Nova Hedwigia, Beiheft 130, p. 285–292 June 2006

A new approach to problems in taxonomy and ecology

Eugene F. Stoermer[1*] & Janice L. Pappas[2]

[1] School of Natural Resources and Environment, University of Michigan, 430 East University Avenue, Ann Arbor, MI 48109-1115, USA, e-mail: stoermer@umich.edu
[2] Museum of Zoology, University of Michigan, 1109 Geddes Avenue, Ann Arbor, MI 48109-1079, USA

With an appendix

Stoermer, E.F. & J.L. Pappas (2006): A new approach to problems in taxonomy and ecology. – Beiheft zur Nova Hedwigia 130: 285–292.

Abstract: Diatom systematics and ecology, two of Frank Round's major research interests, contain much uncertainty. This derives from several sources. The taxonomic diversity of diatoms appears to be much greater than previously thought. Recognition of this diversity has led to realization that population-based ecological studies, although perhaps made more difficult by increased taxonomic resolution, will also likely become even more powerful. There have been significant advances in objective "quantitative" tools useful in understanding diatom taxonomy and parallel advances in tools that allow us to reduce and analyse large, complex, data sets. However, at least partly due to the limited number of investigators involved in research on diatom taxonomy and ecology, these new sources of information have proven difficult to systematize into new and improved understanding.

This has served, in some instances, to sharpen the inherent conflict between ecologists, who need a stable taxonomic framework and taxonomists most interested in exploring and explaining biodiversity, biogeography, and their underlying systematic relationships. We suggest application of methods from the rapidly advancing field of fuzzy logic to diatom studies. These methods allow explicit recognition of uncertainties and systematized incorporation of qualitative information and verbal descriptions into analysis of taxonomic and ecological problems.

Introduction

Most of the time, diatomists use methods and analyses that provide a solution labelled either true or false. On closer inspection, such solutions usually reveal partial truth. Perhaps the model or assumptions used are not completely adhered to so that the analysis encompasses uncertainty. Although diatom taxonomy and ecology do not have formalized objective rules or criteria to determine all cases in a uniform way, the available data do make it plain that revising taxonomy, in particular, is necessary, useful, and ongoing. In taxonomy it is clear that, since characters are not invariant, there will always be some uncertainty in decisions made about species delineations. In taxonomy this uncertainty, at a fundamental level, arises from evolution. The progeny of an ancestor separated in time or space or both will diverge. Human

* Corresponding author

1438-9134/01/0130-285 $ 2.00

frailties in understanding the differences arising serve to compound the basic problem. We would further argue that uncertainty is an inherent characteristic of all complex, self-replicating, systems. Successful understanding of diatom population ecology, one of Round's major interests, thus must contain explicit recognition of uncertainty, and methods for dealing with uncertainty.

In most investigations, some of the desired information about diatoms is incomplete, inaccurate, scant, or absent. There may be abundant morphologic and morphometric information available on some diatom species, but information on species-specific life cycles and cytological or genetic information may be lacking. Even when many data are available, classifying diatoms with regard to relatively simple parameters, such as salinity tolerance, is not necessarily straightforward. For example, taxa from *Surirella* are usually reported as being either freshwater or marine. What does this say about the taxonomy of this genus (Mann 1999b)? Are all species exclusively freshwater or marine? What implication does this have with regard to classifying this genus with related genera? How reliable are available literature reports?

This briefly alludes to some of the problems encountered in diatom research where uncertainty, vagueness, ambiguity, or imprecision exists. Of course all diatomists perform the mental exercise of evaluating such uncertainties in their publications. Such evaluations are dependent on the individual investigator's experience and background and thus introduce an additional degree of uncertainty themselves. The methods we advocate here are, if sufficiently specified, reproducible by other investigators, and their inputs and assumptions are explicit and can be modified to accommodate new information or information unknown to or ignored by the original analyst.

Numerous other examples could be cited. Perhaps one of the most common problems facing diatom ecologists is incorporating information from classic historical sources (e. g. Cholnoky 1968) and modern instrumental measurements. Another typical problem is incorporation of geographic information or habitat variables into ecological analyses. Although not yet, so far as we are aware, applied to diatom studies, Tran et al. (2002) have used fuzzy decision analysis to evaluate the integrated influence of habitat in assessing ecological vulnerability, and Marsili-Libelli (2004) has used fuzzy inferential analysis in the prediction of algal blooms.

