J Forensic Sci, Oct. 1975, Vol. 20, No. 4

J. I. Thornton,[1] *D. Crim.*

The Use of an Agglomerative Numerical Technique in Physical Evidence Comparisons

There is an element of subjectivity to virtually all types of examinations conducted in the forensic laboratory. The level of this subjectivity may range from relatively low, as in the interpretation of certain instrumental results, to relatively high, as in the comparison of handwriting or firearms evidence. It is generally held, however, that comparisons of physical evidence should be as objective as possible.

To achieve a higher degree of objectivity, considerable thought has been given to the use of statistical techniques which presumably would minimize the possibility of bias on the part of the examiner or allow a more facile interpretation of observed data [*1–4*]. Such efforts, however, have not gained currency in most forensic laboratories, particularly those attempting to employ probabilistic models of one type or another. The construction of a formal probability model against which an item of evidence, appropriately analyzed, could be tested for "goodness of fit" in a large population is often rejected for several reasons. In many instances, no justification exists for the assumption of independence of the respective variables. If the variables cannot be treated as independent, then any probability model is necessarily complex [*5*]. Although in some limited cases a model could be constructed to countenance dependent variables, certain arbitrary assumptions may have to be made which may not hold for even slightly differing conditions.

Another consideration, which is legal rather than scientific, is that courts are rather consistent in their disdain for "trial by mathematics," that is, for the application of formal probability models to law-science matters [*6*]. The admission into evidence of such models, however sincere in their presentation, has been held as grounds for reversal of lower court decisions [*7*]. The reluctance of courts to accept statistical evidence, however, may not reflect a bias against statistical evidence in general as much as a recognition that probabilistic models are often less than realistic and are not truly applicable to the situation at hand.

Agglomerative numerical techniques, involving progressive fusion based upon Euclidean distance measurements, would appear to offer certain advantages over other statistical methods in the comparison of physical evidence. It is the suggestion of this author that this *nonprobabilistic* approach may be applicable to the comparison and interpretation of a number of diverse types of physical evidence. These methods, such as those used in numerical taxonomy, are hypothesis generating systems. They may have predictive validity,

Contribution 170 from the Criminalistics-Forensic Science Program, University of California, Berkeley, Calif. Received for publication 13 Feb. 1975; accepted for publication 3 April 1975.

[1] Assistant professor of forensic science, School of Criminology, University of California, Berkeley, Calif.

but differ from probability systems in that probability systems require initial null hypotheses for grouping procedures which are then tested by means of probability theory.

The initial step in this approach is to employ an algorithm to calculate the extent of dissimilarity between evidence and exemplar samples, or the extent of dissimilarity among exemplar samples. A number of similarity coefficients are available [*8*]. In the present study four similarity indices were considered: (1) squared Euclidean distance, (2) mean character distance after character standardization, (3) nonmetric coefficient, and (4) the Canberra metric coefficient. Squared Euclidean distance was eliminated due to its sensitivity to single aberrant character values. The nonmetric coefficient was eliminated since it is not additive over attributes. In this study, the coefficient used was that of Lance and Williams [*9*], now generally known as the "Canberra Metric Similarity Coefficient" [*10*]. The coefficient is an expression of distance in a Euclidean hyperspace, with the dissimilarity D_{ij} between the *i*th and *j*th sites given by

$$D_{ij} = \sum_{k=1}^{p} \left[\frac{|X_{ik} - X_{jk}|}{X_{ik} + X_{jk}} \right] \Big/ p \qquad (1)$$

where X_{ik} and X_{jk} stand, respectively, for the values of the *k*th property for the *i*th and *j*th sites, and p is the number of properties tested. By dividing the summation series by p, the dissimilarity coefficient will lie between 0 and 1. The similarity coefficient S_{ij} is then the complementary of the dissimilarity coefficient.

$$S_{ij} = 1 - D_{ij} \qquad (2)$$

A value approaching unity will then indicate a high degree of affinity between objects or sites being compared, with all properties being considered. A value approaching zero indicates little agreement. This coefficient describes a space whose properties can be explored, and satisfies certain criteria involving symmetry, distinguishability of nonidenticals, and indistinguishability of identicals [*11*]. Since this coefficient defines a model based upon a Euclidean space, several advantages result: (1) the first, and perhaps the ultimate, consideration is that it is generally applicable to a great number of diverse situations, (2) Euclidean models facilitate hierarchical arrangements, and (3) Euclidean systems are represented in our everyday experience, and our developed intuition enables us to grasp their properties more readily than more abstract models.

