# Sur la segmentation des séries chronologiques planctoniques multivariables

Séries chronologiques
Techniques multivariables
Segmentation
Plancton
Chronological series
Multivariate treatment
Segmentation
Plankton

F. Ibanez
Station Zoologique, 06230 Villefranche-sur-Mer, France.

Reçu le 15/3/84, accepté le 4/4/84.

RÉSUMÉ

L'étude des séries en océanographie montre que leur structure et leurs interrelations ne sont pas stables au cours du temps ou sur une radiale, et qu'un traitement global de ces données est mal adapté. L'écologiste doit avant tout identifier des séquences qui correspondent à des états particuliers de l'écosystème pélagique. Nous présentons ici quatre critères de classification : l'estimation de la quantité d'information apportée par le signal à chaque observation; la détection des variations ponctuelles et simultanées des amplitudes et des corrélations des descripteurs; la localisation des changements de sens des gradients d'abondance des populations; la partition optimale de la série multivariable en périodes de dispersion homogène.
Ces segmentations répondent aux impératifs des analyses quantitatives ultérieures, mais elles peuvent également constituer l'objectif fondamental de l'interprétation des séries chronologiques.

*Oceanol. Acta*, 1984, 7, 4, 481-491.

ABSTRACT

Segmentation of planktonic multivariate chronological series

The study of chronological oceanographical series shows that their structure and their relationships vary both with time and along transects, and that the global treatment of data is unsuitable. The ecologist must first identify sequences corresponding to particular properties of the pelagic ecosystem. Four bases for segmentation are presented: estimation of the quantity of information in the signal at each data point; automatic detection of simultaneous variations in amplitude and correlation of the descriptors; location of direction changes in population gradients; optimal segmentation of the multivariate series to find homogeneous zones.
The location of specific types of sequences is necessary for future mathematical treatments, but segmentation can also be considered as the principal object of the investigation of chronological series.

*Oceanol. Acta*, 1984, 7, 4, 481-491.

## INTRODUCTION

L'échelle et la densité des observations sont indissociables de l'interprétation des série chronologiques. En écologie planctonique on distingue trois types de chroniques :

— les séries courtes (moins de 100 prélèvements), qui correspondent soit à un cycle annuel avec des prélèvements hebdomadaires, soit à des radiales avec des observations espacées de plusieurs milles. L'échelle de ces travaux est vaste et les pêches de plancton s'effectuent au filet;

— les séries pluriannuelles : de nombreux laboratoires disposent de comptages réguliers sur un grand nombre d'années avec une fréquence d'échantillonnage hebdomadaire ou bimensuelle;

— les enregistrements en continu par pompage sur des radiales ou en point fixe, qui comprennent plusieurs centaines de récoltes espacées de 1 à 10 minutes.

L'analyse des séries courtes consiste en la description des fluctuations des différents descripteurs et de leurs similitudes. Bien que l'indépendance entre les observations successives ne soit pas réalisée (condition nécessaire à l'inférence statistique), on a transposé pour ce type de séries les techniques multivariables classiques, analyse en composantes principales, analyse des corres-

pondances, analyse discriminantes (Cassie, 1958; 1967; Colebrook, 1964; Ibanez, Dallot, 1969; Ibanez, Seguin, 1972; Binet *et al.*, 1972).

De façon plus précise, Dessier et Laurec (1978) et Estève (1978) ont utilisé l'analyse en composantes principales et l'analyse des correspondances, comme des représentations purement graphiques qui peuvent servir à mettre en évidence les successions de zones faunistiques homogènes et de transitions.

La finalité de ces techniques est de définir à la fois des groupes de descripteurs et des groupes d'observations. Mais, comme la notion d'anisotropie temporelle n'est pas introduite, l'ordination des observations obtenue ne peut respecter exactement la connexité temporelle; ainsi il n'est pas rare de trouver des points très éloignés dans le temps plus rapprochés que des observations successives.

Une technique récente de classification (Legendre *et al.*, 1982), permet une segmentation des séries dans le respect de la relation de connexité. Cependant cette partition peut provenir aussi bien d'un changement des gradients des populations que d'une variation plus ou moins accidentelle de leur hétérogénéité.

Nous montrerons que dans la mesure où on n'utilise pas les données originales, mais strictement les tendances générales, on peut opérer une segmentation des séries courtes.

L'analyse des séries pluriannuelles soulève moins de difficulté, car leur étendue permet une décomposition simple en trois composantes indépendantes, générale, saisonnière et résiduelle. La méthode Census II (Shiskin, Eisenpress, 1957) permet une approximation rigoureuse de la tendance, car le lissage adopté n'introduit aucune périodicité parasite (Bethoux *et al.*, 1980). La segmentation peut s'effectuer à partir d'une analyse d'inertie sur les tendances. Nous proposons ici également une technique adaptée à la partition globale des séries résiduelles.

Pour les enregistrements en continu, nous devons distinguer deux types de traitements : le traitement séquentiel et le traitement *a posteriori* au laboratoire.

Le traitement séquentiel est une approche stratégique; il a avant tout pour objectif la définition immédiate des discontinuités, afin d'improviser au besoin un complément d'échantillonnage (pêches de plancton au filet, coupes hydrologiques verticales).

Certains auteurs proposent d'appliquer les techniques d'analyse en composantes principales (Kelley, 1972), de multirégression (Cruzado, 1971), d'analyse discriminante (Webster, 1973), entre des blocs de stations successifs (environ 10 stations); mais ces procédés ont le désavantage de ne pouvoir repérer exactement la place des discontinuités. Une définition ponctuelle plus précise peut être obtenue si on utilise les modèles de processus adaptatifs prédictifs (Gilchrist, 1976), ou le D 2 au centre (Ibanez, 1981).

Pour économiser le stockage des données en continu, on peut utiliser la notion d'information liée à chaque observation : on ne retiendra que celles dont la quantité d'information est supérieure à un seuil fixé à l'avance (Ibanez, 1982).

Il n'est pas possible de transposer directement les techniques multivariables pour les traitements des séries continues au laboratoire. On devra, d'une part reconnaître globalement les changements de populations, c'est-à-dire effectuer une partition sur l'évolution des gradients, d'autre part étudier les phénomènes de haute fréquence dans des blocs statistiquement homogènes.

Il faut donc d'abord trouver la définition la moins arbitraire possible des tendances générales qui permettront une ordination ou une classification des stations respectant la connexité spatio-temporelle.

On devra ensuite définir des zones où les résidus ont une variance homogène pour pouvoir appliquer l'analyse spectrale (la « fast Fourier transform » si les portions sont longues, l'analyse spectrale dite du maximum d'entropie dans le cas inverse).

Les méthodes exposées dans cet article apportent des éléments de solution à tous ces problèmes.

Pour plus de clarté dans l'illustration, nous avons choisi une application sur des séries courtes, bien que les procédés utilisés conviennent également au traitement des autres types de données.

## MÉTHODOLOGIE

### La quantité d'information liée à un signal multivariable

Il est possible d'estimer, pour une série chronologique multivariable, la probabilité d'apparition d'un point de retournement (un pic ou un creux), compris entre deux situés de part et d'autre (Ibanez, 1982) :

$$P(i) = 2 \times \frac{1}{n(i-1)!(n-i)!},$$

où $n$ correspond au nombre d'observations séparant les deux points de retournement extrêmes, et $i$ la $n$-ième place du point de retournement considéré.

Connaissant la probabilité d'occurrence au hasard d'un point de rupture, on déduira une quantité d'information par la formule de Shannon :

$$I = -\mathrm{Log}_2\, P(i).$$

On peut donc définir pour chaque série une courbe de l'information, si après avoir calculé l'information de chaque point de retournement, on effectue une interpolation linéaire pour les observations intermédiaires. En prenant le seuil de probabilité 5% (4,3 bits) ou 10% (3,32 bits), on reconnaîtra les évènements ayant au plus 5 ou 10% de chances de se produire (ces indices ne peuvent être considérés comme des tests d'hypothèse au sens statistique, mais comme une échelle, un repère descriptif).

