Michael John Lewis,[1] *Ph.D.*

Patterns: Their Storage, Retrieval, and Comparison

REFERENCE: Lewis, M. J., **"Patterns: Their Storage, Retrieval, and Comparison,"** *Journal of Forensic Sciences*, JFSCA, Vol. 32, No. 5, Sept. 1987, pp. 1281-1292.

ABSTRACT: It is recognized that the unique definition of a complex pattern in terms of a numerical descriptor is an unrealistic objective; yet the need exists for an efficient means of handling such information in a variety of fields. A simple procedure is described for abstraction of the essential characteristics of a pattern into a format suitable for automated searching of large reference collections, and potentially, for the interlaboratory exchange of this information. The index developed embodies two distinct aspects of a pattern: "intrinsic" information content (I) and "hierarchical" information (H), calculated on up to 15 of the pattern's elements; the descriptors relate respectively to weighted relative magnitudes and arrangement of elements within the pattern. Significantly, the index is designed to be immune from variability in longitudinal dimensions, both absolute and relative, and to provide for versatile and intelligent searching of a data base.

KEYWORDS: forensic science, patterns, information systems, comparative analysis, data handling, profiling

Comparison of patterns plays an important role in many aspects of the scientific investigation of crime, ranging from the examination of fingerprints and footwear marks through to the chemical profiling of drugs and electrophoretic examination of phenotypes. These patterns can be classified according to one of two broad groups: those which relate to *area*, for example, contact marks; and those which are of a *sequential* univariant form, such as the analog or digital output of a detector. It is the second of these groups that is to be considered here: such patterns are encountered in a diversity of guises.

The essential value of a pattern lies in its capacity to convey complex information about its source; thus, potentially allowing highly specific characterization of that source. It may also contain information of a variable nature arising from its means of production—we shall return to this later. In comparing patterns, therefore, we are involved in the comparison of information from different sources, with the aim of testing for a link between them. When interpreting our findings the inherent significance of the information must be considered: could it, by chance, have originated from another source altogether? An answer to this important question may be provided in one of two ways. If the information relating to the source is of a random nature then statistical probability can be invoked to qualify it. On the other hand, if the information is not of a purely random character, then proper assessment of its significance must necessarily involve reference to a knowledge base. Until now, it has often been necessary to search hard copy collections of patterns; this is a cumbersome procedure which becomes increasingly unsatisfactory as a collection grows.

Commercial interests have led to the development of data handling procedures for certain

Received for publication 8 July 1986; accepted for publication 17 Nov. 1986.

[1]Forensic scientist, Home Office Forensic Science Laboratory, Chorley, Lancashire, United Kingdom.

patterns of the sequential form, notably for mass fragmentation and infrared spectra. However, the techniques used in those cases are not readily transferable to the handling of other sequential patterns which in some respects are less well-defined. Despite the potential value of automated pattern handling in a diversity of areas of forensic science concern, so far very little progress appears to have been made in its development.

In gas chromatography, McCowan and coworkers [*1*] have reported an attempt to remove subjectivity in the comparison of pairs of complex chromatograms. The technique used was to generate for each of the chromatograms two matrices: an "*r*-matrix," composed of relative retentions, and a "*Q*-matrix," describing relative peak areas. Subtraction of the corresponding matrix pairs gave measures of the superimposability of the two chromatograms. More recently, Osman and coworkers [*2*] reported an application of the matrix technique in the comparison of samples of *Cannabis* resin; in that case, only relative retentions were considered. Unfortunately, it would appear that the parameter yielded by this procedure is rather less than definitive of a match or difference; and, in dealing with noncorresponding peaks in a given pair of chromatograms, subjective decisions are not avoided. Another serious limitation is that the method does not lend itself to the storage and retrieval of useful data.

An alternative approach is seen in the work of Huizer [*3*], who was concerned with the characterization of heroin mixtures examined by liquid chromatography. Accurate quantitation of certain peaks in the chromatogram allowed comparisons to be made between different samples of the drug. Although useful, the technique is somewhat time-consuming and the efficient handling of the data generated presents difficulties (problems of this type have been addressed by others [*4*]). More importantly, in a general context, the method is limited to those applications in which specific pattern features can be reliably identified for use.

