Journal of Classification 3:187-224 (1986)

Journal of
Classification

On the Use of Ordered Sets in Problems of Comparison and Consensus of Classifications

Jean-Pierre Barthélemy
Ecole Nationale Supérieure des Télécommunications

Bruno Leclerc
Centre d'Analyse et de Mathématique Sociales

Bernard Monjardet
Centre d'Analyse et de Mathématique Sociales

Abstract: Ordered set theory provides efficient tools for the problems of comparison and consensus of classifications Here, an overview of results obtained by the ordinal approach is presented Latticial or semilatticial structures of the main sets of classification models are described Many results on partitions are adaptable to dendrograms; many results on n-trees hold in any median semilattice and thus have counterparts on ordered trees and Buneman (phylogenetic) trees For the comparison of classifications, the semimodularity of the ordinal structures involved yields computable least-move metrics based on weighted or unweighted elementary transformations In the unweighted case, these metrics have simple characteristic properties For the consensus of classifications, the constructive, axiomatic, and optimization approaches are considered Natural consensus rules (majoritary, oligarchic,) have adequate ordinal formalizations A unified presentation of Arrow-like characterization results is given In the cases of n-trees, ordered trees and Buneman trees, the majority rule is a significant example where the three approaches converge

Résumé: La théorie des ensembles ordonnés fournit des outils utiles pour les problèmes de comparaison et de consensus de classifications Nous présentons une revue des résultats obtenus grâce à l'approche ordinale Les principaux ensembles de modèles de classifications possèdent des structures de treillis ou de demi-treillis, qui sont décrites Le fait que bien des résultats sur les partitions s'adaptent aux hiérarchies indicées provient de la proximité de leurs structures latticielles; de même, des résultats sur les hiérarchies, portant en fait sur les demi-treillis à médianes, ont des équivalents pour les hiérarchies stratifiées et les arbres phylogénétiques de Buneman Pour la

The authors would like to thank the anonymous referees for helpful suggestions on the first draft of this paper, and W H E Day for his comments and his significant improvements of style

Authors' Addresses: Jean-Pierre Barthélemy, E N S T , 46 rue Barrault, F-75634 PARIS CEDEX 13, France; Bruno Leclerc and Bernard Monjardet, C A M S , 54 boulevard Raspail, F-75270 PARIS CEDEX 06, France Please address correspondence to Bruno Leclerc

comparaison des classifications, la semi-modularité des structures ordinales prises en compte permet de définir des métriques de plus courts chemins, basées sur des ensembles de transformations élémentaires, et effectivement calculables Lorsque ces transformations ne sont pas pondérées, ces métriques se caractérisent simplement Pour le consensus de classifications, on considère les approches constructive, axiomatique et par optimisation d'un critère On a de bonnes formalisations ordinales de règles naturelles (majoritaire, oligarchique,), à partir desquelles on obtient une présentation unifiée de divers résultats de type arrowien Dans le cas des hiérarchies, des hiérarchies stratifiées et des arbres de Buneman, un fait important, résultant de leurs structures de demi-treillis à médianes, est que la règle majoritaire peut être obtenue par chacune des trois approches

Keywords: Hierarchical classification, Median; Metric; Numerical taxonomy; Partial order; Partition; Phylogeny; Ultrametric

1. Introduction

The aim of this paper is to give an overview of results obtained by the ordinal approach in problems of comparison and consensus of classifications This ordinal approach relies on one fact and one claim. The obvious fact is that all the sets of usual taxonomic models (partitions, n-trees, ordered trees, ultrametrics, ..) are naturally ordered For instance, a partition can be finer than another one, an n-tree included in another n-tree, the values of an ultrametric always less than the values of another, and so on. The claim is that knowledge of the abstract structure of these orders allows one to solve some problems raised by the definition of suitable methods of comparison and consensus. This claim is now supported by the works of authors such as Day, Janowitz, Margush, McMorris, Neumann, Norton, Schader, and ourselves, and we are hopeful this paper will convince the readers perhaps yet reluctant Notice also this claim does not say the ordinal approach may solve all — or even most — problems of comparisons or consensus of classifications, such an assertion would be insane We shall begin by presenting two general illustrations of our claim, then we shall give more details and other examples while presenting the contents of this paper.

A taxonomic model, such as a partition or an n-tree, is a discrete mathematical object without natural vectorial descriptions. Then in order to compare two such objects we cannot use the classical measures of distance in a vector or Euclidean space, unless we use more or less arbitrary codings But for such objects it is often possible to define elementary transformations of one object into another, and to define the distance between two objects as the minimum number of such transformations needed to obtain the first from the second. As a matter of fact, many such least-move or minimal path length distances have been defined in various fields; for early examples

in the social sciences see Flament (1963) and Boorman and Arabie (1972) Unfortunately this kind of distance may have a serious drawback· it may be very hard to compute effectively (this point has been especially emphasized by Day) On the other hand, when the set of all objects considered is endowed with a partial order, there are natural elementary transformations, i.e , transformations linked with the ordinal structure, and thus natural minimal path length (MPL) metrics For instance, such an elementary transformation is a transformation of x into y when x and y are comparable but do not admit intermediary elements in the partial order For example, in the case of the ordered set of all partitions, such a transformation is the union of two clusters into a unique one or the converse operation A significant class of partially ordered sets, the semimodular posets, have ordinal formulae allowing one to compute easily the associated MPL metric, moreover one can characterize axiomatically such a metric For instance, since the ordered set of all partitions is semimodular one can apply the above results to it.

Let us turn to consensus problems A classical way to define a consensus between n given objects is to define a remoteness index between them and any consensus candidate and to take as consensus objects those minimizing this remoteness A usual way to define remoteness is to use a function of the distances between the given objects and the consensus candidate So the minimal path length metrics described above can be used to define consensus methods But even if the minimal path length metric is easy to compute, the associated (given a remoteness function) consensus method may be problematic since a consensus object may be very hard to compute Such behavior is often the case with the median consensus, defined by taking as a remoteness function the sum of distances, even though this median procedure is often the most relevant (Barthélemy and Monjardet 1981) However, for the class of median semilattices (a significant subset of the modular partially ordered sets), the median consensus becomes easy to compute and moreover this median procedure can be axiomatically characterized. For instance, since the set of all n-trees is a median semilattice (for the inclusion order between n-trees), one can apply the above results to it

This last example also makes clear an advantage of the ordinal approach that is shared by any abstract or axiomatic approach The abstract theory of the median procedure in median semilattices, or equivalently in median graphs, is not quite simple; for instance it requires an unobvious embedding theorem due to Sholander (1954, see Bandelt and Barthélemy 1984). On the other hand it is easy to show that the set of all n-trees is a median semilattice, and so to obtain many results on the median consensus of n-trees.

We now give a description of the contents of this paper, where we make clear the main ordinal structures encountered and the kind of results they allow one to obtain A first remark has already been made all the sets

S of taxonomic models considered here are partially ordered. So the minimal abstract structure is nothing else than that

$$OS_O \ (S,\leqslant) \textit{ is a partially ordered set.}$$

Recall that the relation $\leqslant$ is then reflexive, antisymmetric and transitive, and that the expression *partially ordered set* is often abbreviated as *poset* In Section 2 we systematically investigate the ordinal structure of the five sets of the most usual taxonomic models partitions, ultrametrics (or dendrograms), n-trees, ordered trees, and Buneman (or phylogenetic) trees. In each of these posets there exists a meet or (and) join for any two such models so that

$$OS_1 \ (S,\leqslant) \textit{ is a semilattice.}$$

However OS_1 is not sufficient to describe the common structure of our sets of classifications All these semilattices are semimodular (the precise definition is in Section 3 2 1) from an abstract point of view the significant ordinal structure is that

$$OS_2 \ (S,\leqslant) \textit{ is a semimodular semilattice}$$

Moreover three of these ordered sets of classifications, i e., the sets of n-trees, ordered trees, and Buneman trees, have common additional properties making them very close to distributive lattices the shared ordinal structure of interest is that

$$OS_3 \ (S,\leqslant) \textit{ is a median semilattice.}$$

In Section 2 we do not assume the reader has a basic knowledge of partial orders, so we define through the examples such basic notions as meet, join, covering relation, and irreducible element. Abstract definitions can be found either in the reference books cited or in the other sections

In Section 3 we present results of the ordinal approach for the problem of comparing classifications. First the results already mentioned above are precisely stated For semimodular semilattices (the OS_2 case), we give formulae allowing one to compute the MPL metrics and we present axiomatic characterizations of such metrics. Then we consider more general cases where one copes with minimal weighted path length (MWPL) metrics defined on the ordered set S Here again one uses elementary ordinal

transformations, but now they have an associated weight or cost say there exists a monotone real-valued function v on $(S,\leqslant)$ and the weight of the elementary transformation from x into y is $| v(y) - v(x) |$ Here also the effective computation of this distance may be difficult unless the function v satisfies additional properties and thus constitutes a so-called semivaluation So we obtain our last abstract structure in which

$OS_4(S,\leqslant,v)$ *is a partially ordered set with a semivaluation* v

In this case one obtains also a formula to compute the MWPL metric moreover if S is a semimodular semilattice one has a good local criterion to recognize a semivaluation We end Section 3 by describing the metric approach to consensus problems already mentioned above.

In Section 4 we present results of the ordinal approach to the problem of finding a consensus of classifications First we recall the three main approaches to achieve consensus the axiomatic approach going back to Arrow's theorem; the constructive approach going back at least to Borda (1784) and Condorcet (1785) with their sum of ranks and majoritary rules (see Guilbaud 1952), and the optimization approach in which a remoteness function is minimized. We begin by considering the constructive approach In Section 4 2 we define consensus rules generalizing the classical majoritary or oligarchic rules by taking lattice polynomial functions Such rules can be defined in any semilattice (the OS_2 case) In Section 4 3 we look at the effective computation of these rules The two following sections are devoted to the axiomatic approach First we give an ordinal presentation of Arrow-like axioms, valid in the OS_2 case Next we present Arrow-like results for the consensus of partitions or ultrametrics Finally we show that for median semilattices (the OS_3 case), one can obtain axiomatic characterizations of the lattice polynomial consensus functions Moreover in Section 4 6 we show that in this case, the three approaches to consensus can coincide the median procedure defined by an optimization criterion can be computed by a lattice polynomial rule and can be axiomatically characterized These results can especially be applied to n-trees, ordered trees and Buneman trees

2. Ordinal Structures of Sets of Taxonomic Models

2.1 Natural Orders on Classes of Taxonomic Models

In this section we shall present the latticial or semilatticial structures of sets of the usual taxonomic models We are not here interested by the ordinal structure that a specific taxonomic model, for instance an n-tree, may have, but rather by the ordinal structure of the set of all n-trees, or of subsets of this set. The taxonomic models we shall consider belong to five

classical types (defined below) partitions, dendrograms (valued trees) or ultrametrics, n-trees, ordered trees, and Buneman trees The latticial structure of the set $\mathbf{P}_n$ of all partitions on an n-set is well known, indeed this lattice is a famous example of what is now called a geometric (or matroid) lattice. Moreover, several problems about $\mathbf{P}_n$ have led to significant advances in discrete mathematics (general references on $\mathbf{P}_n$ may be found in Birkhoff 1967, Barbut and Monjardet 1970, Aigner 1979, and Pudlak and Tuma 1980). Recall also that a partition P is finer than a partition P', such order on $\mathbf{P}_n$ being denoted by $P \leqslant P'$, if and only if every cluster of P' is a union of clusters of P.

We begin in Section 2.2 with the study of the lattice $\mathbf{U}_n$ of all ultrametrics on an n-set. This lattice is closely related to $\mathbf{P}_n$ and so may be called a quasi-geometric lattice. Then in Section 2 3 we study the semilattice $\mathbf{T}_n$ of all n-trees, a semilattice belonging to the important class of median semilattices (a generalization of distributive lattices). Finally in Section 2.4 we show that several other sets of taxonomic models are also median semilattices or distributive lattices, a fact that will be very useful for the problems of consensus tackled in Section 4.

2.2 The Lattice of all L-ultrametrics

In this paper X is always a finite set with n elements (an n-set) and L is always the set $\{0<1<...<l\}$ of the first l positive integers, with zero The notions and results below would be exactly the same for $L = \{0<\lambda_1< \ <\lambda_l\}$, λ_i a real number, and they could be adapted for $L = \mathbb{R}^+$.

