



British Ecological Society

Marine Ecology and the Coefficient of Association: A Plea in Behalf of Quantitative Biology

Author(s): Ellis L. Michael

Source: *Journal of Ecology*, Vol. 8, No. 1 (Mar., 1920), pp. 54-59

Published by: [British Ecological Society](#)

Stable URL: <http://www.jstor.org/stable/2255213>

Accessed: 27/11/2013 21:15

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



British Ecological Society is collaborating with JSTOR to digitize, preserve and extend access to *Journal of Ecology*.

<http://www.jstor.org>

MARINE ECOLOGY AND THE COEFFICIENT OF ASSOCIATION: A PLEA IN BEHALF OF QUANTITATIVE BIOLOGY¹

BY ELLIS L. MICHAEL.

*(Scripps Institution for Biological Research of the University of California,
La Jolla, California.)*

CONTENTS.

	PAGE
1. Introductory statement	54
2. The coefficient of association	55
3. The problem of reliability	58
4. A dilemma, and the way out	59

1. INTRODUCTORY STATEMENT.

Investigations of the marine ecologist carry him, so to speak, on to a liquefied land where the vast majority of animals and plants, reduced to invisibility and immersed in an intermixed soil and subsoil, are transported hither and thither by wave, tide, and current. Not only so, but many, perhaps all, of the animal species in this heterogeneous assemblage, called plankton, rise and fall throughout the depths in differential ways. Granting an equivalent of the oak tree or pine tree association, the marine ecologist finds difficulty, not only in describing it, but, even more, in finding it. Since he can not directly witness such an association, he is compelled to rely on indirect evidence furnished by tow-net or similar apparatus. In other words, his only recourse is to measured magnitudes and application of mathematical logic thereto.

Moreover, it seems likely that the world's food supply must be sought, to ever increasing extent, in the resources of the sea. But, in so far as these resources are represented by the fisheries, their extent and value is ultimately dependent upon the plankton—the “meadows” of the sea. Knowledge of the interrelationships among plankton organisms due to food and enemies, coincidence of life-cycles, and similarity in behaviour relative to weather and hydrographic conditions, therefore, assumes an economic as well as a scientific significance. A satisfactory coefficient of association, being the initial step

¹ Read before a joint session of the Ecological Society of America, the Western Society of Naturalists, and the Cooper Ornithological Club, held in Pasadena, California, June 20, 1919.

towards evaluating the whole associational complex, is thus a matter of no small importance.

So the marine ecologist finds himself in a unique position. He is compelled to solve a mathematical problem in order to solve his ecological problems. This fact enables him, perhaps, to appreciate, more fully than his colleagues of the land, the crying need for quantitative method in biology generally. The purpose of this paper, therefore, is threefold: first, to supply what seems to be a suitable coefficient of association; second, to call attention to the necessity for a measure of its reliability, a problem not yet satisfactorily solved; and, lastly, to suggest the only effective means for insuring its solution.

2. THE COEFFICIENT OF ASSOCIATION.

Suppose n random collections be made by means of tow-net or other apparatus, such that an approximately equal volume of water is explored during each collection. If

- x of the collections obtained the two organisms A and B conjointly,
- y of them obtained neither A nor B ,
- z of them obtained A but not B , and
- w of them obtained B but not A ;

every possible associative relationship between A and B is expressible in terms of x, y, z and w . This is better grasped, perhaps, diagrammatically:

$$A \begin{array}{c|c} & B \\ \hline x & w \\ z & y \end{array}, \text{ where } x + y + z + w = n (1).$$

Without entering into detail, suffice the statement that Dr Stephen Forbes¹, Director of the Illinois State Laboratory of Natural History, is, to my knowledge, the only biologist, aside of course from Shelford, who has ever pointed out the need of measuring such interrelationships. His reasoning follows: Since organism A occurred $x + z$ times in the n collections, its chance of occurrence in any particular collection is $(x + z) \div n$. Similarly, $(x + w) \div n$ is the chance of occurrence of organism B in any particular collection. Whence, from the principle of compound probabilities of independent events, the chance of both A and B occurring conjointly in any one collection is

$$\frac{(x + z)(x + w)}{n^2}$$

or, in n collections, the most probable number of accidental conjoint occurrences is $\frac{(x + z)(x + w)}{n}$. But A and B did actually occur conjointly in x collections. Whence, if

$$x \div \left[\frac{(x + z)(x + w)}{n} \right] = \frac{nx}{(x + z)(x + w)} = (\text{Forbes's coefficient}) \quad (2)$$

¹ *Bull. Ill. State Lab. Nat. Hist.* 7, p. 273; also copied by Shelford in this JOURNAL, 3, p. 12.

is significantly greater than one, we conclude there must be a positive association between *A* and *B* unaccountable for by chance; and similarly, if this coefficient is significantly less than one, we conclude there must be a negative association between *A* and *B* unaccountable for by chance.

