APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 2, 13–30 (1986)

Biometrics, medicine and health care
Data aggregation

CLASSIFICATION OF MEDULLARY LESIONS AMONG PARAPLEGIC PATIENTS BY THE LIKELIHOOD OF THE LINKS METHOD

REPLY TO PARISOT'S PROBLEM

(*Applied Stochastic Models and Data Analysis*, **1**,(1), 35–54(1985))

I. C. LERMAN
IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France

SUMMARY

Relative to the problem presented and treated by Parisot in an earlier issue of this journal,[1] we present a new solution based on our approach of hierarchical classification. The methodology used, that we will describe in a concise way, allows us to organize either the set of the descriptive variables or the set of the individuals. Its extreme generality includes a set theoretic coding which respects faithfully the globality of the intrinsic nature of the descriptive variables. Very rich results are obtained in terms of classification trees reduced to their 'significant' levels. One of the most significant levels corresponds to the partition obtained by Parisot. The 'significant' nodes indicate the most relevant associations. The signification is conceived from the empirical distribution — on the successive tree levels — of association coefficients between partition and 'preordonnance' (i.e. total preorder on the set of unordered object pairs).

KEY WORDS Data analysis Hierarchical classification Validation and significance Qualitative ordinal data Medical data Neurology

RECALL OF THE PROBLEM

We will take here the presentation of the problem, given by Parisot.[1]

In order to study a possible link between the characteristics of the spinal medulla and the occurrence of a particular disease, we want to find (as a first step) a classification of medullary lesions among paraplegic patients. First, let us recall some definitions (reproduced, slightly modified, from reference 1, pp. 36–37).

The spinal medulla — the medullary segments

The spinal medulla, an organ which belongs to the central nervous system, is a voluminous fascicule of nerve fibres, below the rachidian bulb. It is situated inside the vertebral column which then normally protects it. At its top, the spinal medulla is connected to the intracranial nervous centres. Along its whole length, it is connected to thirty pairs of rachidian nerves. These rachidian nerves carry the following information between the central nervous system and the striated muscles (those involved in the voluntary movements):

(i) sensory information (upward fibres)
(ii) motor information (downward fibres).

8755–0024/86/010013–18$09.00

Received October 1985
Revised January 1986

The spinal medulla represents the merging of all these nerve fibres, with, schematically, *motor* ones in its *front* part and *sensory* ones in its *back* part. Apart from this role of communication, the spinal medulla is also a nerve centre in the sense that it contains direct connections between some sensory and motor fibres, realizing short circuits called *reflex arches;* these reflexes provide an important part of the involuntary movements.

Because of this last function, the spinal medulla is classically divided into theoretical slices called *medullary segments,* each one containing the reflex arches corresponding to *one* pair of rachidian nerves.

From the top to the bottom of the spinal medulla we find thirty medullary segments, identified in the same terms as vertebrae:

8 cervical segments, from C1 to C8
12 dorsal segments, from D1 to D12
5 lumbar segments, from L1 to L5
5 sacral segments, from S1 to S5.

We can schematically consider that each segment is an area with proper motor and sensory functions.

Traumatic paraplegia — the medullary lesion

A fracture of the vertebral column, due to a trauma, can affect the spinal medulla by direct compression or stretching. The blood vessels running along the medulla and 'feeding' it are often damaged, indirectly affecting lower or upper segments of the medulla.

Finally, a medullary area will be destroyed, with great variability in length, and not necessarily right/left or front/back symmetry. The upper boundaries of the destroyed medullary area determine the upper boundaries of paralysis, producing at least a *paraplegia* which is a paralysis of lower limbs. The lower boundaries of the destroyed medullary area determine the upper boundary of an automatic refunctioning of the medulla: the reflexes remain active, but the voluntary movements obviously do not. These lower boundaries do not always exist, the spinal medulla being possibly destroyed down to its very bottom.

Population — variables

Our study population represents 100 paraplegic patients (so, lesions). Following the patient's identification (002GALG e.g.), we have defined 12 variables for the lesion, since we noticed that the lesion boundaries were not necessarily the same on the right and left sides of the same medullary segment, or on the sensory and motor parts of this segment.

Upper boundaries of the lesion (right and left)

1. First segment with reduced sensitivity (V1R and V1L).
2. First segment with null sensitivity (V2R and V2L).
3. First segment with reduced motor functions (V3R and V3L).
4. First segment with null motor functions (V4R and V4L).

Lower boundaries of the lesion (right and left)

1. First segment where a sensory function restarts (V5R and V5L).
2. First segment where a motor function restarts (V6R and V6L).

For these lower boundaries, it is not possible to make the distinction between 'normal' and 'reduced'; besides, they may not even exist when the lesion spreads down to the bottom of the medulla.

These 12 variables were coded as segments, that is to say as qualitative variables, with $(30 + 1)$ modalities:

(a) C1 to C8 (first to 8th cervical segment)
(b) D1 to D12 (first to 12th lumbar segment)
(c) L1 to L5 (first to 5th lumbar segment)
(d) S1 to S5 (first to 5th sacral segment)
(e) -1 when a (lower) boundary does not exist.

The raw data matrix is given in reference 1, pp. 51–53. In any case, the rows of this table, reordered according to the classification tree, are given in Figure 3 (below).

SOME GENERAL CHARACTERISTICS OF THE METHOD USED

Our solution is based on a general method of hierarchical classification that we have elaborated: the method of the 'likelihood of the links (or associations)'.

We will begin by giving the general framework of the method and will end by mentioning detailed references which justify and develop the approach.

The first characteristic of this method is to adopt a set theoretic representation of the descriptive variables, with regard to the set E of the objects or individuals.

According to these preliminaries we have defined a typology of the descriptive variables depending on the level of their faithful mathematical representation or coding, which may be E, $E \times E = E^2$ or $(E \times E) \times (E \times E) = E^4$, where $\times$ denotes the cross product.

In fact we have — in our experience — considered the following specific types of variables or data:

(i) quantitative variables
(ii) logical variables or attributes
(iii) rows or columns of a juxtaposition of contingency tables
(iv) qualitative nominal variables
(v) qualitative ordinal (partial or total) variables
(vi) weighted graphs on E
(vii) qualitative 'preordonnance' variables.

The first three types ((i), (ii) and (iii)) correspond to a representation at the level of E; the faithful mathematical coding of (iv), (v) and (vi) must at least be at the level of $E \times E$. For the last type — that which we deal with in this real example — the representation, which is originally at the level of $(E \times E) \times (E \times E)$, may be reduced to a 'ranking' function on $E \times E$. More precisely, when the relation defined by the variable is symmetric, we work — for reasons of simplicity — at the level of the set $P_2(E)$ of unordered object pairs instead of the set $E \times E$ of ordered object couples.

Indeed, the principle of the algorithm of hierarchical classification is — by itself — trivial. As a matter of fact, if S denotes the set to be organized and if P denotes a proximity index on any exhaustive system of subsets of S, the algorithm proceeds by consecutive coalescences from the starting state defined by $\{\{t\}/t \in S\}$. More precisely at each step (or level), we aggregate —among the already formed cluster — the pairs of clusters (which represent disjoint subsets of S) corresponding to the maximum of P. The result is a classification tree on S.

