PYRAMIDAL CLUSTERING WITH MISSING VALUES

M. Schader[1] and W. Gaul[2]

1 **Institut für Informatik, Universität der Bundeswehr Hamburg, Germany**

2 **Institut für Entscheidungstheorie und Unternehmensforschung, Universität Karlsruhe, Germany**

ABSTRACT

Pyramidal generalizations of hierarchical clustering can be computed by the PAC (Pyramidal Ascending Classification) algorithm which—in its original form—needs complete dissimilarity data for its application. We discuss how the algorithm can be adapted to what we have called the PACII (Pyramidal Ascending Classification with Incomplete Information) modification and present results for special patterns of missing data. Distances between 15 selected French cities are taken to demonstrate recovery properties when certain distances are missing. The special shape of missing values patterns helps to draw conclusions for the pyramidal cluster analysis of two-mode data, a type of data which is frequently used in applications.

KEY WORDS: Cluster Analysis, Hierarchies, Missing Values, Pyramids, Two-Mode Data.

INTRODUCTION

There are several reasons for tackling the "pyramidal clustering with missing values" problem.

When using hierarchical clustering it is quite obvious to ask for generalizations of hierarchical structures. (See, e.g., Opitz (Hrsg.)(1978), p. 81, and Opitz (1980), p. 68, for what has been called "Quasi-Hierarchie" in this context.)

Meanwhile, the pyramidal generalization of hierarchical clustering, developed by Diday and coworkers, is well-known. (See, e.g., Diday, Bertrand (1986), Bertrand (1986), Diday (1987) for an introduction and Brito, Diday (1990) for a recent application of pyramidal clustering to the representation of symbolic objects.)

Of course, it is of interest to have a lot of examples from different areas to be able to demonstrate the advantages which pyramidal clustering solutions may depict in comparison to hierarchical outputs. (See, e.g., Gaul, Schader, Both (1990) for an interpretation of two-mode data from a marketing application by pyramidal and hierarchical clustering.)

However, it may happen that some of the underlying data sets have missing values or are of such a shape that special patterns of missing values occur in the dissimilarity data to which the original information is transformed. In such cases algorithms are of importance which are able to handle different kinds of incomplete data. (See Schader, Gaul (1991) for a hierarchical clustering approach which can tackle the problem when data are incomplete and references to other approaches known from the literature.)

In the following section, we motivate why the analysis of so-called two-mode data which—for the evaluation by clustering techniques—show special patterns of missing values is of interest for applications.

Next, a short description of the PACII (Pyramidal Ascending Classification with Incomplete Information) algorithm is given which is a modification of the PAC (Pyramidal Ascending Classification) technique. (See, e.g., Diday, Bertrand (1986), Bertrand (1986), Diday (1987).)

Finally, distances of 15 selected French cities are used to demonstrate recovery properties of PACII when certain distances are missing. The special type of missing values patterns is motivated by the two-mode data discussion given below.

In the conclusions further efforts are mentioned to tackle the problem of missing values which occur in a lot of interesting data sets.

TWO-MODE DATA AS AN EXAMPLE FOR DATA WITH MISSING VALUES

Assume, one has a set of elements of a first mode, $M_1 = \{m_{11}, ..., m_{14}\}$, as depicted by stars in Fig. 1 where Fig. 1 should be interpreted as showing a section of a two-dimensional Euclidean space. It is easy to check that the corresponding matrix of Euclidean distances of the first mode (M_1) elements which is given in the upper triangle of Tab. 1 is not pyramidal and, thus, not ultrametric. As this example is so simple, questions like "What are hierarchical or pyramidal clustering solutions of the elements of M_1, respectively?" can be answered immediately. Fig. 2a shows a pyramidal clustering solution obtained by the PAC algorithm (its complete-linkage version).

Assume, there is an additional set of elements of a second mode, $M_2 = \{m_{21}, ..., m_{25}\}$, which is depicted by circles in Fig. 1. Now, the lower triangle of Tab. 1 presents the corresponding Euclidean distances of the second mode (M_2) elements, and Fig. 2b gives the respective PAC solution.

These simple examples of applications of pyramidal clustering already show which advantages pyramids have in comparison to dendrograms which depict the corresponding hierarchies. (The reader is asked to draw her/his dendrogram solutions from the data of Tab. 1 and the visual displays of Fig. 1.)

