REVUE DE STATISTIQUE APPLIQUÉE

C. MICHEL-BRIAND
Y. ESCOUFIER

Segmentation d'un ensemble de courbes

Revue de statistique appliquée, tome 42, n° 4 (1994), p. 5-24
<http://www.numdam.org/item?id=RSA_1994__42_4_5_0>

Rev. Statistique Appliquée, 1994, **XLII** (4), 5-24

SEGMENTATION D'UN ENSEMBLE DE COURBES

C. Michel-Briand, Y. Escoufier

Unité de Biométrie
ENSA-INRA-UM2
9 Place Pierre Viala
34 060, Montpellier cedex

RÉSUMÉ

Les méthodes de segmentation sont des méthodes non paramétriques permettant d'expliquer une variable Y par un ensemble de variables explicatives $\{X^j, j = 1 \dots p\}$.

Nous adaptons ces méthodes dans le cas où la variable Y prend ses valeurs dans un espace de fonctions. Le formalisme proposé inclut celui des méthodes AID (Automatic Interaction Detection) et ELISEE (Exploration des Liaisons et Interaction par Segmentation d'un Ensemble Experimental). Nous présentons les résultats obtenus sur un ensemble de courbes de croissance d'arbres.

Mots-clés : *méthodes de segmentation, arbres de régression, comparaison de courbes, fonctions splines.*

SUMMARY

Classification and regression trees are nonparametric methods which explain one variable Y by a set of variables $\{X^j, j = 1 \dots p\}$. We propose a formalism when the variable Y has values in a functions space including naturally, the methods AID (Automatic Interaction Detection) and ELISEE (Exploration des Liaisons et Interaction par Segmentation d'un Ensemble Experimental). An exemple of application on growth curves is provided.

Keywords : *regression trees, splines, curves comparison..*

1. Introduction

La segmentation s'inscrit dans un ensemble de méthodes non paramétriques qui visent à expliquer une variable Y qualitative ou à valeurs réelles, par un ensemble de variables explicatives $\{X^j, j = 1 \dots p\}$.

Nous nous proposons de mettre en place une méthode de segmentation lorsque la variable Y prend ses valeurs dans un espace de fonctions. Les données à l'origine de cette étude sont des relevés à plusieurs dates, de hauteurs d'arbres ayant poussé sous des climats différents.

Au paragraphe 2, nous rappelons le principe d'une méthode de segmentation dans le cas classique et nous définissons les notations nécessaires pour la suite de notre étude. Nous proposons au paragraphe 3, un formalisme répondant au problème posé et incluant les deux méthodes de segmentation les plus usuelles : AID et ELISEE. Le traitement des données est réalisé au paragraphe 4.

2. Présentation de la méthode de segmentation dans le cas classique

Le problème de la segmentation est de définir une partition des observations sur la base des variables $X^j, j = 1 \ldots p$, de façon à minimiser la variation de Y, la variable à expliquer, à l'intérieur des classes.

Y est soit quantitative uni ou multidimensionnelle, soit qualitative. Quand Y est quantitative, il est naturel de vouloir associer à une classe d'observations la moyenne de Y dans cette classe. Quand Y est qualitative, on est conduit à associer à une classe d'observations, la modalité de Y la plus fréquente.

Initialement, les méthodes de segmentation ont été introduites pour des variables $\{X^j, j = 1 \ldots p\}$ qualitatives. En effet, la fréquence élevée de ce type de problème dans de nombreux domaines : agronomie, aide au diagnostic, gestion, justifiait la mise en place de ce type de méthodes.

Leur utilisation s'est ensuite étendue au cas $\{X^j, j = 1 \ldots p\}$ quantitatives.

Les objectifs essentiels recherchés par les utilisateurs des méthodes de segmentation sont les suivants :

1 – construire une partition de l'échantillon de travail $(X_k, Y_k)_{k=1,\ldots,n}$ au moyen des variables $\{X^j, j = 1 \ldots p\}$ et d'un critère C_n

2 – mettre en évidence les variables les plus explicatives pour Y en donnant une hiérarchie de celles-ci. La hiérarchie n'est pas stricte. Une même variable peut apparaitre à différents niveaux.

3 – fournir une fonction de régression de Y sur X

Les méthodes de segmentation différent uniquement par le choix de C_n, la procédure utilisée étant toujours la même.

Soient $X = (X^j)_{j=1,\ldots,p}$ un p-uplet de variables aléatoires et Y une variable aléatoire définis sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, à valeurs respectivement dans les espaces $E_1 \times \ldots \times E_p$ et G.

Le choix des espaces $(E_j)_{j=1,\ldots,p}$ et G est imposé par la nature qualitative ou quantitative des variables $(X^j)_j$ et Y.

En réalité, un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ i.i.d issu de (X, Y) est observé.

Il induit la probabilité empirique :

$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ où $\delta_{(X_i,Y_i)}$ désigne la mesure de Dirac en (X_i, Y_i).

La présentation des méthodes de segmentation nécessite la définition suivante :

Définition 2.1 : Etant donné un sous ensemble Π de $E_1 \times \ldots \times E_p \times G$, un entier $j = 1, \ldots, p$ et un sous ensemble S de E_j, on appelle *dichotomie* de Π en (j, S), la partition $\{\Pi_1, \Pi_2\}$ de Π déterminée par :

$$\Pi_1 = \{(x^1, \ldots, x^p, y) \in \Pi / x^j \in S\}$$
$$\Pi_2 = \{(x^1, \ldots, x^p, y) \in \Pi / x^j \notin S\}$$

Remarques

1 – Nous nous intéressons uniquement aux dichotomies de la projection canonique de Π notée $\tilde{\Pi}$ sur $E_1 \times \ldots \times E_p$.

2 – La partition $\{\Pi_1, \Pi_2\}$ de Π induit la partition suivante des observations appartenant à Π :

$$\{(X_k, Y_k) \in \Pi / X_k^j \in S\} \ \{(X_k, Y_k) \in \Pi / X_k^j \notin S\}$$

3 – Pour X^j qualitative, S est un sous ensemble des modalités de X^j, pour X^j quantitative, S est un intervalle de R.

La procédure de segmentation est une méthode itérative qui se déroule comme suit :

$C_n(j, S)$ désignera la valeur du critère de sélection de la dichotomie optimale calculé sur la dichotomie engendrée par le couple (j, S);

la procédure démarre pour $\Pi = E_1 \times \ldots \times E_p \times G$ et $\tilde{\Pi} = E_1 \times \ldots \times E_p$

Etape 1 : – rechercher le couple optimal (j_1^*, S_1^*) vérifiant :

$$\text{Arg} \quad \min_{j=1,\ldots,p} \quad \min_{S \subset E_j} \quad C_n(j, S) = (j_1^*, S_1^*)$$

où

$$\text{Arg} \quad \max_{j=1,\ldots,p} \quad \max_{S \subset E_j} \quad C_n(j, S) = (j_1^*, S_1^*)$$

selon la nature du critère C_n;

– effectuer la dichotomie associée à (j_1^*, S_1^*);

Etape k : réitérer le procédé sur *chacun* des sous ensembles issus de l'étape $k - 1$.

