

REVUE DE STATISTIQUE APPLIQUÉE

H. ABIDI

J. PONTIER

**Approche bayésienne pour l'estimation des paramètres
d'un modèle non linéaire. Application à la prédiction
de la taille adulte d'un enfant**

Revue de statistique appliquée, tome 45, n° 1 (1997), p. 41-59

http://www.numdam.org/item?id=RSA_1997__45_1_41_0

© Société française de statistique, 1997, 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/>

APPROCHE BAYÉSIENNE POUR L'ESTIMATION DES PARAMÈTRES D'UN MODÈLE NON LINÉAIRE. APPLICATION À LA PRÉDICTION DE LA TAILLE ADULTE D'UN ENFANT

H. Abidi*, J. Pontier**

* H. Abidi : *Epidémiologie et Information Médicale, Centre Hospitalier Lyon-Sud, 69495 Pierre Bénite cedex, France – E. mail : hassanab@uheim.univ-lyon1.fr*

** J. Pontier : *Centre de Recherche et d'Innovation sur le Sport, Laboratoire de Biométrie et d'Analyse des Données, Université Claude Bernard Lyon I, 69622 Villeurbanne cedex, France*

RÉSUMÉ

Dans les phénomènes de croissance, l'estimation des paramètres d'un modèle non linéaire selon le critère du maximum de vraisemblance est largement utilisé. La précision avec laquelle les paramètres sont calculés est fonction du nombre de mesures et surtout de leur répartition sur toute la période de croissance. Si la courbe de croissance est partiellement connue, l'imprécision sur les paramètres estimés peut être considérable. Mais si on dispose d'une information *a priori* sur la distribution des paramètres du modèle dans la population, la méthode bayésienne doit être utilisée. Dans cet article, nous rappelons le principe de cette approche en modélisation non linéaire. Nous l'avons ensuite testée sur l'exemple de croissance en taille des enfants. Nous avons mis en évidence l'importance de l'information *a priori* sur la qualité de l'estimation des paramètres et sur la prédiction de la taille adulte.

Mots-clés : *Courbes de croissance, modélisation non linéaire, distribution a priori, distribution a posteriori, estimation bayésienne, prédiction.*

ABSTRACT

The estimation of the parameters of a nonlinear model by means of the maximum likelihood (ML) procedure is widely used in the study of growth phenomena. The accuracy with which these parameters are calculated is a function of the number of measures taken and particularly, of their distribution across the growth period. If the growth curve is partially known, this accuracy can be decreased considerably. However, if we have a prior information on the distribution of the parameters of a model in the population, the bayesian method can be usefully used. In this paper, we have recolled the principle of this approach for nonlinear modeling. The method is then applied on data of human height. The results show the importance of a prior information on the quality of estimates of growth parameters and consequently on the prediction of adult height.

Keywords : *Growth curves, Nonlinear modeling, a prior distribution, a posteriori distribution, bayesian estimation, prediction.*

Introduction

On suppose qu'une série $(y_j, j = 1 \text{ à } p)$ de mesures de la taille du même enfant a été effectuée aux instants $t_j, j = 1 \text{ à } p$. Sur un plan cartésien (les t_j en abscisse et les y_j en ordonné), l'ensemble des points (t_j, y_j) a l'allure d'une «courbe» croissante. Ces points sont les résultats d'une discrétisation d'un phénomène continu. Les mesures y_j sont imparfaites, dépendant à la fois de la position de l'enfant mesuré, de la précision de l'appareil de mesure et de la personne qui prend la mesure. La modélisation est l'opération qui permet d'obtenir une trajectoire aussi proche que possible de la trajectoire réelle, toujours inconnue, de l'enfant. Cette approche passe par plusieurs étapes. La première est celle du choix du modèle de croissance. Cette étape semble franchie, puisque de nombreux modèles non linéaires sont proposés pour approcher le phénomène de croissance de la taille. Ce modèle noté $f(t, \theta)$, qui estime la taille de l'enfant à l'âge t , est caractérisé par un certain nombre de constantes appelées paramètres et symbolisé par le vecteur θ . Vient ensuite le choix du critère à optimiser. Ce dernier peut être algébrique comme celui des moindres carrés (ordinaires ou pondérés), ou probabiliste comme celui de la vraisemblance. Le premier critère consiste à chercher parmi l'ensemble des paramètres θ possibles, ceux qui minimisent une distance entre y_j et $f(t_j, \theta)$ pour $j = 1 \text{ à } p$. Le deuxième, suppose que les erreurs de mesures ε_j , dans le modèle d'erreur, $y_j = f(t_j, \theta) + \varepsilon_j$, suivent une certaine loi de probabilité. Dans ce cas, les paramètres à retenir sont ceux qui maximisent la probabilité d'observer les ε_j (ou les y_j) connaissant θ . Cette probabilité est connue sous le nom de la vraisemblance et on parle du maximum de vraisemblance. Dans le cas classique où les ε_j sont linéairement indépendants et suivent la même loi $\mathcal{N}(0, \sigma^2)$, les deux critères sont équivalents. Cette hypothèse simpliste est le plus souvent retenue et de nombreux travaux ont montré sa validité (Hartley et Booker, 1965; Jennrich R., 1969). La troisième étape est celle du choix de la méthode de calcul numérique des paramètres. Du fait que le modèle est non linéaire, toutes ces méthodes sont itératives et se différencient simplement par la façon de passer d'une itération à la suivante. Partant d'une estimation initiale, ces méthodes convergent vers un point optimum (local ou absolu) du critère retenu. En faisant varier l'estimation initiale ou par exploration dans toutes les directions autour de cet optimum, on peut éviter, dans une certaine mesure, qu'il soit local. Ce point correspond à une estimation ponctuelle des paramètres. La précision avec laquelle les paramètres sont calculés est fonction de celle des mesures observées, de leur nombre et surtout de leur répartition sur toute la période de croissance (enfance et adolescence). Le modèle étant construit pour décrire toute cette période, si des mesures manquent sur une longue période, les paramètres correspondant à cette période vont certainement «souffrir» au niveau de la précision et leurs valeurs estimées risquent d'être liées à l'estimation initiale affectée aux paramètres dans le processus itératif.

Dans les enquêtes longitudinales, il est fréquent d'avoir des données incomplètes. En effet, suivre la croissance de la taille d'un enfant pendant 19 ans pour avoir sa courbe complète est une tâche fastidieuse et même irréaliste. Pourtant, l'intérêt d'un tel suivi est important dans la mesure où il permet au pédiatre par exemple d'avoir des prévisions sur l'évolution future de la taille de cet enfant avant que celle-ci n'atteigne son maximum, et décider en conséquence des soins médicaux qui s'imposent. Dans ces conditions, et comme il a été montré dans Abidi (1991), l'estimateur du maximum de vraisemblance ne peut être d'un grand secours.

Dans le cas où l'on dispose d'une information *a priori* sur les paramètres dans la population à laquelle l'enfant est susceptible d'appartenir, la méthode bayésienne peut être utilisée (Beck et Arnold, 1977, p. 109; Theil, 1971, p. 664; Bard, 1974, p. 72). Elle a l'avantage de combiner les deux informations : celle apportée par les données observées de l'enfant et celle concernant l'ensemble de la population, pour donner une estimation des paramètres (dite bayésienne) la plus vraisemblable.