Grâce à la propriété d'additivité de I, on peut calculer une courbe de l'information totale d'une série multivariable, en faisant la somme pour chaque observation des informations correspondant à chaque descripteur. En traçant une ligne parallèle aux abscisses correspondant au seuil statistique, on va découper la série en intervalles de temps où alternent de fortes et de faibles

quantités d'information. Les périodes où l'information est très basse expriment une fréquence trop faible de l'échantillonnage ou l'absence de phénomènes continus; par contre celles pour lesquelles l'information est élevée soulignent les tendances locales, l'intensité des gradients.

**Reconnaissance statistique des hétérogénéités**

Bemont et Waterman (1977) ont montré comment localiser automatiquement les pics importants dans une série univariable par la reconnaissance des segments de variance maximale.

J'ai proposé en 1981 un index permettant de tester, au besoin en temps réel, les discontinuités d'un enregistrement multivariable. Le principe est d'associer à chaque observation sa distance généralisée ($D^2$ de Mahalanobis), vis-à-vis du centre de gravité des $n$ observations précédentes.

$$D^2 = X' R X,$$

où $D^2$ est le carré de la distance généralisée entre un point de coordonnées X et le centroïde des $n$ observations précédentes, et R la matrice de corrélation entre les descripteurs (si les coordonnées sont centrées réduites). Cette expression est identique à un $\chi^2$ à $m$ degrés de liberté pour $m$ descripteurs sous réserve de multinormalité. Si le $\chi^2$ est significatif à 5‰, l'observation considérée a moins de 5 chances sur 100 d'appartenir au groupe des $n$ observations qui la précèdent (là encore ce test n'a qu'un intérêt descriptif).

Les $n$ points retenus définissent une « fenêtre » dont la largeur déterminera la sensibilité de l'indice. Une largeur de fenêtre optimale peut être définie par la fonction Auto D 2 (Ibanez, 1976), qui donne une estimation de l'échelle moyenne des changements temporels d'une série multivariable. Le calcul de l'Auto D 2 est comparable à celui de la fonction d'autocorrélation avec la métrique du $D^2$ de Mahalanobis. On considère la distance du $D^2$ entre la série originale et celle-ci décalée pas à pas d'une unité d'échantillonnage. Pour une série de T valeurs et pour un décalage $\theta$, on aura :

$$B = \frac{1}{T-\theta}\left(\sum_{t=1}^{T-\theta} X_t - \sum_{t=\theta+1}^{T} X_t\right),$$

avec $B = 0, 1, 2, \ldots, T/2$, et le premier terme correspondant au vecteur des moyennes d'un premier groupe d'observations allant de $t$ à $T-\theta$, le deuxième terme correspondant au vecteur des moyennes d'un second groupe allant de $\theta+1$ à T. On a alors :

$$\text{Auto D 2}(\theta) = B' S^{-1} B,$$

S étant la matrice de variance-covariance intragroupe (ou la matrice de corrélation si les variables sont standardisées). La largeur de la fenêtre à utiliser pour le calcul du $D^2$ au centre doit être au moins égale au décalage qui correspond à un maximum de la fonction Auto D 2, autrement dit supérieure à l'échelle moyenne des changements dans la série multivariable.

Compte tenu des propriétés de la métrique, les observations à fort $D^2$ correspondront à un changement important dans l'amplitude des variations, et également des corrélations entre les différents descripteurs.

**Décomposition d'une série multivariable et nécessité des segmentations**

Les séries chronologiques planctoniques correspondent à des fluctuations cycliques et/ou aléatoires autour d'une tendance générale. L'interprétation doit se faire en deux étapes :

— extraction des tendances globales. Leur interprétation et leur modélisation doivent se faire après qu'aient été définies les observations correspondant au changement de sens des gradients;

— recherche des périodicités sur les séries résiduelles. Pour se rapprocher de la stationnarité, on doit découper la série (ou les séries), en segments de variance homogène.

*La partition sur les tendances*

L'approximation de celles-ci par un modèle linéaire ou polynominal reste arbitraire sans hypothèse de départ; ces fonctions seront en outre différentes d'un descripteur à l'autre. Le filtre classique par moyenne mobile présente également deux inconvénients majeurs : le choix de l'intervalle de lissage n'est pas déterminable *a priori*, et dans le cas où on veut faire apparaître une tendance générale bien nette, on va perdre un nombre d'observations non négligeable aux extrémités de la série. Notons également qu'un lissage important entraîne l'apparition de phénomènes périodiques parasites (effet Slutsky-Yule).

C'est pourquoi je propose une technique empruntée à l'économétrie : la méthode graphique dite des points médians (Ibanez, 1984). Elle consiste à définir la tendance par la courbe lissée qui correspond au lieu des points situés à une distance moyenne de l'enveloppe de la courbe originale (*voir* plus loin figure 2).

Dans une étude préliminaire, nous avons constaté que les phénomènes de basse fréquence sont intégralement conservés sur ces séries : la densité spectrale est identique à celle obtenue sur les données originales pour des fréquences inférieures à 0,2.

La figure 1 montre les cohérences entre les séries lissées et originales pour trois catégories planctoniques : nauplii, copépodites, copépodes adultes. Ces séries correspondent à 280 prélèvements par pompage (filtration de 1 $m^3$ par minute; Ibanez, 1976). On note que les cohérences sont toutes significatives lorsque la fréquence est inférieure à 0,2, qu'elles sont parfois significatives jusqu'à 0,3, et négligeables au-delà.

Ainsi notre lissage élimine les oscillations dont la période est inférieure à 5 observations.

Notre filtrage a les propriétés suivantes :

— conservation intégrale du nombre d'observations;

— élimination des phénomènes aléatoires de haute fréquence;

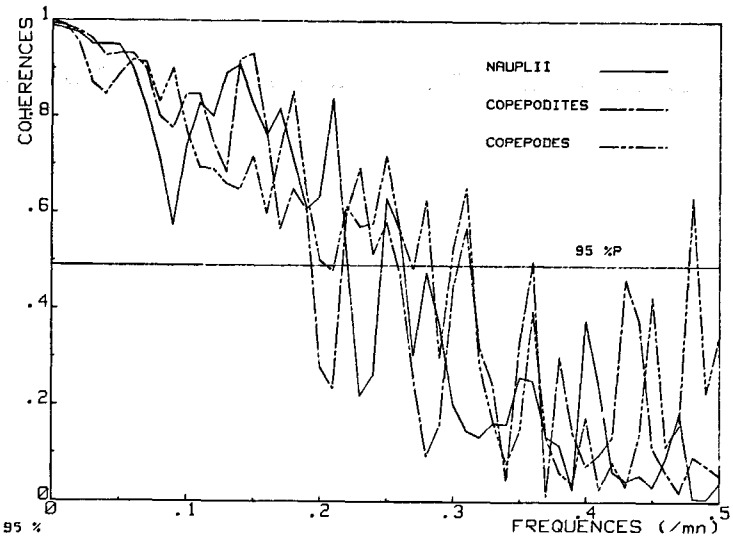


Figure 1
Spectre de cohérence entre les données lissées et originales pour 3 catégories planctoniques.
*Coherence spectrum: smoothed and original data for 3 planktonic populations.*

— identité du processus de lissage pour tous les descripteurs;
— pas d'introduction de phénomènes périodiques artificiels.

Une étude comparative des résultats obtenus à partir de ce lissage au lieu des données originales, a été réalisée en considérant l'analyse en composantes principales. Par rapport aux données brutes, nous avons constaté les propriétés empiriques suivantes :

1) Les moyennes des données lissées sont plus faibles et les variances sont plus homogènes.

