilize single linkage rather than complete linkage and WPG method, since the value of r^* in (E), (F), and (G) are relatively smaller that the value of (A), (B), and (D). Thus those properties discussed above agree with the fact known empirically and subjectively.

Moreover, observing the behavior of r^*'s with the change of γ's, we can see relationships between the obtained relations and the given data. For illustration, in the data (A), (B), (C), it is seen easy that the maximum distance between clusters is significant to investigate the difference of clusters which are well separable or compact in shape. On the other hand, each value of r^* in the data (E), (F) and (G) indicates that each data is spread in shape which is relatively fuzzy. And it is seen that the minimum distance between objects (i.e. the nearest neighbor) plays an important role about construction of clusters. Furthermore, observing a degree of convergence to the transitive closure by the number of reachability, say k, we can understand that the values of k represent a degree of closeness between the hierarchical structure (i.e. dendrogram) and the configuration of original data set. That is, it is clearly indicated that the data (A), (B), (C) and (D) are gathered in spherical compact shape, but that the data (E), (F) and (G) cannot be distinguished groups in the sense of well separable shape.

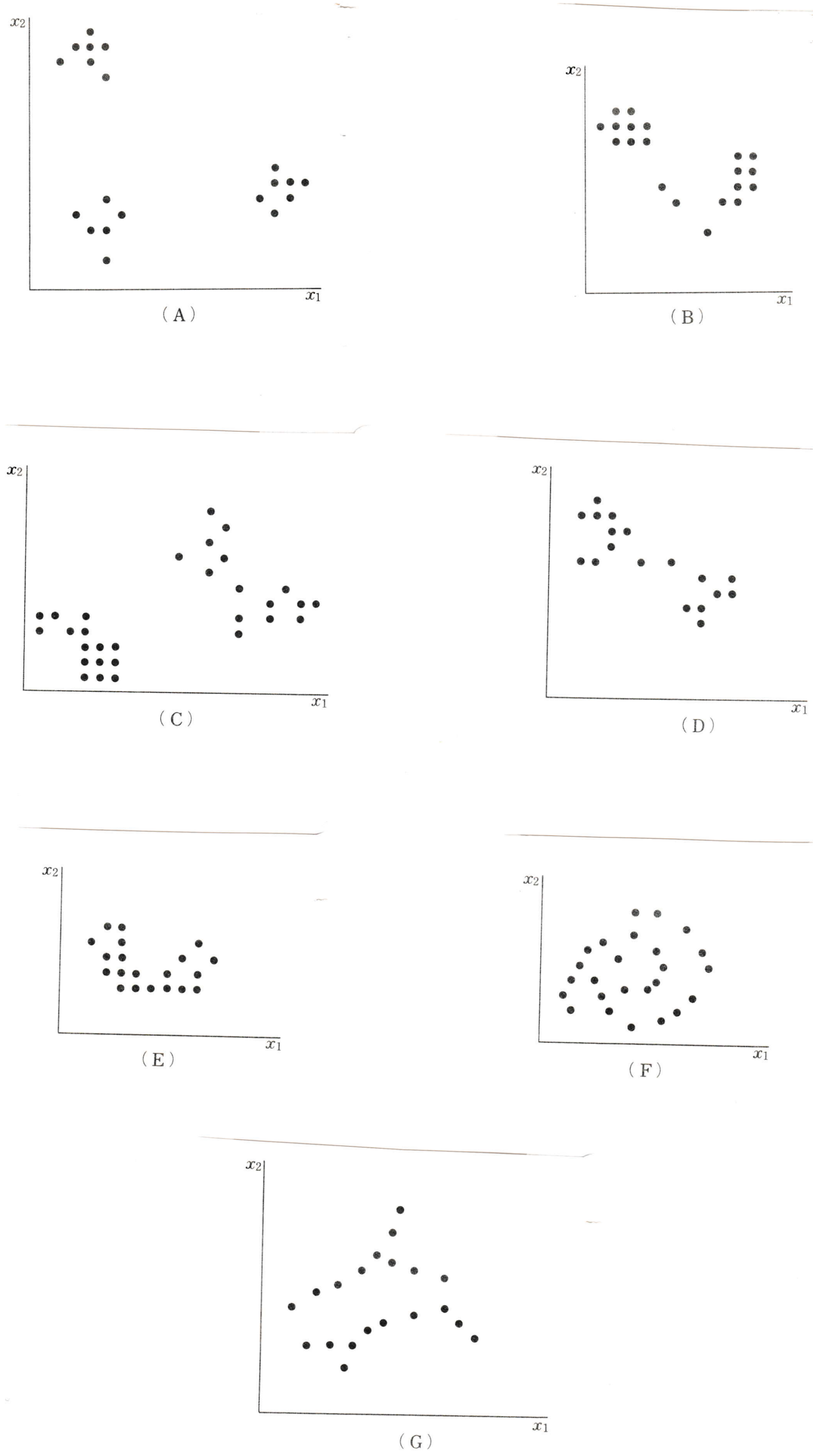


Figure 5.   Artificial data.

Table 4. The r*'s computed for each data set in Figure 5.

	Data set							
γ (X10^{-1})	(A)	(B)	(C)	(D)	(E)	(F)	(G)	
0	0.022	0.118	0.121	0.134	0.272	0.326	0.305	(complete linkage)
1	0.019	0.108	0.102	0.128	0.268	0.311	0.294	
2	0.022	0.100	0.089	0.136	0.254	0.312	0.286	
3	0.031	0.099	0.084	0.156	0.242	0.308	0.279	
4	0.047	0.107	0.093	0.173	0.239	0.314	0.280	
5	0.069	0.173	0.119	0.211	0.202	0.324	0.284	(WPG method)
6	0.100	0.206	0.153	0.256	0.231	0.337	0.302	
7	0.140	0.259	0.191	0.306	0.269	0.352	0.337	
8	0.189	0.314	0.209	0.356	0.291	0.366	0.367	
9	0.245	0.368	0.245	0.395	0.320	0.385	0.375	
10	0.305	0.415	0.275	0.421	0.344	0.394	0.390	(single linkage)
# of reachability (*k*)	4	6	6	7	11	10	13	
# of size (*n*)	20	20	30	18	20	25	20	

4. *Comparing partitions obtained by clustering*

Though there are many problems to be faced in using cluster analysis in practical, the most important and difficult are to handle the following situations:

i) examining the two dendrograms obtained by applying different clustering algorithms to the same data.

ii) comparing and evaluating between two dendrograms based on different sets or same set of data, and examining partitions generated from those dendrograms.

iii) evaluating the number of clusters, that is, comparing partitions specified on a dendrogram.

In short, there exist always the problems of comparing between dendrograms and investigating the partitions formed on dendrograms.

Nevertheless, most studies in the past were mainly concerned with the proposal of clustering algorithm, but in recent years there are an increasing number of studies which have utilized and emphasized the evaluation and stability of clustering techniques as mentioned above.

The present section aims to describe and discuss more fully for practical use the procedure of evaluating the stability of clusters or the sensitivity analysis, and to propose some procedures of the comparing the partitions by clustering or of the estimating the number of clusters.

4.1 *Comparison between two dendrograms*

Firstly we shall consider to examine the two dendrograms obtained by applying distinct clustering algorithm to the same data set. We now denote two dendrograms by $\langle H_A, h\rangle$, $\langle H_B, t\rangle$, and represent the relations (i.e. *similitude relations*) given by the both dendrograms by R_A, R_B, respectively. Then it is natural to apply the conception of fuzzy symmetric difference described in the section 3.2 to this case. Namely, we can investigate the relative difference of two dendrograms by the measure $\rho(R_A, R_B)$. We shall verify the validity of our consideration by simple illustrations.

Example 7.

We put $E=\{1, 2, 3, 4\}$ and denote two dendrograms by $\langle H_A, h\rangle$, $\langle H_B, t\rangle$ namely,

$$H_A = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,4\}, \{2,3\}, \{1,4,2,3\}\}$$
$$\{h_\alpha\} = \{h_0, h_1, h_2, h_3\} = \{1.0,\ 0.8,\ 0.6,\ 0.4\}$$

and

$$H_B = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,3\}, \{2,4\}, \{1,3,2,4\}\}$$
$$\{t_\alpha\} = \{t_0, t_1, t_2, t_3\} = \{1.0,\ 0.7,\ 0.5,\ 0.3\}\ .$$

Accordingly, we can obtain two relations, R_A and R_B, from $<H_A,h>$ and $<H_B,t>$, respectively.

$$R_A = \begin{bmatrix} 1.0 & 0.4 & 0.4 & 0.8 \\ & 1.0 & 0.6 & 0.4 \\ & & 1.0 & 0.4 \\ & & & 1.0 \end{bmatrix} \quad \text{for } <H_A , h>$$

$$R_B = \begin{bmatrix} 1.0 & 0.3 & 0.3 & 0.5 \\ & 1.0 & 0.7 & 0.3 \\ & & 1.0 & 0.3 \\ & & & 1.0 \end{bmatrix} \quad \text{for } <H_B , t>$$

Therefore, using the expression (39),

$$\rho(R_A,R_B) = (\rho_{ij}) = \begin{bmatrix} 0.0 & 0.4 & 0.4 & 0.5 \\ & 0.0 & 0.4 & 0.4 \\ & & 0.0 & 0.4 \\ & & & 0.0 \end{bmatrix}$$

and

$$r = \sum_{i<j}\sum \rho_{ij} = 2.5$$

Thus we can see the relative difference or association between R_A and R_B.

