Control
and Cybernetics

VOL. 10 (1981) No. 1–2

Intuition vs. formalisation in clustering: local and global criteria of grouping

by

JAN W. OWSIŃSKI

Polish Academy of Sciences
Systems Research Institute
Warszawa, Poland

The methods of clustering are classified according to explicit or implicit criteria of grouping. These criteria are derived from a number of basic, intuitively obvious assumptions. In classifying the methods the local or global nature of criteria and the facility of algorithmisation are considered. A new form of global criterion is proposed which makes it possible to solve the clustering problem explicitly both for the optimal number of groups and for their composition. The criterion possesses also the essential local properties. Appropriate simple algorithm is outlined.

1. Introduction

Mathematics is a tautological system. Hence, it is often conjectured that when taking intuitively obvious elementary assumptions and rules of reasoning one should reach equally intuitively obvious results. Experience shows that this is not true. The main reason is the incapacity of predicting the far-off consequences of initial assumptions, especially when there are a number of qualitatively similar assumptions, among which a choice should be made. This is especially true when the results have highly complex and multidimensional nature.

For simple cases, however, it is possible to formulate assessments concerning the outlook of the results. When these assessments can take on a more precise and general form, they can be utilised throught the solution of the problem, together with the initial elementary assumptions. Such is, for instance, the correctional sense of some constraints in the economic problems formulated as mathematical programming tasks.

The same applies to clustering problems, met in data analysis, taxonomy, classification, pattern recognition, etc. There is a choice of elementary assumptions concerning either local distance or similitude of elements in a population, or more global in-group homogeneity vs. inter-group diversity criteria. The clustering methods in constructing their algorithms base upon these criteria. Over bigger and complex populations it is difficult to assess the adequacy of methods applied, since the results are then by no means intuitively analysable. When analysing simple examples one can easily see the inherent biasses of the algorithms. In order to get rid of these biasses or to make them controllable it is necessary to impose an explicit global criterion. Furthermore, it can be hoped that such a criterion could also help in solving the problem of the optimal number of groups, in addition to the usually solved problem of group composition.

2. The problem

Having n elements indexed i, $i \in I=\{1, ..., n\}$, for which mutual "distances" ("dissimilitudes") $d_{ij}=d_{ji} \geqslant 0$, and/or mutual "proximities" ("similarities") $w_{ij}=w_{ji} \geqslant 0$ are defined, to find partition P of the set I into subsets A_m, $P=\{A_1, ..., ..., A_m, ..., A_p\}$, i.e. $\bigcup_{m=1}^{p} A_m=I$, and $A_{m'} \cap A_{m''}=\emptyset$, $m' \neq m''$, such that the elements i belonging to each one of the groups A_m are more "similar" than those belonging to various groups.

The problem thus outlined leaves a wide margin for interpretation. Indeed, particular methods solving this problem operate on additional assumptions concerning quantitative interpretation of notions of "distance" ("dissimilitude") and "proximity" ("similarity").

To illustrate the question of this interpretation assume the points i to be located in R^r, so that each point is characterized by r quantitative features. Thus, for each pair ij one can define a distance d_{ij}, Euclidean or other, metric or not, provided that, as above, $d_{ij}=d_{ji} \geqslant 0$. There is, however, for any of these distance definitions a whole range of choices with regard to "proximities" w_{ij}.

Certainly, when a distance $d_{ij}=d_{ji}=0$ then also the corresponding dissimilarity is equal to zero, and the proximity, or similarity, $w_{ij}=w_{ji}$ reaches its maximum over all the pairs of ponts. Furthermore,

$$d_{ij} < d_{kl} \Leftrightarrow w_{ij} > w_{kl},$$

and these two seem to be the only a priori acceptable features of "similarity" and "dissimilitude".

It should also be noted that by starting from the matrices D and W, composed of d_{ij} and w_{ij}, as given, one abstracts from the physical sense of a problem, which may necessitate a particular distance/proximity measure and thereafter also a particular clustering method. Thus, the considerations herein relate solely to these problems which, notwithstanding their physical meaning, can be represented via the general formulation shown.

3. The elementary assumptions

These assumptions do in fact constitute an interpretation of "similarity" or "likeness" mentioned in the problem formulation.

A. The elements i, j are more similar than i, k iff $d_{ij} < d_{ik}$.

B. The elements belonging to groups A_m, $A_{m'}$ are more similar than those belonging to A_m, $A_{m''}$ iff $d(m, m') < d(m, m'')$, where [5],

$$\text{a.} \quad d(m, m') = \max_{\substack{i \in A_m \\ j \in A_{m'}}} d_{ij}$$

$$\text{b.} \quad d(m, m') = \min_{\substack{i \in A_m \\ j \in A_{m'}}} d_{ij}$$

$$\text{c.} \quad d(m, m') = \frac{1}{\bar{\bar{A}}_m \bar{\bar{A}}_{m'}} \sum_{\substack{i \in A_m \\ j \in A_{m'}}} d_{ij}, \text{ [4]},$$

or any other distance-like function of i, j, $i \in A_m$, $j \in A_{m'}$, e.g.

$$\text{d.} \quad d(m, m') = \frac{\bar{\bar{A}}_m \bar{\bar{A}}_{m'}}{\bar{\bar{A}}_m + \bar{\bar{A}}_{m'}} (\bar{X}_m - \bar{X}_{m'})^T (\bar{X}_m - \bar{X}_{m'})$$

where $\bar{X}_m = \frac{1}{\bar{\bar{A}}_m} \sum_{i \in A_m} x_i$, $x_i = \{x_{i1}, \ldots, x_{ir}\}$ being the vector characterising point i and $\bar{\bar{A}}_m$ — number of elements in A_m. Medians or centroids can also be used [5].