2) Les corrélations entre les descripteurs sont plus élevées.

3) Les pourcentages d'inertie des axes principaux augmentent [exemple : tableau 1 : ACP données du cycle annuel des cinq catégories de chaetognathes (*voir* l'exemple traité plus loin)].

4) La composition des groupes de descripteurs reste inchangée.

5) Les composantes relatives aux tendances correspondent à un lissage parfait de celles obtenues sur les données originales.

6) L'ordination des observations dans les plans principaux conserve la connexité spatiale et/ou temporelle des observations.

La succession des stations se traduit dans l'espace réduit par un tracé rectiligne marqué par des points d'inflexion très reconnaissables (*voir* plus loin figure 7).

Ces points d'inflexion repérables au niveau de l'ordination, vont donc permettre un autre type de découpage d'une série multivariable, car ils signent les changements de sens des gradients de populations.

Tableau 1
Pourcentages cumulés de variance correspondant aux trois premiers axes d'une analyse sur les données originales et lissées.
*Cumulated variance percentages corresponding to the first three axes of analysis of original and smoothed data.*

| | Données | Tendances |
|---|---|---|
| Axe 1 | 41,20 | 44,18 |
| Axe 2 | 63,07 | 72,51 |
| Axe 3 | 80,20 | 88,61 |

Afin de condenser au maximum l'information, on peut résumer chaque groupe de descripteurs visualisé dans l'espace des axes principaux par une variable modèle, le « pattern » du groupe, défini par le cumul des valeurs de chaque descripteur normées à 100, pour leur donner une égale importance : nous avons en effet remarqué que les tendances des descripteurs d'un même groupe sont extrêmement voisines. La suite de l'interprétation, au lieu de se faire classiquement à partir des composantes, qui ont le désavantage d'être chacune une combinaison des mêmes descripteurs, peut s'effectuer sur ces variables modèles.

Puisque l'ordination permet de reconnaître les observations à partir desquelles s'inversent les gradients, on reportera ces coupures sur les profils de chaque variable-modèle, visualisant ainsi les périodes et la nature des changements de l'écosystème. Cette structure pourra être facilement confrontée avec l'évolution des facteurs externes hydrologiques ou climatologiques.

*La segmentation des séries résiduelles*

Après avoir soustrait les tendances des données originales, on obtient des séries résiduelles de moyenne nulle et proches de la stationnarité. Cependant, il serait incorrect d'effectuer une analyse spectrale sur ces séries (pour mettre en évidence des périodicités), car la variance du processus n'est pas forcément constante au cours du temps. On doit donc préalablement définir des intervalles de temps de variance homogène.

Hawkins et Merriam (1973 et 1974) ont élaboré une technique basée sur la programmation dynamique, pour définir un découpage optimal d'une série uni- ou multivariable. Le principe est de trouver $k$ segments (nombre fixé à l'avance), tels que les valeurs des observations à l'intérieur de chacun d'eux soient les plus voisines possible. L'homogénéité d'un segment allant de $i$ à $j$, sera défini par la somme des carrés des écarts de ses valeurs par rapport à leur moyenne, soit $r(i, j)$.

Si la série est partagée en $k$ segments de 1 à $n_1$, $n_1+1$ à $n_2$, $n_{(k-1)}$ à N, la validité de la partition sera donnée par :

$$W = r(1, n_1) + r(n_1+1, n_2) + \ldots + r(n_{(k-1)}, N).$$

Le problème consiste à trouver $n_1, n_2, \ldots, n_{(k-1)}$ tels que W soit minimum. Un procédé récurrent dérivé de la programmation dynamique permet de résoudre ce problème : si nous connaissons $(k-1)$ segments optimaux concernant les observations de 1 à $m$, pour toutes les valeurs de $m$ jusqu'à N, il est possible de trouver les $k$ segments cherchés. Comme le segment correspondant à $k=1$ est unique, on peut en déduire celui qui correspond à $k=2$ et ainsi de suite. Soit $F(m)$ la valeur de W, lorsqu'un segment optimal allant de 1 à $m$ est trouvé, on a :

$$F(m)=r(1, m), \qquad m=1, 2, \ldots, N,$$

$$F_j(m)=\min[F_{j-1}(n)+r(n+1, m)],$$

$$1<n<m.$$

En calculant tous les $F(m)$ avec $m=1, 2 \ldots N$ et $j=1, 2, \ldots, k$, on peut en déduire W correspondant à $F(n)$. Les limites des segments optimaux seront trouvées par une procédure récursive.

La solution pour le cas multivarié est une extension matricielle de cet algorithme.

Soient $p$ descripteurs observés en $n$ points : les $p$ composantes vectorielles $X_1, X_2 \ldots X_n$ sont les valeurs à chaque station des descripteurs. Si on pose par hypothèse que chaque $X_j$ suit une loi multinormale, on a :

$$X_j \sim \mathcal{N}(\zeta_j, \Sigma);$$

et s'il existe $h$ segments homogènes, alors :

$$\zeta_j=\mu_i, \qquad \theta_{i-1}<j\leqq\theta_i,$$

où $\mu_i$ correspond au vecteur des moyennes dans le $n$-ième segment, avec $\theta_{i-1}$ et $\theta_i$ étant respectivement les limites droite et gauche du $n$-ième segment (avec $\theta_0=0$ et $\theta_k=n$). $m_i=\theta_i-\theta_{i-1}$ désigne le nombre de points contenus dans le $n$-ième segment.

Si on considère que $\mu_i$ est échantillonné au hasard au sein d'une population multinormale :

$$\mu_i \sim \mathcal{N}(\mu, \Gamma),$$

en posant :

$$\bar{X}=\sum_{j=1}^{n} X_j/n,$$

et la matrice :

$$S=\sum_{j=1}^{n}(X_j-\bar{X})(X_j-\bar{X})'.$$

on démontre que l'espérance de S, si $n$ est suffisamment grand, est approximativement égale à :

$$n(\Sigma+\Gamma).$$

Calculons la matrice :

$$S_1=\frac{1}{2}\sum_{j=2}^{n}(X_j-X_{j-1})(X_j-X_{j-1})'$$

(matrice des carrés moyens des différences successives). On démontre que son espérance est proche de $n\Sigma$.

S et $S_1$ vont nous fournir une indication sur la magnitude de $\Gamma$ en considérant les valeurs propres de $SS_1^{-1}$.

Si l'une d'entre elles est nettement supérieure à 1, cela signifie que plus d'un segment est présent. Les $p$ valeurs propres $\lambda_i$, correspondant aux vecteurs propres $a_i$, satisfont l'équation :

$$(S-\lambda_i S_1)\, a_i=0.$$

Si on norme les $a_i$ de telle sorte que : $n^{-1} a_i S_1 a_i'=1$, $n^{-1} a_i S a_i'=\lambda_i$, alors les $a_i$ peuvent être réunis pour former une matrice $\mathcal{A}$ d'ordre $pxp$, et on peut définir une transformation des données $X_j$ par :

$$Y_j=AX_j.$$

Les éléments de $Y_j$ sont sans corrélation et ont pour variances $\lambda_1, \lambda_2 \ldots \lambda_p$; les valeurs des différences successives $Y_j-Y_{j-1}$ sont également sans corrélation et ont pour variance l'unité $(I_p)$.

On pourra réduire les dimensions des données transformées en $Y_j$ ne conservant que celles ayant des $\lambda_i>1$. Les composantes de $Y_i$ maximisent le rapport variance inter-segment/variance intra-segment.

