COMMUNITY ECOLOGY 1(1): 33-41, 2000
1585-8553 © Akadémiai Kiadó, Budapest

Additive trees in the analysis of community data

J. Podani[1,2], P. Csontos[1] and J. Tamás[1,3]

[1] *Department of Plant Taxonomy and Ecology, Eötvös University, Ludovika tér 2, H-1083 Budapest, Hungary. Corresponding author. E-mail: podani@ludens.elte.hu*
[2] *Present address: Collegium Budapest, Institute for Advanced Study, Szentháromság u. 2, H-1014 Budapest, Hungary.*
[3] *Present address: Research Institute for Botany and Ecology, Hungarian Academy of Sciences, H-2163 Vácrátót, Hungary.*

Keywords: Classification, Dendrograms, Four-point metrics, Grasslands, Neighbor joining, Succession, Syntaxonomy, Ultrametrics.

Abstract: The paper advocates a more extensive use of additive trees in community ecology. When the distance/dissimilarity coefficient is selected carefully, these trees can illuminate structural aspects that are not obvious otherwise. In particular, starting from squared distances based on presence/absence data, the resulting trees approximate relationships in species richness, a feature not available through other graphical techniques. The construction of additive trees is illustrated by three actual examples, representing different circumstances in the analysis of grassland community data.

Abbreviations: NJ – Neighbor Joining, PCA – Principal Components Analysis, UPGMA – Unweighted Pair Group Method Using Arithmetic Averages.

Nomenclature: Simon (1992) for both plants and syntaxa.

A critique of 'standardized methodology'

Recent applications of classification and ordination methodology to community analysis are dominated by a narrow selection of procedures. A careful scrutiny of relevant papers would certainly reveal that less than a dozen methods have been used in a vast majority of published work. Suggested advantages of this 'standardized methodology' include the following:

- the properties, the relative merits and potential disadvantages of the preferred procedures are assumed to be widely known;
- when the same method and only a few software packages are used throughout the world, the results are considered comparable and more reliable; and
- the information conveyed by the results is easily understandable, because everyone speaks the 'same language'.

On the other hand, there are some serious risks inherent in such an attitude:

- an unrevealed theoretical misconception or a software bug will 'infest' the results worldwide;
- users become too comfortable with their beloved techniques, and apply them uncritically under all circumstances; and
- fashion-like preference in favor of particular methods hinders the development of new methodology and is an obstacle to scientific advancement.

We feel that community ecologists have uncritically applied recently developed numerical methods, accepting the pros while denying the cons. This is surprising because the situation was very different only a few decades ago: ecologists raised a number of original methodological questions and opened new areas completely unexplored even by mathematicians at that time. The early de-

velopment of divisive classificatory techniques, for example, is rooted in numerical ecology (Goodall 1953, Williams and Lambert 1959) and has contributed implicitly to the recent proliferation of data mining algorithms (e.g., Michalski et al. 1998, Westphal and Blaxton 1998). Also, the combinatorial formula developed by Lance and Williams (1967) for a family of cluster analysis procedures has attracted the attention of many mathematicians interested in the complexity theory of algorithms (e.g., Day and Edelsbrunner 1984). The appearance of sum of squares clustering in ecology (Orlóci 1967) was another significant and original contribution to the algorithmic "explosion" of the late 1960's.

Our view is clear on this point: data analysis methods potentially useful in community studies should be continuously refreshed and reevaluated in a critical manner. This does not necessarily require the development of entirely new techniques. Indeed, there is already an enormous collection of different methods suggested in various disciplines both outside and inside biology, and we only have to "rediscover" and adapt them for use under modified circumstances. We need to look for existing and relevant methodology that can illuminate questions raised in the field of community ecology in a nonstandard way.

In this paper, we focus on a graph theoretical procedure that has thus far received only isolated applications to ecological problems. We begin with a brief overview of tree graphs, with special emphasis on their ultrametric and additive properties. This is followed by some technical details concerning the computation of additive trees. The paper concludes with the analysis of three different data sets to illustrate the utility of this approach in community analysis.

Ultrametric and additive trees

Community ecology is concerned with many types of graphs. Food webs, for example, are illustrated most effectively and most commonly by graph theoretical means. *Minimum spanning trees* also appear occasionally, applied mostly for the clarification of ordination arrangements (Digby and Kempton 1987). In these graphs, the number of vertices (nodes) equals the number of objects, m, while the number of branches is m-1. Hierarchical classifications of communities are more commonly portrayed by some special tree graphs called *dendrograms*, produced by conventional agglomerative or divisive clustering procedures. Dendrograms have m terminal vertices, the objects classified, and – if fully resolved – m-1 interior vertices producing a hierarchical structure. Weighted dendrograms (cf. Podani 2000) possess the very important property of being *ultrametric*. Ultrametricity implies that for any triple of objects, i, j and k, two of the three pairwise distances are identical and no smaller than the third distance (Figure 1ab). Formally, any three objects can always be relabeled to satisfy the following inequality

$$d_{ij} \leq \max \{ d_{ik}, d_{jk} \}. \quad (1)$$

In other words, the fusion level for two pairs is never smaller than for the third pair. (This is sometimes violated by the median and centroid methods when they produce 'reversals' in the dendrogram.) The ultrametric condition is very strong and, therefore, the ultrametric distances implied by a dendrogram rarely fit closely the original distances from which the dendrogram is constructed. In fact, the distortion can be exceedingly high suggesting that even though dendrograms may reflect hierarchical classifications adequately, they are not necessarily good graphical representations of distance structures.

