REVUE DE STATISTIQUE APPLIQUÉE

N. EL FAOUZI
Y. ESCOUFIER

Modélisation I-spline et comparaison de courbes de croissance

Revue de statistique appliquée, tome 39, nº 1 (1991), p. 51-64
<http://www.numdam.org/item?id=RSA_1991__39_1_51_0>

© Société française de statistique, 1991, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1991, **XXXIX** (1), 51-64

MODÉLISATION I-SPLINE ET COMPARAISON DE COURBES DE CROISSANCE

N. EL FAOUZI, Y. ESCOUFIER

Unité de biométrie, ENSAM-INRA-UM.II, 9, place Pierre Viala, 34060 Montpellier

RÉSUMÉ

L'étude porte sur des courbes de croissance; on veut d'une part les soumettre à une étude exploratoire qui cherche à reconnaître des groupes homogènes; d'autre part soumettre à des tests de significativité les différences observées dans des groupes spécifiés.

On propose d'ajuster chaque courbe par un modèle de grande flexibilité, défini par les "I-splines". Le caractère linéaire de ce modèle, qui s'écrit comme combinaison linéaire des I-splines, conduit à une distance euclidienne entre les paramètres ce qui autorise l'emploi des méthodes classiques d'analyse des données pour la partie exploratoire et la mise en oeuvre de tests de permutations pour la partie confirmatoire.

Mots-clés : *courbes de croissance, fonctions splines, modélisation, comparaison de courbes.*

ABSTRACT

This paper studies growth curves; first we want to conduct an exploratory study in order to recognize homogeneous groups, we also want to test the significance of the differences observed between given groups.

We propose fitting each curve by a flexible model defined by "I-splines". The linearity of this model, which is a linear combination of I-splines, leads to an euclidean distance between the adjusted parameters and this allows using the classical data analysis methods for the exploratory part of the study, and a permutation test for the confirmatory part of this work.

Key-words : *courbes de croissance, fonctions splines, modélisation, comparaison de courbes.*

1. Introduction : Motivations et objectifs

Les données qui ont motivé cette étude sont des pesées d'agneaux faites pendant 22 semaines sur trois lots de six animaux chacun (*cf.* SAIH A [1985]).

Un premier lot, noté C, est un lot de contrôle; les animaux qui le composent n'ont reçu aucun traitement; le deuxième lot, noté I, a été effectivement traité : le produit objet de l'étude est appliqué aux animaux par injection. Les animaux

du troisième lot, noté T, ont subi une injection du support du produit mais pas du produit lui même.

Dans une telle étude, deux objectifs statistiques peuvent être proposés. On peut vouloir d'abord dans une phase exploratoire visualiser la variabilité des évolutions de poids, c'est à dire identifier des courbes ressemblantes et caractériser les éléments de dissemblances des courbes qui ne se ressemblent pas.

Dans une seconde étape, on voudra apprécier l'effet du traitement. On pourra ici envisager de comparer les trois groupes entre eux, les deux groupes injectés au groupe témoin, le groupe effectivement traité aux deux groupes non traités.

La première phase sera possible si on est capable d'attacher aux courbes quelques descripteurs succeptibles de supporter une méthode d'analyse des données. Ce point sera étudié dans les paragraphes **2** et **3**.

La seconde phase demande qu'on puisse attacher à l'ensemble des courbes une mesure de variabilité totale, qu'on sache la découper en une variabilité interne aux groupes et une variabilté entre les groupes et qu'on puisse ensuite apprécier leur intensité relative. Cet aspect du travail sera traité dans le paragraphe **4** en cohérence avec la démarche des paragraphes précédents.

2. L'analyse des données lorsque les données sont des courbes

Le sujet n'est pas nouveau il a déja fait l'objet de plusieurs publications, DEVILLE J.C [1974], BESSE P [1979], LENOUVEL J [1981], LIBERT G & DUPUIS Ch [1981], SAPORTA G [1981], RAMSAY J.O [1982b], HOULLIER F [1987]. Avant de faire une nouvelle proposition rappelons rapidement différentes posibilités et les inconvénients qu'on peut leur trouver.

Pour les besoins de l'exposé, on considére qu'on dispose d'un ensemble $\mathcal{C}$ de $\mathcal{N}$ courbes dépendant d'un paramètre t parcourant un intervalle donné. Soit donc $\mathcal{C} = \{C_i; i = 1, \ldots, \mathcal{N}\}$.

2.1. Des approches connues.

2.1.1. Première approche.

Les courbes étant toutes connues aux mêmes points d'observations $\{t_j; j = 1, \ldots, p\}$, les données disponibles peuvent être rassemblées en un tableau X de dimension $\mathcal{N} \times p$, d'élement $X_i^j = C_i(t_j)$. Tout choix de métrique Q sur $\mathbb{R}^p$ et D sur $\mathbb{R}^{\mathcal{N}}$ permet alors de définir un triplet (X, Q, D) qui pourra être le point de départ d'une analyse en composantes principales (ACP) ou d'une méthode de classification. Cette méthode ne pourra évidemment pas être utilisée lorsque les courbes ont été observées à des instants $\{t_{i,j} : j = 1, \ldots, p_i\}$ différents par leur localisation et leur nombre. De plus, la succession naturelle des temps d'observation n'est pas prise en compte dans la méthode : La permutation des colonnes du tableau est sans effet sur la représentation et la comparaison des lignes.