Dans la mesure où on a décelé la présence de plus d'un segment, les limites pourront être estimées par la méthode du maximum de vraissemblance. Si $\Sigma$ est connu, les $\theta_i$ doivent minimiser :

$$\sum_{i=1}^{k}\sum_{j=\theta_{i-1}+1}^{\theta_i}(X_j-\bar{X}_i)'\Sigma^{-1}(X_j-\bar{X}_i),$$

où :

$$\bar{X}_i=\sum_{j=\theta_{i-1}+1}^{\theta_i} X_j/m_i.$$

Si $\Sigma$ n'est pas connu, on en prendra une estimation sous la forme de $S_1/(n-1)$. Nous avons alors à minimiser :

$$D=\sum_{i=1}^{k}\sum_{j=\theta_{i-1}+1}^{\theta_i}(X_j-\bar{X}_i)'S_1^{-1}(X_j-\bar{X}_i).$$

En remplaçant les $X_j$ par les $Y_j$, on ne changera pas la valeur de la distance euclidienne D. Comme la matrice $S_1$ des $Y_j$ est la matrice unité $I_p$, on obtient :

$$D=\sum_{i=1}^{k}\sum_{j=\theta_{i-1}+1}^{\theta_i}(Y_j-\bar{Y}_i)'(Y_j-\bar{Y}_i).$$

La localisation des $\theta_i$ peut alors être trouvée par l'algorithme de programmation dynamique suivant : soit $F_k(r)$ la valeur minimale possible de D, que l'on peut trouver avec $k$ segments; alors les $F_k(r)$ seront définis par récurrence par :

$$F_1(r)=\sum_{j=1}^{r}(Y_j-\bar{Y}_1)'(Y_j-\bar{Y}_1),$$

$$F_k(r)=\min_{1<s<n}\left[F_{k-1}(s)+\sum_{j=s+1}^{r}(Y_j-\bar{Y}_k)'(Y_j-\bar{Y}_k)\right].$$

La valeur minimale de D sur le transect entier est donnée par $F_k(n)$, et les limites des segments peuvent être trouvées en revenant en arrière (backtracing).

Nous avons programmé les méthodes uni- et multivariables de Hawkins et Merriam sur un microordinateur Hewlett Packard, 16 bits (modèle 16). Le temps calcul varie de la façon suivante :

| Opérations | Temps proportionnels à |
|---|---|
| Calcul de S et $S_1$ | $np^2$ |
| Diagonalisation de $SS_1^{-1}$ | $p^3$ |
| Localisation des limites $\theta_i$ | $n^2(p+k)$ |

Ces contraintes limitent l'emploi de cette technique pour des séries contenant plusieurs milliers d'observations si on ne peut accéder à un ordinateur très puissant.

Une solution approximative consiste à garder uniquement le premier axe $Y_1$, et on peut utiliser alors l'algorithme du cas univarié. Cette technique implique que la valeur de $k$ soit connue à l'avance. Bien que les tests statistiques soient inapplicables, on peut sélectionner la valeur de $k$ telle que le gain de la contribution pour W ou D d'une partition supplémentaire soit négligeable.

Dans la pratique, nous devrons nous fixer également un nombre $w$, tel que les segments trouvés contiennent au moins $w$ observations. La fonction Auto D2 nous indiquera clairement jusqu'à quel décalage la série reste en moyenne identique à elle-même. Dans le cas multivarié, $w$ doit être choisi très grand si on veut raccourcir le temps calcul très important (*voir* tableau précédent).

La technique de Merriam et Hawkins est très efficace pour la recherche des oscillations internes d'une série univariée. Mais l'extension aux séries multiples peut s'envisager directement sur les données originales si on désire distinguer des périodes pour lesquelles la variabilité reste globalement la même.

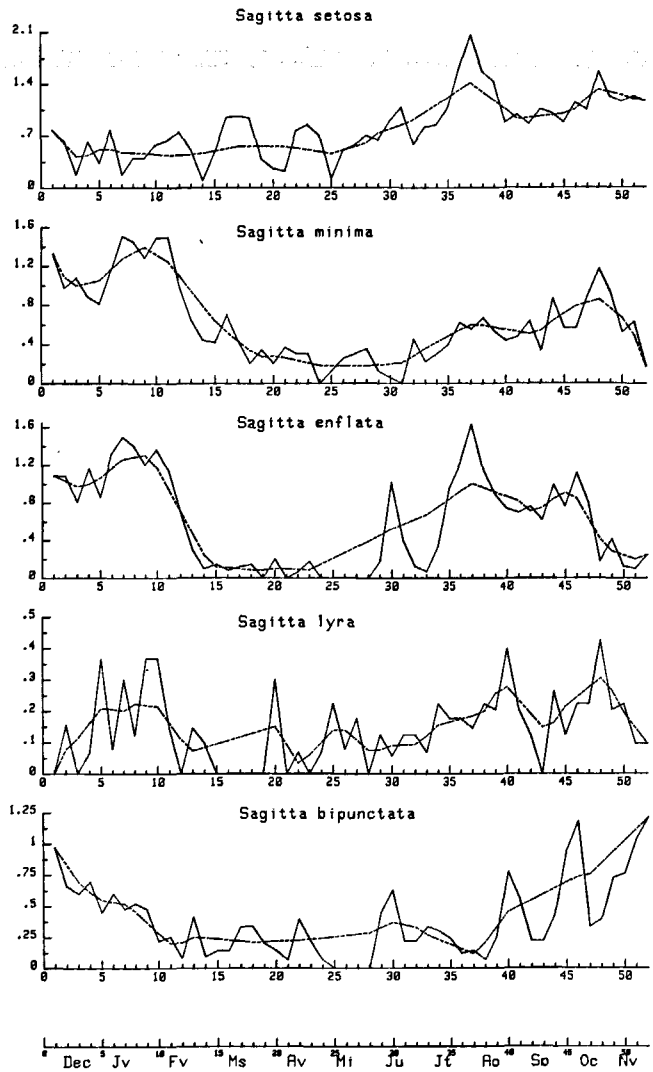


Figure 2

Cycles annuels de 5 jeunes stades de chaetognathes. Les données originales sont en logarithmes et correspondent à 52 moyennes hebdomadaires. La tendance générale est figurée en pointillés.

*Annual cycles for 5 juvenile chaetognath populations. Original data are logarithmic and correspond to 52 weekly averages. Dotted line indicates the general trend.*

## APPLICATION : SEGMENTATION D'UN CYCLE ANNUEL D'ABONDANCE DES CHAETOGNATHES PLANCTONIQUES DE LA BAIE DE VILLEFRANCHE- SUR-MER

Les données correspondent aux comptages quotidiens de 5 catégories de chaetognathes. Pour régulariser le pas d'échantillonnage, on a considéré les 52 moyennes hebdomadaires. Comme l'interprétation écologique de 13 catégories a déjà été réalisée (Ibanez, Dallot, 1969), et afin d'éviter un trop grand nombre de graphiques, nous avons retenu ici 5 descripteurs relativement peu corrélés : *Sagitta setosa, S. minima, S. enflata, S. lyra, S. bipunctata*, aux stades 1 de l'espèce. La figure 2 montre les profils d'abondances en logarithmes et les tendances générales obtenues par la méthode des points médians.

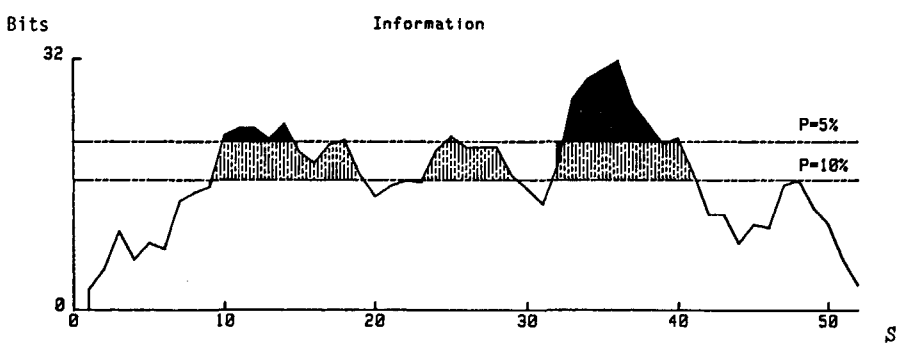


Figure 3

Évolution de la quantité d'information globale relative aux cinq espèces de chaetognathes. Les probabilités 5 et 10% correspondent à des quantités d'information moyenne par espèce de 4,3 et 3,32 bits.