Currently, the most active area of biological data analysis has been phylogenetic reconstruction, radically changing our views on the evolution of many groups of living organisms. These studies rely upon two basically different sources of information. Conventional taxonomic characters ranging from ultrastructure morphology to phenology represent one group and protein and nucleic acid sequences represent the other. Evolutionary systematics utilizes two strategies of tree-building techniques. More often, parsimony analysis is applied directly to the data to minimize the character changes along the tree such that certain requirements are also satisfied (e.g., minimum homoplasy). These procedures cover the domain of classical cladistic methodology. Phylogenetic reconstruction, however, may also be launched from distances calculated from both kinds of data, or obtained directly from immunological or serological experiments (Swofford and Olsen 1990, Nei 1996, Page and Holmes 1998). The ultrametric property that all objects in a dendrogram are equidistant from the root is a major obstacle to interpreting dendrograms as reconstructions of phylogenetic pathways. Therefore, clustering procedures are deemed irrelevant in phylogenetic reconstruction, with the exception of the group average method (UPGMA, Sneath and Sokal 1973) having limited applicability to cases when the molecular clock ("constant rate of evolution") is assumed (Page and Holmes 1998). Instead, distance-based phylogenetic analyses focus on trees in which inter-object distances are not ultrametric but are as close to the input distances as possible. In such trees, the original distance between any two objects is approximated by the sum of branch lengths along the path between these objects in the graph, hence the term *additive trees*. If the distances are accepted as reasonable estimators of evolutionary distances, then the additive trees pro-

vide a more faithful phylogenetic reconstruction than ultrametric trees.

Additivity of distances implies that the condition of being a *four-point metric* is satisfied. This is weaker than the ultrametric property: in fact all ultrametrics are four point metrics at the same time. In order to examine the four point additivity condition, consider any four objects, labeled *h, i, j*, and *k* in the graph. Then, this relationship is expressed by the inequality

$$d_{hi} + d_{jk} \leq \max \{ d_{hj}+d_{ik}, d_{hk}+d_{ij} \} \quad (2)$$

(Buneman 1971, Patrinos and Hakimi 1972, Sattath and Tversky 1977, Shepard 1980, de Soete 1983). Figure 1c illustrates that the six interpoint distances may be expressed according to five components, *a, b, c, d* and *e.* When c=0, we have a star-tree (Fig. 1d). Furthermore, if the four points are considered as tips of a tetrahedron, with the edge lengths proportional to the respective distances, then the sums of opposite edge lengths provide an isosceles triangle. Comparison of inequalities (1) and (2) reveals immediately that in fact these pairwise sums do obey the ultrametric condition.

Approximation of distances by additive trees

Actual distances for a set of ecological objects extremely rarely if ever satisfy condition (1). The chance that the distances meet the requirements of a four-point metric is of course higher, because this is a much weaker condition. Nevertheless, distances coming from actual studies almost always violate this condition as well, so that additive trees can only be approximations to the real distance structure.

The suggestion to approximate distances by additive trees was raised first in the psychometric literature (Carroll and Chang 1976, Cunningham 1978). Sattath and Tversky (1977) proposed a fairly complex algorithm to maximize the fit of within-graph distances to an input matrix. Further algorithms and programs were developed by Corter (1982) and de Soete (1983, 1988). In phylogenetic analysis, the neighbor joining (NJ) technique proposed by Saitou and Nei (1987) has received more attention, especially in the past five years (see also Nei 1996). The principle of this method is to find neighbors sequentially such that the total length of the tree is minimized. The observation that the Sattath-Tversky and the Saitou-Nei algorithms often give identical or very similar results has been confirmed on theoretical grounds by Gascuel (1994). Thus, the computationally much more efficient NJ algorithm is recommended in practice, especially if the number of points is large. In this study, we use the NJ algorithm to generate additive trees for distance matrices,

