XXX^{èmes} journées de statistique 25-29 mai 1998, Rennes
Session : Biométrie et halieutique

UTILISATION DES MÉTHODES DE SEGMENTATION POUR LA CONSTRUCTION DE CARTES ILLUSTRANT LES RELATIONS RESSOURCES HALIEUTIQUES / ENVIRONNEMENT

S. SALAUN [1][2] et J. FERRARIS [1][3]
(1) Antenne ORSTOM, IFREMER, BP 21105, 44311 NANTES Cedex 03
(2) Adresse actuelle : IFREMER, Laboratoire ECOHAL, BP 21105, 44311 NANTES Cedex 03
(3) Adresse actuelle : ORSTOM, B.P A5, 98848 NOUMEA, Nouvelle-Calédonie

MOTS CLÉS : Segmentation, analyse de données, cartographie, ressources halieutiques, environnement

INTRODUCTION:

Les données géoreférencées collectées lors de campagnes océanographiques en vue d'étudier les relations entre les ressources biologiques et l'environnement peuvent être analysées en tenant compte de la dimension spatiale grâce notamment à l'utilisation d'un Système d'Information Géographique. Un tel logiciel permet de plus de restituer les résultats de ces analyses sous forme de cartes. Mais la construction de cartes implique de faire de nombreux choix quant aux variables à représenter et à leur mode de représentation sous forme de gradient (exprimant un continuum) ou sous forme discontinue par la construction de classes. Ces choix sont souvent plus ou moins arbitraires mais peuvent aussi être orientés par une analyse statistique préalable. Les données biologiques et environnementales étant souvent multidimensionnelles, toutes les variables ne sont pas visualisables telles quelles sur la même carte. Une solution pour réduire la dimension du problème consiste à construire des indicateurs synthétiques par exemple grâce à la réalisation d'analyses factorielles et/ou de méthodes de classification qui fournissent une représentation condensée des données (FERRARIS et PELLETIER, 1997).

Les méthodes de segmentation quant à elles fournissent des résultats beaucoup plus explicites pour l'interprétation de la structure et l'identification des facteurs discriminants ; et ceci particulièrement lorsque l'on souhaite expliquer une variable caractérisant la ressource par des variables environnementales.

L'objectif de cette étude est de montrer l'apport de telles méthodes pour la construction de cartes thématiques mettant en relation un type de données biotiques (biomasse, présence/absence d'une espèce, etc.) et une combinaison de facteurs environnementaux (caractérisant la sédimentologie, l'hydrologie, la bathymétrie).

Après une présentation **générale** des méthodes de segmentation et de cartographie **envisagées dans une telle problématique,** nous nous intéresserons à une application concernant

l'analyse des relations entre les poissons récoltés par campagnes de chalutages et leur environnement sur le plateau continental guinéen.

MÉTHODES :

- La segmentation :

→ Principe général :

Le principe de base des méthodes de segmentation est de partir de la population totale et de procéder à des dichotomies successives de celle-ci, de façon à obtenir au fur et à mesure du processus, des populations qui soient le plus homogènes possible vis à vis des classes d'une partition a priori (BREIMAN et al, 1984 ; PERINEL, 1996). Ces méthodes permettent ainsi de hiérarchiser les variables les plus discriminantes mais aussi de mettre en évidence des effets de seuil.

Ainsi, lorsque l'on est en présence d'un tableau de données contenant une variable privilégiée **y** à expliquer par les autres variables du tableau x_1, x_2,...., x_p, la méthode de segmentation consiste à rechercher d'abord la variable x_j qui explique le mieux la variable **y**. Cette variable définit la première division de l'échantillon en deux sous ensembles appelés segment. Puis on réitère cette procédure à l'intérieur de chacun des deux segments et ainsi de suite (LEBART et al, 1995). A partir de la population totale on construit ainsi un arbre de décision binaire (*figure 1*). Chaque chemin de l'arbre (de la racine à la feuille) est donc une combinaison de propriétés décrivant une sous population. Dans un arbre binaire complet, chaque segment contient un seul individu.

