Professional Meetings

Statistical Society of Canada

The 1982 Annual Meeting of the Statistical Society of Canada will take place at the University of Ottawa as part of the Learned Societies Meetings. The dates for the SSC Meeting are presently scheduled to be 3–5 June 1982. The program chairman is:

Dr. William G. Warren
Forintek Canada Corporation
Western Forest Products Laboratory
6620 North West Marine Drive
Vancouver, British Columbia
Canada V6T 1X2.

The local arrangements chairman is:

Professor Mayer Alvo
Department of Mathematics
University of Ottawa
Canada K2H 8S3.

Summer Institute in Survey Research Techniques

For the thirty-fifth consecutive year, the Survey Research Center of the Institute for Social Research Center of the Institute for Social Research at The University of Michigan will hold a Summer Institute in Survey Research Techniques from June 28, 1982, to August 20, 1982.

The Institute is designed to meet some of the educational and training needs of men and women engaged in business, government research and other statistical work, and also to meet the needs of graduate students and university instructors interested in quantitative research in the social sciences.

For further information concerning specific dates, courses, fees, housing, etc., please write to Helene J. Hitchcock, Administrative Manager, Office of the Director, Survey Research Center, Institute for Social Research, Post Office Box 1248, Ann Arbor, Michigan 48106.

SPATIAL VERSUS TREE REPRESENTATIONS OF PROXIMITY DATA

SANDRA PRUZANSKY
BELL LABORATORIES

AMOS TVERSKY
STANFORD UNIVERSITY

J. DOUGLAS CARROLL
BELL LABORATORIES

In this paper we investigated two of the most common representations of proximities, two-dimensional euclidean planes and additive trees. Our purpose was to develop guidelines for comparing these representations, and to discover properties that could help diagnose which representation is more appropriate for a given set of data. In a simulation study, artificial data generated either by a plane or by a tree were scaled using procedures for fitting either a plane (KYST) or a tree (ADDTREE). As expected, the appropriate model fit the data better than the inappropriate model for all noise levels. Furthermore, the two models were roughly comparable: for all noise levels, KYST accounted for plane data about as well as ADDTREE accounted for tree data. Two properties of the data proved useful in distinguishing between the models: the skewness of the distribution of distances, and the proportion of elongated triangles, which measures departures from the ultrametric inequality. Applications of KYST and ADDTREE to some twenty sets of real data, collected by other investigators, showed that most of these data could be classified clearly as favoring either a tree or a two-dimensional representation.

Key words: multidimensional scaling, clustering, tree structures, additive trees.

Most representations of proximity data belong to one of two families of models: continuous spatial models and discrete network models. Spatial models embed the objects (e.g., colors, words, emotions, people or countries) in some coordinate space so that the metric distances between points represent the observed proximities between the respective objects. Network models represent each object as a node in some graph so that the relations among the nodes in the graph reflect the proximities among the objects.

The most widely used spatial representation is the two-dimensional euclidean space, or plane for short. The popularity of this model is due, largely, to its map-like character which provides a convenient visual representation of the relations among the objects, and enhances the interpretation of dimensions and clusters.

Perhaps the most common network model is a tree (i.e., a connected graph without cycles) in which the terminal nodes represent objects, and the distance between two objects is the length of the path that joins them. A tree representation has several attractive formal properties (e.g., each pair of points is joined by a unique path) which permits a parsimonious description and a convenient graphical display. Furthermore, a tree structure lends itself to a natural interpretation as a hierarchical clustering or in terms of common and distinctive features.

A portable PASCAL program implementing the Sattath and Tversky [1977] ADDTREE algorithm is available from J. Corter, Department of Psychology, Stanford University, Stanford, California 94305.

Requests for reprints should be sent to Sandra Pruzansky, Bell Laboratories, 2C-552, Murray Hill, New Jersey 07974.

0033-3123/82/0300-3081
© 1982 The Psychometric Society

The present paper is concerned with the comparison of plane and tree representations of proximity data. We chose to focus on these models because:

a) they are widely used in psychology as well as in other fields,
b) they are simple prototypical examples of spatial and network representations, –
c) they are roughly comparable in complexity; both trees and planes have about $2n$ parameters for a set of n objects,
d) there are fairly efficient scaling methods for constructing planes and trees from a set of similarities or distances between objects, and
e) they lead to different interpretations: planes suggest continuous dimensions, whereas trees suggest discrete clusters.

Several authors [e.g., Fillenbaum & Rapoport, 1971; Shepard, 1974; Carroll, 1976; Sattath & Tversky, 1977; Carroll & Pruzansky, 1980] have constructed both trees and planes from the same sets of proximity data. In this paper, we develop guidelines for comparing trees and planes and explore properties that could help diagnose which of these models is more appropriate for a given set of data. In the first part of the paper we describe a simulation study in which artificial data generated either by a plane or by a tree are scaled using either a plane or a tree model. In the second part of the paper we fit the plane and tree models to 20 sets of proximity data. The implications of the study are discussed in the third and last section.

Simulation Study

Design

The present simulation was based on a factorial design with three factors: generating model (plane or tree), amount of error, and set size.

Generating model. Plane data were generated by selecting n points at random from a uniform distribution over the unit square, and computing euclidean distances between all pairs of points. Tree data were generated by selecting n points at random from the 2^{10} terminal nodes of a 10-level complete binary tree and then eliminating redundant links. This section defines a new tree with the same root, consisting of n terminal nodes and $2n - 2$ links or edges. The length of each link in the selected tree was chosen randomly from a uniform distribution over the unit interval. The path-length distances for all pairs of selected nodes were computed. For a discussion of different schemes for generating random trees see Furnas (Note 1).

Each set of plane and tree distances was normalized to unit variance. All of the analyses discussed in this paper are unaffected by the addition of a constant to all distances. Hence, for present purposes the mean of the distribution of distances is immaterial.

Amount of error. Fallible data were generated by adding to each distance an error component drawn at random from a normal distribution with zero mean and variance σ_e^2. Three noise levels were used:

(i) error-free data ($\sigma_e^2 = 0$),
(ii) small error ($\sigma_e^2 = .25$),
(iii) moderate error ($\sigma_e^2 = .50$).

In addition, random "distances" were generated for each n from a normal distribution with unit variance.

Set size. Each data set consisted of 36, 24, or 12 points. The smaller data sets were nested within the larger data sets. The entire factorial design consisted of two generating models, three error levels and three set sizes, yielding a total of 18 conditions, plus three random conditions—one for each set size. There were 11 replications of each condition resulting in a total of 231 data sets. Thus, each data set consisted of all interpoint distances among 36, 24 or 12 points, with or without added error, generated either by a plane or by a tree.

Analysis

Scaling models. Each data set was analyzed using two different scaling programs: KYST and ADDTREE. KYST is a multidimensional scaling procedure yielding a spatial solution [Kruskal, Young & Seery, Note 2.] Two KYST analyses were performed, one using linear regression and one using monotone regression. KYST analyses were done in two dimensions using euclidean distance, stress formula 1 (S_1), a TORSCA start, and the primary approach to ties.

ADDTREE is a clustering procedure that represents a distance matrix as an additive or "path length" tree, which generalizes the familiar ultrametric tree. ADDTREE first constructs a tree structure based on a partition of all quadruples into pairs of immediate neighbors and then computes least-squares estimates of the lengths of the links [Sattath & Tversky, 1977]. The distances in an additive tree do not depend on the choice of root. However, to enhance the interpretation of the structure ADDTREE selects a root for the tree on the basis of the data.

Goodness-of-fit. For each scaling solution, we have computed two measures of goodness-of-fit: r_L^2 and r_M^2. r_L^2 is the square of the product-moment correlation between the solution and the data, i.e., the proportion of linearly explained variance. r_M^2 is the square of the product-moment correlation between the solution and the "best" monotone transformation of the data, i.e., the proportion of monotonically explained variance. Note that r_M^2 equals $1 - S_2^2$, where S_2 is stress formula 2 defined by

$$\left[\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}(d_{ij} - \bar{d})^2}\right]^{1/2}$$

d_{ij} is the fitted distance between objects i and j, $\bar{d}$ is the average of the fitted distances, and $\hat{d}_{ij}$ is the value of the best fitting monotone function relating given and fitted distances.

