No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Modèles souples pour l'analyse des données longitudinales sur la santé des populations

Joel A. Dubin[1]

Résumé

L'étude de données longitudinales est essentielle si l'on veut observer correctement l'évolution des variables d'intérêt chez les personnes, les collectivités et les populations plus importantes au cours du temps. Les modèles linéaires à effets mixtes (pour les réponses continues observées au fil du temps), ainsi que les modèles linéaires généralisés à effets mixtes et les équations d'estimation généralisées (pour les réponses plus générales, telles que les données binaires ou les dénombrements observés au fil du temps) sont les méthodes les plus répandues pour analyser les données longitudinales provenant d'études sur la santé, même si, comme toute méthode de modélisation, elles ont leurs limites, dues en partie aux hypothèses sous-jacentes. Dans le présent article, nous discutons de certains progrès, dont l'utilisation de méthodes fondées sur des courbes, qui rendent la modélisation des données longitudinales plus souple. Nous présentons trois exemples d'utilisation de ces méthodes plus souples tirés de la littérature sur la santé, dans le but de démontrer que certaines questions par ailleurs difficiles peuvent être résolues raisonnablement lors de l'analyse de données longitudinales complexes dans les études sur la santé des populations.

MOTS CLÉS : longitudinales; courbes; modèles non linéaires; splines pénalisées; lissage.

1. Introduction

La modélisation des données longitudinales est décrite dans divers traités (par exemple, Verbeke et Molenberghs, 2000; Diggle et coll., 2002, Fitzmaurice et coll., 2004). La plupart des discussions (mais non toutes) présentées dans ces textes et la plupart des analyses longitudinales publiées dans les revues appliquées (comme celles traitant de la biomédecine et de la santé des populations) sont axées sur des modèles linéaires et des modèles linéaires généralisés pour données longitudinales. La corrélation entre les mesures répétées au cours du temps chez un même sujet est traitée de manière appropriée à l'aide de ces modèles. Dans le cas des modèles linéaires, l'approche habituellement utilisée est le modèle linéaire à effets mixtes (Laird et Ware, 1982), et pour les modèles linéaires généralisés, la majorité des chercheurs utilisent des modèles marginaux avec équations d'estimation généralisées (Liang et Zeger, 1986) ou à une approche particulière au sujet en utilisant des modèles linéaires généralisés à effets mixtes (par exemple, Breslow et Clayton, 1993; McCulloch et Searle, 2000). Cependant, de nombreux ensembles de données requièrent des approches de modélisation plus souples grâce à une paramétrisation différente des modèles populaires susmentionnés ou bien grâce à des modèles entièrement nouveaux. Nous donnerons trois exemples d'ensemble de données longitudinales qui requièrent ces types d'approches souples.

Tous les exemples présentés dans cet article de synthèse ont été publiés sous diverses formes (et dans diverses revues) dans la littérature biostatistique et biomédicale. À la section 2 qui suit, nous décrivons les exemples, un à la fois, ainsi que le problème d'intérêt défini et le modèle subséquent spécifié. Chaque approche est caractérisée par le fait que le taux de linéarité de la trajectoire d'intérêt au cours du temps n'est pas constant pendant toute la durée du suivi, cette trajectoire étant même relativement non linéaire pour certains exemples. À la section 3, nous présentons quelques remarques sommaires.

[1]Joel A. Dubin, University of Waterloo, 200 University Ave W., Waterloo, ON, Canada, N2L 3G1, jdubin@uwaterloo.ca

