Art. N° 113
Contribution COB N° 104

Union Océanogr. France - Bull. N°9 - 1.972 p.40-45

LE CHOIX DE LA METRIQUE ET LE PROBLEME DES ABSENCES SIMULTANEES DANS LE TRAITEMENT DES DONNEES BENTHIQUES *

par

Pierre CHARDY

Centre Océanologique de Bretagne - BREST - 29N

—La plupart des systèmes de classification et d'ordination (GREIG-SMITH, 1964 ; WHITTAKER, 1967) sont fondés sur la comparaison de variables prises deux à deux et sur le regroupement de ces variables en catégories. En écologie, les variables peuvent être des espèces ou des prélèvements selon qu'il s'agit d'analyse en mode -Q ou en mode -R.—

Le traitement des données écologiques suppose donc un double choix :

- Choix de la métrique qui permet de mesurer le degré de liaison entre toutes les paires de variables et conduit à l'élaboration d'une matrice de similitude (par métrique nous entendons coefficient de corrélation, distance, indice de similarité, de liaison, etc...).

- Choix de l'algorithme (ou analyse mathématique) qui conduit à l'obtention de structures, hiérarchiques ou multidimensionnelles, permettant la reconnaissance des groupes de variables.

La multiplicité des indices de liaison proposés dans la littérature rend le premier choix particulièrement difficile. Cependant, il ne doit pas être négligé car chaque métrique met en valeur une certaine qualité d'information et détermine en grande partie le type de structure extraite par l'analyse. Il importe donc de choisir une métrique servant au mieux

Contribution n°104 du Département scientifique du Centre Océanologique de Bretagne.

les données. Cette préoccupation dépasse très largement le cadre de la benthologie, les quelques lignes qui suivent n'ont d'autre ambition que de fournir une base de réflexion à ce problème.

Il est hors de question de reprendre ici la liste de toutes les métriques utilisées en écologie ou en taxinomie (voir DAGNELIE, 1960 ; DA FONSECA, 1966 ; SOKAL et SNEATH, 1963, etc...). On distingue classiquement les indices de liaisons calculés sur les données qualitatives (présence - absence) et ceux qui sont calculés sur les données quantitatives (abondance, fréquences, dominance). Ajoutons les indices "semi-quantitatifs" constitués par les coefficients de corrélations de rang. A ce niveau, le choix est imposé par la nature des données (qui peuvent être composées de variables nominales, ordinales, repérables, mesurables...). Le but de l'étude peut également être déterminant (aspect strictement faunistique, aspect économique, etc...).

Qu'ils soient qualitatifs ou quantitatifs les indices de similarités appartiennent en fait à 2 grandes familles :

- Les indices issus des statistiques classiques (coefficient de corrélation de Bravais-Pearson, χ^2, coefficients de corrélation de rang de Spearman, de Kendall).

- Les indices que l'on peut qualifier d'empiriques, et qui portent généralement le nom de leur auteur (JACCARD, KULCZYNSKI, ODUM, AGRELL, DICE, SOKAL et MICHENER, etc...).

Les premiers justifient le recours aux tests statistiques, ce qui leur confèrent une valeur absolue. Les seconds ne peuvent être testés et n'ont qu'une valeur relative liée au nombre d'observations. Ce premier argument semble à l'avantage des métriques purement statistiques d'autant qu'elles offrent la possibilité de traiter toutes les catégories de données. En particulier les coefficients de rang peuvent être appliqués aux données non gaussiennes (au prix, évidemment, d'une perte d'information).

Ce choix purement statistique rejette tous les coefficients empiriques qui, eux pourtant, sont fondés sur des considérations écologiques. En effet ces derniers ont généralement été établis dans le but de traduire au mieux certains types de liaisons observées ou pressenties dans le milieu étudié.

Quelques considérations probabilistes sur les ensembles finis, permettent de mieux saisir la nature des divergences existant entre ces deux catégories de métrique.