Example 8.

We shall examine a degree of the association between the methods. Citing again the relation R in Example 4-(2), we compare the results of three methods, that is, single linkage, complete linkage and weighted-pair group. Let us represent each dendrogram formed from these three methods by R_s, R_c, and R_w, respectively.

$$R_s = \begin{bmatrix} 1.0 & 0.7 & 0.7 & 0.8 \\ & 1.0 & 0.9 & 0.7 \\ & & 1.0 & 0.7 \\ & & & 1.0 \end{bmatrix}$$

$$R_c = \begin{bmatrix} 1.0 & 0.5 & 0.5 & 0.8 \\ & 1.0 & 0.9 & 0.5 \\ & & 1.0 & 0.5 \\ & & & 1.0 \end{bmatrix}$$

$$R_w = \begin{bmatrix} 1.0 & 0.6 & 0.6 & 0.8 \\ & 1.0 & 0.9 & 0.6 \\ & & 1.0 & 0.6 \\ & & & 1.0 \end{bmatrix}$$

Thus

$$\rho(R_s\ ,R_c) = \left[\begin{array}{cccc} 0.0 & 0.5 & 0.5 & 0.2 \\ & 0.0 & 0.1 & 0.2 \\ & & 0.0 & 0.5 \\ & & & 0.0 \end{array}\right]$$

Therefore $\rho_1 = \|\rho(R_c, R_s)\| = 2.3$.

Furthermore, in like manner,

$$\rho_2 = \|\rho(R_s, R_w)\| = 1.9 \ .$$

since $\rho_2 < \rho_1$, finally, it may be observed that R_s is close to R_w rather than R_c.

Secondly, we shall think a procedure which compares the set of partitions generated from the two dendrograms, which are obtained by applying different methods to the same data.Let us again denote two dendrograms by $<H_A,h>$, $<H_B,t>$ and represent those relations (i.e. similitude relations) by R_A, R_B. Then these similitude relations may be decomposed in the following form

$$\begin{aligned} R_A &= \bigvee_{h_l} h_l \cdot \underline{R}_A(l) \qquad (\ 0 \le h_l \le 1\ ) \\ R_B &= \bigvee_{t_m} t_m \cdot \underline{R}_B(m) \qquad (\ 0 \le t_m \le 1\ ) \end{aligned} \tag{46}$$

where $\underline{R}$ are equivalence relations in the sense of ordinary set theory, and $h_l\underline{R}_A$ or $t_m\underline{R}_B$ shows that all the elements of the ordinary relation $\underline{R}_A$ or $\underline{R}_B$ are multiplied by h_l or h_B. For example, if

$$R = \left[\begin{array}{cccc} 1.0 & 0.3 & 0.2 & 0.5 \\ & 1.0 & 0.2 & 0.3 \\ & & 1.0 & 0.2 \\ & & & 1.0 \end{array}\right]$$

Then,

$$h_0 = 1.0,\quad h_1 = 0.5,\quad h_2 = 0.3,\quad h_3 = 0.2\ .$$

Thus,

$$R = \bigvee_{h_\alpha} h_\alpha \cdot \underline{R}(\alpha)$$

$$= \max \left\{ 1.0 \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.0)} , \; 0.5 \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.5)} \right.$$

$$\left. 0.3 \cdot \underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 \\ & 1 & 0 & 1 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.3)} , \; 0.2 \cdot \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.2)} \right\} \tag{47}$$

Especially, we try to cut the two dendrograms at a level $\alpha (0 \le \alpha \le 1)$. And we assume $h_l > \alpha > h_{l+1}$, $t_m > \alpha > t_{m+1}$ for the cut at the level α. Then we can obtain two partitioning sets,

$$\begin{aligned} C_A^l &= \{A_1, A_2, \cdots, A_K\} \quad \text{where } K = n - l \\ C_B^m &= \{B_1, B_2, \cdots, B_L\} \quad \text{where } L = n - m \end{aligned} \tag{48}$$

This situation may be shown schematically as Figure 6.

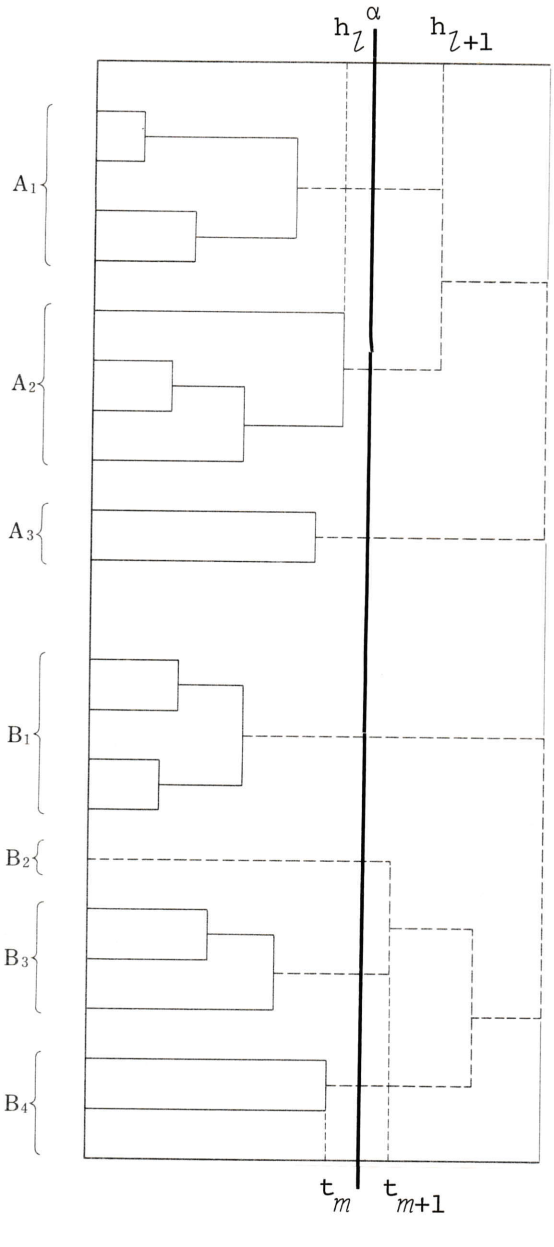


Figure 6. To compare the set of partitions produced from the two dendrograms.

Therefore, using the relationship of decomposition for a similitude relation, namely (46), we can generate the two relations

$$R^*_A = \bigvee_{h_l} h_l \underline{R}_A(l) \qquad (\alpha < h_l) \tag{49}$$

$$R^*_B = \bigvee_{t_m} t_m \underline{R}_B(m) \qquad (\alpha < t_m) \tag{50}$$

In this case, firstly, it is reasonable to consider $\rho(R^*_A, R^*_B)$ as an index for the comparison between two partitions. However, if we turn our attention to the connectedness between objects rather than the difference between trees, it may be seemed that it is natural to use the intersection of two relations, say R^*_A and R^*_B (of course in the sense of fuzzy set theory). Thus the next relationship can be defined,

$$\tau(R^*_A, R^*_B) = R^*_A \wedge R^*_B . \tag{51}$$

And let τ_{ij} denote an element of matrix $\tau(R^*_A, R^*_B)$,

$$\tau^* = \sum_{i<j}\sum \tau_{ij} \tag{52}$$

or

$$\tilde{\tau}^* = \bigvee_{i<j}\bigvee \tau_{ij} \tag{53}$$

To examine clearly what has been described previously, we shall illustrate the following example.

Example 9.

Let R_A and R_B cite from Example 7 and set the level of cut at $\alpha = 0.45$. Then,

$$R^*_A = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.8 \\ & 1.0 & 0.6 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix}$$

$$R^*_B = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.5 \\ & 1.0 & 0.7 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix}$$

accordingly, by (51),

$$\tau(R_A^*, R_B^*) = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.5 \\ & 1.0 & 0.7 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix},$$

and we can obtain $\tau^* = 1.3$ and $\tilde{\tau}^* = 0.7$.

On the other hand, if we calculate $\rho(R_A^*, R_B^*)$ using (39)

$$\rho(R_A^*, R_B^*) = (\rho_{ij}) = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.5 \\ & 0.0 & 0.4 & 0.0 \\ & & 0.0 & 0.0 \\ & & & 0.0 \end{bmatrix}$$

and $r = \sum\sum_{i<j} \rho_{ij} = 0.9$.

In addition, we shall consider another relation

$$R_C = \begin{bmatrix} 1.0 & 0.3 & 0.5 & 0.3 \\ & 1.0 & 0.3 & 0.7 \\ & & 1.0 & 0.3 \\ & & & 1.0 \end{bmatrix}.$$

Then, $\tau^* = 0$, and yet $\tilde{\tau}^* = 0$, moreover, $r = 2.6$, by using $\rho(R_A^*, R_C^*)$.

