REVUE DE STATISTIQUE APPLIQUÉE

H. ABIDI
J. PONTIER
J. BORMS
W. DUQUET

Courbes de croissance : intérêt de la modélisation pour l'analyse de données longitudinales

Revue de statistique appliquée, tome 43, n° 3 (1995), p. 55-72
<http://www.numdam.org/item?id=RSA_1995__43_3_55_0>

Rev. Statistique Appliquée, 1995, **XLIII** (3), 55-72

COURBES DE CROISSANCE : INTÉRÊT DE LA MODÉLISATION POUR L' ANALYSE DE DONNÉES LONGITUDINALES

H. Abidi (1), J. Pontier (2), J. Borms et W. Duquet (3)

(1) Unité d'Hygiène, Épidémiologie et Information Médicale
Centre Hospitalier Lyon Sud
69495 Pierre Bénite cedex, France

(2) Centre de Recherche et d'Innovation sur le Sport
UFR STAPS Université Claude Bernard Lyon I
69622 Villeurbanne cedex France

(3) Laboratoire de Biométrie Humaine et de Biomécanique
Vrije Universiteit Brussel,
Pleinlaan 2 1050 Brussel, Belgique

RÉSUMÉ

L'analyse des courbes de croissance a en général pour objectif de visualiser une typologie de courbes afin de déceler les groupes d'individus ayant une croissance particulière. L'Analyse en Composantes Principales (ACP), comme méthode d'analyse factorielle, semble bien adaptée à ce genre de problème. Mais cette analyse n'est pas possible si les individus ne sont pas mesurés aux mêmes âges, ou si des mesures sont manquantes. Dans ces conditions, nous montrons comment par l'intermédiaire de la modélisation on peut atteindre le même objectif.

Mots-clés : *Courbes de croissance, modèles non linéaires, modélisation, analyse en composantes principales.*

ABSTRACT

In general a main objective of the study of growth curves is to show a typology of curves for the detection of individuals having a particular growth. The Principal Component Analysis (PCA) is generally applied in order to solve this problem. When individuals are not measured at the same age or some measures are missing, this analysis is not possible. Then we show that we can attain this objective by means of the modelization of the growth curves.

Keywords : *Growth curves, nonlinear models, modelization, principal component analysis.*

1. Introduction

L'évolution individuelle d'un caractère quantitatif variant au cours du temps (la taille, le poids, etc.) est connue, de manière très partielle, grâce à une suite finie

de couples (t_i, y_i) de valeurs numériques croissantes en t_i (l'âge, en général), y_i étant la valeur mesurée à l'instant t_i du caractère étudié sur l'individu considéré. La représentation graphique de cette suite de couples, dans un plan muni d'un repère cartésien, est une succession de points jalonnant la «courbe de croissance», continue, mais inconnue faute d'un enregistrement permanent. De telles données, résultant de plusieurs observations échelonnées au cours du temps et réalisées sur un même individu, sont qualifiées de *longitudinales*.

Généralement, une campagne d'acquisition de telles données longitudinales est réalisée sur plusieurs individus, l'objectif étant d'étudier la variabilité inter-individuelle des courbes de croissance, par exemple aux fins de typologie, de détection de comportements anormaux, voire de prédiction. Si, d'un individu à l'autre, les conditions d'acquisition des données sont standardisées (tous les individus sont mesurés le même nombre de fois, et exactement aux mêmes âges), alors les méthodes traditionnelles d'analyse de données peuvent être utilisées directement pour répondre aux questions posées : les données d'observation sont récapitulées dans un tableau Y de valeurs réelles y_{ij}, dont la ligne j correspond à l'individu numéro j, et dont la colonne i correspond à la valeur du caractère y mesurée à l'âge t_i. Dès lors, des méthodes comme l'analyse en composantes principales ou l'analyse discriminante peuvent être utilisées sans problème (Estève et Schifflers 1976, Pernin 1986).

Dans la pratique, s'agissant de l'acquisition de données longitudinales sur des enfants, cette situation idéale est loin d'être toujours réalisée. En effet, même si cette acquisition a été soigneusement planifiée, c'est-à-dire si les instants d'observation prévus sont identiques pour tous les individus, la contrainte imposée aux individus par cette planification (qui peut s'étaler sur une vingtaine d'années, durée totale de la croissance chez l'enfant) peut se révéler insupportable. D'où des défaillances ponctuelles (ce qui se traduit par des données manquantes) ou définitives (abandon en cours d'enquête : séries incomplètes). Le taux de données manquantes peut être si élevé que les techniques habituelles d'estimation de ces données manquantes sont impuissantes à remédier à cette situation.

Ainsi, le concept même d'observation longitudinale de longue durée, sur la croissance de l'être humain, se prête très mal à une planification. Ce «défaut» ne diminue en rien l'intérêt de réaliser une telle étude. Aussi, renonçant à imposer aux individus une standardisation trop contraignante des conditions de leur observation, nous reportons notre effort de standardisation sur les données une fois acquises, c'est-à-dire *après* que les observations aient été réalisées, mais *avant* toute application d'une méthode classique d'analyse de données. Cela signifie que nous supposons seulement :

- que chaque individu a été observé plusieurs fois au cours de sa croissance,
- que les instants (âges) d'observation ne sont pas forcément les mêmes d'un individu à l'autre,
- que les nombres d'observations peuvent différer d'un individu à l'autre.

Ces conditions d'observation sont beaucoup moins contraignantes qu'une planification rigide, et tout à fait compatibles avec la vie habituelle des enfants.

Pour réaliser la standardisation souhaitée, nous proposons d'utiliser la modélisation des courbes de croissance. Une telle modélisation permet de reconstituer une situation standard, c'est-à-dire un tableau complet lignes × colonnes, de deux façons :

- chaque courbe individuelle étant modélisée, le modèle correspondant permet de reconstituer (approximativement) les valeurs du caractère à p instants convenus, les mêmes pour tous les individus;
- chaque courbe étant modélisée, la liste des valeurs des q paramètres caractérise d'une certaine manière l'ensemble de cette courbe.

Dans les deux cas, on associe à chaque individu un «profil» numérique standard, et par conséquent l'ensemble de ces profils constitue un tableau complet susceptible d'être traité par une méthode classique d'analyse d'un tableau de données. Cependant, face à deux réponses possibles à notre problème de standardisation, nous nous interrogeons tout naturellement sur les relations susceptibles d'exister entre les deux. Plus précisément, ces deux analyses, celle sur le tableau des données reconstituées, et celle sur le tableau des paramètres, produisent-elles des résultats entièrement redondants (auquel cas l'une des deux suffit), ou complémentaires (alors il peut être intéressant de les réaliser toutes les deux). C'est l'objet du texte qui suit, de tenter d'apporter une réponse à cette question, à la fois sur un plan théorique et sur un plan appliqué à un cas concret.

