A “Natural” Agglomérative Clustering Method for Biology

L. Dragomirescu and **T. Postelnicu**

Centre of Anthropological Research, Bucharest, Romania
Centre of Mathematical Statistics, Bucharest, Romania

Summary

The authors consider that cluster analysis does not objectivize but represents the biologist’s subjectivity as to (1) characters considered to be significant and to (2) the way of classification. The latter, however, authors’ opinion, must be specific to the field of application. To this effect some methods are suggested for biology. The methods originate in improvements or transformation of Buser and Baroni-Urbani’s method, as well as Watanabe’s method, and have the property of processing overall information with no loss or distortion. An agglomerative method which yields a necessarily unique result is suggested, being considered by the authors as a homologue of Watanabe’s divisive method. The methods proposed are studied using examples logically constructed. These examples can provide from biology, especially from ecology.

Key words: Cluster analysis; Statistics of binary data; Subadditive measure; Ecology.

1. Introduction

This paper is intended to present some hierarchical clustering methods which hold, in our opinion, a privileged position at least with respect to applications in biology. The superiority of these methods arises from their quality of being *natural*. We understand by naturalness here either (1) *the necessary uniqueness* of the result obtained, or (2) the use of *overall information* and, particularly, the nondistortion of the latter, as well as (3) the strong adequacy to biological thinking. The suggested methods result from transformations of the (a) divisive method, yielding a necessarily unique result from Watanabe, 1969 (the method hereafter referred to as *the W method*) or of (b) the agglomerative method, nondistorting the overall information, from Buser and Baroni-Urbani, 1982 (hereafter referred to as *the B method*). To allow for easy understanding of the suggested methods the W method and the B method are briefly presented in paragraph 2.

Procesat de Alexandru Florentin în data de 8 decembrie 2002

In section 3 some extentions and improvements of the B method are given, and in section 4 is presented an agglomerative method yielding a necessarily unique result which was obtained by formulating some *principles of a natural agglomerative classification.*

2. The *W* and *B* Methods

Suppose we have to classify an I set of OTU ((Operational Taxonomic Unity)s, cf. SNEATH and SOKAL, 1973) described by the rows of an OTU-character table, made up of n columns, i.e. n descriptive characters.

a) *The W method* yields a necessarily unique result on the basis of a cohesion value defined for any subset of the I set of OTUs, and of a procedure, made up of three so-called "strategies of a natural classification" (WATANABE, 1969). *Cohesion* is a supra-additive set function, denoted by c i.e.:

(1.1) $c(\bigcup_{k=1}^{r} A_k) \geq \sum_{k=1}^{r} c(A_k)\ (\forall)\{A_k\}_{k=1,2,\ldots,r}$ - family of mutually disjoint sub-sets,

since by "splitting" of a whole $\left(\bigcup_{k=1}^{r} A_k\right)$ into several parts ($\{A_k\}_{k=1,2,\ldots,r}$), possibly, cohesion may be lost. The lost cohesion:

(1.2) $$a\left(\bigcup_{k=1}^{r} A_k \mid A_1, A_2, \ldots, A_r\right) = c\left(\bigcup_{k=1}^{r} A_k\right) - \sum_{k=1}^{r} c(A_k) \geq 0$$

is called "*branching cost of the polychotomy*" $\left(\bigcup_{k=1}^{r} A_k \mid A_1, A_2, \ldots, A_r\right)$. WATANABE (1969) insists on the unicity of a classifications result, viewed as subjacent to the natural situation and proposes us *unique* branching cost the "*interdependence*" defined for any subset $\left(\bigcup_{k=1}^{r} A_k\right)$ of OTUs which is divided into r parts (the OTUs are described in a binary table or a table of several distinct values) as follows:

(1.3) $$W\left(\bigcup_{k=1}^{r} A_k \mid A_1, A_2, \ldots, A_r\right) = \sum_{k=1}^{r} H(A_k) - H(\bigcup_{k=1}^{r} A_k)$$