After all, we can find that τ^* or $\tilde{\tau}^*$ indicates a kind of *degree of agreement between two partitions*. That is, if τ^*'s is large then the construction of two partitions is similar to each other, if τ^*'s is small then it may be consider as the opposite. In addition, r indicates a deviation or a kink of error between relations formed by two partitions.

4.2 *Evaluation of clustering process by the sensitivity analysis*

The most fundamental problem in cluster analysis is the absence of a satisfactory definition as to what the term cluster means. Of course, as most clustering techniques highly depend upon the definition of the term *cluster* and the *nature of data*, that is, the form the type and the quantity and so on, implicit assumptions are set there about the structure present in data. In this section, the main purpose is to describe and investigate more fully for practical use the procedure of evaluating the clustering process, and to propose a procedure of the estimating the number of clusters. As we described in the previous section, single linkage, complete linkage and modified linkage are more flexible and suitable for practical use in many AHC methods. It is actually interest to indicate the number of clusters, but usually it is more difficult to do so. Empirically, it is said that examination of the dendrogram for large changes of fusion level would be useful. But it is statistical informative to determining the number of clusters based on the observation of a dendrogram. Therefore we shall serve this section to discuss or examine the connection between a brief statistical approach and our consideration described already in the previous sections.

For this purpose, firstly, we may use various consideration described the above section, namely comparing procedure between partitions. However, if a little restriction has been admitted, we can take account of another useful approach for the evaluation of clustering process. Though various attempts have been made to handle this kind of problem, the most important point is how to define the term cluster.

Most proposed definitions consist of statements such that a cluster is a set of objects which are similar to each other, and therefore objects from different clusters are dissimilar. But these definitions are fuzzy and vague. Accordingly, we have to fix a paticular definition about the shape of a cluster. Yet, partitioning the data set into various combinations of clusters, we must examine and assess the data. In this section, therefore, we shall merely make an assumption that a cluster is *spherical* and *relatively compact* in shape and that the dissimilarity or similarity is *a metric*.

Next we suppose that notations and criteria of the number of clusters are derived from the following basical relationship

$$T = W + B$$

where T is the total dispersion matrix, W is the matrix of within cluster dispersion, that is, $W = \sum_{i=1}^{k} W_i$ where W_i is the dispersion matrix for the i th cluster C_i and B is the between clusters dispersion matrix.

Then we consider four criterion which are given as follows:

[Cl] Beale's F statistic

Beale (1969) gives a criterion defined by the following expression

$$F(k_1, k_2) = [\frac{R(k_1) - R(k_2)}{R(k_2)}] / A \qquad (54)$$

$$\text{where} \quad A = \frac{n - k_1}{n - k_2} (\frac{k_2}{k_1})^{2/m} - 1$$

with $m(k_2-k_1)$ and $m(n-k_2)$ degrees of freedom.

In this expression, $R(k)$ is the residual sum of squares when the data set is divided into k clusters, namely $R(k)=tr(W)$ according to our notation.

[C2] Calinski and Harabasz's variance ratio criterion

Calinski and Harabasz (1971) suggest the use of the variance ratio criterion (VRC) given by

$$\mathrm{VRC} = \frac{\mathrm{tr}(B)}{k-1} \Big/ \frac{\mathrm{tr}(W)}{n-k}$$

$$= \mathrm{tr}(B) \cdot (n-k) / \mathrm{tr}(W) \cdot (k-1). \qquad (55)$$

In this formula, if a value of VRC monotonically increases with k, then cluster in the sense of the previous definition does not exist. If the VRC is decreasing monotonically with k, it suggests the existence of a nearly hierarchical structure. When the VRC is attaining a maximum at k, it shows the presence of k clusters. Most criteria of this kind are suggested by many authors.

[C3] Marriott's determinant criterion

Marriott (1971) has examined the properties of the following determinant criterion by experiments.

$$C = k^2 \,|W| \;/\; |T| \qquad (56)$$

where k is the number of clusters. When C is a minimum value, its value shows a desirable number of clusters.

[C4] Maronna and Jacovkis' criterion

Maronna and Jacovkis (1974) have suggested a criterion ψ^* which depends upon the within cluster covariance matrix normalized for unit determinant and investigated the property of ψ^* by a number of experiments.

$$\psi^* = \psi / m(n-k) \qquad (57)$$

$$\text{where } \psi = m \sum_{i=1}^{k} (n_i - 1) \left| W_i \right|^{\frac{1}{m}}$$

n_i is the cluster size of i th cluster and the other symbols are the same as in the previous definitions. They have proved that the value of ψ or ψ^* is monotonically decreasing. It is obvious that ψ^* is the geometrical mean of the cluster spread as measured by determinants of each cluster.

[*A procedure of the sensitivity analysis*]

We can examine the behavior of clustering processes by observing the effects of a little change in the data set to which noise is added. A procedure of the sensitivity analysis proposed here is simply summarized as follows:

[Step 1] In the first place, apply the two clustering methods, single and complete linkage , to the initial data set $X=(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)'$ and compute the four criteria at each stage of fusing.

[Step 2] Next disturb the original data set by adding a (multivariate normally distributed) noise to each data $\underline{x}_i$. Disturbed data may be written by

$$\underline{y}_i = \underline{x}_i + \underline{e}_i \qquad (i = 1, 2, 3, \cdots, n) \qquad (58)$$

where $\underline{e}_i$ is a random number generated from $N(\underline{0}, \varepsilon I)$, the constant ε being given as a value keeping of suitable size a suitable measure of rank correspondence between d_{ij} and d^*_{ij}, where d_{ij}'s are the original m dimensional data, d^*_{ij}'s added noises.

[Step 3] Carry out a clustering based on the $\underline{y}$'s and compute the four criteria. Thus the original data become vague by this procedure.

[Step 4] Repeat [Step 2] and [Step 3] until the specified number of iterations is accomplished.

[Step 5] Finally, examine the behavior of each criterion.

Thus we can decide roughly the number of clusters, say k, by examination of the behavior of these criteria following the change of k. Moreover, the reason for the use *a measure of rank correspondence* (for example, rank correlation coefficient) is mainly due to the following property.

Property 7.

In the configuration of data disturbed by adding noises, we can determine almost uniquely the region which remains unchanged the construction of hierarchical structure produced by the initial configuration of data. That is, there exists almost surely the region which keeps the order of the successive fusion level between objects or clusters monotonically invariant. In such a case, it is said to be *global order equivalent*. If a property of global order equivalence collapses rapidly by adding noise, there do not exist the clusters that are well separable and compact. On the other hand, in spite of disturbance in data, when the dendrograms keep approximately global order equivalent, we can interprete that there exist clusters whose cohesion are stable and tight.

Here, to verify the above property, we shall illustrate a brief example as follows.

Example 10.

We shall now consider three configurations in two dimensional space as shown in Figure 7. Then the dendrogram obtained by applying single linkage to these data is really identical each other in the sence of rank order for fusing each object. Actually, it is seen that the dendrogram shown in Figure 8 indicates clearly such situation.

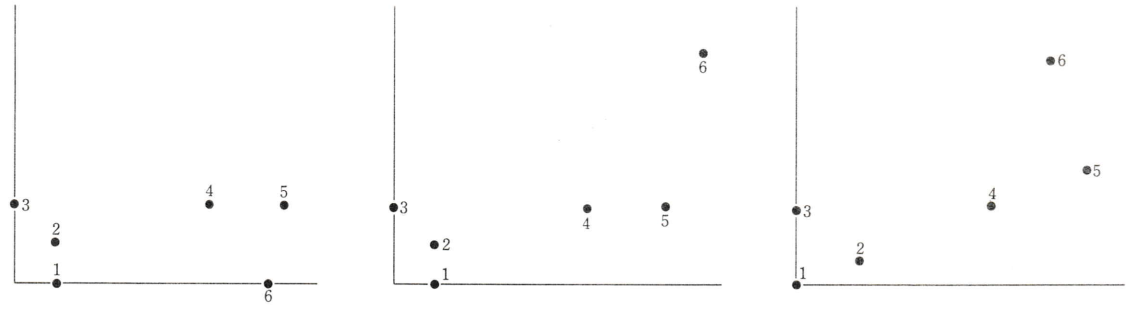


Figure 7. The sets of data which generate a dendrogram with the global order equivalent property.

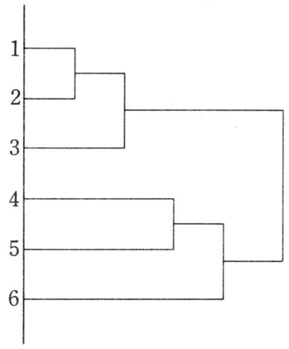


Figure 8. A dendrogram formed by applying single linkage to the data in Figure 7.

In addition, let now E={A, B, C, D} denote a data set consisted of four objects. Then the relationship between E and the dendrogram formed from E is shown as Figure 9. More precisely speaking, this figure illustrates that there exists surely a region which the object D can move freely keeping the order of the successive fusion level global equivalent.