C. The elements i, j are similar iff $d_{ij} \leqslant \varepsilon$, where:

a. ε is arbitrarily chosen, [1, 6, 16],

b. ε is a function of d_{ij}, $i, j \in I$, e.g., [1],

$$\varepsilon = \frac{2}{\bar{\bar{I}}(\bar{\bar{I}} - 1)} \sum_{\substack{i, j \in I \\ i < j}} d_{ij} = \bar{d}$$

D. The elements i, j are more similar than i, k iff $\delta_{ij} < \delta_{ik}$, where for $\forall (i, j, k)$, $i, j, k \in I$, $\delta_{ij} \leqslant \sup(\delta_{ik}, \delta_{jk})$, and δ_{ij} can be obtained from d_{ij} through a simple algorithm given in [4] or in [7].

E. The elements belonging to the same groups A_m are similar, while those belonging to various groups are dissimilar, [4], iff

$$\forall A_m \in P,\ i, j \in A_m,\ k \notin A_m\colon d_{ij} \leqslant d_{ik} \wedge d_{ij} \leqslant d_{jk}.$$

F. The elements belonging to a group A_m are similar, while those belonging and not belonging to it are dissimilar, [9, 10], iff

$$W(A_m) = \sum_{\substack{i \in A_m \\ j \in A_m}} w(d_{ij}) > \sum_{\substack{i \in A_m \\ j \notin A_m}} w(d_{ij}) = w(A_m)$$

where $w(d_{ij})$ is a function of "similarity" or "linkage" decreasing in R^1_+.

4. Construction of a procedure

The initial assumptions formulated here are for the most part intuitively acceptable, if not obvious — with exception perthaps of assumption D, which refers to the notion of ultrametric [4]. Having these assumptions, being in fact local similarity/dissimilarity criteria, one can proceed to construction of groupings $A_m \in P$.

Qualification of locality refers to the fact, that the criteria cited base upon "likeness" or "dissimilitude" between individual elements i, j or groups A_m, $A_{m'}$. No criterion cited before bases upon functions of "likeness" or "dissimilitudes" defined for greater aggregates.

The groupings can be constructed directly on the basis of assumptions A.B. The method is very simple and, again, intuitively obvious. At each step two elements or groupings are merged for which appropriate d is minimal. Or those are separated for which appropriate d is maximal — see [2, 3] for some such algorithms. This way a hierarchy H is obtained.

A hierarchy H is a subset of the power set 2^I, such that

1° $I \in H$,

2° $\forall i \in I$, $\{i\} \in H$, $i = A_i^1$,

3° $\forall A_m^p, A_{m'}^{p'} \in H$ if $A_m^p \cap A_{m'}^{p'} \neq \emptyset$ then either $A_m^p \subset A_{m'}^{p'}$, for $p \leqslant p'$, or $A_{m'}^{p'} \subset A_m^p$, for $p' \geqslant p$, where $A_m^p \in P^p$ and $A_{m'}^{p'} \in P^{p'}$.

In case of the direct pair-wise application of A, B there is additionally

4° $\forall p \, \overline{\overline{P}}{}^p = \overline{\overline{P}}{}^{p-1} + 1$

where, therefore, P^p is a partition of I containing A_m^p, $m = 1, ..., p$.

Obviously, it follows directly from the above definition that for each $A_m^p \in H$ there exists a partition $P \ni A_m^p$, $P \subset H$. Condition 4° stipulates a series $\{P^p\}_{p=1}^n$ of partitions, $\{P^p\}_{p=1}^n = H$, $P^1 = \{I\}$, $P^n = I$.

Thus, for given I and D, for each of the local criteria B. a different hierarchy H (B) is obtained. The methods, related to assumptions B. are referred to as "complete linkage" (B.a), "single linkage" (B.b), "average linkage" (B.c) and Ward technique (B.d). Their generalisation was proposed in [11] and broadened in [17] to include the Ward technique. Originally the Ward technique proceeded by joining these groups, for which the increase of

$$S = \sum_{i \in A_m^p} (x_i - \bar{X}_m)^T (x_i - \bar{X}_m)$$

resulting from joining was minimal. By denoting the above value of "error sum of squares" as S_m, analogous for $A_{m'}^p$ as $S_{m'}$, the one for $A_m^p \cup A_{m'}^p$ as $S_{mm'}$, and the one from B.d. as $S_{m/m'}$ it can be shown that

$$S_{mm'} = S_m + S_{m'} + S_{m/m'}$$

Hence, the original Ward approach is equivalent to sequential hierarchical grouping with B.d.