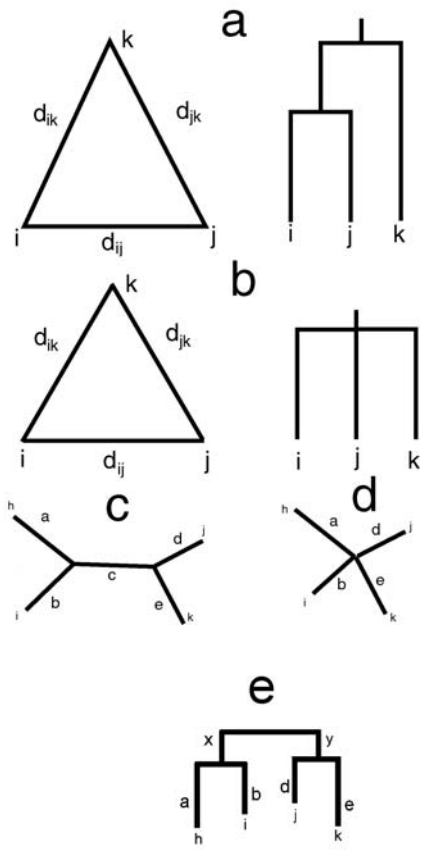

Figure 1. Ultrametrics and additive metrics. **a:** isosceles triangle and the associated dendrogram illustrating the case $d_{ik} = d_{jk} > d_{ij}$, **b:** equilateral triangle and the associated dendrogram for $d_{ik} = d_{jk} = d_{ij}$, **c:** unrooted additive tree for four points, *h, i, j* and *k*, **d:** equality in formula (2) holds when $c = 0$, **e:** rooted version of the tree in Fig.1**c**, note that $x+y=c$.

using the SYNTAX 5.1 program package (Podani 1997). The fit of within-tree distances to the input distances is measured by the matrix correlation, *r*, (Sneath and Sokal 1973) as computed by the same software.

Rooting

It has to be pointed out that additive trees are not rooted *a priori* and therefore do not imply classifications directly. Given *m* points, an additive tree has *m*-2 interior nodes, and *m* external nodes. In phylogenetic systematics, an additive tree can only be used as an hypothesized summary of evolutionary relationships if it is rooted by some external criterion, for which no general rules apply, however. Two rooting options deserve our attention here:

- *Outgroup rooting* is commonly used in phylogenetic studies and is achieved if one of the *m* objects is considered as the sister group of the common ances-

tor of all the other *m*-1 objects. The tree is rooted on the branch connecting this special object, the outgroup, with the nearest interior node.

- In *midpoint rooting*, the two nodes for which the path is the longest in the graph are selected, and the root is placed right halfway between them. Figure 1d illustrates the midpoint-rooted version of the tree in Figure 1c.

Positioning the root is arbitrary, so that one must give serious consideration to why preference is given to any particular method. Arbitrariness implies that an additive tree cannot be used as a starting point to define classifications in the same way as dendrograms. Nevertheless, as we shall demonstrate, it is worthwhile to examine how an additive tree relates to some classification obtained by other numerical or traditional methods.

Materials and methods

Three data sets gathered from rock grassland communities serve as illustrative examples. These data were chosen so as to show the performance of additive trees under different circumstances. In example 1, a relatively small geographic area is surveyed in one year, with a relatively large sample size. Example 2 demonstrates the utility of the method to examine revegetation processes at a similar spatial scale, whereas the last example relates to a broader analysis of rock grassland types at the syntaxonomical level of community types (or "associations" in the terminology of the Zürich-Montpellier school of phytosociology).

Example 1. Dolomite grasslands of Sas-hill. A total of 80 vegetational quadrats (objects) represent a sample of dolomite grasslands of Sas-hill, Budapest, Hungary, at 200-260 m above sea level. The plots of size 4 by 4 m^2 are described in terms of the presence/absence of 123 vascular plant species. Sampling is non-random, as quadrats were placed to avoid overlap with excessive areas showing different degrees of disturbance (e.g., invasion by the horticultural shrub, *Syringa vulgaris*). The vegetation of the area has been roughly divided into three groups, one corresponding to open vegetation of steep, south-facing slopes (type A), the second representing a more closed hilltop vegetation (type B), and the third comprising a completely closed, species rich community on north-facing slopes (type C; for more details, see Podani 1985, 1998). In a sense, the plots of this area may be considered as members in a chronosequence, starting from almost bare rocky ground with very sparse vegetation and ending with complete plant cover on a 20 cm thick rendzina topsoil.

Example 2. Post-fire recolonization. This data set contains information on post-fire successional processes. The study area is located on Zsíros-hill, Budai Mts, at an elevation of 390-410 m above sea level. The original vegetation of the area is rock grassland community which was replaced by *Pinus nigra* plantations around 1950. In 1993, these plantations were completely destroyed by fire, allowing the possibility to evaluate natural regeneration of the dolomite vegetation. Permanent plots 2 x $4m^2$ in size were located in the burnt area immediately after the fire, five on the south-facing and five on the adjacent north-facing slope. The sample plots were investigated for the presence of vascular plants for five consecutive years after the fire, thus yielding a total of 50 objects. More information on the study area is presented in Tamás and Csontos (1998).