In the above example shown in Figure 9, there are four objects in two dimensional space, but it is obvious that this situation is satisfied yet to the case of configuration consisted of many objects and higher dimensions by using the same consideration as induction. Thus we can obtain the following property.

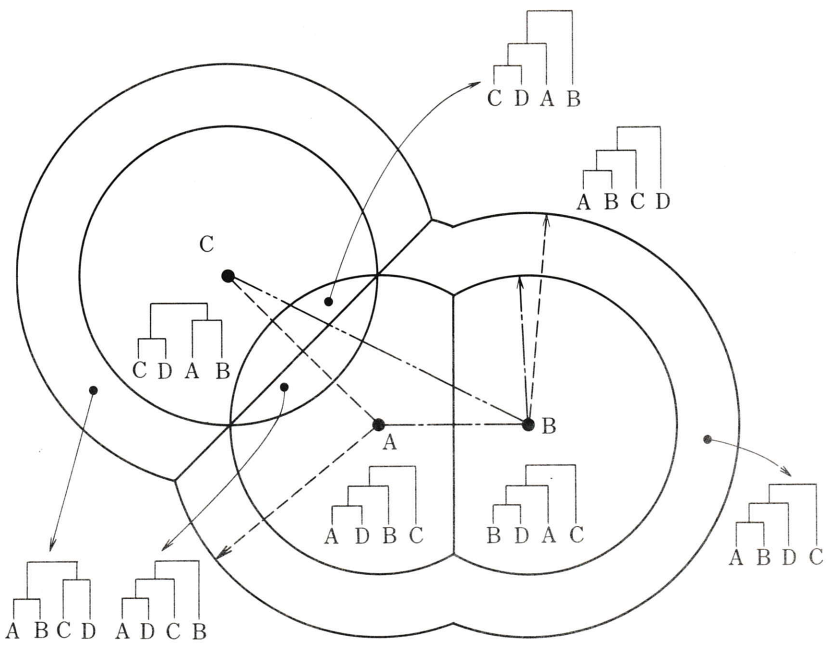


Figure 9. An example of the region which generates the property of the global order equivalence.

Property 8.

Let us consider the *n* objects consisted of multidimensional measurements, we can observe that there exists the region which generates the property of the global order equivalence in space formed by bisecting vertically between hypersphere and any two objects.

However, even though dendrograms are mutually global order equivalent, each value of a criterion for the number of clusters variously changes. Changing the number of clusters, we relatively compare the rank correspondence with the behavior of each criterion. Thus, the stability or robustness of clustering process may be evaluated more objectively. But the number of partitions of the given data is enormous, and it is really impossible to check all of them.

Since we have investigated approximately and locally a partition produced by applying a specified algorithm to the data, which is represented as a dendrogram, we must surely need the *sensitivity analysis* which could enable us to reach a reasonable interpretation as to whether there is any structure in our data or not.

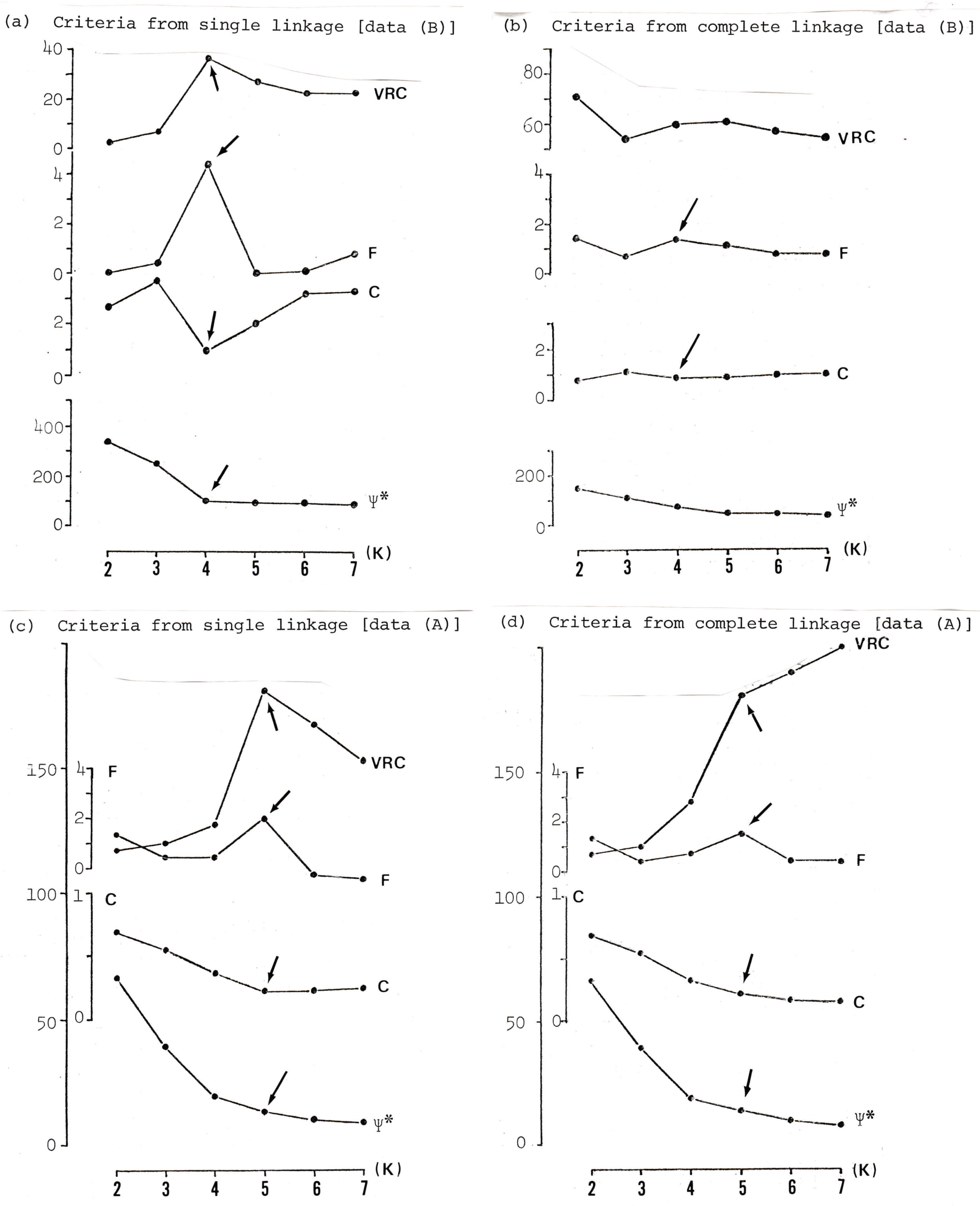
After all, the most advantage of consideration described previously is to explore more reasonable some partitions without searching among all possible partitions under a restricted situation.

Example 11.

We have prepared two sets of artificial data to examine the above procedure of sensitivity analysis. These data is the same one which used already in Example 3. Namely, the one is a data set in which there exist clusters in the sense of the assumption as shown previously, and the other is not a such one. The both data are shown in Figure 4 . All the sets of data used in our experiments consist of 50 objects and are in two-dimensional space (i.e. n=50, m=2). Firstly, the results of computation with the original data sets can be represented by Figures 10-(a), (b), examination of each criterion for data (B) suggests the following features:

(i) When the single linkage method was used, each criterion attains its optimal value for the 4 groups. When the complete linkage method is used, the results in Figure 10-(b). Even if we consider 4 groups, each criterion does not give a clear result.

(ii) In Figure 10-(b), the behavior of each criterion changing with k does not give similar results. Especially, the VRC's and ψ^*'s show