We chose r_M^2 as a measure of fit because it has a natural interpretation as a squared correlation coefficient, or as the proportion of explained variance, and because S_2, unlike S_1, is not altered by adding a constant to all the distances. Since ADDTREE often adds a positive constant to all distances to satisfy the triangle inequality, it should be evaluated using measures that are unaffected by such a transformation. In addition, we found in a pilot study, that the obtained values of r_M^2 were hardly affected by the stress formula (S_1 or S_2) used in the optimization process. This result is not very surprising since S_1 differs from S_2 only in the normalizing factor; S_1 is obtained from S_2 by replacing the denominator by $\sum_{i<j} d_{ij}^2$.

It should be noted that the selected indices of goodness-of-fit, r_L^2 and r_M^2, are not identical to the loss functions used in either KYST or ADDTREE. However, r_M^2 is closely related to S_1, which is minimized by KYST, while the estimation of link lengths in ADDTREE (though not the construction of the tree's topology) maximizes r_L^2. One may expect, therefore, that r_M^2 will favor KYST, while r_L^2 will favor ADDTREE. Both artificial and real data, however, showed little or no bias of this kind.

Results. Since the KYST results using monotone regression and linear regression were highly similar, we report only results using monotone regression. The average values of r_L^2 and r_M^2, respectively, are displayed in Figures 1 and 2 as a function of noise level, for all

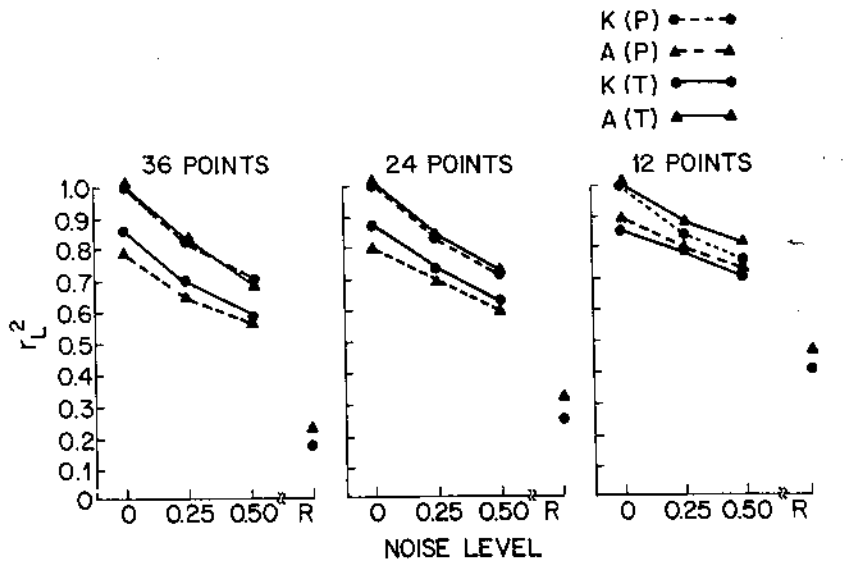

FIGURE 1.
Goodness-of-fit measure, r_L^2, between data, generated by a tree (T), or a plane (P), and distances computed from KYST (K) and ADDTREE (A) solutions to these data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

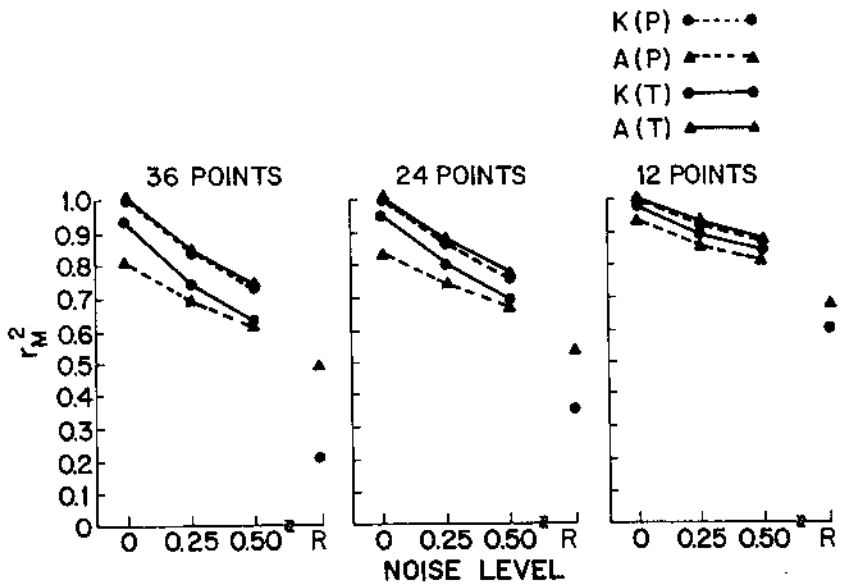

FIGURE 2.
Goodness-of-fit measure, r_M^2, between data, generated by a tree (T), or a plane (P), and distances computed from KYST (K) and ADDTREE (A) solutions to these data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

three set sizes. Unless specified otherwise, we discarded the two extreme values in each condition so the reported averages are based on nine (rather than 11) replications. The medians and interquartile ranges (the sixth, third and ninth observations) for r_L^2 and r_M^2 are presented in Appendix A and Appendix B respectively. These appendices provide information about the expected variations of the two correlational measures for different set sizes and error levels.

Each panel in Figures 1 and 2 includes four curves. Each curve describes the correlations between the data, generated either by a tree (T) or by a plane (P), and the distances computed from the solutions obtained either by KYST (K) or by ADDTREE (A). Thus, the curves labelled K(T), for example, describe the fit of KYST solutions to tree data as a function of noise level.

Figures 1 and 2 show that, under both r_L^2 and r_M^2, KYST fits plane data better than ADDTREE while ADDTREE fits tree data better than KYST, for each of the three error levels. (In fact, the appropriate model fit *each* of the 231 individual data sets better than the inappropriate model did.) The correlations between KYST solutions and plane data are practically identical to the correlations between ADDTREE solutions and tree data, as evinced by the coincidence of the top curves in Figures 1 and 2. However, KYST appears to accommodate tree data better than ADDTREE accommodates plane data. As expected, the correlations between solutions and data decrease as noise level increases. The appropriate model accounts for about 10–20 percent more variance than the inappropriate model. In general, r_L^2 and r_M^2 behave similarly with one notable exception. For random data, the discrepancy between ADDTREE and KYST for r_L^2 is quite small and roughly independent of set size. In contrast, the discrepancy between the models for r_M^2 increases substantially with set size. This effect deserves further analysis.

Examples of ADDTREE and KYST solutions for both tree and plane data are displayed in Figures 3 and 4. For graphical convenience ADDTREE solutions are displayed in rectangular form with dummy vertical links. In this representation the distance between any two objects is the length of the path that joins them excluding the vertical components. Figure 3a shows an ADDTREE solution for a data set generated by a tree, consisting of 24 points without error. Figure 3b shows a two-dimensional KYST solution for the same data. The ADDTREE solution recovered the data perfectly; for the KYST solution, r_L^2 was .88. Figure 4a shows an ADDTREE solution for a data set generated by a plane, again with 24 points and no added error. Figure 4b shows a two-dimensional KYST solution for the same data. The KYST solution recovered the data perfectly; for the ADDTREE solution, r_L^2 was .80. Note that the tree fitted to plane data (Figure 4a) is less balanced than the tree fitted to tree data (Figure 3a) in the sense that it bifurcates into branches with an unequal number of terminal nodes. The points displayed in Figure 3b, based on tree data, are more clustered than the points in Figure 4b, based on plane data.

In addition to plane data we generated five replications of three-dimensional spatial data for the error free and small error conditions. The application of ADDTREE and (two-dimensional) KYST to these data showed that the two-dimensional KYST solution produced a better fit than did ADDTREE for all data sets.

