L'ANALYSE FACTORIELLE EST-ELLE HEURISTIQUE EN ÉCOLOGIE DU PLANCTON ?

SERGE FRONTIER

Océanographe biologiste de l'O.R.S.T.O.M.

RÉSUMÉ

L'hypothèse d'indépendance des observations est rarement réalisée dans les échantillonnages planctoniques, à cause des contraintes particulières au travail en mer.

Une série d'analyses en composantes principales sur 20 variables, appliquée à un ensemble de données relativement homogènes, aboutit seulement à déterminer deux vecteurs significatifs ; la signification du premier est triviale ; celle du deuxième n'a pas encore été exactement élucidée.

Les analyses rencontrées dans la littérature planctologique aboutissent également à déterminer un très petit nombre de vecteurs significatifs. L'écologiste utilise ses connaissances antérieures pour les interpréter. Le résultat est souvent trivial, quand il ne coïncide pas exactement avec les directions a priori *de l'échantillonnage.*

Des efforts sont entrepris (IBANEZ) pour rendre la méthode plus heuristique.

SUMMARY

Recalling basis hypothesis of factorial analysis make aware that these hypothesis are not realized in the main applications to planktological data. In fact, they are hardly feasible, because of the particular hindrances of sampling at sea. Observations are made along a ship route, that introduces stochastic relations between the various spatiotemporal directions and periodicities along which the ecological processes arise : for this, observations are not independant. After the analysis, it is difficult or impossible to distinguish the new information from the information which is initially put in with these stochastic relations. Furthermore, the numerous nul values in the initial data are biasing the analysis results.

Twenty principal component analysis upon 20 variates were applied to a set of data collected in relative homogenous conditions. Only two eigenvectors are found significative (50 to 60 % of the total variance). The signification of the first vector is trivial : it is the ecological gradient between shore and open sea, which was evident before the analysis. The second one is not yet precisely explained : perhaps it corresponds to the first factor too, taking into account its possibly non-linear effect and time lag.

Factorial analysis found in the planktological works generally starts from data observed in much more heterogenous conditions ; much more initial stochastic relations are introduced. For this, here too, very few eigenvectors are found significative. That signification is often trivial, if not exactly coincides with the sampling directions. In the interpretation of the results of analysis, the ecologist uses his previous knowledge on the subject, and find again (sometimes foggier) that he already knew. In this case, new *knowledges are not supplied by the analysis.*

Endeavour is undertaken (IBANEZ) to determinate the application conditions which are necessary to make factorial analysis heuristic in ecology of plankton.

Cah. O.R.S.T.O.M., sér. Océanogr., vol. XII, n° 1, 1974 : 77-81.

Pratiquant depuis deux ans, en collaboration avec F. IBANEZ (Station Zoologique de Villefranche-sur-Mer) le traitement des données planctoniques par l'analyse factorielle, je crois aujourd'hui utile de faire part de quelques réflexions.

Par souci d'employer des « techniques de pointe » on utilise depuis quelques années, en planctologie, des méthodes d'analyse multivariée et en particulier d'analyse factorielle. L'écologiste prend alors généralement l'attitude classique consistant à « faire confiance aux mathématiciens » pour ce qui est de la justification de la méthode, dont le principe lui échappe parfois en dehors de quelques points très généraux :

— projection d'un ensemble de données dans un espace vectoriel de dimension réduite, judicieusement choisi; on obtient ainsi une description simplifiée mais suffisante de la situation observée, la part de variance abandonnée étant assimilée à un bruit ininterprétable;

— assimilation des vecteurs principaux à des processus indépendants les uns des autres, auxquels on cherche à faire correspondre des facteurs agissant sur l'écosystème;

— groupement des variables initiales (espèces biologiques) en sous-ensembles plus ou moins distincts, supposés représenter des groupes naturels d'espèces.

Cette attitude présente des dangers.

L'application de la méthode suppose réalisées des hypothèses, que l'on vérifie rarement ou que la vérification infirme. Nous ne prendrons pour exemple que les hypothèses de l'analyse des composantes principales.

*
* *

Elles stipulent tout d'abord la *multinormalité des données*. En fait, l'analyse se révèle assez robuste vis-à-vis d'une déviation par rapport à cette première condition, et une normalisation approximative des données suffit (les transformations log, $\log^2$, racine cubique, etc. donnent les mêmes résultats finaux : IBANEZ, 1971).