$(\forall)\ \{A_k\}_{k=1, 2,\ldots, r}$ family of mutually disjoint parts, $A_k \subseteq I$, where

(1.4) $$H(A_k) = -\sum_m p(v_m) \cdot log\ p(v_m)$$

is the calculated entropy for the field of probability formed by the distinct column-vectors v_m from the OTU-character subtable which is made up only by the rows from A_k, where $p(v_m)$ are their relative frequencies. Then *natural classification strategies (divisive strategies)* are applied:

Ist Strategy: Strategies II and II are applied to the whole I set of OTUs. Then, they are applied to each A cluster thus obtained. This process is iterated until only one OTU-cluster is obtained.

IInd Strategy: Being given an A cluster, all the possible dichotomies (A | A_1, A_2)(where $A=A_1 \cup A_2$; A_1, $A_2 \neq \emptyset$, $A_1 \cap A_2 = \emptyset$) are considered. If there is only one minimum branching cost dichotomy, it will be the prescribed "splitting". Other-wise, IIIrd strategy is applied.

IIIrd Strategy: The system of all minimum branching cost dichotomies is considered. The prescribed "splitting" will be the last sharp polychotomy, sharper than every dichotomy of the system. Intuitively, the clusters will stand for the "pieces" obtained by cutting the A subset according to all the minimum cost dichotomies.

We note that the *W* method works up overall information (not only the mutual one, as common methods do), *interclusters* information.

b) *The B Method* makes use of overall information as the *W* method does, but uses intraclusters information. Moreover, the method does not distort this information in any way as opposed to the common agglomerative methods described by SOKAL and SNEATH (1973) which proces mutual information, that is additionally distorted. Instead of similarity coefficients, defined in the common methods only for pairs of OTUs, the authors propose (i) two so called "homogeneities" (noted h_I and h_{II}) defined for sets of OTUs described by binary OTU-character tables and (ii) a unique clustering procedure, founded on one of the two homogeneities.

Further on, we shall deal only with *homogencity* h_I, defined in the quoted work, as follows:

(1.5) $$h_I(A) = \left(\sum_{j=1}^{n} s_j \right) / (n \cdot q) \qquad 0 \leq h_I \leq 1$$

where $A \subseteq I$ and the latter is the OTUs set, q is the number of OTUs in A, (n is the number of description binary characters) and s_j is the number of 1-s of the j character, for the OTUs in A.

In DRAGOMIRESCU (1987) is described the clustering procedure as follows: "Consider a table which has L rows and n columns, representing a set of L OTUs, described in n characters. (1) A LIST of all the subsets of OTUs is made up, calculating a homogeneity for each subset. (2) A subset of the LIST which has the maximum homogeneity from among the other subsets of the LIST is considered a cluster. (3) If the cluster formed is the whole set then the clustering is over, otherwise the subsets which contain only strict parts of the already formed cluster(s) are eliminated from the LIST and the item (2) is applied to the new LIST"

3. *B* Method Extension and Improvement

3.1. *The h^* Homogeneity for Binary Tables*

Instead of the h_I homogeneity in (1.5) formula is suggested (in DRAGOMIRESCU, CONSTANTINESCU and BĂNĂRESCU, 1985) the h^* *homogeneity*:

(2.1) $$h^*(A) = \left(\sum_{j=1}^{n} s_j \right) / (n^* \cdot q) \quad 0 \leq h^* \leq 1$$

where n^* is the number of non-identically vanishing columns for the lines, corresponding to the OTUs in A set, while the rest of the notation is the same as that of formula (1.5). In DRAGOMIRESCU (1987) is showed that, in a certain way, h^*, (1) generalizes the famous Jaccard similarity coefficient, (2) originates in the particularization to binary tables of a homogeneity noted by H^* and which is defined in the same work, for tables of any number of values, where H^*, (3) models the terms formulated by BECKNER (1959) on a "polythetic group" or "natural" taxa (cf. SNEATH and SOKAL, 1973).