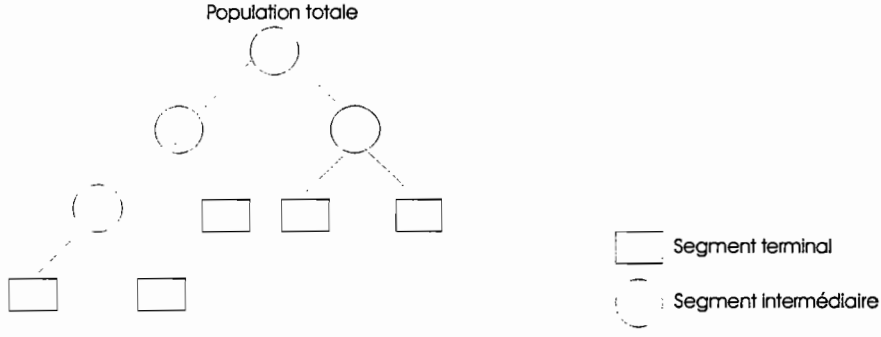

figure 1 : Arbre de décision binaire

Outre cet objectif descriptif, les techniques de segmentation fournissent une règle de décision utile si l'on cherche à prévoir la place d'un nouvel individu **i**. Ainsi, suivant les valeurs que **i** présente pour x_1, x_2, ..., x_p, il va parcourir tel ou tel chemin de l'arbre et tombera dans un des segments terminaux. Ce nouvel individu sera alors affecté à l'un des groupes constitués (dans le cas d'une variable nominale à expliquer) ou sera affecté d'une valeur y_i qui sera la moyenne dans le segment terminal (l'écart-type correspondra aussi à celui du segment).

La segmentation par arbre de décision binaire présente de nombreux avantages tels que la lisibilité des règles d'affectation et la facilité d'interprétation des résultats. De plus, elle

permet d'utiliser en même temps comme variables explicatives des variables continues et nominales et d'analyser soit une variable nominale, soit une variable continue.

Il faut néanmoins prendre garde à cette simplification des résultats et tenir compte, lors de l'interprétation des coupures, de la corrélation qui peut exister entre les variables discriminantes. La variable discriminante qui a entraîné la coupure de l'arbre peut être fortement corrélée avec une autre variable en fonction de laquelle il paraît plus pertinent d'interpréter une coupure. Afin de palier à ces inconvénients, la procédure offre la possibilité de décrire les divisions « equi-réductrices » qui correspondent aux meilleures divisions successives du segment après celle qui a provoqué la partition. De plus, l'existence de divisions « equi-divisantes » permet de classer un nouveau sujet présentant des données manquantes concernant les variables explicatives.

→ Critères de division des nœuds :

Lorsque la variable à expliquer **y** est continue, le critère de sélection de la meilleure division d'un nœud est tel que la variance résiduelle du nœud, qui correspond à la moyenne pondérée des variances de **y** à l'intérieur de chacun de ses segments descendants, soit minimale. Ainsi, pour toute division d_j^m d'un nœud t par une variable x_j, on calcule la moyenne pondérée des variances de y à l'intérieur de chacun de ses segments descendants t_g et t_d, c'est à dire la variance résiduelle du nœud t :

$$\text{var}(d_j^m, t) = (\frac{n_g}{n_t} s_g^2) + (\frac{n_d}{n_t} s_d^2)$$

où n_g, n_d, n_t sont respectivement les effectifs des segments t_g, t_d, t et s_g^2, s_d^2 , sont les variances de la variable continue y à l'intérieur des segments t_g et t_d. On retient la meilleure division d_j^* réalisée par la variable x_j qui correspond à la variance résiduelle minimale :

