Art. N° 298
Contribution COB N° 449

DE L'HEURISTIQUE AU THAUMATURGIQUE EN TRAITEMENT DES DONNÉES D'ÉCOLOGIE MARINE (1)

François Blanc* et Alain Laurec**

* *Maître assistant au Centre Universitaire de Marseille-Luminy, laboratoire d'Hydrobiologie marine, Marseille-Cedex*
** *Chercheur C.N.E.X.O. au Centre Océanographique de Bretagne à Brest.*

Résumé

Dans cet article les auteurs définissent leur philosophie quant au traitement mathématique des données en écologie marine. Le texte est articulé en deux grandes parties qui déballent successivement de l'apport potentiel des méthodes mathématiques en écologie marine et de la stratégie d'utilisation de ces techniques.

Dans la première partie, ils insistent sur la nécessité de ne pas considérer ces méthodes comme miraculeuses (thaumaturgiques), créatrices d'une information absente des données récoltées, mais comme des outils forts utiles (heuristiques) permettant, par une utilisation correcte, une description et une structuration de l'information recueillie. Elles permettent ainsi à l'écologiste de formuler d'une manière explicite et rigoureuse d'une part ses choix fondamentaux comme ses hypothèses, d'autre part ses conclusions. Dans ce chapitre, ils insistent sur deux points essentiels :

— les problèmes de reconnaissance de structures sur un ensemble de points qui peuvent être par exemple des prélèvements, n'ont de sens que pour une topologie donnée ;

— la possibilité de progresser dans la compréhension d'un phénomène écologique passe par le choix d'un modèle (être mathématique que l'on substitue à la réalité parce que plus maniable), qui peut être simplement descriptif ou explicatif selon les ambitions de départ.

Dans la seconde partie, les auteurs mettent l'accent sur l'importance d'une bonne stratégie d'utilisation des méthodes. Il convient donc de connaître leurs limites d'emploi et notamment de bien sérier les hypothèses requises. Ils insistent à propos des problèmes d'inférence sur la nécessité d'une définition claire du doublet échantillon-population. Enfin l'importance d'une planification de l'échantillonnage est évoquée.

Abstract

In an answer to FRONTIER (1975) the authors try to define their philosophy about data processing by mathematical methods in marine ecology. They point out that the ecologist may not expect miracles from these methods, but that, correctly used, they provide an especially heuristic tool if not a thaumaturgic one. They first discuss the exact nature of the benefits that can be brought to ecology and the intrinsic limits of these benefits. On a second step they suggest the main features of what should be any correct strategy in data processing, and try to give FRONTIER an answer on certain precise topics.

Un récent et intéressant article de Frontier (1975) sur l'heuristicité de l'analyse factorielle en écologie planctonique a permis d'ouvrir un débat que nous nous proposons de prolonger (2).

Qu'il nous soit permis toutefois d'élargir notre propos au plan plus général du traitement des données.

Le traitement mathématique ne peut jamais être

(1) Contribution n° 449 du Département Scientifique du Centre Océanologique de Bretagne.
(2) S'il n'apporte pas autre chose, ce débat enrichira peut-être le vocabulaire en Océanographie.

Cah. O.R.S.T.O.M., sér. Océanogr., vol. XIV, n° 2, 1976 : 101-107.

qu'un outil au service de l'écologiste. Aussi celui-ci doit-il d'abord savoir :

(a) quels services il peut, a priori, en attendre ;

(b) comment en retirer le bénéfice maximal, c'est-à-dire :

— quelles sont les limites d'utilisation de l'outil ?

— comment choisir l'outil le mieux adapté à son problème ?

Nous allons discuter successivement ces divers points.

1. L'APPORT DU TRAITEMENT DES DONNÉES EN ÉCOLOGIE MARINE

Il existe deux attitudes contradictoires et également dangereuses de l'écologiste devant les méthodes mathématiques :

— les considérer comme un instrument de la thaumaturgie, manière de « pierre philosophale » transmutant le vil plomb de données partielles et / ou mal récoltées en le métal précieux du résultat écologique définitif et inattaquable.