Definition 2.1 An *L-ultrametric* on the set X is a map U from the set X^2 of all ordered pairs of X to L which satisfies the following conditions

for every x in X, $U(x,x) = 0$,
for every x,y in X, $U(x,y) = U(y,x)$,
for all x,y,z in X, $U(x,y) \leqslant max\,[U(x,z), U(z,y)]$

We denote by $\mathbf{U}_{L,n}$ the set of all L-ultrametrics on the n-set X. It is well known (Benzécri 1967, Johnson 1967) that there is a natural bijection between $\mathbf{U}_{L,n}$ and the set, denoted by $\mathbf{D}_{L,n}$ of all L-dendrograms.

Definition 2.2 An *L-dendrogram* is a map D with domain L and codomain $\mathbf{P}_n$ which satisfies the following conditions

$D(l) = \{X\}$;
$\lambda \leqslant \lambda'$ implies $D(\lambda) \leqslant D(\lambda')$.

This definition is valid with L finite, in the infinite case ($L = \mathbb{R}^+$) one has to add a continuity condition (see Jardine and Sibson 1971) If an

ultrametric U satisfies the condition that $U(x,y) = 0$ implies $x = y$, then in the associated dendrogram $D(0)$ is the finest partition in which every cluster is a single element of X. If not, $D(0)$ may be any partition of X.

The set $\mathbf{U}_{L,n}$ of all L-ultrametrics is naturally pointwise ordered

$$U \leqslant U' \text{ if and only if for all } x,y \text{ in } X,\ U(x,y) \leqslant U'(x,y)$$

$\mathbf{U}_{L,n}$ endowed with this partial order is a lattice First, if U and U' are two L-ultrametrics, an L-ultrametric is obviously defined by taking $max[U(x,y),U'(x,y)]$ for every pair (x,y). This ultrametric is the least upper bound, or *join*, of U and U' and is denoted by $U \vee U'$ Second, it is easy to see that the set of lower bounds of the ultrametrics U and U' has always a greatest element, or *meet*, denoted by $U \wedge U'$ Indeed, $U \wedge U'$ is the classical subdominant ultrametric (Jardine and Sibson 1971) associated with the dissimilarity $\Delta(x,y) = min[U(x,y),U'(x,y)]$ This lattice $\mathbf{U}_{L,n}$ has been studied by Leclerc (1979, 1981, see also Barthélemy, Leclerc and Monjardet 1984a). It is worth noticing, since the set $\mathbf{D}_{L,n}$ of all L-dendrograms is one-to-one with the set $\mathbf{U}_{L,n}$ of all L-ultrametrics, that $\mathbf{D}_{L,n}$ may be also endowed with a latticial structure Indeed, the natural order for dendrograms is also the pointwise order·

$$D \leqslant D' \text{ if and only if for every } \lambda \text{ in } L,\ D(\lambda) \leqslant D'(\lambda),$$

and it is easy to see that $\mathbf{D}_{L,n}$ endowed with this partial order is a dual (anti-isomorphic) lattice of the lattice $\mathbf{U}_{L,n}$. More precisely, the meet and join operations in $\mathbf{D}_{L,n}$ are defined for every λ in L by $D \wedge D'(\lambda) = D(\lambda) \wedge D'(\lambda)$ and $D \vee D'(\lambda) = D(\lambda) \vee D'(\lambda)$ This lattice has been considered by Boorman and Olivier (1973) in the case where $L = \mathbb{R}^+$

From a more abstract point of view, one can notice (Barthélemy *et al.*, 1984a), following Janowitz (1978), that an L-dendrogram is a special case of a residuated map between two partially (or linearly) ordered sets, then $\mathbf{U}_{L,n}$ is the dual of the lattice of all residuated maps between L and $\mathbf{P}_n$, a fact allowing one to use powerful tools of the theory of partially ordered sets (Blyth and Janowitz 1972, Leclerc 1984, Barthélemy *et al.*, 1984a). Moreover this fact accounts for several properties of $\mathbf{U}_{L,n}$ Here we shall only give three properties (Leclerc 1981) corresponding to well-known properties of the lattice of partitions $\mathbf{P}_n$. Indeed, if $L = \{0 < 1\}$, $\mathbf{U}_{L,n}$ is the dual lattice of $\mathbf{P}_n$, and for $|L| \geqslant 3$ we obtain dual generalizations of properties of $\mathbf{P}_n$

First, we need the definition of the covering relation associated with the partial order in $\mathbf{U}_{L,n}$ (such a notion may be defined in any finite partial order)· the ultrametric U is *covered* by the ultrametric U' if and only if $U < U'$ and $U < U'' \leqslant U'$ implies $U'' = U'$ If U is covered by U' we write

$U \prec U'$; in other words, $U \prec U'$ if there exists no ultrametric between U and U'.

In order to characterize this covering relation one uses the well-known notion of the minimum spanning tree (MST) of an arbitrary dissimilarity (Gower and Ross 1969, Hartigan 1975, Chap. II). Then $U \prec U'$ if and only if U and U' have a common MST such that $U'(x,y) = U(x,y) + 1$ for the two ordered pairs associated with a unique edge $\{x,y\}$ of this MST, and $U'(z,t) = U(z,t)$ for all the ordered pairs associated with the other edges of the MST So in this case there is a subset of X^2 on which U' takes the value $U + 1$, whereas on the other ordered pairs U' equals U

Proposition 2.1 *The lattice* $\mathbf{U}_{L,n}$ *of all L-ultrametrics on X is lower semimodular, i.e., for every U and U' in* $\mathbf{U}_{L,n}$, $U \prec U \vee U'$ *and* $U' \prec U \vee U'$ *implies* $U \wedge U' \prec U$ *and* $U \wedge U' \prec U'$.

In order to make more intuitive this property of semimodularity, we give the following definitions (valid for any partially ordered set) An *upper triangle* in $\mathbf{U}_{L,n}$ is a 3-tuple (U_1, U_2, U_3) with $U_1 \prec U_3$ $U_2 \prec U_3$ and $U_1 \neq U_2$. A *quadrilateral* in $\mathbf{U}_{L,n}$ is a 4-tuple (U_0, U_1, U_2, U_3) with $U_0 \prec U_1$, $U_0 \prec U_2$, $U_1 \prec U_3$, $U_2 \prec U_3$ and $U_1 \neq U_2$. Then to say $\mathbf{U}_{L,n}$ is lower semimodular is equivalent to saying that in $\mathbf{U}_{L,n}$ each upper triangle (U_1, U_2, U_3) completes into a quadrilateral (U_0, U_1, U_2, U_3). Notice the lattice $\mathbf{P}_n$ of all partitions of X is upper semimodular, i.e , in $\mathbf{P}_n$ each lower triangle $(P_1 \prec P_2, P_1 \prec P_3)$ completes into a quadrilateral $(P_1 \prec P_2, P_1 \prec P_3, P_2 \prec P_4, P_3 \prec P_4)$

It is well known that a lower or upper semimodular poset S is ranked, i.e., one may assign to each element s of S an integer $r(s)$ such that $s \prec t$ (s is covered by t) implies $r(t) = r(s) + 1$. Such a function r is called a *rank* function. If the ranked poset has a least element ϕ, the rank function assigning the value zero to this least element is called the height function. then $h(s)$ is just the number of nonzero elements in any covering sequence $\phi \prec s_1 \prec \cdots \prec s_p = s$ from ϕ to s. Our second result gives the height function in $\mathbf{U}_{L,n}$.

Proposition 2.2 *The height* $h(U)$ *of an L-ultrametric U is* $h(U) = \sum U(x,y)$, *where the sum is taken over all the pairs of a minimum spanning tree of U.*

We end this section by results on the decomposition of an ultrametric into simple ultrametrics. We use two kinds of simple ultrametrics An L-ultrametric U is said to be *elementary* if and only if there exists $\lambda > 0$ in L and a bipartition $\{Y,Z\}$ of $X(Y \cup Z = X, Y \cap Z = \varnothing)$ such that for all (x,y) in $Y^2 \cup Z^2$, $U(x,y) = 0$, for every x in Y and y and Z, $U(x,y) = \lambda$. Such ultrametrics have been defined by Hubert (1977) An

ultrametric U is said to be *coelementary* if and only if there exists $\lambda < l$ in L and a pair (x_0,y_0) of X^2 such that $U(x_0,y_0) = \lambda$, and for all the other pairs (x,y), $x \neq y$, $U(x,y) = l$

Proposition 2.3 *Every L-ultrametric is a join of at most* $n - 1$ *elementary L-ultrametrics and a meet of at most* $n - 1$ *coelementary L-ultrametrics.*

The fact that every ultrametric is a join of elementary ultrametrics or a meet of coelementary ultrametrics is a simple consequence of the latticial structure of $\mathbf{U}_{L,n}$ Indeed, in any finite lattice, every element is a join of join-irreducible elements, and a meet of meet-irreducible elements, where a *join-irreducible* (resp *meet-irreducible*) element is an element covering (resp covered by) a unique element One may show that the elementary (resp. coelementary) ultrametrics are exactly the join-irreducible (resp. meet-irreducible) ultrametrics of the lattice $\mathbf{U}_{L,n}$. The fact that the decompositions in Proposition 2 3 use at most $n - 1$ simple ultrametrics follows from the fact that an ultrametric is determined by any one of its minimum spanning trees

For proofs of the above propositions and more results on the lattice $\mathbf{U}_{L,n}$ see Leclerc (1979, 1981)

2.3 The Semilattice of all n-trees

Definition 2.3 An *n-tree* (or a *bare tree*) T on X is a family of subsets, called *clusters*, which satisfies the following conditions

$X \in T, \varnothing \notin T$,

for every x *in* X, $\{x\} \in T$,

for all C, C' *in* T, $C \cap C' \in \{C,C',\varnothing\}$.

The set $\mathbf{T}_n$ of all n-trees has on it a natural partial order

$$T \subseteq T' \text{ if and only if } C \in T \text{ implies } C \in T'$$

The intersection of two n-trees, i.e , the set of their common clusters, is obviously an n-tree; so $T \cap T'$ is the meet $T \bigwedge T'$ of the n-trees T and T' Thus the set $\mathbf{T}_n$ of all n-trees on X is a meet semilattice The semilattice $\mathbf{T}_4$ is represented in Figure 1 The least element of $\mathbf{T}_n$ is the *bush* $T_\phi = \{X\} \cup \{\{x\} \quad x \in X\}$, the clusters of the bush will be called the *trivial*

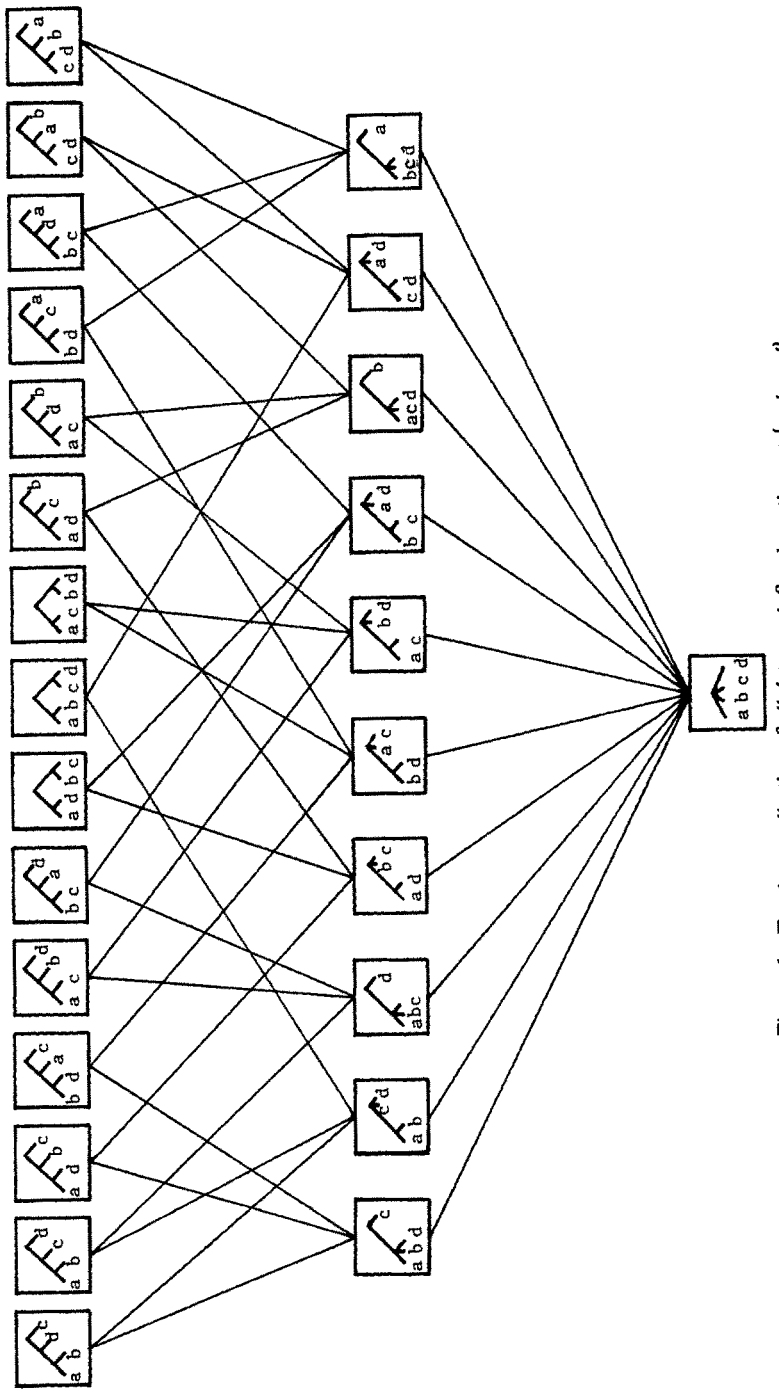

Figure 1. $\mathbf{T}_4$, the semilattice of all 4-trees defined on the set $\{a,b,c,d\}$.