3.2. *The Improved B Method*

In both works, quoted in § 3.1, an improvement of the B Method is described: it is obtained by adding to the (2) condition in § 2 the (2') condition "and it has a maximum number of OTUs". In DRAGOMIRESCU (1987) is shown that the B method thus improved may correctly classify the

Procesat de Alexandru Florentin în data de 8 decembrie 2002

wellknown Watanabe's example (WATANABE, 1969), provided that the h^* homogeneity is used; such a performance had been reached before then only by the *W* method.

Example 1 (WATANABE, 1969)

"Suppose that four girl students live in a dormitory. Three of them are bound by a peculiar mixture of friendship and jealousy, so that none of the group wants to sit alone in the lounge without another member, yet none wants to sit there with both of the remaining two because she cannot stand seeing the evidence of friendship between these latter two. The fourth girl is entaitly neutral to these three and sits in the lounge no matter who else may or may not be sitting there; reciprocally these three pay no attention to the fourth girl. Suppose that x_1, x_2, x_3, and x_4 represent these four girls, and y_j stands for the predicate "is sitting in the lounge at the *j*-th observation".

Table 1a
OTU-Character Table, corresponding to Example 1 (Watanabe)

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
x_1	1	1	0	0	1	1	0	0
x_2	1	1	1	1	0	0	0	0
x_3	0	0	1	1	1	1	0	0
x_4	1	0	1	0	1	0	1	0

The example 1 is "non-trite" in the sense that it cannot be analiyzed by any common agglomerative method as each pair of OTUs has the same number of pairs 1 – 1, 1 – 0, 0 – 1 and 0 – 0.

Obviously, the correct result will be the hierarchical clustering $((x_1, x_2, x_3)\ x_4)$, the fourth girl student being independent while the first three are inseparable. Calculating the homogeneities h_I and h^* the results given in table 1b are obtained.

Table 1b
The values of homogeneities h_I and h^* for example 1 (Watanabe)

Submulţimi UTO	h_I	h^*	Submulţimi UTO	h_I	h^*
$\{x_1,x_2\}$	8/16=1/2	4/6	$\{x_1,x_2,x_3\}$	12/24=1/2	4/6
$\{x_1,x_3\}$	8/16=1/2	4/6	$\{x_1,x_2,x_4\}$	12/24=1/2	4/7

$\{x_1,x_4\}$	8/16=1/2	4/6	$\{x_1,x_3,x_4\}$	12/24=1/2	4/7
$\{x_2,x_3\}$	8/16=1/2	4/6	$\{x_2,x_3,x_4\}$	12/24=1/2	4/7
$\{x_2,x_4\}$	8/16=1/2	4/6			
$\{x_3,x_4\}$	8/16=1/2	4/6	$\{x_1,x_2,x_3,x_4\}$	12/24=1/2	4/7

It may be observed from table 1b that the h_I value (namely 1/2) is the same for any OTUs subset. On the other hand, the original *B* method when applied to h^* may yield several results, by grouping at the first step any of two OTUs, or the first three ones. Only *the improved B method* gives the needed hierarchical clustering, namely (x_1, x_2, x_3) x_4). The result is unique, but only conjuncturally (i.e. as regards table 1a). Indeed, here we give the following example:

Example 2

Let us consider two clusters of girl students $\{x_1, x_2, x_3\}$ and $\{x_3, x_4, x_5\}$, every group behaving like the first 3 girl students from example 1. This behaviour is described in table 2. A similar example can be provided from ecology.

Table 2
OTU-character table as per example 2

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
x_1	1	1	0	0	1	1	1	0	1	0	1	0
x_2	1	1	1	1	0	0	1	0	1	0	1	0
x_3	0	0	1	1	1	1	1	1	0	0	1	1
x_4	1	0	1	0	1	0	1	1	1	1	0	0
x_5	1	0	1	0	1	0	0	0	1	1	1	1

Calculating the h^* homogeneity the maximum (0.6) value is obtained for the $\{x_1, x_2\}$ and $\{x_4, x_5\}$ respectively disjunct clusters, then the following next (0.(4)) maximum is reached by the non-disjunct clusters: $\{x_1, x_2, x_3\}$ and $\{x_3, x_4, x_5\}$. Therefore, in this case, even the improved *B* method will not yield a unique solution.