Now, it suggests itself to ask whether there are relations between first mode (M_1) and second mode (M_2) elements and whether there would be possibilities to perform a joint (pyramidal) clustering of the elements of both sets.

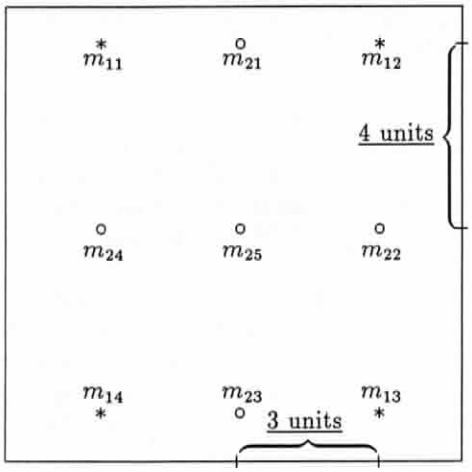


Fig. 1: Graphical Display of Sets of Different Modes in a Two-Dimensional Euclidean Space
($*$ = First Mode (M_1) Elements,
$\circ$ = Second Mode (M_2) Elements)

	m_{11}	m_{12}	m_{13}	m_{14}	m_{21}	m_{22}	m_{23}	m_{24}	m_{25}
m_{11}	0								
m_{12}	6	0							
m_{13}	10	8	0						
m_{14}	8	10	6	0					
m_{21}	■	■	■	■	0				
m_{22}	■	■	■	■	5	0			
m_{23}	■	■	■	■	8	5	0		
m_{24}	■	■	■	■	5	6	5	0	
m_{25}	■	■	■	■	4	3	4	3	0

Tab. 1: Euclidean Distances of the First Mode (M_1) and Second Mode (M_2) Elements of Fig. 1

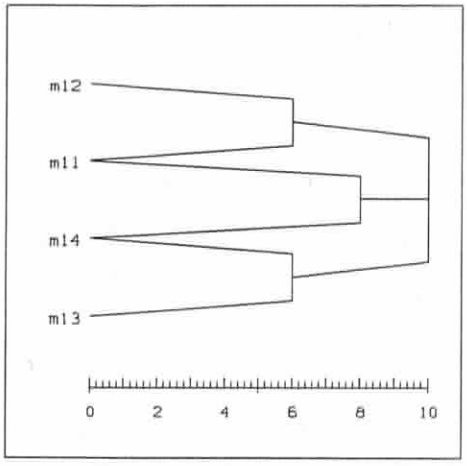


Fig. 2a: Pyramid for the First Mode (M_1) Elements Obtained by PAC

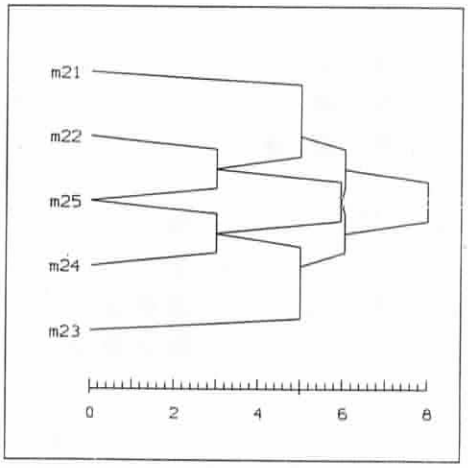


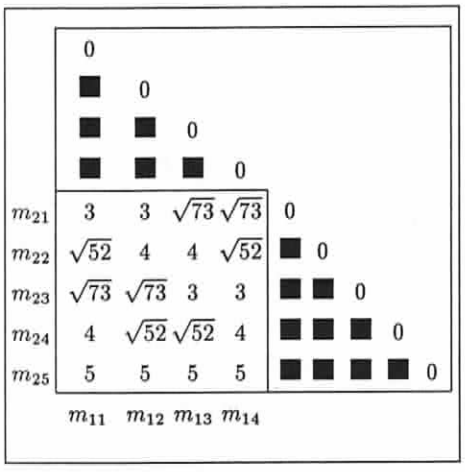
Fig. 2b: Pyramid for the Second Mode (M_2) Elements Obtained by PAC

Having a look at Tab. 1 we see that there are no data which would allow us to draw conclusions with respect to relations between elements of M_1 and M_2 unless some entries in the area indicated by the black boxes are available. (See, however, the recent contribution by Brossier (1990) for computing hierarchies on an overall set if hierarchies on different, possibly overlapping, subsets of the overall set are known.)