The set S may be either the set V of the descriptive variables (represented by the set of the columns of the data table) or the set E of objects (represented by the set of the rows of the data table).

Nevertheless, the crucial problem concerns the definition of the proximity index P; that is to say, the formalization of the notion of 'resemblance' ('similarity'), between elements and disjoint subsets of the set S to be organized, taking into account the mathematical structure of the data table.

The second and the most important distinctive aspect of our methodology concerns a very synthetic, unified and fruitful approach for constructing association coefficients between variables and similarity indices between objects. This approach allows us to introduce, in an original way, a probability scale for measuring the proximities or associations between disjoint subsets of S, in terms of likelihood of the observed links.

More precisely, we start our construction at the level of the definition of an association or similarity coefficient between two given elements t and u of S. This construction includes two steps. The former is the proposition of a 'rough' index $s(t, u)$ which seems 'very natural' from a formal point of view. The latter — and the most important step — consists of a standardization of $s(t, u)$ with respect to a 'hypothesis of non-link' which corresponds to an *adequate* random model, associating with the pair (t, u) a pair (t', u') of independent random elements. The randomness takes into account the formal and statistical structure of the data. On the other hand, it is of importance to note that the problem of defining similarity indices between objects is different in nature from that of defining association coefficients between variables. In both cases, the random model of non-link associates the random variables with the observed ones. The coefficient or index defined by the following formula is called the 'locally' standardized coefficient or index:

$$Q(t, u) = \frac{s(t, u) - E[s(t', u')]}{\sqrt{\text{var}(s(t', u'))}} \tag{1}$$

where E and var denote the mathematical expectation (or mean) and the variance of the random variable (r.v.) $s(t', u')$, respectively. Then, we denote by

$$\{Q(t, u) \mid \{t, u\} \in P_2(S)\} \tag{2}$$

the table — indexed by the set of unordered pairs of elements of S — of these 'similarities' (or 'associations').

The local standardization is followed by a global one defined at the level of (2), where $Q(t, u)$ is replaced by

$$Q_s(t, u) = \frac{Q(t, u)\text{-moy.}_e(Q)}{\sqrt{\text{var.}_e(Q)}} \tag{3}$$

where $\text{moy.}_e(Q)$ and $\text{var.}_e(Q)$ indicate the mean and the variance of (2), respectively (see Table I.).

Then, the new table of the similarities may be written as follows:

$$\{Q_s(t, u) \mid \{t, u\} \in P_2(S)\} \tag{4}$$

The final table of the similarity (or association) indices, interpreted as the likelihood of the observed relationships, is given by

$$\{P(t, u) = \Phi[Q_s(t, u)] \mid \{t, u\} \in P_2(S)\} \tag{5}$$

where Φ is the normal N(0, 1) cumulative distribution function.

Hence a probability scale is introduced to measure the proximities between the elements of the set S to be organized.

To build the classification tree on S, it is necessary to extend this similarity measure between elements to those between disjoint subsets of S. Let B and C be two disjoint subsets of S, then the association measure that we have invented (1970)[3] is defined by the likelihood of the strongest link — measured by P (cf.(5)) — between two elements belonging to B and C, respectively. The index obtained is

$$\pi(B, C) = (\max\{P(b,c)/(b, c) \in B \times C\})^{|B| \times |C|} \tag{6}$$

where $|\cdot|$ indicates the cardinality of $\cdot$.

This index underlies the 'likelihood link algorithm' (L.L.A.) in which — for accurate computing — we consider the strictly increasing transformation $-\log[-\log(\cdot)]$. Thus, we directly use

$$\chi(B, C) = -\log\{-\log(\pi(B, C))\} \tag{7}$$

So that the reactualization formula becomes

$$\chi(B, C \cup D) = -\log(|C| + |D|) + \max\{\chi(B, C) + \log(|C|), \chi(B, D) + \log(|D|)\} \tag{8}$$

With this formula we construct — step by step — the Polish representation of the classification tree (see Table III.).

A decisive stage of the method consists of a dynamical interpretation of the sequence of the tree levels, which leads to the recognition of the most significant levels and of the significant nodes. These interpretations and recognitions are based upon a statistical association coefficient between the last partition to have emerged — at a given level — and an adequate proximity structure retained from the resemblances between the elements of the set S to be classified.

Let T be the set $P_2(S)$ of unordered element pairs from S (i.e. the set of all subsets of S with two elements). The mathematical representation of a partition π (which may be determined at a given level of the classification tree) that we consider here is the Cartesian product:

$$S(\pi) \times R(\pi) \subset T \times T \tag{9}$$

where $S(\pi)$ is the set of pairs of objects separated in different classes by the partition π and where $R(\pi)$ is the set of pairs of objects (gathered) by the partition into the same class.

Two main types of proximity structure on S are considered. The former has an ordinal characteristic and is defined by the 'preordonnance' associated with the similarity index (or association coefficient) Q_s (cf.(3) above) and the latter has a numerical characteristic, and is thus directly defined by the table (4) above.

The 'preordonnance' is a total preoder ω on T, defined as follows:

$$(\forall (p, q) \in T \times T), p \leq q \Leftrightarrow Q_s(p) \leq Q_s(q) \tag{10}$$

In the case where ω is reduced to a total and strict order on T, we represent ω by

$$\mathrm{gr}(\omega) = \{(p, q) \mid (p, q) \in T \times T, p < q \text{ and not } q < p \text{ for } \omega\} \tag{11}$$

The 'rough' index between ω and π is then defined by

$$s(\omega, \pi) = \mathrm{card}(\mathrm{gr}(\omega) \cap (S(\pi) \times R(\pi))] \tag{12}$$

This index is standardized with respect to the hypothesis of non-link, where we associate with π a random element π' in the set, provided by a uniform measure of probability, of all

partitions having a given type, the type of a partition being defined as the sequence of the cardinalities of its classes.

The resulting association coefficient

$$\sum(\omega, \pi) = \frac{s(\omega, \pi) - E(s(\omega, \pi'))}{\sqrt{\text{var}[s(\omega, \pi')]}} \tag{13}$$

is what we call the 'global statistic'.

In fact, and with respect to the standardization followed, the rough index $s(\omega, \pi)$ is equivalent to

$$s_1(\omega, \pi) = \sum\{\xi(p)k(p)/p \in T\} \tag{14}$$

where $\{\xi(p)/p \in T\}$ is the indicator function of $R(\pi)$ and where $\{k(p)/p \in T\}$ is the ranking function on T defined by the total and strict order ω.

This last form (14) may be generalized to the case where the data are a total preorder on T, by defining the ranking function in a suitable way. Afterwards the standardization is done in an analogous way.