L'approche bayésienne est largement utilisée dans l'estimation des paramètres du modèle linéaire, (Potthoff et Roy, 1964; Geisser, 1970; Rao, 1975; Fearn, 1975; Strenio *et al.*, 1983). Elle a été discutée et appliquée par Bock et Thissen (1980) sur leur modèle non linéaire (triple logistique 1976). Berkey (1982) l'a décrite pour l'estimation des paramètres du modèle linéaire de Jents et Bayley (1937). Shohoji et Sasaki (1987) et Kanefuji et Shohoji (1990) l'ont appliquée au modèle de Shohoji et Sasaki (1987). Dans cet article nous nous plaçons dans le cas d'un modèle non linéaire, pour rappeler le principe de cette méthode et montrer son intérêt dans la prédiction de la taille adulte d'un enfant. Nous verrons également que cette méthode peut être utilisée avec profit pour la reconstitution des mesures manquantes durant la période de l'enfance (prédiction de la taille d'un enfant à la naissance par exemple).

Parmi les modèles non linéaires susceptibles de décrire la totalité de la courbe de croissance en taille d'un enfant (Abidi *et al.*, 1996), nous choisirons le deuxième modèle de Jolicœur *et al.* (1992) noté JPA2. Les performances de ce modèle seront discutées sur le plan descriptif et prédictif. Nous montrerons également l'effet des erreurs de mesure sur la qualité de la prédiction. On montrera l'apport de l'information *a priori* au niveau de la qualité prévisionnelle du modèle.

NOTATIONS ET HYPOTHÈSES

– On note $y = {}^t(y_1, y_2, \dots, y_p)$ ¹ le résultat de l'observation de la variable aléatoire Y (la taille de l'enfant) aux différents instants t_j .

– $f(t, \theta)$ note le modèle mathématique choisi (supposé adéquat) pour décrire toute la période de croissance, $\theta = {}^t(\theta_1, \theta_2, \dots, \theta_q)$ le vecteur de q paramètres du modèle à estimer, et $F(\theta) = {}^t(f(t_1, \theta), f(t_2, \theta), \dots, f(t_p, \theta))$.

– Le modèle d'erreur s'écrit donc : $Y = F(\theta) + \varepsilon$ où $\varepsilon = {}^t(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ vecteur résiduel de $\mathbb{R}^p$.

– Il est naturel de supposer que $E(Y) = F(\theta)$ donc $E(\varepsilon) = 0$ ². On suppose connue et inversible la matrice de covariances de ε et on la note Σ . Dans notre cas (modélisation de la taille humaine), il semble légitime de supposer que la distribution de ε est normale :

$$L(\varepsilon) = \frac{1}{\sqrt{(2\pi)^p \sqrt{\det(\Sigma)}}} \text{Exp} \left\{ -\frac{1}{2} {}^t\varepsilon \Sigma^{-1} \varepsilon \right\}^3$$

¹ ${}^t A$ représente la transposée de la matrice A .

² $E(X)$ indique l'espérance de X .

³ $\det(\Sigma)$ et Σ^{-1} notent respectivement le déterminant et l'inverse de la matrice Σ .

Dans ces conditions et pour θ donné, la vraisemblance des erreurs de mesures observées a donc pour expression

$$L(\varepsilon/\theta) = \frac{1}{\sqrt{(2\pi)^p} \sqrt{\det(\Sigma)}} \text{Exp} \left\{ -\frac{1}{2} {}^t(Y - F(\theta))\Sigma^{-1}(Y - F(\theta)) \right\}$$

Pour simplifier le développement qui suit, nous supposons que les ε_j sont linéairement indépendants et suivent la même loi $\mathcal{N}(0, \sigma^2)$ (c'est-à-dire $\Sigma = \sigma^2 I$). Notons qu'on peut toujours se ramener à cette situation grâce à une transformation linéaire judicieuse du vecteur $(Y - F(\theta))$.

– On suppose qu'une information *a priori* sur la distribution de θ en tant que variable aléatoire de $\mathbb{R}^q$, sous forme d'une loi de probabilité notée $L(\theta)$, est disponible. On suppose en plus qu'elle suit une loi normale de moyenne $\bar{\theta}$ et de matrice de covariances V :

$$L(\theta) = \frac{1}{\sqrt{(2\pi)^q} \sqrt{\det(V)}} \text{Exp} \left\{ -\frac{1}{2} {}^t(\theta - \bar{\theta})V^{-1}(\theta - \bar{\theta}) \right\}$$

ESTIMATION BAYÉSIENNE

Le théorème de Bayes relatif aux distributions conditionnelles permet d'écrire la distribution conjointe $L(\theta, \varepsilon)$ de θ et ε sous deux formes :

$$L(\theta, \varepsilon) = L(\theta/\varepsilon)L(\varepsilon) \text{ et } L(\theta, \varepsilon) = L(\varepsilon/\theta)L(\theta)$$

et par conséquent $L(\theta/\varepsilon) = \frac{L(\varepsilon/\theta)L(\theta)}{L(\varepsilon)}$,

$L(\theta/\varepsilon)$ est dite distribution *a posteriori* de θ après observation des résultats de l'expérience. L'estimateur bayésien de θ est celui qui maximise $L(\theta/\varepsilon)$ donc $L(\varepsilon/\theta)L(\theta)$ puisque $L(\varepsilon)$ est indépendant de θ . Dans les conditions des hypothèses citées ci-dessus, la distribution *a posteriori* $L(\theta/\varepsilon)$ aura la forme suivante :

$$L(\theta/\varepsilon) = k \text{Exp} \left\{ -\frac{1}{2\sigma^2} {}^t(Y - F(\theta))(Y - F(\theta)) - \frac{1}{2} {}^t(\theta - \bar{\theta})V^{-1}(\theta - \bar{\theta}) \right\}$$

où $k = \frac{1}{\sqrt{(2\pi\sigma^2)^p}} \frac{1}{\sqrt{(2\pi)^q} \sqrt{\det(V)}} \frac{1}{L(\varepsilon)}$

La valeur de paramètres θ qui maximise $L(\theta/\varepsilon)$ (le mode) est celle qui minimise la fonction $h(\theta) = -\sigma^2 \text{Log}(L(\theta/\varepsilon))$. Comme k et $L(\varepsilon)$ sont indépendants de θ , $h(\theta)$ possède les mêmes minimums que :

$$g(\theta) = \frac{1}{2} {}^t(Y - F(\theta))(Y - F(\theta)) + \frac{1}{2} \sigma^2 {}^t(\theta - \bar{\theta})V^{-1}(\theta - \bar{\theta})$$