— La seconde attitude aussi néfaste, mais souvent issue de la première à la suite d'une déception, conduit à méconnaître l'heuristicité des méthodes.

1.1. Les limites intrinsèques du traitement des données

1.1.1. La valeur du résultat d'un traitement de données va reposer avant tout sur la quantité d'information contenue dans ces données. La méthode mathématique ne saurait être tenue pour responsable d'un plan d'échantillonnage mal conçu pour raisons matérielles, administratives ou autres.

Ainsi, si l'on doit comparer deux prélèvements faits à des endroits et des instants distincts, il est impossible d'attribuer les différences observées à des effets spéciaux ou temporels. Aucune méthode mathématique ne créera l'information qui n'existe pas à ce propos dans les données.

1.1.2. La méthode mathématique ne se substitue en aucun cas à la réflexion écologique. Discutons ce point sur un exemple, celui des valeurs nulles. Le vrai problème n'est pas de nature mathématique, mais biologique. Passons rapidement sur la possibilité que l'espèce présente à la station n'ait pas été récoltée, à cause de sa rareté, de ses possibilités d'évitement, etc., cette observation pouvant d'ailleurs illustrer le paragraphe précédent. Considérons maintenant, le problème des doubles absences. Au plan mathématique, le problème des valeurs nulles n'est pas sans solution. Mais il est certain que c'est l'écologiste qui devra décider si une double absence a une signification écologique réelle ou non. Par exemple si l'on considère des variables qualitatives, les indices ne considérant que les doubles présences traduisent mathématiquement le choix de la deuxième option.

1.2. Les potentialités de la réflexion conjuguée mathématique et écologique

Trois étapes apparaissent dans le traitement d'un problème écologique :

— la traduction mathématique du problème écologique,

— la mise en œuvre des moyens mathématiques susceptibles de résoudre le problème,

— l'interprétation écologique et l'exposé des résultats.

1.2.1. La traduction mathématique.

On discutera ici de deux points importants : la définition d'une topologie et le choix d'un modèle.

Au moment de formuler ses choix écologiques, le praticien prend souvent conscience de l'imprécision de ceux-ci. Trop souvent en effet, en écologie, les choix sont implicites voire inconscients. Prenons l'exemple cher à la biocoenotique, celui de l'influence relative des espèces rares ou fréquentes. Lorsqu'il convient de définir une distance coenotique entre prélèvements, il n'est plus possible de tergiverser et le choix doit être tranché. On quantifie exactement l'importance de chaque espèce par le poids qu'on lui donne. Lorsque l'on se propose d'étudier la structure d'un ensemble de stations ou de prélèvements, les mathématiques soulignent qu'une telle structure n'existe qu'une fois choisie une distance (ou de façon plus générale une topologie (1)). Toute la biocoenotique repose sur la notion de proximité entre prélèvements et l'on se demande si bien des querelles comme celles opposant les tenants de la biocoenotique quantitative à ceux de la qualitative, ne reposent pas sur une méconnaissance par l'écologiste de ce fait fondamental : une structure n'existe qu'une fois la distance choisie. Les tenants du quantitatif et du qualitatif n'étudient pas tout à fait le même problème.

La première idée que nous venons de développer ici correspond en mathématique à la définition d'une topologie à partir de laquelle on pourra raisonner. Le mathématicien, en affirmant qu'il n'existe de structure qu'au moment où on aura défini une topologie, ramène les divergences d'opinions des écologistes à leur juste valeur. Il n'existe pas une vérité, mais plusieurs facettes de celle-ci. A partir de cela, il n'existe pas une méthode optimale d'accès à la vérité, mais un ensemble de problèmes écologiques intéressants.

