stichting
mathematisch
centrum

MC

REKENAFDELING NR 11 MAART

M.H. VAN EMDEN
INTERACTION ANALYSIS: AN APPLICATION OF INFORMATION THEORY IN PHYTOSOCIOLOGY

2e boerhaavestraat 49 amsterdam

INTERACTION ANALYSIS,

AN APPLICATION OF INFORMATION THEORY IN PHYTOSOCIOLOGY

paper presented by M.H. van Emden

at the 14th International Symposium on

Basic Problems and Methods in Phytosociology

organized by Prof. Dr. Dr. h.c. R. Tüxen

at Rinteln, Germany, March 23-26, 1970.

ABSTRACT: This paper introduces Interaction Analysis, a method similar to the Association Analysis introduced by Williams and Lambert in 1959. Its purpose is the analysis of a complex pattern of interactions, a basic problem in phytosociology and also in many other subjects. The method is founded on an intuitively appealing definition of complexity as the sum of amounts of interaction. A definition of such an amount is given which is mathematically equivalent to that for the capacity of a noisy channel in information theory.

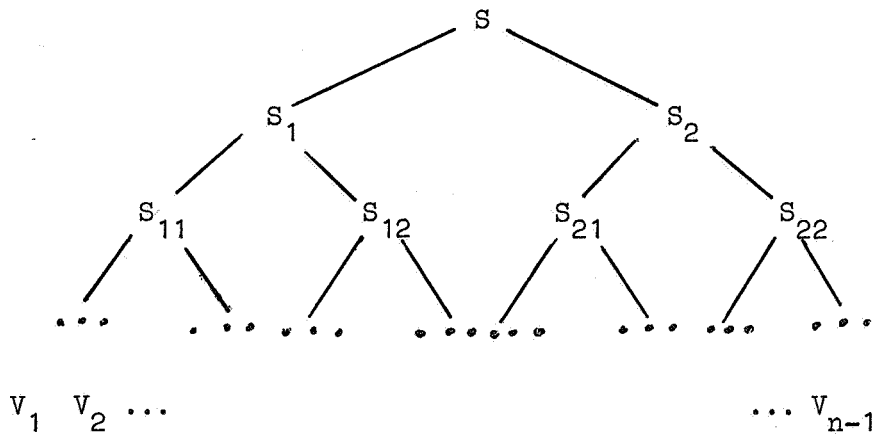
Interactions as additive components of complexity

It is necessary to make precise the concept of *complexity*, and the investigations to be described in this paper are motivated by the desire to see whether entropy is as useful for the measurement of complexity as it has proved to be (in the mathematical theory of communication, see Shannon [2]) for the measurement of the possible information content of a signal. A definition of complexity at once simple and yet, as will be shown later, precise enough to serve as the basis of a mathematically defined measure, is:

The way in which a whole is different from the set of its parts.

We should also be able to say by how much the whole is different from the set of its parts. If it is not at all different, the complexity of the whole is just the sum of complexities of the parts. If it is, we say there is *interaction* between the parts and the complexity of the whole is more than this sum, the difference being the amount of interaction. Therefore, if we can find a suitable measure of these interactions, they should be regarded as additive components of complexity. Such components should somehow take into account the overall intensity of interactions, not just those between pairs of parts.

Let us call the whole a *system*, its parts *subsystems*, or, if they cannot be subdivided any more, *variables*. Suppose that the system S is divided into subsystems S_1 and S_2, which are respectively subdivided into S_{11}, S_{12} and S_{21}, S_{22}, and so on until only variables are left. This may be represented by a hierarchical scheme:



... (1)

According to our definition the complexity of S should equal the total amount of interaction present in S, that is,

$$C(S_1,S_2)+C(S_{11},S_{12})+C(S_{21},S_{22})+\dots ,$$

where C(X,Y) stands for the amount of interaction between X and Y. It may be defined as follows.

Suppose that for every possible subsystem of S there is defined a *potential function* H, and that this function has the properties:

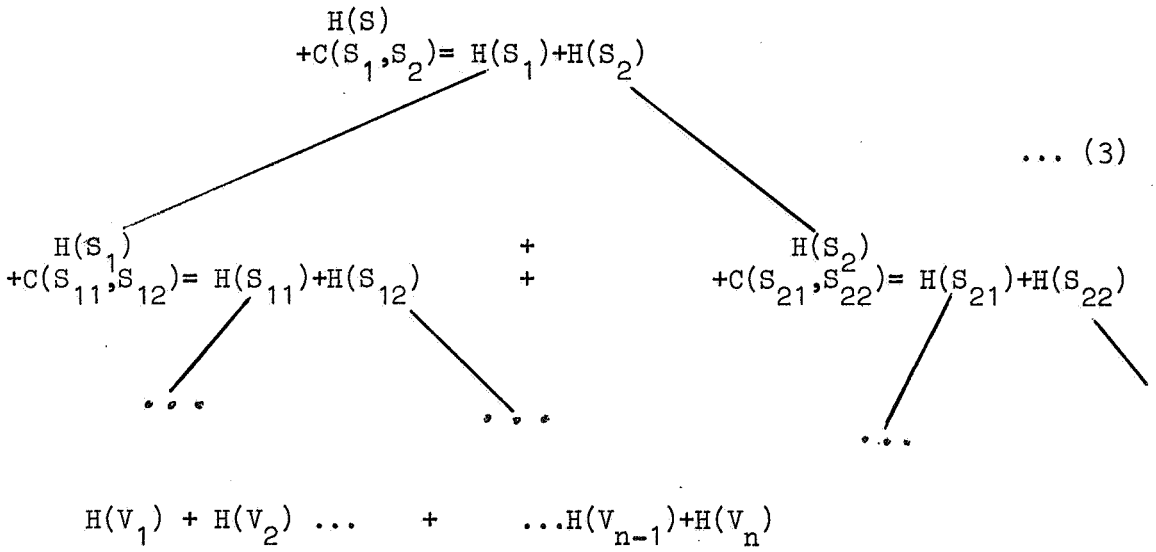
monotonicity: $H(S) \geq H(S_1)$ and $H(S) \geq H(S_2)$
sub-addivity: $H(S) \leq H(S_1)+H(S_2)$. ... (2)

Here S_1 and S_2 are subsystems obtained by a single division of S. The sub-additivity property allows a measure of interaction to be defined as follows:

$$C(S_1,S_2)= H(S_1)+H(S_2)-H(S) \geq 0.$$

This explains why we have chosen the name "potential function" for H.

Interaction so defined is a difference of potentials, which is a straightforward generalization of the corresponding situation in physics where a force of attraction is the gradient of the corresponding potential. The hierarchical scheme (1) may now be accompanied by the corresponding potentials and interactions:



Let us consider the difference between the highest and the lowest levels of potential in the hierarchical diagram:

$$C = H(V_1)+\ldots+H(v_n)-H(S), \qquad \ldots (4)$$

where $V_1,\ldots,V_n$ are the individual variables of which the system consists. The definition of complexity we proposed as: "The way in which a whole is different from the set of its parts". Suppose we want to compare the complexity of different systems: to be able to say that one system is more complex than another. One way to do this is to represent complexity as a real number. The formula (4) is a direct representation of the definition in numbers; it says in effect: "The measure C of complexity is the difference between the potential of the whole and the sum of potentials of its parts".

So much for the definition of complexity. Let us now consider an application. Especially interesting is the case where the set of individual variables $V_1,\ldots,V_n$ is large and where many interact with many others. In such a case we would like to know whether a simpler model would be able to give a satisfactory summary of the given system. Let us consider the sums L_i of the potentials of the subsystems of the same level in the scheme (3):

$L_1 = H(S)$, $L_2 = H(S_1)+H(S_2)$, $L_3 = H(S_{11})+H(S_{12})+H(S_{21})+H(S_{22})$, $\ldots$, $L_k = H(V_1)+\ldots+H(V_n)$, and $L_1 \leq L_2 \leq L_3 \leq \ldots \leq L_k$.

Suppose that L_i, the general term in this sequence, is already almost as large as L. Then we can say, that the subdivisions above level i account for almost all the interactions present in the system, and these subdivisions represent a useful summary of all subdivisions.

In a system S there are many ways to subdivide successively to arrive at the variables $V_1,\ldots,V_n$. Yet, for all these L_k-L_1 is the same and equal to the complexity of S. We want to do few subdivisions and yet capture as large an amount of interaction as possible. Some ways of dividing achieve this better than others. A criterion for judging how well this is achieved is the following quantity:

$$U = L_2-L_1+2(L_3-L_2)+3(L_4-L_3)+\ldots+(k-1)(L_k-L_{k-1})$$
$$=(L_k-L_1)+(L_k-L_2)+\ldots+(L_k-L_{k-1}). \qquad \ldots (5)$$

Apparently, the weights $2,3,\ldots,k-1$ punish large interactions for being low in the scheme. Therefore, the best division scheme is the one for which U is minimum.

Interaction Analysis

In this section we shall show how to apply the results of the previous section to a situation often encountered in phytosociology.