*Global information development for five chaetognath species. 5 and 10% probabilities correspond to 4.3 and 3.32 bit averages for information quantity.*

### La quantité d'information liée aux cycles annuels

Le cumul des quantités d'information des différentes espèces pour chaque station, permet de définir l'évolution globale de l'écosystème (fig. 3). Nous avons tracé les seuils correspondant aux probabilités 5% (4,3 × 5 bits), et 10% (3,32 × 5 bits) d'apparition des évènements au hasard.

L'information est élevée fin janvier et au mois de février d'une part, et pendant les mois d'été d'autre part. Ces périodes sont marquées par une structuration durable des populations, extinction progressive en hiver ou prolifération croissante en été. Les autres saisons semblent marquées par des événements beaucoup plus sporadiques ou aléatoires.

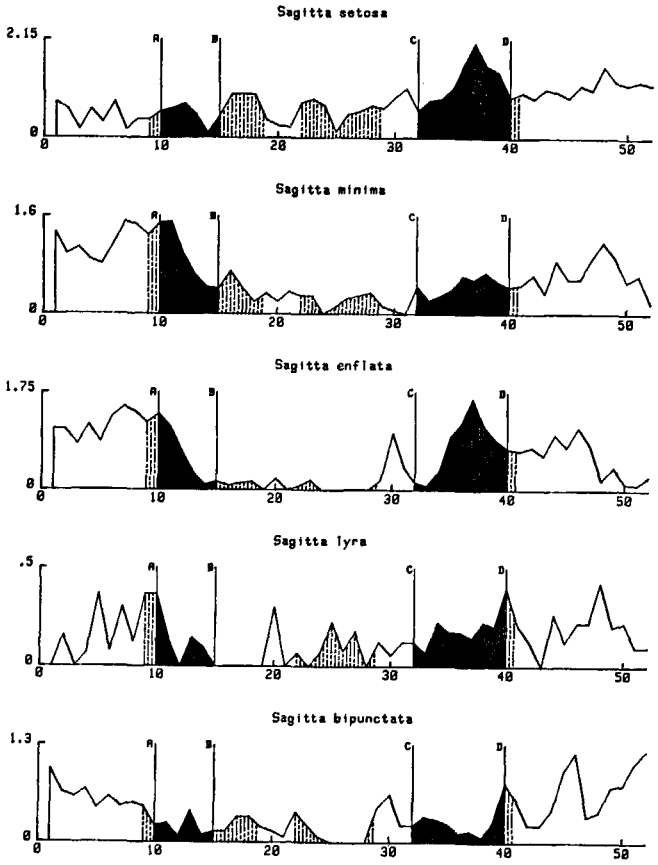


Figure 4
Localisation sur les profils d'abondance, des périodes du cycle annuel correspondant à des quantités d'information supérieures à 4,3 bits (hachures) et 3,32 bits (tirets).
*Location on abundance curves of annual cycle periods corresponding to information quantity in excess of 4.3 bits (hachured) and 3.32 bits (broken lines).*

La figure 4 montre le découpage des courbes d'abondances par la quantité d'information. Les très fortes valeurs de l'information en hiver coïncident avec la disparition en automne d'au moins 3 catégories (*S. minima, S. enflata, S. lyra*), provoquée par la raréfaction de la nourriture disponible. En été, l'information très forte indique la prolifération des juvéniles de *S. setosa, S. enflata et S. lyra*, favorisée par la stabilité verticale des eaux et la richesse trophique.

La période intermédiaire (fin de l'hiver et printemps) présente une information non négligeable; c'est l'indice d'une tendance à une structuration des peuplements, mais non de façon synchrone : le développement de certaines populations pendant quelques semaines succède à leur absence plus ou moins longue. Cette période correspond à une forte hétérogénéité hydrologique, mélanges verticaux, réchauffement progressif des eaux superficielles, début de la production phytoplanctonique.

L'automne et l'hiver n'ont pas d'information significative. Pourtant les chaetognathes ne sont ni absents, ni peu nombreux, mais les fluctuations des jeunes stades sont assez imprévisibles. La remontée de la thermocline d'août à novembre crée des cycloses verticales rendant plus aléatoire la nourriture disponible; des pluies intermittentes peuvent amener une multiplication rapide de certaines catégories; des courants vers la côte peuvent entraîner dans la rade des espèces du large, *etc.* Ces conditions instables se traduisent par des enregistrements assez irréguliers, où chaque événement doit être confronté avec les impulsions ponctuelles climatiques ou hydrologiques.

### La détection des hétérogénéités ponctuelles par le D2 au centre

Le choix préalable d'une fenêtre optimale a été réalisée à partir de la fonction Auto D2 (fig. 5). Comme la distance décroît après un décalage de 18 semaines, nous prendrons une fenêtre de 18 observations pour calculer le D2 au centre. La figure 6 montre les observations pour lesquelles les valeurs du D2 sont significatives à 5%. Certaines de ces coupures (30, 37, 40, 48, en gros traits) coïncident avec celles obtenues à partir des tendances (*voir* plus loin), alors que d'autres en pointillés restent isolées (35, 36, 45). Le sychronisme éventuel de ces deux segmentations n'est pas surprenant : si des inversions de gradients modifient à la fois rapidement l'amplitude des effectifs et les corrélations entre les descripteurs, les deux techniques vont se rejoindre.

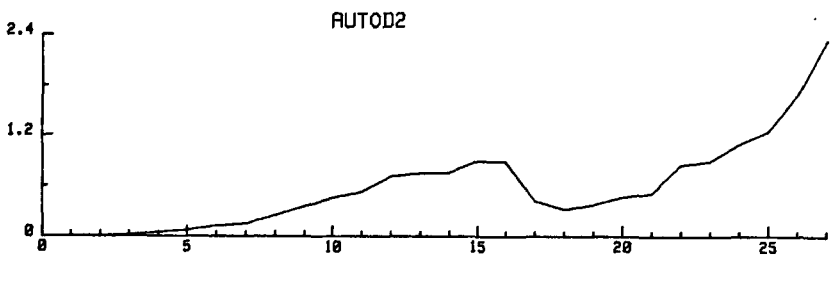


Figure 5
Variation de la fonction Autó D2 jusqu'à un décalage de 26 semaines.
*Variation of the function Auto D2 over a 26-week phase.*