clusters. It is easy to find two n-trees T and T' having no common upper bound (take C in T and C' in T' with $C \cap C' \notin \{C, C', \varnothing\}$) So $\mathbf{T}_n$ is not a lattice, indeed $\mathbf{T}_n$ has $1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-3)$ maximal n-trees (Harding 1971), which are exactly the n-trees with $2n-1$ clusters (often called binary trees) Notice also that the join $T \vee T'$ of two n-trees exists only if $T \cup T'$, the set of clusters belonging to T or T', is an n-tree The meet semilattice $\mathbf{T}_n$ of all n-trees on X has been studied by Leclerc (1985a). Using the remark just above one obtains the following property of $\mathbf{T}_n$.

Proposition 2.4 *For all $T_1, T_2, T_3 \in \mathbf{T}_n$ if $T_1 \vee T_2$, $T_2 \vee T_3$ and $T_3 \vee T_1$ exist, then $T_1 \vee T_2 \vee T_3$ exists.*

Since it is obvious that one obtains an n-tree by deleting any set of nontrivial clusters from an n-tree, one gets the following result.

Proposition 2.5 *For every T in $\mathbf{T}_n$ the set of all $T' \subseteq T$ is a boolean lattice.*

Now, the above two properties imply that $\mathbf{T}_n$ is a median semilattice, a fact having significant consequences.

Definition 2.4 A meet semilattice S is a *median semilattice* if and only if it satisfies the following conditions

for all s, t, u in S, if $s \vee t$, $t \vee u$ and $u \vee s$ exist, then $s \vee t \vee u$ exists,
for every s in S, $\{t \in S \cdot t \leqslant s\}$ is a distributive lattice.

By the second condition, a median semilattice with a greatest element is a distributive lattice Using this same condition, one can prove (Sholander 1954) that a median semilattice has a good canonical embedding into a distributive lattice it is obtained from this lattice by a "good beheading," i.e., a deletion of elements at the top of the lattice, preserving the above first condition. An obvious consequence is that a median semilattice is lower semimodular So, in the case of the median semilattice $\mathbf{T}_n$, we have: $T \prec T \cup T'$ and $T' \prec T \cup T'$ imply $T \cap T' \prec T$ and $T \cap T' \prec T'$ Here, $\prec$ is the covering relation in $\mathbf{T}_n$: $T \prec T'$ if and only if T' has all the clusters of T and just one cluster more. Notice also that the height function $h(T)$ of $\mathbf{T}_n$ is nothing more than the number of nontrivial clusters in the n-tree T.

Another significant property of median semilattice is the existence of generalized majoritary polynomials Here we state the property in the special case of n-trees, a more abstract point of view being developed in Section 4.2 Let $T_1, \ldots, T_i, \ldots, T_v$ be n-trees; we set up $V = \{1, \ldots, i, \ldots, v\}$ and we define a *generalized majoritary family* as a nonempty family $\mathbf{W}$ of subsets of V satisfying the following conditions: if $W \in \mathbf{W}$ and $W \subseteq W'$, then

$W' \in \mathbf{W}$; if $W, W' \in \mathbf{W}$, then $W \cap W' \neq \varnothing$ Then we have the following result

Proposition 2.6 *For every generalized majoritary family* $\mathbf{W}$ *of* V, *the n-tree* $T(\mathbf{W}) = \cup_{W \in \mathbf{W}} (\cap_{i \in W} T_i)$ *exists.*

Notice the definition of $T(\mathbf{W})$ is equivalent to saying that a cluster C belongs to $T(\mathbf{W})$ if and only if there exists W in $\mathbf{W}$ such that C belongs to all the T_i's, $i \in W$. One obtains the usual majoritary rule by taking $\mathbf{W} = \{W \in V \mid |W| \geqslant (v+1)/2\}$. In this case, we say $T(\mathbf{W})$ is an algebraic median of the T_i's $(i \in V)$ The significant fact about this algebraic median is that it can be also defined as an n-tree whose distance from the given T_i's is a minimum (see Sections 3 4 and 4.6). This unobvious fact results from the general theory of medians in median semilattices, such a theory having essentially been initiated by Barbut (1961) (see also Barbut and Monjardet 1970, Monjardet 1980) in the case of distributive lattices, then extended to the general case by Bandelt and Barthélemy (1984) Notice also we have here defined median semilattices as meet semilattices, but corresponding results are obtained with join semilattices (by duality)

2.4 Median Semilattices of Other Sets of Taxonomic Models

2 4.1 Ordered Trees. Recall that a binary relation is called a weak order (or a complete preorder) if it is transitive and complete, so the symmetric part of a weak order is an equivalence relation, and the set of its equivalence classes is linearly ordered In the definition just below, we denote by $\leqslant$ such a weak order and by $<$ its asymmetric part.

Definition 2.5 An *ordered tree* on X is an ordered pair $O = (T, \leqslant)$ where T is an n-tree and $\leqslant$ is a weak order on T satisfying the following conditions

> for all clusters C, C' in T, $C \subset C'$ implies $C < C'$;
> all the 1-clusters of T (i e, the singletons $\{\{x\} \quad x \in X\}$) are equivalent according to this weak order

The clusters of an ordered tree are ranked according to the levels of the weak order, from the zero level (formed by all the 1-clusters) to the top level (formed by the cluster X) More generally the maximal clusters less than or equal to a given level of the weak order form a partition of X Thus we can associate with an ordered tree O a chain $\{\{x\} \cdot x \in X\} = P_O < P_1 < .. < P_i < .. < P_k = \{X\}$ of partitions of X Conversely it is easy to see that such a chain defines an ordered tree (the clusters of T are given by the clusters of the P_i's and the weak order between

two clusters, by the order between the indices of the first partition where they appear) Thus this correspondence between the sets of ordered trees and chains of partitions is one-to-one Now there is a natural partial order on the set of all chains of partitions· a chain is contained in another chain if and only if each partition of the first chain belongs to the second chain. We define the order relation on $\mathbf{O}_n$ the set of all ordered trees on X as the order induced by this inclusion order between partition chains $(T,\leqslant) \leqslant (T',\leqslant')$ if and only if each partition of the chain associated with T belongs to the chain associated with T'

So $(\mathbf{O}_n,\leqslant)$ is order isomorphic to the set of all partition chains endowed with the inclusion order Now we can use the easily proved result that for a partially ordered set with least and greatest elements, the set ordered by inclusion of all its chains (i e , its linearly ordered subsets) containing these two elements is a median semilattice (see Barthélemy *et al*, 1984a, 1984b, for a more general result due to Bandelt). So finally we obtain the following result

Proposition 2.7 *The set* $\mathbf{O}_n$ *of all ordered trees on a set* X *is a median semilattice.*

Especially the majoritary properties stated in Section 2 3 may be translated for $\mathbf{O}_n$

The order between two ordered trees has been defined above by using the order between the associated partitions. One can also give the following direct characterization of this order $(T,\leqslant) \leqslant (T',\leqslant')$ if and only if the three following conditions are satisfied

$T \subseteq T'$,
for all $C,C' \in T$, $C < C'$ implies $C <' C'$;
for all $C \in T$ and $C' \in T'$, $C' \leqslant' C$ implies that there is a $C'' \in T$ such that $C' \subseteq C'' \leqslant C$

2 4.2 Buneman Trees

Let a *tree* be a connected, acyclic, undirected graph (see any book on graph theory, e.g., Harary 1969) A *Buneman* (or *phylogenetic*) tree on X is an ordered pair $B = (T,\psi)$ where $T = (V,E)$ is a tree with vertex set V and edge set E, and where ψ is a map from X to V such that each element in $V - \psi(X)$ is linked with at least three other vertices

We denote by $\mathbf{B}_n$ the set of all Buneman trees on X Buneman trees appear in the recovery of bifurcation processes (phylogenetic trees) and in the problem of fitting an additive tree metric to a dissimilarity measure (see Barthélemy and Luong, 1986, for a review and Guénoche, 1986, for an analysis of recent algorithms). The equality between two Buneman trees is defined by convention as $(T,\psi) = (T',\psi')$ if and only if there exists a tree

isomorphism f from $T = (V,E)$ to $T' = (V',E')$ such that for each x in X, $f\psi(x) = \psi'(fx)$, so we consider that only the vertices in $\psi(X)$ are labeled in the Buneman tree (T,ψ).

Consider the set $\mathbf{P}_{n,2}$ of all bipartitions of X We know from Buneman (1971) that Buneman trees are one-to-one with the subsets H of $\mathbf{P}_{n,2}$ such that: if $\sigma = \{E,E'\}$ and $\tau = \{F,F'\}$ are in H, then one of the intersections in $\{E \cap F, E \cap F', E' \cap F, E' \cap F'\}$ is empty. In this case we say that σ and τ are *noncrossing* (compatible according to Buneman's terminology) In this bijection we associate with $B = (T,\psi)$ in $\mathbf{B}_n$ the set $\mathbf{S}(B)$ of bipartitions of X induced by deletions of single edges of T, i.e., the so-called set of *splits* of B. So, if we consider the graph $G(X)$ whose vertices are all the bipartitions of X, two vertices σ and τ being linked by an edge if and only if σ and τ are noncrossing, the Buneman trees are one-to-one with the complete subgraphs of $G(X)$

Now it is easy to establish that all the complete subgraphs of a graph constitute a median semilattice. So the set $\mathbf{B}_n$ may be ordered as a median semilattice. It is also easy to ascertain that the corresponding order $\leqslant$ may be interpreted as resulting from contraction of edges. $(T,\psi) \leqslant (T,\psi')$ if and only if T is obtained by contracting one or several edges in T' (Figure 2 illustrates the contraction of edges.) So we obtain this result

Proposition 2.8 $(\mathbf{B}_n \leqslant)$ *is a median semilattice.*

Consequently, the results of Section 2.3, and more generally of Sections 4.5 and 4.6, apply to $\mathbf{B}_n$. More details about Buneman trees and the structure of $(\mathbf{B}_n, \leqslant)$ may be found in Barthélemy (1985)

2.5 The Landscape of Ordered Sets of Taxonomic Models

This result ends our study of the ordinal structure of sets of taxonomic models. We have seen that these partially ordered sets belong to two kinds of ordinal structures. geometric or geometric-like lattices (for partitions and ultrametrics), and median semilattices (for n-trees, ordered trees, and Buneman trees). Given the significance, for comparison and consensus problems, of the median-semilatticial or distributive-latticial structure, it is worth pointing out that subsets of $\mathbf{P}_n$ or $\mathbf{U}_n$ can have this structure. We give one nontrivial example. Let $\mathbf{U}(M)$ be the set of all L-ultrametrics on X having a given tree M as minimum spanning tree; then $\mathbf{U}(M)$ partially ordered by the pointwise order is a distributive lattice isomorphic to the pointwise ordered set L^M of all maps $M \to L$.