2. Un problème de métrique

Nous supposons que la modélisation proprement dite ne pose pas problème. Cela signifie que nous disposons d'un «bon» modèle, c'est-à-dire d'une fonction continue $t \rightarrow f(t, \theta)$ (où θ est un ensemble de q valeurs numériques $\theta_1, \theta_2, \ldots, \theta_q$, constantes pour un individu donné, mais variables d'un individu à l'autre; θ est le «vecteur des paramètres»), telle que pour tout individu j et tout instant d'observation t_i, la valeur «théorique» $f(t_i, \theta_j)$ soit une excellente approximation de la valeur observée (ou observable) y_{ij}. Dans ces conditions, l'identification du modèle pour chaque individu (c'est-à-dire la détermination du vecteur θ^j correspondant à l'individu j) nous conduit à construire deux tableaux, chacun de n lignes (le nombre d'individus), l'un noté Y ayant p colonnes (les valeurs du caractère étudié y estimées par le modèle à p instants choisis t_i, $i = 1$ à p), l'autre noté P ayant q colonnes (les valeurs des q paramètres constituant le vecteur θ).

Ayant choisi ici l'analyse en composantes principales (A.C.P.) comme moyen d'explorer l'ensemble des n courbes de croissance, notre objectif est alors de faire en sorte qu'une A.C.P. réalisée sur le tableau P, et une A.C.P. réalisée sur le tableau Y, produisent des résultats «les plus proches» possible. Plus précisément, cette proximité des résultats devra se traduire en termes de proximité des nuages de points, respectivement dans l'espace engendré par les composantes principales de l'une et de l'autre de ces deux A.C.P. (cartes factorielles). Les métriques associées à ces deux A.C.P. sont notées respectivement M_p et M_q. Notre problème peut alors s'énoncer ainsi : l'une des deux métriques étant choisie (arbitrairement), comment construire l'autre métrique pour réaliser la proximité souhaitée?

Le problème du choix de M_q a été soulevé par Houllier (1986), Pernin (1986). Houllier (1987) l'avait résolu dans un contexte mathématique rigoureux, imposant à la métrique M_p de R^p (espace des données d'observation) d'être égale à l'Identité ($M_p = I_p$). Il a ainsi abouti à la *métrique de sensibilité*. Caussinus et Ferré (1989) ont donné une justification du choix de Houllier, en considérant le problème comme une application intéressante du modèle à effet fixes (Caussinus 1986a, 1986b, Besse *et al.* 1987, 1988, Ferré 1989). Dans ce paragraphe nous proposons une démonstration plus simple et plus générale à ce problème. Il est tout d'abord nécessaire de préciser quelques notations et hypothèses.

2.1 Notations

. Y^j note la matrice colonne transposée de la ligne j du tableau Y; ses coefficients sont les p mesures y_{ij} de l'individu j (c'est aussi la «courbe» observée ou reconstituée de cet individu).

. $f(t, \theta)$ note le modèle de croissance; il dépend du temps t, et du vecteur des paramètres θ, de dimension q.

. $F(\theta)$ matrice colonne dont les p coefficients sont les valeurs $f(t_i, \theta)$ du modèle aux âges (dates) t_i.

. θ^j : matrice colonne transposée de la ligne j du tableau P; ses coefficients sont les valeurs estimées des q paramètres de la courbe de croissance de l'individu j.

. $\overline{\theta}$: matrice moyenne des θ^j.

2.2 Hypothèses

Nous conviendrons que l'espace des variables (R^n) est muni de la métrique identité, cas auquel on peut toujours se ramener.

Hypothèse H_0 : Les âges d'observation sont les mêmes pour tous les individus, ce qui permet de calculer et de noter $\overline{Y}$ le vecteur moyen des Y^j (courbe moyenne des n courbes de croissance).

Hypothèse H_1 : La qualité de l'ajustement du modèle aux données observées est jugée satisfaisante (par différents moyens de validation du modèle).

Hypothèse H_2 : $f(t, \overline{\theta})$ est une approximation acceptable de la courbe théorique de la courbe moyenne $\overline{Y}$. Cela implique que les paramètres du modèle sont dans un domaine convexe.

Hypothèse H_3 : Les courbes observées ne sont pas trop dispersées par rapport à la courbe moyenne, ou mieux encore, les paramètres moyens sont suffisamment proches de ceux de tous les individus. Ceci permettra d'approcher la quantité $(F(\theta^j) - F(\overline{\theta}))$ par son développement limité au premier ordre au voisinage des paramètres moyens (tous les individus sont dans un domaine où le modèle est presque linéaire par rapport à ses paramètres).

A noter que si le modèle est linéaire par rapport à ses paramètres (exemple : modèles polynomiaux), les deux dernières hypothèses sont automatiquement vérifiées.

2.3 Construction de la métrique M_q à partir de la métrique M_p

Si $d^2(Y^j, Y^{j'})_{M_p}$ note le carré de la distance entre les individus (ou courbes) j et j' au sens de la métrique M_p, alors :

$$d^2(Y^j, Y^{j'})_{M_p} = {}^t(Y^j - Y^{j'})Mp(Y^j - Y^{j'})$$

On peut écrire la matrice $(Y^j - Y^{j'})$ sous la forme :

$$(Y^j - Y^{j'}) = (Y^j - \overline{Y}) - (Y^{j'} - \overline{Y})$$

d'où $(Y^j - Y^{j'}) \approx (F(\theta^j) - F(\overline{\theta})) - (F(\theta^{j'}) - F(\overline{\theta}))$ (hypothèses H_1 et H_2)

L'hypothèse H_3 permet l'approximation :

$(F(\theta^j) - F(\overline{\theta})) \approx \overline{S}(\theta^j - \overline{\theta})$, où $\overline{S}$ est la matrice à p lignes et q colonnes connue sous le nom de *matrice de sensibilité*, d'élément courant $\overline{S}_{ik} = \dfrac{\partial f(t_i, \overline{\theta})}{\partial \theta_k}$. La notation $\overline{S}$ rappelle que cette matrice est calculée sur le vecteur des paramètres moyens $\overline{\theta}$. Après remplacement et simplification dans l'expression précédente on trouve :

$(Y^j - Y^{j'}) \approx \overline{S}(\theta^j - \theta^{j'})$ et par conséquent :

$$d^2(Y^j, Y^{j'})_{M_p} \approx {}^t(\theta^j - \theta^{j'}){}^t\overline{S}M_p\overline{S}(\theta^j - \theta^{j'}) = d^2(\theta^j, \theta^{j'})_{{}^t\overline{S}M_p\overline{S}}$$

où $d^2(\theta^j, \theta^{j'})_{{}^t\overline{S}M_p\overline{S}}$ note le carré de la distance entre θ^j et $\theta^{j'}$ selon la métrique ${}^t\overline{S}M_p\overline{S}$. Il suffit donc, pour respecter approximativement les inter-distances des individus, de choisir $M_q = {}^t\overline{S}M_p\overline{S}$ comme matrice de la métrique dans l'espace des paramètres.