(1) Si le terme de distance est trop restrictif, celui de topologie est trop général. Nous l'utiliserons, faute de mieux.

Outre la définition d'une topologie, un autre problème de traduction apparaît, celui du choix d'un modèle (1). Celui-ci peut être défini comme un être mathématique que l'on substitue à la réalité, parce que plus maniable. On fait couramment une telle modélisation, ainsi parlera-t-on de largeur et de longueur d'une table, en opérant ainsi, on utilise un modèle mathématique : le rectangle. Dans le choix d'un modèle, la démarche du mathématicien est modeste ; il n'est pas manichéen dans son choix, il ne demande pas aux modèles d'êtres vrais (par opposition à des modèles faux) mais d'être plus ou moins complets, sachant qu'ils ne seront jamais exhaustifs. Par exemple, le modèle biocoenotique n'est ni vrai ni faux, c'est un modèle, son adéquation est plus ou moins satisfaisante ; c'est en effet une transcription abstraite de la réalité, il peut, bien entendu, en laisser certains aspects dans l'obscurité.

Ainsi, si l'on décrit les peuplements par des systèmes de tâches (biocoenoses), il existe certainement des transitions dans les zones frontières, lorsque les zones de transition occupent un pourcentage de surface très important, le modèle du continuum sera certainement d'un emploi plus profitable.

Après cette discussion écologique du modèle, essayons de passer à un formalisme plus mathématique. Prenons pour cela un exemple familier aux halieutes, le phénomène de la croissance pondérale. La formulation mathématique la plus fréquemment utilisée pour ce phénomène est le modèle de Von Bertalanffy

$$W(t) = W\infty\ [1 - e^{-k(t-to)}]^3$$

Nul n'a prétendu, et son auteur moins que tout autre, que cette équation épuisât toutes les modalités du phénomène de croissance. Elle se révèle dans de nombreuses situations pratiques, d'une utilisation satisfaisante. Mais il peut arriver dans d'autres cas, qu'une autre modélisation se révèle plus performante ; ainsi, dans le cas des croissances larvaires, le modèle de Gompertz est-il souvent préférable.

$$\text{Log } (W(t)) = \text{Log } (W\infty)\ (1 - ae^{-kt})$$

Aucun des deux n'est jamais exhaustif en pratique.

1.2.2. Les moyens mathématiques.

La réflexion conjuguée mathématique et écologique a conduit au choix d'un type de modèle. Dans la catégorie de modèles choisie, il faut encore délimiter celui qui est le plus en concordance avec les données dont on dispose. Pour ce faire, il convient de définir un critère d'ajustement, quantité que l'on cherche à minimiser ou maximiser. Pour en citer un des plus classiques, on utilise couramment la méthode des moindres carrés. En termes écologiques, quels sont les a priori de cette méthode ? Il convient de savoir que quelques points en désaccord marqué avec le modèle modifient le résultat plus sûrement qu'un grand nombre de points en désaccord léger. Cela peut aller ou non dans le sens du désir de l'utilisateur.

Une fois le critère choisi, les méthodes d'ajustement relèvent davantage des mathématiques que de l'application écologique (2). Pour revenir au traitement des données, qui était notre préoccupation fondamentale, il s'agit de trouver un moyen d'ajuster des structures observées, forcément plus ou moins complexes, à un modèle relativement simple et manipulable. Ainsi dans le cas des analyses factorielles, procède-t-on à la représentation d'un ensemble de points dans un espace euclidien de dimension réduite (modèle simplifié avec une perte d'information) avec un critère d'ajustement du type moindres carrés. Ce faisant on bénéficie de l'apport des techniques de reconnaissance des formes mises au point entre autres, par les mécaniciens (techniques d'analyse d'inertie).

On peut dire d'une manière générale, que l'écologiste en formulant mathématiquement son problème, bénéficie de recherches extérieures à son champ d'activité. Cela peut d'ailleurs évoquer des analogies avec d'autres disciplines, telles la physique et la chimie.