Example 3. Grassland syntaxonomy. A large phytosociological data set describing different grassland communities in Hungary is used to contrast earlier findings in numerical syntaxonomy with the present graph theoretical analysis. The data table comprises presence/absence data for 130 phytosociological relevés, 127 of them collected by Zólyomi between 1930 and 1950 (see Török and Zólyomi 1998). Quadrat size was variable, the majority being 16 or 25 m^2, which is admittedly a possible source of data heterogeneity and is not recommended for numerical analysis in general. In this study, however, the additive tree method is used to test hypotheses raised and investigated earlier based on the very same sampling strategy, so differences between plot size need not concern us here. Following the guidelines of the Zürich-Montpellier school, these communities have been classified into five "associations", two of them having two geographical variants:

1a Open dolomite rock grassland (*Seseleo leuco- spermi-Festucetum pallentis*) in the Budai Mts (*SF-A*);

1b The same association from the Vértes – Bakony – Keszthelyi Mts Range (*SF-B*);

2a Closed dolomite rock grassland (*Festuco pallenti-Brometum pannonici)* in the Budai Mts (*FB-A*);

2b The same association from the Vértes – Bakony – Keszthelyi Mts Range (*FB-B*);

3 Closed mountain-grass (*Sesleria*) community (*Seslerietum sadlerianae*) from the Budai Mts (*Ss*);

4 Carpathian limestone grassland (*Campanulo divergentiformis-Festucetum pallentis)* from the Bükk Mts (*CF*); and

5 Mountain-grass community *(Seslerietum heuflerianae)* from the Bükk Mts (*Sh*).

Regarding the syntaxonomic nomenclature of these communities, their localities and the classification and ordination analyses of relevés the reader should consult Török and Zólyomi (1998).

Distances and the method of rooting

Neighbor joining analysis was performed on the squared Euclidean distances computed between the quadrats for all the three examples. In the first two example data sets, a completely empty quadrat (i.e., no plant cover) was used as an outgroup. This corresponds to the precolonization stage. The advantage of using squared distances in the presence/absence case is that each pairwise resemblance value is simply the number of species in which the two sites differ. If an empty site is included in the graph, tree shape and the distances read from the branches of the tree will inform the reader directly about the species richness of the whole study region, and will also allow comparative evaluation along different lineages in the diagram. This empty site represents a successional stage with bare rock only, which is not merely hypothetical because the study areas have fairly large rocky surfaces still uninhabited by vascular plants (example 1). Using the empty site as an outgroup, regeneration changes can be contrasted with the plantless, immediate post-fire condition (example 2). In the third example, the tree was rooted using the midpoint method, because the choice of an empty site as an outgroup were less logical in this case.

Results

Grassland succession in Sas-hill

The tree suggests the existence of two major developmental lineages in the Sas-hill grasslands (Figure 2, r = 0.845). The common ancestor of all objects is quadrat 0, located at a distance of about 20 from quadrats no. 28, 32, 33 and 50. These four sites represent the most species-poor stages in the community, with relatively large bare surfaces inside the plots. There are a few more objects joined at increasingly longer distances; all of these species-poor quadrats were previously classified into the open grassland type (A). Then, we find the breakpoint of the two major "clades" (marked by an asterisk in the diagram). The larger, right-side clade contains plots from the open grassland (all from community type A) at the base. From this basis emerges a group characterized by the richest flora in the study area, with plots representing the *Sesleria*-dominated species-rich community at the end (type C). The other major clade in the tree roughly corresponds to the hilltop vegetation (type B). Although this comparison may give the first impression that the additive tree confirms previous classification and ordination analyses, there is a substantial difference that should not be overlooked. In all classifications (Podani 1985, 1998), type B proved to be transitional between A and C, taking intermediate position in the ordinations, and classified together either with A or C at the two-cluster level in the dendrograms. None of the previous analyses indicated any close affinity between types A and C!