Still another hierarchy can be obtained for a given set of elements and "distances" when using the ultrametric δ_{ij}, as defined in assumption D. In fact, the ordering of δ_{ij}, when obtained already from d_{ij}, yields directly a hierarchy.

The methods mentioned are numerically very simple, what constitutes their main merit. They do not provide, however, any measure for determining the "best" p among $p \in N[1, ..., n]$, for $\{P^p\}_1^n$. In setting up a hierarchy these methods utilise the wholly local criteria and therefore do not make it possible to compare their P^p's for any p. Thus, in spite of the obvious biasses of some of these methods in terms of "propensities" to form e.g. bigger or smaller groups, the only intuitively tangible comparisons could be made for simple, unrealistically small examples. Furthermore, many of the sequential hierarchic grouping methods cannot be easily used as classifying devices, i.e. when after *a* $\{P^p\}_1^n$ had been found, to locate an $n+1-$st element of $I^1=I \cup \{n+1\}$, and so on.

A step away from locality, with preservation of intuitiveness, although at a loss of numerical efficiency, is introduced via methods based upon assumptions C. These methods may utilise, in addition to C.a,b, other particular assumptions in order to operationally define appropriate algorithms. The most "conservative" assumption following literally C.a would be that a group A_m can be considered as such iff

$$\forall i, j \in A_m,\ k \notin A_m: d_{ij} \leqslant \varepsilon \wedge d_{ik} > \varepsilon \wedge d_{jk} > \varepsilon \qquad (C')$$

i.e. all points in an A_m are similar and no point similar to them belongs to an other group. Groups formed that way can be called ε—homogeneous. Application of the criterion of ε—homogeneity has the advantage of clear intuitive meaning for the global solution, and it also yields such unique global solution, a partition $P(\varepsilon)$, composed of $p(\varepsilon)$ groups. There is, on the other hand, the additional important burden of computations. Moreover, the criterion is rational for points $i \in I$ in a metric space, while d_{ij} may not have anything to do with formal distances. The same can be said of absolute homogeneity, i.e. the criterion E.

In practice both C.a with C′ and E are rarely used for reasons mentioned above. Homogeneity, as defined above, introduces certain globality into construction of P since in order to obtain this unique global solution all d_{ij} have to be examined at each step of the iterative procedure. The requirements of C′ and E can again be relaxed if an ultrametric δ_{ij} is introduced to C′ and E instead of d_{ij}.

Still, intuitive simplicity and obviousness of assumptions C causes a number of applications based upon them to appear. To operationalise these assumptions in the effective algorithms additional assumptions are introduced. Thus, in the FARRELL and FARRELL-mod [1] methods it is attempted to locate the spheres A_m of a given predefined radius, so that

$$\forall i \in I \quad \exists A_m \ni i,$$

cross-sections $A_m \cap A_{m'}$, $m \neq m'$, are possibly "small" in terms of $\overline{\overline{A \cap A_{m'}}}$ and the centres of A_m approximate local gravity centres.

Another simple and intuitive method thus derived is the "percolation method" [16], which defines for each i a set $V(i, \varepsilon)$,

$$V(i, \varepsilon) = \{j \in I | d_{ij} \leqslant \varepsilon\} \qquad (C''.1)$$

and then a density coefficient $v(i, \varepsilon)$, e.g.

$$v(i, \varepsilon)=\overline{\overline{V}}(i, \varepsilon) \tag{C''.2}$$

The procedure there of chooses sequentially the maximal $v(i, \varepsilon)$ and formes the $A_m \in P$ out of the non-grouped $i's$ corresponding to these $v(i, \varepsilon)$. Again, special assumptions are necessary for classification of boundary points.

Variable ε is also utilised in the method proposed by Slater [14]. This method analyses d_{ij} which are not necessarily symmetric. It postulates, however, that these values be doubly standardised, i.e. row and column sums be equalised. Thus defined distance matrix serves to develop the hierarchy by sequentially analysing links whose values are higher than the ε, decreasing from $d_{ij}^{\max}$ to zero. This way consecutive grouping patterns appear, analogously to "single linkage" or "nearest neighour" procedure. Soundness of individual groups appearing in the hierarchy is checked via special additional techniques.

The gravity method [1] starts from the assumption C.b, and groups the non-grouped $i's$ which fulfill it. Special classificatory assumptions for $i's$ to be added to existing $A'_m s$ are based upon either variance or arithmetic average statistics.

The local criterion F does not yield by itself a unique partition P, but rather a family of groups A_m^p which can be used to form a hierarchy of partitions P^p. Additional criterion, i.e. that a group A_m contains similar elements, [9, 10], iff

$$\forall a_m \subsetneqq A_m, \; w(a_m) > w(A_m) \tag{F'}$$

is used to operationalise a method called minimally interconnected subnetworks technique. The criterion F′ can hardly be referred to as intuitive. It yields, however, very good computational properties. A hierarchy $H^w=\{P^p\}_1^q$ obtained via this method, is much "flatter" than the "full" hierarchies obtained from assumptions B, i.e. $q<n$. The method has, in fact, a "bias" towards "greater" groups (see Appendix 1).