Although this reasoning seems sound, it breaks down when closely examined. Forbes deserves credit for having clearly recognized the quantitative nature of this problem, but he did not fully appreciate its complexity. It is easy to see that for a positive association to be perfect, such as is the case in commensalism, no collections should contain organism *A* without also containing organism *B*, and none should contain *B* without also containing *A*. In other words, both *z* and *w* must be zero, and *x* + *y* must equal *n*. Furthermore, it can be demonstrated that the only case of perfect positive association is had when $x = y = \frac{n}{2}$. Without introducing this proof, attention is called to the obvious fact that, if say fifty random collections were made for a particular species of anemone and crab, and no collection obtained either, no information concerning the relationship between anemone and crab would be had. Similarly, if both anemones and crabs were obtained in each of the fifty collections, the coefficient would tell nothing because it would remain wholly uncertain whether such complete coincidence were due say to commensalism between anemone and crab, or to the fact that the unit of territory explored was so large as to necessitate obtaining at least one anemone and one crab in each collection. Forbes's coefficient takes cognizance of neither of these facts. Thus, when $x = y$ and $z = w = 0$,

$$\frac{nx}{(x+z)(x+w)} = \frac{2x^2}{x^2} = 2 \quad (3),$$

which, as just implied, should be the maximum value—that due to perfect positive association. But, for example, when $x = z = w = 1$ and $y = 997$, clearly an extremely weak association, Forbes's coefficient reduces to 250. Moreover, failure of the single conjoint occurrence *x* makes it zero. These three incompatible values are better visualized diagrammatically:

$$\text{Case 1.} \quad A \begin{array}{c|c} \overbrace{500}^B & 0 \\ \hline 0 & 500 \end{array}, \frac{nx}{(x+z)(x+w)} = 2 \quad (4).$$

$$\text{Case 2.} \quad A \begin{array}{c|c} \overbrace{1}^B & 1 \\ \hline 1 & 997 \end{array}, \frac{nx}{(x+z)(x+w)} = 250 \quad (5).$$

$$\text{Case 3.} \quad A \begin{array}{c|c} \overbrace{0}^B & 1 \\ \hline 1 & 997 \end{array}, \frac{nx}{(x+z)(x+w)} = 0 \quad (6).$$

The first chapter of Yule's *Elements of Statistics* is devoted to the problem of association, and, although his discussion is general and devoid of reference to the particular problem of organismal interrelationship, its applicability to this problem is obvious. After stating that any number of coefficients might be devised, Yule gives

$$\frac{xy - wz}{xy + wz} = (\text{Yule's coefficient}) \dots \dots \dots (7)$$

as the simplest. But, as pointed out by a German—reference to whom I have mislaid—it too breaks down when closely examined. For it gives perfect positive association whenever either z or w is zero, and perfect negative association whenever either x or y is zero. For illustration, Yule's coefficient fails to distinguish between

$$A \overbrace{\left\{ \begin{array}{c|c} 1 & 0 \\ \hline 998 & 1 \end{array} \right\}}^B \text{ and } A \overbrace{\left\{ \begin{array}{c|c} 500 & 0 \\ \hline 0 & 500 \end{array} \right\}}^B \dots \dots \dots (8),$$

giving + 1 in both instances. Likewise, it fails to distinguish between

$$A \overbrace{\left\{ \begin{array}{c|c} 998 & 1 \\ \hline 1 & 0 \end{array} \right\}}^B \text{ and } A \overbrace{\left\{ \begin{array}{c|c} 0 & 500 \\ \hline 500 & 0 \end{array} \right\}}^B \dots \dots \dots (9),$$

giving - 1 in both instances.

At this point the problem was taken up at the Scripps Institution, and Dr G. F. McEwen, the institution's hydrographer, devised the coefficient

$$\frac{xy - zw}{\sqrt{(x + w)(x + z)(y + w)(y + z)}} = (\text{McEwen's coefficient}) \dots (10),$$

which gives + 1 only when both z and w are zero, and - 1 only when both x and y are zero. But, although an improvement over Yule's coefficient, it too breaks down; giving + 1 and - 1 too often, more often in fact than it gives zero—the symbol of an absence of association, which, of course, should be obtainable in the greatest number of ways.

How overcome this defect? As stated above, the required index of perfect association is that $x = y = \frac{n}{2}$ or that $z = w = \frac{n}{2}$. Recognizing the similarity of this requirement to one of the basic facts of algebra, namely, that when $a + b$ is constant, the product ab is greatest when $a = b$, we devised

$$\frac{xy - zw}{\left(\frac{x + y}{2}\right)^2 + \left(\frac{z + w}{2}\right)^2} = (\text{McEwen and Michael's coefficient}) \dots (11).$$

Inspection demonstrates that this coefficient gives + 1 only when $x = y$ and both z and w are zero; and - 1 only when $z = w$ and both x and y are zero. Apparently, therefore, it satisfies all requirements.

