Statistica Applicata Vol. 10, n. 1, 1998

INTERPOINT DISTANCES IN THE MULTIVARIATE DATA ORDERING

Sergio Scippacercola

Dipartimento di Matematica e Statistica, Università "Federico II" di Napoli.

Summary

Aim of the present paper is to propose a new method to ordering a set of n statistical units described by k variables. By a geometrical point of view, any statistical unit is associated to a point in the space of k (k >1) dimensionality. The object of this paper is the ordering of the n statistical units by interpoint ultrametric distances defined in the Minimum Spanning Tree (MST). In the MST, among all possible paths, we search the M path of Maximum length. This M path is the basis to define an ordering function. The ordering criterion is found on the similarity relations of the contiguous points in the MST. We suggest the local criteria to graduate the statistical units not in the M path. Also, to evaluate the quality of the obtained ordering, we use some similarity measures. The method is applied with success to sets of multidimensional data: one of these sets is in a natural ordering.

Key Words: Multivariate Data, Order Statistics, Interpoint Distances, Tree.

1. INTRODUCTION

The ordering of a multidimensional data set is a problem which arises in a lot of researches. Aim of the present paper is to introduce a new method for the ordering of n statistical units by interpoint ultrametric distances derived from the Minimum Spanning Tree (MST).

Preliminarily, we present the principle of the MST and a brief review of the sub–ordering principles (Section 2). In Section 3 are shortly reminded the basic ideas to be used for ordering multidimensional data by interpoint distances. We single out that a Path of Maximum length in the Minimum Spanning Tree (MST) suggests an ordering criterion (global criterion) and we define a function (Section 4) to obtain a sub–ordering of n "k–dimensional" points. Section 5 presents some local criteria to get an order of the points located in the edges of MST according to a Path of Maximum length. In Section 6 we evaluate the similarity between the original point configuration in the k–dimensional space and the correspondent sub–ordering derived according to a proposed method. Finally, (Section 7) the method is applied on real cases of multidimensional data.

2. THE ORDERING OF MULTIVARIATE DATA

Let $\mathbf{X}_{n,k}$ be the data matrix. Let $\Im$ be the set of the n statistical units ($\boldsymbol{p}_i \in \Im$, $i = 1,2,\ldots,n$), and let J, ($\mathbf{x}_j \in$ J, $j = 1,2,\ldots,k$), be the set of k–dimensional quantitative centred variables. From a geometrical point of view, to each $\boldsymbol{p}_i$ statistical unit corresponds a point in the $\Re_k$ sub–space ($k > 1$). Let $\mathbf{D} = \{d_{ij}\}$ ($\mathbf{D} \subseteq \Re_1$) be the interpoint distance matrix , based on a given metric, of the n points.

The aim of the Minimum Spanning Tree (Kruskal, 1956; Prim, 1957; Gower, Ross, 1969) is to search in **D** the minimum distances for some pairs of points. The spanning tree is any set of straight–line segments joining pairs of points so that (Dunn, Everitt, 1982) no closed loops occur, each point is visited by at least one line and the tree is connected. The tree is a set **A** of non empty parts such that it satisfies the relation of inclusion and such that the predecessor vertices of each part of **A** are in total ordering (Chandon, Pinson, 1981). The MST is a partial graph: each vertex is associated to a point. Any similarity or dissimilarity is equivalent to a complete weighted graph: the weights associated with edges are given by the similarities or the dissimilarities (Murtagh, 1993).

The ordering problem is to search a criterion to build an ordered series $\boldsymbol{q}_i$ of the n k–dimensional $\boldsymbol{p}_i$ points such that:

$$\boldsymbol{q}_i \rightarrow \boldsymbol{q}_{i+1} \qquad (i = 1,2,\ldots,n-1) \tag{1}$$

where the arrow is read as "precedes". Let $\boldsymbol{q}_i$, $\boldsymbol{q}_j$, $\boldsymbol{q}_k$ be three elements of the series. The series is ordered if it satisfies the asymmetry and transitivity properties (Crescimanni, 1979):

– $\boldsymbol{q}_i \rightarrow \boldsymbol{q}_j$ or $\boldsymbol{q}_j \rightarrow \boldsymbol{q}_i$ (2)

– $\boldsymbol{q}_i \rightarrow \boldsymbol{q}_j$ and $\boldsymbol{q}_j \rightarrow \boldsymbol{q}_k \Rightarrow \boldsymbol{q}_i \rightarrow \boldsymbol{q}_k$. (3)

The ordering is clear and unambiguous only if the points are on a straight line (in $\Re_1$ subspace) or if the points in $\Re_k$ have a reference to a time series.