Si le modèle est linéaire par rapport aux paramètres, les approximations ($\approx$) figurant dans les relations ci-dessus deviennent des égalités, et ainsi les inter-distances des individus sont identiques, qu'elles soient mesurées avec l'une ou l'autre des deux métriques. On s'attend alors à ce que les deux A.C.P. donnent les mêmes résultats, alors qu'on peut s'attendre à des résultats plus ou moins différents si le modèle n'est pas linéaire.

2.4 Cas particuliers intéressants : $M_p = I_p$ et $M_p = \mathrm{Diag}(1/\mathrm{Var}\,(t_i))$

Un premier cas particulier de la relation précédente est celui où $M_p = I_p$, métrique identité définie dans R^p (ce qui correspond à une A.C.P. sur matrice des covariances). Dans ce cas la métrique M_q de R^q sera égale à ${}^t\overline{S}\,\overline{S}$, notée Q et appelée métrique de sensibilité. Son nom découle de celui de la matrice S définie précédemment. Le terme général de Q est donc :

$$Q_{kh} = \sum_{i=1}^{p} \left[\frac{\partial f(t_i, \overline{\theta})}{\partial \theta_k} \frac{\partial f(t_i, \overline{\theta})}{\partial \theta_h} \right]$$

Il est somme finie du produit deux à deux de fonctions de sensibilité $\frac{\partial f(t, \overline{\theta})}{\partial \theta_k}$, dérivées du modèle par rapport aux paramètres $\theta_k (k = 1 \text{ à } q)$. Ces fonctions jouent un rôle crucial lors de l'identification des paramètres (Bard 1974, Beck et Arnold 1977, Abidi 1991). La métrique de sensibilité tient compte de l'effet de la moyenne d'un paramètre sur la forme de la courbe, du nombre d'observations et surtout de leur répartition sur la période de croissance. Caussinus et Ferré (1989) ont préféré une métrique moyenne C (compromis) des Q_j, où Q_j est la matrice Q définie ci-dessus mais dans laquelle le vecteur moyen $\overline{\theta}$ est remplacé par le vecteur θ^j associé à la courbe j. Cette métrique a été signalée auparavant par Houllier (1987), et Abidi (1991) a obtenu sur un exemple les mêmes résultats en utilisant ces deux métriques. Notons que si le modèle est linéaire par rapport à ses paramètres, C et Q sont identiques.

Un autre cas particulier est celui où $M_p = \text{Diag}\,(1/\text{Var}\,(t_i))$, où $\text{Var}\,(t_i)$ est la variance (entre individus) de la taille à l'âge t_i. Ce cas, classique, correspond à une analyse en composantes principales centrée réduite, ou sur matrice de corrélation, dite aussi analyse en composantes principales normée (A.C.P.N.). Notons D la matrice associée à cette métrique. Compte tenu du développement du paragraphe précédent, la métrique M_q à utiliser dans l'espace des paramètres est donc égale à ${}^t\overline{S}D\overline{S}$. Cette expression peut s'écrire sous la forme ${}^t(\sqrt{D}\ \overline{S})(\sqrt{D}\ \overline{S})$ où $\sqrt{D} = \text{Diag}(1/\sqrt{\text{Var}\,(t_i)})$. Sous cette forme, il est facile de remarquer que la réduction des variables (âge t_i) du tableau Y, produit une transformation de la métrique de sensibilité : pondération des fonctions de sensibilité $\frac{\partial f(t_i, \theta)}{\partial \theta_k}$ par l'inverse de l'écart type de la taille à l'âge t_i (cf. Abidi 1991).

Dans le paragraphe suivant, nous appliquerons à un cas concret d'observation de plusieurs croissances staturales individuelles, la modélisation des courbes de croissance et les deux analyses en composantes principales évoquées ci-dessus : sur paramètres du modèle, et sur valeurs reconstituées de la stature. Nous discuterons les relations entre résultats de ces deux analyses. Nous avons choisi de nous placer dans le premier des deux cas particuliers évoqués ci-dessus, soit celui où M_q est la métrique de sensibilité Q, et par conséquent M_p est la métrique identité I_p (donc A.C.P. sur matrice de covariances des données reconstituées).

3. Application

3.1. Les données longitudinales et leur modélisation

Nous avons travaillé sur un jeu de données concernant la croissance en taille debout (stature) de 59 enfants belges, soit 35 garçons et 24 filles. Ces enfants font partie de l'enquête LEGS (Hebbelinck *et al.* 1980). Ils ont été suivis au cours du temps entre l'âge de 6 ans et l'âge de 19 ans. Les observations ne sont pas toutes effectuées aux mêmes âges, et les 59 enfants n'ont pas non plus le même nombre de mesures, mais pour chacun on dispose d'au moins 21 mesures réparties de façon plus au moins régulière entre 6 et 19 ans. L'âge est exprimé en années et la taille en centimètres.

Sur cette période de la croissance du caractère taille, plusieurs modèles sont disponibles (voir Hauspie 1989, Abidi 1991). Nous avons choisi ici le premier des

trois modèles proposés par Preece et Baines (1978). Ce modèle (noté PB1) dépend de 5 paramètres qu'on notera θ_1, θ_2, θ_3, θ_4, θ_5 ; il a pour expression :

$$f(t,\theta) = \theta_1 - 2\left\{\frac{\theta_1 - \theta_2}{\mathrm{Exp}\,[\theta_3(t-\theta_5)] + \mathrm{Exp}\,[\theta_4(t-\theta_5)]}\right\}$$

où le paramètre θ_1 estime la taille adulte, θ_5 et θ_2 estiment respectivement l'âge et la taille au pic d'adolescence, θ_3 et θ_4 sont des paramètres d'échelle. Une étude détaillée sur ce modèle a été effectuée par Pernin (1986).