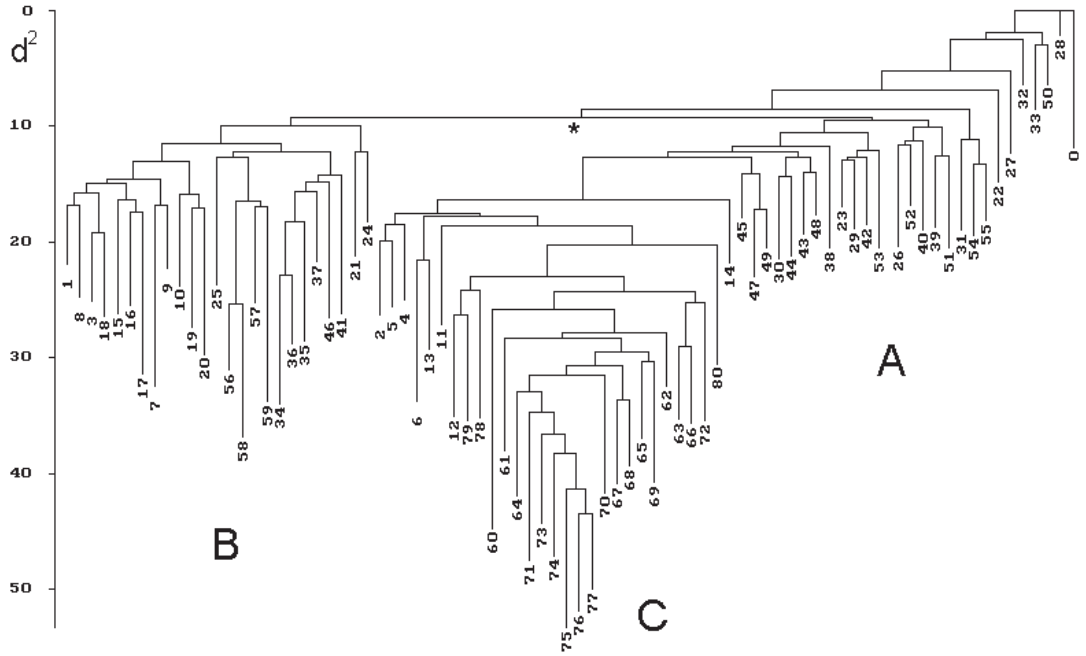

Figure 2. Neighbor joining tree for the Sas-hill data. A, B and C denote major community types identified earlier. The branch labeled by 0 represents the hypothetical precolonization site.

Based on the joint appearance of type A (the open rock grassland) and type C (the most closed grassland type) in the same clade, we put forward the following hypothesis. Even though classification and ordination results and the actual geographic positions support the alternative view, in terms of successional changes type B is *not* an intermediate stage between A and C, but rather a closing grassland type that characterizes hilltops and moderate slopes. It develops from the same starting condition as type C, but then they run through different successional stages. Type A is undoubtedly the pioneering stage in dolomite rocks, as generally acknowledged (Zólyomi 1958), and this stage is still present under more severe ecological conditions (steep slopes, relatively long and direct exposition to sunlight). On north-facing slopes, succession proceeds faster into a closed, species-rich type (C), because of more favorable climatic conditions which in turn give rise to thicker rendzina soils than in any other dolomite grassland.

The performance of the method was further tested by the midpoint rooting option in two separate analyses (no figures shown). In the first case, the empty site was retained in the sample and the result was essentially the same as in Figure 2. The main difference is that type A plots, plus the empty site are moved to the base of the clade of type B plots, thus taking the same intermediate position on the tree as in Fig. 2. In other words, the hypothesis formulated according to the outgroup-rooted analysis was not affected. However, the situation changed completely when the midpoint-rooted analysis did not use the empty plot. Although types A and B got onto the same clade as above, their order was reversed: type B plots appeared at the base and type A plots in terminal positions. The result corresponds to the previous ultrametric classifications suggesting that type B is intermediate in species composition between the other two.

The above results demonstrate that the method of rooting and the inclusion of the empty site greatly affect the shape of the tree, although the major groups remain intact. Thus, the choice of the analytical design has far-reaching consequences in the biological interpretation of results, as is the case in all phylogenetic studies. This is not so in conventional clustering which was also tested for stability upon the addition of the empty site. The result was that the relative positions of the groups and their within-cluster structure did not change at all.

Post-fire changes as traced by permanent plots

A most striking feature of the additive tree (Figure 3, $r = 0.870$) is its ability to show the increase of species richness through time, a possibility unavailable through any other multivariate methods (perhaps with the exception of

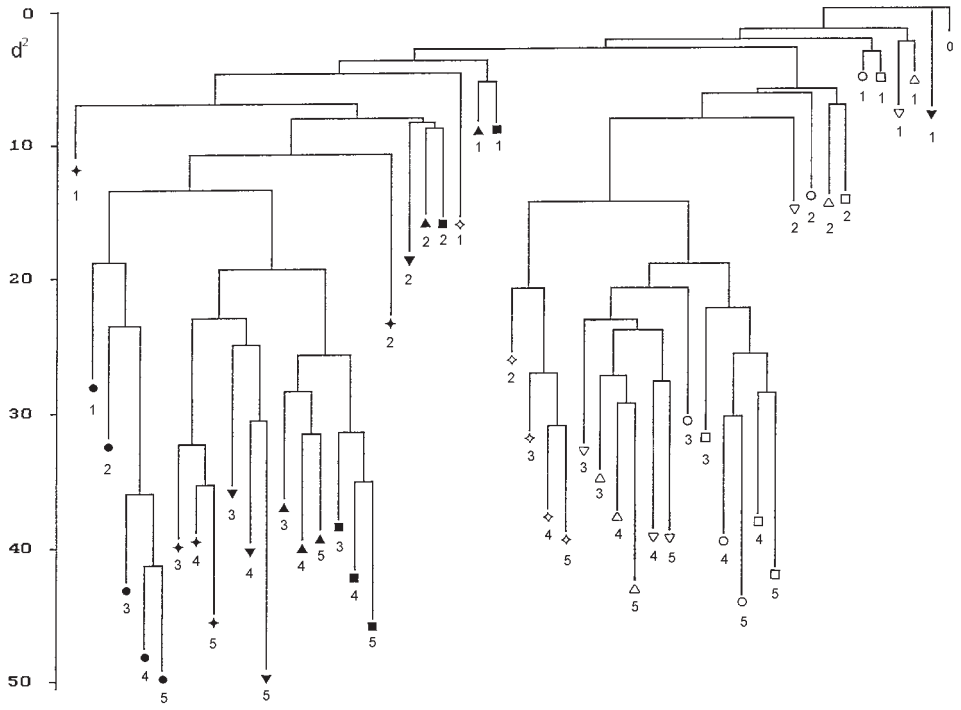

Figure 3. Neighbor joining tree for the post-fire succession data. A permanent plot is marked with the same symbol, temporal changes are indicated by umbers referring to the years of succession. Empty symbols: south-facing slope, full symbols: north-facing slope. The branch labeled by 0 is the empty site.

non-centered PCA, Carleton 1980) and definitely impossible to show by conventional cluster analyses. The symbols and numbers used in this figure allow us to examine whether spatial identity (i.e., data pertaining to the same site but taken in different time are grouped) or temporal coincidence (i.e., grouping according to year, irrespective of spatial position) affects the arrangement of terminal nodes. Let us now examine the graph from right to left.