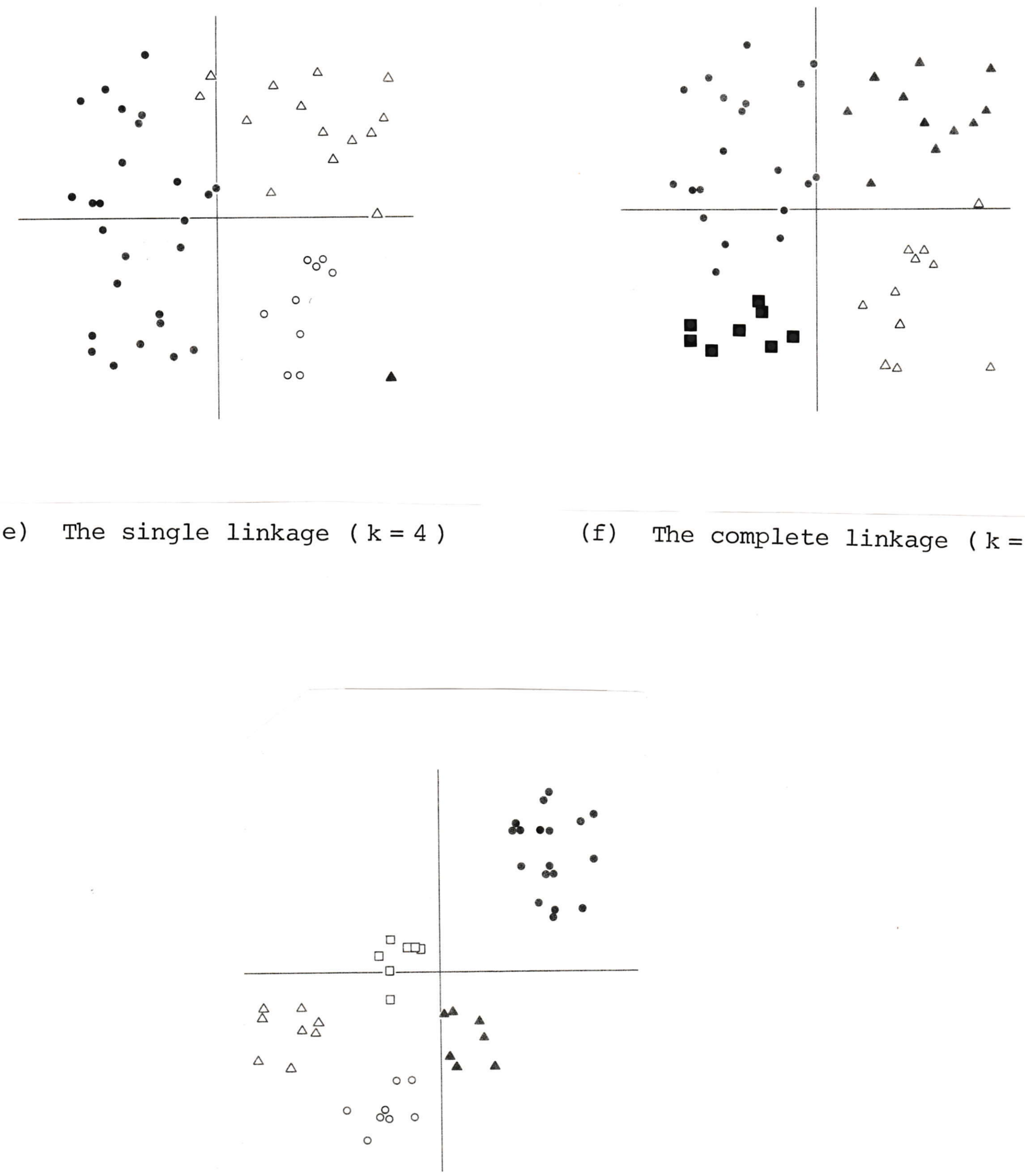
(a) Criteria from single linkage [data (B)]

(b) Criteria from complete linkage [data (B)]

(c) Criteria from single linkage [data (A)]

(d) Criteria from complete linkage [data (A)]

Figure 10. Behavior of each criterion with the change of the number of clusters, say k.



(e) The single linkage (k = 4)

(f) The complete linkage (k = 4)

(g) The single and complete linkage (k = 5)

Figure 10. (continued)

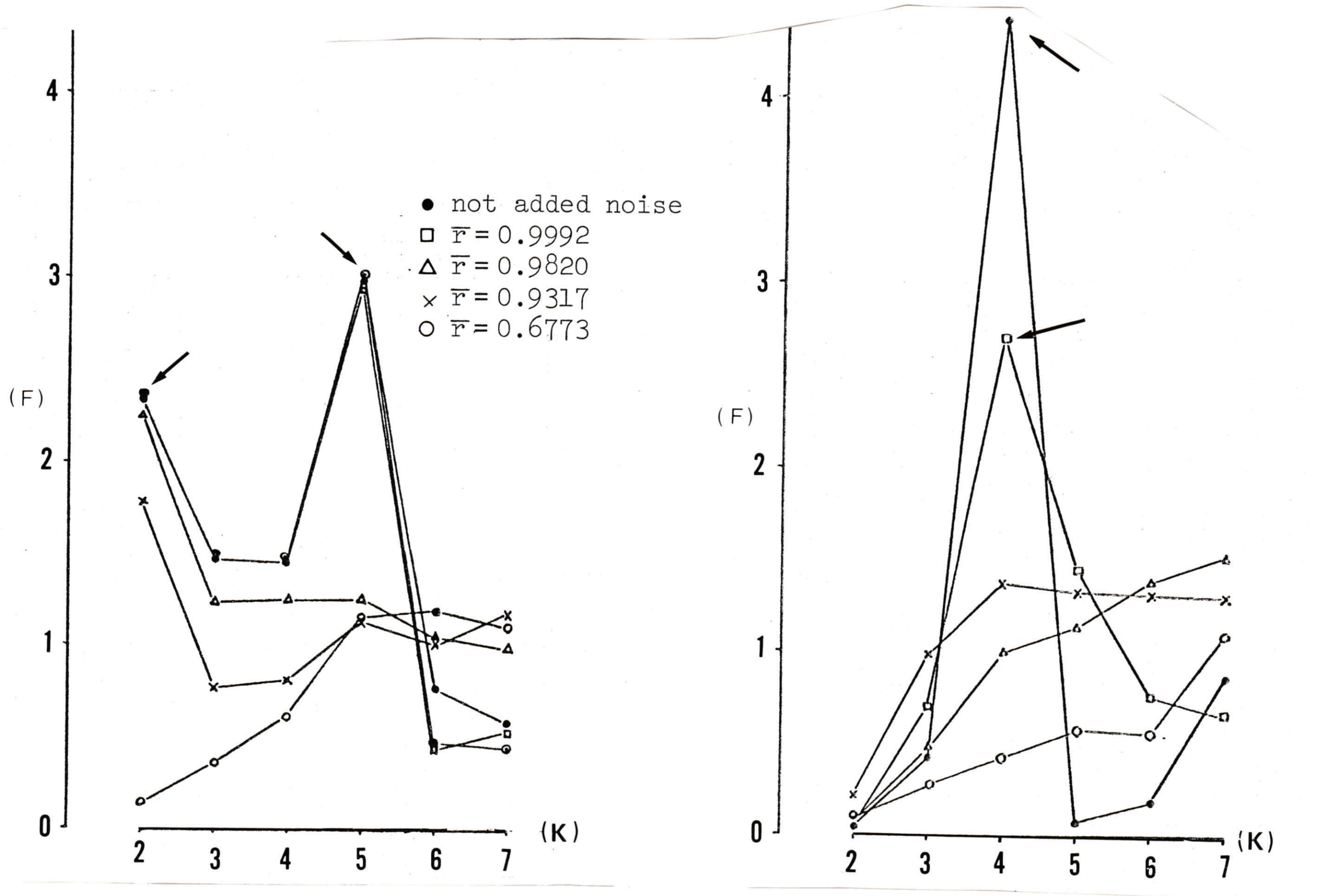
a mutually different tendency. But the F's and C's clearly attain their optimal values for 4 groups.

On the other hand, the results of analysis of the data (A) are shown in Figures 10-(c), (d). Then we can observe the following features:

(i) The number of clusters is clearly five.

(ii) The difference between the two methods is not found and the four criteria are seen to be similar in behavior. In other words, the clusters formed by the two methods show no large difference in the form. In practice, the results attained for k=5 are shown in Figure 10-(e) (single linkage), Figure 10-(f) (complete linkage). In spite of the fact that the two methods give the same decision and that the data consist of four groups, the two figures are only a little different in shape for each other. However, in the case of data (A), the two methods give almost the same results and the results are as expected when the clusters are clearly separated (Figure 10-(g)).