4. Agglomerative Method With a Necessarily Unique Result

Similar to WATANABE (1969) we formulate *principles of a Natural Agglomerative Classification*: (A) By clustering, *homogeneity* (h) may be lost. (Meaning that the homogeneity must be a subadditive set function:

$$h\left(\bigcup_{k=1}^{r} A_k\right) \leq \sum_{k=1}^{r} h(A_k)$$

$(\forall)$ $\{A_k\}_{k=1,2,\ldots,r}$ family of mutually disjoint sub-sets). (B) We shall prefer maximum homogeneity clusters. (C) In order that a classification should be natural, it must be unique (WATANABE, 1969). (D) In consequence of the *C* principle the necessity of a unique decision at every stage of the method arises.

The *B* principle will grant priority to maximum homogeneity clusters while the *C* principle calls for forming a unique cluster (at the current) even though there are several maximum homogeneity clusters. The solution will be supported within the method by the following natural, in our opinion, decision:

Let A, B te two sets of OTUs which have the maximum homogeneity at the respective stage (h_{max}): (i) if $A \cap B = \emptyset$, then both A and B subsets are considered to be clusters at the h_{max} level (within the same stage); (ii) if $A \cap B \neq \emptyset$, then the $A \cup B$ subset is considered to be a cluster at the $h\ (A \cup B)(\leq h_{max})$ level. These principles bring forth the following agglomerative method, which yields a necessarily unique result.

Method Description

A homogeneity is selected (h_I – (1.5) formula, h^*– (2.1) formula) or other homogeneities which are adequate to the problem and to the OTU-character table. Then there are applied what we might call paraphrasing WATANABE (1969) the strategies of natural *agglomerative* classification:

I^st^ Strategy: Strategies II and II are applied to the LIST of all OTUs subsets. The LIST is reduced by eliminating all the subsets which include only strict parts of clusters already formed by applying these strategies. The same process is undergone by *the new* LIST and so on, until LIST= Ø.

II^nd^ Strategy: Being given a LIST of OTUs subsets, the $\mathcal{M}$ family of subsets is considered, ($\mathcal{M} \subseteq$ LIST), with the h_{max} maximum homogeneity within the LIST. If the subsets in $\mathcal{M}$ are mutually disjoint, these subsets are considered to be clusters at the respective level (h_{max} , within the current stage). Otherwise, III^rd^ strategy is applied.

III[rd] Strategy: All the *complete families* (*in* $\mathcal{M}$) are considered (see Appendix for theoretical details or, further on, the commentary on fig. 1 for an intuitive understanding): $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s$ and their unions U_1, $U_2, \ldots, U_s$ which are mutually disjoint (as Corollary 1.2 in Appendix) and with homogeneities less than or equal to the maximum homogeneity ($h_{\max}$) at the current stage. These unions are considered to be clusters in descending order of the homogeneity levels, substages within the current stage.

Remark: Although this method is brought about by the transformation of the *B* method, we formulated it into analogy to the *W* method, since we consider it to be the agglomerative method corresponding to the *W* divisive method.

In figure 1 the $\mathcal{M}$ family is given, as it is made up of 8 subsets of OTUs noted by M_1, M_2, …, M_8 of $h_{\max}$ maximum homogeneity (from a fictitious example; for analytical details see Appendix). The complete families (in $\mathcal{M}$) are $\mathcal{M}_1 = \{M_0\}$, $\mathcal{M}_2 = \{M_2, M_3\}$ and $\mathcal{M}_3 = \{M_4, M_5, M_6, M_7, M_8\}$ and their unions marked by hatchings of different inclination are: M_1, M_3 and $\bigcup_{i=4}^{8} M_i$ accordingly.