Having gone so far, we can start to tackle the reverse situation where we assume that only data are available which describe distances (relations, similarities, or dissimilarities) between pairs of first mode and second mode elements from the different sets M_1 and M_2. Data of this kind are called two-mode data. Tab. 2 gives the Euclidean distances between all pairs of elements from the different sets M_1 and M_2 on the basis of their positioning as presented in Fig. 1.

Now, we can ask whether the information of "between" relations between pairs of elements from the different sets M_1 and M_2 as given in Tab. 2 would allow to recover (better: to approximate) the missing triangle pattern of "within" relations for the elements of the single sets M_1 and M_2 which we already know from Tab. 1 and, again, whether a pyramidal clustering of both sets of first mode and second mode elements in a joint pyramid would be possible.

Obviously, algorithms are needed which are able to handle such situations because there are lots of problems where two-mode data, i.e., data on relations between pairs of elements from different sets, occur. Fig. 3 already shows a two-mode pyramidal clustering solution on the basis of the data of Tab. 2 yielded by application of the PACII modification to be described in the following.



Tab. 2: Two-Mode Data Depicting Relations Between the First Mode (M_1) and Second Mode (M_2) Elements of Fig. 1

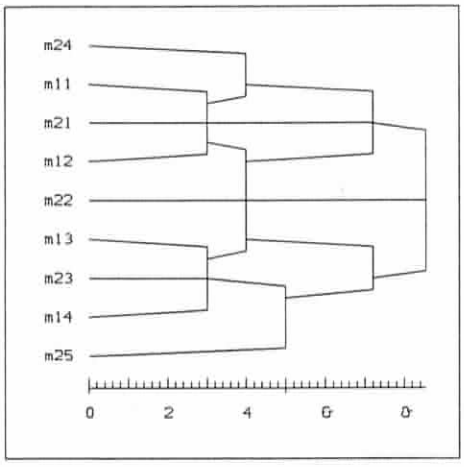


Fig. 3: Two-Mode Pyramidal Clustering on the Basis of the Data of Tab. 2 Obtained by PACII

Furthermore, an algorithm designed to handle missing data should allow for arbitrary kinds of patterns of missing values as the PACII modification of PAC will do. The underlying data of Tab. 1 and Tab. 2 were just chosen to allow an easy check of the positions of the elements in Fig. 1 and the corresponding Euclidean distances.

THE PACII ALGORITHM

For an explanation of the PACII (Pyramidal Ascending Clustering with Incomplete Information) modification we think that it would be easiest to take the PAC description of Diday (1987) and add/change those parts of sentences—indicated by letters in italics—which refer to alterations needed.

This also means that we do not have to repeat all the definitions and notations given in Diday (1987), e.g., that Ω denotes the finite set of objects (perhaps of different modes) which we want to cluster, e.g., that a part (a subset) of Ω is called connex with respect to an order θ on Ω, if it is an interval of this order.

Several versions of the PACII algorithm (single-linkage, complete-linkage, ...) are available dependent on the "philosophy" used in the aggregation step.

If $d(i,j)$ is used to denote the dissimilarities between elements $i, j \in \Omega$ the PACII modification of the PAC algorithm (see, e.g., Diday (1987), p. 14) can be described as follows:

a) Each element of Ω is called "group"; *we calculate dissimilarities* $D(\{i\},\{j\})$ *between groups by setting*

$$D(\{i\},\{j\}) := \begin{cases} d(i,j), & \textit{if } d(i,j) \textit{ is known;} \\ \textit{missing}, & \textit{otherwise.} \end{cases}$$

b) We aggregate the two nearest groups K and L among the groups which have not been aggregated twice, *and for which* $D(K,L)$ *is non-missing.* $D(K,L)$ *is then replaced by a "missing" value. New dissimilarities* $D(K \cup L, J)$ *are computed for each group* J *different from* K *or* L.

For the average-linkage version of PACII the corresponding formula would be

$$D(K \cup L, J) := \begin{cases} \frac{1}{|(K \cup L) \times J - M|} \sum\limits_{(i,j) \in (K \cup L) \times J - M} d(i,j), & \textit{if } |(K \cup L) \times J - M| \neq 0; \\ \textit{missing}, & \textit{otherwise,} \end{cases}$$

where $M := \{(i,j) \in \Omega^2 \colon i = j$ *or* $d(i,j)$ *missing*$\}$.