$$\text{var}(d_j^*, t) = \min[\text{var}(d_j^m, t)]$$

où *m* appartient à d_j et où d_j est l'ensemble des divisions de la variable x_j.
Parmi toutes les meilleures divisions d_j^* obtenues à partir des p variables explicatives, la meilleure division (globale) du nœud t est effectuée à l'aide de la variable qui assure :

$$\text{var}(d^*, t) = \min_{j=1,\dots p} [\text{var}(d_j^*, t)]$$

Lorsque la variable **y** est nominale et répartit les individus en **k** classes, il faut que le mélange des classes soit moins important dans les segments descendants que dans le nœud parent (critère de la pureté maximale). En effet, à chaque segment est associée une mesure de l'impureté i(t) définie par :

$$i(t) = \sum_r^k \sum_s^k P(r|t) P(s|t)$$

avec $r \neq s$ et où $P(r|t)$ et $P(s|t)$ sont les proportions d'individus dans les classes c_r et c_s dans le segment t. Un segment est pur s'il ne contient que des individus d'une seule classe,

dans un tel cas : i(t) = 0. Plus le mélange des classes dans le segment t est important, plus l'impureté i(t) est élevée.

Chaque division d_j^m du nœud t par la variable x_j entraîne une réduction de l'impureté qui s'exprime par :

$$\Delta_j^m = i(t) - p_g i(t_g) - p_d i(t_d)$$

où p_g et p_d sont les proportions d'individus du nœud t dans les segments descendants t_g et t_d.
Ainsi, pour chaque variable x_j, la meilleure division d_j^* est telle que la réduction de l'impureté Δ_j^* est maximale :

$$\Delta_j^* = \max[\Delta_j^m]$$

où m appartient à d_j, l'ensemble des divisions de la variable x_j.
Sur l'ensemble des p variables, la division du nœud t est effectuée à l'aide de la variable qui assure :

$$\Delta^* = \max_{j=1,\ldots,p} [\Delta_j^*]$$

(LEBART et al, 1995).

→ Construction du meilleur sous-arbre:

À partir de l'arbre binaire complet, une phase d'élagage permet de produire un sous arbre optimal exploitable en se fondant sur l'estimation de l'erreur d'affectation ou de prévision à l'aide, soit d'un échantillon test, soit de la validation croisée.

La méthode des échantillons tests, la plus couramment utilisée, consiste à effectuer la discrimination sur une partie seulement de la population initiale (80% souvent) et de tester les règles d'affectation sur les 20% restant.

Dans le cas de petits échantillons, il est conseillé d'utiliser la validation croisée car elle permet de prendre en compte tous les sujets de l'échantillon à la fois pour construire et pour tester l'arbre (GUEGUEN et NAKACHE, 1988). En effet, dans ce cas, l'échantillon est divisé en **m** parties égales ; la discrimination se fait sur l'échantillon de **m-1** parties et le taux d'erreur sur la partie restante, ce qui peut être fait de **m** façons différentes (GUEGUEN et al, 1996).

Les analyses ont été ici réalisées grâce aux procédures REGAR (dans le cas de **y** continue) et DISAR (dans le cas de **y** nominale) du logiciel SPAD.S (GUEGUEN et al, 1996).

- Réalisation des cartes :

Les cartes ont été réalisées par le biais d'un Système d'Information Géographique (logiciel SAVANE, © ORSTOM, version 5.02). Un système d'information géographique (SIG) est un système informatisé d'acquisition, de gestion, d'analyse et de représentation de données à référence spatiale. Il permet de croiser des données de natures diverses telles que les données d'enquêtes, des cartes thématiques, des données topographiques, des images satellites, des modèles numériques de terrain. Un SIG stocke les deux composantes de l'information décrites par une carte : la description des objets spatiaux (coordonnées géographiques en longitude et en latitude) et leurs caractéristiques thématiques (valeur de chaque objet pour telle caractéristique).