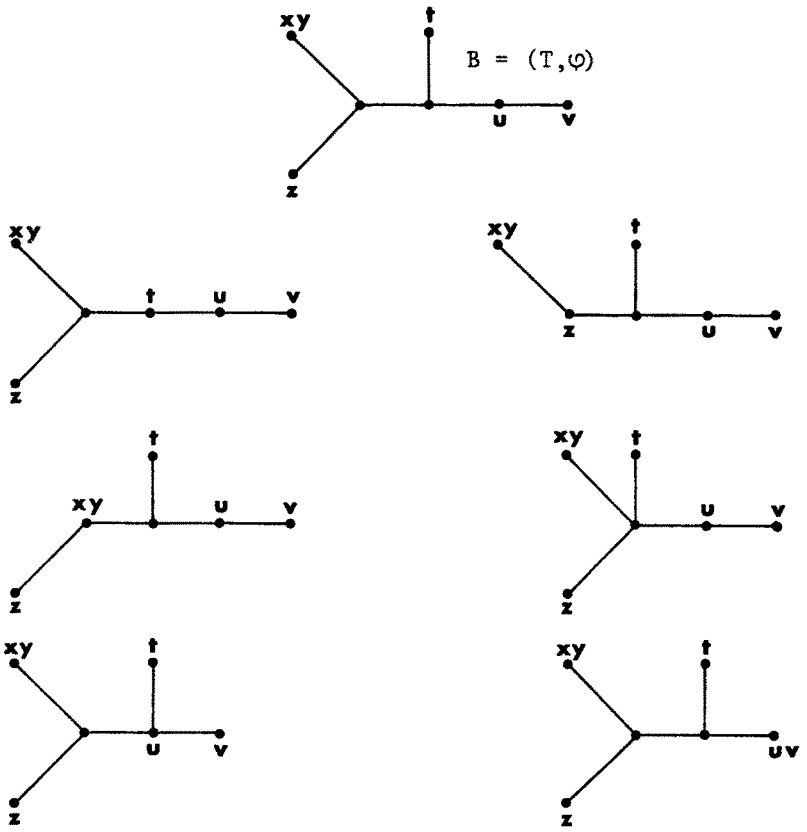

Figure 2 A Buneman tree $B = (T,\psi)$ and the Buneman trees obtained by the contraction of one edge of T Notice that ψ is not necessarily bijective: a vertex may admit multiple labels

3. Ordinal Results on the Comparison Problem: Metric Aspects

3.1 Least Moves for Taxonomic Models

Classifications on the same set X of objects can be compared using the classical least-move approach (Flament 1963, Robinson 1971, Arabie and Boorman 1973) First, admissible elementary transformations between classifications are defined. The least-move metric between two classifications D and D' is defined as the minimum sum of costs of a sequence of transformations connecting D and D'. That is, we construct a graph whose vertices are classifications of a given type (partitions, dendrograms, ordered trees, n-trees, Buneman trees, .) and whose edges are

admissible elementary transformations. Second, we compute a shortest path between two vertices in such a graph, where the length of each edge is the cost of the corresponding elementary transformation. The possibility of inherent intractability of such a computation has been studied by Day and Wells (1984) (cf also Day 1981, 1983a, 1983b, for other complexity results). Here, we just mention that the efficient, classical algorithm of Dijkstra (1959) computes the minimal path length in $O(m^2)$ time, with m as the number of vertices of the graph. Table 1 gives the values of m as a function of $|X| = n$ for our taxonomic models. It illustrates that the complexity of such an algorithm increases exponentially with n. In it, the $p(n)$ are the classical Bell numbers. The $t(n)$ are the numbers occurring in Schröder's fourth problem, and they can be computed from the 2-associated Stirling numbers of the second kind Recall that the number $S_2(n,k)$ enumerates the partitions of an n-set into k classes of cardinality at least equal to 2

$$t(n) = \sum_{k=0}^{n-2} S_2\,(n + k, k + 1) \quad ,$$

(Comtet 1970, 1974; see Leclerc 1985a). The $o(n,i)$ are the numbers of ordered trees with i levels on an n-set and the $S(n,i)$ are the well-known Stirling numbers of the second kind enumerating the number of partitions of an n-set into k classes. In the formulae for $t(n)$ and $b(n)$, $\rho = \log 2 - 1/2$. Notice also that $p(n) \leqslant t(n) \leqslant o(n)$, $p(n) \leqslant u(l,n)$, and $t(n) < b(n + 1)$.

In fact, from Day and Wells (1984), we know that the computations of some distances between classifications are NP-complete (Garey and Johnson 1979). However, consider for example the case of the latticial metric on $\mathbf{P}_n$, i.e., the minimal path length distance on the covering graph of $\mathbf{P}_n$ This metric can be computed with the help of the formula

$$d(P,Q) = 2r(P \quad Q) - r(P) - r(Q) \quad ,$$

with r as some rank function of the lattice $(\mathbf{P}_n, \leqslant)$. The time to compute the join of two partitions is bounded by a polynomial in n. If, as is often the case (for instance for the height), the rank function is also polynomially computable, then the formula ensures that the metric d can be computed in a polynomial time.

Table 1

Set	Number	Formula	Source
$\mathbf{P}_n$	$p(n)$	$p(n) = \frac{1}{e}\sum_{h=0}^{\infty}\frac{h^n}{h!}$	Dobinsky (1877)
$\mathbf{U}_{L,n}$	$u(l,n)$	$u(l,n) = \sum_{i=1}^{\min(l,n-1)}\left\{{l \atop i}\right\} o(n,i)$	Barthelemy, Leclerc, Monjardet (1984a)
$\mathbf{O}_n$	$o(n)$	$o(n) = \sum_{i=1}^{n} S(n,i)\, o(i) - \sum_{i=1}^{n} o(n-i)$	Schader (1980)
$\mathbf{T}_n$	$t(n)$	$t(n) \sim \left[\frac{\rho}{n}\left(\frac{n}{2e\rho}\right)^n\right]^{*}$	Comtet (1970)
$\mathbf{B}_n$	$b(n)$	$b(n) \sim n!\rho^{3/2-n}\left(\frac{2}{\pi n^5}\right)^{*}\left[1+\frac{45-2\rho}{24n}+O\left(\frac{1}{n}\right)\right]$	Foulds and Robinson (1980)

In this section, we shall use the structures of posets of taxonomic models to obtain metrics between classifications. Section 3.2 deals with the case of semimodular posets with nonweighted edges. In Section 3.3 we consider weights on the edges that are compatible with the order Section 3.4 introduces the consensus problem from a metric point of view

3.2 Semimodularity and Metrics for Comparing Classifications: Rank Functions

3.2.1 Semimodularity

We introduced semimodularity in Section 2.2 and we established that the order structures of all the sets studied in Section 2 ($\mathbf{P}_n$, $\mathbf{U}_{L,n}$, $\mathbf{D}_{L,n}$, $\mathbf{T}_n$, $\mathbf{O}_n$, $\mathbf{B}_n$) are semimodular Now we present a somewhat more abstract point of view Consider a poset $(S,\leqslant)$, i e a set S with a transitive, reflexive, and antisymmetric relation $\leqslant$. We say that $s \in S$ *covers* $t \in S$ if $s < t$ and $s < u \leqslant t$ implies $u = t$. In this case we write $s \prec t$ This covering relation has been studied in Section 2 2 for ultrametrics, in Section 2 3 for n-trees, and in Section 2 4.2 for Buneman trees. In the example of the partition lattice $\mathbf{P}_n$, we easily see that $P \prec Q$ if and only if one cluster of Q is the union of two clusters of P, whereas the other clusters are the same in P and in Q.

Semimodularity defines links between triangles and quadrilaterals. As in Section 2 2, we define a *lower triangle* in $(S,\leqslant)$ as a 3-tuple (t,u,v) such that $u \neq v$, $t \prec u$ and $t \prec v$ Dually, an *upper triangle* is a 3-tuple

(u,v,w) *with* $u \neq v$, $u \prec w$ *and* $v \prec w$ *A quadrilateral* in $(S,\leqslant)$ is a 4-tuple (t,u,v,w) such that $u \neq v$, $t \prec u$, $t \prec v$, $u \prec w$ and $v \quad w$.

Definition 3.1 A poset $(S,\leqslant)$ is *lower semimodular* whenever each upper triangle (u,v,w) *completes into a quadrilateral* (t,u,v,w). *Dually* $(S,\leqslant)$ *is upper semimodular* when each lower triangle (t,u,v) completes into a quadrilateral (t,u,v,w).

In the special case of a lattice or of a semilattice, we recapture the definitions of Section 2. As in Section 2, a rank function on the poset $(S,\leqslant)$ is an integer-valued map r defined on S such that $r(u) = r(t) + 1$ if $t \prec u$. We already mentioned in Section 2.2 that a semimodular poset with a least element (or with a greatest element) is ranked. When S is ranked and has a least element ϕ, we call, following the terminology introduced in Section 2.2, the rank function h with $h(\phi) = 0$ the *height function.*

The covering graph of $(S,\leqslant)$ has S as its vertex set and has $\{u,v\}$ as an edge if and only if $u \prec v$ or $v \prec u$. We denote by d the *minimal path length* (MPL) *metric* in this covering graph: $d(s,t)$ is the length of a minimal path between s and t. In the semilatticial case, we also call d the *latticial metric* of $(S,\leqslant)$. Theorem 3.1 generalizes a well-known result in lattice theory (cf. Grätzer, 1978, for instance). It indicates that in the semimodular case the latticial metric is computable by an efficient algorithm when the bounds and rank function are computable by efficient algorithms. This situation is, in some sense, characteristic of semimodularity.

Theorem 3.1 *Let* $(S,\leqslant)$ *be a meet semilattice with a rank function* r. *The following assertions are equivalent:*

(i) $(S,\leqslant)$ *is lower semimodular;*
(ii) r *is such that for each* u, v, w *with* $u \leqslant w$ *and* $v \leqslant w$, $r(u) + r(v) \leqslant r(u \wedge v) + r(w)$;
(iii) the latticial metric on $(S,\leqslant)$ *is given for each* u,v *in* S *by* $d(u,v) = r(u) + r(v) - 2r(u \wedge v)$.

Theorem 3.1* *Let* $(S,\leqslant)$ *be a join semilattice with a rank function r. The following assertions are equivalent:*

*(i)** $(S,\leqslant)$ *is upper semimodular;*
*(ii)** r *is such that for each* t,u,v *with* $t \leqslant u$ *and* $t \leqslant v$, $r(u) + r(v) \geqslant r(u \vee v) + r(t)$;
*(iii)** *the latticial metric on* $(S,\leqslant)$ *is given for each* u,v *in* S *by* $d(u,v) = 2r(u \vee v) - r(u) - r(v)$.

Table 2

Poset	semi modularity	height	latticial metric
$P_n^{(\wedge \vee)}$	upper	$h(P) = n$ − number of clusters in P	$d(P,P') = 2h(P \vee P') - h(P) - h(P')$
$U_n^{(\wedge \vee)}$	lower	$h(U) = \sum_{x\,y \in A} U(x,y)$ A any minimal spanning tree of U	$d(U,U') = h(U) + h(U') - 2h(U \wedge U')$
$D_n^{(\wedge \vee)}$	upper	by duality	by duality
$T_n^{(\wedge)}$	lower	$h(T)$ = number of nontrivial clusters in T	$d(T,T')$ = number of nontrivial clusters in $T \Delta T'$
$O_n^{(\wedge)}$	lower	$h(O)$ = number of nontrivial partitions of O	$d(O,O')$ = number of nontrivial partitions in $O \Delta O' = O \cup O' - O \cap O'$
$B_n^{(\wedge)}$	lower	$h(B)$ = number of edges in B	$d(B,B')$ = number of bipartitions in $S(B) \Delta S(B')$

Assertion (iii) (resp (iii)*) implies that at least one minimal path between u and v passes through $u \wedge v$ (resp. $u \vee v$ Theorems 3.1 and 3.1* extend to arbitrary posets (Monjardet 1976). They hold for our sets of taxonomic models since they are all semimodular semilattices. Table 2 summarizes some results of Section 2 by giving for each poset of taxonomic models its operations, its kind of semimodularity, and formulae for its height function and for its latticial metric.

3.2.2 Axiomatic Characterization of Latticial Metrics

The use of semimodularity allows an axiomatic characterization of the MPL metric (Barthélemy 1979a) The problem concerns whether there exists a metric on a set of taxonomic models satisfying a given set of conditions; and if such a metric exists, is it unique? This kind of study goes back to Kemeny (1959) in the field of preference analysis. The Kemeny approach has stimulated some people working in taxonomy Mirkin and Chernyi (1970) in the case of partitions, Margush (1982) and Leclerc (1985b) in the case of n-trees. Here we give an abstract characterization of the latticial metric in a semimodular semilattice (Barthélemy 1979a).