Intuitively, the complete families unions (in $\mathcal{M}$) are the “complete islands”, made up of “(land) territories” (i.e. subsets in $\mathcal{M}$) “neighbouring by land” (i.e. non-disjoint).

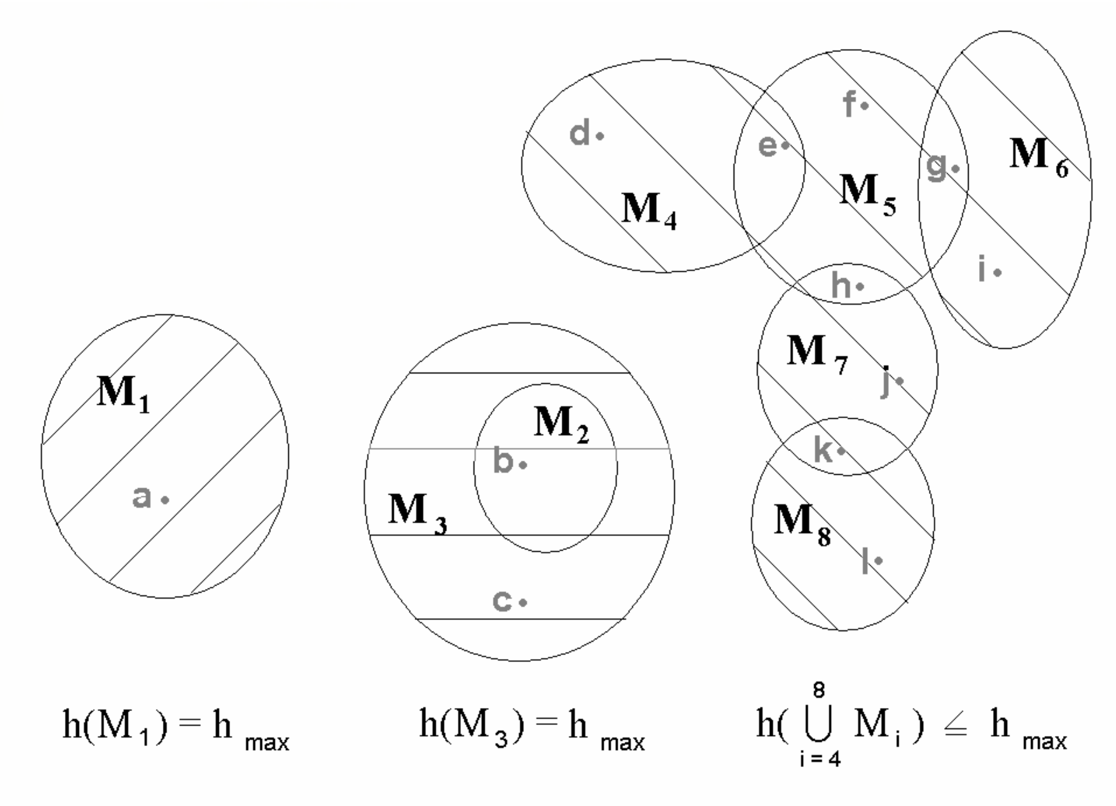

Fig. 1. Three complete families (in M)

Fig. 2. Dendrogram yielded by the § 2.2. method on Example 2

Coming back to § 3.2. example 2, we apply this method (to the h^* homogeneity) and we find out that at the second stage the family of M maximum homogeneity subsets includes only one complete family, having for a union the whole set to be classified as a union. Therefore, the result is unique and, namely, the one in figure 2.