The five quadrats closest to the root are derived from the first year, four from the southern and one from the northern slope, showing temporal aggregation, the most moderate success of regeneration under southern exposition. Then, two main branches develop, clearly distinguishing between the successional process of north-facing and south-facing slopes (left and right clades, respectively). The effect of aspect is thus manifested very early. The only exception is the appearance of a first-year southern slope quadrat on the northern clade. The reason is that this quadrat is richer in species than the other plots from the same exposition, due to the closeness of a grassland patch to the former pine stand. This patch may have served as a refugium of propagules for a more rapid recolonization on this site.

Increase in species richness is slower on the south-facing slope, shown by the shorter branches of this clade. In this, all but one second-year plots form the next group in the tree, indicating the relative homogeneity of plots in the second year of post-fire regeneration. Then, for years 3 to 5, the temporal factor becomes immaterial, and the chaining of nodes is more influenced by the spatial constraint. For the north-facing slope there is an addition of second year plots at the base of the clade. Then, for the subsequent three years, temporal sequences for the same plots are recognizable in form of small clades, with increasing branch lengths. That is, spatial rather than temporal identity starts to dominate the analysis, a tendency more expressed here than in the south-facing slope. The flora of plots is now clearly inherited from the flora of the previous year.

Syntaxonomy of grassland communities

The additive tree (Figure 4, $r = 0.905$) confirms the separation of three community types (*SF, FB* and *Ss*), whereas the other two (*CF* and *Sh*) are mixed up in a single clade (dots in the figure). There are relatively few 'misplacements' in the tree, four relevés are misplaced from *FB*, and three others fall outside the four main branches. The appearance of four clear clades is a more refined result than the classification obtained by Török and Zólyomi (1998). Their overall analysis revealed only three groups, without separating *FB* from *Ss*. Our tree in-

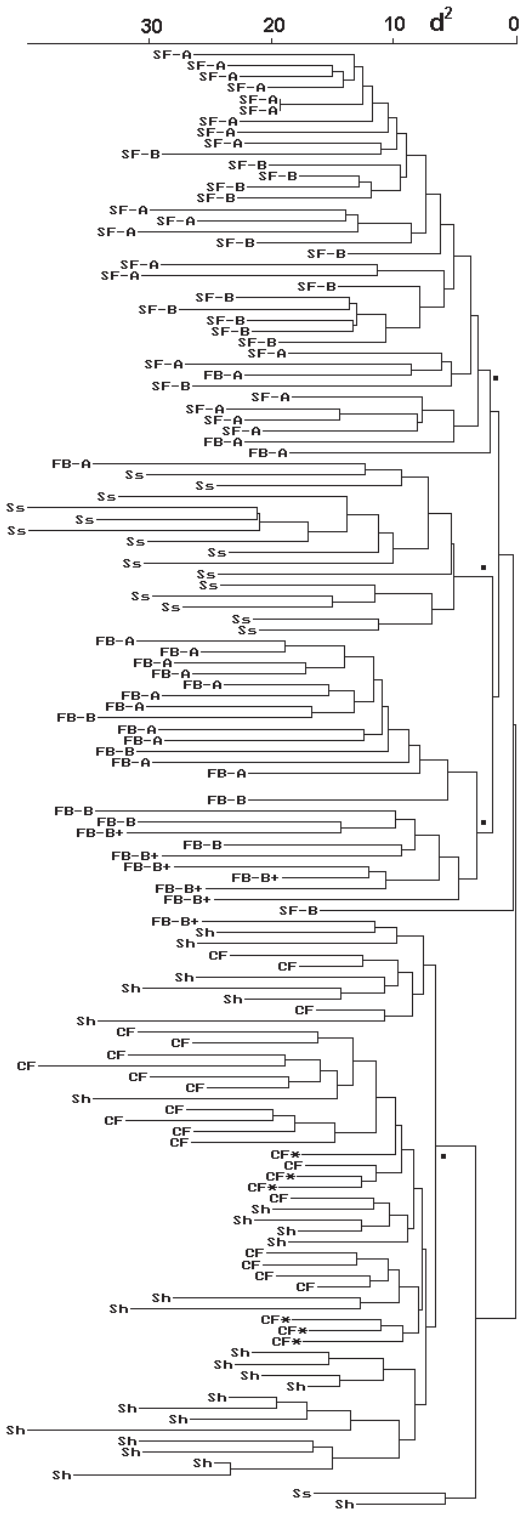

Figure 4. Neighbor joining tree for the syntaxonomic data set for several grasslands (example 3). Labels indicate community types as in the text. * - subass. "*saxifragetosum aizooni*", + - subass. "*primuletosum auriculae*".