For multidimensional points a total ordering is impossible to reach and our interest is restricted to the forms of sub–ordering that Barnett (1976) distinguishes in four principles: *marginal* ordering, *reduced* (aggregate) ordering, *partial* ordering and *conditional* (sequential) ordering. These principles are shared and not mutually exclusive.

The aim of the *marginal* ordering is that of ordering the n k–dimensional points by the ranked values of one of the k–variables. By this way we substitute the original ordering with the one of the ranked values. Obviously the marginal ordering may interest any of the k variables.

In the *reduced* ordering each point is reduced to a single value by means of some combination of the k component values. The intention of the *reduced* ordering

is to synthetically express overall characteristics by some sort of restricted overall ordering for the data set. In this type also, the accumulated distances of each point from all other points are considered.

The principle of the *partial* ordering is to locate the sample space regions including points with the same characteristics with regard to order or rank or extremeness.

The *conditional* sub–ordering is an ordering on a marginal set of points conditional on selection or ordering or ranking as regard to other blocks of the same observations. Marginal samples can be the same original data or the data derived by coordinate transformations.

Particularly, the method proposed in this paper can be classified as *reduced* sub–ordering since it uses interpoint distances to generate a pre–ordering function.

3. THE ORDERING BY INTERPOINT DISTANCES

A proximity index on the set of the pairs of the statistical units makes it possible to associate an ordering to $\boldsymbol{p}_i$ $(i = 1, ..., n)$ statistical units (Chandon, Pinson, 1981).

For ordering the points we consider very interesting the reciprocal relations of the points in the $\Re_k$ space. It is just that set of reciprocal relations between the points which gives useful information for ordering the n statistical units.

Indeed, the application of the MST on the $\mathbf{D}$ set of distances produces a $\mathbf{D}' = \{d'_{ij}\}$ subset ($\mathbf{D}' \subseteq \mathbf{D}$) such that:

$$\forall i, j, k \in \Im \Rightarrow d'_{ij} \leq max\left(d'_{ik}, d'_{jk}\right) \tag{4}$$

The set $\mathbf{D}'$ satisfies the ultra metric axioms. Generally, a system of distances does not define a total ordering on $\Im$ while it defines a pre–ordering since two distances can be equal but the extremes of these distances are not coincident. In this case the anti–symmetric property:

$$\forall a, b \in \Im \;\; \mathrm{R}(a,b) \text{ and } \mathrm{R}(b,a) \Rightarrow (a = b) \tag{5}$$

is not verified.

The succession of $\mathbf{D}'$ segments is a monotone non decreasing succession and it induces a pre–ordering on $\Im$. This set of segments allows to build the graph (MST). In the MST, for each pair of vertices, the conditions of symmetry and transitivity are satisfied.

In the MST, from any point $\boldsymbol{p}_a$ to any point $\boldsymbol{p}_b$, there exists only one path. For non contiguous $\boldsymbol{p}_a$ and $\boldsymbol{p}_b$ points, we can single out, in an ordered way, all

intercurrent points placed along the path joining $\boldsymbol{p}_a$ to $\boldsymbol{p}_b$. This suggests an ordering criterion based on the similarity of the contiguous points in the MST.