A l'issue de la procédure, est obtenue une partition $\mathcal{P}_n^*$ de l'espace $E_1 \times E_2 \times \ldots \times E_p \times G$ induisant une partition de l'échantillon ainsi qu'une hiérarchie des variables explicatives $(X^j)_{j=1,\ldots,p}$. Les résultats sont habituellement représentés par un arbre binaire.

Cette façon de procéder permet également la construction d'une fonction de prédiction du type :

$$m_n(.) = \sum_{\Pi \in \mathcal{P}_n^*} f_n(Y, X, \Pi) 1_{\tilde{\Pi}}(.)$$

l'allure de f_n dépend du critère C_n et de la nature de Y.

Exemples :

(j, S) engendrera la partition Π_1,Π_2 de Π.

Méthodes de segmentation aux moindres carrés :

Le critère C_n caractérisant ces méthodes, qui sont les plus usuelles, est la variance intraclasse de Y, associée à la partition $\{\Pi_1, \Pi_2\}$ des observations de (X, Y) représentées dans la classe Π. C'est donc un critère qu'on cherchera à minimiser.

Plus précisément, soient :

$$G = R^r$$

M une métrique sur R^r

E_n l'espérance par rapport à la probabilité empirique P_n

$C_n(j, S)$:

$$E_n||Y1_{\Pi_1}(X,Y) - \frac{E_n(Y1_{\Pi_1}(X,Y))}{P_n(\Pi_1)}1_{\Pi_1}(X,Y)||^2_M + E_n||Y1_{\Pi_2}(X,Y) - \frac{E_n(Y1_{\Pi_2}(X,Y))}{P_n(\Pi_2)}1_{\Pi_2}(X,Y)||^2_M$$

L'optimalité est obtenue en minimisant le critère C_n.

Pour Y quantitative unidimensionnelle ($r = 1$), M coincide avec l'identité. La méthode associée est connue sous la dénomination A.I.D. (Automatic Interaction Detection).

Pour Y quantitative multidimensionnelle ($r > 1$), la métrique M choisie est celle induite par Γ^{-1}, l'inverse de la matrice des covariances de Y.

La fonction de prédiction associée, dans les deux cas s'écrit :

$$m_n(x) = \sum_{\Pi \in \mathcal{P}_n^*} \frac{E_n(Y1_\Pi(X,Y))}{P_n(\Pi)} 1_{\tilde{\Pi}}(x)$$

(Baccini 1975).

Pour Y qualitative, (les r modalités de Y sont codées par les r vecteurs canoniques de R^r), la métrique M est celle induite par l'inverse de la matrice diagonale des poids des modalités de Y.

La fonction de prédiction est définie par :

soit $P_n(.|X = x)$ la loi empirique conditionnelle de Y sachant $X = x$

$$m_n(x) = \sum_{\Pi \in \mathcal{P}_n^*} f_n(Y, X, \Pi) 1_{\tilde{\Pi}}(x)$$

où $f_n(Y, X, \Pi) = k^*$ avec $k^* = \underset{k=1,\ldots,q}{\text{Arg}} \; P_n(Y = k | X \in \tilde{\Pi})$

La méthode associée est usuellement nommée E.L.I.S.E.E. (Exploration des Liaisons et Interaction par Segmentation d'un Ensemble Expérimental) (Baccini 1975).

Règles d'arrêt :

Les méthodes de segmentation sont des méthodes itératives et nécessitent donc le choix d'une règle d'arrêt. Il existe des règles d'arrêt intuitives qui sont déterminées par un seuil portant sur le nombre d'observations par classe ou la valeur optimale du critère. La plupart du temps, elles conduisent à arrêter trop tôt le découpage ou inversement à accepter des découpages trop fins qui sont ensuite à l'origine de problèmes d'instabilités.

Pour le cas Y unidimensionnelle, il existe une procédure efficace qui permet d'éviter le choix d'un seuil. Elle a été mise au point par Breiman *et al.* (1984). Elle fait partie du catalogue des fonctions disponibles dans le logiciel S (J.M. Chambers *et al.* 1992).

Par ailleurs, il est souhaitable qu' une stabilité de la partition et de la hiérarchie sélectionnées soit assurée par la règle d'arrêt utilisée. L'étude de ce dernier aspect pour des variables explicatives à valeurs réelles n'est pas immédiate, la procédure de segmentation n'étant alors pas récursive : la partition obtenue sur un échantillon de taille $n' > n$ n'est pas forcément plus fine que celle obtenue sur celui de taille n. Elle peut être structurée tout à fait différemment. Cette étude est réalisée dans la thèse de Michel-Briand C. (1993).

3. Courbes et segmentation

Y est désormais une variable aléatoire prenant ses valeurs dans un espace H de fonctions réelles définies sur un intervalle $T = [a, b]$ de R. En réalité, on dispose de q_i observations $i = 1, \ldots, n$ résultant d'une discrétisation des fonctions en des instants $(t_k)_{k=1,\ldots,q_i}$. Au paragraphe 3.1, nous étendons la définition du critère des méthodes de segmentation aux moindres carrés au cas où Y prend ses valeurs dans un espace de fonctions. Au paragraphe 3.2, sont exposés des choix d'espaces H permettant de ramener dans un espace de dimension fini le problème initialement posé dans H.

3.1. Proposition d'un formalisme

Nous avons choisi de sélectionner la partition optimale par un critère mesurant la variance intraclasse de Y dans H, prolongeant ainsi les choix les plus usuels, ceux des moindres carrés. Quelques définitions sont nécessaires pour écrire clairement le critère. Afin d'alléger les notations, nous le présentons lorsque la loi du couple (X, Y) est connue. La procédure de segmentation, quand à elle, ne change pas. C'est celle présentée au paragraphe 2.

Soit H un Hilbert séparable, muni d'un produit scalaire noté $<, >_H$, d'une base orthonormée $(e_i)_{i \in N}$ et de sa tribu borélienne $\mathcal{B}_H$.

Y est désormais une variable aléatoire de $L^2(\Omega, \mathcal{A}, P, H, \mathcal{B}_H)$ espace des variables aléatoires hilbertiennes Z définies sur $(\Omega, \mathcal{A}, P)$ à valeurs dans H et

vérifiant : $\int ||Z||_H^2 dP < \infty$. Notons $\xi(.)$ l'opérateur espérance défini sur l'espace des v.a. hilbertiennes.