Theorem 3.2 *Let $(S,\leqslant)$ be a semimodular meet semilattice. The latticial metric d of $(S,\leqslant)$ is the unique real-valued function defined on S^2 and satisfying the following conditions.*

(1) *For each s,s' in S with $s \leqslant s'$, $d(s,s') = d(s',s)$.*
(2) *For each s,t,u in S with $s \leqslant t \leqslant u$, $d(s,u) = d(s,t) + d(t,u)$.*
(3) *For each s,s' in S, $d(s,s') = d(s,s \wedge s') + d(s',s \wedge s')$.*
(4) *For each s,t,u,v in S such that $s \prec t$ and $u \prec v$, $d(s,t) = d(u,v)$.*
(5) *The smallest strictly positive value of d is 1.*

The following statements are worth noticing Symmetry is only partially required (condition (1)) Conditions (2) and (3) are weak forms of the so-called intermediarity condition in a distributive lattice· $s \wedge s' \leqslant t \leqslant s \vee s'$ implies $d(s,s') = d(s,t) + d(t,s')$ The values assigned to d are not *a priori* assumed to be positive (condition (5)). These rather weak conditions induce a strong result· in particular d will be full symmetric, positive, and will satisfy the triangle inequality. Of course, Theorem 3.2 dualizes to upper semimodular join semilattices provided only that condition (3) is changed to

*(3)** *For each s,s' in S, $d(s,s') = d(s,s \vee s') + d(s',s \vee s')$.*

In Section 2, the expression of the meet and/or join has been indicated for each ordered set of taxonomic models, as well as the covering relation The pleasure of formulating Theorem 3.2 for $\mathbf{P}_n$, $\mathbf{U}_{L,n}$, $\mathbf{D}_{L,n}$, $\mathbf{T}_n$, $\mathbf{O}_n$ and $\mathbf{B}_n$ is left to the reader.

3.3 Semimodularity and Metrics for Comparing Classifications: Weighted Edges

3.3.1 Valuation Theory for Posets

Theorems 3.1 and 3 2 hold whenever all admissible transformations between any two classifications have the same cost. In the general case we try to assign costs compatibly with the order. That is, we consider an isotone function v on the poset $(S,\leqslant)$, i e , $s \leqslant s'$ implies $v(s) \leqslant v(s')$, and we assign the weight $|v(s) - v(s')|$ to the edge $\{s,s'\}$ of the covering graph of $(S,\leqslant)$. We shall denote by d_v the minimal weighted path length (MWPL) metric on this covering graph $d_v(s,s')$ is the minimum length of any weighted path between s and s'. Our remarks in Section 3.1 on the possible difficulty of the computation of d_v are concerned with this general case However, for a class of isotone functions, the so-called valuations, the computation of d_v will be as easy as the computations of the bounds and of the values of v.

Definition 3.2 A real-valued isotone function v on a meet semilattice $(S,\leqslant)$ is a *lower valuation* if for each u,v,w in S with $u\leqslant w$ and $v\leqslant w$. $v(u)+v(v)\leqslant v(u \wedge v)+v(w)$

Theorem 3.3 *Let v be a real-valued isotone function on a meet semilattice $(S,\leqslant)$ The following assertions are equivalent:*

(i) v is a lower valuation;
(ii) the MWPL metric d_v is given by $d_v(s,s') = v(s)+v(s')-2v(s \wedge s')$ for each s,s' in S.

This theorem (Bordes 1976) extends to arbitrary posets (Barthélemy 1978) and dualizes An isotone function v on a join semilattice $(S,\leqslant)$ is an *upper valuation* if for each u,v,t with $t\leqslant u$ and $t\leqslant v$ $v(u)+v(v)\geqslant v(t)+v(u \vee v)$ The dual of Theorem 3 3 then asserts that v is an upper valuation on $(S,\leqslant)$ if and only if $d_v(s,s') = 2v(s \vee s')-v(s)-v(s')$ for each s,s' in S In addition, Theorems 3 1 and 3 3 assert that a meet semilattice is lower semimodular if and only if it is ranked by a lower valuation.

3.3.2 The Use of Semimodularity

So, in the case of an easily computable valuation, the corresponding MWPL metric is easily computable for our taxonomic models. However, to verify that a given isotone function is, or is not, a valuation, may not be easy In case of semimodularity, the quadrilateral condition (Barthélemy 1978) provides a simple device to solve this task

Theorem 3.4 *Let v be an isotone function on a lower semimodular meet semilattice $(S,\leqslant)$. Then v is a lower valuation if and only if for each quadrilateral (t,u,v,w) of $(S,\leqslant)$. $v(u)+v(v)\leqslant v(t)+v(w)$*

Notice that this result dualizes and extends (but only partially) to arbitrary posets (Barthélemy 1978) Many such metric results on posets may be found in Monjardet (1981)

3 3 3 The Determination of Distances Between Classifications

Many distances between partitions derive from lower or upper valuations on the poset $(\mathbf{P}_n,\leqslant)$. cf. Boorman and Arabie (1972), Arabie and Boorman (1973), Barthélemy (1979b), Day (1981). MWPL metrics have also been studied in the case of n-trees (Boorman and Olivier 1973, Hubert and Baker 1977, Margush 1982, Day 1985, Day and Faith 1985, Leclerc 1985b) Here, we only indicate some general ways of obtaining such metrics

for $\mathbf{T}_n$, $\mathbf{U}_{L,n}$, and $\mathbf{B}_n$. First, consider a cluster index f, where f is an isotone real-valued function on 2^X, and for an n-tree T define

$$v(T) = \sum_{C \in T} a_{T,C}, f(C) \ ,$$

where the $a_{T,C}$ are positive real numbers (Leclerc 1985b).

Proposition 3.5 *v is a lower valuation on $\mathbf{T}_n$ if and only if for each n-tree T and each pair C, C' of nontrivial clusters in T:*
$A \neq C$ implies $a_{T,A} - a_{T-\{C\},A} \geqslant 0$; and
$A \neq C$ and $A \neq C'$ implies $a_{T,A} + a_{T-\{C,C'\},A} - a_{T-\{C\},A} - a_{T-\{C'\},A} \geqslant 0$.

Consider, as examples, the following cases. When $f = 1$ and $a_{T,C} = 1$, d_v is simply the latticial metric on $\mathbf{T}_n$. When $f(C) = |C|$ and $a_{T,C} = 1$, d_v is given by $d_v(T,T') = \sum_{C \in T \Delta T'} |C|$. When $f(C) = 1$ and $a_{T,C} = |\{C' \in T: C \subseteq C'\}|$, d_v is the Margush metric (1982).

The case of ultrametrics, or dendrograms, seems to have been less intensively studied. However, one can describe two general ways of obtaining upper (resp. lower) valuations on $\mathbf{D}_{L,n}$ (resp. $\mathbf{U}_{L,n}$). Any upper valuation μ on $\mathbf{P}_n$ induces an upper valuation v on $\mathbf{D}_{L,n}$ defined by

$$v(D) = \sum_{\lambda \in L} \mu(D(\lambda))$$

(Boorman and Olivier 1973). Let v_0 be a positive real-valued function defined on the $(n-1)$-fold Cartesian product L^{n-1}, and assume that v_0 is invariant under all permutations of the coordinates. We define a real-valued function v on $\mathbf{U}_{L,n}$ by

$$v(U) = v_0(U(a_1), U(a_2), \ldots, U(a_{n-1})) \ ,$$

with the a_i's as the pairs of some minimum spanning tree of U.

Proposition 3.6 *v is a lower valuation on $\mathbf{U}_{L,n}$ if and only if v_0 is a lower valuation on L^{n-1}.*

On the other hand, a simple upper valuation on $\mathbf{U}_{L,n}$ is the function v defined by $v(U) = \sum_{x,y \in X} U(x,y)$.

The comparison of Buneman (or phylogenetic) trees has been frequently considered, even in the context of n-trees or dendrograms, as a way of comparing the shapes of the trees; see the references in Robinson and Foulds (1981) The distance they proposed in this last paper, in particular, is simply the latticial metric on $\mathbf{B}_n$. We remark that, more generally, Proposition 3 5 can be reformulated for Buneman trees As before, let $\mathbf{S}(B)$ denote the set of all the splits σ of B, i.e., the set of bipartitions of X induced by deletions of single edges of B Consider a *split index* f, i e , a positive function defined on the set of all bipartitions of X, and for a Buneman tree B define

$$v(B) = \sum_{\sigma \in \mathbf{S}(B)} a_{B,\sigma} f(\sigma) \ ,$$

where each $a_{B,\sigma}$ is a positive real number

Proposition 3.7 *v is a lower valuation on $\mathbf{B}_n$ if and only if for each $B \in \mathbf{B}_n$ and each pair τ,τ' of nontrivial splits of B*
$\sigma \neq \tau$ implies $a_{B,\sigma} - a_{B-\{\tau\},\sigma} \geqslant 0$; and
$\sigma \neq \tau$ and $\sigma \neq \tau'$ implies $a_{B,\sigma} + a_{B-\{\tau,\tau'\},\sigma} - a_{B-\{\tau\},\sigma} - a_{B-\{\tau'\},\sigma} \geqslant 0$.

The case where $a_{B,\sigma} = f(\sigma) = 1$ corresponds to the latticial metric. More generally for $a_{B,\sigma} = 1$, we get $d_v(B,B') = \sum_{\sigma \in \mathbf{S}(B)\Delta\mathbf{S}(B')} f(\sigma)$. Other cases are left to the imagination of the reader

Now, we come back to the difficulty of computing a distance between classifications. Notice, for the lattices or the semilattices $\mathbf{P}_n$, $\mathbf{U}_{L,n}$, $\mathbf{D}_{L,n}$, $\mathbf{T}_n$, $\mathbf{O}_n$, $\mathbf{B}_n$, the bounds are computable polynomially in n (cf Day 1981, 1985, for fine studies of complexity emphasizing several linear cases). Thus, whenever a lower or upper valuation is polynomial in n too, we get a polynomial time (in n) algorithm to compute the related MWPL metric.

3.4 The Metric Approach to Consensus

The landscape of ordinal results in problems of consensus of classifications will be described in Section 4, where the main results will be given. Here we just indicate how the metric approach can be applied to this problem.

Metrics between classifications may be used to obtain consensus in the same way that statistical measures of remoteness may be used to identify a central value in a series of numbers Generally, consider a metric space (S,d), a ν-tuple $\pi = (s_1, \ . \ , s_\nu) \in S^\nu$ and an integer p Considering $s_1, \ldots, s_\nu$ as a statistical series, the central value of order p will be obtained by the minimization of $D_p(\pi,s)$ where

$$D_p(\pi,s) = \sum_{i=1}^{v} d^p(s,s_i)$$

For $p = 1$ we get the so-called median of $(s_1, \ldots, s_v)$ By applying this paradigm to sets of taxonomic models, and by matching the solutions (or one solution) of the problem $\min_s D_p(\pi,s)$ with the number

$$I(\pi) = 1 - [(\min_s D_p(\pi,s)) \,/\, \max_s D_p(\pi,s)] \ ,$$

we get a so-called *consensus index method* (Day and McMorris 1985, see Rohlf 1982 for the notion of consensus index) Alternatives for $I(\pi)$ are, for instance,

$$I(\pi) = 1 - [(\min_s D_p(\pi,s)) \,/\, \max_\pi \min_s D_p(\pi,s)]$$

and

$$I(\pi) = 1 - [(\min_s D_p(\pi,s)) \,/\, ((1/|S|) \sum_{s\in S} D_p(\pi,s))] \ .$$

Such a consensus index method may be difficult to compute efficiently because it may require the solution of difficult optimization problems. A well-known example is the median procedure for some sets of binary relations, coupled with the first alternative for $I(\pi)$ above· here d is the symmetric difference distance between binary relations, $p = 1$, and S is, for instance, the set of all equivalence relations or the set of all linear orders (see Barthélemy and Monjardet 1981) On the other hand, these optimization problems are simple to solve for some metric spaces (S,d) The following section will show this to be the case when S is a median semilattice and d is its associated metric, as when S is $\mathbf{T}_n$, $\mathbf{O}_n$, or $\mathbf{B}_n$

4. Ordinal Results in the Consensus Problem

4.1 Three Main Approaches to Achieve Consensus

Let S be a set of classifications on X and let $V = \{1, \quad . , v\}$ be an index set The *consensus problem* is the problem of defining, for any v-tuple (or *profile*) $\pi = (s_1, \ldots, s_v) \in S^V$, one or several consensus classifications. Each of them is expected to be a good representative of the entire profile, i.e., to agree as well as possible with all the components of π. Especially, one may seek a consensus function $c\cdot S^V \to S$ that assigns a single consensus classification $C(\pi)$ to any profile π.