2.1.2. Deuxième approche

Un modèle général $C(t, \theta_1, \theta_2, \ldots, \theta_k)$ dépendant de k paramètres ayant été choisi, on peut caractériser chacune des courbes $C_i(t)$ par les valeurs $(\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik})$ des paramètres obtenues à l'issue d'une procédure d'ajustement des observations $\{C_i(t_{i,j}) : j = 1, \ldots, p_i\}$ par le modèle $C(t, \theta_1, \theta_2, \ldots, \theta_k)$. On associe alors aux $\mathcal{N}$ courbes un tableau Y de dimension $\mathcal{N} \times k$ d'élement $Y_i^l = \theta_{il}$. Reste à choisir la métrique Q sur $\mathbb{R}^k$ à appliquer aux lignes de Y pour qu'une comparaison des lignes Y_i et $Y_{i'}$ puisse être considerée comme une comparaison réaliste des courbes $C_i(t)$ et $C_{i'}(t)$.

HOULLIER F [1987] a étudié ce problème. Pour des modèles $C(t, a_1, a_2, \ldots, a_k)$ polynomiaux, la solution est simple lorsque toutes les courbes ont été observées aux mêmes instants. Pour des modèles non polynomiaux, cas des modèles usuels de croissance, la distance à utiliser au point Y_i dépend de ce point. C'est à dire qu'on sort du contexte de la géometrie euclidienne qui sert de base à l'analyse linéaire des données. On peut s'y ramener en prenant une métrique moyenne mais ceci n'est qu'une heuristique.

- **Remarque :** Une similarité quelconque sur l'ensemble des courbes conduira à une matrice $S_{\mathcal{N} \times \mathcal{N}}$, de similarité entre courbes. La méthode d'ACPVI[1], (*cf.* RAO C.R [1964], ROBERT P & al [1976], BONIFAS L & al [1984]) permettra de trouver la métrique Q à appliquer à Y pour que la quantité $\|SD - Y'QYD\|^2$ soit minimale. Q présente l'inconvénient de dépendre des observations faites et sera difficilement utilisable pour d'autres courbes.

On doit remarquer pour l'opposer à la partie **2.2** de ce paragraphe, que tous les paramètres du modèle ajusté dépendent de toutes les observations.; Ceci entraine qu'une erreur d'échantillonnage importante sur l'une des observations faites perturbe globalement le modèle alors qu'on pourrait espérer que cette perturbation ne soit que locale.

2.1.3. Troisième approche

On peut associer à chaque courbe un vecteur de descripteurs $(b_1, \ldots, b_k)$ considérés comme importants ou caractéristiques dans le contexte de l'étude faite : Ce seront des temps mis à atteindre certains valeurs; des temps passés aux dessus d'un certain palier; le nombre de franchissement d'un palier dans un temps donné,.... Ne nions pas qu'une bonne connaissance du contexte de l'étude peut conduire à une description des courbes pertinentes pour l'application faite. Comparer les courbes entre elles conduira tout de même à calculer des distances ou des dissimilarités sur la bases de ces descripteurs et il peut s'avérer difficile de restituer les ressemblances et dissemblances des courbes elles mêmes.

Soit Y le tableau $\mathcal{N} \times k$ des descripteurs choisis. La remarque du paragraphe **2.1.2** est valable ici aussi.

---

[1] Analyse en Composantes Principales par rapport à des variables instrumentales.

2.2. Une approche fondée sur les I-splines

L'approche proposée reléve de la famille des méthodes qui consistent à approcher toute fonction par combinaison linéaire des éléments d'une base de fonctions convenablement choisie. Suivant les travaux de RAMSAY J.O [1982a, 1988], nous choisissons la base des I-splines.

- Les M-splines ont été introduites par CURRY H.B & SCHOENBERG I.J [1946] qui en ont détaillé les proprietés en [1966]. Une étude complète des B-splines, fortement liées aux M-splines, a été faite par DEBOOR [1978].

Considérons pour les présenter un intervalle fermé $[L, U]$ et $s = (s_l)_{l=1\ldots n}$ un ensemble de point intérieurs à l'intervalle tels que pour tout $l = 1, \ldots, n$; $s_l \leq s_{l+1}$.

On notera $1_{[a,b[}$ la fonction indicatrice de l'intervalle $[a, b[$.

Pour tout k supérieur ou égal à 1, la base des M-splines d'ordre k est la famille notée $(M_{k,l})$ définie par la famille récurrente suivante :

$$M_{1,l}(x; s) = \frac{1_{[s_l, s_{l+1}[}}{(s_{l+1} - s_l)}$$

Pour tout $k > 1$, et pour tout l,

$$M_{k,l}(x; s) = \frac{k\{(x - s_l)M_{k-1,l}(x; s) + (s_{l+k} - x)M_{k-1,l+1}(x; s)\}}{(k-1)(s_{l+k} - s_l)}$$

On voit que les éléments des bases des M-splines sont des fonctions positives sur un intervalle et nulles ailleurs. Pour tout couple (k, l) , $\int_L^U M_{k,l}(x; s)dx = 1$, si bien que toute spline de base a les caractéristiques d'une densité de probabilité.