Although most diatom ecologists tend to take instrumental measurements at face value, it is clear that such quantities contain uncertainties of various kinds and magnitudes. For example, for measurements of phosphorus, a nutrient element recognized as highly important in diatom ecology, the literature will quickly show that phosphorus concentrations are often reported to be "below the limit of detection". What does this mean? The answer contains a number of uncertainties including, but not necessarily limited to, when the measurement was taken, what instruments were used, what laboratory or investigator made the measurement reported, what were the sampling, preservation, and storage protocols employed. All these factors, and others, can affect the meaning of the lack of a reported value. Of course, the same caveats can also be applied to actual values reported. Further, there are real questions concerning the form of phosphorus being reported, and what fraction of the quantities reported are actually useful to diatoms or other algae (Tarapchak & Herche 1988, 1989)

Use of fuzzy logic

Categorizing information from diatom studies is quite suitable as subject matter in using fuzzy logic (Zadeh 1965, Dubois & Prade 1980, Zimmermann 1991) One way to view fuzzy logic is the application of degrees of truth with regard to measurement and decision-making. By measurement we mean that diatoms can be described with language (e.g., planktonic, pan-

duriform, epipelic), or numbers can represent various aspects of a diatom (e.g., length, number of areolae, number of plastids).

In descriptive analyses, fuzzy logic employs the use of logical constructs or if-then statements, often using the generalized *modus ponens* or "mode that affirms" rule of inference. Determinations are made using linguistic modifiers or hedges as descriptors. For example, a centric diatom that exhibits a variable range of diameters may be described by the linguistic hedges, "very large," "not as large", and "not large at all." Taxonomic decision-making can be formalized using fuzzy logic to determine degrees of truth of outcomes, say, in species designations. Fuzzy logic provides a model of approximate reasoning about a problem and the possible outcomes. That is, if a fuzzy proposition is devised, the truth-value inference about the outcome is that it is approximately true. Informally, for example, for a group of *Stephanodiscus* specimens, the value of their diameter might be described as, "their size is small." The fuzzy proposition might be: "if the specimens are very small, they may belong to another species; so, additional specimens that are found to be very small might be another species."

More formally the generalized *modus ponens* can be described by the following:

Premise: x is A'
Implication: *If x is A, then y is B*
Conclusion: y is B'

where A, A', B, and B' are fuzzy propositions. Note that A' and B' are similar to, but not necessarily identical to, A and B, respectively. The descriptive equivalent would be:

"This freshwater centric diatom is one with *very* noticeable hyaline ridges."
"If the freshwater centric diatom is one with hyaline ridges,
the freshwater centric diatom is a *Cyclotella*."
"This freshwater centric diatom is *Cyclotella*."

The linguistic modifier "very" is used in the premise.

Another obvious application is evaluating commonly used morphological terms, such as "rostrate", "capitate", "lanceolate". The fact that such terms are not precise descriptors is clear from the number and kinds of descriptive modifiers found in the literature, such as "subrostrate", "broadly capitate", "narrowly lanceolate", and innumerable other variants. When referring to data from other sources it may even be useful to qualify statements further in that not all authors apply morphological terms, of any degree of complexity, uniformly. Thus, most appropriate descriptive terms make the form of "narrowly lanceolate *sensu* Hustedt" which might differ from, for example "narrowly lanceolate *sensu* Skvortsov".

Most pennate diatoms also subtly change shape during size diminution resulting from vegetative growth following sexual reproduction. For example, it is common to find descriptions in the literature in the form of "lanceolate, becoming elliptical lanceolate in smaller specimens". It is easy to see that additional linguistic hedges could be used to refine information about shapes observed in various size categories.

Of course, precise mathematical approximations of shape can be obtained by several methods (Stoermer & Ladewski 1982, Pappas et al. 2001, du Buf & Bayer 2002). However, it is often the case that few specimens are available at any given time to an investigator, and those available may not be representative of the species' size range. One of the real advantages of using fuzzy logic, as explained below, is that it is possible to include both numerical data generated in response to a particular question and descriptive data from other sources.