The criterion mentioned above as having numerical character is built — with the same approach — from the starting expression

$$s_2(Q_s, \pi) = \sum\{\xi(p)Q_s(p)/p \in T\} \tag{15}$$

The 'global statistic' criterion denoted by Σ has in practice some different qualities according to its conception: ordinal or numerical. Nevertheless, the most significant levels correspond to the most distinctive local maxima of the following sequence of values (see Figure 1(a)):

$$(\textstyle\sum_j = \sum(\omega, \pi_j)/1 \leq j \leq l\} \tag{16}$$

where π_j is the partition obtained at the jth level of the classification tree.

Another and very important notion that we have introduced (1970)[2] concerns the 'significant nodes', which indicate completion stages of the different classes appearing in the tree. For this, we define a 'local statistic' based on ω(or Q_s) and considering — for each level — the subset of object pairs which have just been joined, with respect to the set of object pairs remaining separated in different classes.

The 'local statistic' θ can be directly deduced from the total one by defining it as the rate of variation of Σ; hence, the value of θ at the jth level is

$$\theta_j = \textstyle\sum_j - \sum_{(j-1)} \tag{17}$$

The significant nodes correspond to the local maxima of the distribution of the local statistic on the increasing sequence of the tree's level (see Figure 1(b)):

$$\{\theta_j/1 \leq j \leq l\} \tag{18}$$

Our technique of reduction of the classification tree consists of retaining the levels where significant nodes are detected. But the tree's representation preserves the totality of the information concerning the sequence of the associations of the detailed tree (cf. the representation tree for the paraplegic patients below).

To complete our presentation we must mention different statistical tools that we have elaborated to control and to make precise the interpretation of the resulting classifications (on the set of the rows (or columns) of the data table).

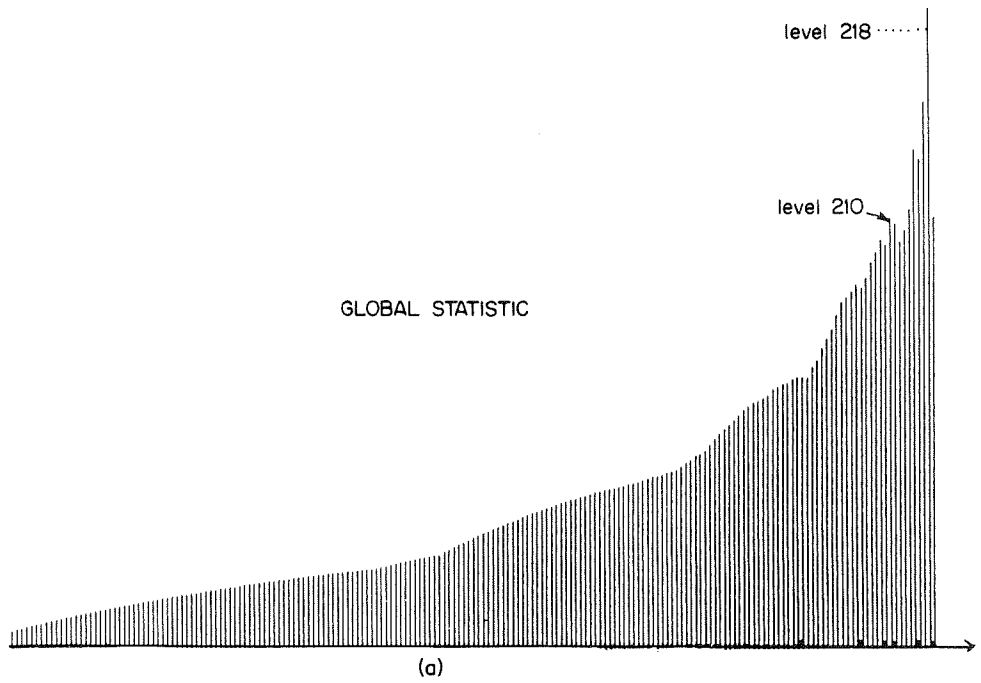


Figure 1(a). Global statistic

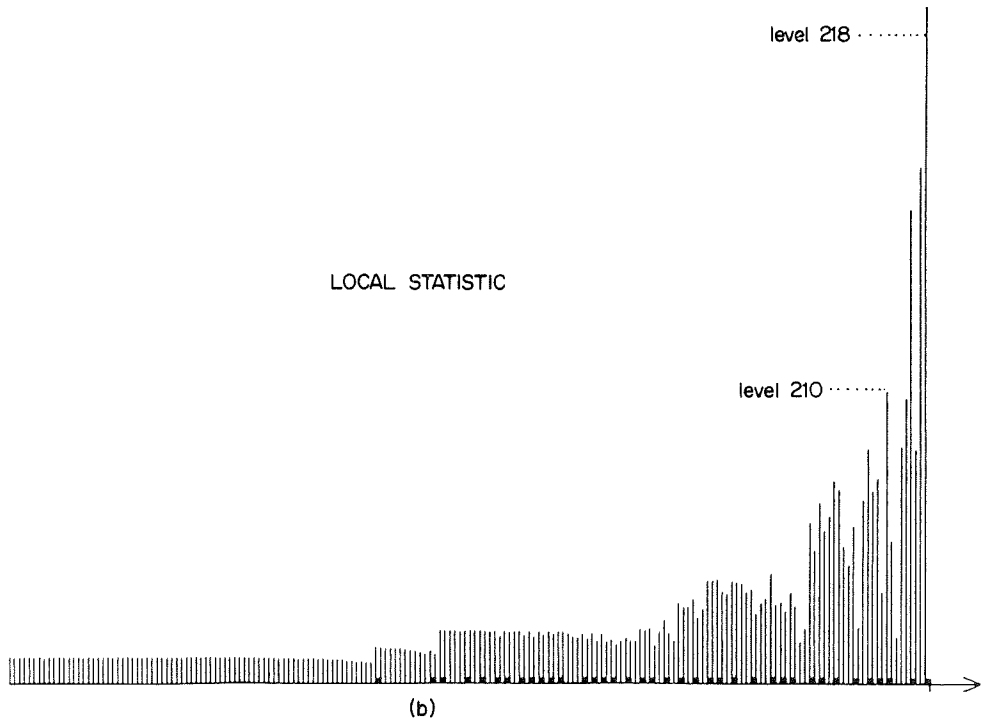


Figure 1(b). Local statistic

The data analysis prior to a classification leads us to measure the neutral character of a given element e of S, with respect to a classificatory goal, by the smallness of the observed variance of the proximities to e, that is to say the variance of the empirical distribution

$$\{Q_s(e,t) \mid t \in S - \{e\}\} \tag{19}$$

which can be written

$$V(e) = \frac{1}{(|S|-1)} \sum \{[Q_s(e,t) - Q_s(e,\cdot)]^2 \mid t \in S - \{e\}\} \tag{20}$$

where $Q_s(e,\cdot)$ is the mean of the distribution (19) (see Table II.).

The bigger the value of $V(e)$ is, the stronger is the role of e in the 'explanation' of a given class containing e.