1.2.3. L'interprétation des résultats.

L'intérêt des conclusions que l'on peut tirer d'un travail scientifique se place sur deux plans : d'une part, mettre en lumière des faits nouveaux, d'autre part, hiérarchiser des faits antérieurement connus (dégager l'essentiel de l'anecdotique).

Il est vrai que, jusqu'à présent peu de phénomènes nouveaux ont été mis en évidence en écologie marine par les méthodes factorielles ; d'une manière générale trois explications sont à avancer :

— ce sont des sujets extrêmement déflorés qui ont été abordés ;

— la collecte des données est fort souvent orientée

(1) On emploie souvent ce terme de modèle dans un sens très restrictif le limitant à son sens explicatif et prédictif, mais les outils de description mathématique supposent eux aussi l'utilisation d'un modèle.

(2) Il convient toutefois de souligner l'objectivité de la procédure d'ajustement d'un modèle. Frontier met lui-même en évidence l'intérêt de cette objectivité lorsqu'il souligne l'aspect peu convaincant de certains regroupements de variables plus établis en fonction des sentiments antérieurs de l'écologiste que de la réalité des données récoltées. Il est ainsi extrêmement utile de s'en remettre à une méthode de partition automatique, lorsqu'on veut regrouper des variables après analyse factorielle.

pour souligner des faits antérieurement connus (ex. : le cycle annuel du plancton ; la stratégie d'échantillonnage est définie le plus souvent pour préciser ce fait en soi trivial) ;

— un certain manque d'audace de l'utilisateur, notamment dans l'interprétation des axes éloignés de l'analyse factorielle.

Très schématiquement, pour regrouper les deux derniers points, on échantillonne souvent de façon à mettre en évidence un phénomène classique, comme la chronologie, le gradient côte-large, etc., qui immanquablement va extraire l'essentiel de la variance et on néglige les axes ultérieurs qui nécessairement ne bénéficient que d'une très faible variance résiduelle, mais qui pourtant, peuvent correspondre aux faits nouveaux attendus par l'écologiste.

La hiérarchisation des faits antérieurement connus a, quant à elle, donné des résultats satisfaisants jusqu'à présent. Ainsi peut-on définir l'importance relative de la salinité ou de la température dans la structure des écosystèmes planctoniques, estuariens ou deltaïques, l'importance relative de tel ou tel polluant, etc.

Si l'on a reproché souvent à l'analyse factorielle de confirmer des évidences, on ne parle jamais des pseudo-évidences que l'écologiste a pu rejeter grâce à elle (influence attendue d'un facteur quelconque non confirmée par l'analyse ou rejetée parmi les phénomènes secondaires). Si l'écologiste a en général assez d'imagination pour entrevoir l'essentiel des paramètres qui pourraient jouer, il lui est beaucoup plus difficile de cerner l'importance relative des différents facteurs.

Pour conclure ce chapitre, nous pouvons dire que, si le cadre contraignant de la publication oblige déjà le chercheur à structurer sa pensée, les exigences du modèle mathématique l'amènent encore davantage à être rigoureux au moment de la formulation du problème et à celui de l'exposé de ses conclusions. Même dans le cas improbable, où aucun fait nouveau ne serait apporté par l'approche mathématique, la transposition des résultats écologiques en une forme facilitant la communication objective serait un acquis essentiel. Un chercheur ne doit pas avoir pour but unique l'acquisition de la connaissance, il doit arriver à transmettre celle-ci. C'est à ce prix seulement que l'on pourra comparer de nombreuses situations écologiques, comparaison d'autant plus réaliste que tous les choix a priori auront été explicites, et qu'on sera assuré de parler rigoureusement de la même chose. Il n'existe plus de critères subjectifs cachés (flair de l'utilisateur) qui viennent fausser les comparaisons.