To implement fuzzy logic, linguistic hedges can be translated into fuzzy sets. Raw numerical data can be converted into fuzzy sets as well. A fuzzy set consists of a number between 0 and 1. The set may include 0 and 1. For example, if 0 represents "absence" and 1 represents "presence," then numbers in between represent degrees of presence. These values constitute a fuzzy set, and the complement of this set is the set of degrees of absence. Each fuzzy set can be represented mathematically by a membership function. Fuzzy set theory is an extension of

classical set theory. As in classical set theory, fuzzy set theory incorporates, for example, conjunctive and disjunctive operators that are among the many ways to analyse such sets.

A fuzzy restriction in a fuzzy logic proposition involves a fuzzy relation that acts as a flexible restraint on the values assigned to a variable. These values are the fuzzy set as a membership function. A composition of functions or composite fuzzy propositions can be translated into a system, consisting of linguistic modifiers, fuzzy constructs, qualifiers, quantifiers, and linguistic truth-values (Bellman & Zadeh 1977).

More formally, adapting an example from Zimmermann (1991), a comparison of the size of one centric diatom to another in the same genus using the linguistic modifier "small" may be presented as the following (with a detailed solution in the Appendix):

For the universe $X = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$; each R' is a fuzzy relation.
$R'(x) = A'$ = "small diameter" = $\{(\mathbf{1},1), (\mathbf{2}, .6), (\mathbf{3}, .2), (\mathbf{4}, 0)\}$ is one fuzzy subset in X.
$R'(x, y) = B'$ = "approximately equal in diameter" = the fuzzy relation:

1	.5	0	0
.5	1	.5	0
0	.5	1	.5
0	0	.5	1

where the rows are x values and the columns are y values. $R'(y) = A'$ composition B'. To solve for the fuzzy relation $R'(y)$, use the max-min compositional rule of inference so:

$$R'(y) = \{(\mathbf{1}, 1), (\mathbf{2}, .6), (\mathbf{3}, .5), (\mathbf{4}, .2)\}$$

with the interpretation in the form of a *modus ponens*:

"x is small in diameter"
"x and y are approximately equal in diameter"
"y is more or less small in diameter."

There are other ways to use fuzzy sets. Data can be fuzzified and used in conventional analyses. This was the case in a diatom ecological study we reported and which involved the use of fuzzy coding of data to be used in canonical correspondence analysis (Pappas & Stoermer 1995). This coding was used to indicate data values near a boundary, but not completely within a given category so that multiple, overlapping fuzzy data resulted. Fuzzified morphologic and morphometric data included multiple ranges of diameters for centric diatoms, lengths and widths for pennate diatoms, surface to volume ratio; and multiple descriptive categories for silicification, symmetry, raphe structure, areolae density and pattern. Multiple ranges of environmental variables that were fuzzified included temperature tolerance, growth rate, and silica to phosphorus ratio. Multiple descriptive categories included pH and salinity tolerance, and light intensity.

Ordination of the fuzzy data produced multiple environmental and taxonomic gradients within the context of the constrained eigenvectors. In addition, within some of the environmental gradients, microgradients of morphologies and morphometries were present. These gradients produced a more detailed picture in which influences on diatom morphology and morphometry are not entirely clear-cut, depicting unsharp boundaries among the influences. Microgradients within gradients depicted absence and presence of influences at the same time in one ordination.

Fuzzy restrictions on fuzzy sets can be determined by devising fuzzy measures. One way to look at fuzzy measures is with respect to the difference between probability and possibility. Probability covers a number of areas including likelihood, frequency, and uncertainty. Possibility also encompasses uncertainty. However, there are differences between what is meant by

uncertainty with regard to probability and with regard to possibility. Uncertainty modelled by possibility involves membership and non-membership in a set simultaneously. Possibility also makes use of linguistic hedges. Sometimes, the underlying structure of numerical data is unknown or not random, so possibility is useful whereas probability does not apply. The degree to which something is possible may or may not be the same thing as the degree to which something is probable.

Rather trivial examples, but ones which occur occasionally, are reports of marine diatoms in oligosaline inland waters. If evaluated strictly in terms of probability the result would be small, but positive. If evaluated in terms of possibility, the correct conclusion, that such occurrences result from contamination due to use of marine diatomites in filters and industrial products, is more easily reached. Again experienced diatomists readily evaluate such trivial examples, but possibility analysis is useful in detecting more subtle cases.