The data analysis posterior to an automatic classification leads us to a family of indices which may have correlational or inertial (variance) nature and which enable us to

(i) cross classifications — even in the case of very large data —
  (i_1) on the set of objects and on the set of variables
  (i_2) on two disjoint subsets of the set of variables, respectively
(ii) measure the degree of responsibility of a given variable (or group of variables) in the definition of a class of objects or individuals
(iii) measure the degree of responsibility of a given object (or class of objects) in the definition of a tendency corresponding to a class of variables.

Interesting problems arise in considering the treatment of these questions in the context of the different types of data.

The justification and development of our methodology is given in References 4–10, where the approach is well situated with respect to others in data classification.

On the other hand, the computing possibilities of the algorithms performing the methods are very large. Thus, for the classical hierarchical algorithm it is possible to reach on a 'large' computer (e.g. HB 68, Multics system) — in a 'reasonable' computing time — a size of 500 for the set to organize. The process is supposed to include the computing of the preordonnance and the distributions of the level statistics. The size of the other side of the data table certainly influences the computing time for establishing the table of similarity indices. But in this case it is a linear function of the size, which can then become very large (many thousands).

For the problem of classifying a large set (some thousands of units), we have recently set up an algorithm and the program supporting it, based on the parallelism idea.[11]

CLASSIFICATION OF THE VARIABLES

We have two main choices in our method. The former and by far the most important concerns the mathematical coding — of a set theoretic nature — of the variables; the latter concerns one of two near hypotheses of non-link which provide the construction of the association's coefficients (see References 4, 5 (Chapter 2) and 6).

According to Parisot's presentation of the problem of paraplegic patients[1], we consider the twelve variables *directly from their definition.*

Two codings respect the nature of the data. The first one — that we adopt for the classification of the variables — consists of considering each (of the 12 variables) as a qualitative ordinal. The second coding — that we adopt for the main problem of the classification of the paraplegic patients — consists of considering each (of the 12 variables) as 'preordonnance'.[9]

This last type of variable — that we take once more below — is very general and very rich. It has been independently considered in a very different context by Chah.[12,13]

However, for the first problem of organizing the set of variables, the qualitative ordinal coding is simpler and seems perfectly adequate.

More precisely, for each variable, we begin by associating with it a qualitative ordinal variable with, *a priori,* 31 ordinal values coded $0, 1, 2, \ldots, 29, 30$, respectively, where $0, 1, \ldots, 7$ correspond to $C1, C2, \ldots, C8$; $8, 9, \ldots, 19$ to $D1, D2, \ldots, D12, 20, 21, \ldots, 24$ to $L1, L2, \ldots, L5$, $25, 26, \ldots, 29$ to $S1, S_2, \ldots, S5$ and 30 corresponds to -1.

But in fact, for a given variable $\omega_j (1 \leq j \leq 12)$, some of the ordinal values $0, 1, 2, \ldots, 29, 30$, are never reached by any individual (paraplegic patient). Then a recoding of each variable is considered in such a way that — for $\omega_j (1 \leq j \leq 12)$ and from the top to the bottom — 0 is the first ordinal value reached, 1 is the second, and so on, up to r_j, which is the last ordinal value reached.

The interpreting of this representation of the variables can be expressed as follows:

(a) For the first 8 variables: the higher the severity of the lesion is, the lower will be the ordinal value of the variable.

(b) For the last 4 variables: the higher the severity of the lesion is, the higher will be the ordinal value of the variable.

For homogeneity reasons, we replace each of the first 8 variables by its opposite; that is to say for $1 \leq j \leq 8$, ω_j is replaced by $\bar{\omega}_j$, where

$$\bar{\omega}_j(i) = r_j - \omega_j(i), \; 1 \leq i \leq 100$$

Let us recall here the following set theoretic representation of a qualitative ordinal variable:

$$R(\omega) = \{(x, y)/\omega(x) < \omega(y)\} \subset E \times E$$

For comparing qualitative ordinal variables in the total, partial and fuzzy cases, we refer to Reference 5, Chapter 2, where more complete references are given.

With this coding the results are given below.

Let us examine the classification tree with the aid of Table IV, which gives in its second column the sequence (16) and in its first and third columns two different interpretations of the level's local statistic. The third column gives (18) and the first one is conceived by standardizing the 'rough' index card $[R'(\pi_j) \cap (\mathrm{gr}(\omega))]$, where $R'(\pi_j)$ is the subset of unit pairs, joined for the first time at the jth level. In Table IV, the statistics are based on the preordonnance.

Table I.

CARDI = 100,	CARDJ = 12

The global mean and the global variance of the empirical distribution of the similarities are:

Mean = 3·893,	Variance = 0·680

Table II. Increasing values of the coefficient $V(e)$: variance of the proximities to e

Item e	12	11	4	3	3	5
$V(e)$	0·40184	0·40797	0·70019	0·73304	0·82147	0·82677
Item e	2	8	9	7	10	1
$V(e)$	0·87396	0·88356	0·90315	0·91315	0·95729	0·96125

Table III. Aggregation by L.L.A. with global reduction of the similarities. Polish representation of the tree

−11	−10	−8	−5	1	2	−6	3	4	−7	−3	5	6	−2	7	8	−9	−4	9	10
−1	11	12	0																

Table IV.

	Level	Local Statistic	Global Statistic	Variations of the Global Statistic
	1	1·706	1·706	0·000
	2	1·706	2·394	0·688
	3	1·705	2·909	0·515
	4	1·705	3·332	0·423
	5	1·704	3·695	0·363
	6	1·647	3·993	0·297
1. Maximum	7	3·289	4·990	0·997
	8	2·513	5·302	0·312
	9	1·099	4·924	−0·378
	10	5·599	6·980	2·056

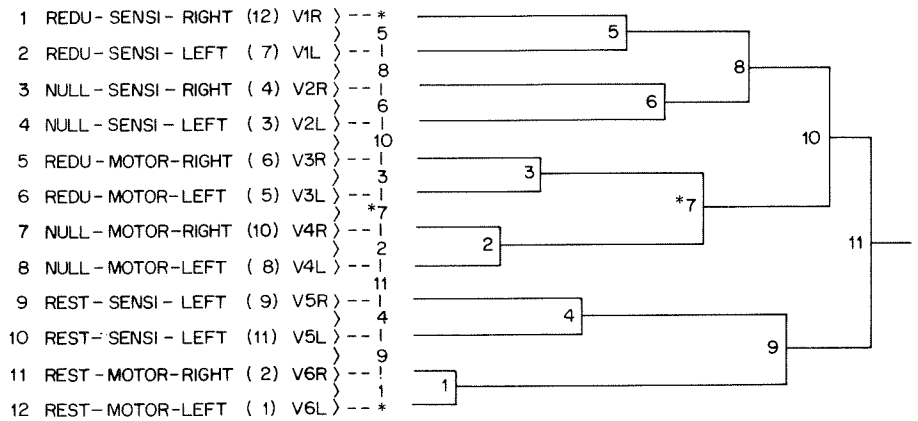


Figure 2. Tree representation

Table II, which gives the increasing sequence of the values of the index (20) will also be of some help in interpreting the results.