Les méthodes de segmentation permettent de sélectionner les variables environnementales les plus discriminantes vis à vis d'une variable biotique. Ce sont ces deux types de variables qu'il s'agit de représenter sur la même carte.

Afin de visualiser les variables environnementales dans l'espace, on utilise ici une interpolation déterministe permettant de connaître en tout point de l'espace considéré les valeurs prises par la variable à cartographier. La méthode d'interpolation utilisée estime la valeur en un point à partir des valeurs mesurées dans un certain nombre de stations voisines. On choisit le nombre de ces stations à retenir en fonction de la densité des stations sur la zone d'étude. Les résultats de ces interpolations sont des modèles numériques de terrain que l'on peut « découper » en un certain nombre de classes afin de les visualiser. Les valeurs de coupure entre ces classes sont celles qui ont provoqué les coupures binaires des arbres de segmentation correspondants.

On peut par la suite superposer plusieurs variables environnementales traitées de la sorte, ceci dans la limite du « visuellement acceptable ». Puis une dernière couche d'information, correspondant à la variable biologique expliquée par ces variables environnementales discriminantes, est ajoutée à la carte.

Il faut aussi noter que les résultats des interpolations des variables environnementales sont aussi utilisés dès la première phase de construction des tableaux de données qui vont être utilisés pour la segmentation. En effet, les données, lorsqu'elles ont été collectées à des moments différents peuvent différer tant au niveau du nombre de stations que de la situation géographique de ces stations. Une solution pour mettre en relation ces données d'origines différentes consiste alors à interpoler la variable environnementale par exemple ; et ceci surtout si les points de sondage auxquels elle correspond sont plus rapprochés que les stations de chalutage correspondant à la variable biologique car on minimise ainsi l'erreur due à l'interpolation. Puis on récupère, au niveau des stations de chalutages biologiques, l'information issue de l'interpolation. Ainsi, les deux types de données deviennent homogènes puisqu'elles concernent le même nombre de stations et la même position géographique.

APPLICATION:

A titre d'illustration, on présentera ici une application de ces méthodes de segmentation à l'analyse des relations entre les ressources démersales du plateau continental guinéen (voir *fig.2*) et leur environnement (SALAUN, 1997).

figure 2 : Position géographique de la Guinée sur le continent africain

Cette démarche a pour but de décrire la répartition des ressources démersales par l'intermédiaire de variables biologiques synthétiques en liaison avec les variables environnementales, et plus particulièrement la sédimentologie. Ce travail, réalisé à l'antenne ORSTOM de Nantes dans le cadre d'un stage de DEA, et fait suite au projet FAO « Systèmes d'information géographique appliqués aux pêcheries de l'Afrique de l'Ouest » (1993-96) qui couvrait le Sénégal, la Mauritanie, le Maroc et la Guinée.

La plupart des données proviennent de campagnes scientifiques de chalutages démersaux sur 150 stations (FONTANA et MORIZE, 1995), seules les données de sédimentologie ont été récoltées lors d'une autre campagne en 800 points de sondage (DOMAIN et BAH, 1993).

Les relations ressources/ environnement peuvent être décrites à différentes échelles de complexité biologique et on traite ici les exemples de présence/absence d'une espèce (variable nominale à deux modalités) puis de taille d'une espèce (variable continue). Les facteurs explicatifs entrant dans l'analyse sont la profondeur, les salinités et températures de fond et de surface, ainsi que les teneurs du sédiment en lutites, sables fins, sables grossiers, graviers et carbonates. Les deux espèces considérées ici : *Pagellus bellottii* (Pageot à taches rouges) et *Sparus caeruleostictus* (Pagre à points bleus). Elles appartiennent à la famille des Sparidae, présente sur l'ensemble de la zone et constituant la plus forte biomasse lors des campagnes.