Perturbing small or large distances. In the above conditions error was added to all distances. We generated additional data where we perturbed the upper or the lower third of the distribution of distances leaving the remaining $\frac{2}{3}$ of the data undisturbed. As in the previous conditions, the distances were transformed (prior to perturbation) to have unit variance. Two noise conditions were used. In the first condition, normal error with variance $\sigma_\epsilon^2 = 1$ was added to the upper or to the lower third of the distribution of distances. In the second condition, the upper or lower $\frac{1}{3}$ of the distances were replaced by distances selected at random from a normal distribution with mean equal to the mean of the original dis-

Noise generation

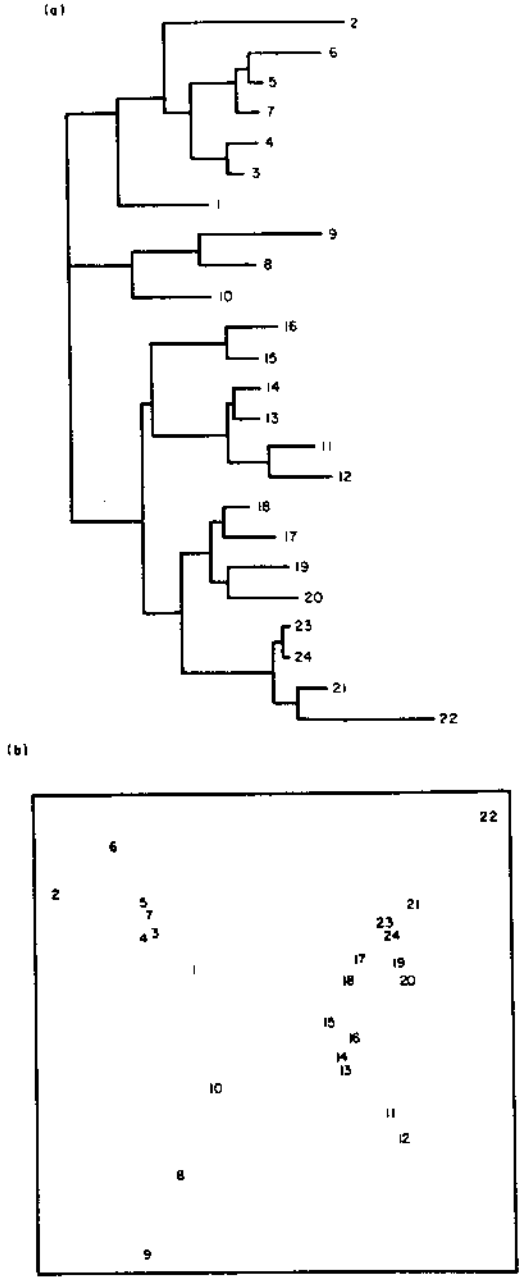

FIGURE 3.
a) ADDTREE solution for a representative data set generated by a tree consisting of 24 points with no error.
b) Two-dimensional KYST solution for the same data.

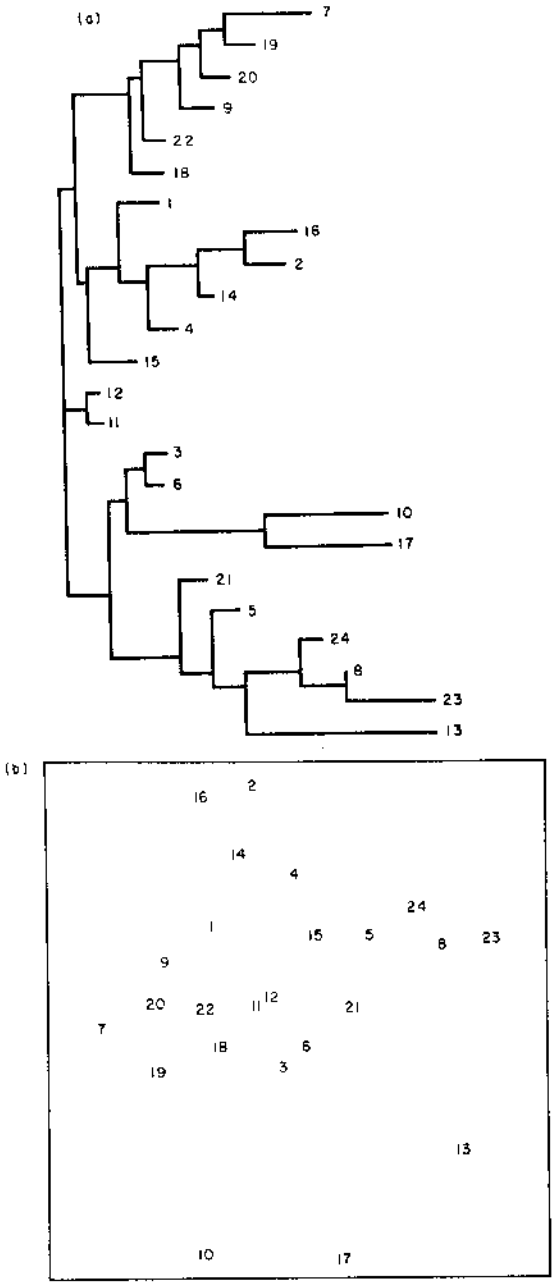

FIGURE 4.
a) ADDTREE solution for a representative data set generated by a plane consisting of 24 points without error.
b) Two-dimensional KYST solution for the same data.

tribution of distances, and unit variance. Note that this operation changes both the mean and variance of distances in the perturbed subset relative to these same parameters of the unperturbed subset.

Five replications were generated for each error level and each set size. Plane data were analyzed by KYST and tree data were analyzed by ADDTREE. The values of r_L^2 and r_M^2 for these analyses are reported in Table 1. Table 1 shows that perturbation of small distances reduces the ADDTREE fit while perturbation of large distances has little effect, except when these distances are replaced by noise. Perturbing either the large or small distances has little effect on KYST. However, KYST is more disturbed by replacing large distances by noise than by replacing small distances by noise.

Diagnostic Measures

The preceding comparisons of planes and trees were based on correlational measures of goodness-of-fit. In this section we examine two properties of interpoint distances, skewness and elongation, that may help to discriminate between trees and planes.

Skewness. A mathematical analysis of Sattath and Tversky [1977] suggests that a convex configuration of points in a plane tend to generate many small distances and fewer large distances, while tree data often produce the opposite pattern. Hence, the skewness of

Table 1: Goodness of fit measures, r_L^2, and r_M^2, between plane data and KYST solutions and between tree data and ADDTREE solutions. Noise was added to or replaced the upper or lower 1/3 of the data distributions. The medians of five replications are listed for three set sizes.

r_L^2					
PORTION OF DISTRIBUTION	*SET*	ADDTREE		KYST	
		NOISE CONDITION		*NOISE CONDITION*	
PERTURBED	*SIZE*	$\sigma_\epsilon^2 = 1.0$	*R*	$\sigma_\epsilon^2 = 1.0$	*R*
UPPER 1/3	36	.98	.61	.97	.54
	24	.98	.70	.96	.44
	12	.92	.79	.90	.38
LOWER 1/3	36	.86	.54	.97	.79
	24	.84	.47	.95	.73
	12	.87	.43	.94	.64

r_M^2					
PORTION OF DISTRIBUTION	*SET*	ADDTREE		KYST	
		NOISE CONDITION		*NOISE CONDITION*	
PERTURBED	*SIZE*	$\sigma_\epsilon^2 = 1.0$	*R*	$\sigma_\epsilon^2 = 1.0$	*R*
UPPER 1/3	36	.99	.71	.97	.65
	24	.98	.86	.98	.54
	12	.95	.92	.93	.59
LOWER 1/3	36	.88	.66	.97	.82
	24	.88	.60	.96	.78
	12	.94	.63	.97	.79

the distribution of distances may help diagnose whether a given data set is more likely to have been generated by a tree or by a plane. We used the standard measure of skewness, that is, the third central moment divided by the cubed standard deviation:

$$\frac{n^{1/2}\sum_{i<j}(\delta_{ij}-\bar{\delta})^3}{\left[\sum_{i<j}(\delta_{ij}-\bar{\delta})^2\right]^{3/2}}$$

where δ_{ij} denotes the data and $\bar{\delta}$ is their average. Note that skewness vanishes if the distribution is symmetric. For a uniform process in the unit square, the skewness of the distribution of distances is .19, as calculated from Ghosh's [1951] results.