2. STRATÉGIE DU TRAITEMENT MATHÉMATIQUE

Après avoir évoqué l'intérêt potentiel de l'outil mathématique pour l'écologiste, nous allons discuter de quelques principes lui permettant d'en tirer le meilleur parti possible. Il faut d'une part connaître les conditions d'application, d'autre part, dans un contexte donné, choisir les outils les plus puissants et les combiner.

2.1. Limite de l'utilisation des méthodes

Dans le paragraphe précédent, nous avons souligné que la démarche normale procédait par le choix d'un type de modèle puis par un ajustement. Le jeu des données qu'on possède n'étant en général qu'un échantillon, un autre problème se pose immédiatement quant à la possibilité de généraliser les résultats obtenus (c'est-à-dire inférer de l'échantillon à la population). Bien des écologistes distinguent mal les hypothèses requises par les procédures d'ajustement de celles nécessaires à l'inférence.

2.1.1. Hypothèses requises par les procédures d'ajustement.

En fait, on peut contester, et le choix du modèle, et la procédure d'ajustement pour le problème traité. En ce qui concerne les hypothèses d'ajustement, on peut distinguer deux cas :

2.1.1.1. *Le modèle descriptif* : dans ce cas, aucune hypothèse *a priori* n'est nécessaire ; on a toujours le droit de chercher, par exemple, les axes d'inertie d'un nuage de points ou de rechercher la partition qui décrit le mieux le nuage.

2.1.1.2. *Le modèle prédictif (explicatif)* : l'exemple même en est la régression ; à ce niveau, on postule l'existence d'un certain nombre de mécanismes obéissant à un formalisme mathématique donné. Les risques d'erreur dans ce cas existent. Ainsi pour la multi-régression, si on a choisi un modèle linéaire, et qu'en définitive des effets quadratiques ou de degré supérieur sont essentiels, toutes les erreurs d'interprétation sont possibles.

ex : $y = ax + bz + \varepsilon$ modèle ajusté

$y = \alpha x^2 + \varepsilon$ modèle réel

avec $z = \beta x^2 + \varepsilon'$ ε, ε' résidus

Dans une telle étude de régression multilinéaire, on risque de conclure à l'action de z sur y et à l'indépendance de y vis-à-vis de x. La distinction que nous avons faite entre modèle descriptif et explicatif n'est pas toujours évidente, en pratique, dans la mesure où on peut, à l'usage, attribuer une valeur

explicative à un modèle au départ descriptif. C'est notamment le cas dans l'analyse factorielle lorsqu'on associe au système d'axes un modèle multilinéaire. Mais il demeure tout à fait légitime d'effectuer une analyse d'inertie lorsqu'un facteur agit de façon non linéaire, et l'apparition d'un nuage parabolique ou autre (effet Guttman simple ou complexe) est un résultat en soi. Ainsi, dans l'exemple de FRONTIER, un simple report dans les deux espaces duaux (variables et observations) lui aurait permis de démontrer l'existence d'une liaison non linéaire entre ses deux premiers facteurs, liaison qu'il ne fait que supposer.

Il convient tout de même de dire que, si les méthodes d'ajustement sont toujours utilisables, elles ne seront optimales que sous certaines hypothèses de nature statistique, mais encore faut-il définir l'optimalité.

Exemple : l'ajustement par les moindres carrés d'une régression multilinéaire. On postule d'abord que le modèle sous-jacent est réellement multilinéaire mais aussi pour justifier le choix des moindres carrés, on fait couramment appel à des hypothèses sur les résidus inexpliqués : on les suppose distribués suivant une loi normale centrée, de même variance et stochastiquement indépendants. Sous cette hypothèse, on aboutit à l'estimation du maximum de vraisemblance donc à un estimateur optimal dans un certain sens.

Que l'on s'écarte de ces hypothèses et cette optimalité disparaît, l'heuristicité restant en général satisfaisante.