Condition similar to F′ serves in [15] to define so called nodal regions-groups, i.e. such that have weaker links (smaller distances) to other regions than the elements being the nodes of the groups. Because of specific formulation this problem is approached similarly as in [14].

5. Objective functions

The method described heretofore were referred to as local and this feature was said to be partly offset by the intuitively obvious nature of most of the assumptions serving to set up the appropriate procedure. This qualification of locality should be again commented upon. Since it is a partition P (or a hierarchy H) that is being sought, the qualifications of locality or globality should in fact refer to a capacity of search in the space of $P's$ (or $H's$). From this point of view the methods mentioned could not even be called local insofar as they do not offer any possibility of comparison and choice among various $P's$ ($H's$). They just determine one P (or one H), and those which determine an H usually do not provide any possibility of choosing a P^p out of $\{P^p\}_1^n = H$. Certainly, each of these latter methods can be complemented with measures of the — relative! — group stability, as it was done by Slater or Tremolières. Thus, through the methodological back door some possibility of comparison is introduced.

Obviously, the possibility of comparison can only be realised through simultaneous explicit accounting for all groupings entering a P. Hence, an overall objective function or its proxy should be constructed.

Another comment refers to the nature of the iterative numerical processes leading to establishment of solution. As we have seen on the example of assumptions B and C the essential change in the character of the solution (H or P) does not necessarily entail a change in the nature of the iterative process (although it may). Thus, in assessing a method more attention should be paid to uniqueness and comparativeness of final results rather than to the course of the procedure.

A number of objective functions have been proposed for solving the grouping problem. Some of them are presented below.

C. Partition P^p is better than partition P^{p^*} iff, [5, 8],

$$\sum_{m=1}^{p} \frac{1}{\bar{A}_m} \sum_{\substack{i,j \in A_m \\ i<j}} d_{ij}^2 < \sum_{m=1}^{p} \frac{1}{\bar{A}_m^*} \sum_{\substack{i,j \in A_m^* \\ i<j}} d_{ij}^2$$

H. Partition P^p is better than partition P^{p^*} iff, [5],

$$\sum_{m=1}^{p} \frac{2}{\bar{A}_m(\bar{A}_m - 1)} \sum_{\substack{i,j \in A_m \\ i<j}} d_{ij}^2 < \sum_{m=1}^{p} \frac{2}{\bar{A}_m^*(\bar{A}_m^* - 1)} \sum_{\substack{i,j \in A_m^* \\ i<j}} d_{ij}^2$$

I. Partition P^p is better than partition P^{p^*} iff, [5],

$$\max_m \left(\max_{i,j \in A_m} d_{ij} \right) < \max_m \left(\max_{i,j \in A_m^*} d_{ij} \right)$$

K. Partition P^p is better than partition P^{p^*} iff

$$\sum_{\substack{m,m'=1 \\ m<m'}}^{p} d(m,m') > \sum_{\substack{m,m'=1 \\ m<m'}}^{p} d(m,m')^*$$

where $d(m,m')$, $d(m,m')^*$ are inter-group distances defined in any of the ways given in assumption B, for partitions P^p and P^{p^*}, respectively.

L. Partition P^p is better than partition P^{p^*} iff, [12],

$$\sum_{i \in I} \sum_{j \in J} d_{ij} x_{ij} < \sum_{i \in I} \sum_{j \in J} d_{ij} x_{ij}^*$$

where $J \subset I$ is a set of eligible centres (usually $J=I$), and

$$\sum_{j \in J} x_{ij} = 1 \ \forall i$$

$$\sum_{j \in J} x_{jj} = p$$

$$x_{ij} \leqslant x_{jj} \ \forall i,j$$

$$x_{ij} \in \{0,1\} \ \forall i,j$$

which indicates a 0–1 programming problem. It is being solved via subgradient method applied to the relaxed Lagrangian form of the initial problem.

M. Partition P^p is better than partition P^{p^*} iff, [5],

$$\frac{S_0^2}{S_I^2} > \frac{S_0^{*2}}{S_I^{*2}}$$

where $S_0^2 = \sum_{\substack{m, m'=1 \\ m<m'}}^{p} d^2(m, m')$, and $S_I^2 = \sum_{m=1}^{p} \frac{1}{\bar{A}_m} \sum_{\substack{i, j \in A_m \\ i<j}} d_{ij}^2$

N. Partition P^p is better than partition P^{p^*} iff, [4],

$$W(L^p, P^p) < W(L^{p^*}, P^{p^*})$$

where L^p is a set $\{1, ..., l_p\} \subset I$ of the "representants" of groups $A_1, ..., A_m$, such that $l_m \subset A_m$ (in particular, $\bar{l}_m = 1$), and the function W can be defined by

$$W(L, P) = \sum_{m=1}^{p} \sum_{i \in l_m} \sum_{j \in A_m} d_{ij}$$

All the objective functions presented provide a comparison of "goodness" of partitions P^p, i.e. partitions of I into a given number of groups p. This is caused by the fact that the functions G through N display similar characteristics as procedures built upon assumptions B through F, i.e. they refer to only one side of the initial problem, either internal homogeneity of groups or their external dissimilitude. The same applies to M since the quotient proposed amplifies the one-sided effect rather than balances the two effects. Thus, the values of these objective functions for optimal P^p are monotone with regard to p.