- RAMSAY J.O [1982a], introduit les I-splines de degré k définies par :

$$I_{k,l}(x; s) = \int_L^x M_{k,l}(u; s)du.$$

La positivité des M-splines entraine la positivité et la monotonie des I-splines. Comme $I_{k,l}(U; s) = 1$, on en déduit que les I-splines sont bornées et donc constantes partout à l'exception d'un intervalle borné. Pratiquement si $s_j \leq x \leq s_{j+1}$ on montre que :

$$I_{k,l}(x; s) = \begin{cases} 0 & \text{si } l > j \\ \sum_{u=l}^{j} (s_{u+k+1} - s_u)M_{k+1,u}(x; s)/(k+1) & \text{si } j - k + 1 \leq l \leq j \\ 1 & \text{si } l < j - k + 1 \end{cases}$$

Cette propriété est importante par la suite car elle montre que dans une combinaison linéaire $F(x) = \sum_{v=1}^{m} a_v I_{k,v}(x; s)$, la fonction $I_{k,v}(x; s)$ n'intervient que localement.

- Considérons l'ensemble $s = (s_l)_{l=1\ldots n}$ de points suivants :

$$s_1 = s_2 = \cdots = s_k = L.$$
$$s_{k+1} < s_{k+2} < \cdots < s_{k+p} < U.$$
$$s_{k+p+1} = s_{k+p+2} = \cdots = s_n = U.$$

On peut définir $m = p + k$ I-splines de degré k non nulles sur $[L, U]$. En effet, sur chaque intervalle de la forme $[s_l, s_{l'}[$, les splines $(M_{k,i})_{i=1}^{l'+k-l-1}$ y sont linéairement indépendantes; en choisissant $l = k$ et $l' = k+p+1$, ce qui revient à prendre $s_l = L$ et $s'_l = U$, le nombre des splines linéairements indépendantes dans cet intervalle est $p + k$; (*cf.* SCHUMAKER L [1981]).

Soit $(a_0, a_1, a_2, \ldots, a_m)$, $m + 1$ coefficients réels. Le modèle proposé consiste à approcher toute courbe $C(t)$ par une combinaison linéaire $\hat{C}(t) = a_0 + \sum_{v=1}^{m} a_v I_{k,v}(t; s)$. On voit que $\hat{C}(t)$ a les propriétés suivantes :

- $t = L \Longrightarrow \hat{C}(t) = a_0$.
- $t = U \Longrightarrow \hat{C}(t) = a_0 + \sum_{j=1}^{m} a_j$.

Si on impose aux coefficients $(a_0, a_1, a_3, \ldots, a_m)$ d'être positifs, $\hat{C}(t)$, combinaisons linéaire à coefficients positifs de fonctions positives croissantes sera croissante. Le modèle sera alors bien adapté à l'approximation d'un phénomène de croissance monotone. En dehors de cette contrainte, on pourra approcher des croissances présentant des régressions locales.

La mise en pratique effective de la méthode demande de déterminer p et k :

De petites valeurs de p et de grandes valeurs de k tendent à augmenter le domaine sur lequel les I-splines ne seront ni nulles, ni constantes. Ces valeurs seraient à exclure pour favoriser l'intervention locale de chacune des I-splines. Or, de manière contradictoire, de petites valeurs de p conduisent à un plus grand nombres d'observations dans chacun des intervalles donc à une meilleure définition des coefficients $(a_0, a_1, a_3, \ldots, a_m)$.

Prenant en compte les remarques de WINSBERG & RAMSAY [1980] et RAMSAY [1988] sur ce point et les ayant confirmées par nos propres expériences, nous avons travaillé dans l'exemple exposé plus loin avec $p = 3$ et $k = 2$.

Les valeurs des trois points intérieurs ont été prises égales aux quartiles de la distribution de l'ensemble des observations faites.

3. Exploration d'un ensemble de courbes

Les données présentées en introduction concernent donc $\mathcal{N} = 18$ courbes de croissances. La courbe C_i est effectivement connue en p_i points $\{t_{i,j} : j = 1, \cdots, p_i\}$.

- ***3.1.*** La première étape de l'étude consiste à substituer à chacun des ensembles $[C_i(t_{i,1}), \ldots, C_i(t_{i,p_i})]$ une courbe $\hat{C}_i(t)$ de la forme $a_0^i + \sum_{v=1}^{m} a_v{}^i I_{k,v}(t; s)$

qui en soit la plus proche au sens des moindres carrés. Les I-splines étant calculées une fois pour toute, le probléme numérique est extrêmement simple à résoudre.