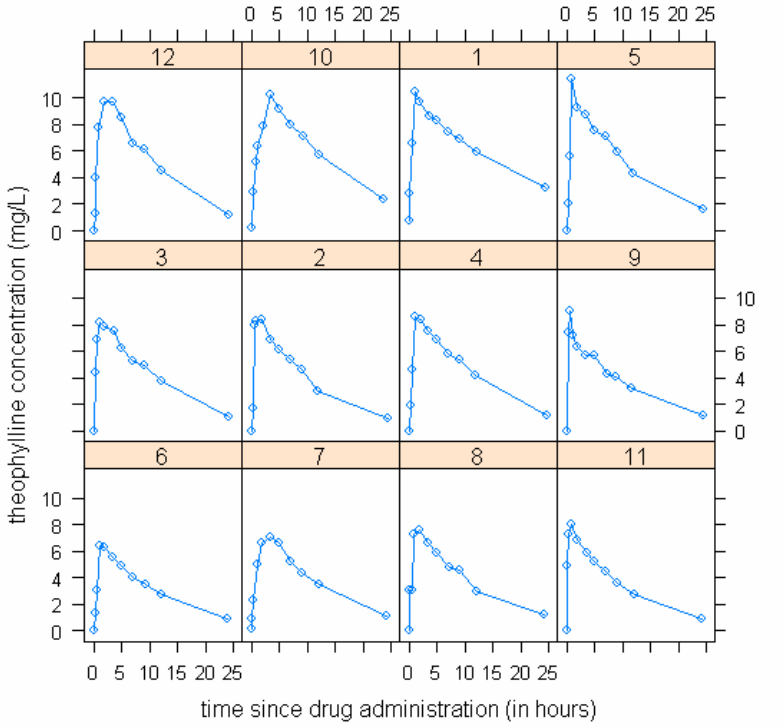
2. Exemples

2.1 Exemple 1 : Pharmacocinétique de la théophylline

La pharmacocinétique, qui est l'étude de la façon dont notre organisme traite la prise d'un médicament, s'intéresse depuis longtemps à des modèles non linéaires. Dans ce domaine, les modèles paramétriques non linéaires sont devenus la norme et les paramètres du modèle représentent des fonctions telles que l'absorption du médicament, ainsi que son élimination et sa clairance. Certains chercheurs se sont concentrés sur la modélisation longitudinale de ces processus, par exemple, par affectation aléatoire de divers niveaux de dosage à un groupe de sujets dans un essai clinique, y compris l'utilisation de modèles non linéaires à effets mixtes (par exemple, Pinheiro et Bates, 2000).

Pinheiro et Bates (2000) présentent un exemple de ce genre, une étude pharmacocinétique à plan expérimental auprès de 12 sujets, avec différences entre les doses reçues par les sujets. Les données longitudinales pour les 12 sujets sont présentées à la figure 1. Les trajectoires non linéaires sont évidentes pour chaque sujet, ainsi que les différences entre les sujets en ce qui concerne le taux et la crête d'absorption, ainsi que l'élimination et la clairance. Des effets aléatoires pourraient être nécessaires pour toute modélisation de ces données, au cas où les dosages ne pourraient expliquer par eux-mêmes ces différences entre les sujets.

Figure 1: pharmacokinetics of Theophylline concentration



Le modèle paramétrique non linéaire longitudinal (à effets mixtes) qui suit, qui est un modèle compartimental de premier ordre, est décrit dans Pinheiro et Bates (2000) pour les données sur la théophylline afin de refléter la concentration Y au temps t après une dose initiale d :

$$Y_{ij} = \frac{d_i E_i A_i}{C_i (A_i - E_i)} [\exp(-E_i t_{ij}) - \exp(-A_i t_{ij})] + \varepsilon_{ij} \qquad (1)_{ij}$$

Dans (1),

Y_{ij} est la concentration de théophylline chez le sujet i au temps t_{ij};

d_i est la dose initiale chez le sujet i;

E_i est le taux d'élimination constant chez le sujet i, où $E_i = \beta_1 + b_{1i}$, β_1 étant un effet d'élimination fixe au niveau de la population et b_{1i}, un effet d'élimination aléatoire propre au sujet de moyenne 0 et de variance σ^2_{b1};