Following the same concept, we obtain additional versions (single-linkage, complete-linkage, ...) of PACII.

With the exception of complete-linkage we do not recommend usage of recurrence formulas, since either they cannot be applied in this situation or they are of no computational advantage.

c) We start again with b) until a group which contains Ω is formed *or no more inter-group dissimilarities D are known.*

d) Each time a group is formed by merging two groups we must associate an order on those two groups. Thus the algorithm builds up an order θ on Ω.

e) Two groups cannot be merged if their union is not connex.

f) Let i and j be extreme elements of the connex part of Ω associated to a group H; no group can be connected to a group included in H which does not contain either i or j.

In a finite number of iterations the algorithm converges to an indexed pyramid on Ω or—if there are too many missing values—to two or more indexed pyramids on disjoint subsets of Ω. The problem of possible inversions remains the same as with PAC.

Of course, when incomplete data occur corresponding considerations are also reasonable with respect to hierarchical clustering applications which could result in what could be called a "HACII" algorithm (See Schader, Gaul (1991)).

Additionally, one could try to take into consideration different approaches to handle missing values situations within data analysis problems, and start to compare recovery properties of the different approaches. (See, e.g., Gaul, Schader (1990) for a least squares based penalty approach called PLSC (Pyramidal Least Squares Classification) to analyze dissimilarity data by pyramidal clustering.)

Next, we present a special example in order to demonstrate recovery properties of the PACII complete-linkage version.

EXAMPLE

To be able to describe salient features of the performance of the PACII algorithm we take as a known set of data the distances between 15 selected French cities and omit special patterns of distances in agreement with what has already been discussed with respect to two-mode data.

Surely, there are lots of possibilities to handle different types of missing values patterns. In the following, we use rectangular shapes for the non-missing data parts because this kind of information has to be considered in the discussion about the evaluation of two-mode data which is another topic of our interest. The term "$m_1 \times m_2$ array" means that the total of cities has been divided into two disjoint subsets of m_1 first mode and m_2 second mode cities so that information is only available for all pairs of elements from the different subsets of cities. Note that in this terminology the matrix representation of a $m_1 \times m_2$ array has m_1 columns and m_2 rows.

Tab. 3 shows the distances between the French cities selected which were taken from a road map of France and normalized for reasons of comparability with other studies. Tab. 4a and Tab. 4b present part of the experimental design which we used for this example.

Taking cities as elements of the subsets of different modes, one could argue that the "between" relations between the two subsets of cities are modeled by a bipartite graph

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Toulouse	0														
2	Bordeaux	.103	0													
3	Limoges	.121	.090	0												
4	Nantes	.226	.136	.130	0											
5	Brest	.339	.240	.252	.122	0										
6	Caen	.305	.236	.187	.113	.155	0									
7	Le Havre	.321	.255	.200	.139	.177	.027	0								
8	Paris	.279	.239	.162	.162	.245	.094	.084	0							
9	Reims	.318	.295	.211	.230	.307	.153	.136	.066	0						
10	Nancy	.327	.327	.237	.290	.379	.233	.214	.137	.080	0					
11	Mulhouse	.312	.335	.245	.323	.424	.282	.268	.186	.140	.066	0				
12	Dijon	.248	.254	.165	.245	.351	.217	.209	.127	.109	.083	.081	0			
13	Lyon	.175	.215	.137	.254	.373	.265	.265	.190	.192	.167	.139	.088	0		
14	Marseille	.155	.246	.208	.338	.460	.379	.385	.315	.323	.292	.249	.217	.131	0	
15	Nice	.227	.164	.256	.383	.506	.408	.407	.330	.323	.274	.218	.215	.141	.074	0

Tab. 3: Distance Matrix for 15 Selected French Cities

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2	■	0													
3	.121	.090	0												
4	.226	.136	■	0											
5	.339	.240	■	■	0										
6	.305	.236	■	■	■	0									
7	.321	.255	■	■	■	■	0								
8	.279	.239	■	■	■	■	■	0							
9	.318	.295	■	■	■	■	■	■	0						
10	.327	.327	■	■	■	■	■	■	■	0					
11	.312	.335	■	■	■	■	■	■	■	■	0				
12	.248	.254	■	■	■	■	■	■	■	■	■	0			
13	.175	.215	■	■	■	■	■	■	■	■	■	■	0		
14	.155	.246	■	■	■	■	■	■	■	■	■	■	■	0	
15	.227	.164	■	■	■	■	■	■	■	■	■	■	■	■	0