En l'absence de contraintes de positivité sur les coefficients a^i. Si on pose :

(•) $$Y'_i = [C_i(t_{i,1}), \ldots, C_i(t_{i,p_i})].$$

(•) X_i, la matrice $p_i \times (m+1)$ d'éléments :

$$\begin{array}{ll} \forall\, r = 1, \ldots, p_i; & (X_i)_{r,1} = 1 \\ \forall\, l = 2, \ldots, m+1; & (X_i)_{r,l} = I_{k,l-1}(t_i; s) \end{array}$$

(•) $$A'_i = (a_0^i, \ldots, a_m^i)$$

On a, selon les résultats bien connus de l'approximation au sens des moindres carrés :

$$A_i = (X_i' X_i)^{-1} X_i' Y_i$$

Le tableau **1** fournit les $(A_i)_i$ obtenus pour les courbes disponibles. La figure **1** donne le tracé de quelques unes des approximations.

TABLEAU I
Coefficients d'ajustement des courbes

Courbes	*Coefficients*						*Facteurs*		
	a_0	a_1	a_2	a_3	a_4	a_5	C	T	I
C1	8.0968	3.1397	5.2750	10.9259	10.1823	2.6760	1	0	0
C2	8.8439	4.4183	8.9986	6.4515	4.1122	1.0217	1	0	0
C3	8.1287	4.1449	6.3337	8.5820	5.8678	2.7829	1	0	0
C4	8.2452	3.5870	6.0607	8.7370	5.4626	2.2195	1	0	0
C5	8.4402	5.2318	7.9183	13.3593	7.8509	3.7014	1	0	0
C6	8.8939	4.4796	8.1314	12.2212	8.0613	2.8019	1	0	0
T1	10.0169	4.0411	5.3853	7.7109	5.5954	0.8430	0	1	0
T2	8.1208	2.0413	6.5712	12.9450	4.3241	2.0175	0	1	0
T3	8.4585	6.1694	3.8638	10.8994	8.1563	2.6182	0	1	0
T4	9.4010	4.9303	5.1843	9.3145	5.7741	2.1019	0	1	0
T5	8.2905	5.4603	3.6293	10.1130	8.1925	5.3429	0	1	0
T6	8.7202	3.3200	4.8826	9.5680	3.5134	2.3261	0	1	0
I1	8.9759	4.1117	6.9987	9.3839	-0.8819	8.9759	0	0	1
I2	8.5009	3.4182	7.0873	8.3117	6.4890	3.3765	0	0	1
I3	8.3516	2.9909	5.3123	6.9604	4.3484	0.9138	0	0	1
I4	8.7835	6.1096	6.1094	10.9150	6.8514	3.1355	0	0	1
I5	9.3523	4.9675	3.7670	9.0587	8.2340	4.3914	0	0	1
I6	9.3517	4.9707	3.7618	9.0627	8.2304	4.3967	0	0	1

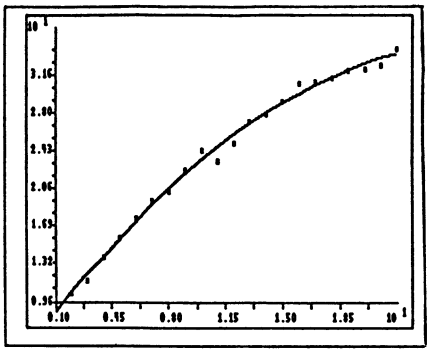


Fig 1.1: Ajustement de la courbe C2.

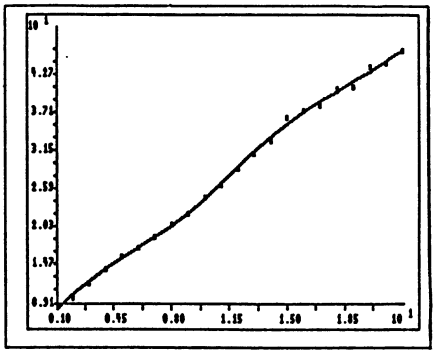


Fig 1.2: Ajustement de la courbe C5.

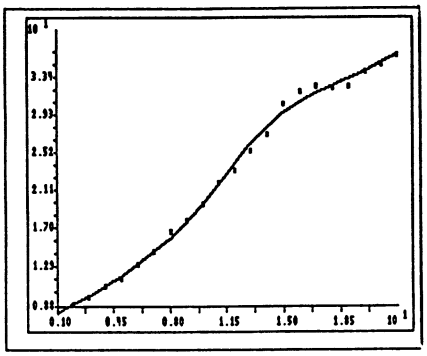


Fig 1.3: Ajustement de la courbe T2.

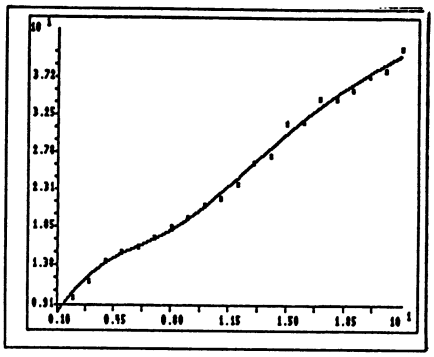


Fig 1.4: Ajustement de la courbe T3.