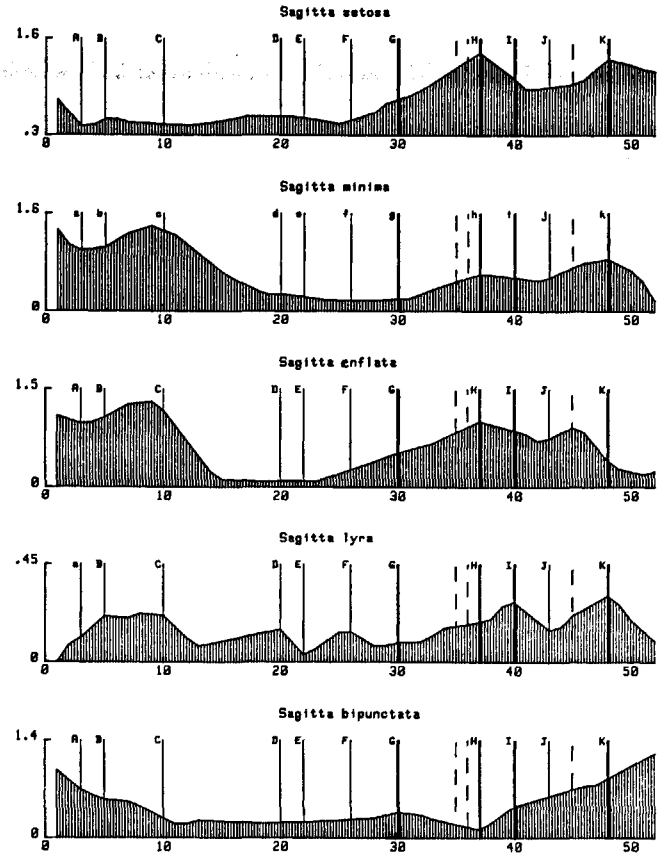


Figure 6

Définition sur les profils des tendances, des discontinuités par la méthode du D 2 au centre et l'analyse en composantes principales sur les tendances générales. Les coupures décelées par le D 2 qui coïncident avec celles de l'analyse d'inertie sont en traits renforcés : G, H, I, K. Les autres sont en tirets. Les traits verticaux surmontés d'une lettre sont tous définis à partir de l'ordination des observations par l'analyse en composantes principales.

*Trend curves showing discontinuities established by the D 2 to the centre method, and principal component analysis for general trends. Breaks identified by D 2 coinciding with breaks identified by inertia analysis are shown in firm characters: G, H, I, K; other discontinuities in broken lines. Vertical dashes with superposed letters are all based on principal component analysis.*

Dans le détail, si on se reporte aux courbes de données originales, chaque observation sélectionnée par le D 2 correspond effectivement à l'apparition relativement soudaine d'une forte biomasse pour quelques catégories, corrélée à une chute notable des autres populations. Le tableau 2 indique le sens des gradients après les coupures du D 2.

Ainsi le D 2 indique clairement les instants où les discontinuités entre les espèces sont les plus accentuées : ce sont les périodes « aberrantes » de l'évolution saisonnière. Ces modifications de structure à petite échelle de temps se situent principalement en été et au début de l'automne. L'interprétation de ces phénomènes doit être recherchée dans les modifications biologiques du peuplement : rythmes de reproduction, influence de la pression des prédateurs, etc.

L'index n'a pas retenu de discontinuités significatives avant la période estivale; en effet les inversions de gradients ont souvent été progressives, affectant dans le même sens la plupart des espèces, ou les fluctuations de biomasse sont restées trop minimes.

**La partition sur les tendances par l'analyse en composantes principales**

L'ordination des observations (fig. 7), selon les axes 1, 2 et 3 (après le lissage) permet de discerner les inver-

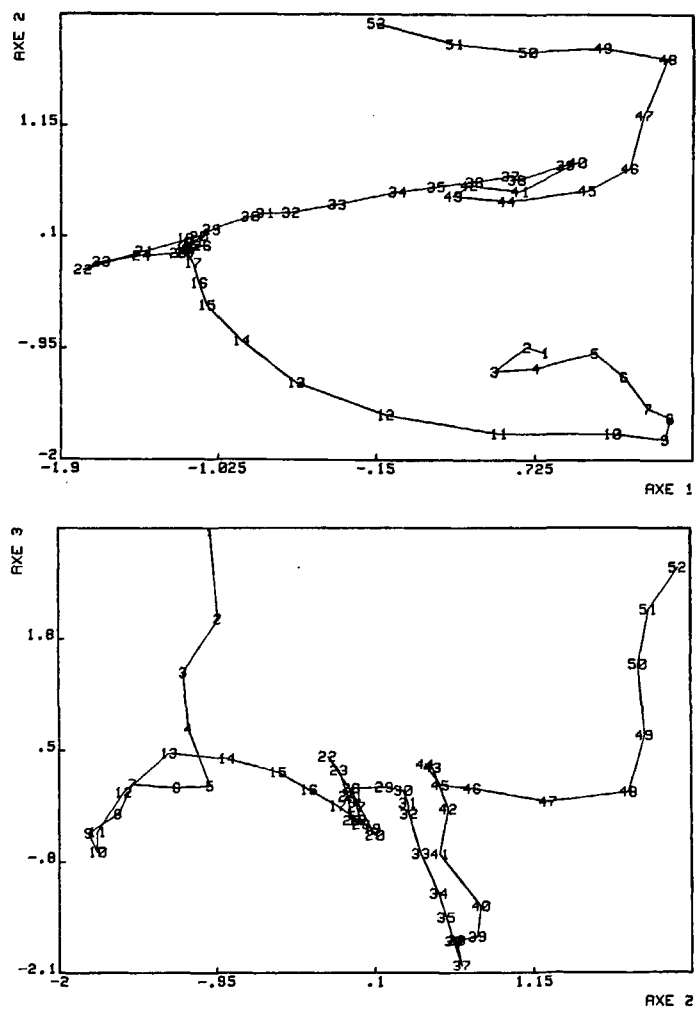


Figure 7

*a*) Ordination des 52 observations dans le plan des axes principaux 1 et 2 extraits d'une analyse sur les tendances.
*b*) Ordination des 52 observations dans le plan des axes principaux 2 et 3.

*a*) *Ordination of 52 observations on principal axes 1 and 2, based on principal component analysis.*
*b*) *Ordination of 52 observations on principal axes 2 and 3.*

Tableau 2

Nouvelle orientation des gradients à partir des observations sélectionnées par le D 2 au centre.

*New gradients based on observations selected by the D 2 to the centre method.*

| Observations | *S. setosa* | *S. minima* | *S. enflata* | *S. lyra* | *S. bipunctata* |
|---|---|---|---|---|---|
| 30 (G) | + | | + | | − |
| 35, 36, 37 (H) | − | − | − | + | + |
| 40 (I) | − | − | − | − | + |
| 45 | + | + | − | + | + |
| 48 | − | − | − | − | + |

Tableau 3
Sens des gradients dans les intervalles reconnus par l'analyse en composantes principales.
*Gradient direction in intervals identified by principal component analysis.*

| Gradients | *S. setosa* | *S. minima* | *S. enflata* | *S. lyra* | *S. bipunctata* |
|---|---|---|---|---|---|
| C : 1-10 | − | + | + | + | − |
| C-E : 10-22 | | − | − | − | |
| E-H : 22-37 | + | + | + | + | − |
| H-J : 37-40 | − | | − | + | + |
| I-J : 40-43 | | | − | − | + |
| J-K : 43-48 | + | + | − | + | + |
| K : 48-52 | − | − | − | − | + |

sions de tendances : observations numéros 3, 5, 10, 20, 22, 26, 30, 37, 40, 43, 48.

En reportant ces coupures sur les courbes d'abondances lissées (fig. 6), on reconnaîtra la nature des gradients exprimés. Si on examine de près les ordinations, on s'aperçoit qu'il est possible de négliger certains points d'inflexion mineurs. On peut également se donner un nombre minimal d'observations par segment. En considérant des intervalles d'au moins 5 observations, on obtient les coupures suivantes : de 1 à 10, de 10 à 22, de 22 à 37, de 37 à 40, de 40 à 48 et de 48 à 52.

Le tableau 3 indique le sens de ces gradients pour chaque descripteur dans ces intervalles.

## La segmentation des séries résiduelles

On obtient les séries résiduelles en soustrayant les tendances des séries originales (fig. 8). Si des écarts présentent des amplitudes positives ou négatives relativement constantes, cela signifie que les séries de départ peuvent se décomposer en un mouvement lent, la tendance générale, et une juxtaposition de phénomènes cycliques de haute fréquence ou d'une variabilité purement aléatoire.