Une condition nécessaire pour que $g(\theta)$ admette un minimum est que $\frac{\partial g(\theta)}{\partial \theta_h} = 0$ pour tout h . Si on retient une méthode analytique pour le calcul des paramètres, θ doit être une des solutions possibles du système de q équations non linéaires : $\frac{\partial g(\theta)}{\partial \theta_h} = 0$ ($k = 1$ à q), soit

$$-\sum_{j=1}^p (y_j - f(t_j, \theta)) \frac{\partial f(t_j, \theta)}{\partial \theta_h} + \sigma^2 (V^{-1}(\theta - \bar{\theta}))_h = 0 \quad (h = 1 \text{ à } q).$$

où $(V^{-1}(\theta - \bar{\theta}))_h$ est la $h^{\text{ème}}$ composante de $V^{-1}(\theta - \bar{\theta})$. Partant d'une estimation initiale des paramètres (ceux de l'information *a priori* par exemple, supposée disponible), la méthode itérative de Newton permet le passage de θ^i obtenu lors de l'itération i à $\theta^{(i+1)}$ par l'intermédiaire de la relation

$$\theta^{i+1} = \theta^i - H^{-1}(\nabla g) \quad (1)$$

où H est le Hessien de $g(\theta)$, matrice dont l'élément de la $k^{\text{ème}}$ ligne et la $h^{\text{ème}}$ colonne est :

$$\left(\frac{\partial^2 g(\theta)}{\partial \theta_k \partial \theta_h} \right) = - \sum_{j=1}^p \left(\frac{\partial^2 f(t_j, \theta)}{\partial \theta_k \partial \theta_h} \right) (y_j - f(t_j, \theta^i)) + \sum_{j=1}^p \frac{\partial f(t_j, \theta)}{\partial \theta_k} \frac{\partial f(t_j, \theta)}{\partial \theta_h} + \sigma^2 (V^{-1})_{kh} \quad (2)$$

et ∇g le vecteur gradient de $g(\theta)$ de composantes $\frac{\partial f(\theta)}{\partial \theta_h}$ dont les dérivées sont évaluées au point vecteur θ^i (valeur de θ à l'itération i) et $(V^{-1})_{kh}$ l'élément courant de V^{-1} .

Si E note la matrice dont l'élément courant est le premier terme du membre de droite de (2), la relation (1) peut s'écrire explicitement sous la forme :

$$\theta^{i+1} = \theta^i - [E + {}^tSS + \sigma^2 V^{-1}]^{-1} (-{}^tS(Y - F(\theta)) + \sigma^2 V^{-1}(\theta^i - \bar{\theta}))$$

où S , matrice de sensibilité associée au modèle f , est la matrice d'éléments $\left(\frac{\partial f(t_j, \theta)}{\partial \theta_h} \right)_{jh}$ ($j = 1, p$ et $h = 1, q$).

REMARQUES

1) Dans la pratique, le calcul des éléments de E est lourd à cause de la présence du terme $\left(\frac{\partial^2 f(t_j, \theta)}{\partial \theta_k \partial \theta_h} \right)$. Du fait qu'en fin de processus itératif, $y_j - f(t_j, \theta^i)$ est de plus en plus faible, cette matrice E est souvent négligée par rapport à $({}^tSS + \sigma^2 V^{-1})$.

2) En général, σ^2 n'est pas connue à l'avance, elle est estimée à chaque itération par la quantité

$$\hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=1}^p [y_j - f(t_j, \hat{\theta})]^2$$

si E est négligée, la matrice H à inverser a donc la même forme que celle proposée dans la méthode de Marquardt (1965) où la matrice identité est remplacée par V^{-1} et λ par $\hat{\sigma}^2$ qui (comme λ) décroît d'une itération à la suivante. En fin de processus, et si le modèle est adéquat, $\hat{\sigma}^2$ est faible donc aussi $\hat{\sigma}^2 V^{-1}$ et la méthode de Newton se réduit à la méthode de Gauss.

3) Comme la méthode de Marquardt, l'estimation bayésienne utilisant l'information *a priori* est un moyen de passer outre la colinéarité des colonnes de S . En présence de données qui ne couvrent pas toute la période de croissance, la matrice S peut avoir des colonnes nulles. En effet, prenons l'exemple du modèle JPA2 dont le paramètre θ_7 dépend exclusivement de la période de l'adolescence (13 ans à 18 ans).

Pour tout t en dehors de cette période on a $\left(\frac{\partial f(t, \theta)}{\partial \theta_7} \right) = 0$ (Abidi, 1991). L'absence de mesures sur cette période donnera des valeurs nulles de la colonne 7 de la matrice de sensibilité S . Ce cas se présente pour nous quand la période de modélisation est de 1 mois à 11 ans par exemple. La matrice ${}^t S S$ sera donc d'ordre inférieur à q , d'où l'utilité du terme $\sigma^2 V^{-1}$ qui jouera le rôle de mesures supplémentaires.

4) L'importance du terme $\sigma^2 V^{-1}$ est fonction à la fois de σ^2 et des éléments de V^{-1} . Dans le cas particulier où V est diagonale (indépendance linéaire entre les paramètres du modèle) et est calculée sur un échantillon très homogène, les éléments de V sont «très faibles» donc ceux de V^{-1} «très grands» d'où une «grande» importance de $\sigma^2 V^{-1}$. Dans ces conditions, l'information *a priori* peut provoquer un conflit avec celle apportée par les mesures de l'enfant si ce dernier a une croissance particulière comparée à celle de la population (qui a permis le calcul de V). Ceci se manifeste par une tendance à la divergence de la méthode itérative. Cette situation est prévue dans le logiciel ACCTM d'analyse de courbes de croissance (Abidi *et al.*, 1996) avec lequel le calcul a été réalisé. Il propose de pondérer l'information *a priori* ($\sigma^2 V^{-1}$) par un coefficient μ choisi entre 0 et 1.

5) La matrice des covariances des paramètres est estimée par $\hat{\sigma}^2 [E + {}^t S S + \hat{\sigma}^2 V^{-1}]^{-1}$. La présence du terme $\sigma^2 V^{-1}$ fait que l'information *a priori* donne des estimations des paramètres plus précises par rapport à l'estimation du maximum de vraisemblance. Cette précision est d'autant plus grande que l'échantillon qui est à la base du calcul de V est homogène et représentatif de la population.

APPLICATION

Les données et le modèle de croissance

On considère 27 enfants «normaux» (13 garçons et 14 filles) provenant de l'enquête auxologique française (1953-1978), et dont la taille a été mesurée régulièrement et sans mesures manquantes, aux âges de un mois, 3 mois, 6 mois,

9 mois, 1 an puis tous les 6 mois jusqu'à l'âge de 19 ans. La modélisation des données de ces enfants a déjà fait l'objet de plusieurs travaux (Pontier *et al.*, 1988; Jolicœur, Pontier *et al.*, 1988; Caussinus et Ferré, 1989; Abidi, 1991). Dans le présent article, nous montrons l'apport de l'estimation bayésienne dans la prédiction de la taille adulte de ces enfants. Pour cela, nous avons voulu reconstituer le suivi de chaque enfant à partir de l'âge de 11 ans et comparer l'évolution réelle de cet enfant et celle prédite par le modèle. Pratiquement, chaque courbe de croissance observée est décomposée en 17 périodes : de 1 mois à 11 ans, de 1 mois à 11,5 ans, ..., de 1 mois à 19 ans (figure 1). Chaque enfant sera donc modélisé 17 fois. L'analyse des 17 estimations des paramètres et leur impact sur les courbes théoriques en fonction de la période utilisée, va nous permettre de déduire à partir de quel âge les estimations relatives au même enfant deviennent stables.