A FORTRAN program (available from the author) was written to compute the similarity coefficients for a maximum of 100 samples and 150 characters. The characters may consist of any numerical property, such as density, refractive index, or percentage element composition. When the coefficients are computed for all sample pairs, a matrix is constructed by an appropriate arrangement of the format of the computer output. This matrix is symmetrical about its principal diagonal, so that only one half of the matrix need be considered. The diagonal itself, consisting of a comparison of a sample with itself, has the numerical value of 1 and may be disregarded.

With the similarity coefficients now computed, the samples are grouped hierarchically according to a "nearest neighbor" sorting strategy, and a dendrogram is constructed. The dendrogram is a useful graphic device for illustrating the divergence of properties among entities (for example, objects, organisms, sites, or samples), thereby illustrating the similarity, or lack thereof, among the entities under consideration.

The process begins with the construction of an array of the coefficients, beginning with the most similar pair. All other coefficients are then examined and subjected to the following strategy [*12*].

1. If neither of the members of the next most similar pair is represented in the previously constructed group, they are designated as forming a new group.

2. If one member of the next most similar pair is already in a group, the other member of this pair is added to the group.

3. If both members of the next most similar pair are in different groups, the two groups are joined.

4. If both members of the next most similar pair are in the same group, the pair is discarded and the next most similar pair is considered.

This process is continued until all samples are fused into a single group. The dendrogram is then constructed according to the similarity coefficient representing the point where a specific sample joins the cluster.

It should be noted that this is not the only sorting strategy which could be employed in the construction of a hierarchical arrangement of sample pairs. Both centroid and nearest-neighbor sorting strategies were considered [*9*]; the nearest-neighbor strategy was selected for its simplicity. For forensic purposes, it is not likely that groups will be picked off the hierarchical diagram at any arbitrary level of similarity, but only at high levels of similarity for which nearest-neighbor sorting procedures are entirely adequate.

Three examples of the utility of this approach to the interpretation of physical evidence will be given. The first is concerned with the similarity of primate hair, the second with glass, and the third with soil.

Hair

Table 1 illustrates the raw data pertaining to the morphological examination of primate hair. The data have been abstracted from the work of Rosen [*13*]. Confining our attention to these species and these measurements, the computation of similarity coefficients and the construction of a dendrogram yield a relationship depicted in Fig. 1. On the basis of these six characters, the primate hair most similar to that of man is that of the baboon, and the least similar is that of the lemur. The length of the vertical shaft of the dendrogram gives an assessment of the *extent* of this similarity or dissimilarity; a more expanded discussion of this aspect will be presented in the glass example following.

TABLE 1—*Metrical data pertaining to primate hair* (*after Rosen* [13]).

Species	Mean Max Diameter, μm	Mean Min Diameter, μm	Mean Hair Index	Mean Scale No. per 40 μm	Mean Scale Width, μm	Mean Scale Index
Man	91.85	64.65	71.65	4.95	12.80	16.00
Gorilla	109.45	86.55	79.28	5.25	9.00	8.50
Chimpanzee	122.96	96.44	79.02	4.22	10.22	8.54
Baboon	103.38	78.62	76.63	3.44	12.00	11.94
Lemur	35.03	23.94	69.28	5.82	8.00	23.76

Glass

Table 2 illustrates the density, refractive index, color, and hardness of the first ten glass samples listed in the study by Nelson [*14*]. Consider now a hypothetical situation in which the first nine of these samples are construed as exemplar samples; the tenth sample will be designated as an evidence sample, E_v. It is examined and the values of 2.537, 1.521, 14, and 4 are obtained for the density, refractive index, color, and hardness, respectively. Inspection of Table 2 shows that the "evidence" fragment is not *identical* to any of the "exemplar" samples with regard to these four properties. The questions which may now be asked, however, are whether the evidence sample is *similar* to any of the exemplar samples and, if so, *how* similar and to *which* samples. The Canberra metric similarity coefficient and the