Le critère à optimiser est celui des moindres carrés ordinaires, la méthode d'identification utilisée ici (calcul numérique des paramètres) est celle de Marquardt (Marquardt 1963). Plusieurs critères peuvent être pris en considération pour apprécier la qualité de l'ajustement d'une courbe. Nous retenons l'Écart Quadratique Moyen (EQM) et la Variance Résiduelle (Var$R$) :

$$\mathrm{EQM} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left[y_i - f(t_i,\widehat{\theta})\right]^2} \qquad \mathrm{Var}R = \frac{1}{p-q}\sum_{i=1}^{p}\left[y_i - f(t_i,\widehat{\theta})\right]^2$$

Dans ces formules, le nombre de paramètres est q, ici égal à 5 pour tous les individus, le nombre de valeurs observées sur l'individu est p, ici variable d'un individu à l'autre, y_i désigne la mesure de la taille de l'individu (en cm) à l'instant t_i (en années) et $f(t_i,\widehat{\theta})$ la valeur théorique de la taille en cet instant t_i d'après le modèle. L'avantage du premier critère est qu'il est indépendant du nombre de paramètres et s'exprime par l'unité de mesure du caractère étudié (ici le cm). Quant au deuxième critère, il a l'avantage d'être une estimation non biaisée de la variance des résidus (ε_i) par rapport au modèle :

$$y_i = f(t_i,\,\theta) + \varepsilon_i.$$

Dans la figure 1 ci-dessous sont représentés les écarts quadratiques moyens pour tous les individus. On notera que tous ces écarts sont inférieurs à 1 cm, ce qui est l'ordre de grandeur de l'erreur de mesure lors de la prise de la taille d'un enfant debout, et ce qui par conséquent permet de considérer que le modèle choisi est très bon (voir hypothèse H_1).

Le tableau 1 suivant donne une idée de la valeur moyenne des paramètres par sexe, et sexes confondus, et de leur dispersion exprimée par l'écart type. Les deux dernières lignes sont relatives aux valeurs moyennes et aux dispersion de l'Écart Quadratique Moyen (EQM) et de la variance résiduelle (Var$R$). Les faibles valeurs de ces deux quantités montrent la grande qualité du modèle choisi dans l'ajustement de ces courbes observées, que ce soit celles des filles ou celles des garçons.

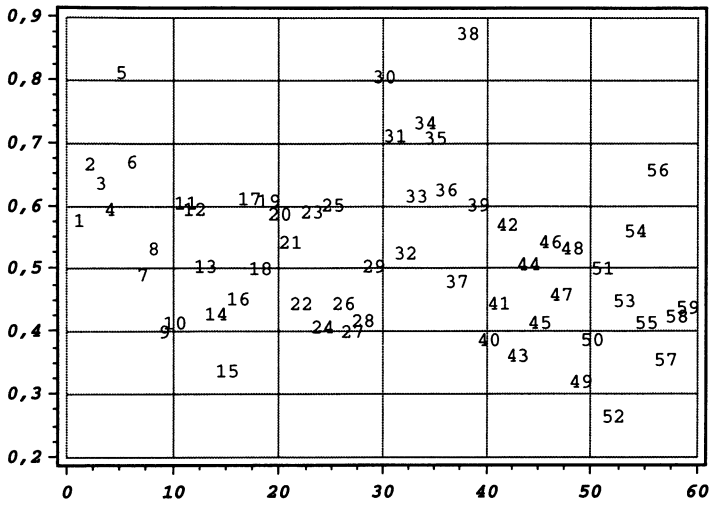

FIGURE 1 : *En ordonnée, valeur de l'écart quadratique moyen (EQM) exprimée en cm. En abscisse sont portés les numéros des courbes : garçons n°1 à 35, filles n°36 à 59.*

TABLEAU 1 :
Moyenne et écart type (entre parenthèse) par sexe et sexes confondus des 5 paramètres du modèle PB1 et des quantités EQM (écart quadratique moyen) et VarR (variance résiduelle).

Paramètres	Moy. et écart type des 35 Garçons	Moy. et écart type des 24 filles	Moy. et écart type des 59 enfants
θ_1	179,983 (6,638)	164,486 (5,887)	173,679 (9,909)
θ_2	167,476 (6,505)	153,179 (5,727)	161,660 (9,369)
θ_3	0,112 (0,010)	0,125 (0,017)	0,117 (0,015)
θ_4	1,274 (0,154)	1,189 (0,128)	1,240 (0,150)
θ_5	14,462 (0,964)	12,517 (1,114)	13,671 (1,403)
EQM (en cm)	0,560 (0,117)	0,486 (0,125)	0,530 (0,126)
Var R (en cm^2)	0,426 (0,117)	0,330 (0,185)	0,387 (0,187)

3.2. A.C.P. sur les données reconstituées

Une fois les 59 courbes modélisées, nous avons créé le tableau Y contenant, pour chacun des 59 enfants, les valeurs, estimées par le modèle, de la stature à 27 âges choisis, soit tous les 6 mois de l'âge de 6 ans à l'âge de 19 ans, inclus. Ayant choisi, dans l'analyse des paramètres (voir ci-après § III.3) de nous placer dans le cas particulier où l'espace R^q des individus est muni de la métrique de sensibilité Q, nous devons munir R^p (espace des courbes de croissance) de la métrique identité (I_p). En d'autres termes, l'analyse en composantes principales sur le tableau Y sera centrée seulement (A.C.P. sur matrice des covariances).

Le résultat de cette analyse est schématisé par le premier plan factoriel des courbes de croissance (reconstituées), figure 2 ci-après. Ce dernier plan représente 96,27% d'inertie totale, dont 80,75% est dû au premier axe et 15,52% au second.

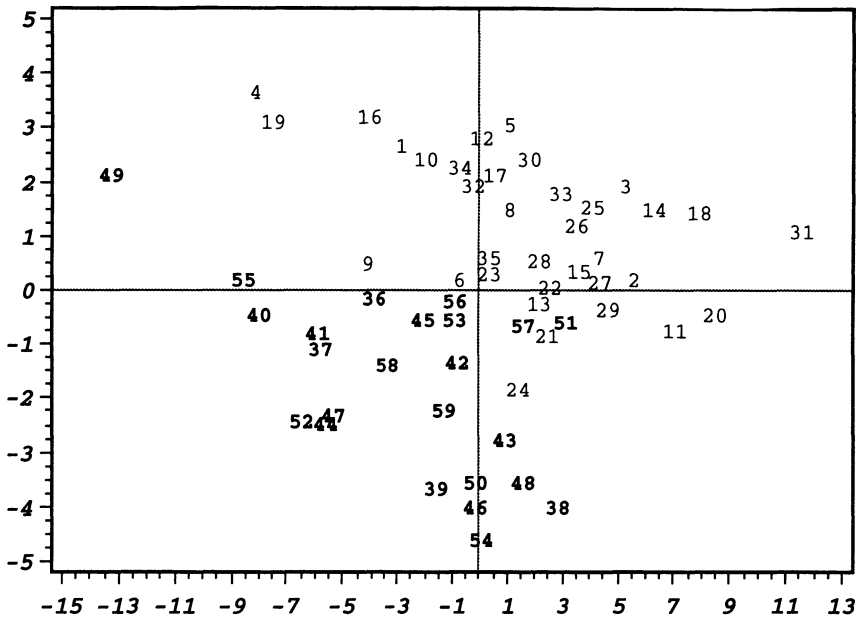

FIGURE 2 : *Premier plan factoriel de l'A.C.P. sur matrice des covariances des données reconstituées par le modèle PB1. Les garçons sont numérotés de 1 à 35, les filles de 36 à 59. Ce plan représente 96,27% d'inertie totale dont 80,75% est dû au premier axe et 15,52% au second.*