Soit Z une v.a. hilbertienne.

ξ est défini par : $\xi(Z) = \sum_i E(< Z, e_i >_H) e_i$

L'opérateur ξ possède les mêmes propriétés que l'opérateur E. Il est indépendant de la base $(e_i)_i$ choisie.

Reprenant les notations introduites au paragraphe 2, le critère $C(j, S)$ s'écrit :

$$E||Y 1_{\Pi_1}(X,Y) - \frac{\xi(Y 1_{\Pi_1}(X,Y))}{P_{X,Y}(\Pi_1)} 1_{\Pi_1}(X,Y)||_H^2 + E||Y 1_{\Pi_2}(X,Y) - \frac{\xi(Y 1_{\Pi_2}(X,Y))}{P_{X,Y}(\Pi_2)} 1_{\Pi_2}(X,Y)||_H^2$$

Il reste à choisir des espaces H convenables pour définir complètement la procédure de segmentation. C'est l'objet du paragraphe suivant.

3.2. Choix d'espaces H

Les choix d'un espace H exposés au cours de ce paragraphe, sont dictés par le désir de ramener le problème initial à un problème de dimension finie. L'analyse des données classique utilisant une représentation matricielle des données, cette volonté se retrouve dans la plupart des travaux d'analyse des données traitant un échantillon de trajectoires de processus. Citons, par exemple : Deville J.C. (1974), Besse P. (1979), Besse P.- Ramsay J.O. (1986), Lenouvel J. (1981), Libert G. et Dupuis Ch. (1981), Saporta G. (1981), Ramsay J.O. (1982), Houllier F. (1987), Virion M.Ch. (1988), El Faouzi N. - Escoufier Y. (1991).

Ces choix d'espaces doivent aussi permettre une comparaison réaliste des fonctions. Nous avons choisi selon la qualité et la signification des observations réalisées, de proposer une reconstitution des fonctions à l'aide d'un opérateur d'interpolation ou d'ajustement; nous verrons alors que comparer deux fonctions reconstituées dans H, revient à comparer deux vecteurs de paramètres dans un espace de dimension fini muni d'une métrique induite par les opérateurs d'interpolation ou d'ajustement.

3.2.1. Espaces H associés à un opérateur d'interpolation

Au cours de ce paragraphe, les observations sont supposées réalisées à des instants identiques : $(t_k)_{k=1,\ldots,q}$. L'observation de fonctions en des instants $(t_k)_{k=1,\ldots,q}$ contient implicitement la notion d'opérateur de discrétisation D_q. De façon classique, celui-ci sera défini sur H, à valeurs dans R^q, linéaire et continu. Nous avons choisi de définir un opérateur Φ_q de reconstitution des fonctions discrétisées de la façon suivante :

Définition : Φ_q défini sur R^q, à valeurs dans H est linéaire, continue. Il vérifie les deux conditions suivantes :

- pour tout x de R^q, $D_q(\Phi_q(x)) = x$ (1)
- pour tout h de H, $\lim_{q \to +\infty} \Phi_q o D_q(h) = h$ (2)

La condition (2) assure une fidélité de la reconstitution par interpolation des fonctions discrétisées.

Si $(b_i)_{i=1,\ldots,q}$ désigne la base canonique de R^q, $(\Phi_q(b_i))_{i=1,\ldots,q}$ est alors une base de $Im\Phi_q$ puisque Φ_q est injectif.

(pour tout x de R^q, $\Phi_q(x)$ est donc défini de façon unique)

Ainsi, les choix de D_q et Φ_q permettent de retrouver de façon immédiate un résultat observé par Saporta en 1981.

Propriété : le choix d'une interpolation dans un hilbert H revient à celui d'une métrique A dans R^q :

$$\forall (x, y) \in R^q, \quad < \Phi_q(x), \Phi_q(y) >_H = {}^t x A y$$

avec $A_{ij} =< \Phi_q(b_i), \Phi_q(b_j) >_H$

Comparer les interpolées dans H, revient donc à comparer les vecteurs d'observations dans R^q muni de la métrique A.

La partition optimale est obtenue en minimisant le critère empirique $C_n(.,.)$ calculé sur les interpolées associées aux réalisations $(Y_i)_{i=1,\ldots,n}$ de Y. Notons le $C_n^q(.,.)$. Compte tenu de la propriété précédente, il mesure la variance intraclasse de $D_q(Y)$ dans l'espace R^q muni de la métrique A.

La condition de convergence sur l'opérateur $\Phi_q o D_q$ assure la convergence de $C_n^q(j, S)$ vers $C_n(j, S)$ lorsque q tend vers $+\infty$, et donc celle des optimaux puisqu'à n fixé, (j, S) décrit un ensemble fini.

Voici deux exemples de choix possibles pour (D_q, Φ_q); on supposera que les instants de discrétisation vérifient : $\lim_{q \to +\infty} \operatorname{Max}_j |t_j - t_{j+1}| = 0$ (avec $t_1 < t_2 < \ldots < t_q$).

Exemple 1 :

$$H = L^2(T, \mathcal{T}, \mu)$$

où rappelons le, T est un intervalle, les t_j appartenant à T, $\mathcal{T}$ étant la tribu borélienne associée et μ la mesure de Lebesgue.

D_q et Φ_q peuvent être définies respectivement sur H et R^{q-1} par :

$$D_q(h) = \left(\frac{1}{\mu([t_{j-1}, t_j[)} \int_{t_{j-1}}^{t_j} h(t) d\mu(t) \right)_{j=2,\ldots,q}$$

$$\Phi_q(x) = \sum_{j=2}^{q} x_j 1_{[t_{j-1}, t_j[}$$

L'interpolation est convergente. En effet, soit : H_q l'espace vectoriel de $L^2(T)$ engendré par les $(1_{[t_{j-1},t_j[})_{j=2,\ldots,q}$. La suite des sous espaces vectoriels $(H_q)_q$ est croissante. La réunion $\underset{q}{U} H_q$ est dense dans $L^2(T)$.

L'opérateur $\Phi_q o D_q$ coïncide avec la projection orthogonale de H sur H_q.

Un théorème issu de celui de Banach-Steinhaus (voir Laurent p. 290) nous permet alors d'obtenir la convergence de $\Phi_q o D_q$ vers Id_H.

Puisque $\Phi_q(b_i) = 1_{[t_{i-1},t_i[}$, A est la matrice de terme général : $A_{ij} = \mu([t_{j-1}, t_j[)\delta_{ij}$ où δ_{ij} désigne le symbole de Kronecker ($\delta_{ij} = 1$ si $i = j$, et 0 sinon).

Ce type de métrique a été utilisé par Deville J.C. (1974) pour l'étude du calendrier de constitution des familles.