4. AN ORDERING FUNCTION

According to the previous idea, we propose to order the points with reference to the Maximum Path in the MST and to one of its extreme points. In the Minimum Spanning Tree, among all possible paths, we search the M Path of Maximum length. Let $\boldsymbol{p}_i$ be a generic point on the MST and let $\boldsymbol{p}_0$ and $\boldsymbol{p}_z$ be extreme vertices of the M Path. To compute the distance, via MST, from $\boldsymbol{p}_0$ to a point $\boldsymbol{p}_m$, we add the lengths of the m consecutive segments d' starting from $\boldsymbol{p}_0$ to reach $\boldsymbol{p}_m$ via MST. Therefore, we define a function g:

$$g(\boldsymbol{p}_0, \boldsymbol{p}_m) = \sum_{i=0}^{m-1} d'(\boldsymbol{p}_i, \boldsymbol{p}_{i+1}) \tag{6}$$

The (6) is a monotone transformation (Chandon, Pinson, 1981) because:

$$\forall i,j,k,l \in \Im \;\; if \;\; d'(i,j) \le d'(k,l) \Rightarrow g(i,j) \le g(k,l) \tag{7}$$

Furthermore, the function g, so defined, satisfies the asymmetry and transitivity conditions (2) and (3). We deduce that the function (6) orders all points on $\Re_1$ from $\boldsymbol{p}_0$. By (6) we obtain a linearisation of the MST: according to a distance from $\boldsymbol{p}_0$, via MST, any point is reported on a straight line equal to a Maximum Path. Thus, the lateral edges of the MST having a vertex on M are squashed on the same straight line. Therefore, we can consider the Maximum Path as the load bearing structure (or *skeleton*) of the original point configuration.

5. THE LOCAL CRITERIA

The function (6) orders all points from extreme $\boldsymbol{p}_0$ of M. If, instead of $\boldsymbol{p}_0$, we consider an ordering from extreme $\boldsymbol{p}_z$ of M, we obtain an ordering generally different from the previous one in presence of points on the aside edges of M. In order to avoid this ambiguity, the method develops two steps. Let $\boldsymbol{p}_k$ be a vertex on M and let $\boldsymbol{p}_h$ be a point not on M. In the first step we order only the points which are vertices of M. The points in M are ordered by function (6) (*global criterion*). In the successive step we order each point $\boldsymbol{p}_h$ not in M. Each vertex $\boldsymbol{p}_h$, on aside edge, connected to a vertex $\boldsymbol{p}_k$ on M is an element of the cluster $\boldsymbol{p}_k$.

We propose two *local* alternative criteria for ordering the vertices on the aside edges:

1. Each vertex $\boldsymbol{p}_h$ not in M is graduated with the same value of the $\boldsymbol{p}_k$ vertex on M:

$$g(\boldsymbol{p}_0, \boldsymbol{p}_h) = g(\boldsymbol{p}_0, \boldsymbol{p}_k) \tag{8}$$

In this case each aside edge with all vertices can be visualised by a local graduation from vertex $\boldsymbol{p}_k$.

2. The $\boldsymbol{p}_h$ point is at the ultra metric distance $d'(\boldsymbol{p}_h,\boldsymbol{p}_0)$ from $\boldsymbol{p}_0$ and it is at distance $d'(\boldsymbol{p}_h,\boldsymbol{p}_z)$ from $\boldsymbol{p}_z$. The point $\boldsymbol{p}_h$ is in ratio $d'(\boldsymbol{p}_h,\boldsymbol{p}_0)/d'(\boldsymbol{p}_h,\boldsymbol{p}_z)$ from $\boldsymbol{p}_0$ and $\boldsymbol{p}_z$ in the MST. In order to preserve this ratio for ordering on M, we propose the following graduation:

$$g(\boldsymbol{p}_0, \boldsymbol{p}_h) = \frac{d'(\boldsymbol{p}_0, \boldsymbol{p}_z) d'(\boldsymbol{p}_0, \boldsymbol{p}_h)}{d'(\boldsymbol{p}_0, \boldsymbol{p}_h) + d'(\boldsymbol{p}_h, \boldsymbol{p}_z)} \tag{9}$$

In this criterion the point $\boldsymbol{p}_h$ is graduated on M at a distance, from $\boldsymbol{p}_0$ and $\boldsymbol{p}_z$, proportional to the ratio in the MST. The (9) preserves the original ratio in the ultrametric distance.