A_i est le taux d'absorption constant chez le sujet i, où $A_i = \beta_2 + b_{2i}$, β_2 étant un effet d'absorption fixe au niveau de la population et b_{2i}, un effet d'absorption aléatoire particulier au sujet de moyenne 0 et de variance σ^2_{b2};

C_i est la clairance chez le sujet i, où $C_i = \beta_3 + b_{3i}$, β_3 étant un effet de clairance fixe au niveau de la population et b_{1i}, un effet de clairance aléatoire particulier au sujet de moyenne 0 et de variance σ^2_{b3};

ε_{ij} est le terme d'erreur intra-sujet chez le sujet i au temps j, supposé conditionnellement indépendant et de loi N(0, σ^2).

Bien que (1) semble être un modèle très complexe, chacun des paramètres a une interprétation biologique, plus précisément l'absorption, la clairance et l'élimination. Il s'agit d'une extension d'un modèle paramétrique non linéaire général : $Y_{ij} = f(X_{ij}, \beta, b_i) + \varepsilon_{ij}$, où les termes β représentent des effets fixes au niveau de la population. En outre, une souplesse suffisante a été intégrée dans (1) pour tenir compte de l'hétérogénéité entre les sujets qui ne peut être expliquée à l'aide des covariables observées, comme le niveau de dosage. Il s'agit d'un modèle complexe et souple, mais ayant une valeur explicative. En ce qui concerne l'ajustement de ce genre de modèles non linéaires paramétriques longitudinaux, il existe dans R et dans S-Plus une excellente fonction appelée *nlme*, rédigée par Pinheiro et Bates (2000), qui provient du progiciel/bibliothèque NLME. En SAS, on peut utiliser *Proc NLMIXED*.

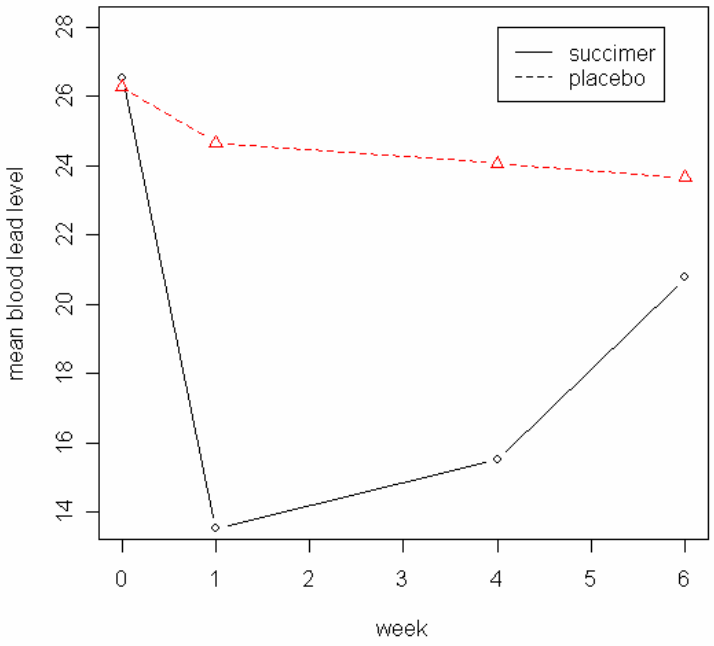
2.2 Exemple 2 : Essai clinique chez les enfants exposés au plomb

Ce deuxième exemple est tiré d'un essai aléatoire contre placebo d'un chélateur, le succimer, chez des enfants ayant une concentration sanguine élevée de plomb comprise entre 20 et 44 µg/dL. L'essai a été réalisé auprès de 100 enfants vivant dans des logements pauvres du centre-ville, âgés de 12 à 33 mois. Cet exemple est bien décrit dans Fitzmaurice et coll. (2004). L'objectif était de déterminer si le succimer réduirait plus que le placebo la concentration sanguine de plomb chez les enfants dont la concentration de plomb était élevée.