→ Présence du Pagre à points bleus sur la zone d'étude :

La procédure de discrimination par arbre de décision binaire appliquée à la variable présence-absence du Pagre, au cours de la campagne correspondant à la saison sèche, aboutit à la création de trois segments terminaux intéressants (*fig.3*)

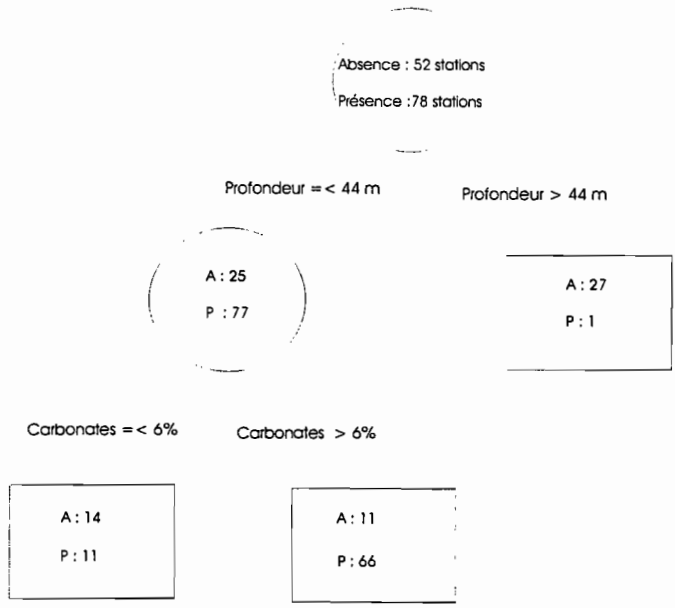

figure 3 : Résultats de la segmentation appliquée à « présence/absence du Pagre »

D'après ces résultats, lorsque le Pagre est présent, il s'agit à 85% de fonds de profondeur inférieure à 44 m et pour lesquels la teneur en carbonates est supérieure à 6%. Afin de représenter ceci sous forme de cartes, les variables environnementales apparaissant les plus discriminantes, c'est à dire la profondeur et la teneur en carbonates, ont été interpolées à l'ensemble de la zone puis scindées en deux classes selon les valeurs de coupure obtenues par la segmentation (44 m pour la profondeur et 6% pour la teneur en carbonates). Enfin, chaque segment terminal de ce sous-arbre correspond à une combinaison de variables environnementales que l'on peut alors représenter sous formes de plages de couleurs lors de la représentation cartographique (*fig.4*).

Le Pagre apparaît peu présent dans les zones les plus profondes. De plus, au niveau de la sédimentologie, le facteur expliquant le mieux sa répartition est la teneur en carbonates et non un type de granulométrie. Ceci peut être lié au fait que les poissons de la famille des Sparidae ne vivent pas directement sur le fond mais s'en rapprochent pour se nourrir. Cette relation avec les sédiments carbonatés trouverait alors une explication dans le régime alimentaire.

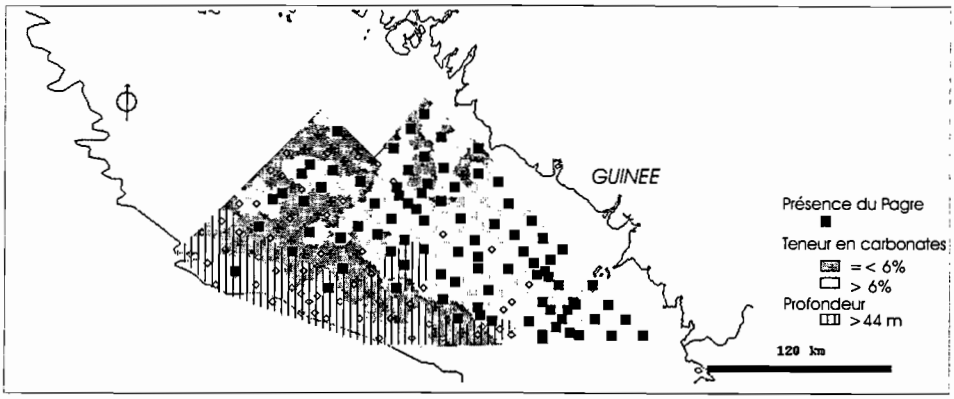

figure 4 : Présence du Pagre sur la zone d'étude en liaison avec l'environnement

→ Taille du Pageot à taches rouges sur la zone d'étude :