Exemple 2 : H est un hilbert à noyau autoreproduisant k, c'est-à-dire un hilbert (ici l'espace des fonctions sur T à valeurs dans R) caractérisé par une fonction k définie sur $T \times T$, à valeurs dans R et vérifiant :

- pour tout t de T $\quad k(t, .) \in H$
- pour tout h de H $\quad < k(t, .), h >_H = h(t)$ (propriété d'autoreproduction)

De la définition de k, on déduit notamment que k est une fonction symétrique et définie positive ou nulle.

L'opérateur D_q est défini par : pour tout h de H, $D_q(h) = (h(t_i))_{i=1,\ldots,q}$. Les espaces à noyau autoreproduisant possèdent de nombreuses propriétés remarquables (*cf.* Duc - Jacquet M., 1973) dont celle-ci : ce sont les seuls Hilbert dont les fonctions d'évaluation δ_t, c'est-à-dire les fonctions qui à tout élément h de H associe $h(t)$, sont continues.

Soit K la matrice d'élément général : $K_{ij} = k(t_i, t_j)$

L'opérateur Φ_q est défini sur R^q par :

$$\Phi_q(x) = \sum_{j=1}^{q} \alpha_j k(t_j, .) \text{ avec } \alpha \text{ le vecteur de coordonnées } \alpha_j \ (j = 1, \ldots, q)$$

vérifiant $\alpha = K^{-1}x$

(cela suppose de s'être restreint à des noyaux k strictement positifs)

Φ_q permet d'associer à chaque vecteur d'observation x, son interpolée spline c'est-à-dire l'interpolée de norme minimale dans H (voir thèse Duc - Jacquet M., Chapitre 2, 1973). Ce type d'interpolation est convergente pour les mêmes raisons que celles données dans l'exemple 1 : on choisira pour H_q, le sous espace vectoriel de H engendré par les $(k(t_i, .))_{i=1,\ldots,q}$.

Nous avons : $\Phi_q(b_i) = \sum_{j=1}^{q} C_{ij} k(t_j, .)$ où $(C_{ij})_{j=1,\ldots,q}$ désigne la $i$ème colonne de K^{-1}.

La métrique induite par Φ_q sur R^q est donc associée à la matrice K^{-1}.

Remarque :

Besse et Ramsay (1986) ont réalisé l'ACP d'un échantillon de fonctions dans des espaces de Sobolev $H^m = \{h/h, h', \ldots, h^{(m-1)}$ absolument continues et $\int (h^{(m)}(t))^2 dt < +\infty\}$ qui sont des espaces à noyau autoreproduisant.

3.2.2. Choix d'espaces H associés à un opérateur d'ajustement

La mise en place d'un opérateur d'ajustement Δ permet d'envisager la situation où les observations $x = (x_k)_{k=1,\ldots,q}$ sont entachées d'erreur $\varepsilon = (\varepsilon_k)_{k=1,\ldots,q}$ supposées i.i.d., centrées et ayant des écarts-types constants soit : $x = D_q oh + \varepsilon$. Elle permet aussi de s'affranchir, le plus souvent, de l'hypothèse concernant l'égalité des instants de discrétisation et ainsi de réduire la dimension de l'espace dans lequel les fonctions sont reconstituées.

Notons H_r un sous espace vectoriel de H, de dimension r ($r \leqslant q$). Eventuellement, r est une fonction de q vérifiant : $\lim_{q \to +\infty} r(q) = +\infty$. De façon classique, nous exigerons que Δ soit un opérateur de R^q dans H_r, linéaire, continu, vérifiant :

$$\operatorname{Arg} \operatorname*{Min}_{g \in H_r} f(D_q(g), x) = \Delta(x)$$

f est une fonction de $R^q \times R^q$ dans R à préciser.

La qualité de l'ajustement sera mesurée, lorsque c'est possible par : $E||h - \Delta(x)||_H^2$ qui se décompose de la façon suivante :

$$E||h - \Delta(x)||_H^2 = ||h - E(\Delta(x))||_H^2 + E||\Delta(x) - E(\Delta(x))||_H^2$$

Le premier terme représente le biais de l'estimateur $\Delta(x)$, le second terme sa variance.

Voici deux exemples classiques de choix possibles pour Δ :

Exemple 1 : *Méthode des moindres carrés*

R^q est munie de sa base canonique. f est ainsi définie :

$$f(D_q(g), x) = ||D_q(g) - x||_q^2$$

où $||.||_q$ correspond à la norme usuelle dans R^q.

Notons $(b_1, \ldots, b_r)$ une base de H_r et G la matrice (q, r) de terme général : $G_{ij} = (D_q(b_j))_i$ Les coordonnées $(\Theta_j)_{j=1,\ldots,r}$ de $\Delta(x)$ dans la base des b_j sont donnés par : $\Theta = (G'G)^{-1}G'x$. La métrique A induite par Δ sur R^r, (c'est-à-dire la métrique induite dans R^r par le choix de la base $b_1, \ldots, b_r$ dans H_r), a pour terme général :

$$A_{ij} = < b_i, b_j >_H$$

Soient $(\Theta^i)_{i=1,\ldots,n}$ les n vecteurs coordonnés des ajustées des réalisations $(Y_i)_{i=1,\ldots,n}$ (avec $Y_i \in R^q$) de Y. La partition optimale est obtenue en minimisant $C_n(.,.)$ calculée sur les ajustées. Il mesure donc la variance intraclasse des $(\Theta^i)_{i=1,\ldots,n}$ calculée dans R^r muni de la métrique A.

Les B-splines de par leurs propriétés peuvent fournir une base intéressante. En particulier, une erreur d'échantillonnage n'aura qu'une répercussion locale, les paramètres du modèle n'étant pas ajustés sur toutes les observations. On trouvera dans Deboor C. (1978) et Schumaker L. (1981) les définitions et propriétés de celles-ci. Les vitesses de convergence du biais et de la variance de $\Delta(x)$ sont données par Girdhar G. Agarwal and Studden W.J. (1980). Elles sont de l'ordre de $\frac{1}{k^{2d}}$ pour le biais et $\frac{k}{n}$ pour la variance, k étant le nombre de nœuds et d l'ordre des B-splines. (la dimension de l'espace H_r est : $r = k + d$).

Cox D.D. (1988) propose des constructions de sous espaces H_r pour lesquels sont obtenus des vitesses de convergence en $r^{-(\rho-\tau)/2}$ pour le biais et en $n^{-1}r^{\tau+1}$ pour la variance; ρ et τ sont des réels intervenant dans la définition de la norme de H. D'autres bases du type B-spline peuvent être utilisées comme les M-splines ou les I-splines (voir Ramsay 1988 El Faouzi - Escoufier 1991).