Notons ici que l'analyse est également très robuste vis-à-vis de la *précision sur les données de départ*. La caractérisation des abondances planctoniques au moyen de classes logarithmiques très larges donne les mêmes résultats, au terme de l'analyse, que les comptages précis (FRONTIER et IBANEZ, 1974; IBANEZ, *en préparation*). Les comptages précis sont donc inutiles; il s'ensuit un gain de temps considérable dans le dépouillement des récoltes.

Plus lourde de conséquences est l'hypothèse d'*indépendance des observations*. L'analyse ne devrait s'appliquer qu'à un ensemble de données recueillies « au hasard » et indépendamment les unes des autres. Ce ne peut être le cas en écologie du plancton : en effet, les contraintes de l'échantillonnage en mer introduisent au départ des *liaisons entre les observations* et des *liaisons entre les directions spatio-temporelles suivant lesquelles se produisent les processus écologiques* que l'on désire étudier. Les résultats de l'analyse comprennent alors, pour une part, l'information introduite artificiellement, au départ, par ces liaisons; l'information réellement nouvelle peut être difficile à discerner.

Concrètement, l'échantillonnage est réalisé à partir d'un navire, qui parcourt un trajet. Ce dernier introduit *a priori* des liaisons entre les coordonnées d'espace et de temps. Pour prendre un exemple caricatural : supposons que le navire aille du nord au sud à une époque où se produisent des variations écologiques rapides, on ne saura alors ce qui revient, dans les variations d'une quantité observée, au changement de latitude et au changement de saison. Deux variables corrélées l'une avec le temps, l'autre avec l'espace, pourront être bien exprimées par le premier axe principal, et cette liaison sera fallacieuse puisqu'uniquement due aux caractéristiques de l'acte d'échantillonnage.

Plus généralement, l'acte d'échantillonnage introduit un ensemble de corrélations entre

— les différentes directions spatio-temporelles;

— les différentes échelles auxquelles se produisent les variations le long d'une même direction (échelle des temps : périodicités nycthémérale, liée à la marée, saisonnière, etc.; échelle des distances : structures à l'échelle du décamètre, du kilomètre, des 100 kilomètres ...). On peut alors effectuer sur l'ensemble des coordonnées des échantillons une analyse en composantes principales ayant pour effet de « résumer » la structure de l'artefact. Après avoir analysé l'ensemble des variables biologiques, on s'aperçoit que les premières composantes biologiques sont plus ou moins corrélées avec les premières composantes spatio-temporelles, ce qui rend l'enchevêtrement des processus écologiques difficile à décomposer. En toute logique, il conviendrait de n'appliquer l'analyse à un ensemble de données planctologiques qu'à la condition que les corrélations entre les directions spatio-temporelles de l'échantillonnage soient négligeables. Cela ne peut guère être réalisé que si un plan d'échantillonnage a été conçu à l'avance pour qu'il en soit ainsi (IBANEZ, 1973). Or les plans de campagnes océanographiques sont généralement conçus en fonction des trajets de navire acceptables par le Commandant et / ou l'Administration, ce qui est très différent ...

Autres causes, plus classiques, d'incertitude : l'arbitraire sur le *nombre d'axes principaux retenus*

comme significatifs; l'arbitraire sur la *délimination des « groupes de variables »* dans les plans factoriels. Sur ces deux points, en l'absence d'algorithme mathématique rigoureux et satisfaisant, l'intuition et l'expérience acquise sont appliquées plus souvent que le raisonnement déductif. Il en est de même de l'« *identification* » *des axes à des facteurs de l'écosystème.*

Examinons des exemples concrets. Nous envisagerons d'abord le traitement, par analyse en composantes principales, des données planctologiques recueillies dans le cadre du programme « Baie d'Ambaro » du Centre O.R.S.T.O.M. de Nosy-Bé (Madagascar) (*). Une baie de 800 km² a été quadrillée vingt-fois en un an, au moyen d'un réseau de 44 stations. Chaque quadrillage a fait l'objet de deux analyses : l'une portant sur les coordonnées des récoltes (latitude, longitude, fond, marée, temps écoulé depuis le début du quadrillage, heure); l'autre sur 20 variables biologiques.

Avant toute interprétation des résultats des analyses, nous avons essayé de déterminer le nombre d'axes « significatifs ». Nous avons pour cela imaginé deux tests empiriques :

1°) TEST « ε » (IBANEZ, 1973). On introduit une vingt-et-unième variable « biologique » fictive, à valeurs calculées en chaque station au moyen d'une table de nombres au hasard, et on refait l'analyse. La variable ainsi fabriquée se trouve bien exprimée à partir d'un certain vecteur propre; on admet alors que ce dernier n'exprime que le hasard. Dès lors, il est légitime de considérer que seuls les vecteurs propres de rang inférieur peuvent avoir une signification.