Tab. 4a: 2 × 13 Array (Configuration 2)

and that from all the information given one tries to conclude how the "within" relations for the elements in the subsets of single mode cities look like and how they fit into an overall pyramidal clustering scheme where the elements of the sets of different modes are jointly presented. In Fig. 4a the location of the French cities are shown on a map of France together with the distances used when the 2 × 13 array of Tab. 4a is selected. The question is, of course, whether from this limited information the PACII algorithm would

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2	■	0													
3	■	■	0												
4	■	■	■	0											
5	■	■	■	■	0										
6	■	■	■	■	■	0									
7	■	■	■	■	■	■	0								
8	■	■	■	■	■	■	■	0							
9	■	■	■	■	■	■	■	■	0						
10	■	■	■	■	■	■	■	■	■	0					
11	.312	.335	.245	.323	.424	.282	.268	.186	.140	.066	0				
12	.248	.254	.165	.245	.351	.217	.209	.127	.109	.083	■	0			
13	.175	.215	.137	.254	.373	.265	.265	.190	.192	.167	■	■	0		
14	.155	.246	.208	.338	.460	.379	.385	.315	.323	.292	■	■	■	0	
15	.227	.164	.256	.383	.506	.408	.407	.330	.323	.274	■	■	■	■	0

Tab. 4b: 10 × 5 Array (Configuration 10)

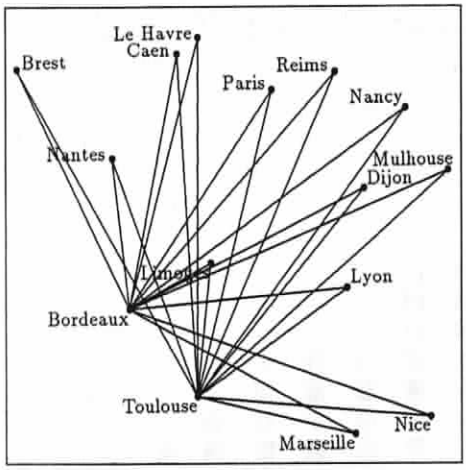


Fig. 4a: Distances known on the Basis of the 2 × 13 Array of Tab. 4a where the Positioning of the Cities Reflect their Locations on a Map of France

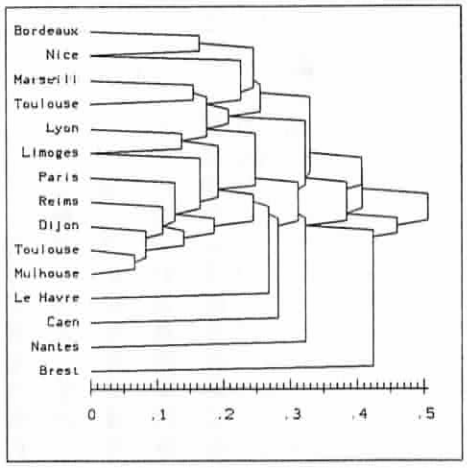


Fig. 4b: PACII Solution on the Basis of the 10 × 5 Array of Tab. 4b (Configuration 10)

be able to construct a pyramidal clustering solution which comes near to the distance structure of the problem without missing data. Fig. 4b gives the PACII solution on the basis of the 10 × 5 array of Tab. 4b.

In Fig. 5, for all $m_1 \times m_2$ arrays of the French cities example the PACII results (dashed line) are presented in terms of CCC (Cophenetic Correlation Coefficient). Additionally, the results of the PLSC algorithm mentioned before (see, e.g., Gaul, Schader (1990) for a description of algorithmic aspects) are displayed (solid line).

For $m_1 = m_2$ (if $m_1 + m_2$ is even) or $m_1 = m_2 \pm 1$ (if $m_1 + m_2$ is odd) it is likely that one gets the best fit in terms of CCC because in such a situation the number of missing values is minimal.