dicates that *Ss* is in fact the most homogeneous community, in an obvious conflict with Török and Zólyomi's study in which several iterative steps were necessary to distinguish between *FB* and *Ss*. A further result is that *SF* and *FB* are each well-defined floristically, because their geographic variants cannot be separated equivocally on the tree, especially for *SF*. On the other hand, our findings support Török and Zólyomi's conclusion that *CF* and *Sh* are not different community types. Those authors extracted the data for these two "communities" from the full table, but neither ordination nor classification analyses were able to make distinction between them. Consequently, *CF* and *Sh* should be treated as representatives of the same association. The careful conclusion of the same authors that the two "subassociations" (*CF* subass. *saxifragetosum aizooni* and *FB* subass. *primuletosum auriculae,* see Figure 4) are not distinct from the respective parental community type is also confirmed by our additive tree. In summary, the present analysis supports the syntaxonomy of Török and Zólyomi (1998, p. 123) which otherwise agrees with the classification proposed earlier by Simon (1992).

Discussion

Graph theoretical analysis is rarely used in contemporary vegetation ecology, because ordinations and classifications predominate in studies of community data. This is so even though there are some applications of minimum spanning trees (e.g., Wildi and Schütz 2000, this issue) and splitstrees (Dale 2000, this issue). In fact, the latter author uses an extended additive tree model to evaluate structure which reappears in several branches of the tree. Earlier, Dale et al. (1986) and Dale (1989) also used the additive tree method for the comparison of different similarity measures. We believe that the present study is the first application of the neighbor joining method, an efficient alternative to the minimum sum of squares optimiziation proposed by Sattath and Tversky (1977) to generate additive trees for actual vegetation data.

In addition to our points related mostly to technicalities, there is some philosophy that underlies our approach. Phylogenetic methods, including the NJ technique, have been specifically designed to reveal temporal branching patterns based on data on the recent status of the study objects. Their use is not restricted to the reconstruction of evolutionary pathways of plants and animals. Indeed, they are applicable to any problems in which historical events occur (e.g., languages, Cavalli-Sforza et al. 1988). They were extensively used in cladistic biogeography (e.g., Rosen 1978), for the purpose of contrasting evolutionary patterns for taxa based on their distributional properites (using the so-called area cladograms). A most successful attempt to apply such methods in a large-scale biogeographical survey of the distribution of fish species is due to Legendre (1986). The natural question to ask is whether these methods can also be applied to small-scale surveys where the historical element, although present, is not always evident.

We feel that the results presented in this paper demonstrate that the arsenal of data analysis methodology deserves continuous renewal, by a feedback from other areas of biology. NJ analysis produced a result in conflict with other analyses in a small-scale vegetation survey, thus allowing a different hypothesis on potential successional trends in the study area. If an empty site were not included in the study, no such hypothesis would be generated, and the tree would have been no more than a confirmation of previous classifications. The tree constructed for data from permanent plots illustrate the ability of the method to track revegetation processes in such a way that floristic changes are directly interpretable from the diagram. An additive tree may also be used as an auxiliary classificatory tool to confirm or reject distinction between proposed syntaxonomical units. Admittedly, the success of additive trees as illustrations of classifications strongly depends on the *a posteriori* positioning of the root. In our examples, midpoint rooting proved to be useful to confirm earlier hierarchical classifications, although it may be only a happy coincidence. In succession surveys, either from chronosequences or from permanent plots, an empty site may serve as an outgroup object if the successional sere involves pioneering stages. In studies of secondary succession, inclusion of the empty site is logical whenever the former vegetation was completely destroyed. As with many other procedures of exploratory data analysis, several sets of data have to be analyzed carefully in order to establish the relative merits of the particular method. Even though the results presented in this paper convincingly demonstrate the potential utility of additive trees in ecology, the present approach also requires further tests and comparisons with standard methodology.

Acknowledgements. We are grateful to M. B. Dale for discussions on the concept of additive trees and to N. C. Kenkel and an anonymous referee for their thorough criticism of the manuscript. This work was supported by Hungarian National Research Fund (OTKA, grant no. T29784) and by the Hungarian Committee for Technological Development (EU-98-C6-072).

References

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. In: F. R. Hodson, D. G. Kendall and P. Tautu (eds). *Mathematics in the Archaeological and Historical Sciences*. Edinburgh Univ. Press, Edinburgh. pp. 387-395.

Carleton, T. J. 1980. Non-centered component analysis of vegetation data: a comparison of orthogonal and oblique rotation. *Vegetatio* 42:59-66.

Carroll, J. D. and J. J. Chang. 1976. Spatial, non-spatial and hybrid models for scaling. *Psychometrika* 41:439-463.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi and J. L. Mountain. 1988. Reconstruction of human evolution: bringing together genetic, archaeological and linguistic data. *Proc. Natl. Acad. Sci. USA* 85:6002-6006.