Exemple 2 : *Ajustement spline*

Soit H un espace à noyau autoreproduisant k, et ρ un réel strictement positif.

$\Delta(x)$ est la fonction de H minimisant parmi les fonctions u de H :

$$||u||_H^2 + \rho \sum_{i=1}^{q} (u(t_i) - x_i)^2$$

Notons : K la matrice de terme général $k(t_i, t_j)$ et $||\ ||_{q,K^{-1}}$ la norme induite par K^{-1} sur R^q. Résoudre le problème précédent revient à rechercher dans R^q le minimum de l'expression suivante :

$$\left\{ \begin{array}{l} f(y,x) = ||y||_{q,K^{-1}}^2 + \rho \sum_{i=1}^{q} (y_i - x_i)^2 \\ (1) \\ \text{avec} \quad y = D_q(g) \quad g = \sum_{j=1}^{q} \alpha_j k(t_j, .) \quad \text{et} \quad \alpha = K^{-1}y \end{array} \right.$$

α désignant le vecteur de composantes α_j $(j = 1, \ldots, q)$ En effet :

$$\underset{u\in H}{\text{Min}}\left[||u||_H^2+\rho\sum_{i=1}^{q}(u(t_i)-x_i)^2\right]=\underset{y\in R^q}{\text{Min}}\left[\underset{u\in H}{\text{Min}}(||u||_H^2+\rho\sum_{i=1}^{q}(y_i-x_i)^2\right]$$

avec $u(t_i)=Y_i$.

Or, nous avons vu au paragraphe 3.2.1, que l'interpolée spline en y_i ($i=1,\ldots,n$) (c'est-à-dire l'interpolée de norme minimale dans H) que nous noterons g, est un élément de H_q, sous espace vectoriel de H engendré par $(k(t_i,.))_{i=1,\ldots,q}$: $g=\sum_{j=1}^{q}\alpha_j k(t_j,.)$ avec α le vecteur de coordonnées α_j ($j=1,\ldots,q$) vérifiant $\alpha=K^{-1}y$. La métrique induite dans R^q étant associée à la matrice K^{-1}, l'expression (1) est ainsi obtenue.

La détermination de Δ se fait en différentiant l'expression (1) par rapport à y_j pour $j=1,\ldots,q$. On obtient :

$$(\Delta(x)(t_i))_{i=1,\ldots,q}=(I+\frac{1}{\rho}K^{-1})^{-1}x$$

d'où :

$\Delta(x)=\sum_{j=1}^{q}\beta_j k(t_j,.)$ avec β le vecteur de coordonnées β_j ($j=1,\ldots,q$) vérifiant $\beta=K^{-1}(\Delta(x)(t_i))_i$ soit encore : $\beta=(K+\frac{1}{\rho}I)^{-1}x$.

Il est à noter que l'utilisation de l'ajustement spline exige l'égalité des instants en lesquels les réalisations Y_i de Y sont observées. La métrique induite par Δ sur R^q a pour terme général :

$$< k(t_i,.),k(t_j,.) >_H = k(t_i,t_j)$$

Elle est associée à la matrice K.

Soient : $(\beta^i)_{i=1,\ldots,n}$ les n vecteurs coordonnées de dimension q des ajustées des réalisations $(Y_i)_{i=1,\ldots,n}$ de Y. La partition optimale est obtenue en minimisant $C_n(.,.)$ calculé sur les ajustées. Compte tenu des développements précédents, il mesure la variance intraclasse des $(\beta^i)_{i=1,\ldots,n}$ calculée dans R^q muni de la métrique K.

4. Un exemple d'application

Ce paragraphe décrit une application des méthodes exposées précédemment à un exemple particulier. Les individus soumis à l'étude sont ici des arbres; la variable à expliquer Y est la hauteur des arbres; celle-ci a été relevée à plusieurs dates; c'est donc par nature une courbe discrétisée. Les variables X^j sont des descripteurs des conditions de culture de ces arbres.

On pourrait envisager pour l'analyse de ces données d'autres approches que la segmentation. On peut par exemple faire des classes sur la base de la variable Y seule puis rechercher les configurations de descripteurs X^j les plus fréquents dans ces classes. On peut symétriquement faire des classes sur la base des X^j seuls puis associer à chaque classe une description résumée raisonnable des courbes Y

associées aux individus de la classe, la moyenne par exemple. La mise en œuvre de ces deux approches peut être multiple : elle peut reposer sur l'utilisation de véritables méthodes de classification ou sur la reconnaissance visuelle de groupes dans des plans factoriels. Dans ce cas la technique des points supplémentaires fournira une approche qualitative de la dépendance de Y par rapport aux X^j. Ces méthodes ont l'avantage de traiter globalement les variables X^j; elles ont l'inconvénient de ne prendre en compte l'objectif de l'étude, l'explication de Y par les X^j, que d'une manière tardive, *à posteriori*. L'analyse en composantes principales par rapport à des variables instrumentales (et ses extensions dans les cas qualitatifs) n'oublie pas l'objectif de l'étude : elle lui donne la forme stricte d'une contrainte d'appartenance de la solution à un espace vectoriel particulier qui peut être jugée moins utilisable que les solutions fournies par la segmentation. Quoiqu'il en soit, il pourra être intéressant dans certaines études de se donner la possibilité de bénéficier de l'éventuelle complémentarité des résultats des différentes approches.

Ayant substitué à une courbe un ensemble de points de discrétisation, une procédure assez fréquente chez les utilisateurs de l'analyse des données est de traiter les données comme si les discrétisations constituaient un vecteur d'observation. On doit souligner qu'une telle approche détruit l'information temporelle contenue dans les données. Il suffit pour s'en convaincre de remarquer que la permutation des deux mêmes points de discrétisation dans toutes les courbes ne perturbe pas les résultats. La représentation des courbes utilisées ici et qui pourrait l'être pour d'autres approches de l'analyse des données n'a pas cet inconvénient et en ce sens parait préférable.

Au paragraphe 4.1, nous présentons les données. Le paragraphe 4.2 détaille la mise en œuvre de la méthode.

4.1 Présentation des données

Les données ont été mises à notre disposition par le laboratoire LECSA du centre de recherche INRA de Montpellier.

Elles concernent 152 arbres dont la hauteur a été relevée durant une période de six mois, à des dates identiques mais non équidistribuées pour chacun des arbres. Neuf mesures ont été ainsi relevées pour chacun d'eux. Les arbres étudiés sont : des noyers hybrides, des noyers communs, des noyers noirs et des merisiers.

Une partie de ces arbres a poussé librement. L'autre partie a été mise sous abri serre. Cinq types d'abri serre ont été utilisés :

1. des abris serre standards : gaine de 10 cm de diamètre en moyenne, étanche.