The average values of skewness for both tree and plane data are displayed in Figure 5 for all noise levels and set sizes; Appendix C shows the medians and interquartile ranges. As seen in Figure 5, the plane data exhibit positive skewness, for practically all set sizes and noise levels. Tree data, on the other hand, exhibit negative skewness but become more symmetric with an increase in added error. The distributions of tree and plane data are presented in Figures 6a and 6b, respectively, for all 11 data sets consisting of interpoint distances between 36 points without error.

Similar results were obtained using other methods for generating tree and plane data. Three-dimensional spatial data, based on a uniform process in the unit cube and two dimensional data based on a bivariate normal process (with zero correlation) exhibited positive skewness. Note that the distribution of euclidean distances between all points generated by the latter process has a chi distribution with two degrees of freedom. Straightforward calculation shows that the skewness of this distribution is .63. In constrast, the observed distributions of distances for ultrametric trees (where all terminal nodes were equidistant from the root) and for fully balanced binary trees yielded negative skewness.

Figure 7 presents the average skewness of the fitted distances obtained by applying KYST and ADDTREE to both plane and tree data; the medians and interquartile ranges are shown in Appendix D. Figure 7 indicates that the skewness of distributions from KYST

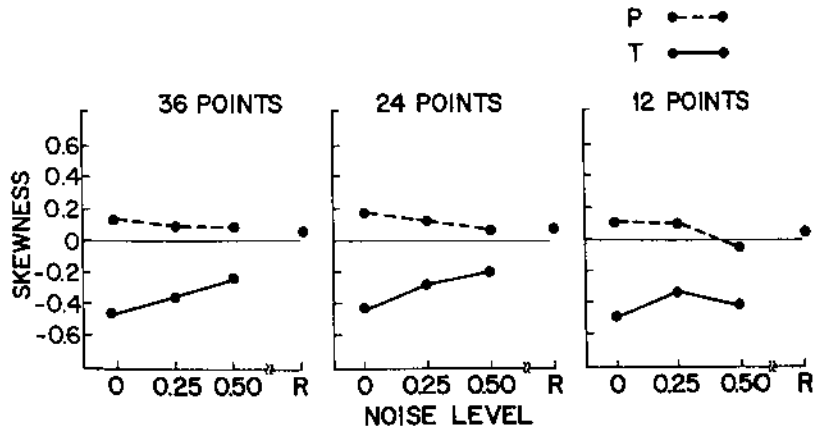

FIGURE 5.
Skewness measure for both tree and plane data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

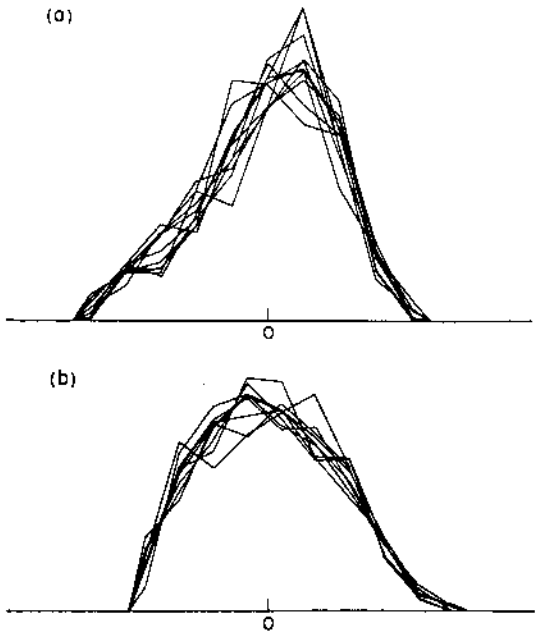

FIGURE 6.
a) Distributions of 11 sets of tree data generated from 36 points without error. b) Distributions of 11 sets of plane data generated from 36 points without error.

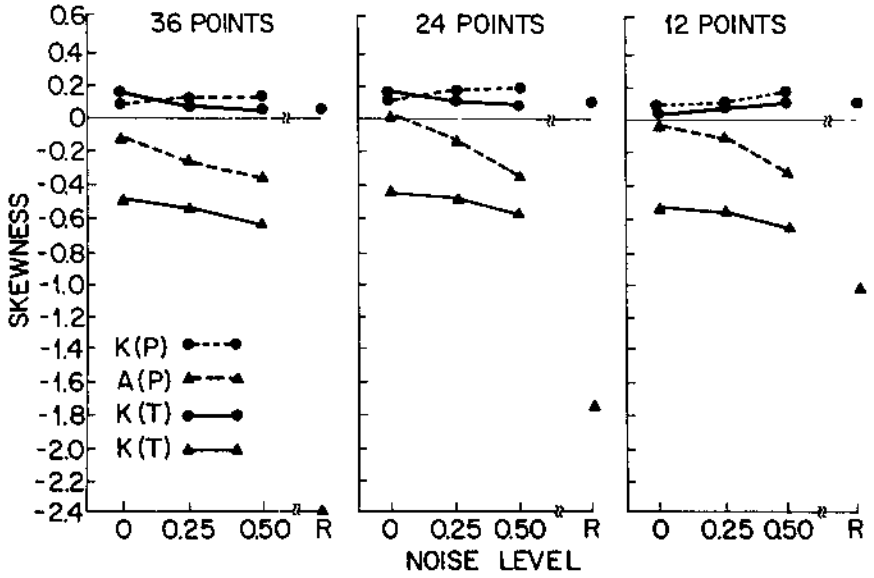

FIGURE 7.
Skewness measure from solutions obtained by applying KYST and ADDTREE to both plane and tree data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

solutions are positive, while those from ADDTREE solutions are negative for both plane and tree data. The skewness of the representations, therefore, tends to reflect the properties of the scaling model rather than the properties of the data.

Elongation. In a binary (rooted) tree, any triple of terminal nodes generally forms a sub-tree say $(i, j)k$ so that i and j belong to a cluster that does not include k. Because intercluster distances, on the average, exceed intracluster distances, δ_{ij} is expected to be smaller than δ_{jk} and δ_{ik}. Furthermore, since there is no basis to order this pair of intercluster distances, it is reasonable to expect if $\delta_{ij} \leq \delta_{jk} \leq \delta_{ik}$ then $\delta_{ik} - \delta_{jk} \leq \delta_{jk} - \delta_{ij}$. Thus, the triangle formed by the distances among any three objects is elongated, that is, the middle side M is closer in length to the long side L than to the short side S, or L-M < M-S. Note that in an ultrametric tree, L = M ≥ S.

The elongation of a data set is defined as the proportion of triangles where L-M < M-S. For ultrametric trees elongation should be one; for random data the expected elongation is .5. Figure 8 presents the average elongation values for both plane and tree data. Appendix C shows the medians and interquartile ranges. Figure 9 displays the average elongation values obtained by applying KYST and ADDTREE to both plane and tree data. Appendix E shows the medians and interquartile ranges. Figure 8 shows that, for all error levels, tree data produce a higher proportion of elongated triangles than plane data. The proportion of elongated triangles decreases with noise level for both planes and trees. (The mean proportion of elongated triangles for three-dimensional spatial data was .61 for the error free condition and .57 for the small error condition.)

Figure 9 shows that ADDTREE solutions to tree data produce the highest proportion of elongated triangles, while KYST solutions to plane data produce the least. This relation holds for all noise levels. However, KYST solutions to tree data produce more elongation than ADDTREE solutions to plane data. Unlike skewness, therefore, the elongation of a representation appears to depend more on the data than on the scaling model.