A point which gains in importance with introduction of the objective functions is numerical efficiency of algorithms. The merit of most of the procedures based upon the local assumptions A through F lied in their simplicity. The same can hardly be said on procedures for optimising with regard to G through N. Some of these procedures refer to dynamic programming philosophy [8], some other utilise special iterative algorithms based upon properties connecting local (with respect to A_m) and global (with respect to P) optima [4]. The algorithms proposed by Diday [4] has essential numerical advantages, especially from the point of view of memory requirements. It necessitates, however, a good initial guess and does not safeguard against cycling phenomena. Algorithm similar in its principle to the one given in [4] is presented in [18] for the objective function similar to K with $(x_i - \bar{X}_m)^T (x_i - \bar{X}_m)$ instead of d_{ij}^2. Dynamic programming, robust in reaching solutions is more cumbersome in calculations.

Thus, in order to solve the grouping problem in its absolute form, i.e. together with p, different objective functions have to be developed. This, however, can make the computational problems even more difficult.

Because of that, it seems, there have been very little efforts aiming at construction of such global objective functions and the corresponding algorithms. An example of such function is cited verbally in [6] after Holzinger and Tryon.

O. Partition P is better than partition P^* iff

$$\frac{\bar{D}_0}{\bar{D}_I} > \frac{\bar{D}_0^*}{\bar{D}_I^*}$$

where

$$\bar{D}_0 = \frac{1}{\sum_{m=1}^{p} \bar{\bar{A}}_m (n - \bar{\bar{A}}_m)} \sum_{m=1}^{p} \sum_{i \in A_m} \sum_{\substack{j \in A_{m'} \\ m' > m}} d_{ij}$$

is the inter-group average distance among pairs ij, and

$$D_I = \frac{2}{\sum_{m=1}^{p} \bar{\bar{A}}_m (\bar{\bar{A}}_m - 1)} \sum_{m=1}^{p} \sum_{\substack{i, j \in A_m \\ i<j}} d_{ij}$$

is the intra-group average distance among pairs ij.

This global function has been abandoned just because of the computational difficulties, resulting from its form, which hardly lends itself to algorithmic simplifications. Hence, Fortier and Solomon [6] have tried another approach, consisting in construction of the objective function which in a way utilises the assumpsion C. Thus

P. Partition P is better than partition P^* iff

$$C(P) > C(P^*)$$

where $C(P) = \sum_{\substack{i, j \in I \\ i<j}} g_{ij}$, and $g_{ij} = (d_{ij} - \varepsilon)\, \gamma_{ij}$, with $\gamma_{ij} = +1$ when i and j are in the same A_m and $\gamma_{ij} = -1$ otherwise.

It can easily be shown that the function C has the basic property that ensures existence of a non-trivial maximum over its maximal values for various p, i.e. $p^{\text{opt}} \neq 1$ and $p^{\text{opt}} \neq n$, where

$$p^{\text{opt}} = \{p_{\text{opt}} | C(p_{\text{opt}}) = \max_p (C^{\max}(p)) = \max_p (\max_{P^p} C(P^p))\}$$

Fortier and Solomon have not proposed any efficient algorithm, analysed a relatively small example ($n=19$) and have in fact restricted their considerations to the case $\varepsilon=0.5$. Further work reported in the paper did not regard elaboration of a more efficient algorithm for optimising the function proposed, but rather its application for a very specific purpose, related to factor analysis.

Thus, the present author proposes another form of the global objective function, first introduced in [13].

Q. Partition P is better than partition P^* iff

$$Q(P) > Q(P^*)$$

where

$$Q(P) = (1-\rho) \sum_{m=1}^{p} \sum_{\substack{i, j \in A_m \\ i<j}} w_{ij} + \rho \sum_{m=1}^{p} \sum_{\substack{i \in A_m \\ j \notin A_m}} d_{ij}^*$$

w_{ij} is a function $w(d_{ij})$, $R_+^1 \cup \{0\} \rightarrow R_+^1 \cup \{0\}$, such that

1° $d_{ij} < d_{ik} \Leftrightarrow w(d_{ij}) > w(d_{ik})$

2° $\bar{d}^* = \bar{w} = 1$, with bars denoting averages,

d_{ij}^*, i.e. elements of the matrix D^*, are obtained from the d_{ij} through

$$d_{ij}^* = \frac{d_{ij}}{\bar{d}_{ij}} = \frac{n(n-1)}{2} \frac{d_{ij}}{\sum\limits_{\substack{i,j \in I \\ i<j}} d_{ij}}$$

and $\rho \in [0, 1]$.

The intuitive sense of this function is obvious — it requires maximisation of intra-group "similarities" together with maximisation of inter-group "dissimilarities".

We can, for instance, set

$$w(d_{ij}) = \frac{d^{*\max} + d^{*\min} - d_{ij}^*}{d^{*\max} + d^{*\min} - 1}$$

where $d^{*\max}$ and $d^{*\min}$ denote maximum and minimum distances in D^*, respectively.