Corter, J. E. 1982. ADDTREE/P: a PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behav. Res. Meth. Instrument.* 14: 353-354.

Cunningham, J. P. 1978. Free trees and bidirectional trees as representations of psychological distance. *J. Math. Psychol.* 17:165-188.

Dale, M. B. 1989. Mutational and nonmutational similarity measures. *Coenoses* 3:121-133.

Dale, M. B. 2000. On plexus representations of dissimilarities. *Community Ecology* 1.

Dale, M. B., M. Beatrice and R. Venanzoni. 1986. A comparison of some methods of selecting species in vegetation analysis. *Coenoses* 1:35-52.

Day, W. H. E. and H. Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* 1:7-24.

de Soete, G. 1983. A least squares algorithm for fitting additive trees to proximity data. *Pychometrika* 48: 621-626.

de Soete, G. 1988. Tree representations of proximity data by least squares methods. In: H. H. Bock (ed.), *Classification and Related Methods of Data Analysis*, North Holland, Amsterdam, pp. 147-156.

Digby, P. G. N. and R. A. Kempton. 1987. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, London.

Gascuel, O. 1994. A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.* 11:961-963.

Goodall, D. W. 1953. Objective methods for the classification of vegetation I. The use of positive interspecific correlation. *Aust. J. Bot.* 1:39-63.

Lance, G. N. and W. T. Williams. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J.* 9:373-380.

Legendre, P. 1986. Reconstructing biogeographic history using phylogenetic-tree analysis of community structure. *Syst. Zool.* 35:68-80.

Michalski, R. S., I. Bratko and M. Kubat (eds). 1998. *Machine learning and data mining: Methods and Applications*. Wiley, New York.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30:371-403.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. *J. Ecol.* 55:193-205.

Page, R. D. M. and E. C. Holmes. 1998. *Molecular Evolution. A Phylogenetic Approach*. Blackwell, Oxford.

Patrinos, A. N. and S. L. Hakimi. 1972. The distance matrix of a graph and its tree realization. *Q. Appl. Math.* 30:255-269.

Podani, J. 1985. Syntaxonomic congruence in a small-scale vegetation survey. *Abstracta Botanica* 9: 99-128.

Podani, J. 1994. *Multivariate Data Analysis in Ecology and Systematics*. SPB Publishing, The Hague.

Podani, J. 1997. SYN-TAX 5.1: A new version for PC and Macintosh computers. *Coenoses* 12:149-152.

Podani, J. 1998. Numerikus cönológiai vizsgálatok a Sas-hegy (Budai-hg.) dolomitsziklagyepjeiben. (A complex numerical analysis of dolomite rock grasslands of the Sas-hegy Nature Reserve, Budapest, Hungary. In Hungarian with English summary) In: P. Csontos (ed.), *Sziklagyepek szünbotanikai kutatása*. Scientia, Budapest, pp. 211-229.

Podani, J. 2000. Simulation of random dendrograms: some comments. *J. Classif.* 17 (in press).

Rosen, D. E. 1978. Vicariant patterns and historical explanation in biogeography. *Syst. Zool.* 27:159-188.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.

Sattath, S. and Tversky, A. 1977. Additive similarity trees. *Psychometrika* 42:319-344.

Shepard, R. N. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* 210:390-398.

Simon, T. 1992. *A magyarországi edényes flóra határozója. Harasztok - Virágos növények*. Tankönyvkiadó, Budapest.

Sneath, P.H.A. and Sokal, R. R. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

Swofford, D. L. and G. J. Olsen. 1990. Phylogeny reconstruction. In: D. M. Hillis and C. Moritz (eds.), *Molecular Systematics*. Sinauer, Sunderland, Mass. pp. 411-501.

Tamás, J. and P. Csontos 1998. A növényzet tűz utáni regenerálódása dolomitra telepített feketefenyvesek helyén. (Early regeneration of dolomite vegetation after burning of *Pinus nigra* plantations. In Hungarian with English summary.) In: P. Csontos (ed.), *Sziklagyepek szünbotanikai kutatása*. Scientia, Budapest, pp. 231-264.

Török, K. and B. Zólyomi 1998. A Kárpát-medence öt sziklagyeptársulásának szüntaxonómiai revíziója. (Syntaxonomic revision of five rocky grassland communities of the Carpathian Basin. In Hungarian with English summary.) In: P. Csontos (ed.), *Sziklagyepek szünbotanikai kutatása*. Scientia, Budapest, pp. 109-132.

Westphal, C. and T. Blaxton. 1998. *Data Mining Solutions*. Wiley, New York

Wildi, O. and M. Schütz. 2000. Reconstruction of a long-term recovery process from pasture to forest. *Community Ecology* 1.

Williams, W. T. and J. M. Lambert. 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* 47:83-101.

Zólyomi, B. 1958. Budapest és környékének növénytakarója. In: M. Pécsi (ed.), *Budapest természeti képe*. Akadémiai Kiadó, Budapest, pp. 508-642.