2.1.2. L'INFÉRENCE.

Lorsqu'on parle d'extrapolation à partir d'un échantillon, ce ne peut être qu'une inférence sur la population que l'échantillon est censé représenter. Cette simple évidence est trop souvent négligée en écologie marine. On ne sait plus ce qu'un échantillon est censé représenter. La validité de l'extrapolation de l'échantillon à la population s'estime au moyen d'un test statistique de significativité. On lit couramment en statistique appliquée à l'écologie : le test suivant est significatif, mais l'opposition avec le terme « fortuit » est très mal perçue.

Le terme « fortuit » se définit également à partir du doublet population-échantillon. La population correspond souvent à un bloc spatio-temporel. Un échantillon est prélevé dans ce bloc spatio-temporel et une corrélation empirique est calculée sur cet échantillon. Un coefficient sera dit significatif, si les théories statistiques permettent de supposer que le coefficient calculé sur « tous » les points du bloc aurait été non nul ; sinon il sera dit fortuit. Cela n'empêche pas que si l'on prend un bloc spatio-temporel beaucoup plus vaste que le précédent, le nouveau coefficient global puisse être ou non nul.

La notion de « significativité » ou de « fortuité » n'a pas de sens dans l'absolu, hors du doublet *échantillon-population*. Les statistiques classiques ont développé une théorie extrêmement sophistiquée en ce qui concerne les inférences ; hélas ces tests reposent sur des hypothèses très strictes et nous tomberons d'accord avec FRONTIER pour reconnaître que le plus souvent, elles ne sont pas satisfaites dans le domaine qui nous concerne. Les tentatives pour élaborer des tests reposant sur des hypothèses moins strictes sont au départ fort sympathiques, encore faut-il noter que la puissance d'un test est, en règle générale, fonction de la restrictivité des hypothèses. En analyse multivariable les tests classiques reposent pour l'essentiel sur :

— la nature des distributions : normales ou multinormales,

— l'indépendance stochastique des échantillons.

La plupart des distributions rencontrées en écologie marine sont non normales voire non normalisables.

La continuité spatio-temporelle de la plupart des phénomènes étudiés rend dépendants les échantillons voisins. Nous nous retrouvons donc en accord avec FRONTIER pour souligner l'inadéquation des théories classiques à notre problème. Nous irons même plus loin avec BENZECRI (1973) pour dire qu'aucune théorie classique ne peut quantifier la probabilité d'obtenir dans les problèmes multidimensionnels une configuration *interprétable* et cependant *fortuite*.

Exemple : on obtient après analyse factorielle, un premier axe extrayant une part de variance, non significative suivant les tests classiques, mais ordonnant d'une manière satisfaisante des espèces suivant leur préférendum thermique connu antérieurement ; il serait ridicule de ne pas accepter ce résultat. Ceci illustre le fait que, si les théories classiques sont quelque peu prétentieuses (normalité, indépendance), elles sont incapables de prendre en compte les connaissances antérieures et perdent ainsi beaucoup de leur puissance.

Les tentatives pour surmonter le premier handicap des théories classiques sont extrêmement prometteuses, notamment par les méthodes non paramétriques (coefficient de rang ou codage).

Les tentatives mentionnées par FRONTIER pour lever d'emblée toutes les hypothèses des tests classiques appellent tout de même quelques commentaires :

— le test ε (IBANEZ, 1975) : Nous avons dit que l'inconvénient des méthodes classiques réside dans les hypothèses de normalité et d'indépendance des échantillons. Dès lors que ces hypothèses ne peuvent

être satisfaites par notre matériel, nous voyons mal l'intérêt d'introduire, dans le lot des variables écologiques ne respectant pas ces hypothèses, une nouvelle variable qui, elle, en respecte certaines.