Analysis of Real Data

In this section we describe the application of KYST and ADDTREE to 20 sets of proximity data that had been collected previously by other investigators. The data con-

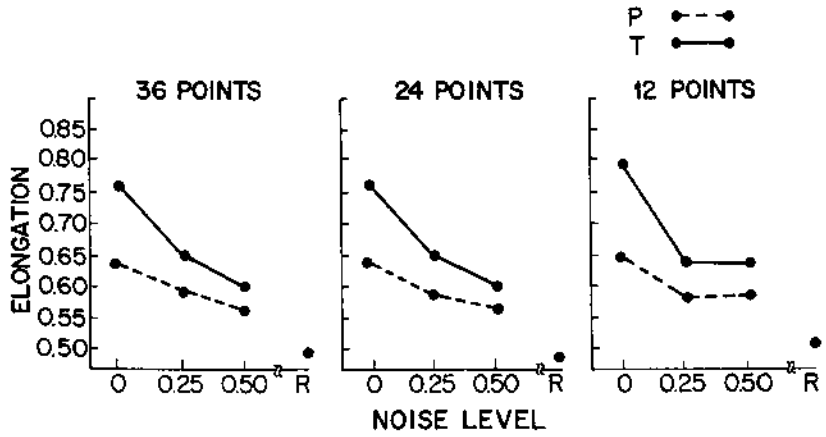

FIGURE 8.
Elongation measure for both tree and plane data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

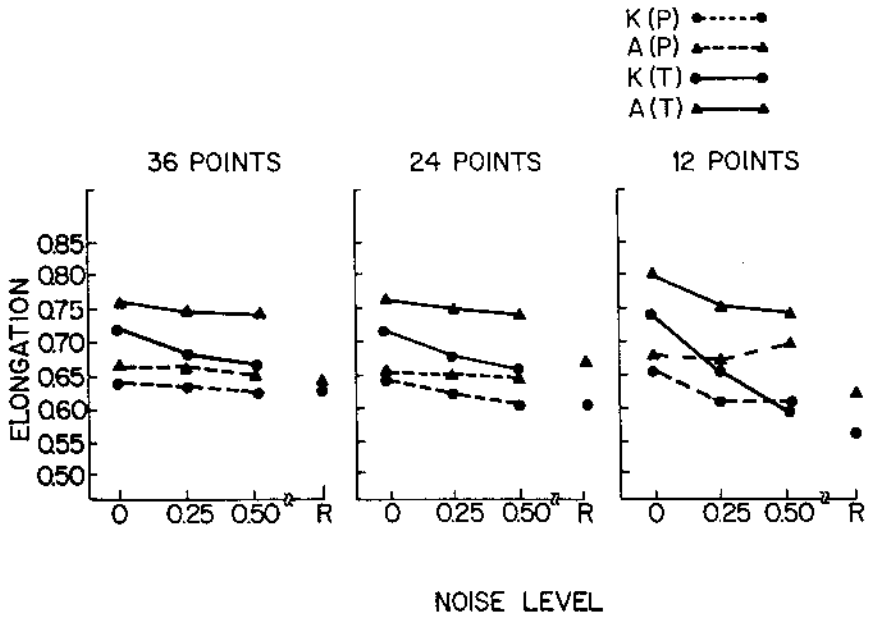

FIGURE 9.
Elongation measure from solutions obtained by applying KYST and ADDTREE to both plane and tree data. The mean over nine replications is plotted as a function of noise level, for three set sizes. R denotes random data.

sisted of both conceptual and perceptual stimuli. The number of stimuli ranged from 12 to 36, the same range used in the simulation study. Conceptual stimuli consisted of lists of concepts from various semantic fields such as fruits, animals and occupations. The perceptual stimuli were divided into three subgroups: forms, colors, and sounds, either speech sounds or complex tones. Table 2 describes the basic features of the various data sets.

We computed goodness-of-fit measures, r_L^2 and r_M^2, of both solutions for each data set. We also estimated skewness and elongation from the data and the solutions. The left-hand section of Table 3 presents the linear and monotone measures of goodness-of-fit for both KYST and ADDTREE. The solution that fits the data better is underlined in each case.

The empirical data reveal a highly consistent pattern. The first ten data sets, which use as stimuli concepts selected from various semantic fields, are better fit by a tree than by a plane. The last six data sets, which use colors and sounds as stimuli are better accounted for by a plane than by a tree. Furthermore, sets 11 and 14, which employ a factorial design, are better fit by a plane than by a tree. (Factorial designs are marked by an F in the description column.) Although the stimuli in set 13 have a factorial structure, one of the factors (size) was completely dominated by the other (shape), and these data are better described by a tree. Set 12 appears inconclusive; this is the only case where the linear and the monotone measures diverge.

The middle and right-hand sections of Table 3 display, respectively, the skewness and elongation measures for the original data and for the two representations. The solution whose value is closest to the data (column labelled *DATA*) is underlined for each data set.

Inspection of the skewness values in Table 3 shows that the skewness for factorial data (11, 13, 14) and the color data (15, 16, 17) is better matched by KYST than by ADDTREE.

Table 2: Source and Description of Data Sets

SET	SOURCE	STIMULUS DESCRIPTION	SET SIZE	DESIGN	COMMENTS	METHOD
	Conceptual					
1	Mervis et al (Note 3)	Vehicles	20	natural selection	twenty most common objects from each category	similarity rating
2	Mervis et al (Note 3)	Sports	20	natural selection		similarity rating
3	Mervis et al (Note 3)	Vegetables	20	natural selection		similarity rating
4	Mervis et al (Note 3)	Furniture	20	natural selection		similarity rating
5	Mervis et al (Note 3)	Tools	20	natural selection		similarity rating
6	Mervis et al (Note 3)	Weapons	20	natural selection		similarity rating
7	Mervis et al (Note 3)	Fruits	20	natural selection		similarity rating
8	Henley (1969)	Animals	30	natural selection	most common objects	dissimilarity rating
9	Kraus (Note 4)	Occupations	35	natural selection	representative sample	sorting
10	Hutchinson and Lockhead (Note 5)	Varied Objects	36	natural selection	6 common items in each of 6 categories	dissimilarity rating
11	Gati (Note 6)	Students	16	4 x 4 factorial design	4 political parties by 4 academic fields	dissimilarity rating
	Perceptual					
	Forms					
12	Kuennapas and Janson (1969)	Letters	25	exhaustive set	lower case Swedish letters	similarity rating
13	Gati (Note 6)	Polygons	16	4 x 4 factorial design	4 shapes by 4 sizes	dissimilarity rating
14	Gati and Tversky (1981)	Plants	16	4 x 4 factorial design	4 pot sizes by 4 leaf sizes	dissimilarity rating
	Colors					
15	Indow and Uchizono (1960)	Chips I	21	Munsell colors	varying in hue and chroma over wide range	distance rating
16	Indow and Uchizono (1960)	Chips II	21	Munsell colors		distance rating
17	Indow and Kanazawa (1960)	Chips III	24	Munsell colors	varying in hue, value and chroma over wide range	distance rating
	Sounds					
18	Terbeek (Note 7)	Vowels	12	varying in four linguistic features	results from English speaking listeners	triadic comparison
19	Bricker and Pruzansky (Note 8)	Sine Waves	12	4 x 3 factorial design	4 modulation frequencies and 3 modulation percentages	dissimilarity rating
20	Bricker and Pruzansky (Note 8)	Squares Waves	12	4 x 3 factorial design	4 modulation frequencies and 3 modulation percentages	dissimilarity rating

In all other sets, the skewness of the observed distances is better matched by ADDTREE than by KYST. A similar pattern emerges from the inspection of the elongation values. The elongation in the nonfactorial conceptual data is better matched by ADDTREE than by KYST, with one exception (3). In contrast, the elongation in six out of nine perceptual data sets is better matched by KYST than by ADDTREE.

In Figure 10 we plotted elongation against skewness for all data sets. The plotting character (circle or square) indicates which scaling *solution* gave a better fit as measured by r_L^2. The empty circle and square indicate the mean skewness and elongation for the plane and tree data obtained in the simulation study for sets of 36 points with no added error. Note that the line in Figure 10 neatly separates the data sets that are better fit by KYST from those that are better fit by ADDTREE.

Summary and Discussion

In this paper we investigated two of the most common representations of proximities, planes and trees, in order: (a) to develop guidelines for comparing these representations using traditional indices of goodness-of-fit and: (b) to discover properties that could help diagnose which representation is more appropriate for a given set of data.