2°) TEST DU « BÂTON BRISÉ » (FRONTIER, *non publié*). On compare la répartition de la variance totale entre les 20 vecteurs propres successifs, non à une équirépartition (extrêmement improbable), mais à la répartition moyenne dans le partage au hasard d'une quantité fixe en 20 termes classés par ordre décroissant (problème classique du « bâton brisé » — voir par exemple PIELOU, 1969, p. 214). Les premiers axes sont considérés comme significatifs s'ils extraient visiblement plus de variance que ne le prévoit le modèle aléatoire. A partir d'un certain rang, les vecteurs cessent d'extraire plus de variance que prévu, et se partagent la variance résiduelle : on estimera qu'ils sont non-significatifs, et ne décrivent que le bruit.

Appliqués à nos données, on constate que les deux tests coïncident (voir figure 1) : ε apparaît presque toujours exprimé par le premier vecteur non significatif au sens du test « bâton brisé ». Dans 11 cas sur 20, seuls les deux premiers axes (totalisant 50 à 60 % de la variance) sont ainsi retenus; dans 9 cas un troisième vecteur est également significatif.

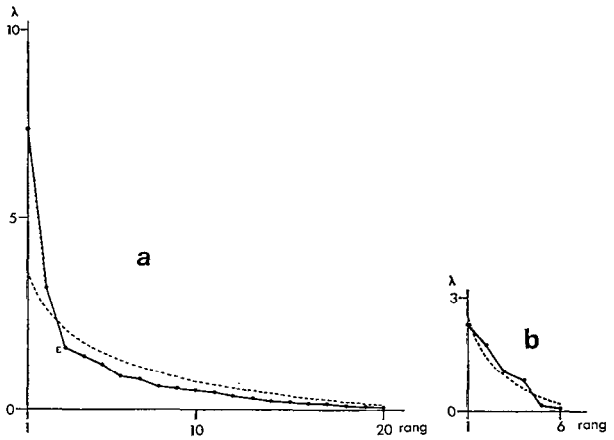

Fig. 1. — *En trait plein* : décroissance des valeurs propres λ en fonction du rang. *a* : analyse en composantes principales de 20 variables biologiques ; ε indique le premier vecteur où la variable fictive apparaît bien représentée ; *b* : analyse en composantes principales de 6 coordonnées spatio-temporelles lors du même quadrillage (test ε non réalisé). *En trait pointillé* est indiquée dans chaque cas la décroissance théorique moyenne dans le modèle du « bâton brisé ».

Nous avons tenté d'« identifier » ces axes en cartographiant les coordonnées des points-observations dans la base des vecteurs propres. Le premier axe (38 %, en moyenne, de la variance totale) prend alors une signification évidente : c'est le gradient côte-large de conditions écologiques; on retrouve les lignes de niveau du fond marin. Par contre, la signification des deuxième et éventuellement troisième axes n'a pas encore été élucidée; la cartographie fournit des images assez cohérentes (contrairement aux axes suivants), mais ne rappelant la répartition d'aucun facteur écologique présumé, et par surcroît très variables d'une date à l'autre. Peut-être le deuxième vecteur décrit-il un facteur trophique général; peut-être s'agit-il simplement du premier facteur, compte tenu d'une non-linéarité de ses effets et / ou d'un délai d'action; ce délai permettrait au peuplement planctonique d'être repris et brassé par les courants locaux : ainsi, la cartographie n'exprimerait que la turbulence hydrodynamique à l'échelle de la baie. Des exemples de cartographie des axes principaux sont donnés dans FRONTIER et IBANEZ, 1974, pour un quadrillage ayant donné trois vecteurs significatifs.

L'analyse des données spatio-temporelles se révèle peu significative : la répartition de la variance entre les vecteurs propres suit de très près le modèle aléatoire du « bâton brisé ». Pourtant, des coefficients

(*) Voir FRONTIER, 1971.

de corrélation significatifs apparaissent fréquemment entre les composantes principales « biologiques » et « spatio-temporelles ». Ces corrélations sont fortuites; toujours est-il qu'elles rendent encore plus délicates l'interprétation des axes.