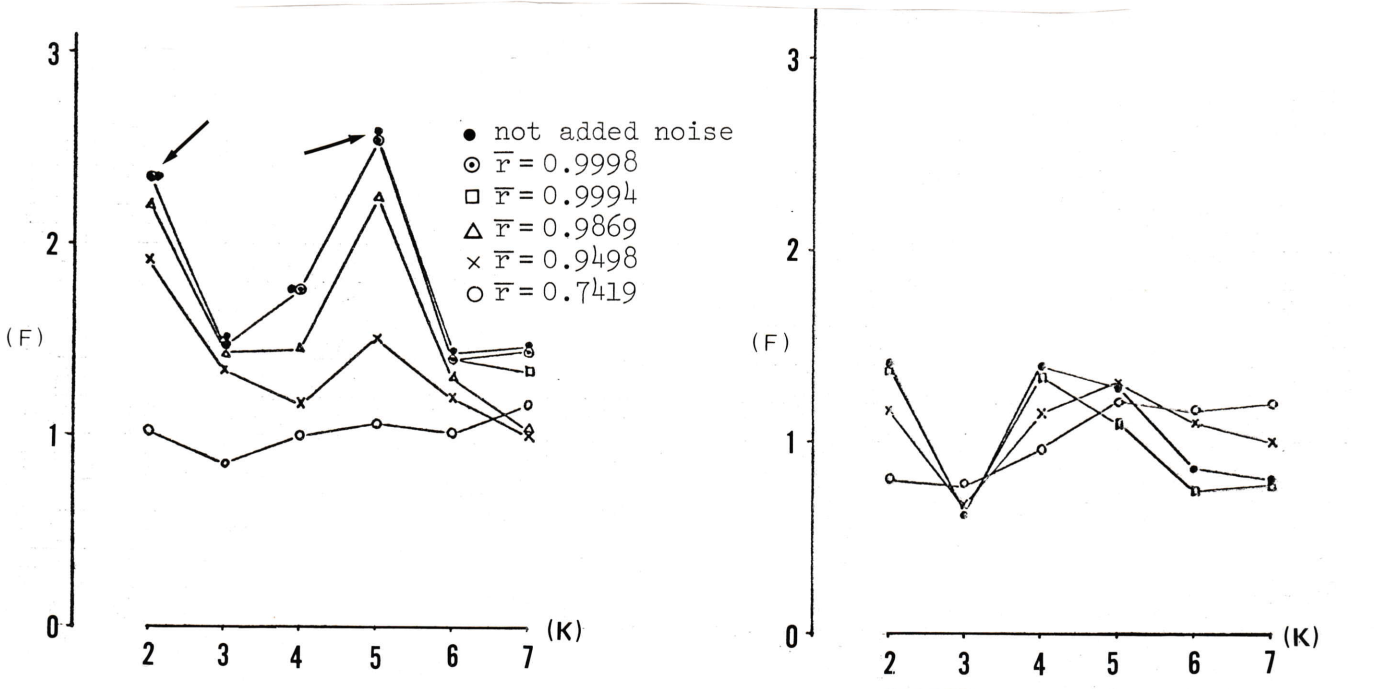
Now, let us apply the sensitivity analysis to the data (A) and (B). Beale's F criterion is mainly employed to investigate its behavior. The reason for taking up Beale's F especially here is that Beale's F criterion represents the relative quantity of the change of k and involves a noticeable effect of the dimension m. The experimental results of the sensitivity analysis are shown in Figures 11 and 12. By comparing the results obtained by the single linkage method with those by the complete linkage method, the single linkage method is seen to give more sensitive results than the complete linkage method. In other words, the complete linkage method is more stable or robust for a little change of structure in data than the single linkage method.



(a) The single linkage for data (A)

(b) The single linkage for data (B)

Figure 11. The behavior of Beale's criterion



(a) The complete linkage for data (A)

(b) The complete linkage for data (B)

Figure 12. The behavior of Beale's criterion.

(*) In these figures $\overline{r}$'s are the average of rank-correlations obtained from experiment of 100 times.

In the case of the single linkage method, the larger the quantity of noises that are added to the original data, the more difficult it becomes to distinguish F's change with k. Figures 12-(a),(b) are typical examples. On the other hand, even though the complete linkage method is a little inferior to the single linkage method as to the ability of detecting the number of clusters, the behavior of F's is quite stable and presents a reasonable basis for detecting compactly spherical clusters. Obseving the change of the criteria with k, we can find a clue to the number of clusters and obtain heuristically a useful procedure which is suitable for investigating the tendency implicitly included in data which is usually unknown.

Finally it is proposed that an extremely useful procedure in exploring data structure is to combine the *graph theoretic method* and the *sensitivity analysis method.* The graph theoretic method rather investigates the link-like relationship between objects than look for clusters. The sensitivity analysis method using the typical criteria (that indicates the existence of clusters, for example, of spherical shape) investigates the stability of the structure in data. We call this combination of the two techniques the *hybrid procedure.* Furthermore considering based on the above discussion, we have suggested a procedure which compares and evaluate, firstly, between partitions or methods, and next estimate the number of clusters by sensitivity analysis under a brief restriction, namely such as shape of clusters, metric relations, and so on.

The sensitivity analysis described above is a powerful tool which gives a clue of solution for case iii) in the problems stated at the beginning of this section. However assumption imposed on the procedure is more strictly and, therefore, is not practical. A disadvantage of traditional or statistical technique is that the use of them is unlikely to succeed in analyzing the given data, since the really data are vague and slightly unreasonable.

But we can now suggest a procedure of examining the number of clusters from a view point of fuzzy theory. Generally we can obtain a solution by applying a clustering method to a given data. Accordingly it is very difficult to investigate a set of partitions based on the only dendrogram. Of course this set of partitions is local optima, since verification of all possible partitions cannot be carried out. Therefore generating many dendrograms by adding noise to a given data, according to the consideration in section 4.1, we can compare those dendrograms with the only dendrogram derived from original data. Thus if we assume that objects have slightly a metric property, we can investigate practically and effectively the property of partitions with the procedure in section 4.1. In other words, comparison between the disturbed dendrograms and original one suggests some procedure of quantitative evaluation of clustering process. Especially, at least, a dendrogram derived from original data is an approximate solution. Therefore, without checking all possible partitions, it is reasonable to enhance the latent tendency of data and to examine more precisely the behavior of many dendrograms obtained by adding noise. Thus we can evaluate and estimate the partitions or the comparisons between several classifications. In order to interprete fits and errors between partitions it is easily understood that the concepts of fuzzy theory are more natural and valid.

5. *Several illustrations and a short discussion*

Several considerations for the AHC methods were described in the previous sections. In this section we shall attempt to illustrate some examples of clustering procedure by applying our proposal to the sets of data obtained practically.

Example 12.

The first analysis is of the set of data used by Peay (1975), which is taken from Parkman and Sawyer (1967). The raw data consisted of the numbers of marriages occuring between members of different ethnic groups in Hawaii. The measure is normalized for overall marriage rates adjusted to indicate a kind of disparity measure. But in our illustration this measure is transformed into an agreement rate. Accordingly the larger the value, the larger the intergroup marriage rate too. The name of ethnic groups included (i.e. objects), and the identified numbers to them are listed as follows:

O_1 : Hawaiian O_2 : Part - Hawaiian

O_3 : Caucasian O_4 : Puerto Rican

O_5 : Fillipino O_6 : Chinese

O_7 : Japanese O_8 : Korean

Thus the given raw data is shown in Table 5.

Firstly we shall examine the results obtained by applying single linkage and complete linkage to the similarity matrix in Table 5. Table 6 shows $\Delta = (\tilde{\delta}_{ij})$ produced by single linkage and Table 7 is $\nabla = (\underset{\sim}{\delta}_{ij})$ by complete linkage. And the dendrograms as shown in Figures 13, 14 are produced from these Δ and ∇ . Furthermore the fuzzy degree of fitness r^* indicate the following values.

Table 5. The similarity matrix S for Peay's example.

	1	2	3	4	5	6	7	8
1	1.00	0.94	0.79	0.70	0.82	0.73	0.67	0.68
2		1.00	0.88	0.79	0.86	0.84	0.77	0.77
3			1.00	0.80	0.78	0.76	0.76	0.80
4				1.00	0.81	0.63	0.59	0.63
5					1.00	0.70	0.70	0.72
6						1.00	0.76	0.79
7							1.00	0.80
8								1.00

Table 6. $\Delta = (\tilde{\delta}_{ij})$ obtained by single linkage.

	1	2	3	4	5	6	7	8
1	1.00	0.94	0.88	0.81	0.86	0.84	0.80	0.80
2		1.00	0.88	0.81	0.86	0.84	0.80	0.80
3			1.00	0.81	0.86	0.84	0.80	0.80
4				1.00	0.81	0.81	0.80	0.80
5					1.00	0.84	0.80	0.80
6						1.00	0.80	0.80
7							1.00	0.80
8								1.00

Table 7. $\nabla = (\underset{\sim}{\delta}_{ij})$ obtained by complete linkage.

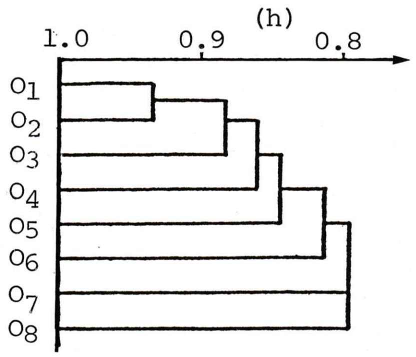
	1	2	3	4	5	6	7	8
1	1.00	0.94	0.70	0.70	0.82	0.59	0.59	0.59
2		1.00	0.70	0.70	0.82	0.59	0.59	0.59
3			1.00	0.80	0.70	0.59	0.59	0.59
4				1.00	0.70	0.59	0.59	0.59
5					1.00	0.59	0.59	0.59
6						1.00	0.76	0.76
7							1.00	0.80
8								1.00

i) if complete linkage, $r^* = 0.173$

ii) if WPG method, $r^* = 0.058$

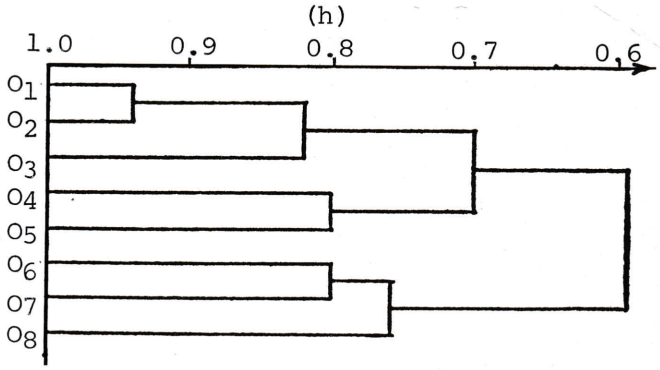
iii) if single linkage, $r^* = 0.000$

Therefore it is easily seen that there exist fairly fitted solution between single linkage and WPG. Especially we can observe the best fitted solution in the case of single linkage. However observing the dendrogram formed by single linkage (i.e. Figure 13), we can detect the existence of so-called *chaining-effect*.