2. des abris serre ventilés : la gaine est percée d'une dizaine de trous de 3 cm de diamètre et est soulevée de 10 cm du sol.

3. des abris serre microperforés : la gaine est percée de petits trous de 1 mm de diamètre sur toute sa surface.

4. des abris serre dont l'atmosphère intérieure est mise continuellement sous pression de CO_2.

5. des abris serre pour lesquels sont placés à leur base, un bloc de carbonate.

Ainsi sont créés différents milieux climatiques caractérisés par :

- le taux de CO_2.

- une température variant à l'intérieur des abris serre de + 5 ° C à + 10 ° C par rapport à la température extérieure.

- la présence plus ou moins importante du vent.

- un taux d'humidité relative qui atteint pour les types d'abri serre 1 - 3 - 4 - 5 le taux de 100 %

- un rayonnement lumineux constamment 60 % plus faible que celui à l'extérieur pour tous les types d'abri serre.

D'autre part, une partie des arbres : une dizaine de témoins et une dizaine par type d'abri serre a été irriguée par un système de goutte à goutte.

4.2 Traitement des données

Les variables explicatives $(X^j)_{j=1,\ldots,8}$ issues du protocole expérimental décrit ci-dessus sont :

l'espèce : variable qualitative dont les modalités sont noyer commun, noyer noir, noyer hybride, merisier respectivement codées 1, 2, 3 et 4

le taux de CO_2 : variable qualitative à 3 modalités taux de CO_2 relevé à l'extérieur pour la modalité 1 puis augmenté pour les modalités 2 et 3

la température : variable qualitative à 3 modalités : t°extérieure, t°extérieure + 5°, t°extérieure + 10°respectivement codées 1, 2 et 3

le vent : variable qualitative à 3 modalités absence de vent, présence partielle de vent, plein vent, respectivement codées 1, 2 et 3

le taux d'humidité relative : variable qualitative à 3 modalités taux d'humidité extérieur, taux moyen, taux de 100 % respectivement codées 1, 2 et 3

le rayonnement lumineux : variable qualitative à 2 modalités rayonnement extérieur, rayonnement de 60 % plus faible respectivement codées 2 et 1

l'irrigation : variable qualitative à 2 modalités, 1 désignera l'absence d'irrigation

la hauteur initiale des arbres : variable quantitative découpée en 7 classes; cette variable correspond à la mesure faite au premier instant. Elle prend les valeurs de 1 à 68 cm.

Au vu des données, il n'y a pas de raison de privilégier un modèle de croissance précis. Aussi, il semble raisonnable d'envisager une interpolation ou un ajustement spline.

Nous nous contenterons de présenter les résultats obtenus pour un espace H, associé à un opérateur d'interpolation. Nous avons retenu un espace H qui permette de mettre en évidence l'influence des variables $(X^j)_{j=1,\ldots,8}$ sur les vitesses de

croissance. Plus précisément, notre choix s'est porté sur un espace de Sobolev dont le produit scalaire prend uniquement en compte les dérivées.

H est l'espace de Sobolev suivant :

$H = H_0^1[0,1] = \{h/h$ absolument continue, $h' \in L^2[0,1]$ et $h(o) = o\}$

Il est muni du produit scalaire :

$$\forall (h_1, h_2) \quad < h_1, h_2 >_H = \int_0^1 h_1' \; h_2' dt$$

Son noyau k est défini par : $k(t_i, t_j) = \min(t_i, t_j)$; il ne dépend que des instants de discrétisation.

On trouvera, par exemple dans Duc-Jacquet 1973, l'explicitation des noyaux des espaces de Sobolev les plus classiques.

La matrice symétrique K d'élément général $K_{ij} = k(t_i, t_j)$ est ici :

$$K = \begin{bmatrix} 0,09 & - & - & - & - & - & - & - \\ 0,09 & 0,21 & - & - & - & - & - & - \\ 0,09 & 0,21 & 0,34 & - & - & - & - & - \\ 0,09 & 0,21 & 0,34 & 0,47 & - & - & - & - \\ 0,09 & 0,21 & 0,34 & 0,47 & 0,61 & - & - & - \\ 0,09 & 0,21 & 0,34 & 0,47 & 0,61 & 0,74 & - & - \\ 0,09 & 0,21 & 0,34 & 0,47 & 0,61 & 0,74 & 0,87 & - \\ 0,09 & 0,21 & 0,34 & 0,47 & 0,61 & 0,74 & 0,87 & 1 \end{bmatrix}$$

Notons $D_8(Y_i)$ le vecteur de composantes $Y_i(t_k)$ pour $k = 1, \ldots, 8$ (où $Y_i(t_k)$ est la longueur de l'arbre i à l'instant t_k).

Le critère de sélection de la partition optimale mesure la variance intraclasse des $D_8(Y_i)$ pour $i = 1, \ldots, 152$ dans R^8 muni de la métrique K^{-1} (voir paragraphe 3.2.1. exemple 2).

La règle d'arrêt utilisée lors de la procédure est la suivante : on refuse les dichotomies induisant deux classes ayant moins de sept observations (5 % de l'effectif total), ou dont la somme des inerties n'est pas suffisamment faible. Plus précisément, une classe n'est plus dichotomisée lorsque la dichotomie optimale sélectionnée est associée à une valeur du critère supérieure à l'inertie de la classe moins 5 % de l'inertie de la population totale. Les seuils choisis sont ceux habituellement utilisés (Baccini A.). Ainsi, à l'issue de la procédure, on obtient un arbre induit par une partition à 4 classes. La première variable sélectionnée oppose les arbres mis sous abri-serre aux arbres témoins.

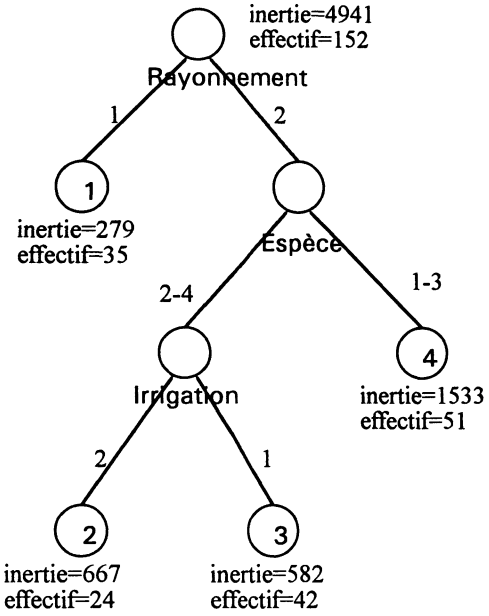
Aux pages suivantes sont présentées : l'arbre de régression (figure 1) et les moyennes des courbes de croissance des arbres appartenant aux classes 1, 2, 3 et 4 (figures 2a, 3a, 4a, 5a), ainsi que quelques courbes de croissance appartenant à chacune de ces classes (figures 2b, 3b, 4b, 5b); la moyenne d'un ensemble de courbes est un résumé qui comporte des limites évidentes. Le produit scalaire choisi ne faisant intervenir que les dérivées des courbes de croissance, on peut dire que :