The three main approaches to defining consensus functions are the constructive, the axiomatic and the optimization ones Very simple illustrations of these approaches are provided by the arithmetic mean, viewed as a consensus function on profiles of numbers. The mean is obtained by sequences of sums and divisions (constructive approach); it is the unique function satisfying linearity, bound conditions and invariance by permutation of indices (axiomatic approach), it minimizes the sum of its squared differences with the elements of the profile (optimization approach)

Furthermore, this example prompts several remarks First, the interest in the mean as a consensus of tuples of numbers is precisely due to this convergence of properties of easy calculability, representativity, and optimality Second, important consensus results concern relations between the three approaches that are consequences of mathematical results derived from the structural properties of the domain S Third, following Neumann and Norton (1985), we emphasize the diversity of consensus problems and consensus functions though the mean has good properties, it is not the only consensus of numbers used in practice

In this section, we present some ordinal concepts and results that turn out to be useful in research on the consensus of classifications Lattice polynomials are typically ordinal construction rules, some of them correspond to well-known consensus rules (Section 4.2), their computational time is a polynomial function in n and, in most important cases, in v (Section 4 3) Join irreducible elements are the abstract ordinal counterpart of various elementary entities (ordered pairs, clusters,) involved in axiomatic Arrow-like approaches (Section 4.4). Their use leads to a unified presentation of a number of axiomatic characterizations of polynomial consensus rules (Section 4 5) In the optimization approach, we consider medians, already defined at the end of Section 3 In median semilattices, like $\mathbf{T}_n$, $\mathbf{O}_n$, and $\mathbf{B}_n$, the majority rule polynomial (the algebraic median of Section 2 3) gives always a median, one that is unique if v is an odd number (Section 4 6). Thus the three approaches converge in this case Indeed, this fact is an extension of the properties of the median of numbers, the most frequently used alternative to the mean as a consensus of numbers Sets of numbers and median semilattices have just enough common structural features to admit similar solutions to the consensus problem, with similar properties

4.2 Consensus Rules and Lattice Polynomial Functions

In this section and the following ones, we assume that S is a meet semilattice A lattice polynomial consensus rule $c_{\mathbf{W}}$ may be associated with any family $\mathbf{W} \subseteq \mathbf{P}(V)$ of subsets of V For any profile $\pi = (s_1, \ldots, s_v)$ in S^V,

$$c_{\mathbf{W}}(\pi) = \bigvee_{W \in \mathbf{W}} \bigwedge_{i \in W} s_i$$

where it is assumed there is at least one nonempty set W in $\mathbf{W}$. Note that rule $c_{\mathbf{W}}$ does not change when $\mathbf{W}$ is replaced by the set of its minimal elements (relative to inclusion order).

Such consensus rules constitute the general formalization of a natural idea. For any s_0 in S, if there is W in $\mathbf{W}$ such that all the i's in W agree about the fact that $s_0 \leqslant s_i$, then the consensus $c_{\mathbf{W}}(\pi)$ is such that $s_0 \leqslant c_{\mathbf{W}}(\pi)$ too. In this sense, any W in $\mathbf{W}$ is a so-called *decisive subset* of indices.

Indeed $c_{\mathbf{W}}$ is defined as a function if and only if the joins involved exist for any profile π. This is the case if S is a lattice. Otherwise, $c_{\mathbf{W}}$ works for some families $\mathbf{W}$ (for instance, if $|\mathbf{W}| = 1$) and does not for other ones (for instance, if there are two distinct singletons in $\mathbf{W}$) It is not difficult to extend to any median semilattice the observation, about n-trees, in Section 2.3 in a median semilattice, $c_{\mathbf{W}}$ is defined as a function when $\mathbf{W}$ is a generalized majoritary family where any two decisive subsets intersect The corresponding polynomial consensus rule $c_{\mathbf{W}}$ is said to be a *generalized majoritary rule.* When $|\mathbf{W}| = 1$, $c_{\mathbf{W}}$ is a meet and corresponds to an *oligarchic,* or strict consensus rule such that, for some nonempty $W \subseteq V$

$$c(\pi) = \bigwedge_{i \in W} s_i \ .$$

Special cases of oligarchic rules occur when $W = V$ or when $W = \{i\}$ for some i in V:

$$c(\pi) = \bigwedge_{i \in V} s_i \qquad (\textit{unanimity} \text{ rule});$$

$$c(\pi) = s_i \qquad (\textit{dictatorial} \text{ rule}).$$

When unanimity is replaced by agreement between a given number w of elements of π, we get the *quota* rule c_w (identical to unanimity for $w = v$)

$$c_w(\pi) = \bigvee_{W \subseteq V, |W| = w} \bigwedge_{i \in W} s_i \ .$$

In a median semilattice, polynomials corresponding to quota rules are defined as functions for $w > v/2$ The case where w is the least integer greater than $v/2$ corresponds to the well-known *majority* rule:

$$m(\pi) = \bigvee_{W \subseteq V, |W| > v/2} \bigwedge_{i \in W} s_i$$

All these definitions dualize in join semilattices by exchanging join and meet operators. We obtain lattice polynomials in the dual form, some of them corresponding to the so-called dual oligarchic rules (joins of elements of π), dual majority rule, and so on In lattices, we have lattice polynomials in both forms. In distributive lattices and in median semilattices for $w > v/2$, the quota rule with fixed w and the dual quota rule with $w' = v - w + 1$ are the same, by laws of distributivity.

$$c_w(\pi) = \bigvee_{W \subseteq V, |W| = w} \bigwedge_{i \in W} s_i = \bigwedge_{W \subseteq V, |W| = v-w+1} \bigvee_{j \in W} s_j$$

$$= c_{v-w+1}(\pi) \ .$$

Especially, for odd v, majority and dual majority rules are the same.

In the following sections, we recall or establish properties of polynomial rules One has to observe that, sometimes, these rules do not give interesting consensus classifications For instance, when the n-trees of a profile $\pi \in \mathbf{T}_n^v$ are too different, the majority rule n-tree is the bush $T_\varnothing$ with only trivial clusters Several nonpolynomial consensus methods on n-trees have been proposed (Adams 1972, Neumann 1983, Stinebrickner 1984, Finden and Gordon 1985, Neumann and Norton 1985). Though the properties of these methods are not always clear, they may give nontrivial consensus n-trees in situations where polynomial rules do not work satisfactorily One may use a consensus index (see Section 3 4) in order to avoid using a consensus n-tree when the elements of the profile are irreconcilable The situation is quite the same in the cases of ordered trees or Buneman trees

4.3 Some Computational Considerations

Clearly, the usefulness of lattice operations in construction rules depends on their computational time complexity It was recalled in Section 3 that the lattice operations in $\mathbf{P}_n$ and $\mathbf{U}_{L,n}$ and the meet in $\mathbf{T}_n$, $\mathbf{O}_n$ and $\mathbf{B}_n$ are polynomial in n. In the three median semilattices, the join is the union of subsets of at most n elements (clusters, partitions, or splits, respectively). Hence, whenever it is defined, each of the rules of the previous paragraph leads to an n-polynomially computable function Nevertheless, the computation of $c_W(\pi)$ by polynomial formula involves $|W| - 1$ join operations and $\Sigma_{W \in W} (|W| - 1)$ meet operations, a number that may be exponential in v, as it is for the majority rule. Fortunately, there is another way to compute any quota rule consensus in a time polynomial both in v and n (Monjardet 1980).

Let $\pi = (T_1, \ldots, T_v) \in \mathbf{T}_n^V$ be a profile of, say, n-trees. Then clusters in $c_w(\pi)$ are exactly those present in at least w of the T_i's This property

follows from distributivity laws and from the fact, that will be emphasized in the next paragraph, that clusters correspond to join-irreducible elements in $\mathbf{T}_n$ For each cluster C present in n-tree T_i, for some i in V, one has to enumerate the n-trees of the profile that contain C Then, the total number of occurrences to examine is $O(v^2n)$. The situation is the same for ordered trees and Buneman trees, with partitions and splits instead of clusters.

Quota rule polynomials in $\mathbf{P}_n$ or $\mathbf{U}_{L,n}$ constitute a more complex case. First, we begin with partitions, indeed equivalence relations will be considered instead of them. We bring the case of equivalence relations closer to the previous distributive ones by using another expression of lattice polynomials of equivalence relations

$$c_{\mathbf{W}}(\pi) = \bigvee_{W \in \mathbf{W}} \bigwedge_{i \in W} R_i = \psi\left(\bigcup_{W \in \mathbf{W}} \bigcap_{i \in W} R_i \right),$$

where $\pi = (R_1, \ldots, R_v)$ is a profile of equivalence relations on X and ψ is the transitive closure on binary relations Because of ditributivity of the boolean lattice of all binary relations on X, the pairs in

$$R' = \bigcup_{W \subseteq V, |W| \geqslant w} \bigcap_{i \in W} R_i$$

are those present in at least w of the R_i's. Then, the computation of quota rule polynomials may be done in two steps The first step is the recognition of the pairs in R' The second step is the determination of the transitive closure of R', $c_w(\pi) = \psi(R')$; this second step requires $O(max(n,|R'|))$ time (see for instance Gibbons 1985). Notice that this method does not work for polynomials in the dual form

Under duality, the situation for ultrametrics is similar First, for each pair (x,y) one finds the w-th greatest value of the numbers $U_i(x,y)$, $i \in V$ This is a well-known v-linear searching problem (Knuth 1973). Then, one uses an efficient single-linkage algorithm on the resulting valued relation in order to compute the dual w-quota rule polynomial, corresponding to the w-quota rule polynomial for dendrograms, in time polynomial in v and n Single linkage is the same as the ultrametric anticlosure which, under duality, generalizes transitive closure to valued relations, its computation may be done with $O(n^2)$ comparisons of numbers (Hartigan 1975).

4.4 An Ordinal Framework for the Arrow-like Approach

The main illustrations of axiomatic characterizations in consensus studies are given by Arrow's famous theorem (1951) and by many results of the same type in various domains Classification takes now a prominent place among them.

One may distinguish three common elements in the general design of these results. the complex objects that have to be aggregated are described by collections of elementary entities, the main condition is an axiom of independence or neutrality that turns out to be strong enough to determine almost the general form of compatible consensus functions, one or several other subsidiary axioms eliminate some *a priori* undesirable functions.

For a unified ordinal presentation of several results, we first consider a class of typically ordinal elementary entities, whose existence is a consequence of latticial structure Let S be a meet semilattice. An element t in S is said to be *join-irreducible* if t cannot be expressed as the join of two elements s, s' in S, both distinct from t $t = s \vee s'$ implies that $s = t$ or $s' = t$. The property given in Section 2 2, that any element of S is the join of a collection of join-irreducible elements, follows from this definition Meet-irreducible elements are defined dually. In Section 2.2, join-irreducible (called elementary) and meet-irreducible (called coelementary) ultrametrics have been described Let J be the set of all join-irreducible elements of S Any element $s \in S$ may be described by a collection of elementary facts $t \leqslant s$ holds or does not hold for each $t \in J$

Table 3 describes join-irreducible elements t and elementary facts $t \leqslant s$ for classification lattices or semilattices (under duality in the case of ultrametrics) Our main Arrow-like axioms are related to these join-irreducible elements and elementary facts In this ordinal framework, the classical *independence of irrelevant alternatives* axiom for a consensus function c $S^V \to S$ becomes

(1) for any $t \in J$, $\pi = (s_1, \quad , s_v) \in S^V$, $\pi' = (s'_1, \quad . . , s'_v) \in S^V$, if $\{i \quad t \leqslant s_i\} = \{i \quad t \leqslant s'_i\}$, then $t \leqslant c(\pi)$ if and only if $t \leqslant c(\pi')$

The meaning of this axiom is that c is decomposable into elementary consensus functions, each of them related to an element of J A stronger axiom is the *neutrality* axiom (N) which states that elementary consensus functions are all the same

(N) for any $t, t' \in J$ and $\pi, \pi' \in S^V$, if $\{i \quad t \leqslant s_i\} = \{i \quad t' \leqslant s'_i\}$, then $t \leqslant c(\pi)$ if and only if $t' \leqslant c(\pi')$

Our strongest main axiom is the axiom (NM) of *monotonic neutrality*

(NM) for any $t, t' \in J$ and $\pi, \pi' \in S^V$, if $\{i \quad t \leqslant s_i\} \subseteq \{i \quad t' \leqslant s'_i\}$, then $t \leqslant c(\pi)$ implies $t' \leqslant c(\pi')$.