Described in the following is a completely generalized approach to the handling of all patterns of a sequential form. It is based on the premise that the comparison of patterns *does* in fact demand careful, albeit subjective, assessment by the experienced eye; and that it can never—with confidence—be completely reduced to the comparison of simple descriptors of some kind. Instead, the objective is to provide an index of sufficient character to allow the rapid abstraction of comparable patterns from a computer data base for further examination.

Information Reduction

In Fig. 1, a pattern is represented as a continuously varying function of some measured characteristic with respect to distance. A numerical approximation to this pattern—including information on profiles—would clearly require a large volume of data; but if we are

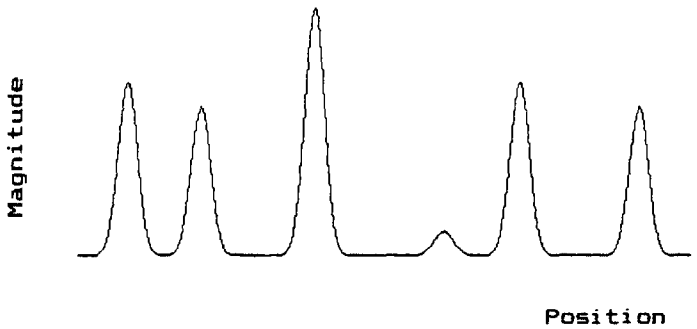

FIG. 1—*An example of a sequential pattern.*

simply concerned with the indexing of a series of similar patterns, much of this information is redundant. As in dealing with spectra, therefore, individual "band" profiles can be ignored and the pattern reduced to a sequence of position (abscissa) and magnitude (ordinate) information describing the maxima. However, whereas spectra are typified and handled by their closely defined positional data, other sequential patterns are not so well-defined in this respect. For a variety of reasons associated with its mode of production, a pattern's overall length and the relative spacing of its elements may be prone to variation. Spacing differences between patterns may not, therefore, denote differences in source related information; accordingly, our system must avoid discrimination in this respect.

It is concluded that the "positional" information, used in spectroscopy, is an encumbrance when dealing with other types of sequential pattern. Bearing in mind again that we seek only to index the pattern, little is to be lost by completely discarding this information, simply retaining the *order* of the pattern's elements. Thus we are left with an ordered set of data (see Fig. 2), whose individual values describe the magnitudes or relative magnitudes of the pattern's features. In practice, the values could represent band intensities, peak heights