Table 3: Goodness-of-fit and diagnostic indices for 20 sets of real data

DATA DESCRIPTION	GOODNESS-OF-FIT r_L^2 K	r_L^2 A	r_M^2 K	r_M^2 A	SKEWNESS DATA	SKEWNESS SOLUTION K	SKEWNESS SOLUTION A	ELONGATION DATA	ELONGATION SOLUTION K	ELONGATION SOLUTION A
CONCEPTUAL										
1. Vehicles	.68	.84	.86	.89	-1.13	.38	-1.19	.68	.59	.66
2. Sports	.65	.81	.84	.86	-1.28	.27	-1.46	.70	.64	.76
3. Vegetables	.74	.81	.85	.85	-0.88	.31	-1.14	.67	.64	.78
4. Furniture	.65	.83	.79	.87	-1.09	.47	-1.0	.64	.54	.60
5. Tools	.60	.78	.72	.82	-0.94	.42	-0.96	.62	.57	.60
6. Weapons	.79	.91	.90	.93	-0.52	.56	-0.26	.73	.63	.66
7. Fruits	.65	.82	.80	.87	-1.03	.48	-1.07	.66	.58	.61
8. Animals	.75	.83	.86	.88	-1.11	-0.07	-1.49	.76	.67	.77
9. Occupations	.76	.93	.92	.95	-1.16	.11	-1.29	.82	.73	.91
10. Varied Objects	.87	.96	.92	.97	-1.09	-0.60	-1.19	.80	.76	.85
11. Students (F)	.82	.59	.89	.72	-0.09	-0.36	-0.89	.53	.68	.79
PERCEPTUAL										
FORMS										
12. Letters	.55	.69	.77	.73	-1.36	.25	-1.31	.73	.62	.73
13. Polygons (F)	.60	.76	.73	.88	-0.24	.18	-0.75	.68	.59	.84
14. Plants (F)	.67	.54	.79	.66	-0.02	.26	-0.47	.57	.56	.76
COLORS										
15. Chips I	.95	.82	.97	.85	.26	.25	.02	.63	.61	.62
16. Chips II	.94	.91	.96	.93	.33	.41	.12	.59	.58	.61
17. Chips III	.94	.80	.96	.85	.13	.01	-0.16	.67	.70	.76
SOUNDS										
18. Vowels	.88	.83	.95	.88	-0.47	-0.02	-0.62	.64	.65	.69
19. Sine Waves (F)	.85	.57	.90	.68	-0.20	.17	-0.44	.52	.57	.73
20. Square Waves (F)	.92	.70	.98	.78	-0.15	.34	-0.30	.62	.52	.59

The first part of the paper dealt with artificial data. We generated, randomly, both additive similarity trees and two-dimensional configurations of points and varied the set size and the amount of added error. Interpoint distances were analyzed using two popular scaling methods KYST [Kruskal, Young & Seery, Note 2] which produces a spatial representation, and ADDTREE [Sattath & Tversky, 1977] which represents such data as an additive tree. Using both linear and monotone indices of fit, we found that for each error level and set size, tree data were *always* fit better by ADDTREE than by KYST and, conversely, plane data were always fit better by KYST than by ADDTREE. In general, the appropriate model accounted for 10 to 20 percent more variance than did the inappropriate model. Furthermore, a two-dimensional KYST solution fit three-dimensional spatial representations better than ADDTREE. For all set sizes and error levels ADDTREE fit tree data as well as KYST fit plane data. Hence, the models are, in some sense, comparable.

When the inappropriate scaling model was applied to the data, KYST appeared to accommodate tree data slightly better than ADDTREE handled plane data. The application of ADDTREE to plane data tended to produce unbalanced trees (see Figure 4a), while the application of KYST to tree data tended to produce clustered configurations (see Figure 3b). Although such configurations could arise naturally, their presence may indicate an application of the inappropriate model. ADDTREE fit was more sensitive to the perturbation of small distances than to the perturbation of large distances. The opposite pattern was observed for KYST. This is consistent with the finding of Graef and Spence

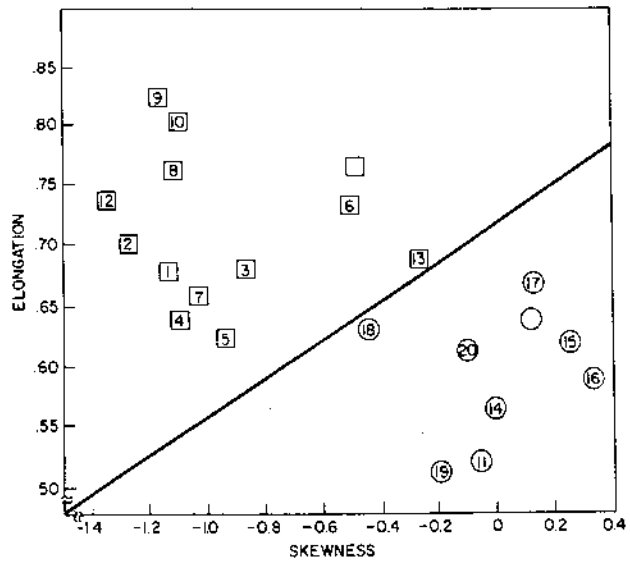

Figure 10.
The measure of elongation plotted against the skewness measure for 20 sets of real data. A circle indicates that the KYST solution gave a better fit as measured by r_L^2; a square indicates that the ADDTREE solution gave a better fit. The number corresponds to the data set number described in Table 2. The open circle and square indicate the mean skewness and elongation for the plane and tree data obtained in the simulation study for sets of 36 points with no added error.

[1979] that large distances are critical to satisfactory performance of a nonmetric multidimensional scaling algorithm while small distances play a less crucial role.

We explored two properties of the *data* that might help distinguish between trees and planes: skewness based on the third central moment of the distribution of distances, and elongation, measured by the proportion of elongated triangles among all triples in a set of interpoint distances. The results of the simulation showed that trees were quite negatively skewed and became more symmetric as noise level increased. In constrast, plane data exhibited slight positive skewness. The elongation measure was higher for trees than for planes; this difference decreased as noise level increased.

We also computed the skewness and elongation of the distances obtained by KYST and ADDTREE. We found that KYST solutions were positively skewed for both plane and tree data and ADDTREE solutions were generally negatively skewed. KYST and ADDTREE also differed in elongation when applied to the appropriate data. The proportion of elongated triangles exceeded .75 for ADDTREE solutions of tree data and was below .65 for KYST solutions of plane data, for all noise levels.

In the second part of this paper we applied KYST and ADDTREE to judgments of proximity between stimuli collected by other investigators. The results showed that most data sets could be clearly classified as favoring either a tree or a two-dimensional spatial representation. In general, colors, sounds and factorial structures were better represented by a plane, whereas conceptual stimuli from various semantic fields were better modelled by a tree. Most of the latter stimuli were generated by selecting the most frequent and/or familiar instances of the respective semantic categories. The sucess of the models, evidently,

depends both on the nature of the stimuli (perceptual vs. conceptual) and on the method used to select them. Factorial designs favor planes while "natural selection" favors trees.

As in the simulation study, the different indices were generally compatible. The data that were better fit by the tree exhibited substantial negative skewness and high elongation while the data that were better described by the plane yielded very small negative or positive skewness and lower elongation. The data sets could be separated by a straight line in the skewness and elongation plane, as shown in Figure 10.

In summary, the present results suggest the following conclusions:

(i) KYST and ADDTREE can be used to compare and test the two-dimensional euclidean model against an additive similarity tree; the appropriate model was always superior to the inappropriate model. Furthermore, over a wide range of set size and error level, KYST fit plane data about as well as ADDTREE fit tree data.

(ii) Two diagnostic measure, skewness and elongation, appear promising for distinguishing between the models. Trees tend to produce negative skewness and a high proportion of elongated triangles. Planes tend to produce lower elongation and positive skewness.

(iii) The proposed measures of goodness-of-fit, of skewness and of elongation can help diagnose whether an observed set of proximity data is better represented as a tree or a plane. The present data support the general hypothesis that the plane is more appropriate for either perceptual or factorial data, while the tree is more appropriate for nonfactorial conceptual stimuli.