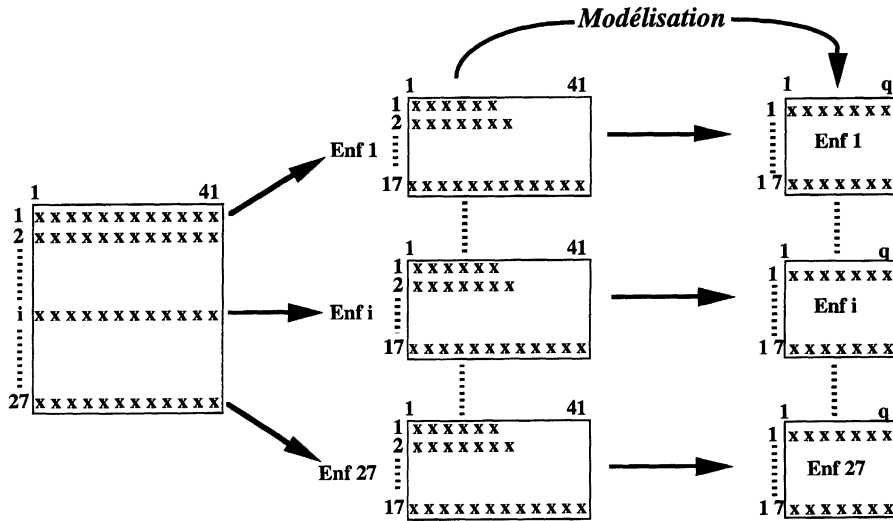


FIGURE 1

La période de croissance (1 mois à 19 ans) de chaque enfant est décomposée en 17 périodes, la première est de 26 mesures et couvre la période de 1 mois à 11 ans, la deuxième couvre la période de 1 mois à 11,5 ans, etc., la dernière contient toutes les mesures de cet enfant (au nombre de 41). La modélisation de l'ensemble des $459 = 17 \times 27$ courbes va nous donner le tableau de 459 lignes et q colonnes correspondant aux paramètres du modèle choisi.

Pour la modélisation de la croissance du caractère taille de l'être humain, plusieurs modèles sont candidats (Abidi, 1991; Hauspie R., 1989). Nous choisissons pour cette application JPA2, c'est-à-dire le deuxième modèle de Jolicœur, Pontier et Abidi (1992) à 8 paramètres ($\theta_1, \theta_2, \dots, \theta_8$) et d'expression analytique

$$f(t, \theta) = \theta_1 \left[1 - \left\{ 1 + \left(\frac{t + \theta_8}{\theta_2} \right)^{\theta_3} + \left(\frac{t + \theta_8}{\theta_4} \right)^{\theta_5} + \left(\frac{t + \theta_8}{\theta_6} \right)^{\theta_7} \right\}^{-1} \right]$$

Le paramètre θ_1 estime la taille adulte et le paramètre θ_8 est lié à la période de gestation.

Pour avoir l'information *a priori*, on a procédé en deux étapes : 1) modélisation des 27 courbes complètes en utilisant le critère du maximum de vraisemblance, 2) calcul des paramètres moyens $\bar{\theta}$ et de la matrice de covariance V des paramètres obtenus par sexe. La loi *a priori* des paramètres étant supposée normale $\mathcal{N}(\bar{\theta}, V)$. Les paramètres moyens des 13 garçons vont servir aussi comme estimation initiale pour 221 courbes (13 garçons $\times$ 17 périodes) et ceux des 14 filles constitueront une estimation initiale pour 238 courbes (14 filles $\times$ 17 périodes).

Qualité globale de l'estimation

Comme critère de jugement de la qualité globale d'ajustement, nous avons retenu l'écart quadratique moyen (EQM)

$$EQM = \sqrt{\frac{1}{p} \sum_{j=1}^p [y_j - f(t_j, \hat{\theta})]^2}$$

Ce critère a le mérite de s'exprimer avec la même unité de mesure que celle du caractère étudié (le cm dans notre cas) et de ne pas dépendre du nombre de paramètres du modèle.

Chaque courbe de la figure 2 représente les différentes valeurs de l'EQM obtenues au cours des différentes modélisations relatives à la courbe de croissance d'un même enfant. Les valeurs des abscisses indiquent la fin de la période utilisée dans le calcul des paramètres : de l'âge d'un mois à 11 ans, à 11,5 ans, ..., à 19 ans. Selon ce critère, la qualité globale de tous les ajustements peut être considérée comme excellente. L'écart entre la mesure observée et celle estimée par le modèle est en moyenne inférieur à 1 cm (moins que l'erreur que l'on peut commettre lors de la prise de la taille d'un enfant). Pour certains enfants la valeur de ce critère a augmenté de façon brutale, comme par exemple le cas du n°5 où l'EQM était inférieure à 0,24 cm, puis il dépasse 0,55 cm en ajoutant les mesures de 12,5 ans à 13,5 ans. Cette remarque laisse entendre que c'est peut être une conséquence d'erreurs de mesure sur cette période. Nous vérifierons par la suite cette hypothèse.

Evolution de l'estimation de la taille adulte d'un enfant

Le paramètre θ_1 du modèle JPA2 estime la taille adulte de l'enfant. La figure 3, montre l'évolution de l'estimation de ce paramètre pour un même enfant au cours des différentes périodes de croissance utilisées. A quelques exceptions près, les courbes sont «horizontales», elles montrent donc une bonne prédiction de la taille réelle de l'enfant dès l'âge de 11 ans. Comme pour l'EQM, certaines courbes ne sont pas régulières (monotones) comme par exemple celles des enfants 4, 5, 6 et 10.

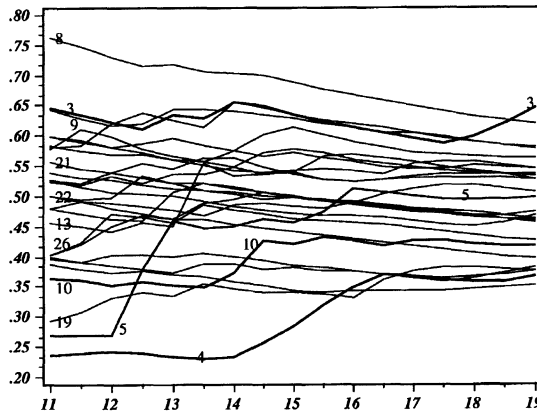


FIGURE 2

Chaque courbe représente l'évolution de l'EQM mesurant la qualité globale des ajustements du même enfant selon la période prise en compte