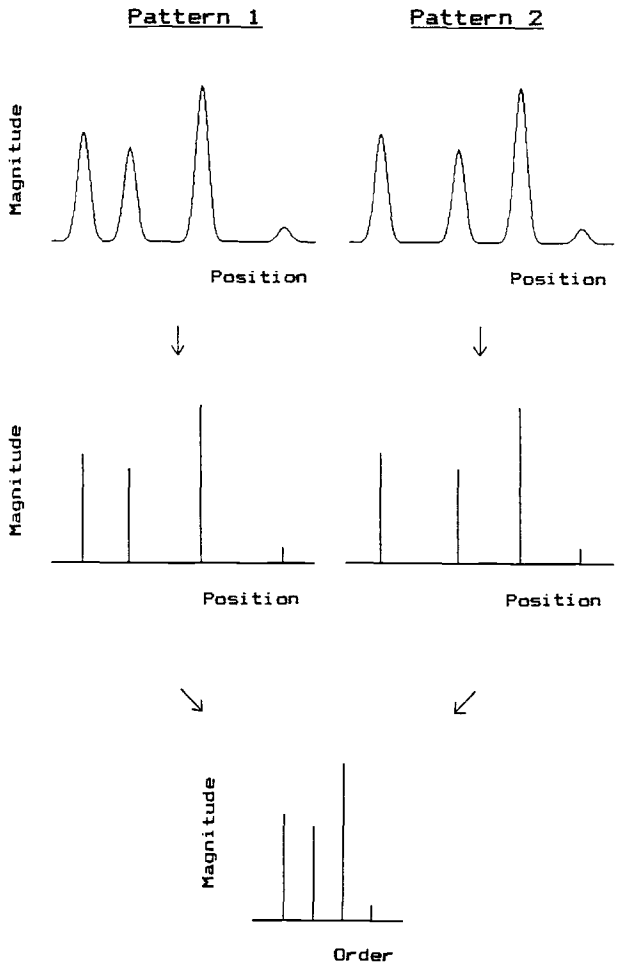

FIG. 2—*Information reduction: pattern to ordered data set, ignoring spatial separation.*

or areas, groove widths, or even—where a convention has been established—an assortment of attributes having no natural order of their own. The diagrams used in this work, although suggestive of spectrometry or chromatography, therefore have no fixed connotation.

Coding

While it would be possible to store the information directly in the form of a series of relative magnitudes, substantial advantages are to be gained by further processing the data to abstract its essential characteristics which may then be concisely expressed in a format suitable for intelligent use.

There are two distinct aspects to the data set that we have obtained from the pattern. One relates to information contained within the data as a whole: namely, the randomness or variation in magnitude present in its component values. This will be referred to as the intrinsic element, I, of our index. It will be shown later how appropriate selection of functions can enable maximum discrimination to be achieved. The second and independent aspect of the set is the order of the data: this can be expressed in the form of a simple hierarchical representation, H, consisting of a sequence of symbols. Effectively, we have separated the two dimensions of our data; Fig. 3 illustrates the principle.

Before we can apply this treatment, account has to be taken of the fact that the total number of elements will vary from example to example. In order that all patterns belonging to a given family can be handled in a uniform manner, a concept of "base" is needed. This is defined to be the maximum number of pattern elements to be considered for the purpose of indexing. Thus, with a base (b) of ten in use, a pattern will be characterized on the ten most major elements within it. In general, the optimum choice of base will be closely related to the number of significant features which are typically encountered in the particular pattern type; it does not, however, mean that patterns having a larger number of features cannot be effectively dealt with using the same base. By limiting the base selectable to fifteen, the hierarchical descriptor can always be represented as a string composed of hexadecimal characters (1 to F). As will be seen later on, this facilitates the identification of similar—but nonidentical—patterns.

Intrinsic Descriptor, *I*

This descriptor is concerned with the variety or diversity present in the component values of the data set. There is no preordained means of quantifying such a notion and our purpose

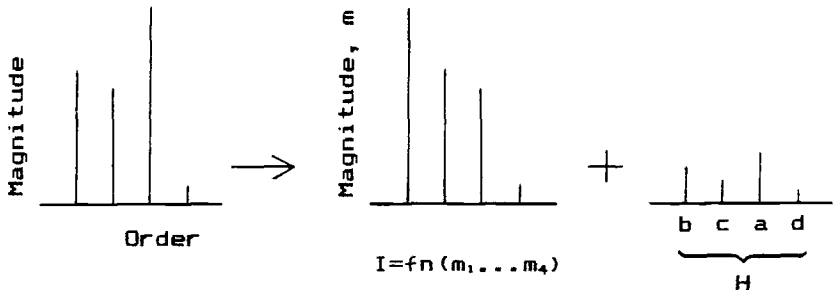

FIG. 3—*Separation of the two components of the data set: magnitude information, to be described by* I, *and hierarachical structure,* H. *In this illustration the position of* a *in the string identifies the position of the major feature in the pattern,* b *the next, and so on.*

could, for example, be served by adopting any of the several measures used by ecologists in describing species diversity [5]. One of the most common used in this field is based on Simpson's index [6]:

$$C = \sum_{i=1}^{i=n} p_i^2$$

where n is the number of species and p_i is the proportional abundance of the ith species relative to the total abundance; diversity is measured as the reciprocal of C. Another measure, taken from information theory, is Shannon's entropy [7]:

$$H = -\sum_{i=1}^{i=n} p_i \log p_i$$

(here, H is to be distinguished from the hierarchical descriptor used in this work). Other formulae exist, but it has been demonstrated by Hill [8] that most are related to one another; and, in the opinion of Routledge [9], little extra information is to be gained by their use. To apply these indices to quantify the intrinsic information content of our data set, p_i would relate the individual magnitudes to their total; but, unfortunately, the formulae are less than satisfactory for our purpose.