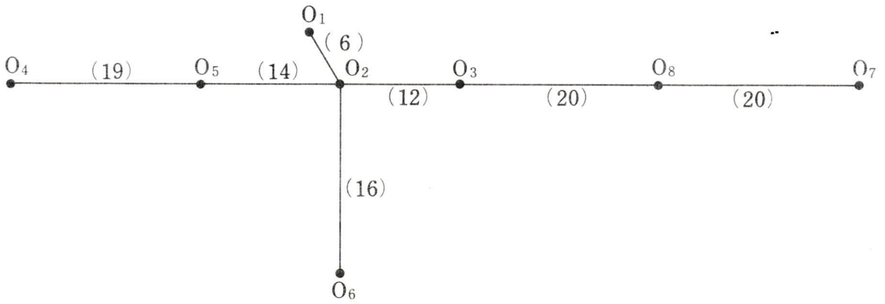
Dendrogram formed from Table 6.

Figure 13.



Dendrogram formed from Table 7.

Figure 14.



Fibure 15. MST generated from the matrix Δ of Table 6.

And that also shows no large changes in hierarchical level. Accordingly, in general, one has been considered that such situation is undesirable in the meaning that the data contains no group structure.

On the other hand, as shown in Figure 14, complete linkage produces a dendrogram which large changes in level, especially going from three groups to two groups. Thus one has determined just like one that there exist explicitly clusters. However the judgement is poor, since we can obtain another information by the investigation of MST formed from Δ. Therefore, in the following, we shall try to make MST based on Δ. The result is shown in Figure 15. This enables us to examine visually and intuitively relationships between objects.

For example, there exists slightly the connectedness between O_4 and O_7. But O_1 and O_2 are very closely related. Furthermore we can observe the similar situation between O_2 and O_5, or O_2 and O_3. After all, in this example, the link-like information between the objects plays an important role for the purpose of interpreting and exploring the data.

Example 13.

We shall illustrate an example that has examined the feature of nucleotide conformations observed in yeast phenylalanine tRNA (i.e. $\mathrm{tRNA}^{\mathrm{Phe}}$) [Kitamura et al. (1977)]. Generally it has been well known that the molecular structure of $\mathrm{tRNA}^{\mathrm{Phe}}$ give a clue to investigation about nucleotide conformations. Therefore it is necessary to think some reasonable procedure for finding the common and specific conformations which may regulate the molecular geormetry and the conformation in tRNA. Hence it is highly natural to apply cluster analysis to the solution of this problem.

We now attempt to classify the conformations consisted of 74 nucleotide units (i.e. objects), and measurements are seven torsion angles observed to each unit. Since the torsion angles are a kind of directional data, we shall denote by $d_{ij}(\omega)$ the length of the chord between i and j on a unit circle for torsion angle ω as shown in Figure 16. Namely,

$$\begin{aligned} d^2_{ij}(\omega) &= (\sin\omega_i - \sin\omega_j)^2 + (\cos\omega_i - \cos\omega_j)^2 \\ &= \{1 - \cos(\omega_i - \omega_j)\} . \end{aligned}$$

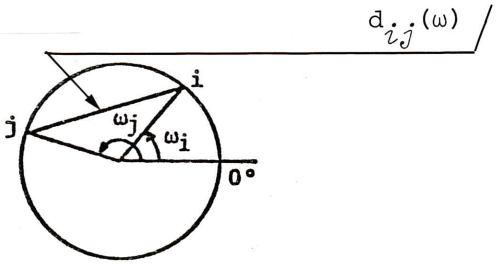


Figure 16.

Therefore let us make the sum of $d^2_{ij}(\omega^t)$ about the seven torsion angles ω^t $(t = 1, 2, \cdots, 7)$. Then we can obtain a kind of distance $d_{ij} = \{\sum_{t=1}^{7} d^2_{ij}(\omega^t)\}^{1/2}$ $(i, j = 1, 2, \cdots, 74)$. Thus the clustering may be carried out by using the distance matrix calculated by the above expression from data set. Here, the distance matrix is normalized and transformed into a similarity matrix by the same procedure as Example 6. Next applying single linkage and complete linkage to the similarity matrix, we can obtain the following result. That is,

i) if complete linkage, $r^* = 0.283$

ii) if single linkage, $r^* = 0.266$.

Thus, in this case, it is seemed that the result of single linkage is slightly better than that of complete linkage, but almost same. But in this case the number of reachability to the transitive closure indicates nine. This value is relatively very small in comparison to the order of initial similarity matrix, say n=74. Thus it is suggested that the given data is very close to the hierarchical structure. Accordingly, it is useful and suitable for the following analysis that MST is chosen rather than the dendrogram. Hence we shall make MST from the result of single linkage or transitive closure. This result are illustrated in Figure 17. Investigating this figure, we can find the significant features that the properties of MST are more reasonable and very well agreed with the suggestions of many experts from an empirical or specialistic point of view.

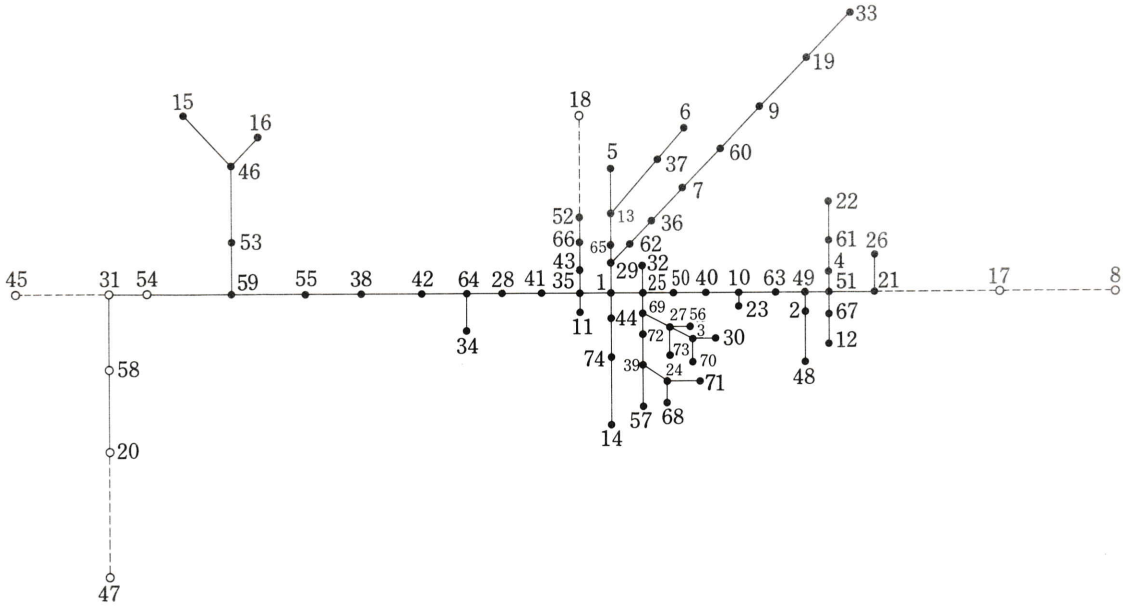


Figure 17. MST produced from nucleotide conformations of $\mathrm{tRNA}^{\mathrm{Phe}}$.

However, at present, we cannot examine the number of clusters based only the above result. Therefore we shall try to estimate the number of clusters by using four criteria described already in the previous section 4.2. The computational results are shown in Table 8. Of course, these results are not conclusive but there is some useful information. For example, it is easily seen that these criteria computed by changing k variously show almost the same behavior except ψ^*, and that they attains also its reasonable value for the seven groups. Forming seven clusters on MST in Figure 17 and classifying by the use of various symbols, we can observe the fact that the configuration of MST indicates that there exist two groups in the data. That is,

Group A : one main cluster consisted of 65 objects (indicated by symbol "•" in Figure 17).

Group B : several small clusters, of which one consists of 4 objects and another consist of singleton mutually. (indicated by symbol "∘" or connected by the dotted line in Figure 17).

Table 8.

# of clusters criteria	2	3	4	5	6	7	8
Calinski's VRC	4.946	5.667	4.177	3.795	4.179	6.216	5.715
Beale's F	0.575	1.141	0.294	0.734	1.725	5.085	0.915
Marriott's C	2.095	1.625	1.666	1.140	0.239	0.027	0.024
Maronna's ψ^*	0.060	0.053	0.053	0.050	0.044	0.033	0.034

If the number of clusters are increasing, then Group A is hierarchically subdivided into several clusters. Hence it is predicted that the general tendency of organization is constitueted by Group A.