Several comments regarding the generality of these conclusions are in order. First, the results obtained in the simulation study depend on the choice of a generating process. Although the configurations generated in the simulation resemble those obtained from real data, different methods for generating planes and trees could produce different results. Second, the majority of real data sets analysed in this paper could be readily diagnosed as favoring a plane or a tree representation. Other data sets may be more difficult to diagnose. In such cases, both representations might be useful if they highlight different aspects of the data. Ultimately, the choice of a representation depends, in addition to goodness-of-fit, on the interpretability and the theoretical interest of the proposed solution. Third, it is important to note that the measures of skewness and elongation are invariant under a linear but not under a monotonic transformation. Consequently, inferences based on these measures rely on interval, and not merely ordinal, information. Finally, we hope that future research will explore, both theoretically and empirically, additional diagnostic properties [e.g., Schwarz & Tversky, 1980], and apply them to richer and more complex proximity models [e.g., Carroll, 1976; Tversky, 1977].

REFERENCE NOTES

1. Furnas, G. W. The construction of random, terminally labelled, binary trees. Unpublished paper, Bell Laboratories, 1981.
2. Kruskal, J. B., Young, F. W., & Seery, J. B. How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Unpublished paper, Bell Laboratories, 1973.
3. Mervis, C. B., Rips, L. J., Rosch, E., Shoben, E. J., & Smith, E. E. Personal communication, 1980.
4. Kraus V. Personal communication, 1976.
5. Hutchinson, W. & Lockhead G. Personal communication, 1979.
6. Gati, I. Personal communication, 1980.
7. Terbeek, D. A cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics, 1977, 37.*
8. Bricker, P. D. & Pruzansky, S. A comparison of sorting and pair-wise similarity judgment techniques for scaling auditory stimuli. Unpublished paper, Bell Laboratories, 1970.

REFERENCES

Carroll, J. D. Spatial, nonspatial and hybrid models for scaling. *Psychometrika*, 1976, *41*, 439–463.

Carroll, J. D. & Pruzansky, S. Discrete and hybrid scaling models. In E.D. Lantermann & H. Feger (Eds). *Similarity and Choice*, Bern: Hans Huber, 1980.

Fillenbaum S. & Rapoport, A. *Structures in the subjective lexicon*. New York: Academic Press, 1971.

Gati, I. & Tversky, A. Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 1982, *8*, 325–340.

Ghosh, B. Random distances within a rectangle and between two rectangles. *Calcutta Mathematical Society Bulletin*, 1951, *43*, 17–24.

Graef, J., & Spence, I. Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, 1979, *86*, 60–66.

Henley, N. M. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Behavior*, 1969, *8*, 176–184.

Indow, T. & Uchizono, T. Multidimensional mapping of Munsell colors varying in hue and chroma. *Journal of Experimental Psychology*, 1960, *59*, 321–329.

Indow, T. & Kanazawa, K. Multidimensional mapping of Munsell colors varying in hue, chroma and value. *Journal of Experimental Psychology*. 1960, *59*, 330–336.

Kuennapas, T., & Janson, A. J. Multidimensional similarity of letters. *Perceptual and Motor Skills*, 1969, *28*, 3–12.

Sattath, S., & Tversky, A. Additive Similarity Trees, *Psychometrika*, 1977, *42*, 319–345.

Schwarz, G., & Tversky, A. On the reciprocity of proximity relations. *Journal of Mathematical Psychology*, 1980, *22*, 157–175.

Shepard, R. N. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 1974, *39*, 373–421.

Tversky, A. Features of Similarity. *Psychological Review*, 1977, *84*, 327–352.

Manuscript received 3/24/81
Final Version received 12/22/81

Appendix A. r_L^2 *measure*: median and interquartile range from KYST and ADDTREE analyses for two data representations, three set sizes, three noise levels and a random condition for each set size. The random condition is listed under the tree representation for display purposes only. It has no particular data representation.

TREE REPRESENTATION									
		ADDTREE				KYST			
SET SIZE	QUARTILE	*NOISE LEVEL*				*NOISE LEVEL*			
		0	.25	.5	RANDOM	0	.25	.5	RANDOM
36 Points	Q_1	1.00	.82	.68	.21	.83	.67	.57	.14
	Q_2	1.00	.83	.70	.24	.86	.71	.58	.17
	Q_3	1.00	.83	.72	.25	.88	.72	.61	.18
24 Points	Q_1	1.00	.82	.70	.31	.85	.71	.61	.24
	Q_2	1.00	.84	.72	.32	.87	.74	.64	.25
	Q_3	1.00	.84	.74	.34	.91	.76	.65	.27
12 Points	Q_1	1.00	.86	.78	.39	.80	.72	.69	.34
	Q_2	1.00	.88	.80	.46	.86	.79	.72	.39
	Q_3	1.00	.89	.83	.55	.92	.83	.76	.48

PLANE REPRESENTATION									
		ADDTREE				KYST			
SET SIZE	QUARTILE	*NOISE LEVEL*				*NOISE LEVEL*			
		0	.25	.5		0	.25	.5	
36 Points	Q_1	.75	.62	.53		1.00	.81	.69	
	Q_2	.80	.65	.56		1.00	.82	.70	
	Q_3	.80	.66	.60		1.00	.83	.72	
24 Points	Q_1	.77	.65	.59		1.00	.81	.69	
	Q_2	.78	.70	.60		1.00	.83	.70	
	Q_3	.83	.72	.63		1.00	.84	.73	
12 Points	Q_1	.81	.77	.69		1.00	.81	.72	
	Q_2	.91	.78	.71		1.00	.83	.75	
	Q_3	.93	.81	.79		1.00	.86	.78	

Appendix B. $r_{\hat{\mu}}^2$ *measure*: median and interquartile range from KYST and ADDTREE analyses for two data representations, three set sizes, three noise levels and a random condition for each set size. The random condition is listed under the tree representation for display purposes only. It has no particular data representation.

TREE REPRESENTATION									
		ADDTREE				KYST			
SET SIZE	QUARTILE	*NOISE LEVEL*				*NOISE LEVEL*			
		0	.25	.5	RANDOM	0	.25	.5	RANDOM
36 Points	Q_1	1.00	.84	.73	.46	.92	.72	.61	.19
	Q_2	1.00	.85	.75	.49	.93	.75	.63	.21
	Q_3	1.00	.85	.75	.53	.95	.77	.65	.23
24 Points	Q_1	1.00	.86	.76	.49	.94	.78	.68	.31
	Q_2	1.00	.88	.77	.53	.95	.80	.70	.34
	Q_3	1.00	.88	.80	.57	.96	.82	.71	.35
12 Points	Q_1	1.00	.91	.85	.62	.96	.84	.82	.53
	Q_2	1.00	.93	.88	.68	.98	.90	.86	.59
	Q_3	1.00	.94	.90	.72	.98	.92	.88	.67

PLANE REPRESENTATION									
		ADDTREE				KYST			
SET SIZE	QUARTILE	*NOISE LEVEL*				*NOISE LEVEL*			
		0	.25	.5		0	.25	.5	
36 Points	Q_1	.79	.68	.59		1.00	.84	.73	
	Q_2	.82	.69	.61		1.00	.84	.74	
	Q_3	.82	.70	.65		1.0	.85	.76	
24 Points	Q_1	.82	.72	.65		1.0	.85	.75	
	Q_2	.83	.75	.67		1.0	.87	.76	
	Q_3	.85	.78	.69		1.0	.87	.77	
12 Points	Q_1	.86	.84	.77		1.0	.91	.84	
	Q_2	.94	.86	.80		1.0	.92	.87	
	Q_3	.96	.87	.85		1.0	.93	.90	

Appendix C. *Skewness and Elongation measures from plane and tree data*: median and interquartile range for three set sizes, three noise levels and a random condition for each set size.