Finally, one has to choose some subsidiary axioms Arrow-like results that differ only on such axioms are generally almost identical. In this study, we shall take a well-known subsidiary axiom, not the weakest possible one in many situations, the *Pareto* axiom (P)·

(P) for any $\pi \in S^V$, $c(\pi) \geqslant \bigwedge_{i \in V} s_i$.

Table 3

Set	Join-irreducible elements t	fact $t \leqslant s$, $t \in J$, $s \in S$
$\mathbf{P}_n$	partitions P_{xy} with a unique cluster $\{x,y\}$ and singletons otherwise	x and y are in the same cluster in P
$\mathbf{U}_{L\,n}$	meet-irreducible ultrametrics $U' = U(\lambda_o, xy)$ such that $U'(x,y) = \lambda_o$ for some $\lambda_o \in L$, $x,y \in X$, $x \neq y$, and $U'(x',y') = 1$ otherwise	$U(x,y) \leqslant \lambda_o$
$\mathbf{D}_{L\,n}$	dendrograms $D' = D(\lambda_o, xy)$ such that $D'(\lambda) = P_o$ if $O \leqslant \lambda, < \lambda_o$ $D'(\lambda) = P_{xy}$ if $\lambda_o \leqslant \lambda < 1$ $D'(1) = \{X\}$	$\lambda \geqslant \lambda_o$ implies that x and y are in the same cluster in $D(\lambda)$
$\mathbf{T}_n$	n-trees T_C with a unique nontrivial cluster C	$C \in T$
$\mathbf{O}_n$	ordered trees associated with chains of partitions containing a unique nontrivial partition P	P is a partition in the chain corresponding to the ordered tree O
$\mathbf{B}_n$	Buneman trees defined by a unique split σ	σ is a split of B

Indeed (P) is an axiom of local unanimity. It is equivalent to the condition that for any t in J, if $t \leqslant s_i$ holds for all the s_i's in the profile, then $t \leqslant c(\pi)$ also holds for the consensus $c(\pi)$.

4.5 Axiomatic Characterizations of Polynomial Consensus Rules

The earliest Arrow-like result in cluster analysis is Mirkin's (1975) characterization of oligarchic rules on partitions. An improved version of his result is the following.

Theorem 4.1 *A consensus function c on partitions is either oligarchic or else the constant function* $c(\pi) = \{X\}$ *if and only if c satisfies axioms (I) and (P).*

This result may be obtained as a consequence of a more general one concerning dendrograms In this case, axiom (I) is equivalent to the threshold binariness axiom on ultrametric consensus functions proposed by Leclerc (1984) It leads to threshold oligarchic rules which do not correspond to polynomials of dendrograms or ultrametrics Consider an arbitrary antitone mapping $W . L \to \mathbf{P}(V)$ (i e, $\lambda \geqslant \lambda'$ implies $W(\lambda) \subseteq W(\lambda')$) and define for any π in $\mathbf{D}^V_{L,n}$ and λ in L

$$c(\pi)\ (\lambda) = \bigwedge_{i \in W(\lambda)} D_i\ (\lambda)\ .$$

By a usual convention, if $W(\lambda) = \varnothing$ then $c(\pi)\ (\lambda) = \{X\}$ If there is some λ in L such that $W(\lambda) \neq \varnothing$, the consensus function c is called *threshold oligarchic* otherwise $c(\pi)$ is the maximum of $\mathbf{D}_{L,n}$ so that $c(\pi)\ (\lambda) = \{X\}$ for all λ From Leclerc's Theorem 8.2, one obtains the following result

Theorem 4.2 *A consensus function c on dendrograms is either threshold oligarchic or else gives always the maximum dendrogram if and only is c satisfies axioms (I) and (P).*

In order to obtain oligarchic consensus functions, one must introduce another axiom that ensures W is constant For instance, the *flatness* axiom (F) specifies a kind of neutrality with regard to L

(F) for any $\pi, \pi' \in \mathbf{D}^V_{L,n}$ and $\lambda, \lambda' \in L$,
if $D_i(\lambda) = D'_i(\lambda')$ for all $i \in V$, then $c(\pi)(\lambda) = c(\pi')(\lambda')$

Corollary 4.3 *A consensus function c on dendrograms is either oligarchic or else gives always the maximum dendrogram if and only if c satisfies axioms (I), (P) and (F).*

Recently, Neumann and Norton (1985) have obtained a dual result on partitions, with an extension to dendrograms This characterization may be restated in a form similar, under duality, to that of Theorem 4 1. Consider $\mathbf{P}_n$ as a join-semilattice and define dual oligarchic rules as those corresponding to joins Let (P') and (I') be the axioms derived from (P) and (I) by dualization. In (I') elementary facts have the form $t \geqslant s$, where s is in S and t is in J', the set of meet-irreducible elements in S Meet-irreducible partitions are those with exactly two clusters. Then, from Neumann and Norton's result, we have the following

Theorem 4.4 *A consensus function c on partitions is either dual oligarchic or else the constant function* $c(\pi) = P_0$ *if and only if c satisfies axioms (I') and (P').*

In these results, it would be easy to add another condition in order to eliminate constant functions

Now we present general Arrow-like results for median semilattices In that case it is possible to have, with axiom (I) only, very different elementary consensus functions related to elements of J. Indeed, with axiom (NM) instead of (I), one gets a characterization of generalized majority rule (Barthélemy *et al.* 1984a, 1984b, Monjardet 1986), with a particularization to n-trees due to McMorris and Neumann (1983).

Theorem 4.5 *Let S be a median semilattice, different from a distributive lattice, and* $c \quad S^V \to S$ *a consensus function on S. Then c satisfies (NM) and (P) if and only if c is given by a polynomial consensus function* $c_{\mathbf{W}}$ *for some generalized majority family* $\mathbf{W}$.

Corollary 4.6 *A consensus function on n-trees corresponds to a generalized majority rule if and only if it satisfies axioms (NM) and (P).*

Similar particularizations of Theorem 4 5 may be stated in the cases of ordered trees and Buneman trees Theorem 4 5 remains true for distributive lattices with the addition that in this case one can also get the constant function c giving always the greatest element of the lattice

With some change of subsidiary axioms, it is possible to obtain a specific characterization of the majority rule polynomial m defined in Section 4 2. In the papers cited above, Theorem 4.5 is followed by a corollary of this type, using axioms of *symmetry* (VS) with regard to V and 2-*idempotence* (2ID)

(VS) c is invariant by permutation of indices in π.

(2ID) If there exists $s,s' \in S$ such that any element of π is either s or s', then $c(\pi) \in \{s,s'\}$

Corollary 4.7 *Let S be a median semilattice and $c \quad S^V \to S$ be a consensus function on S Then c is given by the majority rule if and only if it satisfies axioms (NM), (VS) and (2ID)*

It is possible to characterize majority rule on n-trees, ordered trees, or Buneman trees by particularizations of corollary 4 7

Before concluding this discussion of Arrow-like approaches, we make a brief survey of some results where main axioms other than those defined above have been used Oligarchic rules in median semilattices are characterized by replacing axiom (NM) in Theorem 4.5 by a similar one, where t may be any element of S (Barthelemy *et al* 1984a; Monjardet 1986) In several works, the main axioms impose the stability of consensus functions A consensus function $c : S^V \to S$ is called *stable* when, for any $Y \subset X$, if two profiles π and π' induce pairwise identical classifications on Y, then $c(\pi)$ and $c(\pi')$ induce again the same classification on Y Stability properties may be also defined for classification methods For instance, Régnier (1977) has pointed out the triviality of classification functions with dissimilarities on X as domain and partitions as range, when stability is assumed.

In classification or preference consensus problems, stability axioms are not as related as (I) to the corresponding ordinal structures, but they are quite natural The works on partitions or ultrametrics by Mirkin (1975), Leclerc (1984a), and Neumann and Norton (1985) are based upon stability conditions In the case of partitions, stability is equivalent to (I) In the case of dendrograms, stability may be expressed by Leclerc's binariness axiom, it is then weaker than (I) and leads to a class of consensus functions that includes threshold oligarchic ones On the other hand, in some cases, stability turns out to be stronger than (NM) and leads, as in Arrow's theorem, to dictatorial consensus rules this is the case for tree quasi-orders (McMorris and Neumann 1983), n-trees with a further neutrality condition (Neumann 1983, by theorems 1 and 2), and unrooted n-trees, a special class of Buneman trees (McMorris 1985)

An important role of order theory may be to provide general results on the consensus problem that are particularizable in cluster analysis as well as in other fields Theorem 4 5 is a good illustration of that claim Another one is a recent characterization, due to Janowitz (1985), of generalized majority rules by a property of neutrality that is close to axiom (N) This result applies to many situations, but, in its present form, does not seem adequate for consensus in lattices In the case of n-trees, the particularization of the Janowitz result corresponds to Neumann's (1983) lemma implying Corollary 4 6 above

4.6 Median Semilattices: Obtaining Median Classifications

The determination and the construction of an optimal consensus is generally a difficult problem of combinatorial optimization Restricting the

problem to the structure of median semilattices is a fortunate special case where a median may be always obtained by an algebraic formula Extending a result of Barbut (1961) on distributive lattices, Bandelt and Barthélemy (1984) have shown the following result.

Theorem 4.8 *Let S be a median semilattice. For any profile π in S^V, majority rule provides a median of π for the latticial (MPL) metric related to S. Furthermore, if v is odd, this median is unique.*

In the same paper, Bandelt and Barthélemy show this uniqueness to be a characteristic property of median semilattices. By particularization, and taking into account the construction of majority consensus as a join of majoritary join-irreducible elements (Section 4.3), one obtains a result due to Margush and McMorris (1981).

Corollary 4.9 *Let $\pi \in \mathbf{T}_n^V$ be a profile of n-trees. The set of all clusters present in more than half of the t_i's is the collections of clusters of an n-tree T that is a median of π for the MPL metric on $\mathbf{T}_n$. Moreover, if v is odd, then T is unique.*

Similar results hold for ordered trees and Buneman trees.

It is possible to determine all the n-trees that are medians of an even profile $\pi = (T_1, \ldots, T_v)$ by starting from the majority rule n-tree $m(\pi)$ and completing $M(\pi)$ by clusters present in half of the T_i's in a compatible way. Clearly, distinct maximal median n-trees may exist; the problem of finding a median n-tree with as many clusters as possible may have several solutions.

Similar constructions for median ordered trees, or Buneman trees, exist. These constructions may be derived from the straightforward extension to median semilattices of the following characterization of medians in a finite distributive lattice S (Barbut 1967, Monjardet 1980). an element $t \in S$ is a median of the profile $\pi = (s_1, \ldots, s_v) \in S^V$ for the latticial MPL distance on S if and only if it is contained between majority and dual majority rule consensus where, by distributivity laws, the latter is given by the quota rule for $w = v/2$.

$$
\begin{aligned}
m(\pi) &= \bigvee_{W \subseteq V, |W| = v/2+1} \ \bigwedge_{i \in W} s_i \leqslant t \\
&\leqslant \bigwedge_{W \subseteq V, |W| = v/2+1} \ \bigvee_{i \in W} s_i \\
&= \bigvee_{W' \subseteq V, |W'| = v/2} \ \bigwedge_{i \in W'} s_i = m'(\pi) \ .
\end{aligned}
$$

Theorems 4 5 and 4.8 establish the convergence of the three approaches of the consensus problem in median semilattices· majority rule, given by an algebraic formula, is axiomatically characterized and provides one (sometimes among others) solution of a natural optimization problem. The possible other solutions of this optimization problem (i.e , the set of all the medians) have been described Moreover, the whole median procedure can be also axiomatically characterized. Prompted by Young and Levenglick's characterization (1978) of median linear orderings, Barthélemy and McMorris (1986) give a characterization of all the median n-trees

The research of median partitions, or dendrograms, is a much more difficult problem A study by Régnier (1965) on median equivalence relations, with the symmetric difference metric, was probably the first work on consensus of classifications The determination of these median equivalence relations, probably an NP-complete problem, was transformed by Régnier into the resolution of a linear integer programming problem. Since this work, significant computational progress has been made (Marcotorchino and Michaud 1982), based especially on the use of a performing linear programming code

References

ADAMS, E N , III (1972), "Consensus Techniques and the Comparison of Taxonomic Trees," *Systematic Zoology, 21*, 390-397