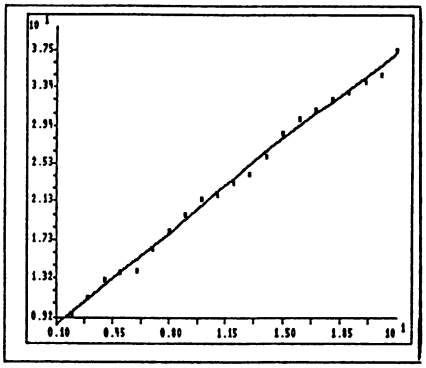


Fig 1.5: Ajustement de la courbe I2.

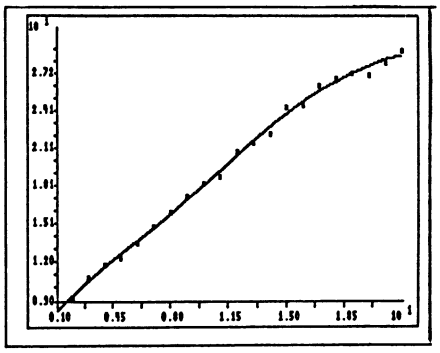


Fig 1.6: Ajustement de la courbe I3.

FIGURE 1

Représentation graphique de quelques courbes

Lorsqu'on introduit les contraintes de positivité sur les élements de A_i, la solution peut être obtenue par des méthodes de pas à pas, par exemple le programme REGCAZ d'ADDAD (*cf.* BERGOUGNAN D & al [1984]), ou encore en utilisant la méthode du gradient projeté (*cf.* MINOUX M [1983]).

• ***3.2.*** La seconde étape de l'étude consiste à comparer les $\hat{C}_i(t)$ entre elles. Choisissant le produit scalaire usuel entre courbes :

$$\langle \hat{C}_{i1}(t).\hat{C}_{i2}(t)\rangle = \int_L^U \hat{C}_{i1}(t)\hat{C}_{i2}(t)dt$$

et posant Q la matrice $(m+1)\times(m+1)$ d'éléments

$$Q_{1,j} = Q_{j,1} = \int_L^U 1_{[L,U]}(t)I_{k,j}(t;s)dt$$

$$Q_{j1,j2} = \int_L^U I_{k,j1}(t;s)I_{k,j2}(t;s)dt$$

on a

$$\langle \hat{C}_{i1}(t)\ .\ \hat{C}_{i2}(t)\rangle = A'_{i1}QA_{i2}$$

• **Propriété :** La matrice Q est définie positive. En effet :

$$A'_iQA_i = \int_L^U [a_0^i + \sum_{j=1}^m a_j^i I_{k,j}(t;s)]^2 dt$$

donc $A'_iQA_i = 0$ implique $a_0^i + \sum_{j=1}^m a_j^i I_{k,j}(t;s) = 0$. On a vu en (**2.2**) que les $(I_{k,v})$ étaient des combinaisons linéaires des $(M_{k,j})_j$ qui sont elles mêmes linéairement indépendantes (*cf.* SCHUMAKER L [1981]), si bien que A'_iQA_i ne peut être nul que si A_i est nul. La conséquence de ce résultat est de permettre la comparaison des courbes observées à travers toute méthode d'analyse des données prenant pour point de départ le tableau Y de ligne $Y_i = A'_i$, et reposant sur une distance entre lignes calculée à partir de la métrique Q. Une matrice D diagonale positive donnera le poids que l'on veut accorder à chaque courbe dans l'analyse.

Le tableau **2** donne la matrice Q obtenue pour les splines calculées. A titre d'exemple nous avons réalisé une classification non hiérarchique des courbes disponibles par la méthode des nuées dynamiques.

Les tableaux **3** et **4**, donnent la constitution des formes fortes avec une hiérarchie réalisée sur les centres de gravité de ces dernières. En coupant l'arbre à un niveau correspondant à la valeur 9 pour l'indice d'agrégation, on peut identifier 4 classes dont les courbes moyennes sont données à la figure **2** .

La classe 1, formée par (FF3,FF7), contient les individus les plus "lourds". La classe 4, formée par (FF8), contient les individus "légers". Les classes 2 et

TABLEAU II
La matrice Q définissant la distance entre courbes

21.000	19.333	15.833	10.500	5.166	1.666
19.333	18.666	15.754	10.500	5.166	1.666
15.833	15.754	14.607	10.452	5.166	1.666
10.500	10.500	10.452	9.216	5.118	1.666
5.166	5.166	5.166	5.118	3.941	1.587
1.666	1.666	1.666	1.666	1.587	5.000

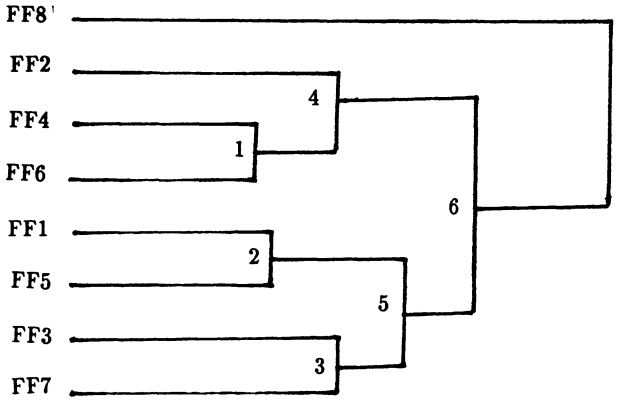
TABLEAU III
Constitution des formes fortes