la classe 1 rassemble les arbres qui ont des vitesses de croissance quasiment nulles sur la période d'observation;

la classe 2 contient les arbres qui ont une croissance assez rapide après un démarrage un peu lent; on ne relève pas la présence nette de palier;

les arbres de la classe 3 sont caractérisés par une croissance peu rapide avec pour la plupart la présence d'un palier en fin de croissance;

la classe 4 est moins homogène que les précédentes : la majeure partie des arbres ont tout d'abord une vitesse de croissance élevée, puis une vitesse nulle en fin de période d'observation (voir figure 5a); la présence d'arbres dont la courbe de croissance présente un palier en début de période ou une absence de palier en fin de période (voir figure 5b) explique l'allure de la courbe moyenne obtenue et la forte inertie de cette classe.



arbre de régression obtenu avec la métrique K^{-1}

figure 1

L'arbre de segmentation obtenu, permet de dire par exemple : à un arbre dont la courbe de croissance est inconnue , mais pour lequel la variable rayonnement prend la modalité 1, la méthode associe pour courbe de croissance, la moyenne des courbes de croissance représentées dans la classe 1 (figure 2a).

Remarque :

Les résultats ont été obtenus grace à un programme spécifiquement écrit pour ce type de données. Il diffère des programmes standard de segmentation par la nature

multidimensionnelle de Y et par la prise en compte de la métrique A. Il a été réalisé en Turbo Pascal 5 sur IBM PC. Une copie du programme peut être obtenue auprès des auteurs.

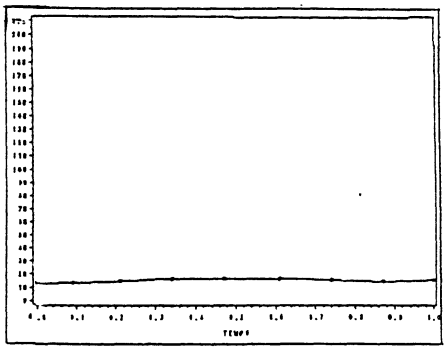


Figure 2a
Moyenne des courbes de croissance dans la classe 1.

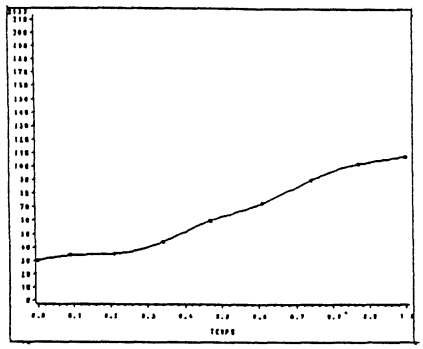
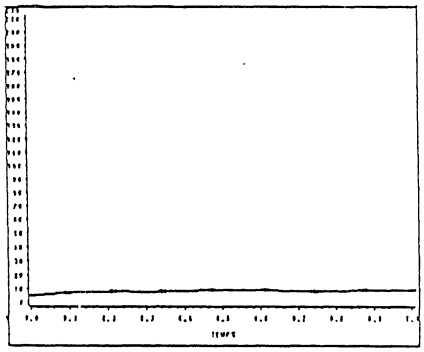


Figure 3a
Moyenne des courbes de croissance dans la classe 2.



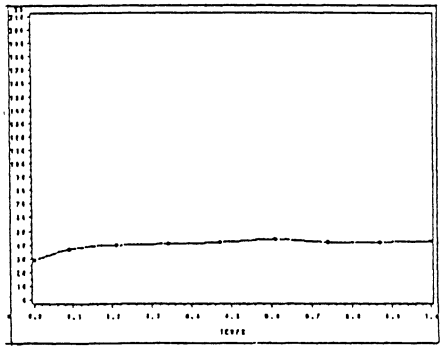
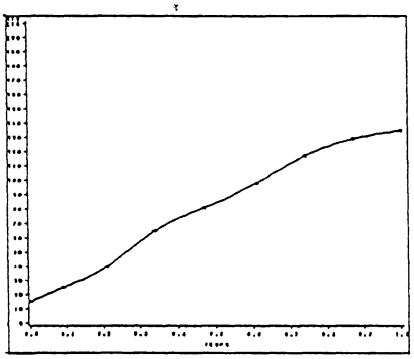


Figure 2b
2 courbes de croissances appartenant à la classe 1.



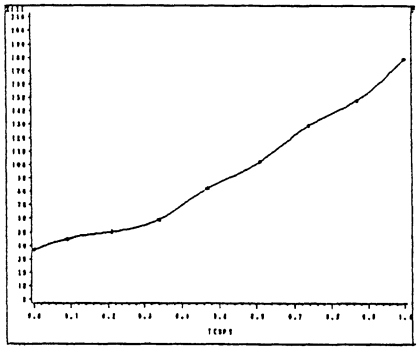


Figure 3b
2 courbes de croissances appartenant à la classe 2.

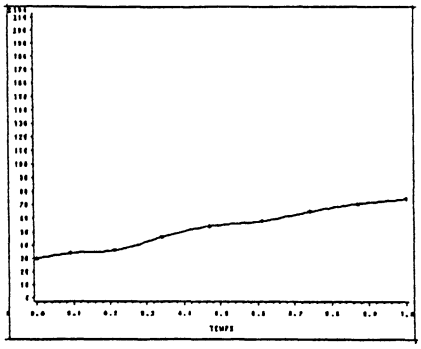
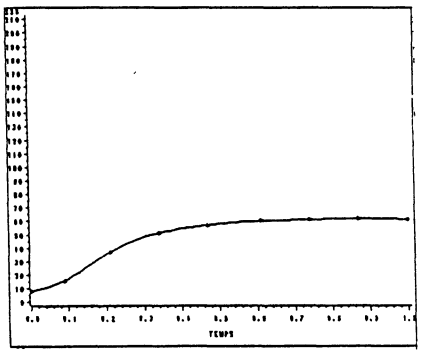
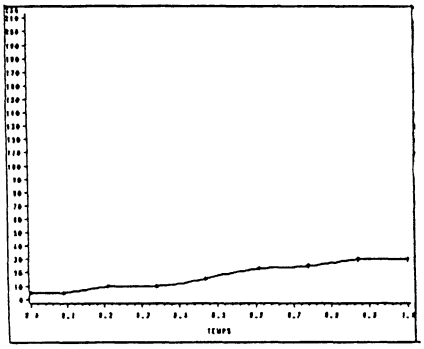


Figure 4a
Moyenne des courbes de croissance dans la classe 3.