6. INDEXES FOR THE ORDERING EVALUATION

In order to evaluate the quality of the obtained ordering, we emphasize that the ordering produces a new structure of points. Let $\boldsymbol{G}_{n,1}$ be the vector containing the distances of the n points $\boldsymbol{p}_i$ from $\boldsymbol{p}_0$ via MST. We can use a measure of the degree to which the two configurations (**X** and **G**) resemble one another, after centring the **X** and **G** matrices. We suggest the following indexes:

1. *RV coefficient* (Robert, Escoufier, 1976)

$$I_E = \frac{tr[(\mathbf{XX}')(\mathbf{GG}')]}{\left[tr(\mathbf{XX}')^2 tr(\mathbf{GG}')^2\right]^{1/2}} \quad (0 \le I_E \le 1) \tag{10}$$

where the n points are disposed in **X** and in **G** with the same order since any permutation of rows unbiases the I_E index. The coefficient measures the similarity between the relative positions of the n points in the subspaces spanned by the columns of **X** and **G**.

2. The *procrustean index* (Mardia *et al.*, 1979; Gordon, 1980) properly normalised:

$$I_p = \frac{2tr(\Gamma^{1/2})}{[tr(\mathbf{XX}') + tr(\mathbf{GG}')]} \quad (0 \le I_p \le 1) \tag{11}$$

where Γ is the eigenvalues matrix of (**X'GG'X**).

In our context, differently from I_E, I_p measures the similarity between two configurations in presence of roto–translation and reflection.

3. The *congruence coefficient* (Borg, Lingoes, 1987):

$$I_c = \frac{\sum_{i=1}^{k} d_{iX} d_{iG}}{\left(\sum_{i=1}^{k} d_{iX}^2 \sum_{i=1}^{k} d_{iG}^2 \right)^{1/2}} \qquad (0 \le I_c \le 1), \tag{12}$$

where d_{iX} is the i–th distance of **X** and d_{iG} is the i–th distance of **G**, and $m = (n)\ (n-1)/2$. This coefficient measures the similarity for the pair of configurations (**X** and **G**) in terms of interpoint distances.

If the indices are close to zero the similarity is low, if they are close to one the similarity is high. The previous measures perform a comparison of two configurations (**X** and **G**) of different dimensionality. These measures lead some information on the quality of the ordering because each coefficient condenses a great deal of different information on the similarity of two configurations.

7. APPLICATIONS

In this Section, we apply the proposed method to order some real data sets of multidimensional points that are elements of an euclidean space. First, we consider a set of points having a natural ordering to verify the validity of the proposed method. In particular, we use the data of an experiment of G. Ekman on colour perception regarding the responses of 31 subjects on "qualitative similarity" of each pair of 14 colours with wavelengths ranging from 434 nm to 674 nm. From dissimilarities matrix (Tab. I), derived from correspondent similarity matrix in Borg, Lingoes (1987; p. 59), it is easy to note an almost perfect rule of the "among" dissimilarities between columns perceived by respondants according to the influence in their wavelengths. In this case, we apply the method starting directly from the euclidean distance matrix (Tab. I).

In Fig. 1a this natural order is shown. By the proposed method, the 14 points are perfectly ordered along the Maximum Path in terms of their wavelengths (Fig. 1b). In Fig. 1 the wavelengths are in bold character and the distances from one of the point extreme of the Maximum Path are in normal character. The variations, as regards to the natural order, depend exclusively on the noise in the experimental data. In the Maximum Path of MST are not present aside edges, so we apply only the global criterion defined in (6). This result is confirmed by similarity measures computed (I_E =0.98, I_p = 0.96 and I_c =0.99) between the original configuration and the correspondent ordering obtained by the global criterion (6).

Tab. I: Dissimilarities of colours for wavelengths ranging from 434 nm to 674 nm (Borg, Lingoes, 1987; pag. 59).