L'examen des résidus et des fonctions d'autocorrélation pour les 3 catégories *S. setosa*, *S. minima* et *S. bipunctata*, indique la présence permanente de phénomènes périodiques (fig. 8). Grâce à l'analyse spectrale dite du maximum d'entropie (Burg, 1967), spécialement adaptée au traitement des séries courtes, on peut voir pour ces catégories qu'un cycle de fréquence 0,15 à 0,2 se superpose à la tendance générale annuelle. Ce cycle, dont la période est de 5 à 6 semaines, semble en accord avec la durée probable liée au passage d'une génération à une autre.

*S. enflata* et *S. lyra* présentent au contraire des accidents qui altèrent la régularité des séries résiduelles et des autocorrélations. Pour *S. enflata*, on remarquera entre les mois d'avril et août (intervalle non hachuré : fig. 8), une tendance locale avec deux pics bien marqués, qui contraste avec les autres observations. D'après la courbe originale, ces pics correspondent à de fortes biomasses étalées au plus sur 2 ou 3 semaines, et alternant avec des périodes d'absence de l'espèce. Ce phénomène semble donc lié à un événement indépendant de l'évolution en place de la population; c'est vraisemblablement une arrivée massive de *S. enflata* à partir d'eau du large qui en est responsable.

Pour *S. lyra*, on note une tendance décroissante entre les observations 15 et 20, qui correspond à la disparition brutale de cette espèce.

C'est précisément lorsque l'on est en présence de séries irrégulières de ce type, que l'on doit obligatoirement avoir recours à une segmentation préalable à l'étude des cycles. Notre exemple est ici tout à fait évident,

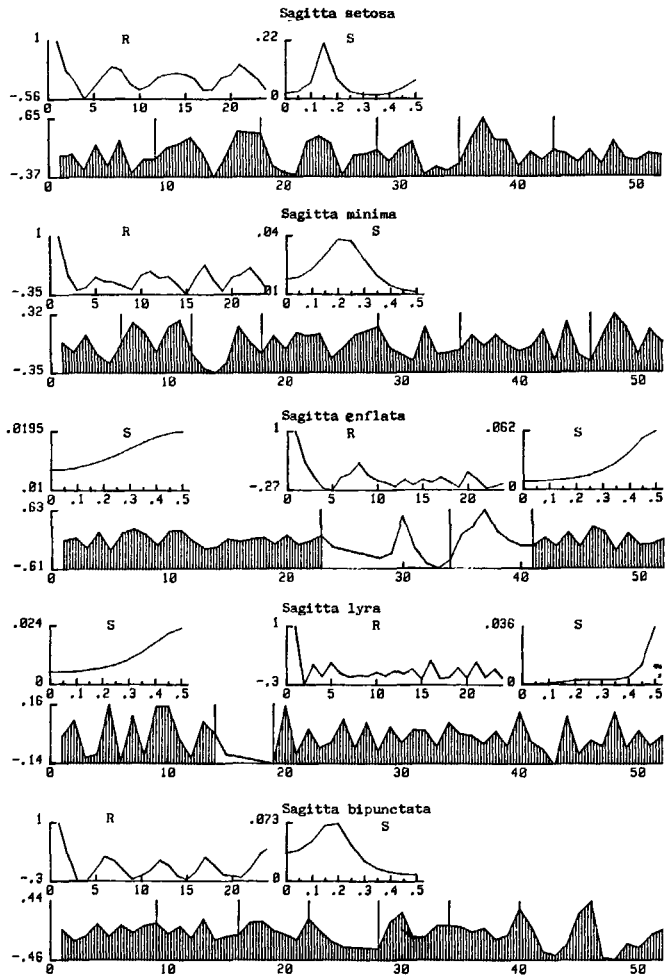


Figure 8
Les courbes présentant des hachures correspondent aux 5 séries résiduelles après élimination de la tendance générale. Les traits verticaux découpent les séries résiduelles en blocs de dispersion homogène par la méthode de Hawkins et Merriam. Les intervalles non hachurés dans les séries résiduelles correspondent à des zones impropres à l'analyse spectrale. La fonction d'autocorrélation marquée d'un R est figurée au dessus de chaque séries ainsi que les spectres marqués d'un S.

*Hachured curves correspond to 5 residual series after elimination of the general trend. Vertical lines seperate the residual series into homogeneous blocks by the Hawkins and Merriam method. Non-hachured intervals in the residual series correspond to zones unsuitable for spectral analysis. Autocorrelation function R is shown above each series, together with spectra S.*

Tableau 4
Variation des limites des partitions en fonction de leur nombre.
*Partition limit variation as a function of number.*

| | Nombre de partitions K | | | |
|---|---|---|---|---|
| | K=2 | K=3 | K=4 | K=5 |
| Pourcentages de variance expliquée | 7,15 | 27,74 | 30,39 | 31,74 |
| Numéros des coupures | 34 | 34 | 41 | 41 |
| | – | 23 | 34 | 34 |
| | – | – | 23 | 23 |
| | – | – | – | 11 |

mais la technique formelle de Merriam et Hawkins est essentielle lorsque les séries sont très longues. En fonction du nombre de partitions, pour *S. enflata* nous obtenons les coupures suivantes (tab. 4).

Comme le gain de variance expliqué est faible à partir de K=5, nous retiendrons les 3 coupures 23, 34, 41 qui correspondent bien à l'hétérogénéité soulignée précédemment (fig. 8).

L'analyse spectrale pour les deux sous-ensembles de 1 à 23, et de 41 à 52, montrent des phénomènes de haute fréquence voisins de la fréquence même de l'échantillonnage, ce qui revient à dire que notre échelle d'observation est sans doute trop grande pour reconnaître des phénomènes cycliques éventuels propres à ce descripteur. Par la même méthode, nous trouvons pour *S. lyra* des coupures aux stations 14 et 19, et on retrouve les mêmes cycles courts dans les intervalles homogènes de la série résiduelle.

Dans le cas précis de cette catégorie, il nous semble que, compte tenu de la taille très faible de ces organismes, un filet à maille plus serrée serait plus adéquat à la capture.

Nous avons également sur la figure 8, les coupures pour les autres séries qui présentent un cycle bien défini. L'intervalle de temps entre deux limites apparaît relativement constant et voisin de la longueur d'onde du phénomène périodique sous-jacent.

Pour illustrer la technique de Hawkins et Merriam dans le cas multivarié, nous avons considéré globalement les 5 séries résiduelles. L'analyse discriminante nous fournit 3 vecteurs canoniques associés à des valeurs propres supérieures à 1, et finalement, tenant compte de ces trois dimensions, on obtient les limites de partition suivantes : 15, 24, 34, 42. Cette classification correspond en fait aux trois étapes fondamentales du cycle annuel des stades juvéniles de chaetognathes dans la baie de Villefranche-sur-Mer :

— automne-hiver (42 à 52 et 1 à 15) : fluctuations automnales rapides, rythme de haute fréquence. Cette instabilité est suivie d'une phase de décroissance progressive de la biomasse;

— début printemps (15-24). Absence ou rareté des jeunes stades. Pas de structurations des peuplements;

— fin printemps-été (24-34; 34-42). La fin du printemps est la phase de reprise plus ou moins visible de l'activité reproductrice. L'été correspond à une époque de prolifération pour la majorité des chaetognathes. Ce développement est favorisé à la fois par la stabilité verticale des masses d'eau et par l'abondance de nourriture.

## CONCLUSION

Les propriétés et les interrelations des processus aléatoires en océanographie sont instables dans le temps et dans l'espace. Quelle que soit l'échelle des expériences, les structures physiques et biologiques se modifient de façon soudaine ou progressive; la description de ce phénomène dynamique est fondamentale pour l'interprétation en écologie pélagique.

Dans les séries pluriannuelles, se succèdent des phases de latence, de croissance puis de mortalité des organismes planctoniques. Mais ce modèle est rarement suffisant pour rendre compte de l'évolution temporelle : les changements climatiques d'une année sur l'autre vont affecter à la fois la biomasse et la variation saisonnière, et de plus le cycle annuel sera masqué sporadiquement par le déplacement des populations.