For the non-trivial case, i.e. when only $d^{\max} > \bar{d} > d^{\min}$, the objective function $P(Q)$ reaches maximum for some $p^{\text{opt}} \in (1, n)$, provided a certain simple condition on ρ holds. When $\rho = 1/2$ and $w(d_{ij})$ is defined as above, $Q(I) = Q(\{I\}) = \frac{1}{4} n(n-1)$.

6. The algorithm

First the triangular matrices D^* and W are formed according to the averaging formula given before. Working of the algorithm which optimises partitions P for subsequent values of ρ can be summarised as follows

```
1. $k=0$; $\rho^\circ=1$; $P^{\text{opt}}(1)=I$
2. $k=k+1$
3. $\rho^{kO} = \max\limits_{A_m, A_{m'} \in P^{\text{opt}}(\rho^{k-1})} \tilde{\rho}(A_m, A_{m'})$
4. if $\bar{L}^i(\rho^{k-1}, \rho^{kO}) = 2 \ \forall i$ then go to 6.
5. $\tilde{\rho}(A_m, A_{m'} | A_m, A_{m'} \in L^i(\rho^{k-1}, \rho^{kO})) = \tilde{\rho}(L^i(\rho^{k-1}, \rho^{kO})) \ \forall i$; go to 3
6. $\rho^k = \rho^{kO}$, $P^{\text{opt}}(\rho^k) = (P^{\text{opt}}(\rho^{k-1}) +$
       $- \{A_m \in L(\rho^{k-1}, \rho^{kO}) \wedge P^{\text{opt}}(\rho^{k-1})\}) \cup L(\rho^{k-1}, \rho^{kO})$
7. if $P^{\text{opt}}(\rho^k) \neq \{I\}$ then go to 2
8. end.
```

In the above summary presentation we have

$$\tilde{\rho}(A_m, A_{m'}) = \frac{W(A_m, A_{m'})}{W(A_m, A_{m'}) + D(A_m, A_{m'})}$$

with

$$W(A_m, A_{m'}) = \sum_{\substack{i \in A_m \\ j \in A_{m'}}} w_{ij} \text{ and } D(A_m, A_{m'}) = \sum_{\substack{i \in A_m \\ j \in A_{m'}}} d_{ij}$$

and

$$L(\rho, \tilde{\rho}) = \{A_m \in P^{\text{opt}}(\rho) | \exists A_{m'} \in P^{\text{opt}}(\rho) : A_m \tilde{\rho} A_{m'}\}$$

where $A_m \tilde{\rho} A_{m'} \Leftrightarrow \tilde{\rho}(A_m, A_{m'}) = \tilde{\rho}$, and

$$L(\rho, \tilde{\rho}) = L^1(\rho, \tilde{\rho}) \cup L^2(\rho, \tilde{\rho}) \cup L^3(\rho, \tilde{\rho}) \cup \ldots$$

$$L^i(\rho, \tilde{\rho}) \wedge L^j(\rho, \tilde{\rho}) = \emptyset, \; i \neq j$$

$$A_m, A_{m''} \in L^i(\rho, \tilde{\rho}) \Leftrightarrow \exists \{A_{m1}, A_{m2}, \ldots, A_{ml(m', m'')}\} \subset$$
$$\subset L^i(\rho, \tilde{\rho}) \cup \{\emptyset\} : A_{m'} \tilde{\rho} A_{m1} \wedge A_{ml(m', m'')} \tilde{\rho} A_{m''}$$

$$\exists A^{m'} \in L^i(\rho, \tilde{\rho}), \; A^{m''} \in L^j(\rho, \tilde{\rho}), \; i \neq j, \; A^{m'} \tilde{\rho} A^{m''}$$

and therefore

$$\tilde{\rho}(L^i(\rho, \tilde{\rho})) = \frac{\sum\limits_{\substack{A_m, A_{m'} \in L^i(\rho, \tilde{\rho}) \\ m < m'}} W(A_m, A_{m'})}{\sum\limits_{\substack{A_m, A_{m'} \in L^i(\rho, \tilde{\rho}) \\ m < m'}} W(A_m, A_{m'}) + \sum\limits_{\substack{A_m, A_{m'} \in L^i(\rho, \tilde{\rho}) \\ m < m'}} D(A_m, A_{m'})}$$

Thus, the algorithm displayed shows certain apparent similarity with the one of Slater in that it analyses patterns of groupings for various "levels of perception" or "linkage" (ε in case of Slater's, ρ in the present method's case). There is, however, an essential difference between the two approaches since values of ε refer to values of direct local inter-element distances d_{ij} or linkages w_{ij} while consecutive values of ρ are determined on the basis of $\tilde{\rho}$, which take into account the overall situation in $Q^{\text{opt}}(P^{\text{opt}})$ over a segment in ρ.

7. Conclusions

The method presented in this paper allows finding of partitions of a set of elements into mutually disjoint subsets through explicit optimisation of a global objective function, so that not only the composition of subsets, but also their number is being optimised.

This global objective function displays intuitively acceptable local properties as well (see Appendix 2). To comment upon the possibility of finding the optimal number of clusters, a sentence from [5], p. 30 can be cited: "This monograph will not be concerned with the very difficult problem of determining the number of clusters".