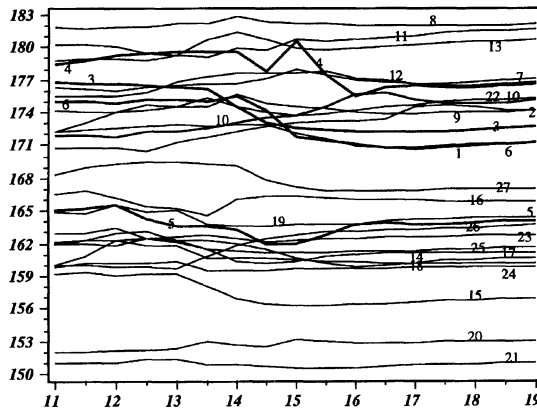


FIGURE 3

Le paramètre θ_1 du modèle JPA2, estime la taille adulte d'un enfant. Chacune des courbes représente l'évolution de l'estimation de ce paramètre selon la période prise en compte dans le calcul des paramètres

Evolution globale des paramètres du modèle

Sur le tableau obtenu par concaténation verticale de ceux de droite de la figure 1, on a pratiqué une analyse en composantes principales avec métrique de sensibilité (Houllier, 1987; Abidi *et al.*, 1995). Les résultats de cette analyse sont résumés par le premier plan factoriel de la figure 4. L'évolution globale des paramètres du même enfant selon la période de croissance utilisée est représentée par une «trajectoire». Le numéro indiqué au bout de chaque trajectoire correspond à la position de l'enfant quand toute sa courbe de croissance observée (de l'âge d'un mois à 19 ans) est prise

en compte dans l'estimation des paramètres. Le fait que ce plan factoriel explique 95,3% de l'inertie totale (dont 75,9% est due au premier axe et 19,4% au deuxième), explique que, les trajectoires les moins étalées sont celles qui correspondent à une stabilité dans l'évolution des paramètres en fonction de la période prise en compte dans leur calcul. Ainsi la non régularité des courbes des enfants 4, 5, 6 et 10 dans la figure 3 (voir remarque ci-dessus), explique pour ces enfants une trajectoire plus étalée que les autres, dans la figure 4.

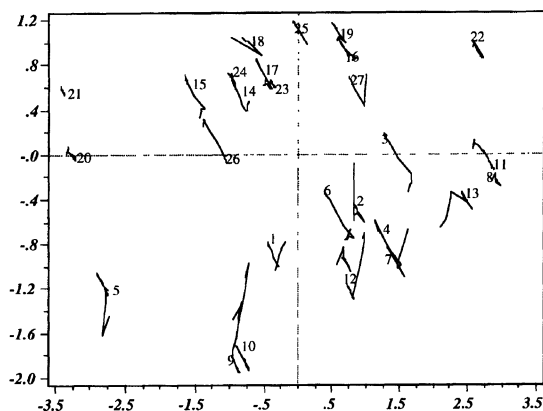


FIGURE 4

Premier plan factoriel de l'ACP sur les paramètres du modèle JPA2 muni de la métrique de sensibilité. Chaque croquis ou trajectoire représente l'évolution globale des paramètres du même enfant selon la période de croissance utilisée.

Ecart entre la taille réelle de l'enfant et celle prédite par le modèle

Dans notre exemple, on connaît la taille de nos 27 enfants à 19 ans. Nous avons calculé l'écart entre la taille de l'enfant à cet âge et celle prédite par le modèle selon les 17 périodes de croissance utilisées. Chacune des courbes de la figure 5 est relative à un enfant. Elle représente l'évolution de cet écart en fonction de la période de croissance prise en compte (d'un mois à 11 ans, à 11,5 ans, ..., à 19 ans). Globalement l'erreur maximale de prédiction est en valeur absolue inférieure à 5 cm. Elle devient inférieure à 2 cm dès que les données de l'enfant couvrent la période d'adolescence (15 ans). L'ajout de certaines mesures modifie la trajectoire de l'erreur de prédiction.

La figure 6, montre l'évolution moyenne par sexe de l'écart absolu entre la taille réelle et celle estimée par le modèle JPA2 selon la période de croissance prise en compte. L'intervalle de confiance au niveau de 95% est calculée directement à partir de 13 (ou 14) erreurs de prédiction relatives à une période d'observation. En moyenne l'erreur de prédiction est inférieure à 1,6 cm chez les filles et à 2,5 cm chez les garçons, et qu'elle est systématiquement supérieure chez les garçons, quelle que soit la période de croissance prise en compte. La taille des échantillons est relativement faible (13 ou 14 selon le sexe), l'hypothèse de normalité ne peut être assurée. Les intervalles de confiance sont donc à prendre avec précaution.

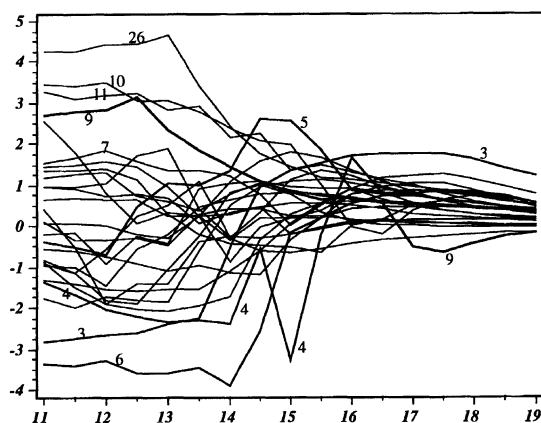


FIGURE 5

Chaque courbe représente l'évolution de l'écart entre la taille réelle et celle prédite par le modèle JPA2 du même enfant selon la période de croissance modélisée.

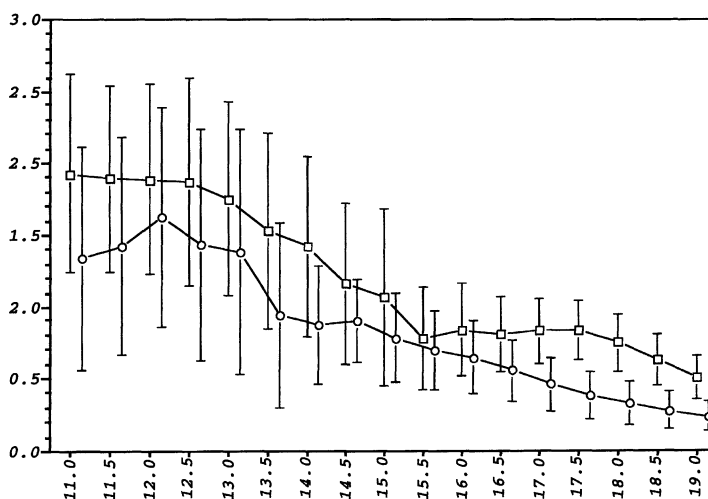


FIGURE 6

Ecart absolu moyen entre la taille réelle et celle prédite par le modèle JPA2 et son intervalle de confiance par sexe

□ moyenne sur les 13 garçons, ○ moyenne sur les 14 filles

Effet des erreurs de mesures sur la prédiction de la taille à 19 ans

On pouvait penser que l'évolution de l'écart entre la taille réelle et celle prédite par le modèle serait régulière et monotone (croissante ou décroissante). La figure 5 ci-dessus, montre que certaines courbes évoluent de manière désordonnée.