References

BUSER, M. and BARONI-URBANI, C., 1982: A direct nondimensional clustering method for binary data. Biometrics **38**, 351-360.
DRAGOMIRESCU, L., 1987: Some extension of Buser and Baroni-Urbani's clustering method. Biom. J. **29**, 1-9.
DRAGOMIRESCU, L., CONSTANTINESCU, V. and BĂNĂRESCU, P., 1985: A numerical taxonomy method adequate to the biological thinking. Applications Watanabe's example and the Achonthobrama Genus (*Pisces*, *Cyprinidae*). Traveaux du Muséum d'Histoire naturelle Grigore Antipa, **27**, 243-265.
SNEATH, A. and SOKAL, R., 1973: *Numerical Taxonomy*. San Francisco: Freeman.
WATANABE, S., 1969: *Knowing and Guessing*. New York, John Wiley.

Appendix

For explaining the complete family notion (in $\mathcal{M}$) we give first:

Definition 1.1. *A conex family* (in $\mathcal{M}$) is called a sub-family $\mathcal{C} \subseteq \mathcal{M}$ with the characteristic that:

(A 1.1.) $(\forall)\, A, B \in \mathcal{C}\ (\exists)\ p \in \mathbf{N}$ and $\{C_0, C_1, \ldots, C_p\} \subseteq \mathcal{C}$ so as $C_{i-1} \cap C_i \neq \emptyset$ $(\forall)\ i = 1, 2, \ldots, p$, where $A = C_0$ and $B = C_p$

In figure 1 (in the text) we have:

$I = \{a, b, c, d, e, f, g, h, i, j, k, l\}$; $\mathcal{M} = \{M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8\}$

where:

$M_1 = \{a\}$, $M_2 = \{b\}$, $M_3 = \{b, c\}$, $M_4 = \{d, e\}$, $M_5 = \{e, f, g, h\}$, $M_6 = \{g, i\}$, $M_7 = \{h, j, k\}$, $M_8 = \{k, l\}$

It is seen that families:

$\mathcal{C}_1 = \{M_2\}$, $\mathcal{C}_2 = \{M_3\}$, $\mathcal{C}_3 = \{M_4, M_5\}$, $\mathcal{C}_4 = \{M_4, M_5, M_6\}$,
$\mathcal{M}_1 = \{M_2\}$, $\mathcal{M}_2 = \{M_2, M_3\}$, $\mathcal{M}_3 = \{M_4, M_5, M_6, M_7, M_8\}$ are conex families while the families: $\mathcal{N}_1 = \{M_1, M_2\}$, $\mathcal{N}_2 = \{M_2, M_5, M_6\}$ are not conex.

It is immediately proved that:

Proposition 1.1. If $\mathcal{C}_1$ and $\mathcal{C}_2$ are conex families (in $\mathcal{M}$) and $(\exists)\ C^1 \in \mathcal{C}_1$ and $C^2 \in \mathcal{C}_2$, so as $C^1 \cap C^2 \neq \emptyset$ then $\mathcal{C}_1 \cup \mathcal{C}_2$ is a conex family (in $\mathcal{M}$).

Corollary 1.1. If conex families (in $\mathcal{M}$) $\mathcal{C}_1$ and $\mathcal{C}_2$ are not disjoint (i.e. $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$), then $\mathcal{C}_1 \cup \mathcal{C}_2$ is a conex family (in $\mathcal{M}$).

Among the conex families (in $\mathcal{M}$) we shall be interested in the maximum ones relative to inclusion, i.e.:

Definition 1.2. *A complete (whole) family* (*in* $\mathcal{M}$) is called a conex family (in $\mathcal{M}$) $\mathcal{C}$, with the characteristic that.

(A 1.2.) $\quad (\forall)\ A \in \mathcal{M}$ so as $(\exists) B \in \mathcal{C}$ with $A \cap B \neq \emptyset \Rightarrow A \in \mathcal{C}$.

In the example in fig.1, families $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are complete (in $\mathcal{M}$), while families $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ and $\mathcal{C}_4$ are not complete (in $\mathcal{M}$).

With a view to foundation of the classification method in § 4. The following is important.