SKEWNESS MEASURE

SET SIZE	QUARTILE	TREE DATA			PLANE DATA			RANDOM
		NOISE LEVEL			*NOISE LEVEL*			
		0	.25	.5	0	.25	.5	
36 Points	Q_1	−.64	−.45	−.36	.05	.02	.03	−.05
	Q_2	−.45	−.41	−.24	.13	.08	.08	.01
	Q_3	−.35	−.30	−.17	.19	.15	.15	.12
24 Points	Q_1	−.52	−.43	−.27	.02	−.03	−.08	−.03
	Q_2	−.47	−.27	−.21	.15	.16	.07	.10
	Q_3	−.32	−.19	−.17	.31	.27	.18	.14
12 Points	Q_1	−.63	−.48	−.59	−.13	−.07	−.26	−.13
	Q_2	−.51	−.43	−.53	.01	.13	−.02	.01
	Q_3	−.41	−.14	−.24	.35	.26	.10	.15

ELONGATION MEASURE

SET SIZE	QUARTILE	TREE DATA			PLANE DATA			RANDOM
		NOISE LEVEL			*NOISE LEVEL*			
		0	.25	.5	0	.25	.5	
36 Points	Q_1	.73	.64	.59	.61	.58	.55	.49
	Q_2	.78	.65	.61	.63	.60	.57	.50
	Q_3	.79	.67	.61	.66	.61	.58	.50
24 Points	Q_1	.74	.61	.59	.61	.56	.55	.48
	Q_2	.77	.64	.61	.64	.59	.57	.50
	Q_3	.80	.70	.62	.67	.62	.59	.51
12 Points	Q_1	.73	.60	.63	.63	.56	.56	.48
	Q_2	.79	.63	.64	.63	.58	.60	.53
	Q_3	.90	.68	.66	.70	.63	.63	.54

Appendix D. *Skewness measure from solutions*: median and interquartile range from KYST and ADDTREE analyses for two data representations, three set sizes, three noise levels and a random condition for each set size. The random condition is listed under the tree representation for display purposes only. It has no particular data representation.

PLANE REPRESENTATION

SET SIZE	QUARTILE	ADDTREE				KYST			
		NOISE LEVEL				*NOISE LEVEL*			
		0	.25	.5	RANDOM	0	.25	.5	RANDOM
36 Points	Q_1	−.64	−.68	−.76	−2.89	.07	.03	−.02	.02
	Q_2	−.50	−.54	−.62	−2.80	.11	.06	.06	.04
	Q_3	−.37	−.46	−.50	−2.54	.24	.15	.15	.06
24 Points	Q_1	−.52	−.53	−.67	−2.19	.03	−.01	.03	.12
	Q_2	−.48	−.47	−.58	−1.61	.15	.09	.08	.13
	Q_3	−.32	−.36	−.49	−1.43	.27	.29	.24	.16
12 Points	Q_1	−.73	−.73	−.94	−1.32	−.16	−.06	−.12	.01
	Q_2	−.51	−.64	−.72	−.97	.07	.08	.18	.16
	Q_3	−.41	−.34	−.24	−.74	.32	.23	.28	.21

TREE REPRESENTATION

SET SIZE	QUARTILE	ADDTREE			KYST		
		NOISE LEVEL			*NOISE LEVEL*		
		0	.25	.5	0	.25	.5
36 Points	Q_1	−.24	−.38	−.54	.02	.02	.02
	Q_2	−.09	−.27	−.36	.10	.14	.13
	Q_3	−.05	−.10	−.24	.17	.19	.23
24 Points	Q_1	−.08	−.18	−.48	−.02	.09	.13
	Q_2	−.05	−.09	−.39	.10	.18	.19
	Q_3	.18	.06	−.16	.25	.30	.32
12 Points	Q_1	−.33	−.31	−.54	−.13	−.07	.07
	Q_2	−.14	−.06	−.28	−.01	.06	.16
	Q_3	.39	.18	−.13	.33	.32	.34

Appendix E. *Elongation measure from solutions*: median and interquartile range from KYST and ADDTREE analyses for two data representations, three set sizes, three noise levels and a random condition for each set size. The random condition is listed under the tree representation for display purposes only. It has no particular data representation.

TREE REPRESENTATION

SET SIZE	QUARTILE	ADDTREE NOISE LEVEL 0	.25	.5	RANDOM	KYST NOISE LEVEL 0	.25	.5	RANDOM
36 Points	Q_1	.73	.71	.72	.62	.68	.66	.64	.63
	Q_2	.78	.75	.74	.64	.72	.67	.66	.63
	Q_3	.79	.78	.77	.67	.75	.70	.69	.64
24 Points	Q_1	.74	.69	.72	.66	.70	.64	.62	.60
	Q_2	.77	.74	.73	.67	.71	.68	.65	.61
	Q_3	.80	.80	.79	.69	.75	.71	.69	.61
12 Points	Q_1	.73	.69	.68	.55	.66	.61	.54	.54
	Q_2	.79	.73	.73	.64	.75	.65	.59	.57
	Q_3	.90	.85	.83	.67	.85	.70	.66	.59

PLANE REPRESENTATION

SET SIZE	QUARTILE	ADDTREE NOISE LEVEL 0	.25	.5		KYST NOISE LEVEL 0	.25	.5	
36 Points	Q_1	.63	.64	.62		.62	.61	.61	
	Q_2	.66	.66	.64		.64	.63	.62	
	Q_3	.68	.69	.68		.66	.66	.65	
24 Points	Q_1	.62	.61	.64		.62	.59	.59	
	Q_2	.64	.65	.64		.64	.62	.60	
	Q_3	.71	.69	.65		.68	.65	.63	
12 Points	Q_1	.65	.62	.65		.63	.59	.56	
	Q_2	.66	.66	.67		.65	.60	.60	
	Q_3	.73	.70	.81		.70	.63	.66	

PSYCHOMETRIKA—VOL. 47, NO. 1.
MARCH, 1982

A MULTIDIMENSIONAL SCALING MODEL FOR THE SIZE-WEIGHT ILLUSION

TERRENCE R. DUNN

UNIVERSITY OF MELBOURNE

RICHARD A. HARSHMAN

UNIVERSITY OF WESTERN ONTARIO

The kinds of individual differences in perceptions permitted by the weighted euclidean model for multidimensional scaling (e.g., INDSCAL) are much more restricted than those allowed by Tucker's Three-mode Multidimensional Scaling (TMMDS) model or Carroll's Idiosyncratic Scaling (IDIOSCAL) model. Although, in some situations the more general models would seem desirable, investigators have been reluctant to use them because they are subject to transformational indeterminacies which complicate interpretation. In this article, we show how these indeterminacies can be removed by constructing specific models of the phenomenon under investigation. As an example of this approach, a model of the size-weight illusion is developed and applied to data from two experiments, with highly meaningful results. The same data are also analyzed using INDSCAL. Of the two solutions, only the one obtained by using the size-weight model allows examination of individual differences in the strength of the illusion; INDSCAL can not represent such differences. In this sample, however, individual differences in illusion strength turn out to be minor. Hence the INDSCAL solution, while less informative than the size-weight solution, is nonetheless easily interpretable.

Key words: individual differences, multidimensional scaling, three-mode factor, INDSCAL, size-weight illusions.

Introduction

Models for three-way MDS

Several models have been proposed for studying individual differences in multidimensional scaling. The first such model was the Tucker and Messick [1963] "points of view" approach, based on an Eckart and Young [1936] resolution of the N by $n(n-1)/2$ matrix of interpoint distances. This model has been superceded by more general models which have overcome weaknesses pointed out by Ross [1966].

The weighted euclidean model. Horan [1969] proposed an individual differences model for multidimensional scaling in which the subjects gave different weights to the axes of a common stimulus space. Thus if d_{jki} is the psychological distance for person i between stimulus j and stimulus k, then the model can be written as

$$d^2_{jki} = \sum_{t=1}^{r} w^2_{it}(b_{jt} - b_{kt})^2 \tag{1}$$

where b_{jt} is the projection of stimulus j on dimension t, and w_{it} is a weight indicating the importance person i gives to dimension t. This representation has come to be known as the weighted euclidean model. Horan [1969] developed a procedure for estimating the configuration of points (i.e., a set of b_{jt} values) given dissimilarity matrices from several subjects,

This paper is based on the first author's doctoral dissertation at the Department of Psychology, University of Illinois at Urbana-Champaign. The aid of Professor Ledyard R Tucker is gratefully acknowledged.

Requests for reprints should be sent to T. R. Dunn, California State University, Division of Information Systems, 5670 Wilshire Blvd., Los Angeles, CA. 90036.

0033-3123/82/0300 – 2666$00.75/0
© 1982 The Psychometric Society