Un effort pour appliquer la méthode des composantes principales à un ensemble relativement homogène de données (secteur marin limité, écologiquement individualisé sinon homogène, quadrillé en un temps court) a donc abouti à déterminer :

— une première composante principale de signification quelque peu triviale : le gradient côte-large avait été mis en évidence depuis longtemps;

— une deuxième composante que nous n'avons pas encore pu identifier avec certitude. Elle ne représente peut-être que l'influence du principal (unique ?) facteur discernable à cette échelle d'observation, compte tenu d'un délai d'action et d'une non-linéarité possible de ses effets (ces caractéristiques pouvant d'ailleurs varier selon les variables biologiques). Rappelons que l'analyse part d'une matrice de corrélations instantanées, et se réfère à un modèle multilinéaire!

— une troisième composante, apparemment significative dans certains cas, mais ininterprétée;

— 17 composantes non significatives, réalisant 40 à 50 % de la variance totale.

Quant au classement des espèces selon leurs positions dans les plans factoriels, il ne reflète pour le premier axe que le classement côte-large, visible avant analyse; pour le second, il tendrait à déterminer des groupes très variables d'une date à l'autre, même rapprochées d'une semaine.

* * *

Si l'on examine maintenant les analyses publiées dans la littérature planctologique (Binet, 1968; Boucher, 1970; Dandonneau, 1971; De Bovée, 1970; Ibanez, 1968; Ibanez et Dallot, 1969; Ibanez et Seguin, 1972; Blanc *et al.*, 1972; Reyssac et Roux, 1972 — pour ne citer que quelques travaux français récents), on relève des constatations pouvant se grouper sous trois rubriques :

(1) Les échantillonnages sont considérablement plus hétérogènes que dans l'exemple donné ci-dessus. Ils permettent ainsi à des processus agissant selon des directions spatio-temporelles différentes, et à des échelles différentes, de se manifester simultanément. Il s'ensuit qu'un plus grand nombre d'axes, associés à autant de « facteurs » indépendants, est généralement retenu. Mais l'échantillonnage est sélectif, certaines directions ou échelles étant mieux échantillonnées que d'autres; enfin, des liaisons importantes sont introduites entre les diverses coordonnées de l'échantillonnage. Les conclusions concernant l'écosystème analysé comporte donc une part d'arbitraire, liée aux caractéristiques de la saisie des données. Ibanez (1973) analyse à cet égard la campagne océanographique « Mediprod 1 ».

(2) Le nombre d'axes principaux retenus fait l'objet d'une décision principalement fondée sur l'« évidence » de l'interprétation ultérieure. Cette évidence est considérée comme se justifiant d'elle-même, et n'est pas autrement vérifiée. Or un test appliqué après-coup (fondé, comme nous l'avons dit, sur une comparaison avec ce que donnerait un tirage au hasard) permet parfois d'établir que ces axes « évidents » étaient fortuits (exemples dans Ibanez, 1973).

(3) L'interprétation des axes retenus, ainsi que les groupements de variables, sont essentiellement réalisés en fonction de l'expérience antérieure de l'écologiste : celui-ci « retrouve » les groupes d'espèces déjà établis, et les délimite dans les plans factoriels à l'aide de lignes sinueuses, voire contournées. Au terme de l'analyse, il n'a rien trouvé de *nouveau*. En ce sens, la méthode apparaît jusqu'ici peu heuristique. Par ailleurs, prétendre qu'on a de cette façon démontré ce qui n'était jusqu'alors qu'intuitif, est un abus de langage. En effet, une démonstration est l'établissement de conclusions sûres à partir de prémisses vérifiées. Tout au plus la méthode permet-elle ici de donner une *description* géométrique simplifiée d'un ensemble complexe de données.

D'autres techniques d'analyse factorielle, parfois moins critiquables que l'Analyse en Composantes Principales, car faisant appel à moins d'hypothèses invérifiables ou irréalisables (l'Analyse des Correspondances par exemple) ne semblent guère plus satisfaisantes à cet égard dans leur application à la planctologie (Binet *et al.*, 1972 a et b; Ibanez et Seguin, 1972; Reyssac et Roux, 1972).

Nous avons passé sous silence d'autres conditions d'application, le plus souvent délibérément transgressées, telle que la nécessaire *rareté des valeurs nulles* dans l'ensemble des données initiales. L'analyse prend en compte les coefficients de corrélation élevés entre deux espèces rares, traduisant seulement le fait que les deux espèces sont simultanément absentes de la plupart des récoltes. De plus, une valeur nulle peut signifier que l'espèce est absente à la station, mais aussi qu'elle est rare et n'a pas été capturée en raison de l'incertitude d'échantillonnage; l'analyse ne distingue en rien ces deux cas. Nous pensons que les problèmes liés à la signification des zéros pourraient être résolus au moyen d'une pondération des

variables en fonction de leur fréquence, pondération aboutissant à donner plus d'importance aux co-occurrences d'événements rares. Encore faudrait-il que la pondération n'altère pas le reste de l'analyse. Le problème est à l'étude (IBANEZ, *en préparation*).