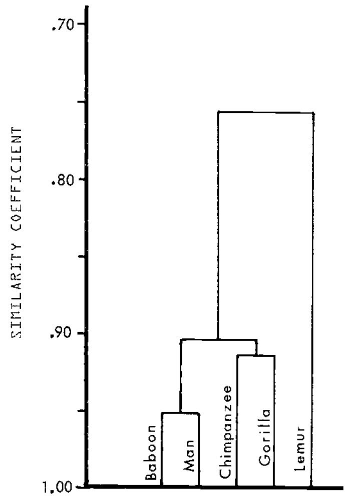

FIG. 1—*Dendrogram illustrating similarity in primate hair, based upon the metrical data of Table 1.*

TABLE 2—*Properties of the first ten glass samples reported by Nelson* [14]. *The tenth sample is arbitrarily considered to represent an evidence* (E_v) *sample.*

Sample	Density at 20°C, g/ml	Refractive Index at 20°C	Color Group No.	Hardness Group No.
1	2.559	1.515	13	2
2	2.526	1.518	14	1
3	2.596	1.518	12	2
4	2.556	1.514	14	5
5	2.536	1.521	13	2
6	2.550	1.513	13	3
7	2.557	1.515	12	6
8	2.521	1.518	15	3
9	2.525	1.519	15	6
10	2.537	1.521	14	4

resulting dendrogram will be of assistance in answering these questions. Computer manipulation of the four variables will result in the matrix of coefficients shown in Table 3. From this table a dendrogram may be constructed, as shown in Fig. 2. Inspection of the dendrogram provides the basis for the objective assessment of similarity. It is noted that the evidence sample is most similar to Exemplar Sample 5, Exemplar Sample 1 is the next most similar, and Exemplar Sample 2 is the least similar. Exemplar Samples 7, 9, and 4 are very similar to one another, but in a relative sense are not particularly similar to Samples 1, 5, and E_v.

Soil

It has been suggested that the complexion of soil enzyme activity may characterize a soil as having originated from a given geographical location. In work reported previously [*15*], the author has illustrated a dendrogram involving five variables (enzyme activity) and 28 sites. For a systematic comparison of this number of samples and properties, the dendrogram appears to be extremely useful. Although an assessment of similarity may be made

TABLE 3—*Similarity coefficients computed from the data of Table 2.*

	Sample									
J/K	1	2	3	4	5	6	7	8	9	E_v (10)
1	1.00000	.89908	.98127	.88268	.99330	.94748	.86455	.98122	.84833	.99426
2	.89908	1.00000	.88184	.82543	.90433	.85895	.79432	.91547	.81233	.90482
3	.98127	.88184	1.00000	.86398	.97603	.92875	.86582	.96398	.83115	.97698
4	.88268	.82543	.86398	1.00000	.87733	.92664	.95757	.88376	.96027	.87828
5	.99330	.90433	.97603	.87733	1.00000	.94484	.85875	.98647	.85407	.99905
6	.94748	.85895	.92875	.92664	.94484	1.00000	.90460	.93276	.89154	.94580
7	.86455	.79432	.86582	.95757	.85875	.90460	1.00000	.84669	.96386	.85970
8	.98122	.91547	.96398	.88376	.98647	.93276	.84669	1.00000	.86518	.98696
9	.84833	.81233	.83115	.96027	.85407	.89154	.96386	.86518	1.00000	.85408
10 (E_v)	.99426	.90482	.97698	.87828	.99905	.94580	.85970	.98696	.85408	1.00000

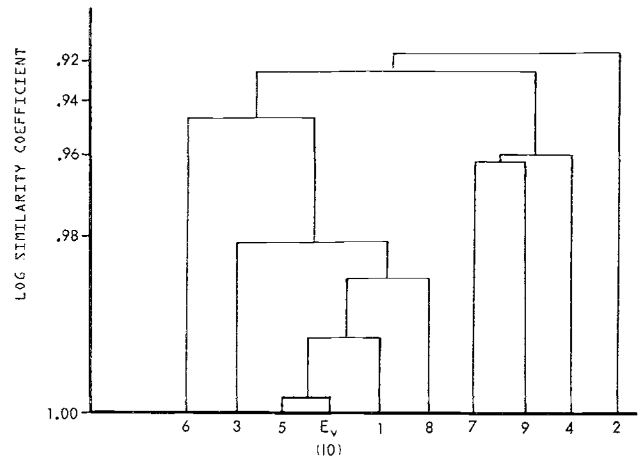

FIG. 2—*Dendrogram illustrating the degree of similarity of the ten glass samples of Table 2. Samples 10* (E_V) *and 5, with a coefficient of 0.99905, are the most similar.*

directly from the raw data, such "eyeball" comparisons are subjective, somewhat nonquantitative, and increasingly difficult to make as the data matrix becomes large.