Le bâton brisé : le test du bâton brisé employé par FRONTIER semble mal utilisé ; en raisonnant sur un exemple, nous allons essayer de démontrer ce défaut d'application :

Supposons que l'on étudie 20 espèces et 5 prélèvements extraits d'un bloc spatio-temporel. La variance ne pourra jamais se répartir qu'entre 4 facteurs (n points sont contenus dans un espace de dimension n — 1). Le test du bâton brisé sur les 20 variables espèces sera significatif. Ceci ne signifie nullement qu'en prenant « tous » les points du bloc spatio-temporel, on aurait obtenu une structure non sphérique. On a simplement mis en évidence une trivialité mathématique. L'objection que nous avons faite précédemment sur la mauvaise définition du doublet échantillon-population s'applique ici. Si le test est pratiqué sur les variables espèces, on ne voit à aucun moment apparaître la taille de l'échantillon (ici les 5 prélèvements). Cette erreur est probablement due pour une grande part, à la méconnaissance de la nature duale des problèmes (espace des espèces, espace des prélèvements) en analyse factorielle. D'autre part, on peut reprocher à FRONTIER le fait suivant : une fois la significativité de la première composante reconnue, il conviendrait de ne plus s'intéresser qu'à la variance résiduelle, et comparer donc, dans son exemple, le pourcentage de variance résiduelle expliquée par le deuxième facteur au plus grand segment d'un bâton brisé en 19 morceaux et non en 20, etc.

On peut dire enfin que ce test n'est pas débarrassé de l'hypothèque que représente la nécessaire indépendance des échantillons.

En l'état des choses, les tests classiques sont très peu utilisables. Cependant des voies de progrès sont apparues. On en est pourtant réduit, dans l'immédiat, à accepter les tests classiques comme de simples minorants ou majorants de l'incertitude, et à s'en remettre à un bon sens que la disponibilité d'un ordinateur n'interdit pas d'utiliser. On a fondé également beaucoup d'espoirs dans les méthodes de simulation de type Monte-Carlo. Elles permettent certes de générer des tests dans des cas où le calcul exact paraît impossible, les distributions étant trop compliquées, mais encore faut-il connaître la nature de la distribution.

Exemple : on pourra bâtir des tests, non plus sur des distributions multinormales, mais sur des distributions quelconques, mais connues, par exemple des distributions admettant des zéros à l'origine. On pourra de même prendre en compte la dépendance des échantillons, à condition de connaître la loi liant stochastiquement les échantillons.

2.2. Le choix des méthodes et leurs combinaisons

2.2.1. ADÉQUATION D'UNE MÉTHODE A UN PROBLÈME DONNÉ.

Il se trouve qu'historiquement la méthode en composantes principales fut une des premières méthodes de traitement des données ; c'est peut être une des raisons qui lui a fait attribuer une place excessive. De nombreuses méthodes plus ou moins récentes ont été mises au point qui permettent de couvrir un champ plus vaste de problèmes. On pourra citer pour exemple les techniques de partition automatique, coordonnées principales, correspondances, traitement des séries temporelles, etc. FRONTIER déplore que la technique factorielle ne travaille que sur des corrélations instantanées, et note que sur les données qu'il traite, il aurait aimé étudier les effets de décalage. Dans ce cas, l'utilisation des méthodes des séries temporelles est à suggérer ; bien évidemment des séries temporelles suffisamment longues sont alors nécessaires, mais on est là ramené aux problèmes d'échantillonnage.

Bien souvent l'outil existant ne répond pas tout à fait aux préoccupations de l'écologiste et certaines modifications de détails peuvent être précieuses. Ainsi, pour reprendre l'exemple de FRONTIER, enlever le gradient côte-large, facteur trivial qui semble l'embarrasser, aurait été probablement possible, par un simple centrage bloc par bloc, ou un codage différent selon l'emplacement géographique, comme l'a très bien illustré LE FOLL (1972). A cet égard, il faut maîtriser suffisamment l'outil pour savoir l'adapter.