	434	*445*	*465*	*472*	*490*	*504*	*537*	*555*	*584*	*600*	*610*	*628*	*651*	*674*
434	.0	.14	.58	.58	.82	.94	.93	.96	.98	.93	.91	.88	.87	.84
445	.14	.0	.50	.56	.78	.91	.93	.93	.98	.96	.93	.89	.87	.86
465	.58	.50	.0	.19	.53	.83	.90	.92	.98	.99	.98	.99	.95	.97
472	.58	.56	.19	.0	.46	.75	.90	.91	.98	.99	1.	.99	.98	.96
490	.82	.78	.53	.46	.0	.39	.69	.74	.93	.98	.98	.99	.98	1.
504	.94	.91	.83	.75	.39	.0	.38	.55	.86	.92	.98	.98	.98	.99
537	.93	.93	.90	.90	.69	.38	.0	.27	.78	.86	.95	.98	.98	1.
555	.96	.93	.92	.91	.74	.55	.27	.0	.67	.81	.96	.97	.98	.98
584	.98	.98	.98	.98	.93	.86	.78	.67	.0	.42	.63	.73	.80	.77
600	.93	.96	.99	.99	.98	.92	.86	.81	.42	.0	.26	.50	.59	.72
610	.91	.93	.98	1.	.98	.98	.95	.96	.63	.26	.0	.24	.38	.45
628	.88	.89	.99	.99	.99	.98	.98	.97	.73	.50	.24	.0	.15	.32
651	.87	.87	.95	.98	.98	.98	.98	.98	.80	.59	.38	.15	.0	.24
674	.84	.86	.97	.96	1.	.99	1.	.98	.77	.72	.45	.32	.24	.0

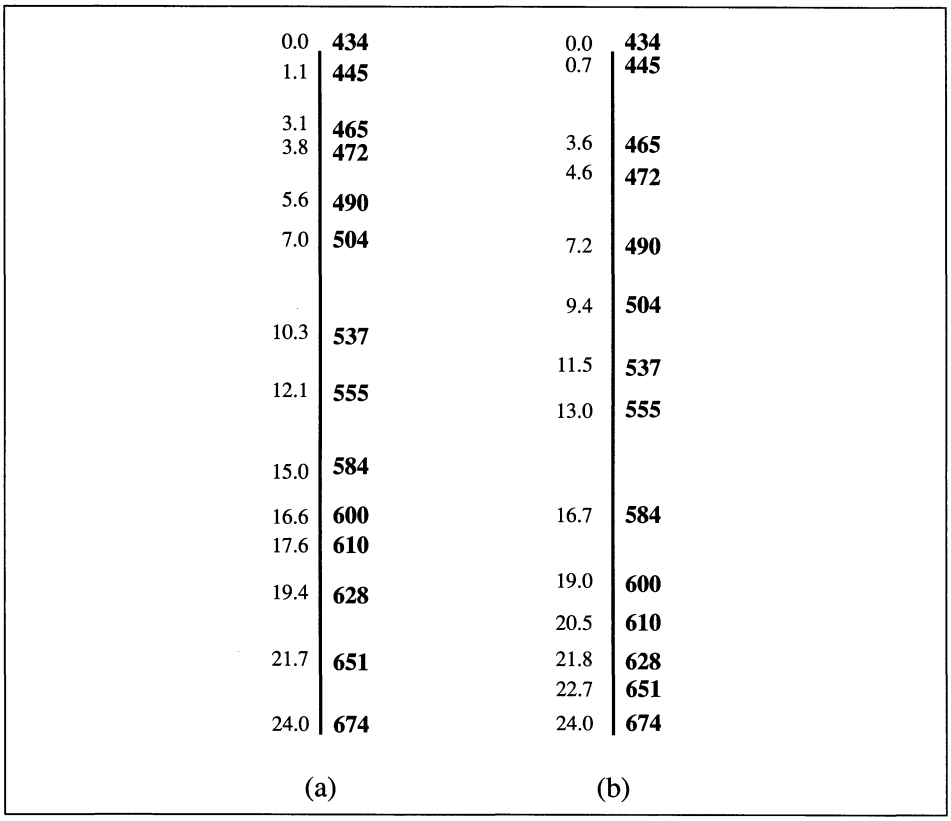

Fig. 1: (a) The natural ordering of the colours by wavelengths – (b) The ordering by the Maximum Path.