Simpson's index gives greatest weight to the major elements in the set, and is therefore regarded as an abundance measure; with Shannon's entropy, the minor elements are weighted most heavily. In practice, it is found that the numerical values obtained in both cases tend to lie towards the upper end of the range that is potentially available for the given data set; this has the detrimental effect of limiting the discrimination afforded. Furthermore, both formulae have the disadvantage of being virtually unresponsive to *variation* in the largest values present, and sometimes, much exaggerated sensitivity to changes in the smallest. A demonstration of these rather subtle and undesirable aspects of the two formulae is included in the Appendix.

Functions

In view of the limited suitability of these standard indices for characterization of our data set, new functions were developed. They permit fuller use of a defined range of values for the intrinsic descriptor and allow the control of sensitivity in selected regions of the magnitude scale. This latter quality will be seen to have considerable utility.

The approach adopted uses *relative*, rather than proportional, magnitudes as a basic parameter. No loss of significance occurs because the total number of elements, which is embodied in the proportional measure, is already known—being given by the hierarchical descriptor. Practically, it places more importance on the accurate measurement of the principle feature of the pattern (since the others are to be related to it), but it offers gains in other respects. The intrinsic descriptor is to be defined by

$$I = \frac{\sum_{i=1}^{i=b} fn(r_i) - 1}{b - 1} \tag{1}$$

where $fn(r_i)$ is some function of the magnitude of the ith element of the set relative to the principle element; and b is the base. Given that $fn(r_i)$ is arranged to have a value in the

range zero to unity, this expression may be interpreted as being a measure of the average of the weighted values of the set, excluding the principle feature (defined to be unity) which in effect is removed before averaging. Thus, a single-feature "pattern" will have an I of zero; and a pattern whose b elements are all of the same size will have an I equal to unity; between the two extremes there will be a steady progression, influenced by the nature of $fn(r_i)$.

If $fn(r_i)$ were to be replaced simply by r_i, then the value of I would vary linearly with change in any r_i, providing uniform sensitivity over the full range. However, considerably more information can be abstracted from the data set by using weighting functions that increase sensitivity in different parts of the magnitude scale. Three functions are suggested that allow this:

$$L(r_i) = \sin \frac{\pi r_i}{2} \tag{2}$$

which enhances discrimination at the low end of the scale; and

$$U(r_i) = 1 + \sin \frac{\pi (r_i - 1)}{2} \tag{3}$$

which accentuates differences at the upper end; and

$$M(r_i) = \frac{1}{2} + \frac{1}{2} \sin \frac{\pi (2r_i - 1)}{2} \tag{4}$$

a function which gives maximum sensitivity in the mid range. These weighting functions are displayed graphically in Fig. 4 together with their first derivatives, which indicate the regions of greatest sensitivity for each.

It will be appreciated that, when functions L, U, and M are used in turn in Expression 1, emphasis is being given to different aspects of the data set; the measures obtained will be denoted I_L, I_U, and I_M. All will tend to increase with increasing total of r, and their values will fall in the order $I_L \geq I_M \geq I_U$. Beyond this latter constraint they are independent, conveying different characteristics of the data set. Consequently, to obtain maximum characterization, use should be made of all three values in defining the intrinsic descriptor. An illustration of their independence of one another will be seen later in Fig. 5.