La figure 2 donne quatre mesures moyennes longitudinales (semaine 0, semaine 1, semaine 4 et semaine 6) et la trajectoire moyenne pour chaque groupe de traitement. Une ou deux caractéristiques clés se dégagent de l'examen du graphique. En premier lieu, on note une chute appréciable de la concentration sanguine du plomb après la première semaine chez le groupe recevant le succimer, comparativement à une diminution faible (éventuellement un effet placebo et(ou) une simple régression vers la moyenne) chez le groupe recevant le placebo. En deuxième lieu, après la chute initiale chez le groupe recevant le succimer, la concentration sanguine du plomb augmente de nouveau, tout en demeurant, en moyenne, plus faible que pour le groupe recevant le placebo au cours de la période de suivi de six semaines. Cette hausse de la concentration est due au rééquilibrage de la concentration sanguine du plomb après la perte initiale déclenché par l'organisme des enfants recevant le succimer, lequel entraîne une libération de plomb dans le sang par les os et les muscles.

Cet exemple requiert un modèle paramétrique nettement plus simple que celui utilisé à l'exemple 1. Ici, on s'intéresse initialement à un simple changement, dans une trajectoire moyenne par ailleurs rectiligne, de la réponse au cours du temps. Donc, on peut utiliser un modèle linéaire à effets mixtes, avec la création d'un point de changement unique. Il s'agit de la forme la plus simple de ce qui est autrement appelé un modèle *spline linéaire par morceaux* longitudinal comprenant un effet aléatoire propre au sujet et un seul nœud au point de changement.

Figure 2: mean blood levels over time from lead study



Fitzmaurice et coll. (2004) ont proposé le modèle spline linéaire par morceaux longitudinal suivant :

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1(\text{semaine}_{ij}) + \beta_2(\text{semaine}_{ij} - 1)_+ + \beta_3(\text{trt}_i * \text{semaine}_{ij}) + \beta_4(\text{trt}_i * (\text{semaine}_{ij} - 1)_+) + \varepsilon_{ij} \quad (2)$$

Dans (2),

Y_{ij} est la réponse longitudinale de la concentration de plomb chez le sujet i au temps semaine_{ij}, où le sujet i = 1 à 100 et le temps de la mesure j = 1 à 4;

β_0 est l'ordonnée à l'origine au niveau de la population;

b_{0i} est l'ordonnée à l'origine aléatoire propre au sujet i, que l'on suppose être indépendante et de loi $N(0, \sigma^2_{b0})$; cet effet aléatoire induit la corrélation entre les mesures répétées au cours du temps chez chaque enfant;

β_1 représente la pente au niveau de la population (pente initiale);

$(\text{semaine}_{ij} - 1)_+ = (\text{semaine}_{ij} - 1)$ si $(\text{semaine}_{ij} - 1) > 0$, et 0 autrement; ici, le point de changement (nœud) est spécifié à la semaine 1;

β_2 représente l'ajustement de la pente au niveau de la population à partir de la semaine 1;

β_3 représente l'interaction entre le groupe de traitement (1 = succimer, 0 = placebo) et la pente initiale au niveau de la population;

β_4 représente l'interaction entre le groupe de traitement (1 = succimer, 0 = placebo) et l'ajustement de la pente au niveau de la population à partir de la semaine 1;

ε_{ij} est le terme d'erreur intra-sujet chez le sujet i au temps j que l'on suppose suivre une loi $N(0, \sigma^2)$ et être conditionnellement indépendant.

Notons les **réponses prévues** qui suivent pour le sujet i au temps j, qui peuvent être inférées à partir de (2) :

$E(Y_{ij}) = \beta_0 + \beta_1(\text{semaine}_{ij}) + \beta_2(\text{semaine}_{ij} - 1)_+$ [pour le groupe du placebo]

$E(Y_{ij}) = \beta_0 + (\beta_1 + \beta_3)(\text{semaine}_{ij}) + (\beta_2 + \beta_4)(\text{semaine}_{ij} - 1)_+$ [pour le groupe du succimer]

Donc, pour le groupe du placebo, cela se traduit par une pente de β_1 avant le point de changement (semaine 1) et de $(\beta_1 + \beta_2)$ après le point de changement.