More formally, for example, devising a fuzzy measure as a possibility measure and distribution may be given by the following:

The fuzzy proposition is: "The diatom *Asterionella ralfsii* W. Smith is tolerant of low pH conditions since it is found in such an environment."

Let N be the fuzzy proposition; then N is the variable named "*Asterionella ralfsii*" where X is on the interval [1,6], and this is the linguistic variable "pH", and "low" is a term of a term set of "pH"; (another term, "high" might be another example of a term set of "pH").

Let $\pi(x)$ be a possibility distribution induced by a fuzzy set F in X

where $\mu_F(x) = \{(1, 0), (2, .1), (3, .2), (4, .6), (5, .55), (6, .4)\}$, and let A' be the non-fuzzy (numerical) set of X on the interval [1, 6]. The possibility that x belongs to A' is $\Pi(A')$ where Π is the possibility measure induced by π:

$$\Pi(A') = \sup_{x \in A} \mu_F(x) = \sup_{x \in A} \pi(x)$$

where *sup* is the supremum (or maximum or least upper bound) (Dubois & Prade 1980). The possibility measure = 0.6, or the possibility that *Asterionella ralfsii* was found in pH 4 waters is 0.6.

As this example shows, possibility theory is related to fuzzy set theory in that membership functions may also be possibility distributions. A possibility measure is the least upper bound of a possibility distribution. An occasionally encountered example is where all reports of rare species may not encompass the actual size range. In such cases it is entirely reasonable to include the possible size range, since this is reasonably well constrained for most diatom species (Edlund & Stoermer 1997, Mann 1999a).

Possibility measures, like other fuzzy measures, have properties that determine their applicability in analyses where uncertainty exists. Fuzzy measures are non-additive while probability is additive. The additivity property does not allow for measurement error, a fact that is "conveniently ignored" in most purely probabilistic approaches. Therefore, fuzzy measures are more characteristic of realistic measurement.

To adapt probability to more realistic measurement capacity, superadditivity and subadditivity properties have been applied to devise belief and plausibility measures (Dempster 1967). These measures have been interpreted to be upper (belief) and lower (plausibility) probabilities or interval-valued probabilities. There are also special belief and plausibility measures associated with necessity and certainty measures, and some with even less restrictive properties (Wang & Klir 1992).

Fuzzy measures with less restrictive properties can be used with a variety of functions and operators. One application in diatom research has been classification integration (Pappas & Stoermer 2001), which involves using a composition of functions. In a study of *Asterionella* from the Great Lakes, fuzzified Fourier shape coefficients were ordinated. Seven shape groups were delineated, and six of those groups were divided into two regions for further

analysis. Each region of three shape groups was analysed using the fuzzy integral to determine degree of fuzzy shape group overlap and degree of certainty in assignment of specimens to shape groups. The composition of functions used was the minimum and supremum (maximum). Results indicated that shape groups overlapped to the degree of approximately the threshold value of 0.5 or slightly greater. This was expected since overlap among these similar-sized taxa indicates a similar stage in the vegetative reproductive life cycle. Degree of certainty in specimen assignment was at the threshold value of 0.5 for one specimen and greater than this value for all other specimens. Degree of certainty is more indicative of a realistic assessment of species designation than saying that a specimen "is" or "is not" definitely a given species.

Classification integration is one of many techniques in information fusion for the purpose of decision-making. In taxonomic decision-making, various pieces of information, when aggregated, produce species identification, to some degree. This process contains a good deal of uncertainty, particularly in groups such as diatoms, as evident from the employment of terms such as "*aff.*", "*cf.*", "*sensu*" and "*vide*". Although taxonomists understand such terms (however, we would not argue that all taxonomists attach precisely the same meaning) the meaning contained is not available to formal analysis. The application of fuzzy logic allows appropriate incorporation of this type of information. The "species problem" is particularly acute in diatoms, as we seem to be moving from extremely compressed and artificial taxonomic treatments towards much more expanded and natural classification schemes. It is obvious that treatments of taxonomic data that allow incorporation of more data types and explicit treatment of uncertainties are useful in constructing useful classifications. What may not be so clear is that these approaches can also help to bridge the gap between modern and classical taxonomic treatments.