Hierarchical Descriptor, *H*

In the section that introduced coding it was pointed out that, having obtained a measure of the relative magnitudes within the data set, the only independent information remaining is the actual order of the data. Patterns such as chromatograms have a sense or direction by which that order can be unambiguously defined. Other patterns lack a natural sense but usually may be ordered according to some characteristic; for example, by beginning at the end closest to the data set's center of gravity.

The hierarchical descriptor is simply an ordered string of b hexadecimal characters; a value of hex(b) being assigned to represent the principle element, hex($b-1$) the next, and so on. For example, working with a base 10, a descriptor of "97634A2518" would indicate that the largest feature (A being hexadecimal for 10) is the sixth element, the next largest the first element, and so on. In coding a pattern in which two elements happen to be of equal magnitude, the first one encountered is accorded superiority; and, if fewer than b features are available in a particular example, the balance of characters is made up of terminal zeros.

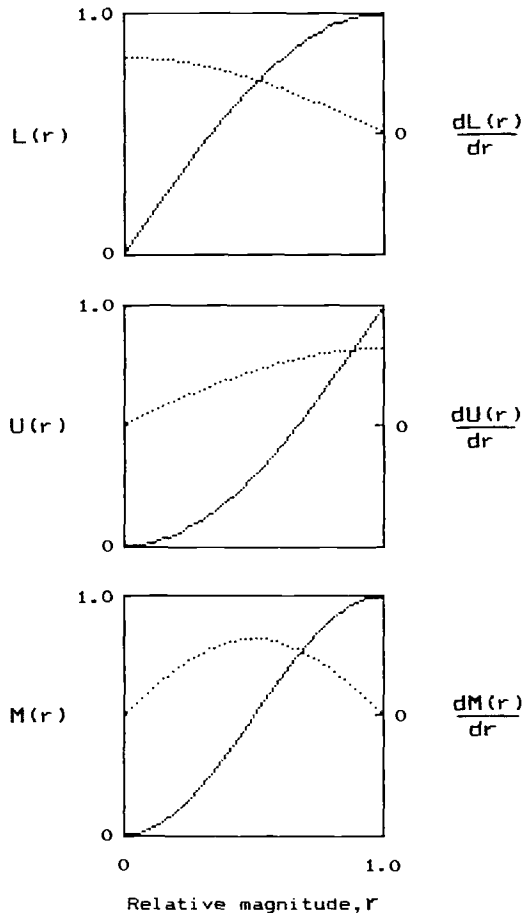

FIG. 4—*Weighting functions* L, U, *and* M *employed in characterizing the relative magnitude data; for formulae see Expressions 2, 3, and 4. The solid curves relate to the functions and the dotted curves to their first derivatives (indicating change of weighting with* r*).*

The Data Base, Practical Aspects

Our sequential pattern has now been reduced to simple parameters: three numbers, to describe intrinsic information content, and a sequence of hexadecimal characters describing H. For convenience, these will be combined into one string to give a composite index (IH_b); at the same time allocating contiguous space to contain general information concerning the pattern's origin, and so forth. Thus, the whole record of index plus data is contained in a single block of string space. The most logical way of filing it is to place the records in order of one of the three I values, usually choosing that which affords best discrimination for the family of patterns concerned.

To optimize the use of data space and to produce a standard format for the index, the number of digits used for the intrinsic descriptors needs to be designated—the length of H is already fixed. Conveniently, the value given by Expression 1 may be multiplied by a factor of 999, so that each I will now run from 000 to 999. Examples of this working index are given in Fig. 5 together with the data sets from which they are derived.