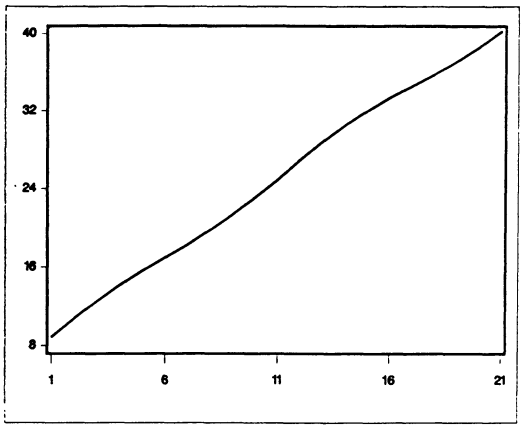
Forme forte N°	*Cardinal*	*Individus*
FF1	5	C1, C3, C4, T2, I2
FF2	3	C2, T6, I6
FF3	2	C5, C6
FF4	1	T1
FF5	2	T3, T5
FF6	1	T4
FF7	3	I1, I4, I5
FF8	1	I3

TABLEAU IV
Hiérarchie des formes fortes (FF)

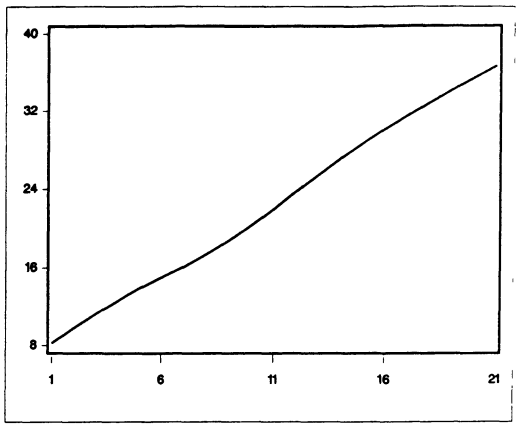
Niveau	*Indice*	*Effectif*	*Effectif ponder*	*Description des classes*
1	5.03	2	0.25	FF4, FF6
2	5.48	2	0.25	FF1, FF5
3	7.20	2	0.25	FF3, FF7
4	7.46	3	0.37	FF2, FF4, FF6
5	9.15	4	0.50	FF1, FF3, FF5, FF7
6	10.94	7	0.87	FF1, FF2, FF3, FF4, FF5, FF6, FF7
7	14.77	8	1.00	FF1, FF2, FF3, FF4, FF5, FF6, FF7, FF8

Présentation de la classification hiérarchique

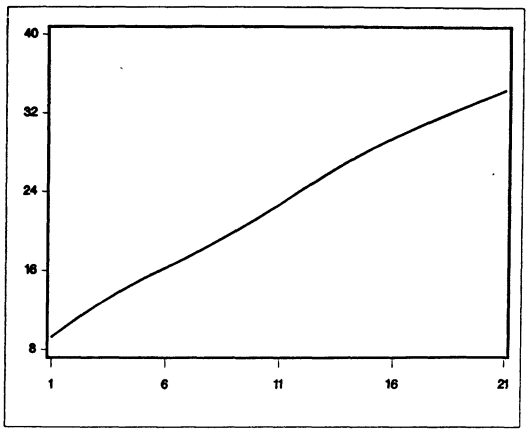




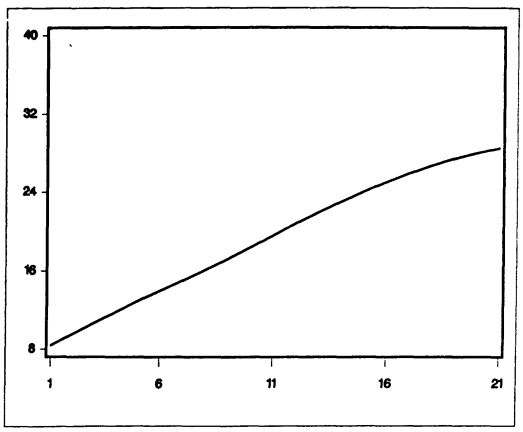
Courbe moyenne de la classe:1



Courbe moyenne de la classe:2



Courbe moyenne de la classe:3



Courbe moyenne de la classe:4

FIGURE 2
Représentation graphique des courbes moyennes.

3 formées respectivement par (FF2,FF4,FF6) et par (FF1,FF5), semblent différer essentiellement par une régression de croissance entre la 6eme et 11eme semaines.

Le tableau **5** redonne les coefficients d'ajustement pour les courbes réorganisées selon leur appartenance aux classes. Cette réorganisation permet de mieux appréhender la similitude entre coefficients pour des courbes appartenant à la même classe.