The similarity coefficient method may also be used without the construction of a dendrogram; a matrix of similarity coefficients may clearly suggest groupings of entities possessing similar properties. Table 4 depicts phosphatase and arylsulfatase K_m (Michaelis constant) values developed from the examination of ten soils from two separate locations, one group being a Columbia sandy loam and the other a Hanford sandy loam. Six of the most dissimilar Columbia soils studied [*15*] were compared with four Hanford soils by means of the Canberra metric similarity coefficient. Figure 3 illustrates the matrix of similarity coefficients. In this instance an inspection of the matrix shows that two distinct groups of soils are represented, with no overlap. The most dissimilar site-pair within either group is between C^{26} and C^{28}, with a coefficient of 0.74. The highest coefficient between any two soils of opposite group (that is, Columbia versus Hanford) is 0.59, observed in the C^{28}-H^G pair.

TABLE 4—*Phosphatase and arylsulfatase* K_m *values developed from examination of ten soils from separate locations* [15].

Soil	K_m Values	
	Phosphatase[a]	Arylsulfatase[b]
Columbia		
01	2.02×10^3M	3.14×10^4M
23	2.83×10^3M	4.12×10^4M
24	3.37×10^3M	3.04×10^4M
25	2.66×10^3M	2.86×10^4M
26	3.42×10^3M	1.88×10^4M
28	2.39×10^3M	3.86×10^4M
Hanford		
38	4.24×10^3M	3.23×10^4M
46	4.02×10^3M	3.86×10^4M
48	4.38×10^3M	3.14×10^4M
G	2.34×10^3M	3.44×10^4M

		Columbia Soils						Hanford Soils			
		C^{01}	C^{23}	C^{24}	C^{25}	C^{26}	C^{28}	H^{38}	H^{46}	H^{48}	H^{G}
Columbia Soils	C^{01}	1.0									
	C^{23}	.85	1.0								
	C^{24}	.87	.88	1.0							
	C^{25}	.91	.89	.93	1.0						
	C^{26}	.75	.77	.88	.83	1.0					
	C^{28}	.91	.94	.85	.90	.74	1.0				
Hanford Soils	H^{38}	.41	.51	.53	.47	.50	.47	1.0			
	H^{46}	.41	.51	.53	.47	.51	.46	.94	1.0		
	H^{48}	.40	.51	.52	.46	.49	.46	.98	.93	1.0	
	H^{G}	.55	.56	.49	.54	.46	.59	.84	.84	.83	1.0

FIG. 3—*Matrix of similarity coefficients of a Columbia soil-Hanford soil composite sample, based on the* K_m *values indicated in Table 4. Six Columbia soils and four Hanford soils are compared by means of the Canberra metric similarity coefficient; the sole characters used in the comparison are phosphatase and arylsulfatase Michaelis constants. The soils clearly form two discrete groups, with no overlap.*

If the number of characters or properties being observed in the comparison is large, the resulting hierarchical arrangement will represent what Sneath and Sokal [*8*] term a "natural" classification, and may have the greatest predictive powers for general purposes. If the number of characters is small, however, or if the characteristics are redundant or highly correlated, then the dendrogram represents a convenient but arbitrary contrivance. This by no means destroys the value of the technique, but it should be recognized that other, equally valid, techniques of assessing similarity may exist.

With this technique, it is not necessary that the sampling be representative of a general population or constitute a carefully randomized sample. Also it is not necessary that the properties measured possess any *known* frequency distribution. Since the hierarchical technique does not countenance the relationship of dependent and independent variables, as is necessary for regression analysis, characters may be added to the matrix which have an undetermined correlation with one another. Hence, in the example of soil comparisons, color as measured by Munsell value and chroma, percentage composition, or any other character capable of being reduced to a numerical value may be accommodated in the Canberra metric similarity coefficient. The resulting dendrogram may be refined to the extent necessary by adding additional characters.