3. THE PROBLEM OF RELIABILITY.

Immediately another difficulty pops up. What use is a coefficient without a measure of its reliability? Stated more concretely: given a coefficient of + 0.2, does it mean anything unless one can determine the number of ways a coefficient as large or larger than + 0.2 could have arisen by chance? It would seem not.

To state the problem is simple. Assuming all possible coincidences in the distribution of organism *A* and organism *B* as equally likely to occur by chance, how many ways may the numbers between 1 and *n*, taken one, two, three and four at a time, be arranged so as to add to *n*? Answer to this question is the denominator of the required probability, and is given by the equation

$$D = 4 + 6(n - 1) + 2[(n - 1)(n - 2)] + \frac{(n - 1)(n - 2)(n - 3)}{6} \quad (12),$$

where the successive terms denote the sum of arrangements of the numbers taken one, two, three and four at a time. As a matter of fact this equation is strictly accurate only when *n* is some multiple of four, but, for all practical purposes, solution for intermediate values may be had by interpolation. Thus, when *n* is ten, *D* = 286: actually, there are 308 arrangements; but, assuming *n* first as eight and then as twelve, and taking the mean of the two results, gives 310—an error of but two.

The problem of evaluating the numerator is more difficult, and, as yet, remains unsolved. It is readily formulated: on the same assumption as that involved in solution for the denominator, how many ways may the numbers between 1 and *n*, taken two, three and four at a time, be arranged so that the coefficient (equation 11) will have a value numerically equal to or exceeding any prescribed value? Once this question receives a general algebraic answer, the ratio between numerator and denominator gives the required measure of reliability—the measure without which any coefficient of association must remain, for the most part, meaningless.

In lieu of direct solution of this problem, there are four tedious and unsatisfactory methods of obtaining, in special cases, an uncertain substitute. But the conditions permitting employment of such methods are so seldom fulfilled that, except for the last mentioned, which has an intrinsic value of its own, they can rarely be of more than theoretical interest. The four methods are: (1) utilization of the binomial expansion $(\frac{1}{2} + \frac{1}{2})^n$; (2) averaging a number, say thirty, of coefficients obtained from similar but independent series of collections, and evaluating the dispersion about the average by means of the normal probability integral; (3) employment of Pearson's general frequency method based upon *Bayes Theorem* and the hyper-geometric series; and (4) determining the nature and degree of correlation of each of the two organisms with the inorganic elements of their environmental complexes, the method

thus far made most use of at the Scripps Institution, and brought into indirect relation to the present problem with especial clearness in *Differentials in behaviour of the two generations of Salpa democratica relative to the temperature of the sea*¹.

4. A DILEMMA, AND THE WAY OUT.

How is the problem here raised to be solved? More than this: how are the mathematical problems permeating the whole field of quantitative biology to be solved? "For evident reasons," say Kapteyn and Uven², "a [mathematical] theory for the benefit of biologists would be best worked out by a biologist. . . . About the worst possible thing will be to put the task wholly on a mathematician." The biologist thus finds himself in a dilemma. Like myself, he faces important biological problems demanding mathematical tools for their solution. But, owing to inadequate training, he finds that, like myself, he is incapable of making these tools. Even in the case of close cooperation between biologist and mathematician, the latter, as a rule, knows too little of biology to visualize the problem, unless it be translated into mathematical terminology, and the biologist, as a rule, finds himself unable to make this translation. The problem cries for solution. What is to be done?

The only way out of this dilemma that I can conceive is to insist upon proficiency in mathematics, just as we now insist upon proficiency in English, as prerequisite to a major in biology. *This means reform.* Is it necessary? It seems so: for numbers are the words without which exact description of any natural phenomenon is impossible. They, and their graphical equivalents, are the only means of describing the *how much*, just as words, and their pictorial equivalents, are the only means of describing the *what kind*. Whence, to hold proficiency in English essential, and proficiency in mathematics non-essential to healthy biology, is equivalent to claiming that the biological sciences, unlike all others, are unconcerned with the *how much* of their phenomena. Assuredly, every objective phenomenon, of whatever kind, is quantitative as well as qualitative; and to ignore the former, or to brush it aside as inconsequential, virtually replaces objective nature by abstract toys wholly devoid of dimensions—toys that neither exist nor can be conceived to exist. The truth of this is unescapable in spite of the emphatic assertion, with which I am in full accord, that biology, confronted as it is with so huge a variety of things, must ever continue the qualitative science *par excellence*.

¹ *Univ. Calif. Pub. Zool.* **18**, p. 239.

Skew frequency curves in biology and statistics. Groningen, 1916, p. 5.

[We are glad to publish Mr Michael's plea. But we may question if it is practicable to insist on "proficiency in mathematics" for *all* biologists at this stage of the development of science. Many biologists must still for a long time be concerned with studies in which mathematics have little part, though this will become less and less the case. The immediate need seems provision of adequate training in certain branches of mathematics, e.g. the calculus and the theory of statistics, for those biologists who need them. The same is true of physical chemistry in relation to biology. EDITOR, JOURN. ECOL.]