À une échelle biologique plus fine, on peut raisonner au niveau des tailles des individus. Le Pageot à taches rouges (Pagellus bellottii) est abondamment présent sur le plateau continental guinéen mais sa faible taille reste bien souvent un frein à sa commercialisation. On dispose des données de capture en poids et en nombre de poissons pour chaque espèce et grâce à une clé longueur/poids disponible (FONTANA et MORIZE, 1995), on peut facilement calculer la taille moyenne individuelle du Pageot par station. Cette taille est une variable quantitative que l'on peut expliquer par les variables environnementales grâce à la segmentation.

Il se dégage alors trois segments terminaux intéressants (*fig.5*). Ces résultats montrent que les plus grandes tailles apparaissent sur des profondeurs inférieures à 58.5 m et où la teneur en sables gros des sédiments est supérieure à 25%.

Les variables explicatives (sables grossiers et profondeur) ont été là aussi interpolées à l'ensemble de la zone d'étude avant d'être partitionnées en fonction des valeurs de coupure données par les résultats de la segmentation.

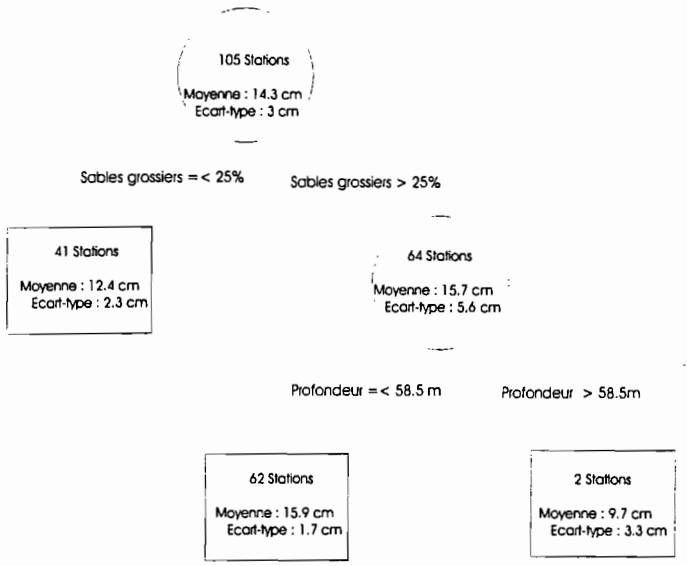

figure 5 : Résultats de la segmentation appliquée à la variable « taille du Pageot »

Par rapport à la question qui nous intéressait ici, c'est à dire identifier les zones où la taille du Pageot est maximale, on a choisit une représentation cartographique des tailles qui doit permettre une interprétation rapide (*fig.6*). On voit ainsi que les fortes tailles, correspondant comme nous l'avons vu au plus fortes teneurs en sables gros sont présentes au centre du plateau continental. De plus, cette carte, associée aux résultats de la segmentation, met en évidence l'effet de seuil dû à la profondeur : la taille du Pageot augmentant de la côte vers le large jusqu'à une certaine profondeur (autour de 58 m) à partir de laquelle la tendance s'inverse, ce qui est confirmé par la *figure 7.*