Le cercle des corrélations des 27 variables (âges de 6 ans à 19 ans tous les 6 mois) du tableau Y, avec les 2 premières composantes principales de l'A.C.P. sur matrice des covariances est représenté figure 3. Pour alléger le dessin, nous n'avons explicité que 5 points (correspondant à 6, 9, 13, 15 et 19 ans). Les points non représentés sont sur la courbe tracée les reliant tous dans l'ordre chronologique. Cette courbe part du point «6 ans», monte jusqu'à «9 ans», redescend à «13 ans», enfin monte jusqu'à «19 ans» en passant par l'intermédiaire «15 ans».

L'interprétation des deux premières composantes principales, à partir du cercle des corrélations de la figure 3, ne paraît pas évidente si l'on veut l'exprimer à partir des «variables», c'est-à-dire en termes de valeurs de la taille aux différents âges. La première composante principale s'identifie presque (au signe près) à la taille à l'âge de 15 ans. La deuxième composante principale met en évidence une sorte d'oscillation au cours de laquelle les âges 9 ans et 13 ans paraissent jouer des rôles de «charnières», découpant ainsi la durée de croissance en périodes dont l'interprétation ne peut résulter que d'un retour aux données, peut-être d'un examen plus approfondi de la forme des courbes de croissance (nous pensons à des périodes d'accélération ou de décélération par exemple). Quoi qu'il en soit, arrivés au stade de ce cercle de corrélation, nous ne sommes pas capables sans information complémentaire de fournir une interprétation des positions des individus dans la carte factorielle de la figure 2. Tout au plus remarque-t-on une séparation nette entre les deux sexes, mais ceci n'a rien d'original, ce phénomène apparaissant dans la plupart des analyses de caractères biométriques dans lesquelles figurent des filles et des garçons.

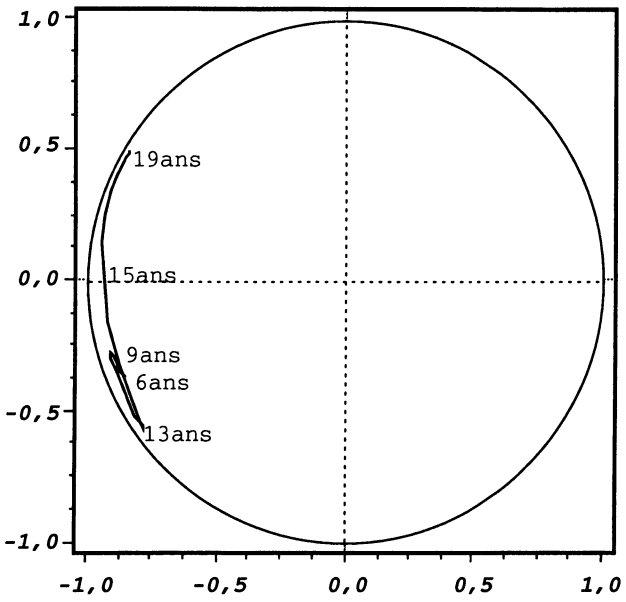

FIGURE 3 : *Corrélations des variables (âges) du tableau Y avec les deux premières composantes principales de l'A.C.P. sur matrice des covariances. Voir commentaire dans le texte.*

3.3. A.C.P. sur les 5 paramètres du modèle $PB1$

L'étape de modélisation nous a permis d'obtenir le tableau P à n lignes (ici, les 59 courbes de croissance) et q colonnes (ici, les 5 paramètres du modèle PB 1). Comme précisé plus haut (§ II,4 et § III,2), nous avons choisi pour des raisons de simplicité de munir l'espace R^q des individus de la métrique de sensibilité Q. Sur

le tableau P des valeurs des paramètres, nous avons donc réalisé une analyse en composantes principales en utilisant cette métrique. Sur la figure 4 on a représenté le premier plan factoriel des individus (courbes de croissance). Ce plan représente 97,77% de l'inertie totale, soit 75,80% pour le premier axe, et 21,97% pour le second. Comme dans l'analyse précédente, nous notons immédiatement une séparation des deux sexes.

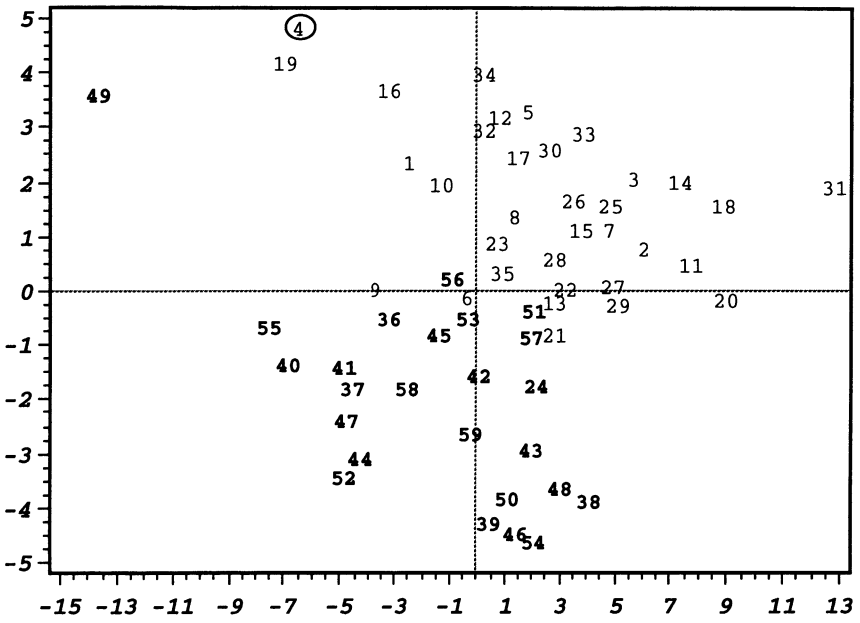

FIGURE 4 : *Premier plan factoriel de l'A.C.P. sur les 5 paramètres du modèle $PB1$ avec métrique de sensibilité. Les garçons sont numérotés de 1 à 35, les filles de 36 à 59. Ce plan représente 97,77% d'inertie dont 75,80% pour le premier axe et 21,97% pour le second. L'enfant n°4 sort des limites du graphique, avec 7,5 pour ordonnée.*