Proposition 1.2. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are complete families (in $\mathcal{M}$) and they are distinct (i.e. $\mathcal{M}_1 \neq \mathcal{M}_2$), then $U_1 \cap U_2 = \emptyset$, where

$$U_1 = \bigcup_{M \in \mathcal{M}_1} M \qquad \text{and} \qquad U_2 = \bigcup_{M \in \mathcal{M}_2} M$$

(Unions of two distinct complete families are disjoint).

Proof:

Suppose, per absurdum, that $U_1 \cap U_2 \neq \emptyset$. So $(\exists)\ x \in U_1 \cap U_2$. It means that: $(\exists)\ M_1 \in \mathcal{M}_1$, so that $x \in M_1$ and $(\exists)\ M_2 \in \mathcal{M}_2$, so that $x \in M_2$. Therefore $M_1 \cap M_2 \neq \emptyset$. But $\mathcal{M}_1$ is a complete family and $M_1 \in \mathcal{M}_1$, therefore $M_2 \in \mathcal{M}_1$. It follows that $M_2 \in \mathcal{M}_1 \cap \mathcal{M}_2$ and, taking into account that $\mathcal{M}_1$ and $\mathcal{M}_2$ are conex families, it follows, as per corollary 1.1., that $\mathcal{M}_1 \cup \mathcal{M}_2$ is a conex family. Thus $(\forall)\ A \in \mathcal{M}_1$ and $B \in \mathcal{M}_2$, $\mathcal{M}_1 \cup \mathcal{M}_2$ being conex family, it follows that $(\exists)\ p \in N$ and $\{C_0, C_1, \ldots, C_p\} \subseteq \mathcal{M}_1 \cup \mathcal{M}_2$, so that:

(A 1.3.) $\quad C_{i-1} \cap C_i \neq \emptyset \quad (\forall)\ i = 1, 2, \ldots, p$ where $A = C_0$ and $B = C_p$

Writing the (A 1.3.) conditions in detail and in inverse order and taking into account that $B \in \mathcal{M}_2$, $\mathcal{M}_2$ being a complete family it follows that:

$\emptyset \neq C_{p-1} \cap C_p = C_{p-1} \cap B \Rightarrow C_{p-1} \in \mathcal{M}_2$

$\underline{\emptyset \neq C_{p-2} \cap C_{p-1} \qquad\qquad \Rightarrow C_{p-2} \in \mathcal{M}_2}$

$\emptyset \neq C_0 \cap C_1 = A \cap C_1 \qquad \Rightarrow A \in \mathcal{M}_2$

Therefore we took $(\forall)\ A \in \mathcal{M}_1$ and we obtained that $A \in \mathcal{M}_2$, i.e. $\mathcal{M}_1 \subseteq \mathcal{M}_2$.

In a completely analogous, but using the fact that $\mathcal{M}_1$ is a complete family, it follows that $\mathcal{M}_2 \subseteq \mathcal{M}_1$. Therefore $\mathcal{M}_1 = \mathcal{M}_2$ which is at variance with the hypothesis ($\mathcal{M}_1 \neq \mathcal{M}_2$). It follows that $U_1 \cap U_2 = \emptyset$.

Corollary 1.2. Let $\{\mathcal{M}_k\}_{k=1, 2,\dots, s}$ be a system of s distinct complete families (in $\mathcal{M}$). If we note,

$$U_k = \bigcup_{M \in \mathcal{M}_k} M$$

Then the family $\{U_k\}_{k=1, 2,\dots, s}$ is mutually disjunct. (The unions of the distinct complete families (in $\mathcal{M}$) are mutually distinct).

Received July 1990

Dr. LIVIU DRAGOMIRESCU
Centre of Anthropological Research
Romanian Academy
Bd. Eroilor Sanitari nr. 8
72421 Bucharest, Romania

Prof. Dr. TIBERIU POSTELNICU
Centre of Mathematical Statistics
Romanian Academy
Bd. Magheru nr.22
70158 Bucharest, Romania

A “Natural” Agglomérative Clustering Method for Biology
L. Dragomirescu and T. Postelnicu

Procesat de Alexandru Florentin în data de 8 decembrie 2002