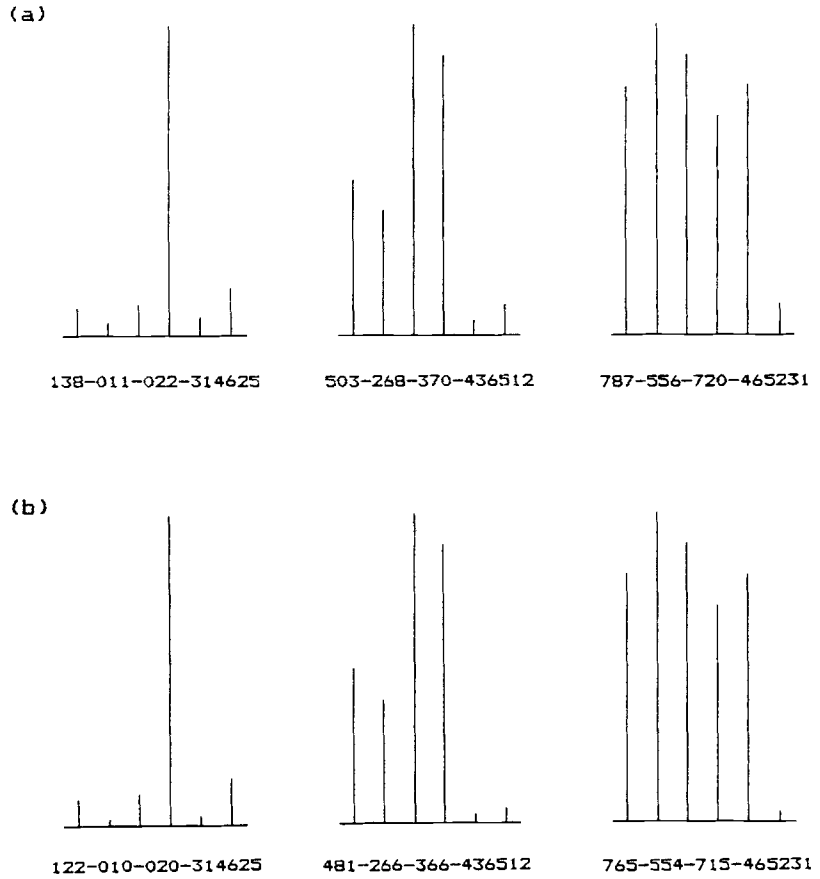

FIG. 5—*Examples of patterns and their descriptors* IH_b *(composed of* I_L, I_U, I_M, *and* H, *computed on the six elements shown). Set* (b) *differs from set* (a) *in the relative sizes of some of the minor features, illustrating the consequent changes in intrinsic descriptors;* H *is unchanged in the examples given.*

Searching

The nature of the index described permits various means of conducting a library search. Most simply, the data base will first be searched according to the value of one of the intrinsic descriptors; those records falling within a selected window then being further examined for match with the other *I*s and for correspondence of *H*. Another of several possible options is to commence with *H*, in which case the full data base must be searched. The flexibility available allows the user to locate, not only matching patterns, but patterns which may differ in any of several different respects. For example, by placing less reliance on I_L, patterns which differ only in the smallest features will not be excluded from consideration.

To allow criteria to be set for identification of hits, the discrepancy between the indices of subject and library records needs to be quantified. This is done by comparing corresponding parts of the pair concerned: for the intrinsic descriptors it is given simply by the difference between two numbers; for the hierarchical descriptor the discrepancy is measured by sum-

mation of the absolute differences in the individual hexadecimal components at corresponding positions in the two Hs. The process is as follows:

	I_L	I_U	I_M	H
Subject index:	426	122	223	97634A2518
Library index:	429	144	255	A763492518
Discrepancies:	3	22	32	1000010000 = 2

In the case of intrinsic descriptors, the acceptable discrepancy in each will be governed by pattern reproducibility (that is, in the relative magnitudes of its features) and the accuracy of its measurement. With the hierarchical descriptor, a perfect match of order will clearly result in a discrepancy of zero. Where a reversal in the relative sizes of two features occurs, the difference will be 2, while changes in three of them will give a value of 4. The maximum possible difference in H is $b(b + 3)/2 - 2$, which can occur when a pattern of b elements is compared with one having a single feature.

Thus far, the encoding of a pattern and its comparison with a reference collection can be accomplished rapidly and *fully automatically;* the product being a short list of candidate matches. It is at this point that system operator becomes actively involved.