The differences between the dashed and solid lines in Fig. 5 indicate the extent (in terms of CCC) by which PACII is outperformed by PLSC. However, in terms of CPU time PACII is several orders of magnitude faster than PLSC. The computing time for being able to draw the dashed line and the solid line was 1–2 minutes and several hours, respectively. PACII seems to react to a greater extent (in terms of CCC) to the special shape of missing values patterns. Clearly, a PACII output can be used as starting solution for the PLSC algorithm if one wants to improve the CCC-fit. But CCC-fit is just one criterion. Whether recovery properties of the PACII algorithm are already satisfactory is a question which also depends on the order constructed during the computation of the solution. The PACII solution based on the 8×7 array (configuration 8) which is depicted in Fig. 6a and which has a "good" CCC-fit (see Fig. 5) is still "similar" to the PACII solution based on the 10×5 array (configuration 10) which we already know

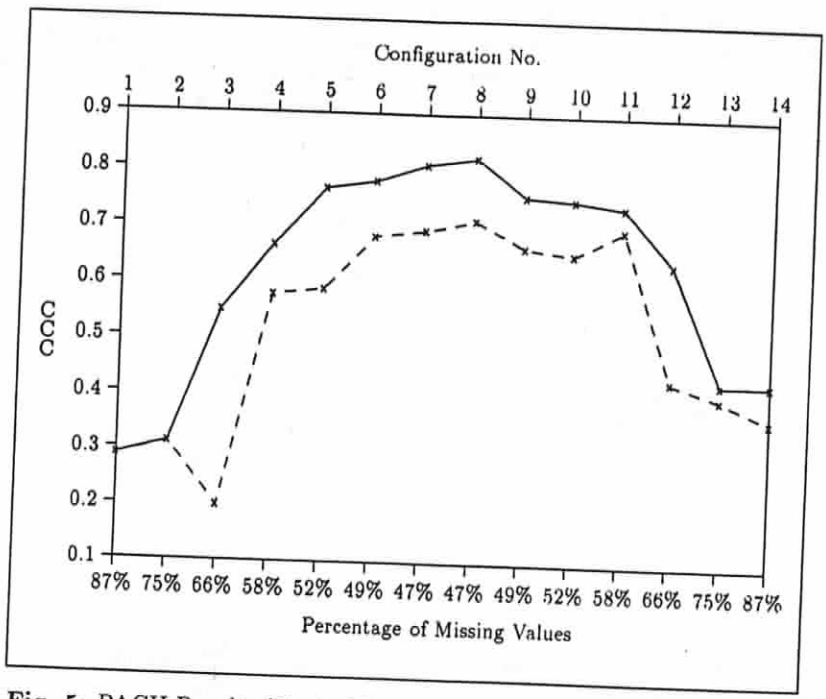


Fig. 5: PACII Results (Dashed Line) and PLSC Results (Solid Line) in Terms of CCC

from Fig. 4b. However, the PACII solution based on the complete data gives a totally different ordering as can be seen from Fig. 6b.

When comparing the results of Fig. 4b, Fig. 6a, and Fig. 6b with the locations of the cities on a map of France as depicted in Fig. 4a the (pyramidal) clusters obtained by PACII are not too bad but the different orderings of the different solutions may be somehow confusing to those not familiar with pyramidal clustering.

However, orders are just used by the algorithm for the special kind of overlapping of clusters which pyramidal clustering generates. When interpreting the pyramidal clustering results for application purposes the ordering of the objects clustered is of no importance (e.g., the reversed order with respect to the clustering results would give the same solution, e.g., an alteration of suborders with respect to those objects which belong to non-overlapping clusters does not effect the overall solution). Indeed, in the PLSC approach orderings are rearranged in trying to find a best solution, and as we have seen from Fig. 5, the solutions of the PACII modification can still be improved. At the moment, the simplicity of the alteration of the PAC algorithm to adjust it to what we have called the PACII modification is one of the advantages which should be stressed.

CONCLUSIONS

In this paper we have presented the PACII generalization of the well-known PAC algorithm for constructing pyramidal clustering solutions when the underlying dissimilarity data are incomplete. With respect to the suitability of the PACII algorithm we have argued on the basis of visual inspection and in terms of CCC. Other goodness-of-fit criteria, as, e.g., VAF (Variance Accounted For) and TIC (Theil's Inequality Coefficient) gave similar results.