2.2.2. LA PLEINE UTILISATION DE L'OUTIL.

Par-delà les aspects bien connus des différents outils, utilisation du pourcentage de variance, saturations (factor loadings), etc., il existe en général certaines grandeurs moins connues et qui peuvent éclairer les résultats. A notre avis, dans toutes les méthodes d'analyse d'inertie, l'utilisation des structures duales doit être systématique. Celle des contributions relatives (BENZECRI, 1973) pourra grandement éclairer les interprétations. L'introduction de variables supplémentaires à masses nulles, le report de barycentres, etc., si l'on considère les méthodes de régression, l'étude fine des résidus (distribution, indépendance, variance) apporteront souvent des résultats substantiels.

2.2.3. LA COMBINAISON DES OUTILS.

Il est essentiel de disposer d'un large éventail de méthodes, non seulement pour choisir la mieux adaptée, mais encore pour combiner ces méthodes.

Ainsi une étude de partition automatique dans les premiers axes permettra de faire des groupements objectifs, l'expérience prouvant que la limitation aux premiers axes clarifie les structures. Il peut être intéressant d'utiliser des codages et des distances différentes.

2.2.4. La planification de l'échantillonnage.

Trop souvent, l'écologiste ne se pose le problème du traitement des données qu'une fois celles-ci récoltées. Si les contraintes matérielles ne permettent certainement pas d'avoir une stratégie idéale, une réflexion a priori permettrait certainement une grande amélioration.

CONCLUSIONS

Nous tomberons d'accord avec Frontier en disant que l'analyse factorielle a été mal utilisée pour le traitement des données en écologie marine, mais cette phase d'apprentissage était certainement nécessaire. L'essentiel est d'en bien tirer les fruits et non pas de condamner la méthode. Il importe de souligner une évolution de point de vue, à notre sens, essentielle. De tentatives, probablement présomptueuses, pour identifier des modèles écologiques sous-jacents (analyse factorielle sensu stricto), on en revient à un point de vue plus mesuré, où la priorité est donnée à l'aspect descriptif, l'explication écologique étant accueillie avec plaisir lorsqu'elle se présente avec clarté. Dans le même temps, la gamme des outils disponibles s'est enrichie. Il reste à espérer que la communication se fasse assez bien entre statisticiens mathématiciens et écologistes pour que l'abondance de progrès méthodologiques ne submerge pas l'utilisateur, porté alors, à n'y voir qu'une suite de recettes. L'interprétation devra se faire encore plus avant pour que la collecte des données soit faite dans le sens du traitement spécifique qui suivra. Ceci nous semble essentiel pour le progrès de l'écologie, où l'information recueillie est trop importante pour être transmise sans une mise en forme sachant dégager les structures essentielles et les exprimer sous une forme constante.

Manuscrit reçu au S.C.D. de l'O.R.S.T.O.M. le 13 avril 1976.

BIBLIOGRAPHIE

Benzecri (J. P.) *et al.*, 1973. — L'analyse des données. I. La taxinomie : 624 pp ; II. L'analyse des correspondances. Dunod, Paris, 624 pp.

Frontier (S.), 1974. — L'analyse factorielle est-elle heuristique en écologie du plancton? *Cah. O.R.S.T.O.M.*, Sér. Océanogr., vol. XII, n° 1 : 77-81.

Ibanez (F.), 1975. — Contribution à l'analyse mathématique des évènements en écologie planctonique, optimisations méthodologiques, étude expérimentale en continu à petite échelle de l'hétérogénéité du plancton côtier. Tome 2, texte et travaux : 154 pp. Thèse Doct. État, Université de Paris VI.

Le Foll (Y.), 1972. — L'analyse factorielle des évolutions. Compte tenu du petit séminaire de statistique. Grenoble.

Imprimé par
le Service de Documentation
Centre d'Etudes Nucléaires de Saclay
Août 1977

ISSN 0336 - 3112

A RETOURNER

22/12/87

13. 11. 84