Not only is information combined from, say, morphology, developmental and reproductive biology, and ecology, but also such information may originate from a number of sources, including experts in the field. For example, in taxonomy it is often difficult for beginners in the field to differentiate between an abnormal specimen of one morphological series and a specimen actually belonging to another morphological series. An expert, through experience, will recognize a number of subtle, and not necessarily uniform, clues that characterize abnormal specimens.

Fuzzy decision-making can be used in the context of multiple types of information with the decision being made by one individual or by a group of individuals (Zimmermann 1987). Any combination of types of information and involvement by individuals can be accommodated in the fuzzy decision-making process. The only limit is the amount of time needed to complete the decision-making process.

Although diatomists are trained to use quantitative methods and reject "old fashioned" opinion-based conclusions, it has become evident that there is a good deal of experiential learning that is not easily captured by "objective" measurements and observations, and considerable effort is being made to capture experience-based expertise before it is lost. This is especially important in taxonomically neglected groups, such as diatoms. We believe that fuzzy logic is one way to capture and sustain this type of information.

In fuzzy decision-making, each expert's assessment can be used as input and determined to be reliable or tested for deficiencies with regard to inaccuracies, over-cautiousness, and over-confidence (Dubois et al. 1999). Decision-making involves weighting alternatives using linguistic hedges, for example, and translating these hedges into fuzzy values for use in further analysis (Dubois & Prade 1984). If-then rules in various iterations are used in the decision-making process. Each fuzzy set can be translated into a membership function that has a graphical form. From this, some form of information fusion can be used, and the result can be translated into a linguistic solution (Tong & Bonissone 1984). The process can be simple

by using few options or complex requiring programming. Solutions mirror reality as the best explanation at the time in a form that accounts for uncertainty, vagueness, or ambiguity embodied in the data used. Fuzzy decision-making is the amalgamation of fuzzy logic, fuzzy set theory, and fuzzy measure theory.

Conclusions

Although not yet widely used in taxonomic or ecological studies, analyses using fuzzy logic have been developed and used to address problems in engineering and other applied fields where qualitative inputs and data with significant inherent uncertainties must be addressed. We argue that these approaches are particularly well suited to many problems in diatom research. They can serve to deal effectively with inherent uncertainties in data and observations derived from complex adaptive systems, be they organisms or ecosystems, and more easily incorporate the background of observational and semi-quantitative data into a systematic analytical framework.

Acknowledgements

This paper is dedicated to Frank Round, a real pioneer who explored more aspects of diatom studies than most of us can comprehend, and left large footprints on the entire field.

References

BELLMAN, R.E. & L.A. ZADEH (1977): Local and fuzzy logics. – In: DUNN, J. M. & G. EPSTEIN (eds): Modern uses of multiple-valued logic: 105–165. D. Reidel Publishing, Dordrecht, The Netherlands.

CHOLNOKY, B.J. (1968): Die Ökologie der Diatomeen in Binnengewässern. – J. Cramer, Lehre, Germany.

DEMPSTER, A.P. (1967): Upper and lower probabilities induced by a multivalued mapping. – Ann. Math. Statistics **38**: 325–339.

DUBOIS, D. & H. PRADE (1980): Fuzzy sets and systems: Theory and applications. – Academic Press, New York.

DUBOIS, D. & H. PRADE (1984): Criteria aggregation and ranking of alternatives in the framework of fuzzy set theory. – In: ZIMMERMANN, H.-J., L.A. ZADEH & B.R. GAINES (eds): Fuzzy sets and decision analysis: 209–240. Elsevier Science Publishing, Amsterdam.

DUBOIS, D., H. PRADE & R. YAGER (1999): Merging fuzzy information. – In: BEZCEK, J.C., D. DUBOIS & H. PRADE (eds): Fuzzy sets in approximate reasoning and information systems: 335–401. Kluwer, Boston.

DU BUF, H. & M.M. BAYER (2002): Automatic diatom identification. – Series in Machine Perception and Artificial Intelligence **51**: 1–328. World Scientific Publishing Co., Singapore.