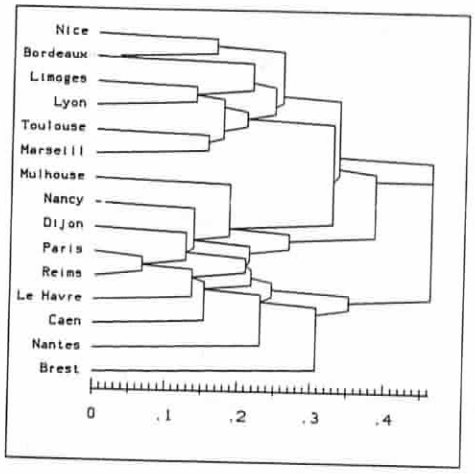


Fig. 6a: PACII Solution on Basis of the 8×7 Array (Configuration 8)

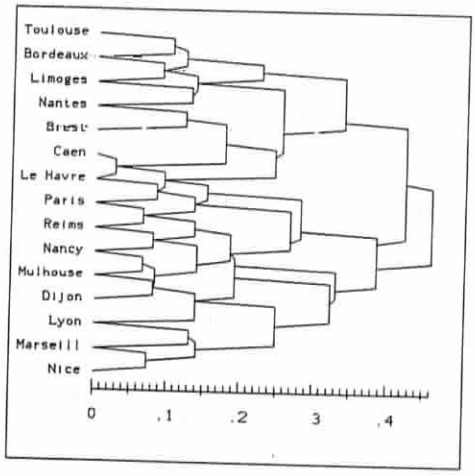


Fig. 6b: PACII Solution on the Basis of the Complete Data

As already indicated in the example above, the consideration of strategies for alterations of orderings of the underlying objects, so that "best" overlappings in the sense of pyramidal clustering are yielded, is of interest. Here, further research could be devoted to the problem of how to perform step d) of the algorithm described before by taking into account alteration possibilities of (sub)orderings to improve fit criteria.

In the underlying paper the discussion on missing values which appear in data to be analyzed was restricted to two-mode data. Here, first experiences have already been gathered. (In Gaul, Schader, Both (1990) pyramidal clustering was already combined with applications of other data analysis techniques, e.g., with the CAR (Clusterwise Aggregation of Relations) approach (Gaul, Schader (1988)) to support interpretation of consumer behavior on the basis of two-mode data.)

Of course, the analysis of other types of data in which missing values could occur is of interest. This was just one study of several others (see, e.g., Gaul, Schader (1990) and Schader, Gaul (1991)) to find out the reaction of data analysis techniques to incomplete input, how algorithms have to be modified or newly developed to cope with situations when data are incomplete, and how data with special shapes of missing values patterns can be evaluated.

REFERENCES

[1] P. Bertrand, Étude de la Représentation Pyramidale, Thèse, Université Paris-Dauphine, (1986).

[2] P. Brito and E. Diday, Pyramidal Representation of Symbolic Objects, In Knowledge, Data and Computer-Assisted Decisions, (Eds. M. Schader and W. Gaul), Springer, N.Y., (1990), 3–16.

[3] G. Brossier, Piecewise Hierarchical Clustering, J. of Classification 7, (1990), 197–216.

[4] E. Diday, Orders and Overlapping Clusters by Pyramids, Rapports des Recherche, N° 730, INRIA, (1987).

[5] E. Diday and P. Bertrand, An Extension of Hierarchical Clustering: The Pyramidal Presentation, In Pattern Recognition in Practice II, (Eds. E.S. Gelsema and L.N. Kanal), North-Holland, Amsterdam, (1986), 411–424.

[6] W. Gaul and M. Schader, Clusterwise Aggregation of Relations, Applied Stochastic Models and Data Analysis, 4, (1988), 273–282.

[7] W. Gaul and M. Schader, Pyramidal Classification Based on Incomplete Dissimilarity Data, (1990), submitted.

[8] W. Gaul, M. Schader and M. Both, Knowledge-Oriented Support for Data Analysis Applications to Marketing, In Knowledge, Data and Computer-Assisted Decisions, (Eds. M. Schader and W. Gaul), Springer, N.Y., (1990), 259-271.

[9] O. Opitz, Numerische Taxonomie, Fischer, Stuttgart, (1980).

[10] O. Opitz, (Hrsg.), Numerische Taxonomie in der Marktforschung, Vahlen, München, (1978).

[11] M. Schader and W. Gaul, The MVL (Missing Values Linkage) Approach for Hierarchical Classification when Data are Incomplete, to appear in Proceedings of the 15th Annual Conference of the GfKl, Springer, N.Y., (1991).