It is of some interest to notice that the order on the leaves of the hierarchical proximity structure given by the classification tree is exactly that directly given in the presentation of the problem. Then, the mutual relations between the variables are compatible with this order.

The behaviour of the 'global statistic' Σ through the last levels shows that the main discrimination at the 10th level ($\Sigma(\pi_{10}, \omega) = 6{\cdot}98$) consists of two classes separating the beginning and the end of the lesion.

The class concerning the starting of the lesion is more important in size and more clearly

structured: effectively we notice a significant node at the 7th level, when the variables indicating 'null motor functions' join those indicating 'reduced motor functions'.

It is interesting to point out the separation between the 'sensitivity' and the 'motor functions' in both classes. On the other hand, the behaviour of the right side is directly associated with that of the left side for all the aspects: restarting motor functions, starting null motor functions, starting reduction motor functions, starting reduction sensitivity, starting null sensitivity. The order of starting of these variables is reshaped according to the decreasing values of the strength of the association between right and left sides.

The most neutral parameters [with respect to their associations with the others (cf. Table II)] concern the restarting of motor functions. But as far as the variables of sensitivity functions are concerned, we observe high discrimination in the beginning of the reduction and in the restarting. A rather high degree of discrimination is also noticed for the variables which concern the beginning of the nullity of the motor functions.

Let us now consider the main problem for this study of classifying the set of individuals (paraplegic patients). Beforehand we recall that in our approach — and whichever is the more important of the two problems: classification of the variables (as in the inquiry data) or classification of the objects, individuals — we systematically propose to begin by the classification of the variables before classifying the set of objects. The end of the treatment consists of reorganizing rows and columns of the data table, according to the crossing of two reduced trees, obtained by the classifications of the variables and the objects, respectively.

CLASSIFICATION OF THE PARAPLEGIC PATIENTS

Coding of the variables

It would be perfectly possible to consider here for the descriptive variables exactly the same coding (qualitative ordinal) as that adopted above for the treatment of the variables.[9] But as we have mentioned previously we will consider here a 'preordonnance' coding of the variables.

Let us now define this type of qualitative variable. Let $J = \{1, 2, \ldots, j, \ldots, m\}$ be the set denoting the m modalities of a given qualitative variable q. We suppose that the perception of the mutual resemblances between the modalities enables the expert to give a total preorder on the set of pairs of modalities:

$$J^{(2)} = \{jh = (j, h)/1 \leq j \leq h \leq m\} \tag{21}$$

This ranking (with ties) is established in such a way that the higher the rank of a pair is, the greater is the degree of the perceptual similarity between the two components of the pair.

More precisely, we introduce a ranking function r, defined as follows: for each $jh \in J^{(2)}$

$$r(jh) = \sum_{1 \leq i \leq (k-1)} l_i + (l_k + 1)/2 \tag{22}$$

if jh belongs to the kth class of the total preorder on $J^{(2)}$, then the classes having the cardinalities $l_1, l_2, \ldots, l_k, \ldots, l_n$ are linked by the relation

$$\sum_{1 \leq k \leq n} l_k = m(m+1)/2 = \text{card } (J^{(2)}) \tag{23}$$

For each $j \in J, r(jj)$ is maximum and equals $[l_1 + l_2 + \ldots + l_{(n-1)} + (l_n + 1)/2)$.

This variable determines a preordonnance on E defined by a total preorder on $P_2(E)$, where E is the set of objects (or individuals). If we denote by $E_j (1 \leq j \leq m)$, the subset of the objects

which possess the jth modality of the variable q, we have the following set decomposition of $P_2(E)$:

$$P_2(E) = \sum_{1 \le j \le m} P_2(E_j) + \sum_{1 \le j \le h \le m} E_j * E_h \tag{24}$$

where

$$E_j * E_h = \{\{x, y\}/x \in E_j, y \in E_h\} \tag{25}$$

is the subset of unordered object pairs belonging to E_j and E_h, respectively.

The place of $E_j * E_h$ (resp. $P_2(E_j)$) with respect the ranking determined on $P_2(E)$, is compatible with the value of $r(jh)$ [resp. $r(jj)$], $1 \le j \le h \le m$.

If we denote by σ_{jh} $(1 \le j \le h \le m)$ the proportion of unordered object pairs which possess the jth and the hth modalities of the qualitative variable q, respectively, we can give the expression of the standardized index of similarity between two given objects x and y:[9,11]

$$Q(x, y) = \frac{\sum \{\sigma_{jh}[r(q(x), q(y)) - r(jh)]/1 \le j \le h \le m\}}{\left\{\sum_{l \le k} \sigma_{lk}\left[\sum_{j \le h} \sigma_{jh}(r(lk) - r(jh))\right]^2\right\}^{1/2}} \tag{26}$$

This index will play the role of that in equation (1).

In our problem of classification of paraplegic patients, to each of the twelve variables considered above, we associate a preordonnance coding, where the rank of a given pair of modalities (affected vertebrae) is a strictly increasing function of the number of strictly intermediate vertebrae. This last number is *a priori* within the interval $[0, 28]$, but in fact, for statistical reasons, the number of classes of the observed preordonnance — for each of the twelve variables — is very limited with respect to the range of formal possibilities.

Interpretation of the results

The Table V gives the empirical distribution of levels for the 'global statistic' conceived from equation (15). It would be perfectly possible and of some interest to consider also that based on the preordonnance (formula (13)).

We will focus our attention on the last levels which give the most relevant partitions according to the behaviour of the 'global' and 'local' level statistics.

The most and very clearly significant classification is determined at the level 91, where $\Sigma_{91} = 37 \cdot 007$ and $\theta_{91} = 2 \cdot 499$. The consecutive level corresponds to an abrupt fall of Σ:

$$\theta_{92} = \Sigma_{92} - \Sigma_{91} = -7 \cdot 506$$

The significant node 91 results from the association of the two nodes 84 and 87 which, respectively, underly two subclasses that we consider below. Each significant node is marked by an asterisk in our tree representation (cf. Figures 2 and 3). Then, we will denote (*91) instead of 91.

Another distinctive partition is given at the 88th level, where $\Sigma_{88} = 35 \cdot 946$ and $\theta_{88} = 2 \cdot 143$. A fall of Σ is associated with the following level:

$$\theta_{89} = \Sigma_{89} - \Sigma_{88} = -2 \cdot 680$$

Then, we will briefly interpret the 91th level by taking into account the 88th level.

As we have just mentioned, the node (*91) determines a highly significant class. If we consider the interval of vertebrae going from D2 to L3, the starting of the lesion in this class occurs

Table V.