Une comparaison visuelle de la disposition des individus (courbes de croissance) dans les deux plans factoriels (figures 2 et 4), nous montre déjà que ces deux représentations sont très proches l'une de l'autre. Le calcul des corrélations entre les composantes principales issues des deux analyses confirme cette remarque, En effet si $C1$ et $C2$ notent les deux premières composantes principales issues de l'A.C.P. centrée sur les courbes reconstituées, et $C'1$ et $C'2$ celles qui découlent de l'A.C.P. sur les paramètres, on a $\text{cor}(C1, C'1) = 0,99684$ et $\text{cor}(C2, C'2) = 0,96596$, valeurs suffisamment élevées pour que l'on puisse considérer que $C'1$ est approximativement proportionnel à $C1$, $C'2$ est approximativement proportionnel à $C2$. Ainsi, à un éventuel facteur d'échelle près sur chacun des deux axes, les configurations des deux nuages de points dans ces deux plans factoriels sont presque en coïncidence.

Le cercle de corrélation (figure 5) correspondant à la carte factorielle de la figure 4, permet une interprétation concrète, biologique, des positions des individus sur cette carte factorielle. En effet, la première composante principale (axe horizontal),

en corrélation positive élevée avec θ_1 (taille adulte) et θ_2 (taille au moment du pic d'adolescence), est un facteur de taille. Cette composante oppose les enfants de petite taille (à gauche dans la carte factorielle) à ceux de grande taille (à droite). La deuxième composante principale (axe vertical) s'identifie presque au paramètre θ_5 (âge au moment du pic d'adolescence). Elle oppose les enfants ayant un pic d'adolescence précoce (en bas dans la carte factorielle) à ceux qui l'ont plus tardivement (en haut).

Remarque. L'examen du cercle des corrélations de la figure 5 permet de se rendre compte de la forte corrélation existant entre les paramètres θ_1 et θ_2 (cette corrélation est égale à 0,988). Sur le plan de la technique de modélisation, ces deux paramètres peuvent être considérés comme redondants, situation qui pourrait justifier si besoin était une reparamétrisation du modèle. Sur le plan biologique, cette forte corrélation est intéressante dans la mesure où elle autorise une bonne estimation de la taille adulte à partir de la taille au moment du pic d'adolescence.

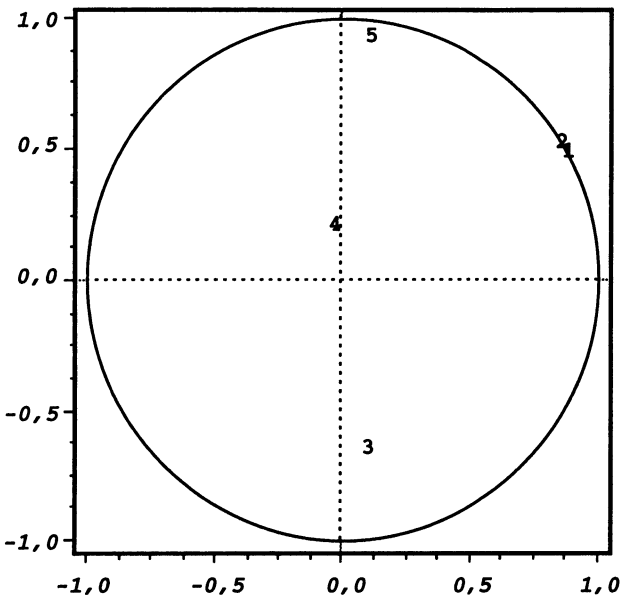

FIGURE 5 : *Corrélations des 5 paramètres du modèle PB1 avec les deux premières composantes principales de l'A.C.P. avec métrique de sensibilité.*

Pour illustrer notre commentaire ci-dessus sur l'interprétation des composantes principales à partir des paramètres, nous donnons dans le tableau 2 les valeurs des paramètres pour quelques enfants répartis en quatre groupes, qui sont opposés d'une part selon le premier axe (groupes A et B), d'autre part selon le second (groupes C et D). Afin de comparer ces groupes par rapport à l'individu moyen, nous avons ajouté à ce même tableau, les paramètres moyens des 35 garçons et ceux des 24 filles.

TABLEAU 2 :

Valeurs des paramètres pour certains enfants, regroupés selon leur position sur le première plan factoriel (voir figure 4). Les deux dernières lignes contiennent les paramètres moyens des 35 garçons et des 24 filles. Voir commentaire dans le texte.

Goupes	**N°Cb**	θ_1	θ_2	θ_3	θ_4	θ_5
A	31	198,645	185,575	0,113	1,266	15,045
A	18	191,188	177,796	0,116	1,308	14,493
A	11	186,854	176,086	0,123	0,880	14,763
A	20	187,973	173,630	0,113	1,117	13,372
B	49	155,052	146,100	0,108	1,291	15,806
B	55	156,909	144,469	0,097	1,117	12,920
B	40	156,591	144,418	0,105	1,201	12,529
C	4	177,696	163,955	0,087	1,127	17,192
C	34	181,403	169,696	0,106	1,070	16,104
C	32	178,980	167,174	0,106	1,527	15,311
C	12	180,584	167,665	0,107	1,303	15,147
D	50	165,796	153,127	0,131	1,088	11,199
D	39	162,839	151,454	0,122	1,173	11,383
D	46	164,855	151,163	0,118	0,959	10,754
D	54	163,819	155,820	0,178	1,417	11,488
Moyenne	**Garçons**	179,983	167,476	0,112	1,274	14,460
Moyenne	**Filles**	164,486	153,179	0,125	1,189	12,517

3.4. Bilan de ces deux A.C.P.

Nous venons de réaliser deux A.C.P., toutes les deux subséquentes à une même opération de modélisation de 59 courbes de croissance :

- l'une des deux A.C.P. porte sur le tableau des valeurs de la stature des 59 enfants, valeurs reconstituées à l'aide du modèle, à des âges convenus toujours les mêmes d'un enfant à l'autre;
- l'autre A.C.P. porte sur le tableau des valeurs, enfant par enfant, des paramètres du modèle.

Les métriques respectives intervenant dans ces deux A.C.P. ont été choisies pour que les inter-distances des individus soient aussi proches que possible dans les deux analyses. De fait, les cartes factorielles concernant les deux premières composantes principales issues de ces deux analyses montrent peu de différences dans les configurations des deux nuages de points (figures 2 et 4).