TABLEAU V
Coefficients d'ajustement des courbes classés suivant les classes

Courbes	*Coefficients*						*Classes*
	a_0	a_1	a_2	a_3	a_4	a_5	
C5	8.4402	5.2318	7.9183	13.3593	7.8509	3.7014	1
C6	8.8939	4.4796	8.1314	12.2212	8.0613	2.8019	1
I1	8.9759	4.1117	6.9987	9.3839	-0.8819	8.9759	1
I4	8.7835	6.1096	6.1094	10.9150	6.8514	3.1355	1
I5	9.3523	4.9675	3.7670	9.0587	8.2340	4.3914	1
C2	8.8439	4.4183	8.9986	6.4515	4.1122	1.0217	2
T1	10.0169	4.0411	5.3853	7.7109	5.5954	0.8430	2
T4	9.4010	4.9303	5.1843	9.3145	5.7741	2.1019	2
T6	8.7202	3.3200	4.8826	9.5680	3.5134	2.3261	2
I6	9.3517	4.9707	3.7618	9.0627	8.2304	4.3967	2
C1	8.0968	3.1397	5.2750	10.9259	10.1823	2.6760	3
C3	8.1287	4.1449	6.3337	8.5820	5.8678	2.7829	3
C4	8.2452	3.5870	6.0607	8.7370	5.4626	2.2195	3
T2	8.1208	2.0413	6.5712	12.9450	4.3241	2.0175	3
T3	8.4585	6.1694	3.8638	10.8994	8.1563	2.6182	3
T5	8.2905	5.4603	3.6293	10.1130	8.1925	5.3429	3
I2	8.5009	3.4182	7.0873	8.3117	6.4890	3.3765	3
I3	8.3516	2.9909	5.3123	6.9604	4.3484	0.9138	4

4. Mise à l'épreuve d'une hypothèse

A une analyse en composantes principales du triplet (Y, Q, D) sera associée l'inertie totale $Tr(YQY'D)$. On sait que cette quantité est une mesure de la variabilité totale des données disponibles.

Des travaux récents de SABATIER R & al [1988], LEBRETON J.D & al [1988], et TERBRAAK C.J.F [1986] ont montré qu'il était possible de décomposer la variabilité mise en évidence par une méthode d'analyse des données et d'essayer de porter un jugement de significativité de chacune de ses parts . Dans cet esprit, toute partition des $\mathcal{N}$ courbes en classes disjointes peut être décrite par la donnée du tableau disjonctif, notons le Z, de ces classes. Soit $P_Z = Z(Z'DZ)^{-1}Z'D$, le

projecteur $D-$orthogonal sur le sous espace de $R^{\mathcal{N}}$ engendré par les colonnes de Z.

Un résultat classique établit que :

$$Tr(YQY'D) = Tr[P_ZYQ(P_ZY)'D]+Tr[(I_{\mathcal{N}\times\mathcal{N}} - P_Z)YQ((I_{\mathcal{N}\times\mathcal{N}} - P_Z)Y)'D]$$

$Tr[P_ZYQ(P_ZY)'D]$ est la part de la variabilité totale expliquée par l'appartenance aux classes. C'est une variabilité interclasse. S'inspirant des approches usuelles en analyse de variance on peut envisager de quantifier la part prise par les classes dans la variabilité par le critère :

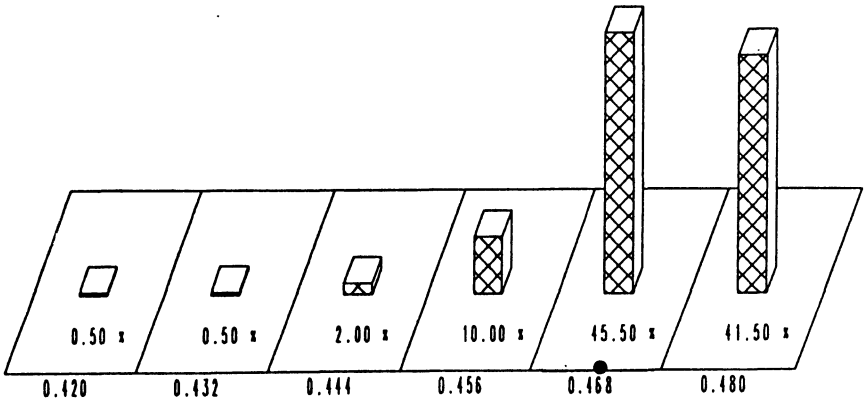
$$S_Z = \frac{Tr[P_ZYQ(P_ZY)'D]}{Tr[YQY'D]}$$

Un test de permutation, consistant à attribuer aléatoirement chaque courbe à l'une des classes permettra d'apprécier la signification du résultat obtenu.

- **Résultat du test :**

On considère les trois groupes initiaux comme trois traitements, la distribution de la valeur test S_Z après 200 affectations au hasard dans ces 3 classes est donnée ci-dessous. On a situé par (•) la valeur du critère pour la répartition initiale en groupes. On voit que cette valeur est dépassée plus de 129 fois par des affectations au hasard, soit dans 64.5 % des cas. On ne peut donc pas considérer qu'il y a un effet significatif des traitements.

Distribution du critère



Références

[1] BESSE P, 1979 "*Etude descriptive d'un processus : Approximation et interpolation*" Thése de 3 ° cycle, U.P.S Toulouse.