The second example concerns a set of bidimensional data showing the relationship between percentage of unemployed and education level (% graduates) in the districts of Naples in 1989 (Regione Campania, Yearbook 1990). In Fig. 2, are presented both the MST (left side) and the Maximum Path (right side) on the same data. In this example an aside is evidentiated along the Maximum Path going from Chiaia (vertex 2) to Montecalvario (vertex 4).

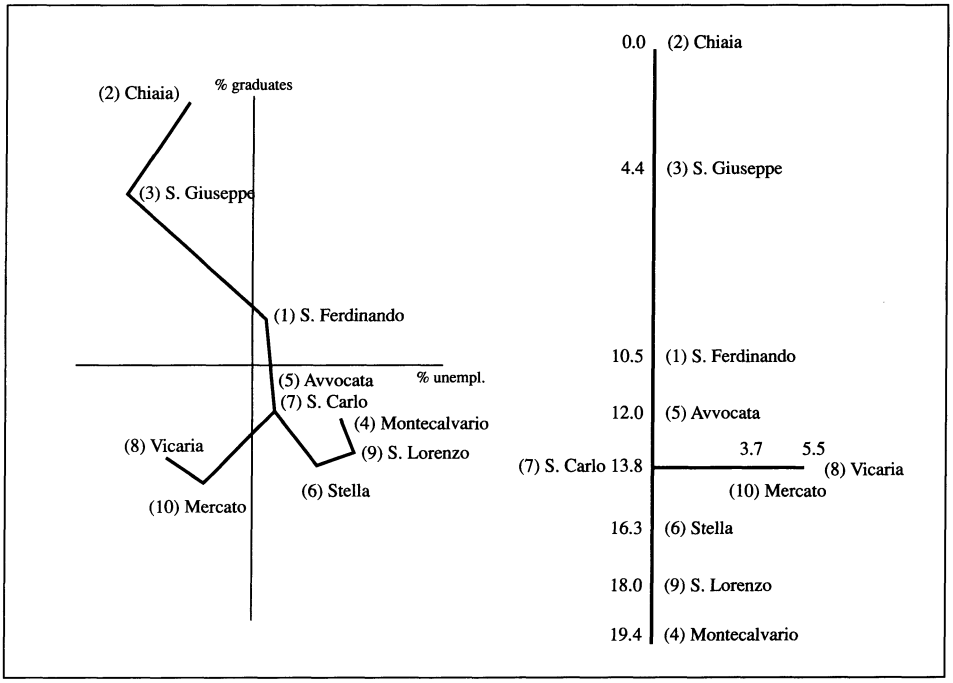

Fig. 2: MST and the Maximum Path: unemployement and graduation on ten Districts of Naples.

The linearisation of the tree (Fig. 3 – left side) produced according to global criterion (6) allows to locate the vertices not in the Maximum Path (*skeleton*). Also, in Fig. 3 (in the middle) is applied the local criterion (8) and, on right side, the local criterion (9). The vertices ordered by global or local criteria are in bold character.

It is possible to produce a detailed order of the vertices on the reduced edges. In this application, the measures of similarity (I_E =0.94, I_p = 0.74 and I_c =0.92) for global criterion (6) suggest to prefer the first order (Fig.3 – left side).

0.0	(2) Chiaia	0.0	(2) Chiaia	0.0	(2) Chiaia
4.4	(3) S. Giuseppe	4.4	(3) S. Giuseppe	4.4	(3) S. Giuseppe
10.5	(1) S. Ferdinando	10.5	(1) S. Ferdinando	10.5	(1) S. Ferdinando
				11.2	**(10) Mercato**
12.0	(5) Avvocata	12.0	(5) Avvocata	12.0	(5) Avvocata
				12.5	**(8) Vicaria**
13.8	(7) S. Carlo	13.8	(7) S. Carlo **(10) Mercato (8) Vicaria**	13.8	(7) S. Carlo
16.3	(6) Stella	16.3	(6) Stella	16.3	(6) Stella
17.5	**(10) Mercato**				
18.0	(9) S. Lorenzo	18.0	(9)S. Lorenzo	18.0	(9) S. Lorenzo
19.3	**(8) Vicaria**				
19.4	(4) Montecalvario	19.4	(4) Montecalvario	19.4	(4) Montecalvario

Fig. 3: Orderings by the global and local criteria of the ten referred Districts (unemployement and graduation data).