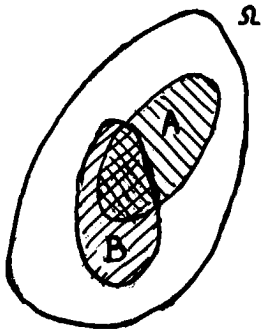

Fig. 1

La formulation des métriques statistiques repose sur la définition et les propriétés des probabilités des ensembles finis (les variables sont des variables aléatoires sur un ensemble fini). Soit Ω un tel ensemble dont la probabilité P (Ω) est égale à 1 (fig. 1). Pour tout couple d'évènements (A, B) appartenant à Ω, on peut définir les probabilités P (A) ; P (B) ; P (A ∩ B) et P ($C^{\Omega}_{(A \cup B)}$)

Prenons le cas ou (A) et (B) sont des ensembles de prélèvements contenant respectivement les espèces "a" et "b" ; (A∩B) sera l'ensemble des prélèvements contenant à la fois "a" et "b" et $C^{\Omega}_{(A \cup B)}$ l'ensemble des prélèvements ne contenant ni "a" ni "b". Qu'elles soient qualitatives ou quantitatives les métriques diffèrent par l'importance qu'elles accordent au terme $C^{\Omega}_{(A \cup B)}$, c'est-à-dire à l'absence simultanée des variables. On peut donc distinguer deux grandes catégories de métriques en fonction de ce critère :

- Les métriques qui se refèrent à un ensemble fini ou dénombrable et intègrent les doubles absences dans leur formulation.

- Les métriques qui rejettent l'hypothèse de l'ensemble fini, et ignorent les doubles absences.

D'un point de vue pratique, comparons deux espèces "a" et "b" mal représentées dans un ensemble important de prélèvements : si l'on utilise une métrique qui tient compte des doubles absences la corrélation $r_{(a, b)}$ sera positive. Inversement si l'indice de liaison ne tient pas compte des doubles absences la corrélation $r_{(a, b)}$ peut être négative. A titre d'exemple, les mêmes données ont été traitées simultanément par les deux types de métriques, et une analyse factorielle a été appliquée sur chacune des matrices de similitudes obtenues. Il s'agit de 30 prélèvements benthiques effectués sur le plateau continental du Golfe de Gascogne

données brutes extraites des travaux de GLEMAREC, 1969). Les prélèvements sont regroupés de la façon suivante : infralittoral (A_1 et A_2), circalittoral côtier (B), circalittoral du large (C). Dans le premier cas la matrice de similitude est obtenue par le coefficient de BRAVAIS-PEARSON et analysée par la méthode des composantes principales. Dans le second cas le même coefficient est utilisé, mais en éliminant systématiquement les doubles zéros à chaque calcul de corrélation entre couple de variables. L'analyse factorielle est effectuée selon la méthode de GOWER (1966) pour des raisons purement formelles (voir IBANEZ, 1969).

Pour apprécier l'importance des divergences enregistrées, examinons la visualisation des 30 prélèvements dans l'espace défini par les 2 premiers axes de chaque analyse (Fig. 2 et 3).