	Level	Local Statistic	Global Statistic
	1	1·847	1·847
	2	0·766	2·613
	3	0·569	3·183
	4	0·487	3·669
1. Maximum	5	1·174	4·844
	6	0·326	5·170
	7	0·310	5·480
	8	0·296	5·776
2. Maximum	9	0·554	6·330
	10	0·264	6·594
	11	0·249	6·843
	12	0·243	7·086
	13	0·238	7·324
	14	0·233	7·557
	15	0·229	7·786
	16	0·222	8·007
	17	0·215	8·222
	18	0·208	8·431
	19	0·202	8·633
	20	0·199	8·832
	21	0·194	9·026
	22	0·192	9·219
	23	0·188	9·407
	24	0·182	9·589
	25	0·179	9·768
	26	0·175	9·942
	27	0·171	10·114
	28	0·169	10·283
	29	0·168	10·451
	30	0·166	10·617
	31	0·165	10·782
	32	0·164	10·946
	33	0·160	11·106
	34	0·158	11·264
	35	0·155	11·420
	36	0·152	11·572
	37	0·147	11·719
	38	0·142	11·861
3. Maximum	39	0·143	12·004
	40	0·130	12·134
4. Maximum	41	0·221	12·355
	42	0·214	12·569
	43	0·213	12·782
	44	0·213	12·995
	45	0·100	13·094
	46	0·026	13·120
	47	0·164	13·285

Table Continued

Table V. *(Continued)*

	Level	*Local Statistic*	*Global Statistic*
5. Maximum	48	0·216	13·501
	49	0·203	13·703
6. Maximum	50	0·409	14·113
	51	0·372	14·484
	52	0·323	14·807
	53	0·100	14·907
	54	0·340	15·247
7. Maximum	55	0·376	15·623
	56	0·349	15·972
	57	0·341	16·313
8. Maximum	58	0·401	16·715
	59	0·099	16·814
9. Maximum	60	0·381	17·195
	61	0·370	17·565
	62	0·323	17·888
	63	0·350	18·238
	64	0·381	18·619
	65	0·407	19·026
10. Maximum	66	0·436	19·462
	67	0·403	19·864
	68	0·198	20·063
	69	0·415	20·478
11. Maximum	70	0·441	20·918
	71	0·408	21·327
12. Maximum	72	0·519	21·845
	73	0·514	22·360
	74	0·465	22·825
	75	0·426	23·251
13. Maximum	76	0·475	23·726
	77	0·211	23·937
	78	0·655	24·592
14. Maximum	79	0·991	25·583
	80	0·923	26·507
	81	0·495	27·002
	82	0·468	27·470
15. Maximum	83	1·204	28·674
	84	−0·152	28·522
	85	1·679	30·201
	86	1·791	31·992
	87	1·811	33·803
16. Maximum	88	2·143	35·946
	89	−2·680	33·265
	90	1·242	34·507
17. Maximum	91	2·499	37·007
	92	−7·506	29·501
18. Maximum	93	1·596	31·096
	94	−1·069	30·027
	95	−30·021	0·006

within the interval [D2, D7] or — except for the only individual 183 GREG — [D4, D7]. The lesion ends within the interval [D6, L3] and mostly between D8 and L3.

This class is obtained by aggregating two subclasses — mentioned above — and underlain by the nodes 84 and 87. For the second one, the starting and the end occur globally later than for the first one. Thus, for the first subclass (84) the beginning of the lesion is around D4 and D5, but for the second class, the beginning of the lesion is around D6. More precisely, to be exhaustive in our description and analysis we have — simply by counting — to determine on each subclass the empirical frequency distribution of each of the three groups of variables: 'sensory functions', 'motor functions' (starting of the lesion), 'restart sensory or motor functions'. The common scale is ordinal, the modalities of which go from C1, C2, ..., to S4, S5, − 1. As we saw above, most often the support of the distribution on a given class is an interval of little size.

We leave to the interested reader the determination and the histogram drawing of these distributions on each class of the 91th (or 88th) level partition.

We can note that the subclasses underlain by the nodes (84) and (87) are two classes of the 88th level partition. The node (77) achieves a class of both partitions (the 88th and the 91th). For this class, the lesion starts mostly in the cervical vertebrae (from C5 to C8) and, sometimes, at the beginning of the dorsal vertebrae (D1, D2 or D3). The end of the lesion occurs within the last half of the dorsal vertebrae or the first half of the lumbar vertebrae.

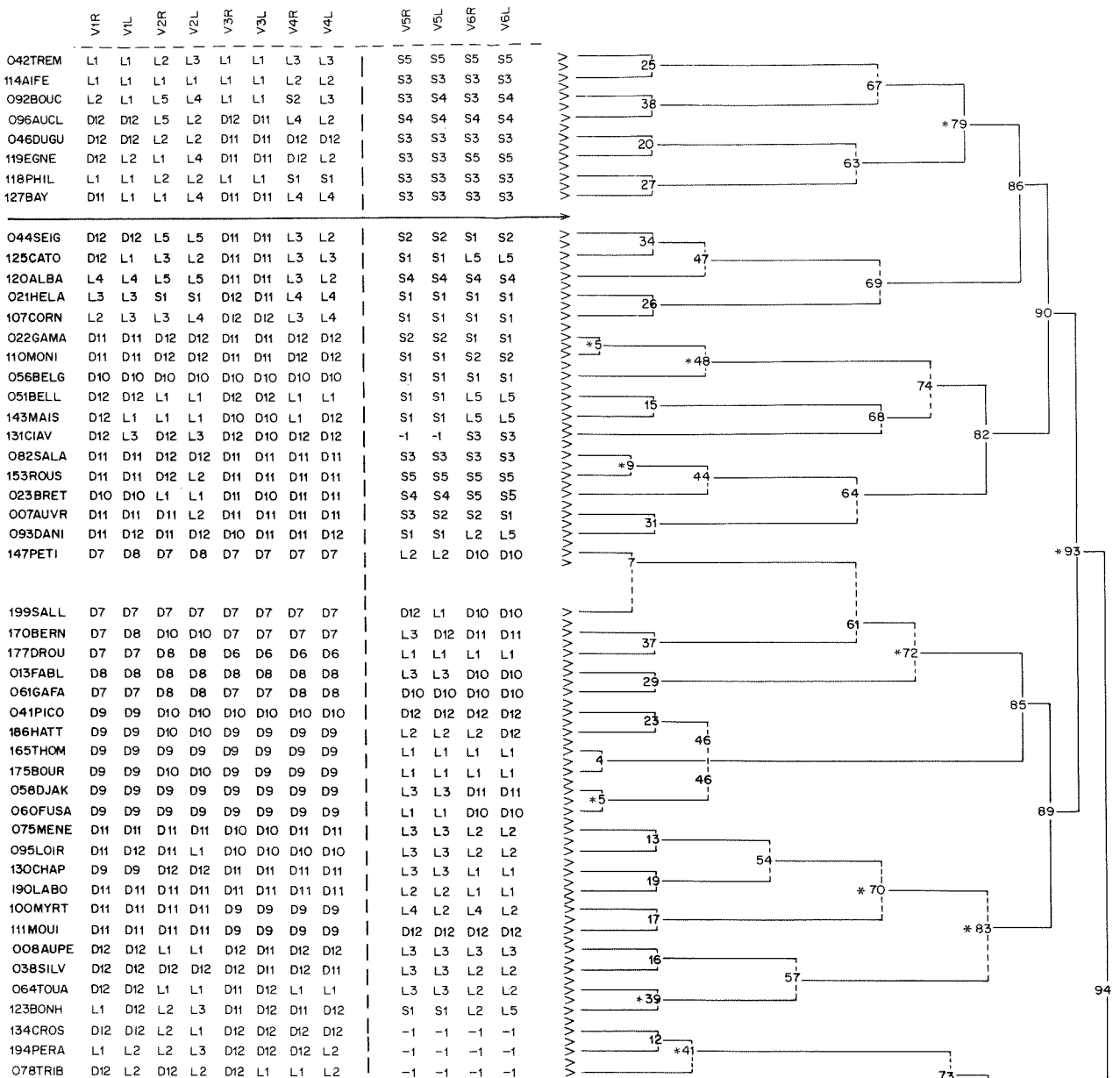
We can order the two classes (77) and (*91) or the three classes (77), (84) and (87), with respect to the position of the starting of the lesion: (77) < (*91) or (77) < (84) < (87). Henceforth, we will try to present the subsequent classes according to this order. On the other hand, and for simplicity reasons, we will denote a given class by the number of the node which ends it.