Furthermore, the algorithm resulting thereof is relatively simple and its complexity is comparable to the one of methods basing upon iterative merging of the two closest neighbours.

Rerefernces

[1] BIELECKA K., PAPRZYCKI M., PIASECKI Z. Proposal of New Taxonomic Methods for Agricultural Typology. *Geographia Polonica*, 40. Warszawa 1979.

[2] BIELECKA K., SZCZOTKA F. A. Study on Evaluation of Applicability of Quantitative Methods in Agricultural Typology (in Polish). *Biuletyn Informacyjny* z. 23, Instytut Geografii i Przestrzennego Zagospodarowania PAN, Warszawa 1972.

[3] BYFUGLIEN J., NORDGARD A. Region-Building — A Comparison of Methods. *Norsk Geografisk Tijdsskrift* **27**, 2(1973).

[4] DIDAY E., SIMON J. C. Clustering Analysis, in: Digital Pattern Recognition. Ed. K.S. Fu. Berlin, Springer Verlag 1976.

[5] DURAN B. S., ODELL P. L. Cluster Analysis, A Survey, Berlin, Springer Verlag 1974.

[6] FORTIER, J. J., SOLOMON H. Clustering Procedures, in: Multivariate Analysis I. Ed. P.R. Krishnaiah. New York, Academic Press 1966.

[7] HARTIGAN J. A. Clustering Algorithms. New York, J. Wiley 1978.

[8] JENSEN R. E. A Dynamic Programming Algorithm for Cluster Analysis. *Operational Research* **12** (1969), 1034–1057.

[9] KACPRZYK J., STAŃCZAK W. Application of the Method of Minimally Interconnected Subnetworks to Division of a Group of Enterprises into Subgroups (in Polish). *Archiwum Automatyki i Telemechaniki* **20**, 4 (1975).

[10] KACPRZYK J., STAŃCZAK W. On a Further Extension of the Method of Minimally Interconnected Subnetworks. *Control a. Cybernetics* **7**, 2 (1978).

[11] LANCE G. N., WILLAMS W. T. A General Theory of Classificatory Sorting Strategies. *Computer Journal* **10**, 3, 4 (1967).

[12] MULVEY J. M., CROWDER H. P. Cluster Analysis: An Application of Lagrangian Relaxation. *Management Science* **25**, 4 (1979), 329–340.

[13] OWSIŃSKI J. W. The Problem of Optimal structure Partitioning with Allocation of Resources, in: Design Principles for the Interactive Planning Systems and Some Models of Decision Systems (in Polish). Systems Research Institute, Internal Report ZPZC 34–41/79, Warszawa 1979.

[14] SLATER P. B. A Hierarchical Regionalization of Japanese Prefectures Using 1972 Interprefectural Migration Flows. *Regional Studies* **10** (1976), 123–132.

[15] SLATER P. B. A Multiterminal Network-Flow Analysis of an Unadjusted Spanish Interprovincial Migration Table. *Environment a. Planning A.* **8** (1976), 875–878.

[16] TREMOLIÈRES R. The Percolation Method for an Efficient Grouping of Data, *Pattern Recognition* **11** (1979).

[17] WISHART D. Some Problems in the Theory and Application of the Methods of Numerical Taxonomy. Ph. D. Thesis. University of St. Andrew's, Scotland 1971.

[18] YOUNG T. V., CALVERT T. W. Classification, Estimation and Pattern Recognition. New York, American Elsevier 1974.

Received, October 1980.

APPENDIX 1. BIASSES OF A METHOD

Consider minimally interconnected subnetworks technique, working according to conditions F and F′, and described in [9, 10].

A1.1. Figure A1.1 presents an example of a graph $\{I, D\}$ which is indivisible in terms of F and F′. Appropriate (partial) conditions of indivisibility for this technique are given in [13]. Obviously, the result illustrated in Fig. A1 is highly counterintuitive.

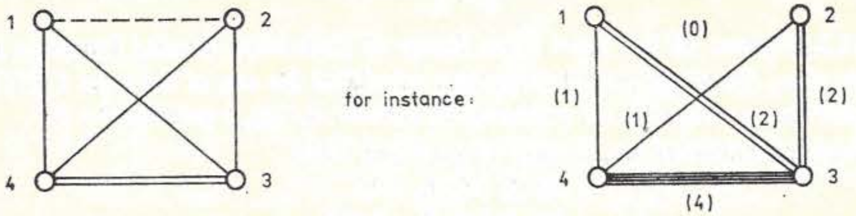

Fig. A1.1. A case of structure indivisible with the minimally interconnected subnetworks techniques. Numbers in brackets denote w_{ij}