Par exemple, celle de l'enfant n°5, dont l'erreur de prédiction était faible à 11 ans, a augmenté à 12 ans, atteint un maximum en changeant de signe à 14,5 ans, puis devient pratiquement horizontale à partir de 16 ans. Pour vérifier si des erreurs de mesure éventuelles peuvent expliquer ce changement, la figure 7 présente les résidus (écarts entre les 41 mesures de la taille de cet enfant et celles estimées par le modèle JPA2 sur la courbe entière). Jusqu'à 11 ans ces résidus sont relativement faibles, puis deviennent très importants à 12 ans et à 13 ans. On a fait la même chose pour d'autres enfants, ceux pour lesquels les courbes de la figure 2 et 3 sont non régulières et celles de la figure 4 sont plus étalées. On arrive à la même conclusion : là où il y a un changement de direction, il y a un résidu anormalement important.

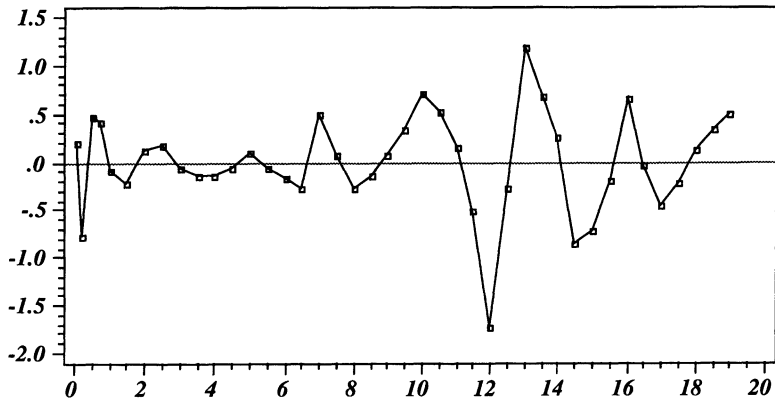


FIGURE 7

Courbe des résidus (écart entre la taille réelle et celle estimée par le modèle JPA2) de l'enfant n°5

Pour essayer de préciser l'effet des erreurs de mesure sur la qualité de la prédiction, à partir des paramètres de l'enfant n°5, on a remplacé les mesures de cet enfant par celles estimées par le modèle. A chaque estimation, un bruit aléatoire de moyenne nulle et d'écart type e (égal à 0,01, 0,05, 0,1 et 0,5 cm) a été ajouté selon la méthode de Box et Müller (1958). On a donc créé quatre courbes théoriques qui se distinguent uniquement par l'ordre de grandeur de l'erreur ajoutée à la mesure estimée. Sur ces courbes, une modélisation évolutive a été réalisée : exactement de la même façon que pour les données réelles, on a effectué 17 modélisations pour chaque courbe. Ensuite a été calculé l'écart entre la taille à 19 ans et celle prédite par le modèle en fonction de la période de croissance prise en compte. Sur la figure 8, on a superposé l'erreur de prédiction pour l'enfant n°5, et celles de ce même enfant reconstitué par le modèle plus un bruit aléatoire dont l'écart type est indiqué sur la courbe. Les erreurs de prédiction de la taille à 19 ans de cet enfant à partir de 14,5 et 15 ans, sont très supérieures aux erreurs obtenues en ajoutant le bruit le plus fort. On peut donc noter que l'erreur de prédiction est fonction de l'importance de l'erreur qui entache la mesure.

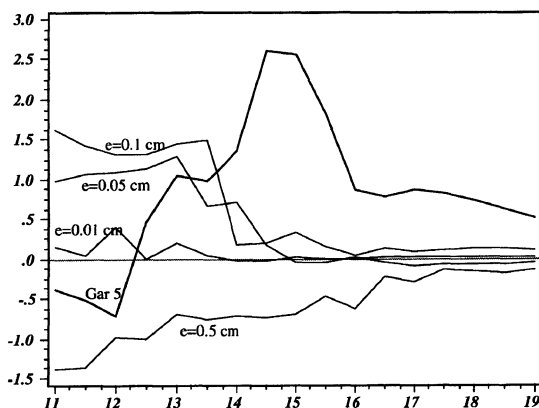


FIGURE 8

La courbe «Gar 5» est extraite de la figure 5 ci-dessus, elle représente l'évolution de l'écart entre la taille réelle et celle prédite par le modèle de l'enfant n°5 selon la période de croissance prise en compte dans la modélisation. Pour les autres, les données de cet enfant sont remplacées par leur estimation par le modèle plus une erreur aléatoire normale de moyenne nulle et d'écart type e indiqué sur la courbe.

Impact de la période d'observation sur les courbes théoriques

a) En présence d'information a priori

Sur la figure 9 sont superposées la courbe observée de l'enfant n°5, ses 17 courbes théoriques qui reconstituent l'évolution théorique de sa taille (échelle à gauche en cm), et les 17 courbes de vitesse instantanée de croissance (échelle à droite en cm/an) en fonction de la période d'observation. Le choix de cet enfant se justifie par les remarques précédentes : présence d'une probable erreur de la mesure à 12 ans, et mauvaise prédiction de sa taille à 19 ans. Que ce soit pour la taille théorique ou la vitesse instantanée, les 17 courbes sont pratiquement identiques sur la période d'observation commune (1 mois à 11 ans). C'est à partir de 12 ans, que l'impact sur les courbes théoriques, et surtout sur les courbes de vitesse devient important.

b) En l'absence d'information a priori

Si on applique la méthode du maximum de vraisemblance au lieu de la méthode de Bayes, qu'advient-il de la qualité de la prédiction de la taille adulte ?

A cette question on peut trouver une réponse détaillée dans Abidi (1991). Dans cet article, on présente le cas du même enfant numéro 5. Une modélisation évolutive a été appliquée sur la courbe observée de cet enfant en choisissant comme critère à optimiser celui du maximum de vraisemblance sous l'hypothèse que les erreurs sont indépendantes et suivent la même loi normale de moyenne nulle et de variance constante (ce qui correspond au critère des moindres carrés ordinaires). Les résultats de cette modélisation sont résumés par les graphiques de la figure 10 où sont

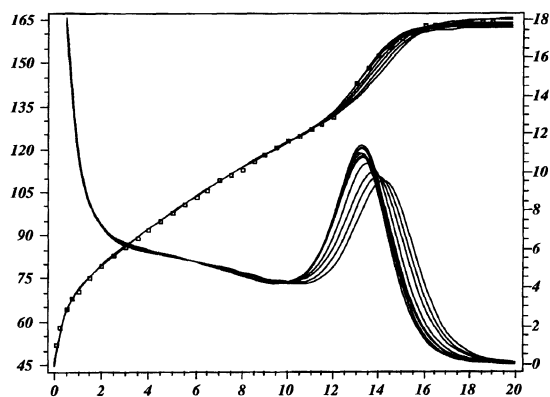


FIGURE 9

Superposition de la courbe observée (□) de l'enfant n° 5, de ses 17 courbes théoriques (échelle à gauche) et ses 17 vitesses instantanées de croissance (échelle à droite), selon la période de modélisation utilisée.