Bien que les différences entre les deux cartes factorielles soient faibles, on peut tenter de les expliquer. Ainsi, il est important de noter que ne sont pas rigoureusement respectées les hypothèses qui ont permis de justifier le choix du couple de métriques, aboutissant à cette proximité aussi grande que possible des résultats des deux A.C.P. Si H_0 et H_1 sont vérifiées par construction, puisque les données constituant le tableau Y sont calculées à partir du modèle, les hypothèses H_2 et H_3 ne le sont pas, comme en témoigne le graphique de la figure 6 ci-dessous. L'hypothèse H_2 est légèrement en défaut, à cause de la non linéarité du modèle. Ceci se traduit par le fait que la courbe théorique des paramètres moyens (Para Moy) et la moyenne des courbes modélisées (Moy Cb) ne coïncident pas tout à fait. Quant à l'hypothèse H_3, elle n'est pas vérifiée, comme le montrent les cas individuels choisis, dont les courbes diffèrent nettement de la courbe moyenne. Remarquons que, malgré un écart important entre la courbe n°31 et la courbe moyenne, la position du point 31 est pratiquement la même sur les deux cartes factorielles (figures 2 et 4); peut-être cette stabilité est-elle due au fait que la courbe n°31 reste constamment à peu près parallèle à la courbe moyenne? Quant aux courbes n°4 et n°32, qui sont des cas extrêmes, leurs écarts par rapport à la courbe moyenne peuvent expliquer les différences de position absolue des points correspondants, dans les deux cartes factorielles, mais les positions relatives de ces points ne sont toutefois pas bouleversées. Ces constatations nous permettent de supposer que la méthode proposée possède une certaine robustesse vis à vis des écarts par rapport à l'hypothèse H_3.

L'interprétation des positions des individus dans la carte factorielle résulte classiquement de l'interprétation, grâce au cercle des corrélations, des relations existant entre chaque composante principale et l'ensemble des «variables actives». Or ici la première A.C.P., réalisée sur 27 variables actives, montre un cercle des corrélations sans interprétation immédiate évidente des composantes principales. Par contre, la deuxième A.C.P., réalisée sur 5 variables actives, bénéficie immédiatement d'une interprétation biologique claire, grâce à la signification concrète et connue par avance des paramètres du modèle utilisé.

Nous pouvons donc dire que nous sommes en présence

- d'une unique carte factorielle, issue indifféremment de l'une ou l'autre des deux A.C.P.,
- et de deux cercles des corrélations, dont l'un est d'interprétation non évidente sans analyse plus approfondie de la situation, alors que l'autre est d'interprétation immédiate compte tenu des informations disponibles.

Ainsi, pour répondre au problème posé initialement, nous pouvons dire que ces deux analyses sont très liées entre elles : sur le plan des cartes factorielles, leurs résultats sont presque identiques, ce qui permet de conclure à une redondance très forte; sur le plan des cercles de corrélation, elles sont remarquablement complémentaires, la difficulté, voire l'impossibilité, d'interprétation des composantes

principales de l'une étant compensée par la facilité d'interprétation des composantes principales de l'autre.

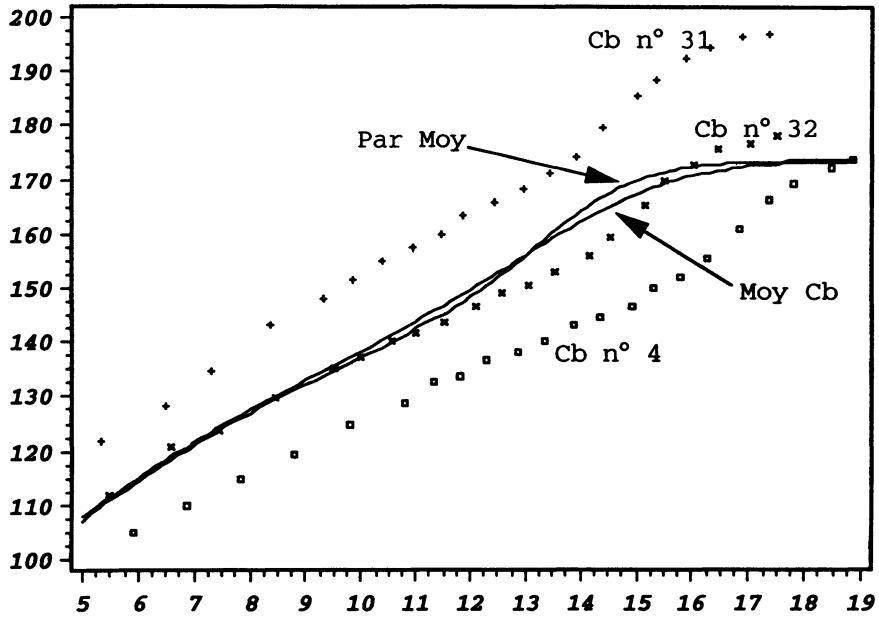

FIGURE 6 : *Courbes de croissance staturale des enfants n°4, 31 et 32 (garçons), courbe théorique des paramètres moyens (Par Moy) et moyenne des courbes théoriques (Moy Cb).*

4. Conclusion

Une étude longitudinale de la croissance se heurte presqu'inévitablement, au moins lorsqu'il s'agit de l'enfant et de l'adolescent, à deux défauts majeurs interdisant l'utilisation directe des méthodes d'analyse des données : l'hétérogénéité, entre individus, du nombre d'observations et des âges auxquels sont effectuées celles-ci.

La modélisation préalable des courbes de croissance individuelles permet de se ramener au cas standard d'un tableau complet lignes × colonnes se prêtant à l'analyse. En fait, nous obtenons à partir de la modélisation deux tableaux : l'un (Y) contenant les valeurs reconstituées de la stature à des âges choisis, l'autre (P) contenant les valeurs des paramètres du modèle. Nous nous sommes donc interrogés sur les relations susceptibles d'exister entre résultats d'une même méthode d'analyse des données pratiquée sur l'un et sur l'autre de ces deux tableaux. Plus précisément, nous avons ici choisi comme méthode l'analyse en composantes principales.