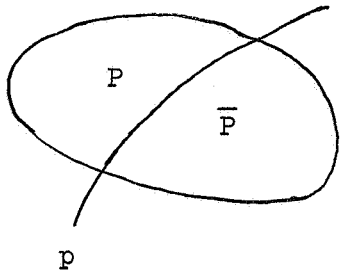
There it often happens that data are collected in such a way that numerous small areas of the surface of the earth, called *quadrats*, are scrutinized for the occurrence of *species* of plants. Similar data are also studied in other subjects, so, instead of "quadrats" and "species", we shall use the more general terms "objects" and "predicates". The data are arranged in an object-predicate table, which is a rectangular array of noughts and crosses. The j^{th} cell of the i^{th} row shows whether the i^{th} object does (when it contains a cross) or does not (when it contains a nought) possess the j^{th} predicate.

	objects ⟶					
predicates	1	2	3	4	5	6
↓ 1	0	0	x	0	0	x
2	x	0	x	x	x	0
3	x	x	0	0	0	x
4	0	x	0	x	0	0
5	0	0	0	0	x	0

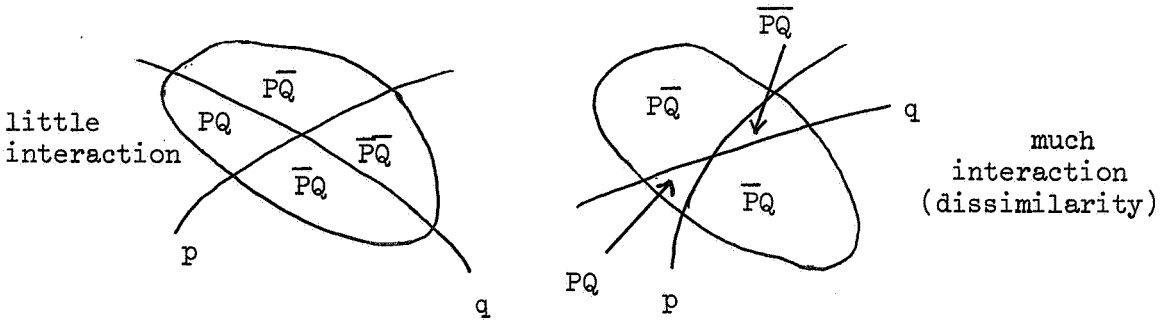
From the mathematical point of view, which we adopt here, there is only a relative difference between objects and predicates: objects may be regarded as predicates to the predicates, which then feature as objects. In other words, if we stand the object-predicate table on its side, it is an object-predicate table again. In Association Analysis, if "normal analysis" corresponds to the one position of the table, "inverse analysis" corresponds to the other.

We can now give an interpretation of the abstract notion of a "system" introduced in the previous section: the system is the set of predicates and the predicates represent the variables of the system. In that way the Analysis will be concerned with interactions among predicates. The only other thing that is still unspecified is the potential function H, and we proceed to show that the entropy of a set of partitions can play this rôle.

Predicate p divides the objects into two subsets, P and $\overline{P}$, those that have p and those that do not have p, respectively.



Likewise, predicate q effects a partition. When this partition is "similar" or "dissimilar" to the previous one, we say that there is interaction between the two.



The amount of interaction is conveniently measured by a difference of entropies. The entropy of a partition is defined as follows. Let the relative frequencies of the k cells of the partition r be $f_1,\ldots,f_k$, then the entropy of the partition equals

$$H(r) = - f_1 \log(f_1) - \ldots - f_k \log(f_k).$$

Each of the predicates p and q separately define a partition, and so also does the simultaneous application of p and q. The interaction between p and q is then defined to be

$$H(p) + H(q) - H(p,q) \qquad \ldots (6)$$

In information theory p represents the message going into a noisy channel and q the message coming out of it. Shannon defined (6) as the information transmitted by the channel. McGill [1] noticed the applicability of the information-theoretical concept in a situation more similar to the one discussed here. The quantity (6) is 0 in the case of no interaction and positive otherwise.

Any number of predicates may be applied simultaneously to the set of objects, and in a similar way an interaction may be defined for them:

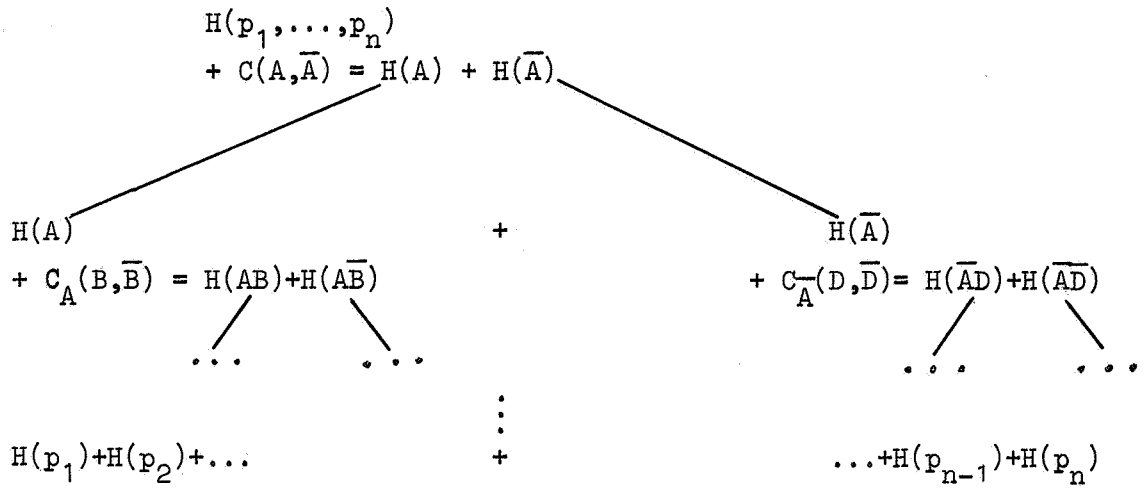
$$H(p_1)+\ldots+H(p_n)-H(p_1,\ldots,p_n) \qquad \ldots (7)$$

This quantity and its decomposition has been studied by Watanabe [4]. If $p_1,\ldots,p_n$ are all the predicates of the object-predicate table, we may interpret this to be total amount of interaction present in the table. The entropy possesses the properties (2) and therefore qualifies as a potential function. This is explained in detail in an earlier paper [3]. The quantity (7) is the total amount of complexity involved in the interactions of the system of variables, which are, in this case, predicates.

As shown before, the amount of complexity may be hierarchically decomposed. In the case of the object-predicate table, this is done as follows. We saw that any set of predicates defines a partition in the set of objects. That is, to each object, say b, two sets of predicates are assigned: the set B of predicates that object b has and the set $\overline{B}$ that b does not have. Again, the difference in entropies is not negative:

$$H(B) + H(\overline{B}) - H(p_1,\ldots,p_n) \geq 0 \ .$$

This quantity may be called the entropy loading of object b. Thus, just like in Association Analysis, the set of predicates is successively split up:



Why should this decomposition, out of all possible ones, be especially meaningful? Such is the case, if a few splits already push up the level of the potential function near its maximum. Then a few objects already characterise almost all interactions between predicates. By how much this is the case, is indicated by the quantity U (5) of the corresponding scheme. Ideally, we want to choose the splitting objects so, that U is minimum. At this time, we cannot make a better guess than to follow Williams and Lambert in choosing the object with largest entropy loading available at each step.

This ends the present description of Interaction Analysis. Much of it is implicit in Association Analysis, as described by Williams and Lambert [5]. They used "association" instead of interaction, which is, perhaps, unfortunate, because most people think of it as something like positive correlation, while the chief merit of their work lies in the fact that they include the positive as well as the negative, in short, what we call interaction. They did not derive interaction from a potential function; indeed, they did not give a numerical definition of interaction, but used instead a numerical criterion for deciding which subdivision to effect. This criterion involves computing tail probabilities for testing independence in large numbers of 2x2 contingency tables.

Unfortunately, these are about the only contingency tables where the asymptotically approximating chi-squared distribution gives poor results; so either prohibitively laborious exact probabilities are called for, or corrections must be applied. Without such corrections, a mathematical analysis of the properties of this criterion seems rather formidable; with them, it seems hopeless.

Our method seems to represent a significant advance compared to Association Analysis in two respects:
1) Interaction Analysis has a clear conceptual basis. Its purpose is of central interest in phytosociology: the analysis of a complex pattern of interactions. The most important concept, that of complexity, is given a numerical representation that corresponds in a simple way to an intuitively appealing definition of complexity.
2) Entropies of partitions are simple to compute. There is hardly any scope for messy approximations that would make it difficult to understand what is going on.

Acknowledgement

The research reported here was carried out at the Mathematical Centre and entirely supported by the Hugo de Vries Laboratory for Systematic Botany in the University of Amsterdam. The Netherlands Organization for the Advancement of Pure Research (Z.W.O.) has provided support for the author's attendance at the Rinteln Symposium.

References

[1] McGill, W.J. Multivariate information transmission. Psychometrika, 19(1954), 97-116.

[2] Shannon, C.E. The Mathematical Theory of Communication. University of Illinois Press, Urbana (Ill), 1949.

[3] Van Emden, M.H. The hierarchical decomposition of complexity. in: Machine Intelligence 5(1970), B. Meltzer and D. Michie (eds), 361-380.

[4] Watanabe, S. Information theoretical analysis of multivariate correlation. IBM J.Res.Dev., 4(1960), 66-82.

[5] Williams, W.T. & J.M. Lambert. Multivariate methods in plant ecology I.J. Ecology, 47(1959), 83-101.