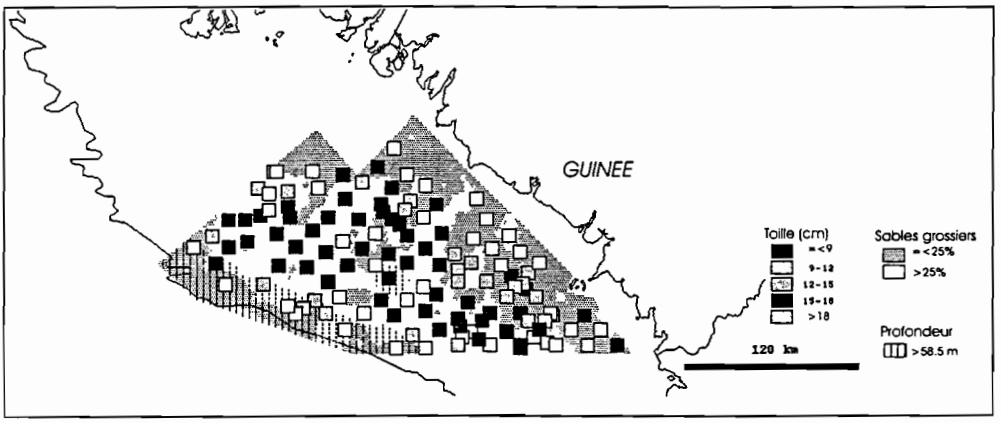

figure 6 : Tailles moyennes individuelles du Pageot en relation avec l'environnement

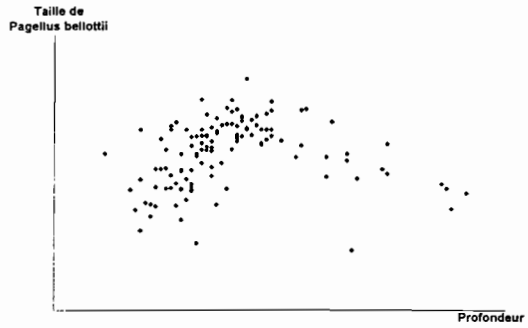

figure 7 : Tailles moyennes individuelles du Pageot en fonction de la profondeur

CONCLUSION:

L'objectif de cette étude était de présenter un exemple de complémentarité entre analyse de données et cartographie, et plus particulièrement de décrire l'apport que peuvent représenter les méthodes de segmentation lorsque l'on souhaite expliquer un facteur biotique par un ensemble de facteurs abiotiques. L'approche spatiale de ces relations ressources-environnement a été facilitée ici au niveau méthodologique par l'utilisation d'un Système d'Information Géographique. Ce logiciel a été utilisé tout au long de cette analyse. Une des fonctions les plus utiles dans cette étude est la possibilité d'effectuer des interpolations. Cette interpolation est utile tant pour la représentation des variables environnementales sur toute la zone d'étude que lors de l'étape préliminaire de construction de tableaux de données où à chaque variable biologique correspond un ensemble de variables discriminantes. Le SIG est bien sûr aussi indispensable pour la visualisation sur une carte des résultats de la segmentation (représentation des segments sous forme de plages de couleurs, superposition des représentations des différentes variables). Il y a donc un va-et-vient permanent entre le logiciel d'analyse de données et le logiciel SIG et on peut donc regretter que l'analyse de données proprement dite ne se fasse au sein même du logiciel SIG, ce qui faciliterait l'analyse. De plus, l'analyse pourrait être optimisée en prenant en compte la dépendance spatiale des données notamment par exemple par l'utilisation de méthodes de géostatistique ce que ne permettait pas ce logiciel SIG.

Quoiqu'il en soit, l'application développée ici montre l'importance de se baser sur une démarche statistique tant pour identifier les variables environnementales les plus intéressantes à représenter que pour construire une partition de ces variables environnementales basé sur une réalité biologique. Ce découpage de l'espace pourrait aussi conditionner le plan d'échantillonnage de campagnes à venir si l'on souhaite, par exemple, décrire un paramètre biologique précis. Ce travail montre aussi l'intérêt de construire des cartes thématiques différentes en fonction du phénomène biologique à interpréter.