Les deux tableaux Y et P, bien que relatifs aux mêmes individus, et bien que tous les deux issus de la même modélisation, sont de nature complètement différente. Le premier contient, pour chaque individu, les valeurs d'un même caractère estimées

à des âges successifs choisis (données longitudinales). Le second associe à chaque individu une liste de valeurs de paramètres d'un modèle, valeurs résultant d'un calcul numérique (processus d'identification du modèle, à partir des données d'observation) qui tient compte de la sensibilité du modèle par rapport à chacun de ses paramètres, du nombre d'observations et surtout de leur répartition sur la période de croissance. De ce fait, ces paramètres ne sont pas calculés avec la même précision, ni entre eux, ni d'un individu à l'autre. Ces circonstances font que, si le choix de la métrique pour l'analyse du tableau Y est assez naturelle (en fait on analysera la matrice de covariance, ou la matrice de corrélation), il n'en est pas de même de la métrique pour l'analyse du tableau P. Nous avons vu que, moyennant un choix judicieux du couple de métriques, choix justifié par quelques hypothèses, l'A.C.P. de l'un et l'A.C.P. de l'autre de ces tableaux aboutissent à deux nuages de points ayant à peu près la même configuration. Si ce résultat met en évidence une certaine redondance entre les deux A.C.P., en permettant de conclure qu'on peut se livrer indifféremment à l'une ou l'autre, il reste que, sur le plan des cercles de corrélations, donc de l'interprétabilité des composantes principales, les résultats sont sans lien évident.

Comme exemple numérique, nous avons réalisé ces deux A.C.P. sur des données de croissance staturale d'enfants belges. Pour ajuster les courbes de croissance, nous avons choisi le modèle PB1 de Preece et Baines (1978). Nous avons vérifié que les résultats de la modélisation sont très satisfaisants, l'écart entre le modèle et les observations étant inférieur à l'erreur de mesure de la taille debout d'un enfant. Comme attendu, les cartes factorielles obtenues respectivement à l'issue de l'A.C.P. centrée sur les données reconstituées, et de l'A.C.P. avec métrique de sensibilité sur les paramètres du modèle, se sont révélées très similaires, et ceci bien que deux des hypothèses ne soient pas rigoureusement respectées. S'il ne nous a pas été possible d'interpréter, à partir du cercle des corrélations, les composantes principales issues de la première A.C.P., en revanche l'interprétation des composantes principales de la deuxième A.C.P. à partir du cercle des corrélations correspondant a été immédiate.

A partir de là, il serait hasardeux de conclure, à partir de ce seul exemple numérique, qu'il suffit dans tous les cas de ne pratiquer que l'A.C.P. sur les paramètres du modèle. Si dans cet exemple l'A.C.P. sur données reconstituées peut apparaître superflue, puisque l'A.C.P. sur paramètres a donné pratiquement les mêmes résultats, avec en plus une facilité d'interprétation biologique des composantes principales, ce fait ne peut être tenu pour une loi générale : nous avons ici bénéficié de l'interprétabilité biologique des paramètres du modèle choisi. Or les paramètres d'un modèle de croissance (en général une fonction déterministe), n'ont pas forcément une interprétation biologique claire, et dans ces conditions la considération du seul cercle des corrélations relatif à l'A.C.P. sur les paramètres peut n'être d'aucun secours dans l'interprétation des composantes principales. On peut même imaginer que, à l'inverse du cas concret examiné ici, la considération simultanée des deux cercles de corrélation puisse être un moyen d'obtenir une interprétation biologique des paramètres d'un modèle.

Tous les calculs et graphiques du présent article ont été réalisés grâce au logiciel ACC^{TM} (Analyse des Courbes de Croissance), créé par Hassane Abidi, perfectionné grâce aux suggestions des membres du laboratoire de Biométrie Humaine et de Biomécanique de l'Université Libre de Bruxelles. Plusieurs options dans ce logiciel résultent de différentes remarques de Roland Hauspie, du Laboratoire d'An-

thropogénétique de cette même université, que nous remercions ici. Le logiciel, ainsi que les données brutes à partir desquelles a été réalisée la présente étude, sont disponibles auprès du premier auteur.

Référence

ABIDI Hassane (1991). Contribution à la méthodologie de la modélisation des courbes de croissance. Exemple de la croissance staturale chez l'être humain. Lyon, Thèse Doct., Univ. Claude Bernard, 232 p.

BARD Y. (1974). *Nonlinear parameter estimation*. New York, Academic press, 341 p.

BECK J. V., ARNOLD K. J. (1977). *Parameter estimation in engineering and science*, New York, Wiley

BESSE P., CAUSSINUS H., FERRE L., FINE J. (1987). Sur l'utilisation optimale de l'analyse en composantes principales. *C. R. Acad. Sci. Paris*, **304**, I, 15, 459-462

BESSE P., CAUSSINUS H., FERRE L., FINE J. (1988). Principal component analysis and optimisation of graphical displays. *Statistics*, **19**, 2, 301-312

CAUSSINUS H. (1986a). Models and uses of Principal Component Analysis, *Multidimensional Data Analysis* (J. de Leeuw ed.), D.S.W.O. Press, Leiden, 149-170

CAUSSINUS H. (1986b). Quelques réflexions sur la part des modèles probabilistes en Analyse des Données. *Data Analysis and Informatics*, IV, (Diday ed.), North-Holland, Amsterdam, 151-165

CAUSSINUS H., FERRE L. (1989). Analyse en composantes principales dindividus définis par les paramètres dun modèle. *Statistique et Analyse des Données*, **41**, 19-28

ESTEVE J., SCHIFFLERS E. (1976). Discussion et illustration de quelques méthodes d'analyse longitudinale. *Proceedings of the 9th International Biometric Conference*, Vol. 1, Boston August 22-27, 463-480

FERRE L. (1989). Choix de la dimension optimale pour certains types d'analyses en composantes principales. *C. R. Acad. Sci. Paris, 309*, I, 959-964

HAUSPIE R. C. (1989). Mathematical models for the study of individual growth patterns. *Rev. Epidém. et Santé Publ.*, **37**, 461-476

HEBBELICK M., BLOMMAERT M., BORMS J., DUQUET W., VAN DER MEER J. (1980). A multidisciplinary longitudinal growth study. Introduction of the projects LEGS. (Ostyn M., Beunen G., Simens J., eds.), Kinanthropometry II, Int. Series Sports Sciences, IX, Baltimore, University Park Press, 317-325

HOULLIER F. (1986). *Echantillonnage et modélisation de la dynamique des peuplements forestiers. Application à l'Inventaire Forestier National.* Lyon, Thèse Doct., Univ. Claude Bernard, 267 p.

HOULLIER F. (1987). Comparaison de courbes et de modèles de croissance; choix d'une distance entre individus.*Statistique et Analyse des Données*, **12**, 17-36

MARQUARDT D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, **11**, 2, 431-441

PERNIN M.-O. (1986). *Contribution à la méthodologie d'analyse de données longitudinales. Exemple de la croissance chez l'être humain (auxologie).* Lyon, Thèse doct., Univ. Claude Bernard

PREECE M. A., BAINES M. J. (1978). A new family of mathematical models describing the human growth curve. *Annals of hum. Biology*, **5**, 1-24