As discussed earlier, no parameter can be relied upon to define uniquely a complex pattern, and, in the view of the author, final comparison properly falls to the experienced eye—advantage was taken of that in formulating the index. However, before consulting hard copy patterns, a preliminary inspection of the candidate matches may usefully be performed on the computer. Graphical representation of a pattern requires, of course, that the record associated with its index must include the necessary data. Although the index itself is calculated on a limited number of elements, many more may be included here for the purpose of visual inspection: an allocation of 100 bytes, for example, will enable the display of up to 50 pattern elements detailing both their magnitude and position (discarded for the index), making rapid visual comparison practicable. For certain purposes this may suffice without the need to proceed to definitive paper records.

In any area in which decisions have to be taken, inevitably borderline cases will arise; that is not least true in pattern handling. Criteria need to be set to determine the presence or absence of pattern elements, and attention given to the question of resolution. When faced with a pattern for which the prescribed limits are approached, reassessment should be made against the data base following suitable editing of the elements on which the index is produced. In this way, the fullest possible search can be ensured.

Conclusion

By appropriate management of the information contained within a pattern, concise and meaningful descriptors can be derived, enabling its automated storage, retrieval, and comparison. Application to a drug profiling system confirms that the index described in this paper provides a powerful means of pattern characterization, allowing simple and effective handling of complex patterns. Details of that work will be fully reported elsewhere.

APPENDIX

Efficacies of Mathematical Functions

Simpson's index and Shannon's entropy are both influenced by the number of data being considered. To facilitate comparison between them they require to be normalized in respect of their maximum permissible values. In the present context, this can be regarded as simply scaling the results, because the number of data in use will be known.

Shannon's entropy maximizes when all values in the data set are the same ($p_1 = p_2 = \ldots = p_n = 1/n$), when it becomes equal to $\log n$. The function itself, divided by this maximum, has been used to measure evenness (J), which has a maximum value of 1. It will be recalled that with Simpson's index (C), the reciprocal is used to measure diversity; that is, $D = 1/C$. This also maximizes when all the data are equal, the value then being simply n. D/n will therefore be used in like manner to J in the comparison to be made.

In Fig. 6, examples of the behavior of these functions are shown. On each diagram two plots are given, relating to two—rather different—sets of data: each of them is composed of five fixed values, with the sixth (r) being given by the abscissa. This enables observation of the change in function value (ordinate) with the variation of a single member of the set in all regions of the magnitude scale. First derivatives (dotted curves) are included to highlight regions of sensitivity. It is seen that, although the two sets of data chosen represent near extremes of character that could exist in a set, the spans of D/n and J are limited to about three quarters and two thirds of the scale, respectively. More seriously, both functions can be recursive (seen in this example with J), and they are always most susceptible to influence from the smaller members of the set. Indeed, in the case of Shannon's index, there is hypersensitivity in this region, and conversely, complete insensitivity towards differences among the larger values of the set.

For direct comparison, Fig. 7 shows the result of inserting the same data in the expressions I_L, I_U, and I_M introduced in this paper. These functions are seen to provide progressive rather than recursive change with r and to use the whole of the available range of values. Furthermore, the controlled weighting provided in three different regions allows comprehensive characterization of all the relative magnitude information within the data set.