In conclusion, we shall attempt to summarize some of the suggestions already described in the previous sections. Above all our main purpose has been to examine several properties which characterize the cluster analysis, especially the hierarchical clustering.

Firstly, arrangement of AHC methods suggests the fact that many methods have a similar feature in common. We have discussed consistently the generalized extension of these properties by the aid of the fuzzy set theory. Thus, it has been shown that our approach includes a natural generalization and extension for many AHC methods, especially which are similar to single linkage and complete linkage.

Next, we proposed that a degree of fitness between solutions of AHC methods and the similarity or distance of original data is investigated by a fuzzy symmetric difference. And an indicator, say fuzziness r^*, derived from a fuzzy symmetric relation makes possible comparisons among the methods.

Finally we discussed the problems of comparing between dendrograms and investigating the partitions formed on dendrograms, and proposed a practical procedure, which observes the correspondance between the dendrograms (i.e. equivalence relations) and which examines the goodness of fit between partitions generated from dendrograms. And to compare our consideration with some traditionally statistical procedures of evaluating the clustering process, we proposed an experimental procedure, say sensitivity analysis. Thus we can obtain also a procedure of estimating the number of clusters and of detecting the clusters. Examination of several experiments and practical applications showed that our proposal is available and useful. Thus we have overcome systematically many difficult problems included in most of AHC methods which have been said to be empirical and subjective up to now.

[References]

[1] Anderberg, M.R. (1973): *Cluster Analysis for Applications*, Academic Press.

[2] Bailey, K.D. (1974): Cluster Analysis in *Sociological Methodology*, (ed.) Heise, D.R., Jossey-Bass.

[3] Beale, E.M.L. (1969): Euclidean Cluster Analysis, *Bull. I.S.I.*, vol. 43, Book 2, pp. 92-94.

[4] Benzécri, J. (1973): *L'Analyse des Données:* Tom I (La Taxinomie), Dunod.

[5] Bijnen, E.J. and Stouthard, Ph. C. (1973): *Cluster Analysis: Survey and Evaluation of Techniques*, Tiburg Univ. Press.

[6] Blashfield, R.K. (1976): A Consumer Report on the Versatility and User Manuals of Cluster Analysis Software , *Proc. of the Statistical Computing Section*. ASA, pp. 31-37.

[7] Blashfield, R.K. and Aldenderfer, M.S. (1977): A Consumer Report on Cluster Analysis Software: (1)-(4).

[8] Bock, H.H. (1974): *Automatische Klassifikation*, Vandenhoeck & Ruprecht.

[9] Calinski, T. and Harabasz, J. (1974): A Dendrite Method for Cluster Analysis, *Communication in Statistics*, vol.3 , No.1, pp. 1-27.

[10] Clifford, H.T. and Stephenson, W. (1975): *An Introduction to Numerical Classification*, Academic Press.

[11] Cole, A.J. (ed.) (1969): *Numerical Taxonomy*, Academic Press.

[12] Cormack, R.M. (1971): A Review of Classification, *Journal of Royal Statistical Society, Series A*, 134, No.3, pp. 321-367.

[13] Duran, B.S. and Odell, P.L. (1974): *Cluster Analysis: A Survey*, Springer Verlag.

[14] Enslein, K., Ralston, A. and Wilf, H.S. (1977): *Statistical Methods for Digital Computers*, Vol. III of *Mathematical Methods for Digital Computers*, John Wiley.

[15] Everitt, B. (1974): *Cluster Analysis*, John Wiley.

[16] Everitt, B. (1977): Cluster Analysis, in *Exploring Data Structures of the Analysis of Survey Data*, Vol. 1, John Wiley.

[17] Farris, J.S. (1969): On the Cophenetic Correlation Coefficient, *Systematic Zoology*, Vol. 18, pp. 279-285.

[18] Hartigan, J.A. (1967): Representation of Similarity Matrices by Trees; *J. Am. Statist. Ass.*, Vol. 62, pp. 1140-1158.

[19] Hartigan, J.A. (1975): *Clustering Algorithm*, John Wiley.

[20] Hubert, L.J. (1974): Some Applications of Graph Theory to Clustering, *Psychometrika*, Vol. 39, No. 3, pp. 283-309.

[21] Hubert, L.J. (1975): Hierarchical Clustering and the Concept of Space Distortion, *The Br. J. Math. Statist. Psychol.*, Vol. 29, No. 2, pp. 121-133.

[22] Jardine, J. and Sibson, R. (1971): *Mathematical Taxonomy*, John Wiley.

[23] Johnson, S.C. (1967): Hierarchical Clustering Schemes, *Psychometrika*, Vol. 32, No. 3, pp. 241-254.

[24] Kaufmann, A. (1973): *Introduction à la Théorie des Sous-ensembles Flous*, Tome I, Masson et Cie.; Tome II, Tome III, Tome IV (1975,1975,1977).

[25] Kitamura, K., Wakahara, A. et al. (1977): Classification of Nucleotide Conformations Observed in Yeast Phenylalanine tRNA: Application of Cluster Analysis.

[26] Kruskal, J.B. (1964): Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, *Psychometrika*, Vol. 29, pp. 1-27.

[27] Lance, G.N. and Williams, W.T. (1967): A General Theory of Classificatory Sorting Strategies: I. Hierarchical Systems, *The Computer Journal*, Vol. 9, pp. 373-380.

[28] Lerman, I.C. (1970): *Les Bases de la Classification Automatique*, Gauthier-Villars.

[29] Maronna, R. and Jacovkis, P.M. (1974): Multivariate Clustering Procedures with Variable Metrics, *Biometrics*, Vol. 30, pp. 499-505.

[30] Marriott, F.H.C. (1971): Practical Problems in a Method of Cluster Analysis, *Biometrics*, Vol. 27, pp. 501-514.

[31] Matusita, K. and Ohsumi, N. (1978): Evaluation Procedure of Clustering Techniques, *France-Japan Seminar*, Paris, March. pp. 13-20.

[32] McQuitty, L.L. (1956): Agreement Analysis, Classifying Persons by Predominant Patterns of Response, *British J. Stat. Psych.* Vol. 9, pp. 5-16.

[33] McQuitty, L.L. (1957): Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies, *Educational Psychological Measurement*, Vol. 17, pp. 207-229.

[34] McQuitty, L.L. (1960 a): Hierarchical Linkage Analysis for the Isolation of Types, *Educational Psych. Measurement*, Vol. 20, pp. 55-67.

[35] McQuitty, L.L. (1960 b): Hierarchical Syndrome Analysis, *Educational Psych. Measurement*, Vol. 20, pp. 293-304.

[36] McQuitty, L.L. (1961): Typal Analysis, *Educational Psych. Measurement*, Vol. 21, pp. 677-696.

[37] McQuitty, L.L. (1962): Multiple Hierarchical Classification of Institutions and Persons with Reference to Union-management Relations and Psychological Well-being, *Educational Psych. Measurement*, Vol. 22, pp. 513-531.

[38] McQuitty, L.L. (1963): Rank Order Typal Analysis, *Educational Psych. Measurement*, Vol. 23, pp. 55-61.

[39] Peay, E.R. (1975): Nonmetric Grouping; Clusters and Cliques, *Psychometrika*, Vol. 40, Vol. 3, pp. 297-313.

[40] Rohlf, F.J. and Sokal, R.R. (1962): The Comparison of Dendrograms by Objective Method; *Taxon*, Vol. 11, pp. 33-40.

[41] Roux, M. (1969): An Algorithm to Construct a Particular Kind of Hierarchy, in *Numerical Taxonomy*, (ed.) Cole, A.J., Academic Press.

[42] Ryszin, J. (ed.) (1977): *Classification and Clustering*, Academic Press.

[43] Sneath, P.H.A. and Sokal, R.R. (1973): *Principles of Numerical Taxonomy*, Freeman.

[44] Sokal, R.R. and Sneath, P.H.A. (1963): *Principles of Numerical Taxonony*, Freeman.

[45] Späth, H. (1975): *Cluster Analyse Algorithmen*, R. Oldenbourg Verlag.

[46] Tyron, R.C. and Bailey, D.E. (1970): *Cluster Analysis*, New York; McGraw-Hill Book Co.

[47] Wishart, D. (1969): An Algorithm for Hierarchical Classifications, *Biometrics*, Vol. 25, pp. 165-170.

[48] Zadeh, L.A. (1965): Fuzzy sets, *Information and Control*, Vol. 8, pp. 338-353.

[49] Zadeh, L.A. (1971): Similarity Relations and Fuzzy Ordering, *Inf. Sciences*, Vol. 3, pp. 177-200.

[50] Zahn, C.T. (1969): Approximating Symmetric Relations by Equivalence Relations, *J. Soc. Ind. Appl. Math.*, Vol. 12, No. 4, pp. 840-847.

[51] Zahn, C.T. (1971): Graph-theoretical Methods for Detecting and Describing Gestalt Clusters, *IEEE Transactions on Computers*, Vol. C-20, No. 1, pp. 68-86.