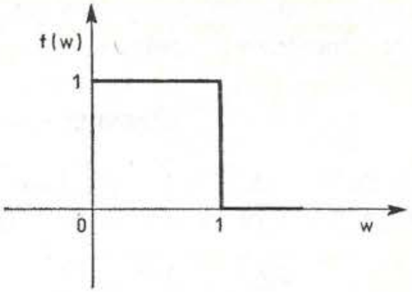

Fig. A1.2. Example of probability density function for values of w_{ij}

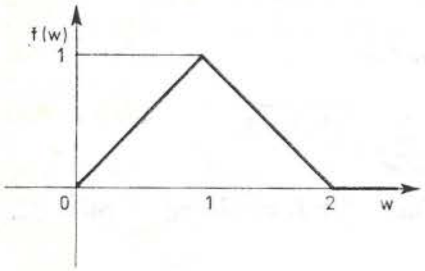

Fig. A1.3. Example of probability density function for values of w_{ij}

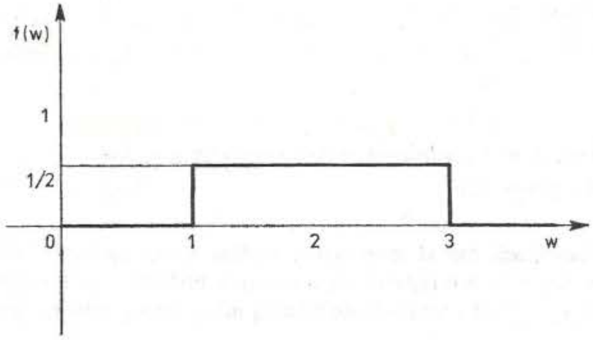

Fig. A1.4. Example of probability density function for values of w_{ij}

A1.2. Figures A1.2–4 show three forms of probability density functions over values of w_{ij} for which probabilities were calculated for any of the pairs i, j, $i, j \in I$ to fulfill the condition F′, i.e. to constitute subsets $A_m=\{i, j\}$ entering $P \in H^w$.

These probabilities are, in case of $n=4$ (see [13]):

for $f_1(w): P_1=0.3$,
for $f_2(w): P_2=0.095$,
for $f_3(w): P_3=0.0094$.

Since for $n=3$ the probability equals 1 for any form of the density function, the above values indicate clearly the bias of the technique towards "greater" sets to enter P.

The above comments do not apply solely to the technique considered. They could be extended to any of the "local" methods, having implicit biasses built into the appropriate procedures.

APPENDIX 2. SOME PROPERTIES OF $Q(P)$

Suppose there is $P^{\text{opt}}(\rho^*)$ known for given D, i.e. also $Q^{\text{opt}}(\rho^*)$. For each set $L \subset P^{\text{opt}}(\rho^*)$ of subsets $A_m \in P^{\text{opt}}(\rho^*)$ a function can be defined

$$\Delta Q_L(\rho)=(1-\rho)\sum_{\substack{A_m, A_{m'} \in L\\ m<m'}} W(A_m, A_{m'})-\rho \sum_{\substack{A_m, A_{m'} \in L\\ m<m'}} D(A_m, A_{m'})$$

linear in ρ. Assuming that ρ is being decreased from ρ^* we are interested in condition $\Delta Q_L(\rho) \geqslant 0$. It is equivalent to

$$\rho \leqslant \frac{\sum\limits_{\substack{A_m, A_{m'} \in L\\ m<m'}} W(A_m, A_{m'})}{\sum\limits_{\substack{A_m, A_{m'} \in L\\ m<m'}} W(A_m, A_{m'})+\sum\limits_{\substack{A_m, A_{m'} \in L\\ m<m'}} D(A_m, A_{m'})}$$

Obviously, the value of $\tilde{\rho}$ reaches its maximum for L consisting of pairs A_m, $A_{m'}$, for which

$$\tilde{\rho}(\{A_m, A_{m'}\}, \rho^*)=\max_{L \subset P^{\text{opt}}(\rho^*)} \tilde{\rho}(L, \rho^*)$$

For maximal value of $\tilde{\rho}$ there occurs a merger of $A_m, A_{m'}$ so that $P^{\text{opt}}(\tilde{\rho}_{\max})$ differs from $P^{\text{opt}}(\rho^*)$ by this merger. The case of multiple pairs corresponding to $\tilde{\rho}_{\max}$ is dealt with by introduction of the set $L(\rho^*, \tilde{\rho}_{\max})$, as defined in the text.

Intuicja i formalizacja w zagadnieniach grupowania: lokalne i globalne kryteria grupowania

Dokonuje się klasyfikacji metod grupowania według tego, czy kryterium grupowania jest jawne, czy niejawne. Kryteria te otrzymuje się z pewnych podstawowych, oczywistych intuicyjnie założeń. Przy klasyfikacji metod rozpatruje się lokalną lub globalną naturę kryteriów i możliwość algorytmizacji.

Zaproponowano pewną nową postać kryterium globalnego umożliwiającego jawne rozwiązanie zagadnienia grupowania zarówno dla optymalnej liczby grup, jak dla ich składu. Kryterium charakteryzuje się także istotnymi właściwościami lokalnymi. Podano prosty przykład.

Интуиция и формализация в вопросах группирования: локальные и глобальные критерии группирования

Проводится классификация методов группирования согласно тому является ли критерий группирования явным или же неявным. Эти критерии получаются из некоторых основных, интуитивно очевидных предпосылок. При классификации методов рассматривается локальная или глобальная природа критериев и возможности алгоритмизации.

Предлагается некоторый новый вид глобального критерия, позволяющего явно решать задачи группирования как для оптимального числа групп, так и для их состава. Критерий характеризуется также существенными локальными свойствами. Приведен простой пример.