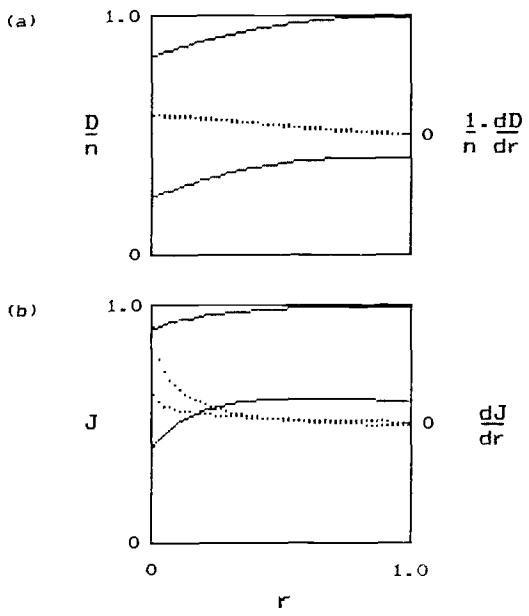

FIG. 6—*An illustration of the characteristics of standard diversity measures:* (a) *Simpson's index and* (b) *Shannon's entropy—both normalized with respect to* n *(see text). In each diagram, the lower of the two solid curves relates to the data set 1.0, 0.05, 0.05, 0.05, 0.05,* r; *and the upper one to the set 1.0, 0.9, 0.9, 0.9, 0.9,* r. *The dotted curves are their derivatives, shown to indicate the responsiveness of the functions at different values of* r.

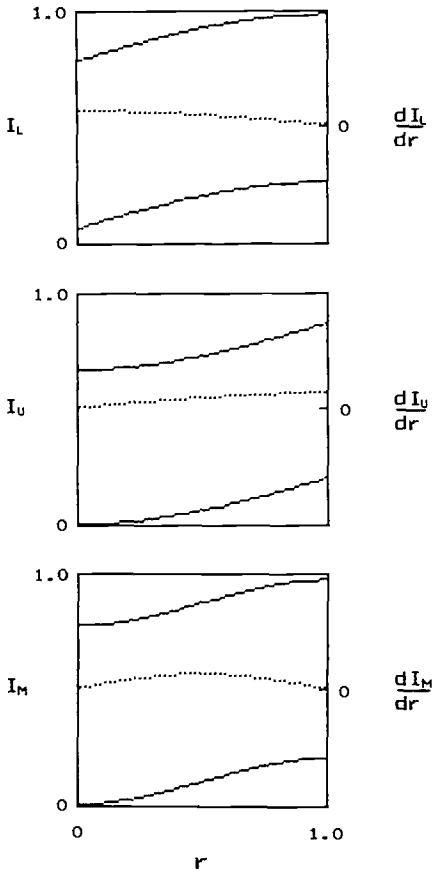

FIG. 7—*The behavior of* I_L, I_U, *and* I_M. *The data used are those used in Fig. 6.*

References

[1] McCown, S. M., Manos, C. G., Pitzer, D. R., and Earnest, C. M., "The (r,Q) Matrices—A Tool for the Manipulation of Chromatographic Patterns," *Analyst,* Vol. 107, 1982, pp. 1393-1406.

[2] Osman, A., Thorpe, J. W., and Caddy, B., "Comparison of Cannabis Samples from Different Origins by the Headspace Technique and an Assessment of Chromatographic Traces Using the r-Matrix," *Journal of the Forensic Science Society,* Vol. 25, 1985, pp. 427-433.

[3] Huizer, H., "Analytical Studies on Illicit Heroin. II. Comparison of Samples," *Journal of Forensic Sciences,* Vol. 28, No. 1, Jan. 1983, pp. 40-48.

[4] Smalldon, K. W. and Brown, C., "The Evidential Value of Multiple Continuous Measurements—a Simplified Approach to Data Analysis," *Medicine, Science and the Law,* Vol. 20, 1980, pp. 154-162.

[5] Putman, R. J. and Wratten, S. D., *Principles of Ecology,* Croom Helm Ltd., Beckenham, United Kingdom, 1984.

[6] Simpson, E. H., "Measurement of Diversity," *Nature,* Vol. 163, 1949, p. 688.

[7] Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication,* The University of Illinois Press, Urbana, 1962.

[8] Hill, M. O., "Diversity and Evenness: a Unifying Notation and its Consequences," *Ecology,* Vol. 54, 1973, pp. 427-432.

[9] Routledge, R. D., "Diversity Indices: Which Ones Are Admissible?," *Journal of Theoretical Biology*, Vol. 76, 1979, pp. 503–515.

Address requests for reprints or additional information to
Michael John Lewis
Home Office Forensic Science Laboratory
Washington Hall, Euxton
Chorley, Lancashire, PR7 6HJ United Kingdom