AIGNER, M (1979), *Combinatorial Theory*, Berlin: Springer-Verlag

ARABIE, P , and BOORMAN, S A (1973), "Multidimensional Scaling of Measures of Distances Between Partitions," *Journal of Mathematical Psychology, 17*, 31-63

ARROW, K J (1951), *Social Choice and Individual Values*, New York: Wiley

BANDELT, H J , and BARTHELEMY, J P (1984), "Medians in Median Graphs," *Discrete Applied Mathematics, 8*, 131-142

BARBUT, M (1961), "Médiane, distributivité, éloignements," repr (1980), *Mathématiques et Sciences humaines, 70*, 5-31

BARBUT, M , and MONJARDET, B (1970), *Ordre et Classification, Algèbre et Combinatoire*, Paris: Hachette

BARTHELEMY, J P (1978), "Remarques sur les propriétés métriques des ensembles ordonnés," *Mathématiques et Sciences humaines, 61*, 39-60

BARTHELEMY, J P (1979a), "Caractérisations axiomatiques de la distance de la différence symétrique entre des relations binaires," *Mathématiques et Sciences humaines, 67*, 85-113

BARTHELEMY, J P (1979b), "Propriétés métriques des ensembles ordonnés Comparaison et agrégation des relations binaires," Thése, Université de Besançon

BARTHELEMY, J P (1985), "From Copair Hypergraphs to Median Graphs with Latent Vertices," submitted

BARTHELEMY, J P , and LUONG, X (1986), "Mathématique, algorithmique et histoire des représentations arborées," submitted

BARTHELEMY, J P , and MCMORRIS, F R (1986), "The Median Procedure for n-Trees," *Journal of Classification, 3*,329-334

BARTHELEMY, J P , LECLERC, B , and MONJARDET, B (1984a), "Ensembles ordonnés et taxonomie mathématique," in *Orders Description and Roles*, eds M Pouzet and D Richard, *Annals of Discrete Mathematics, 23*, Amsterdam: North-Holland, 523-548

BARTHELEMY, J P , LECLERC, B , and MONJARDET, B (1984b), "Quelques aspects du consensus en classification," in *Data Analysis and Informatics III*, eds E Diday *et al*, Amsterdam: North-Holland, 307-316

BARTHELEMY, J P , and MONJARDET, B (1981), "The Median Procedure in Cluster Analysis and Social Choice Theory," *Mathematical Social Sciences, 1*, 235-268

BENZECRI, J P (1967), "Description mathématique des classifications," repr (1973) in *L'analyse des données, la Taxinomie*, J P Benzécri *et coll* , Paris: Dunod, 119-152

BIRKHOFF, G (1967), *Lattice Theory*, 3rd ed , Providence: American Mathematical Society

BLYTH, T S , and JANOWITZ, M F (1972), *Residuation Theory*, Oxford: Pergamon Press

BOORMAN, S A , and ARABIE, P (1972), "Structural Measures and the Method of Sorting," in *Multidimensional Scaling, Vol 1, Theory and Applications in the Behavioral Sciences*, eds R N Shepard, A K Romney and S B Nerlove, New York: Seminar Press, 226-249

BOORMAN, S A , and OLIVIER, D C (1973), "Metrics on Spaces of Finite Trees," *Journal of Mathematical Psychology, 10*, 26-59

BORDA, J C (1784), *Mémoire sur les élections au scrutin, Histoire de l'Académie Royale des Sciences pour 1781*, Paris

BORDES, G (1976), "Métriques bornées définies par des valuations sur un demi-treillis," *Mathématiques et Sciences humaines, 56*, 89-96

BUNEMAN, P (1971), "The Recovery of Trees from Measures of Dissimilarity," in *Mathematics in Archaeological and Historical Sciences*, eds F R Hodson, D G Kendall and P Tautu, Edinburgh: Edinburgh University Press, 387-395

COMTET, L (1970), "Sur le quatrième problème et les nombres de Schröder," *Comptes-Rendus Académie des Sciences de Paris, A-271*, 913-916

COMTET, L (1974), *Advanced Combinatorics*, Dordrecht: Reidel

CONDORCET, M J A (1785), *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris

DAY, W H E (1981), "The Complexity of Computing Metric Distances Between Partitions," *Mathematical Social Sciences, 1*, 269-287

DAY, W H E (1983a), "The Role of Complexity in Comparing Classifications," *Mathematical Biosciences, 66*, 97-114

DAY, W H E (1983b), "Computationally Difficult Parsimony Problems in Phylogenetic Systematics," *Journal of Theoretical Biology, 103*, 429-438

DAY, W H E (1985), "Optimal Algorithms for Comparing Trees with Labelled Leaves," *Journal of Classification, 2*, 7-28

DAY, W H E , and FAITH, D P (1986), "A Model in Partial Orders for Comparing Objects by Dualistic Measures," *Mathematical Biosciences, 8*, 179-192

DAY, W H E , and MCMORRIS, F R (1985), "A Formalization of Consensus Index Methods," *Bulletin of Mathematical Biology, 47*, 215-229

DAY, W H E , and WELLS, R S (1984), "Extremes in the Complexity of Computing Metric Distances Between Partitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*, 69-73

DIJKSTRA, E W (1959), "A Note on Two Problems in Connection with Graphs," *Numerische Mathematik, 1*, 269-271

DOBINSKY, T (1877), "Summirung der Reihe $\Sigma n^m/n!$ fur $m = 1,2,3,4,5,$," *Grunert Archiv, 61*, 333-336

FINDEN, C R , and GORDON, A D (1985), "Obtaining Common Pruned Trees," *Journal of Classification, 2*, 255-276

FLAMENT, C (1963), *Application of Graph Theory to Group Structure*, New York: Prentice Hall

FOULDS, L R , and ROBINSON, R W (1980), "Determining the Asymptotic Number of Phylogenetic Trees," in *Combinatorial Mathematics VII, Lecture Notes in Mathematics 829*, Berlin Springer-Verlag, 110-126

GIBBONS, A (1985), *Algorithmic Graph Theory*, Cambridge (UK): Cambridge University Press

GOWER, J C , and ROSS, G J S (1969), "Minimum Spanning Tree and Single Linkage Cluster Analysis," *Applied Statistics, 18*, 54-64

GRATZER, G (1978), *General Lattice Theory*, Basel: Birkhaüser Verlag

GUENOCHE, A (1986), "Etude comparative de cinq algorithmes d'approximation des dissimilarités par des arbres à distances additives," *Mathématiques et Sciences humaines*, to appear

GUILBAUD, G Th (1952), "Les théories de l'intérêt général et le problème logique de l'agrégation," *Economie Appliquée, 5*, 501-551, repr (1968) in *Eléments de la Théorie des Jeux*, Paris: Dunod

HARARY, F (1969), *Graph Theory*, Reading, Mass Addison-Wesley

HARDING, E S (1971), "The Probability of Rooted Tree-Shapes Generated by Random Bifurcations," *Advances in Applied Probability, 3*, 44-77

HARTIGAN, J A (1975), *Clustering Algorithms*, New York Wiley

HUBERT, L (1977), "Data Analysis Implications of Some Concepts Related to the Cuts of a Graph," *Journal of Mathematical Psychology, 15*, 199-208

HUBERT, L , and BAKER, F B (1977), "The Comparison and Fitting of Given Classification Schemes," *Journal of Mathematical Psychology, 16*, 233-255

JANOWITZ, M F (1978), "An Order Theoretic Model for Cluster Analysis," *SIAM Journal on Applied Mathematics, 34*, 55-72

JANOWITZ, M F (1985), "A Generalized Setting for Consensus Functions," University of Massachusetts, Department of Mathematics and Statistics

JARDINE, N , and SIBSON, R (1971), *Mathematical Taxonomy*, London: Wiley

JOHNSON, S C (1967), "Hierarchical Clustering Schemes," *Psychometrika, 32*, 241-254

KEMENY, J G (1959), "Mathematics without Numbers," *Daedalus, 88*, 575-591

KNUTH, D E (1973), *The Art of Computer Programming, Vol 3, Sorting and Searching*, Reading, Mass: Addison-Wesley

LECLERC, B (1979), "Semi-modularité des treillis d'ultramétriques," *Comptes-Rendus Académie des Sciences de Paris, A-288*, 575-577

LECLERC, B (1981), "Description combinatoire des ultramétriques," *Mathématiques et Sciences humaines, 73*, 5-37

LECLERC, B (1984a), "Efficient and Binary Consensus Functions on Transitively Valued Relations," *Mathematical Social Sciences, 8*, 45-61

LECLERC, B (1984b), "Indices compatibles avec une structure de treillis et fermeture résiduelle," Technical Report P 011, Centre d'Analyse et de Mathématique Sociales

LECLERC, B (1985a), "Les hiérarchies de parties et leurs demi-treillis," *Mathématiques et Sciences humaines, 89*, 5-34

LECLERC, B (1985b), "La comparaison des hiérarchies: indices et métriques," *Mathématiques et Sciences humaines, 92*, 5-40

MARCOTORCHINO, F , and MICHAUD, P (1982), "Agrégation de similarités en classification automatique," *Revue de Statistique Appliquée, 30*, 21-44

MARGUSH, T (1982), "Distances Between Trees," *Discrete Applied Mathematics, 4*, 281-290

MARGUSH, T , and MCMORRIS, F R (1981), "Consensus n-Trees," *Bulletin of Mathematical Biology, 43*, 239-244

MCMORRIS, F R (1985), "Axioms for Consensus Functions on Undirected Phylogenetic Trees," *Mathematical Biosciences, 74*, 17-21

MCMORRIS F R , and NEUMANN, D A (1983), "Consensus Functions on Trees," *Mathematical Social Sciences, 4*, 131-136

MIRKIN, B G (1975), "On the Problem of Reconciling Partitions," in *Quantitative Sociology, International Perspectives on Mathematical and Statistical Modelling*, New York: Academic Press, 441-449

MIRKIN, B G , and CHERNYI, L B (1970), "On Measurement of Distance Between Partitions of a Finite Set of Units," *Automation and Remote Control, 31*, 786-792

MONJARDET, B (1976), "Caractérisations métriques des ensembles ordonnés semi-modulaires," *Mathématiques et Sciences humaines, 56*, 77-87

MONJARDET, B (1980), "Théorie et applications de la médiane dans les treillis distributifs finis," *Annals of Discrete Mathematics, 9*, 87-91

MONJARDET, B (1981), "Metrics on Partially Ordered Sets, A Survey," *Discrete Mathematics, 35*, 173-184

MONJARDET, B (1986), "Characterizations of Polynomial Functions in Median Semilattices," Technical Report, Centre d'Analyse et de Mathématique Sociales

NEUMANN, D A (1983), "Faithful Consensus Methods for n-Trees," *Mathematical Biosciences, 63*, 271-287

NEUMANN, D A , and NORTON, V (1985), "Consensus of Partitions with Applications to Other Structures," Technical Report n° 85-20, Bowling Green State University, Department of Mathematics and Statistics

PUDLAK, P , and TUMA, J (1980), "Every Finite Lattice Can be Embedded in the Lattice of all Equivalences Over a Finite Set," *Algebra Universalis, 10*, 74-95

REGNIER, S (1965), "Sur quelques aspects mathématiques des problémes de classification automatique," *ICC Bulletin, 4*, 175-191, repr (1983), *Mathématiques et Sciences humaines, 82*, 13-29

REGNIER, S (1977), "Stabilité d'un opérateur de classification," *Mathématiques et Sciences humaines, 60*, 21-30

ROBINSON, D F (1971), "Comparison of Labelled Trees with Valency Three," *Journal of Combinatorial Theory, 11*, 105-119

ROBINSON, D F , and FOULDS, L R (1981), "Comparison of Phylogenetic Trees," *Mathematical Biosciences, 53*, 131-147

ROHLF, F J (1982), "Consensus Indices for Comparing Classifications," *Mathematical Biosciences, 59*, 131-144

SCHADER, M (!980), "Hierarchical Analysis: Classification with Ordinal Object Dissimilarities," *Metrika, 27*, 127-132

SHOLANDER, M (1954), "Medians, Lattices and Trees," *Proceedings of the American Mathematical Society, 5*, 808-812

STINEBRICKNER, R (1984), "s-Consensus Trees and Indices," *Bulletin of Mathematical biology, 46*, 923-935

YOUNG, H P , and LEVENGLICK, A (1978), "A Consistent Extension of Condorcet's Election Principle," *SIAM Journal on Applied Mathematics, 35*, 285-300