Nous ne conclurons pas en déclarant injustifiée l'application de l'analyse factorielle à l'écologie du plancton, mais en déclarant que la méthode est encore insuffisamment maîtrisée par ses utilisateurs, au nombre desquels nous figurons. Un effort s'impose donc (que nous avons entrepris, mais que nous souhaitons voir se généraliser) afin d'établir plus rigoureusement les conditions d'application, compte tenu des contraintes particulières à l'échantillonnage en mer.

Manuscrit reçu au S.C.D. le 13 mars 1974.

BIBLIOGRAPHIE

BINET (D.), 1968 — Variations saisonnières du zooplancton et plus particulièrement des Copépodes du plateau continental de Pointe-Noire (Congo). Thèse 3e Cycle, Paris, *multigr.* 145 p.

BINET (D.), DESSIER (A.), GABORIT (M.), ROUX (M.) — 1972 a — Premières données sur les Copépodes pélagiques de la région congolaise. II) Analyse des Correspondances. *Cah. O.R.S.T.O.M., sér. Océanogr.* 10 (2) : 125-138.

BINET (D.), GABORIT (M.), ROUX (M.) — 1972 b — Copépodes pélagiques du plateau ivoirien. Utilisation de l'Analyse des Correspondances dans l'étude des variations saisonnières. *Doc. scient. Centre Rech. Océanogr. Abidjan*, 3 (1) : 47-79, *multigr.*

BLANC (F.), LEVEAU (M.), BONIN (M.-C.), LAUREC (A.) — 1972 — Écologie d'un milieu eutrophique : traitement mathématique des données. *Mar. Biol.* 14 (2) : 120-129.

BOUCHER (J.) — 1970 — Écologie et relations trophiques du zooplancton en Méditerranée nord-occidentale. Première partie. Thèse 3e Cycle, Paris, *multigr.* 130 p.

DANDONNEAU (Y.) — 1971 — Étude du phytoplancton sur le plateau continental de Côte d'Ivoire. I) Groupes d'espèces associées. *Cah. O.R.S.T.O.M. sér. Océanogr.* 9 (2) : 247-266.

DE BOVÉE (F.) - - 1970 — Écologie et relations trophiques du zooplancton en Méditerranée nord-occidentale. Deuxième partie. Thèse 3e Cycle, Paris, *multigr.* 130 p.

FRONTIER (S.) — 1971 — Présentation de l'étude d'une baie eutrophique tropicale : la baie d'Ambaro (côte nord-ouest de Madagascar). *Cah. O.R.S.T.O.M., sér. Océanogr.* 9 (2) : 147-148.

FRONTIER (S.) et IBANEZ (F.) — 1974 — Utilisation d'une cotation d'abondance fondée sur une progression géométrique, pour l'analyse des composantes principales en écologie planctonique. *J. exp. mar. Biol. Écol.* 14, *sous presse.*

IBANEZ (F.) — 1968 — Application de l'Analyse factorielle en écologie : écologie et taxinomie numérique. Thèse 3e Cycle, Paris, *multigr.* 130 p.

IBANEZ (F.) et DALLOT (S.) — 1969 — Étude du cycle annuel des Chaetognathes planctoniques de la rade de Villefranche par la méthode de l'Analyse des Composantes Principales. *Mar. Biol.* 3 (1) : 11-17.

IBANEZ (F.) — 1971 — Effet de la transformation des données dans l'analyse factorielle en écologie planctonique. *Cah. Océanogr.* 23 : 68-80.

IBANEZ (F.) et SEGUIN (G.) — 1972 — Étude du cycle annuel du zooplancton d'Abidjan. Comparaison de plusieurs méthodes d'analyse multivariable : Composantes Principales, Correspondances, Coordonnées Principales. *Inv. Pesq.* 36 (1) : 81-108.

IBANEZ (F.) — 1973 — Méthode d'analyse spatio-temporelle du processus d'échantillonnage en planctologie, son influence dans l'interprétation des données par l'analyse en composantes principales. *Ann. Inst. Océanogr.*, n.s., 49 (2) : 83-111.

PIELOU (E. C.) — 1969 — An Introduction to mathematical Ecology. Wiley Intersci. 235 p.

REYSSAC (J.) et ROUX (M.) — 1972 — Communautés phytoplanctoniques dans les eaux de la Côte d'Ivoire. Groupes d'espèces associées. *Mar. Biol.* 13 (1) : 14-33.