Un échantillonnage en continu sur une radiale va faire ressortir l'alternance de populations de composition faunistique différente, mais également aux propriétés statistiques hétérogènes.

De nombreux auteurs (Platt, Denman, 1975; Fasham, 1976; Steele, Henderson, 1979; Laurec, 1979) ont montré par l'analyse spectrale au-dessous de quelle fréquence critique les facteurs biologiques contrôlaient la distribution du plancton, se substituant aux phénomènes physiques de la turbulence. Avec la partition dans le domaine spatial et/ou temporel, il ne s'agit plus de définir une échelle moyenne, mais de localiser des zones qui correspondent à des états particuliers de l'écosystème pélagique.

Nous avons présenté des méthodes de segmentation pour des séries multivariables, car les descripteurs hydrologiques ou biologiques ne varient pas de façon indépendante dans le temps et dans l'espace : les masses d'eau se succèdent avec des variations d'amplitude des paramètres et de leur corrélations : sur une radiale pour des groupes d'espèces planctoniques, on peut définir une référence spatiale suivant les critères hydrologiques, à laquelle se superpose une fluctuation temporelle traduisant un comportement migratoire ou un transport passif du plancton.

En raison de la complexité des structures, des changements dans la nature même des processus sous-jacents, la segmentation doit s'opérer en fonction de différents critères, liés soit à des impératifs statistiques, soit à la vérification de certaines hypothèses écologiques.

L'estimation de la quantité d'information liée à un signal multivariable permet de reconnaître les événements significatifs au sens statistique, ceux qui traduisent un certain niveau de structuration. Des quantités d'informations faibles marquent la réalisation de phénomènes purement aléatoires ou l'inefficacité de l'échantillonnage. Les discontinuités plus ou moins ponctuelles comme les fronts hydrologiques, la prolifération subite d'une espèce, seront détectées par la distance au centre, au besoin en temps réel. Des phénomènes macroscopiques comme les gradients principaux des populations peuvent être mis en évidence par l'analyse en composantes principales sur les tendances générales estimées par la technique des points médians. La partition en blocs

homogènes due à Merriam et Hawkins, est un outil fiable si l'on veut reconnaître les limites de structures homogènes destinées à des analyses quantitatives ultérieures.

Certes, ces quatre méthodes ne constituent pas l'ensemble des partitions possibles. On pourrait, par exemple, tout en gardant la démarche récursive de Merriam et Hawkins, modifier le critère de classification : maximiser les corrélations entre les descripteurs, maximiser la diversité, *etc.* Mais nous avons voulu souligner ici la nécessité des segmentations, tant pour l'analyse mathématique des processus que pour définir les bases de l'interprétation. La segmentation n'est pas une technique unique; c'est en fonction des objectifs fixés par sa planification que l'océanographe devra élaborer des modèles de partition spécifiques.

## RÉFÉRENCES


**Bemont T. R., Waterman M. S.,** 1977. Locating maximum variance segments in sequential data, *J. Inter. Assoc. Math. Geol.*, **9,** 55-61.

**Béthoux N., Étienne M., Ibanez F., Rapaire J. L.,** 1980. Spécificités hydrologiques des zones littorales. Analyse chronologique par la méthode Census II et estimation des échanges océan-atmosphère appliquées à la baie de Villefranche-sur-Mer, *Ann. Inst. Océanogr.*, **56,** 81-95.

**Binet D., Gaborit M., Dessier A., Roux M.,** 1972. Premières données sur les copépodes pélagiques de la région congolaise. II. Analyse des correspondances, *Cah. ORSTOM, Ser. Oceanogr.*, **10,** 127-137.

**Burg J. P.,** 1967. Maximum entropy spectral analysis, Paper presented at the *37th annual international meeting, Soc. Explor. Geophys., Oklahana city, Oklahana, October* 31, 1967.

**Cassie M.,** 1958. Canonical discrimination analysis and the zooplankton cycle in lake Maggiore, 1957-1958, *Inst. Ital. Idrobiol.*, **25,** 33-40.

**Cassie M.,** 1967. Principal component analysis of the zooplankton of lake Maggiore, *Mem. Inst. Ital. Idrobiol.*, **21,** 129-144.

**Colebrook J. M.,** 1964. Continuous plankton records: a principal component analysis of the geographical distribution of plankton, *Bull. Mar. Ecol.*, **6,** 78-100.

**Cruzado A.,** 1971. Sequential statistical analysis of continuous underway oceanographical data, *Invest. Pesq.*, **35,** 261-267.

**Dessier A., Laurec A.,** 1978. Le cycle annuel du zooplancton à Pointe-Noire (Congo). Description mathématique, *Oceanol. Acta*, **1,** 3, 285-304.

**Fasham M. J. R.,** 1976. Observations on the horizontal coherence of chlorophylle *a* and temperature, *Deep-Sea Res.*, **23,** 527-583.

**Gilchrist W.,** 1976. *Statistical forecasting*, Wiley and Sons, N.Y., 308 p.

**Hawkins D. M., Merriam D. F.,** 1973. Optimal zonation of digitized sequential data, *Jour. Inter. Assoc. Math. Geol.*, **5,** 389-397.

**Hawkins D. M., Merriam D. F.,** 1974. Zonation of multivariate sequences of digitized geologic data, *Jour. Inter. Assoc. Math. Geol.*, **6,** 263-271.

**Ibanez F.,** 1976. Contribution à l'analyse mathématique des événements en écologie planctonique, *Bull. Inst. Océanogr. Monaco*, **72,** 1-96.

**Ibanez F.,** 1981. Immediate detection of heterogeneities in continuous oceanographic recordings. Application to time series analysis of changes in the bay of Villefranche-sur-Mer, *Limnol. Oceanogr.*, **26,** 336-349.

**Ibanez F.,** 1982. Sur une nouvelle application de la théorie de l'information à la description des séries chronologiques planctoniques, *J. Plankton Res.*, 619-632.

**Ibanez F.,** 1983. Optimisation de la représentation des séries chronologiques planctoniques multivariables, *Rapp. Comm. Int. Mer Médit.*, **28,** 113-115.

**Ibanez F., Dallot S.,** 1969. Étude du cycle annuel des chaetognathes planctoniques de la rade de Villefranche-sur-Mer par l'analyse en composantes principales, *Mar. Biol.*, **3,** 11-17.

**Ibanez F., Seguin G.,** 1972. Étude du cycle annuel du zooplancton d'Abidjan. Comparaison de plusieurs méthodes d'analyses multivariables, *Invest. Pesq.*, **36,** 81-108.

**Kelley J. C.,** 1972. A strategy for continuous multivariate analysis in oceanography, *Invest. Pesq.*, **36,** 175-178.

**Laurec A.,** 1979. Analyse des données et modèles prévisionnels en écologie marine, *Thèse Univ. Aix-Marseille*, 405 p.

**Legendre P., Dallot S., Legendre L.,** 1982. Description numérique de l'évolution d'un peuplement : le groupement chronologique (en prép.); communication Congrès Classification Society, Toronto, Juin 1981.

**Platt T., Denman K. L.,** 1975. Spectral analysis in ecology, *Ann. Rev. Ecol. System.*, **6,** 183-187.

**Shiskin J., Eisenpress H.,** 1957. Seasonal adjustment by electronic computer methods, *J. Am. Stat. Assoc.*, **52,** 415-449.

**Steele J. H., Henderson E. W.,** 1979. Spatial patterns in North Sea plankton, *Deep-Sea Res.*, **26,** 955-963.

**Webster R.,** 1973. Automatic soil-boundary location from transect data, *J. Inter. Assoc. Math. Geol.*, **5,** 27-37.