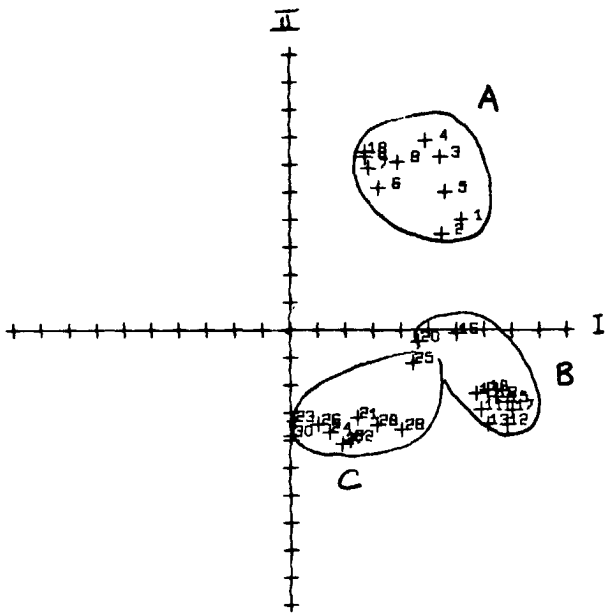

Fig. 2 : Analyse factorielle tenant compte des doubles absences

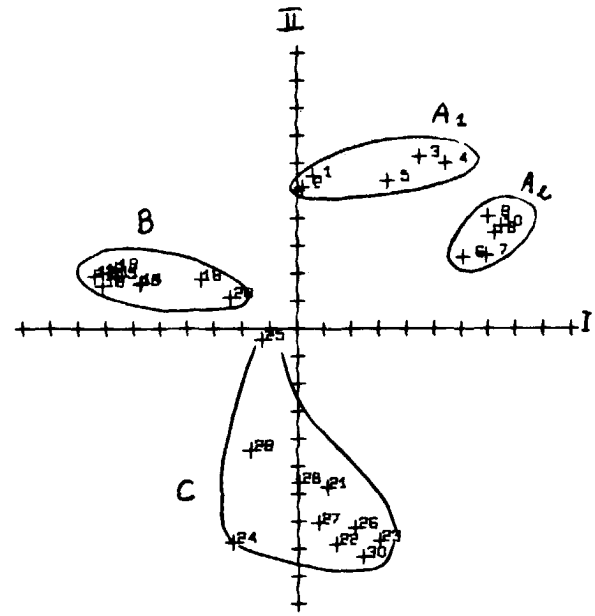

Fig. 3 : Analyse factorielle excluant les doubles absences

Dans le premier cas (avec les doubles absences) le premier axe n'est pas discriminant et rassemble tous les prélèvements dans les saturations positives (fig. 2). Bien que cette première composante extrait une fraction importante de la variance totale (25 %) elle n'a aucune signification écologique : il s'agit d'un artefact créé par les nombreuses absences simultanées des espèces. Dans le deuxième cas (sans les doubles

absences) le premier facteur est bipolaire et la discrimination entre les étages beaucoup plus nette (fig. 3). On remarque également que l'homogénéité faunistique du circalittoral côtier est beaucoup plus forte que celle du circalittoral du large. En conséquence, l'information écologique apportée par les deux premiers axes est beaucoup plus importante dans le second cas que dans le premier.

Le choix de la métrique implique donc une réflexion à la fois statistique et écologique. En benthologie, les cas assimilables à des ensembles finis (ou fermés) sont en définitive assez rares. C'est avant tout un problème d'homogénéité faunistique, c'est-à-dire d'échelle d'observation et d'échantillonnage (dont les difficultés en benthos sont certaines).

Toutes les fois que cela est possible, il est évidemment souhaitable d'utiliser les métriques issues des statistiques, notamment le coefficient de corrélation de Bravais-Pearson ou la covariance, qui sont les pivots de tous modèles linéaires (régressions multilinéaires, corrélations canoniques, analyse en composantes principales, etc...). Mais ce n'est pas toujours des coefficients statistiques qu'il faut attendre une solution à nos problèmes, surtout lorsque la matrice de données à analyser comporte de nombreux zéros.

INDEX BIBLIOGRAPHIQUE.

CANCELA DA FONSECA, J.P., 1966. L'outil statistique en biologie du sol - III - Indices d'intérêt écologique. Rev. Ecol. Biol. Sol., 3 : 381-407.

DAGNELIE, P., 1960. Contribution à l'étude des communautés végétales par l'analyse factorielle. Bull. Serv. Carte Phytogeogr., série B, 5 (1) : 7-71.

GOWER, J.C., 1966. The distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325-338.

GLEMAREC, M., 1969. Les peuplements benthiques du plateau continental Nord - Gascogne. Thèse Etat Univ. Paris.

GREIG-SMITH, P., 1964. Quantitative plant ecology. Butter worths, London. 256 pp.

IBANEZ, J.J., 1969. Application de l'analyse factorielle en planctonologie écologie et taxinomie numérique. Thèse 3ème Cycle, Fac. Sci. Paris, 130 pp.

SOKAL, R. et P.H.A. SNEATH, 1963. Principles of numerical taxonomy. San Francisco. London.

WHITTAKER, R.H., 1967. Gradient analysis of vegetation. Biol. Rev., vol. 49, pp. 207-264.

* Manuscrit reçu le 17-6-1972