[2] BERGOUGNAN D, & CAZES P, & MULLON Ch, 1984 "*Présentation de deux programmes de régression*" In Data analysis and informatics III, Diday & al (Eds), pp 163-176.

[3] BONIFAS L & ESCOUFIER Y & GONZALEZ P.L & SABATIER R, 1984 "*Choix de variables en analyse en composantes principales.*", Rev.Stat.Appl., vol.32, $N°$ 2, pp 5-15.

[4] CURRY H.B & SCHOENBERG I.J, 1946 "*On Polya frequency functions IV :The spline functions and their limits*", Bull. Amer. Math. soc. (Résumé), 53, pp 1114.

[5] CURRY H.B & SCHOENBERG I.J, 1966 "*On Polya frequency functions IV : The fundamental spline functions and their limits*", J. d'analyse Mathématique, 17, pp 71-107

[6] DEBOOR C, 1978 "*A practical guide to splines*" New York Springer-Verlag.

[7] DEVILLE J.C, 1974 "*Méthodes statistiques et numériques de l'analyse harmonique.*" Annales de l'INSEE, 15, pp 3-101.

[8] EL FAOUZI N, 1987 "*Modélisation des courbes de croissance par les fonctions splines : Estimation et calcul de distance entre courbes.*" DEA, USTL, Montpellier.

[9] ESCOUFIER Y, 1977 "*Operator related to data matrix in recent development in statistics*" J.R. Barra et al Editors, North-Holland, Publishing company, pp 125-131.

[10] FLETCHER R, 1981 "*Practical methods of optimization 2*", Wiley New York.

[11] FREIDMAN H & SILVERMAN B.W, 1989 "*flexible parsimonious and additive modeling*", Technometrics, (avec discussion), vol 31, N°1, pp 1-40.

[12] 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, pp 17-36.

[13] LAWSON CL & HANSON RJ, 1974 "*Solving least squares problems*", Prentice-hall.

[14] LEBRETON J.D & CHESSEL D & PRODON R & YOCCOZ N, 1988 "*L'analyse des relations espèces milieu par l'analyse canonique des correspondances -1- variables de milieu quantitatives.*" Acta oecologica - Oecologica Generalis, 9, pp 53-67.

[15] LENOUVEL J , 1981 "*Etude d'une famille de courbes par des méthodes d'anlyses de données : Application à l'analyse morphologique des courbes provenant de données médicales*" Thése de 3° cycle, Université de Rennes I.

[16] LIBERT G & DUPUIS Ch, 1981 "*Comparaison de courbes de thermoluminiscence de quartz par l'analyse des coefficients d'autocorrélation*", Rev. Stat. Appl., vol.29, $N°$ 4, pp 51-59.

[17] MINOUX M, 1983 "*Programmation mathématique*" Tome 1, Dunod.

[18] RAMSAY J.O, 1982a "*Some statistical approches to Multidimentional Scaling Data*", J.R.S.S ,145, part 3, pp 285-312.

[19] RAMSAY J.O, 1982b "*When data are functions*", Psychometrika, vol.47, pp 379-396.

[20] RAMSAY J.O, 1988 "*Monotone régression splines in action*", Statistical Sciences Vol 3, N °4, pp 425-461

[21] RAO C, 1964 "*The use and interpretation of principal component analysis in applied research*" Sankya, Ser. A, 26, pp 329-359.

[22] ROBERT P & ESCOUFIER Y, 1976 "*A unifying tool for linear multivariate statistical methods : The Rv-coefficient*" App. Stat, 15, 3, pp 257-265.

[23] SABATIER R & LEBRETON J.D & CHESSEL D, 1988 "*Principal component analysis with instrumental variables as a tool for modelling composition data*" In Multiway Data Analysis, pp. 341-352, Coppi & al (Eds), Amsterdam.

[24] SAIH A, 1985 "*Analyse et comparaison des courbes de croissance : Méthodes et programmes.*", Thèse de 3° cycle, USTL Montpellier.

[25] SAPORTA G , 1981 "*Méthodes exploratoires d'analyses des données temporelles* " Thèse d'etat, U.M.C Paris VI.

[26] SCHUMAKER L, 1981 "*Spline functions : Basic theory*", Wiley Interscience.

[27] TER BRAAK C, 1986 "*The analysis of vegetation-environement relationship by canonical correspondence analysis*", Vegetatio 69, pp 69-77.

[28] WINSBERG S & RAMSAY J.O, 1980 "*Monotonic transformations to additivity using splines*", Biometrika 67, pp 669-676.

[29] WINSBERG S & RAMSAY J.O, 1981 "*Analysis of pairwise preferences data using integrated B-splines.*" Psychometrika, 46, pp 171-186.

[30] WRIGHT W & WEGMAN E, 1980 "*Isotonic convex and related splines*", annals of statistics, 8, pp 1023-35.

[31] WRIGHT W & WEGMAN E, 1983 "*Splines in statistics*" J.A.S.A , 78, pp 351-65.

[32] WOLD S, 1974 "*Spline functions in data analysis*" Technometrics, vol.16, pp 1-11.