In the assessment of dendrograms for similarity of properties in samples, however, two important considerations should be recognized. First, nothing in the dendrogram indicates sampling density. If the sampling density is not uniform, the dendrogram will be skewed accordingly. Second, while the length of the stem in the dendrogram denotes dissimilarity, independent information would be needed to translate this dissimilarity into subjective reality. Hence, green beer bottle glass and clear window glass could join a given cluster at a similarity coefficient of 0.99 or 0.09, depending upon the characteristics selected for comparison. For proper interpretation, the data must originate from the same data matrix also.

The author believes that the Canberra metric similarity coefficient approach has many attractive features for the comparison of diverse types of physical evidence in the forensic laboratory. As this statistical approach is nonprobabilistic in nature, it may find the approbation in its application to law-science matters that probability models have been consistently denied.

Acknowledgment

The author is grateful to Professor A. D. McLaren for helpful discussion.

References

[1] Kingston, C. R. and Kirk, P. L., "The Use of Statistics in Criminalistics." *Journal of Criminal Law, Criminology and Police Science,* Vol. 55, No. 4, Dec. 1964, pp. 514-521.
[2] Biasotti, A. A., "A Statistical Study of Individual Characteristics of Fired Bullets," *Journal of Forensic Sciences,* JFSCA, Vol. 4, No. 1, Jan. 1959, pp. 34-50.
[3] Parker, J. B., "The Mathematical Evaluation of Numerical Evidence," *Journal of the Forensic Science Society,* Vol. 7, No. 3, July 1967, pp. 134-144.
[4] Parker, J. B., "A Statistical Treatment of Identification Problems," *Journal of the Forensic Science Society,* Vol. 6, No. 1, Jan. 1966, pp. 33-39.
[5] Schlesinger, H. L., Lukens, H. R., Guinn, V. P., Hackleman, R. P., and Korts, R. F., "Special Report on Gunshot Residues Measured by Neutron Activation Analysis," Report AT (04-3)-167, Division of Isotope Development, U.S. Atomic Energy Commission, Accession No. GA-9829, National Technical Information Service, Springfield, Va., 1970.
[6] Tribe, L. H., "Trial by Mathematics: Precision and Ritual in the Legal Process," *Harvard Law Review,* Vol. 84, No. 6, April 1971, pp. 1329-1393.
[7] *People vs. Collins,* 68 Cal. 2d 319, 438 P. 2d 33, 66 Cal Rptr 497 (1968).
[8] Sneath, P. H. A. and Sokal, R. R., *Numerical Taxonomy,* Freeman & Co., San Francisco, 1973.
[9] Lance, G. W., and Williams, W. T., "Computer Programs for Hierarchical Polythetic Classification ('Similarity Analyses'), *Computer Journal,* Vol. 9. No. 1, May 1966, pp. 60-64.
[10] Webster, R. and Burrough, P. A., "Computer-Based Soil Mapping of Small Areas from Sample Data. I. Multivariate Classification and Ordination," *Journal of Soil Science,* Vol. 23, No. 2, 1972, pp. 210-221.
[11] Williams, W. T. and Dale, M. B., "Fundamental Problems in Numerical Taxonomy," *Advances in Botanical Research,* Vol. 2, 1965, pp. 35-68.
[12] Williams, W. T., and Lambert, J. M., and Lance, G. N., "Multivariate Methods in Plant Ecology. V. Similarity Analyses and Information-Analysis," *Journal of Ecology,* Vol. 54, No. 2, July 1966, pp. 427-445.
[13] Rosen, S. I., "Identification of Primate Hair," *Journal of Forensic Sciences,* JFSCA, Vol. 19, No. 1, Jan. 1974, pp. 109-112.
[14] Nelson, D. F., "The Identification of Lucas 700 Headlampglass Fragments by their Physical Properties," *Analyst,* Vol. 84, No. 999, June 1959, pp. 388-392.
[15] Thornton, J. I. and McLaren, A. D., "Enzymatic Characterization of Soil Evidence," *Journal of Forensic Sciences,* JFSCA, Vol. 20, No. 4, Oct. 1975, pp. 674-692.

School of Criminology
University of California
Berkeley, Calif. 94720