The class (89) which belongs to the partition of the 91th level is composed of two subclasses (85) and (*83) which are two classes of the 88th level partition.

For the class (85), the starting of the lesion occurs within the interval [D6, D7, D8, D9, D10] of the dorsal vertebrae, and for the other class (*83) which ends with a significant node, the beginning of the lesion starts later and mainly in D11 and D12. The end of the lesion for this last class (*83) has some tendency to occur later than for the first mentioned class (87), since, for this last (87) the end is around D10, D11, D12 and L1, whereas the end for (*83) is around D12, L1, L2, L3, reaching in some cases L4, L5 and even S1.

What characterizes very clearly the class (*88) — which ends with a very significant node — is the fact that all of the bottom of the spinal medulla is damaged. On the other hand, the lesion does not start before the last half of the dorsal vertebrae or — generally — after the lumbar vertebrae.

For a deeper analysis, it is now possible to follow step by step the building of each class roughly described above. Finally, we can give a brief look at the level 93 which corresponds to a significant node. At this last level, we distinguish three classes (*88), (92) and (*98) which can be ordered according to the position of the damaged section of the spinal medulla: (92) < (*93) < (*88). But the drop of the local statistic between the nodes (*91) and (92) suggest to us to retain the partition with four classes, where (92) is replaced by its two component classes (77) and (*91). Then the order considered on the four classes is: (77) < (*91) < (*93) < (*88).


V1R V1L V2R V2L V3R V3L V4R V4L V5R V5L V6R V6L
042TREM L1 L1 L2 L3 L1 L1 L3 L3 S5 S5 S5 S5
114AIFE L1 L1 L1 L1 L1 L1 L2 L2 S3 S3 S3 S3
092BOUC L2 L1 L5 L4 L1 L1 S2 L3 S3 S4 S3 S4
096AUCL D12 D12 L5 L2 D12 D11 L4 L2 S4 S4 S4 S4
046DUGU D12 D12 L2 L2 D11 D11 D12 D12 S3 S3 S3 S3
119EGNE D12 L2 L1 L4 D11 D11 D12 L2 S3 S3 S5 S5
118PHIL L1 L1 L2 L2 L1 L1 S1 S1 S3 S3 S3 S3
127BAY D11 L1 L1 L4 D11 D11 L4 L4 S3 S3 S3 S3
044SEIG D12 D12 L5 L5 D11 D11 L3 L2 S2 S2 S1 S2
125CATO D12 L1 L3 L2 D11 D11 L3 L3 S1 S1 L5 L5
120ALBA L4 L4 L5 L5 D11 D11 L3 L2 S4 S4 S4 S4
021HELA L3 L3 S1 S1 D12 D11 L4 L4 S1 S1 S1 S1
107CORN L2 L3 L3 L4 D12 D12 L3 L4 S1 S1 S1 S1
022GAMA D11 D11 D12 D12 D11 D11 D12 D12 S2 S2 S1 S1
110MONI D11 D11 D12 D12 D11 D11 D12 D12 S1 S1 S2 S2
056BELG D10 D10 D10 D10 D10 D10 D10 D10 S1 S1 S1 S1
051BELL D12 D12 L1 L1 D12 D12 L1 L1 S1 S1 L5 L5
143MAIS D12 L1 L1 L1 D10 D10 L1 D12 S1 S1 L5 L5
131CIAV D12 L3 D12 L3 D12 D10 D12 D12 -1 -1 S3 S3
082SALA D11 D11 D12 D12 D11 D11 D11 D11 S3 S3 S3 S3
153ROUS D11 D11 D12 L2 D11 D11 D11 D11 S5 S5 S5 S5
023BRET D10 D10 L1 L1 D11 D10 D11 D11 S4 S4 S5 S5
007AUVR D11 D11 D11 L2 D11 D11 D11 D11 S3 S2 S2 S1
093DANI D11 D12 D11 D12 D10 D11 D11 D12 S1 S1 L2 L5
147PETI D7 D8 D7 D8 D7 D7 D7 D7 L2 L2 D10 D10
199SALL D7 D7 D7 D7 D7 D7 D7 D7 D12 L1 D10 D10
170BERN D7 D8 D10 D10 D7 D7 D7 D7 L3 D12 D11 D11
177DROU D7 D7 D8 D8 D6 D6 D6 D6 L1 L1 L1 L1
013FABL D8 D8 D8 D8 D8 D8 D8 D8 L3 L3 D10 D10
061GAFA D7 D7 D8 D8 D7 D7 D8 D8 D10 D10 D10 D10
041PICO D9 D9 D10 D10 D10 D10 D10 D10 D12 D12 D12 D12
186HATT D9 D9 D10 D10 D9 D9 D9 D9 L2 L2 L2 D12
165THOM D9 D9 D9 D9 D9 D9 D9 D9 L1 L1 L1 L1
175BOUR D9 D9 D10 D10 D9 D9 D9 D9 L1 L1 L1 L1
058DJAK D9 D9 D9 D9 D9 D9 D9 D9 L3 L3 D11 D11
060FUSA D9 D9 D9 D9 D9 D9 D9 D9 L1 L1 D10 D10
075MENE D11 D11 D11 D11 D10 D10 D11 D11 L3 L3 L2 L2
095LOIR D11 D12 D11 L1 D10 D10 D10 D10 L3 L3 L2 L2
130CHAP D9 D9 D12 D12 D11 D11 D11 D11 L3 L3 L1 L1
190LABO D11 D11 D11 D11 D11 D11 D11 D11 L2 L2 L1 L1
100MYRT D11 D11 D11 D11 D9 D9 D9 D9 L4 L2 L4 L2
111MOUI D11 D11 D11 D11 D9 D9 D9 D9 D12 D12 D12 D12
008AUPE D12 D12 L1 L1 D12 D11 D12 D12 L3 L3 L3 L3
038SILV D12 D12 D12 D12 D12 D11 D12 D11 L3 L3 L2 L2
064TOUA D12 D12 L1 L1 D11 D12 L1 L1 L3 L3 L2 L2
123BONH L1 D12 L2 L3 D11 D12 D11 D12 S1 S1 L2 L5
134CROS D12 D12 L2 L1 D12 D12 D12 D12 -1 -1 -1 -1
194PERA L1 L2 L2 L3 D12 D12 D12 L2 -1 -1 -1 -1
078TRIB D12 L2 D12 L2 D12 L1 L1 L2 -1 -1 -1 -1
25
38
67
20
63
27
*79
86
34
47
69
26
90
*5
*48
74
15
68
82
*9
44
64
31
7
*93
61
37
*72
29
85
23
46
4
46
*5
89
13
54
19
*70
17
*83
16
57
*39
94
12
*41
73