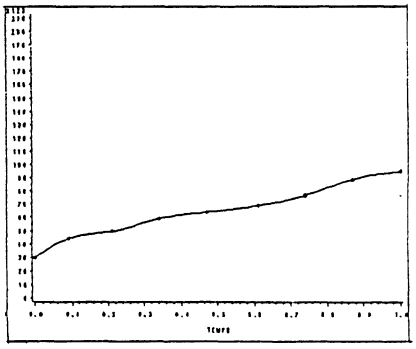


Figure 4b
3 courbes de croissances appartenant à la classe 3.

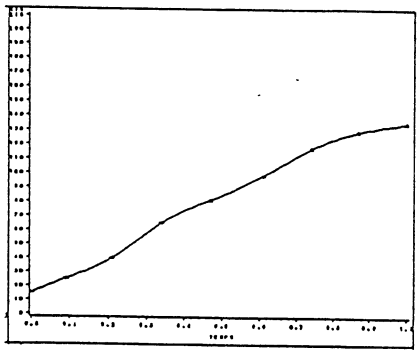


Figure 5a
Moyenne des courbes de croissance de la classe 4.

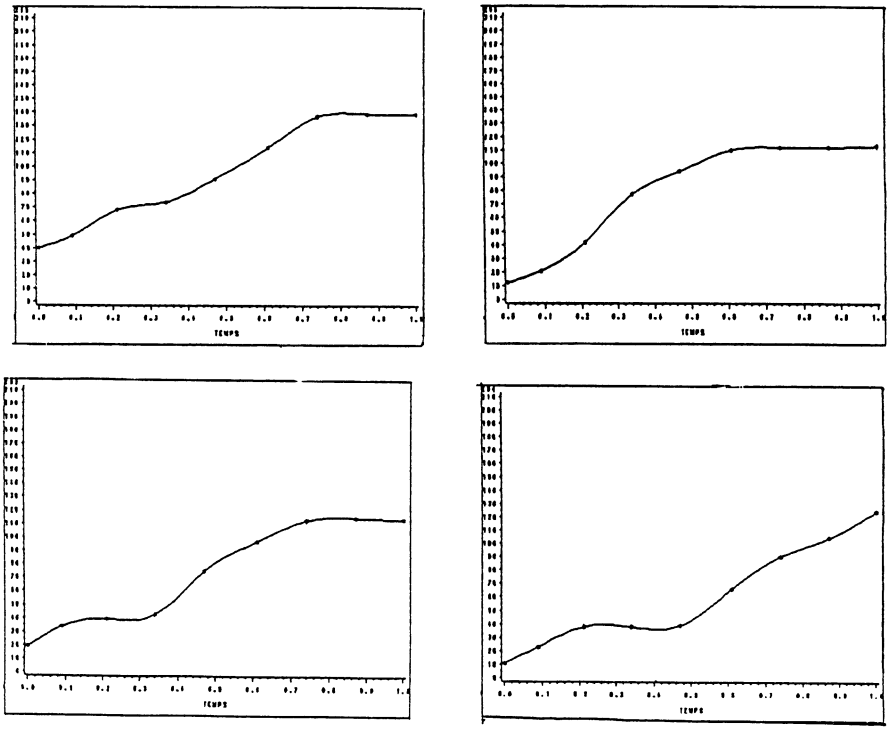


Figure 5b
4 courbes de croissances appartenant à la classe 4.

Références Bibliographiques

BACCINI A. (1975). «Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées». Thèse de 3e cycle. Université P. Sabatier. Toulouse.

BESSE PH. (1979). «Etude descriptive des processus : Approximation et interpolation». Thèse de 3e cycle. Université P. Sabatier. Toulouse.

BESSE PH.,RAMSAY J.O. (1986). "Principal components analysis of sampled functions" Juin. Psychometrika. Vol. 51, n° 2, p. 285-311.

BREIMAN L.,FRIEDMAN J.H., OLSHEN R.A., STONE C.J. (1984). "Classification and regression trees" 1984. Belmont C.A. Wadsworth.

CHAMBERS J.M.,HASTIE T.J. (1992). "Statistical models in S" . Wadsworth Brooks/Cole.

COX D.D. (1988). "Approximation of least squares regression on nested subspaces". Annals of Stat. Vol. 16, n° 2, p. 713-732.

DEBOOR C. (1978). "A practical guide to splines". New York - Springer Verlag.

DEVILLE J.C. (1974). «Méthodes statistiques et numériques de l'analyse harmonique». Annales de l'INSEE, 15, p. 3-101.

DUC-JACQUET M. (1973). «Approximation des fonctionnelles linéaires sur les espaces hilbertiens autoreproduisants». Thèse de 3e cycle. Université Scientifique et Médicale de Grenoble.

EL FAOUZI N. ESCOUFIER Y. (1991). «Modélisation I-spline et comparaison de courbes de croissance». Revue Stat. Appliquées. xxxix (1), p. 51-64.

GIRDHAR G., AGARWAL - STUDDEN W.J. (1980). "Asymptotic integrated mean square error using least squares and bias minimizing splines". Annals of Stat. Vol. 8, n° 4, p. 1307-1325.

HOULLIER F. (1987). «Comparaison des courbes et modèles de croissance : choix d'une distance entre individus». Statistique et Analyse des Données. Vol. 12, n° 3, p. 17-36.

LAURENT P.J. (1972). «Approximation et optimisation». Coll. Enseignement des Sciences - Chez Hermann.

LENOUVEL J. (1981). «Etude d'une famille de courbes par des méthodes d'analyse des données : application à l'analyse morphologique de courbes provenant de données médicales". Thèse de 3e cycle. Université de Rennes 1.

LIBERT G., DUPUIS CH. (1981). «Comparaison de courbe de thermoluminiscence de quartz par l'analyse des coefficients d'autocorellation». Revue Stat. Appliquées. Vol. 29, n° 4, p. 51-59.

MICHEL-BRIAND C. (1993). «Etude asymptotique et applications des méthodes de segmentation ». Thèse. Université Montpellier 2.

RAMSAY J.O. (1982). "When data are functions". Psychometrika. Vol. 47, p. 379-396.

RAMSAY J.O. (1988). "Monotone regression splines in action". Statistical Sciences. Vol. 3, n° 4, p. 425-461.

SAPORTA G. (1981). «Méthodes exploratoires d'analyse des données temporelles». Thèse d'Etat. UMC Paris 6.

SCHUMAKER L. (1981). "Spline functions : Basic theory". Wiley Interscience.

VIRION M. CH. (1988). «Méthodologies statistiques de la discrimination : application aux électrophorégrammes des farines de blé». Thèse. USTL Montpellier.