Le critère à optimiser est celui du mode de la distribution a posteriori.

superposées les 17 courbes théoriques de la taille et celles des vitesses instantanées correspondantes de ce même enfant. Chaque courbe est le résultat de la modélisation sur une période de croissance donnée (1 mois à 11, 1 mois à 11,5 ans,..., 1 mois à 19 ans). Les 17 courbes sont presque identiques sur la période de 1 mois à 11 ans, puis se détachent les unes des autres pour donner des pronostics complètement différents de la taille adulte de cet enfant et de sa vitesse instantanée de croissance.

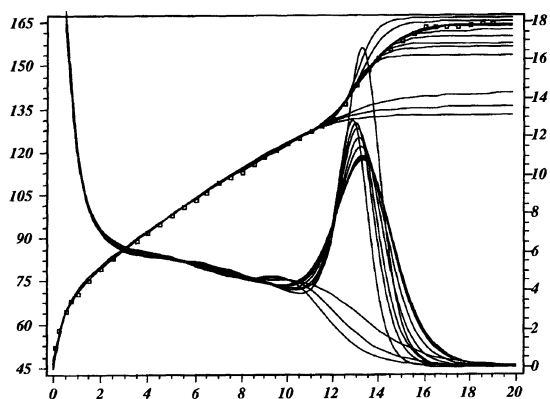


FIGURE 10

Superposition de la courbe observée (□) de l'enfant n° 5, de ses 17 courbes théoriques (échelle à gauche), et ses vitesses instantanées de croissance (échelle à droite) selon la période de modélisation utilisée.

Le critère à optimiser est celui du maximum de vraisemblance.

DISCUSSION

La modélisation d'une courbe de croissance observée comme celle de la taille humaine, pose de nombreux problèmes. Le choix du modèle théorique qui est susceptible d'approcher la croissance réelle vient en premier. Au début de la modélisation de ce caractère, on a commencé par découper la période de croissance en deux : la période de l'enfance et celle de l'adolescence. Des modèles théoriques sont proposés (ou adaptés) pour telle ou telle période. Pour exemple, le modèle de Count (1943) a été utilisé pour décrire la période de l'enfance, le modèle de Gompertz et le modèle logistique pour la période de l'adolescence (Marubini *et al.* 1971, Hauspie 1981). On peut imaginer la difficulté d'interprétation des paramètres de plusieurs modèles décrivant différentes parties du même phénomène de croissance. Plus récemment des modèles plus complets sont construits pour décrire l'évolution du caractère taille de la naissance à l'âge adulte. Dans cet article, nous avons travaillé avec l'un des meilleurs modèles de croissance : le deuxième des deux proposés par Jolicœur *et al.* (1992), noté JPA2.

Vient ensuite le choix du critère à optimiser et la méthode d'identification. Le critère est en général soit algébrique (moindres carrés ordinaires ou pondérés) soit probabiliste (maximum de vraisemblance). Sous la triple hypothèse de normalité, d'indépendance, et d'homoscédasticité des erreurs de mesures, les deux critères sont équivalents. C'est dans ces conditions que nous nous sommes placés dans tous les calculs que nous avons effectués. La méthode de calcul des paramètres peut être exploratrice (la simplexe en est un bon exemple) ou analytique (celles qui découlent de la méthode de Newton). Nos calculs sont effectués selon la méthode analytique proposée par Marquardt (1963) et implantée dans le logiciel ACCTM d'analyse et modélisation des courbes de croissance (Abidi *et al.*, 1996).

Quand les problèmes ci-dessus sont résolus, on s'interroge sur la précision avec laquelle les paramètres sont calculés. On sait qu'elle est fonction de celle des mesures observées, de leur nombre et surtout de leur répartition sur la période de croissance (Bard, 1974; Abidi, 1991). Le manque de mesures dans une zone entraînera forcément une imprécision sur le ou les paramètres sensibles dans cette zone et ce quel que soit le critère à optimiser, la méthode de calcul ou le modèle utilisé. Cette éventualité rend peu fiables l'emploi de ces méthodes dans les problèmes de prévision.

Pour remédier à cette lacune l'estimation bayésienne nous propose d'incorporer une information *a priori* dans le critère à optimiser. Cette information doit provenir d'un échantillon représentatif de l'ensemble des courbes à modéliser. Dans cet article, nous l'avons tirée de l'ensemble des données observées. Ensuite nous avons montré l'intérêt de cette méthode dans, 1) la précision de l'estimation des paramètres et 2) la prédiction de la taille d'un enfant à 19 ans. Nous avons reconstitué l'historique de 27 enfants qui ont été suivis de l'âge d'un mois à 19 ans.

Avant de parler de la prévision, il faut s'assurer de la qualité de l'ajustement aux données observées. Encore une fois le modèle JPA2 montre une grande souplesse pour décrire la croissance en taille d'un enfant.

Chacun de nos 27 enfants ayant été modélisé 17 fois, on se demande si la suite constituée par les 17 vecteurs de paramètres relatifs au même enfant est «numériquement convergente», et à partir de quel âge. Nous avons montré que l'erreur

moyenne de prédiction est en valeur absolue inférieure à 2,5 cm chez les garçons et 1,6 cm chez les filles (figure 6) et qu'elle est très faible quand les mesures couvrent la période d'adolescence. Nous avons montré sur un exemple l'effet des erreurs de mesures sur la qualité de la prédiction (figure 8). Sur le même exemple, nous avons montré, l'avantage de l'estimation bayésienne (figure 9) sur celle du maximum de vraisemblance (figure 10).

Il est vrai que dans notre exemple de la modélisation du caractère taille de l'être humain, il est plus important de pouvoir prédire la taille adulte que la taille à la naissance. La méthode est pourtant applicable dans le deuxième cas comme dans le premier. Pour le montrer, nous avons procédé de manière inverse. Chacun de nos 27 enfants a été modélisé 25 fois. La première est relative à la période de 11 ans à 19 ans, la deuxième de 10,5 ans à 19 ans et ainsi de suite, la 25^{ième} correspond à la modélisation de la courbe observée de cet enfant. Une analyse en composantes principales avec métrique de sensibilité a été appliquée sur le tableau de 675 (25×27) lignes et 8 colonnes (les paramètres du modèle JPA2).

La figure 11 représente le premier plan factoriel de cette analyse. Ce dernier exprime 96,72% d'inertie totale dont 89,37% est due au premier axe et 7,35% au second. C'est presque le même pourcentage d'inertie que celui du plan factoriel de la figure 4. Une simple comparaison visuelle de ces deux figures (même dimension, même échelle), montre que l'évolution globale (trajectoire) des paramètres est beaucoup plus stable dans ce deuxième cas.