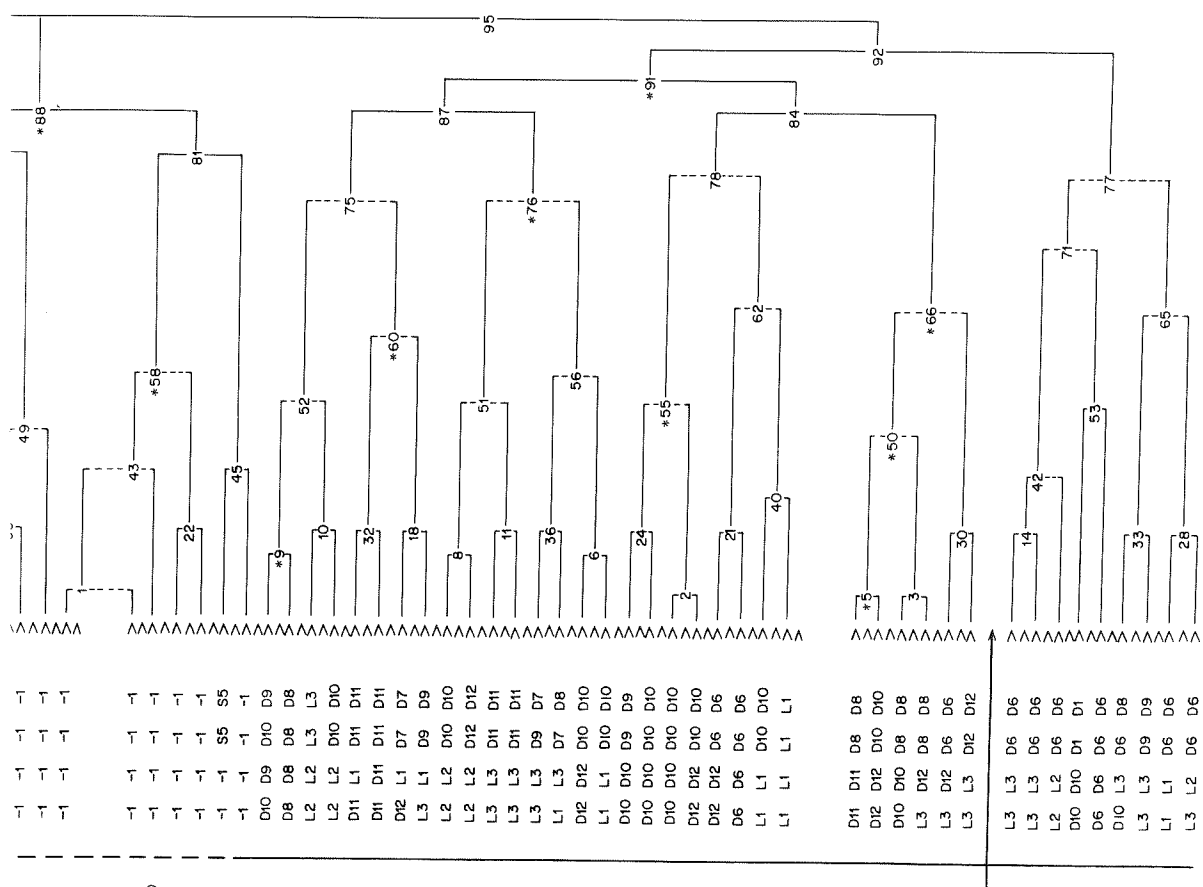


Figure 3.

ACKNOWLEDGEMENT

We are indebted to Philippe Peter for technical assistance and large collaboration at the level of programming and interpretation of the data.

REFERENCES

1. P. Parisot, 'Application of similarity aggregation techniques to a popuation of paraplegic patients', *Applied Stochastic Models and Data Analysis,* **1**, 35–54 (1985).
2. I. C. Lerman, '*Les Bases de la Classification Automatique*', Gauthier Villars, Paris, 1970.
3. I. C. Lerman, 'Sur l'analyse des données préalable à une classification autmatique. Proposition d'une nouvelle mesure de similarité', *Rev. Math. et Sc. Hum.*, No 32, Paris, 1970, pp. 5–15.
4. I. C. Lerman, 'Combinatorial analysis in the statistical treatment of behavioral data', *Quality and Quantity,* **14** 431–469 (1980).
5. I. C. Lerman *Classification et Analyse Ordinale des Données,* Dunod, Paris, 1981.
6. I. C. Lerman, 'Sur la signification des classes issues d'une classification automatique', in J. Felsenstein (ed), *Numerical Taxonomy*, NATO Advanced Studies, Springer, 1983.
7. I. C. Lerman, 'Analyse classificatoire d'une correspondance multiple, typologie et régression', in E. Diday, M. Jambu, L. Lebert, J. Pagès and R. Tomassone (eds) *Data analysis and Informatics*, North Holland, 1984, pp. 193, 212.
8. I. C. Lerman, 'Justification et validité statistique d'une échelle [0, 1] de fréquence mathématique, pour une structure de proximité sur un ensemble de variables observées', *Publ. Inst. Stat. Univ. de Paris, XXIX, fasc. 3–4,* 1984, pp. 27–57.
9. I. C. Lerman and Ph. Peter, 'Elaboration et logiciel d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application à la recherche d'un consensus en classification', *IRISA, Publ. Int. no. 262,* Rennes, 1985.
10. I. C. Lerman, 'Classifications et informatique', *Rev. Informatique et Sc. Humaines,* Paris, no. 66, Septembre 1985.
11. I. C. Lerman and Ph. Peter, 'Organisation et consultation d'une banque de 'petites annonces' à partir d'une méthode de classification hiérarchique en parallèle', *Quatrièmes Journées Internationales Analyse des Données et Informatique*, Versailles 9–11 October 1985.
12. S. Chah, 'Agrégation des préordonnances', *Etude F-063*, Centre Scientifique IBM, Paris, 1984.
13. S. Chah, 'Critères de classification sur des données hētērogènes' *Quatrièmes Journées Internationales Analyse des Données et Informatique*, Versailles 9–11 October 1985.