De plus, comme le soulignent GUEGUEN et NAKACHE (1988), cette méthode basée sur la construction d'un arbre de décision binaire fournit une règle de décision simple, est robuste vis à vis de données aberrantes et tient compte naturellement des interactions qui peuvent exister entre les données.

Les méthodes de segmentation possèdent donc de nombreux avantages lorsque l'on souhaite construire des cartes en se basant sur des critères statistiques, dont la facilité

d'interprétation directe et de visualisation rapide des résultats mais aussi, comme nous l'avons vu, la possibilité de détermination de seuils.

Ces méthodes de segmentation peuvent aussi être utilisées par exemple à la suite d'une typologie afin de retrouver comment sont combinées entre elles les différentes variables caractéristiques de la classe et ainsi de mieux décrire sa variabilité intrinsèque (TONG et PERINEL, 1996).

Enfin, l'approche décisionnelle de la segmentation, qui permet d'affecter de nouveaux individus dans des segments déjà caractérisés, parait intéressante en halieutique lors de l'élaboration de règles de gestion (TAQUET et al, 1997).

REFERENCES BIBLIOGRAPHIQUES :

BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J., 1984 : « Classification and Regression Trees », Wadsworth International Group.

DOMAIN F. et BAH M.O., 1993 : « Carte sédimentologique du plateau continental guinéen à 1:200 000 et notice explicative » n° 108, ORSTOM-CNSHB, Paris, 15p.

FERRARIS J., PELLETIER D., 1997 : « Deuxième table ronde sur les SIG et Journée thématique sur le spatial », Lettre de l'Association Française d'Halieumétrie, n°5.

FONTANA A., MORIZE E., 1995 : « Projet protection et surveillance des pêches de la ZEE guinéenne- Volet scientifique- Rapport de fin d'étude », ORSTOM-CNSHB, 137p.

GUEGUEN A., NAKACHE J. P., NICOLEAU-MOLINA J., 1996 : « SPAD.S, version 3, Procédures de segmentation », CISIA.

GUEGUEN A., NAKACHE J. P., 1988 : « Méthodes de discrimination basée sur la construction d'un arbre de décision binaire », Rev. Stat. Appl. 36, 19-38.

LEBART L., MORINEAU A., PIRON M., 1995 : « Statistique exploratoire multidimensionnelle », Ed. DUNOD, 440p.

PERINEL E., 1996 : « Segmentation et analyse de données symboliques - Application à des données probabilistes imprécises », Thèse, Université Paris Dauphine, 346p.

SALAUN S, 1997 : « Analyse des relations ressources-environnement au sein d'un système d'information géographique - Application aux ressources démersales du plateau continental guinéen », Rapport de DEA, ORSTOM-UBO, 30p.

TAQUET M., GAERTNER J-C. , BERTRAND J., 1997 : « Typologie de la flottille chalutière du port de Sète par une méthode de segmentation », Aquat. Living Resour. , 10, 137-148

TONG H., PERINEL E., 1996 : « Une approche numérique/symbolique pour l'extraction et la formalisation de connaissances : Application à la description de tactiques de pêche artisanale

au Sénégal ». In : Méthodes d'étude des systèmes halieutiques et aquacoles. J. Ferraris, D. Pelletier, M.J. Rochet éds. Colloques et Séminaires, ORSTOM, 157-164.

Salaun S., Ferraris Jocelyne. (1998)

Utilisation des méthodes de segmentation pour la construction de cartes illustrant les relations ressources halieutiques/environnement

In : Duby C. (ed.), Gouet J.P (ed.), Laloë Francis (ed.). *Biométrie et halieutique*

Paris : Société Française de Biométrie, p. 59-69. (Comptes-Rendus des Sessions de la Société Française de Biométrie ; 15)

Journées de la Société Française de Biométrie, Rennes (FRA), 1998/05/26-28.