EDLUND, M.B. & E.F. STOERMER (1997): Ecological, evolutionary, and systematic significance of diatom life histories. – J. Phycol. **33**: 897–918.

MANN, D.G. (1999a): The species concept in diatoms. – Phycologia **38**: 437–495.

MANN, D.G. (1999b): Crossing the Rubicon: the effectiveness of the marine/freshwater interface as a barrier to the migration of diatom germplasm. – In: MAYAMA, S., M. IDEI & I. KOIZUMI (eds): Proceedings of the Fourteenth International Diatom Symposium, Tokyo: 1–22. Koeltz Scientific Books, Koenigstein.

MARSILI-LIBELLI, S. (2004): Fuzzy prediction of the algal blooms in the Orbetello lagoon. – Environmental Modelling and Software **19**: 799–808.

PAPPAS, J.L. & E.F. STOERMER (1995): Multidimensional analysis of diatom morphologic & morphometric phenotypic variation and relation to niche. – Ecoscience **2**: 357–367.

PAPPAS, J.L. & E.F. STOERMER (2001): Fourier shape analysis and fuzzy measure shape group differentiation of Great Lakes *Asterionella* Hassall (Heterokontophyta, Bacillariophyceae). – In: ECONOMOU-AMILLI, A. (ed.): Proceedings of the 16th International Diatom Symposium: 485–501. Amvrosiou Press for the Faculty of Biology, University of Athens, Greece.

PAPPAS, J.L., G.W. FOWLER & E.F. STOERMER (2001): Calculating shape descriptors from Fourier analysis: Shape analysis of *Asterionella* (Heterokontophyta, Bacillariophyceae). – Phycologia **40**: 440–456.

STOERMER, E.F. & T.B. LADEWSKI (1982): Quantitative analysis of shape variation in type and modern populations of *Gomphoneis herculeana*. – Nova Hedwigia Beih. **73**: 347–386.

TARAPCHAK, S.J. & L.R. HERCHE (1988): Orthophosphate concentrations in lake water: analysis of Rigler's radiobioassay method. – Canad. J. Fish. Aquatic Sci. **45**: 2230–2237.

TARAPCHAK, S.J. & L.R. HERCHE (1989): Phosphate uptake by microbial assemblages: model requirements and evaluation of experimental methods. – J. Environm. Qual. **18**:17–25.

TONG, R.M. & P.P. BONISSONE (1984): Linguistic solutions to fuzzy decision problems. – In: ZIMMERMANN, H.-J., L.A. ZADEH & B.R. GAINES (eds): Fuzzy sets and decision analysis: 323–334. Elsevier Science Publishing, Amsterdam.

TRAN, L.T., C.G. KNIGHT, R.V. O'NEILL, E.R. SMITH, K.H. RIITTERS & J.D. WICKHAM (2002): Environmental assessment: Fuzzy decision analysis for integrated environmental vulnerability assessment of the mid-Atlantic region. – Environm. Managem. **29**: 845–859.

WANG, Z. & G.J. KLIR (1992). Fuzzy measure theory. – Plenum Press, New York.

ZADEH, L. (1965): Fuzzy sets. – Information & Control **8**: 338–353.

ZIMMERMANN, H.-J. (1987): Fuzzy sets, decision making, and expert systems. – Kluwer, Boston.

ZIMMERMANN, H.-J. (1991): Fuzzy set theory and its applications, 2nd revised edition. – Kluwer, Boston.

Received 10 January 2005, accepted in revised form 19 September 2005.

Appendix

The fuzzy relation example—solving for $R'(y)$:

A'	B'			
1, .6, .2, 0	*1*	.5	0	0
	.5	1	.5	0
	0	.5	1	.5
	0	0	.5	1

Take the minimum values of row vector A' and the first column of B':

1 and *1*	1
0.6 and *0.5*	0.5
0.2 and *0*	0
0 and *0*	0

(Note that the rows and columns are the same values since $R'(x, y)$ is symmetric).

Do the same with the 2nd, 3rd and 4th columns of B' with respect to A':

1, 0.5, 0, 0
0.5, 0.6, 0.2, 0
0, 0.5, 0.2, 0
0, 0, 0.2, 0

Now, take the maximum value for each row:

1
0.6
0.5
0.2

This vector is $R'(y)$.