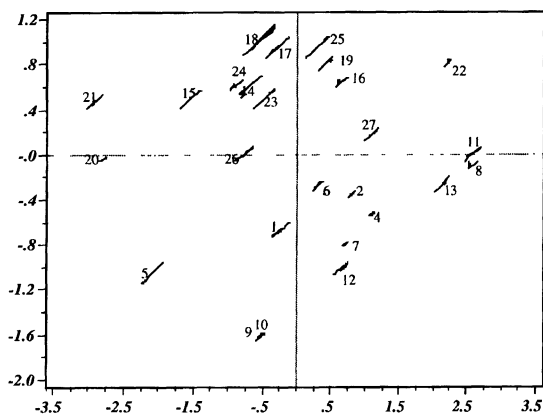


FIGURE 11

Premier plan factoriel de l'ACP sur les paramètres du modèle JPA2 muni de la métrique de sensibilité. Il explique 95,8% de l'inertie totale dont 74,7% est dû au premier axe et 21,1% au second. Chaque trajectoire représente l'évolution globale des paramètres du même enfant selon la période de croissance prise en compte : 11 ans à 19 ans, 10,5 ans à 19 ans etc., de 1 mois à 19 ans

Sur la figure 12, on a superposé trois courbes. Celle en petits cercle représente l'évolution de la taille observée de l'enfant numéro 1 (garçon). La courbe continue (A) est le résultat de la modélisation sur la période de 1 mois à 11 ans. Enfin la courbe

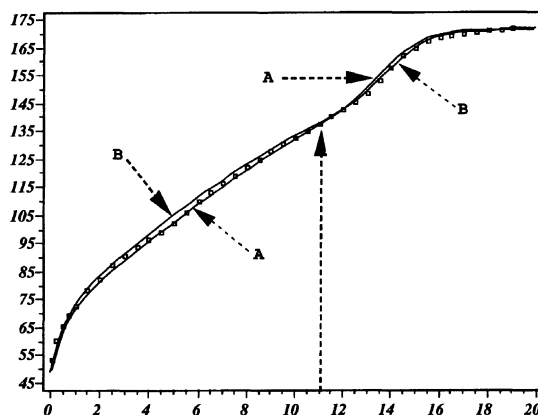


FIGURE 12

Superposition de la courbe observée ($\square$) de l'enfant n°1, de sa courbe théorique (A) si seules les mesures jusqu'à 10,5 ans sont prises en compte et sa courbe théorique (B) si seules les mesures de 11 ans à 19 ans sont prises en compte dans l'estimation des paramètres du modèle JPA2 (avec l'information a priori).

(B), est celle qui correspond à la modélisation sur la période 11 ans à 19 ans. Cette figure nous montre (comme pour les paramètres) que la qualité de la prévision de la période de l'enfance est meilleure que la prédiction de la période d'adolescence.

Références bibliographiques

- Abidi H. (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.
- Abidi H., Pontier J., Borms J., Duquet W. (1995), Courbes de croissance : Intérêt de la modélisation pour l'analyse de données longitudinales, *Rev. Statistique Appliquée*, XLIII, 3, 55-72.
- Abidi H., Borms J., Duquet W., Pontier J. (1996), Bayesian Estimation of the Parameters of a Nonlinear Model, An Application of Human Height, *Growth, Development and Aging*, 60, 3-4, 113-129.
- Abidi H., Borms J., Duquet W. (1996), ACCTM, logiciel de modélisation et analyse des courbes de croissance sur macintosh, *Cahiers d'Antropologie et Biométrie Humaine*, 14, 1-2, 327-342.
- 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, John Wiley & Son, 494 p.
- Berkey C. S. (1982), Bayesian Approach for a Nonlinear Growth Model, *Biometrics*, 38, 953-961.

- Bock R. D., Thissen D. (1976), Fitting Multi-Component Models for Growth in Stature, *Proc. 9th Int. Biometrics Conf.* 1, 431-442.
- Bock R. D., Thissen D. (1980), Statistical Problems of Fitting Individual Growth Curves, in «*Human Physical Growth and Maturation : Methodologies and Factors.*» (Johnston F. E., Roche A. F. and Susanne C., eds.). Plenum Press, New York, 265-290.
- Box G. E. P., Müller M. E. (1958), A Note on the Generation of Random Normal Deviates, *Ann. Math. Stat.*, 29, 610-611.
- Caussinus H., Ferré L. (1989), Analyse en composantes principales d'individus définis par les paramètres d'un modèle, *Statistique et Analyse des Données*, 41, 19-28
- Count E. W. (1943), Growth Patterns of Human Physique : An Approach to Kinetic Anthropometry, Part. I, *Human Biol.*, 15, 1-32.
- Fearn R. (1975), A Bayesian Approach to Growth Curves, *Biometrika* 62, 89-100.
- Geisser S. (1970), A Bayesian Analysis of Growth Curves, *Sankhya, Series A*, 32, 53-64.
- Hartley H. O., Booker A. (1965), Nonlinear Least Squares Estimation, *Annals of Math. Statist.*, 36, 638-605.
- Hauspie R. C. (1981), L'ajustement de modèles mathématiques aux données longitudinales de croissance, *Bull. Royal. Belge Antrop. Préhist.*, 92, 157-165.
- Hauspie R. C. (1989), Mathematical Models for the Study of Individual Growth Patterns, *Rev. Epidém. et Santé Publ.*, 37, 461-476
- 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
- Jennrich R., (1969), Asymptotic Properties of Nonlinear Least Squares Estimators, *The Annals of Mathematical Statistics*. 40, 2, 633-643.
- Jenss R. M., Bayley N. (1937), A Mathematical Method for Studying the Growth of a Child, *Human Biol.* 9, 556-663.
- Jolicœur P., Pontier J., Pernin M. O., Sempé M. (1988), A Lifetime Asymptotic Growth Curve for Human Height, *Biometrics*, 44, 995-1003.
- Jolicœur P., Abidi H., Pontier J. (1991), Human Stature : Which Growth Model? *Growth, Development et Aging*, 55, 129-132.
- Jolicœur P., Pontier J., Abidi H. (1992), Asymptotic Models for the Longitudinal Growth of Human Stature, *American Journal of Human Biology*, 4, 461-468.
- Kanefuji K., Shohoji H. (1990), On a Growth Model of Human Height, *Growth, Developpement & Aging*, 54, 155-165.
- Marquardt D. W. (1963), An Algorithm for Least Squares Estimation of Nonlinear Parameters, *J. Soc. Indust. Appl. Math.*, 11, 2, 431-441.
- Marubini E., Reselle L., Barghini G. (1971), A Comparative Fitting of the Gompertz and Logistic Functions to Longitudinal Height Data During Adolescence in Girls, *Human Bio.*, 43, 237-252.

- Pontier J., Jolicœur P., Abidi H., Sempé M. (1988), Croissance Staturale chez l'enfant : le modèle JPPS, *Biom. Praxim.*, 28, 27-44.
- Potthoff R. F., Roy S. N. (1964), A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems, *Biometrika*, 51, 313-326.
- Rao C. R. (1975), Simultaneous Estimation of Parameters in Different Linear Models and Applications to Biometric Problems, *Biometrics*, 31, 545-554.
- Shohoji T., Sasaki H. (1987), Individual Growth of Stature of Japanese, *Growth*, 51, 432-450.
- Shohoji T., Sasaki H., Kanefuji K. (1987), An Emperical Bayesian Approach for Individual and Average Growth Curves, *Proceeding of ISI Satellite Meeting on Biometry*, September 21, 1987, Osaka, 129-136.
- Strenio J. F., Weiisberg H., Bryk A. S. (1983), Empirical Bayes Estimation of Individual Growth Curve Parameters and their Relationship to Covariates, *Biometrics*, 39, 71-86.
- Theil H. (1971), *Principles of Econometrics*, New York : John Wiley & Son.