8. CONCLUSION

The interpoint distances and the MST have themselves a pre–ordering criterion. In the MST, from any point to any other point, there exists only one path. For non contiguous points, we can single out, in an ordered way, all intercurrent points placed along the path joining the points. This suggests an ordering criterion based on the similarity of the contiguous points in the MST and with reference to a Maximum path.

The Maximum Path shows the hidden basic structure of the configuration of points. The proposed function permits the linearisation of the Maximum Path and so we reach an ordering of the points.The Maximum Path of the MST proves to be an useful tool to order multidimensional data. The results of the applications confirm the validity of the method.

ACKNOWLEDGEMENTS

The research for this paper has been supported by a MURST (60%, 1995) grant (Resp. Prof. N. C. Lauro).

This work was presented, in a preliminary version, at the meeting "Giornate di Analisi di dati Multidimensionali", 30–31 Oct. 95, Napoli.

REFERENCES

BORG I., LINGOES J., (1987), *Multidimensional Similarity Structure Analysis,* Springer–Verlag, 58–64.

BARNETT V., (1976), The ordering of Multivariate Data, *Journal of Royal Statistical Society,* **139**, B, 318–354.

CHANDON J.L., PINSON S., (1981), *Analyse typologique, Théories et applications,* Ed. Masson, Paris.

CRESCIMANNI A., (1979), Oscillazione ed autocorrelazione a più dimensioni, *Metron,* XLIV, 1–4, 179–194.

DUNN G., EVERITT B.S., (1982), *An introduction to mathematical taxonomy,* Cambr. Univ. Press, Cambridge.

GORDON A.D., (1980), *Classification,* Chapman and Hall, New York.

GOWER J.C., ROSS J.S., (1969), Minimum Spanning Trees and Single Linkage Cluster Analysis, *Applied Statistics,* **18**, 54–64.

ISTAT, (1993), *Le Regioni in cifre,* System Data Comp., Tivoli, Roma.

KRUSKAL J.B., (1956), On the shortest spanning subtree of a graph and the travelling salesman problem, *Pro. of the American Math. Soc.,* **7**, 48–50.

MARDIA K.V., KENT J.T., BIBBY J.M., (1979), *Multivariate Analysis,* Ac. Press, London.

MURTAGH F.D., (1993), "Cluster Analysis Using Proximities", in: Van Mechelen I. *et al.* eds., *Categories and Concepts,* Ac. Press, London.

PRIM R.C., (1957), Shortest Connection network and some generalizations, *Bell System Tech. Journal,* **36**, 1389–1401.

REGIONE CAMPANIA, (1990), *Annuario Statistico Campano,* Jannone, Fisciano (SA).

ROBERT P., ESCOUFIER Y., (1976), A unifying tool for linear multivariate statistical methods; The RV coefficient, *Applied Statistics,* **25**.

LE DISTANZE INTER–PUNTI NELL'ORDINAMENTO DI DATI MULTIVARIATI

Riassunto

Viene proposta una metodologia per ordinare un insieme di n unità statistiche (u.s.) descritte da k variabili. Da un punto di vista geometrico, ad ogni u.s. è associata, nello spazio k–dimensionale, un punto. Obiettivo del presente lavoro è l' ordinamento delle n u.s. mediante le distanze ultrametriche inter–punti desunte dal Minimum Spanning Tree (MST).

La metodologia considera, tra tutti i cammini possibili nel MST, il cammino M di lunghezza massima. Tale cammino è usato come riferimento per definire un criterio di ordinamento basato sulle relazioni di similarità dei punti contigui nel MST. Inoltre vengono proposti alcuni criteri da utilizzare quando non tutte le unità si trovano su M. Per valutare la qualità dell'ordinamento ottenuto, si fa riferimento ad alcune misure di similarità. Infine, la metodologia è applicata a due insiemi di dati multivariati: uno di questi possiede un ordine naturale.