Pour le groupe du succimer, cela se traduit par une pente de $(\beta_1 + \beta_3)$ avant le point de changement (semaine 1) et de $(\beta_1 + \beta_3 + \beta_2 + \beta_4)$ après le point de changement.

En résumé, l'équation (2) représente un modèle longitudinal relativement simple, mais qui a été adapté afin de tenir compte d'un changement dans la trajectoire au cours du temps, y compris le fait que ce changement de trajectoire est une fonction du groupe de traitement. On pourrait soutenir qu'un deuxième point de changement, à la semaine 4, offrirait un meilleur ajustement aux données sur la concentration sanguine du plomb. Il n'est pas difficile d'imaginer l'extension de modèles tels que (2) afin d'inclure plusieurs points de changement (nœuds); autrement dit, des modèles tels que (2) peuvent facilement être étendus afin de pouvoir tenir compte d'une plus grande non-linéarité dans la trajectoire au cours du temps. Nous décrirons un modèle complexe de ce genre à l'aide d'un ensemble de données différent, à l'exemple 3 qui suit. Pour ce qui est de la programmation, ces types de modèles à point de changement fixe, tels qu'ils sont décrits par l'équation (2) plus haut, peuvent facilement être ajustés dans un programme de modélisation linéaire avec effets mixtes tels que *lme* (dans R ou S-Plus) ou *lmer* (dans R), ou dans *Proc Mixed* dans SAS.

2.3 Exemple 3 : Étude sur la santé respiratoire

Cet exemple, que nous ne présentons que brièvement ici, provient d'une étude de l'association entre la santé respiratoire et la pollution par des particules inhalables chez un échantillon de 41 écoliers, sur une période de 109 jours consécutifs en 1991. On a mesuré quotidiennement le débit maximal expiratoire (DME), ainsi que la quantité de particules (P) et plusieurs variables météorologiques, comme la température la plus basse de la journée (TB). Malheureusement, aucune figure illustrant cet exemple n'était disponible en vue de son inclusion dans le présent article.

Le modèle linéaire à effets mixtes initial qui suit est une extension du modèle linéaire utilisé par les premiers chercheurs (Pope et coll., 1991):

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1(P)_j + \beta_2(TB)_j + \beta_3(\text{temps})_j + \varepsilon_{ij} \qquad (3)$$

Dans (3), la réponse Y_{ij} représente la mesure du DME chez l'enfant i (i = 1 à 41) au temps j (j = 1 à 109), et b_{0i} est un effet aléatoire unique entre sujets représentant la variabilité inexpliquée (hétérogénéité) entre les enfants aux niveaux de référence du DME. En outre, notons que l'indice i est supprimé pour P, TB et le temps, car ces variables ne sont pas particulières à un individu dans la présente étude.

Coull et coll. (2001) se sont rendus compte que le modèle susmentionné était inadéquat, car des relations plus souples étaient nécessaires entre DME et TB, ainsi qu'entre DME et le temps. En outre, il fallait tenir compte de la variabilité inexpliquée en ce qui concerne la relation entre DME et P, ainsi que de la corrélation sériale; donc, Coull et coll. (2001) ont proposé le *modèle additif mixte* suivant pour ces données :

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})(P)_j + f(TB)_j + g(\text{Temps})_j + \varepsilon_{ij} \qquad (4)$$

Le modèle (4) reflète les changements ci-après comparativement au modèle (3) :

b_{1i} a été ajouté pour tenir compte de la variabilité inexpliquée entre les enfants dans la relation entre DME et P;

ε_{ij} tient maintenant compte de la corrélation sériale, de sorte que $\varepsilon_{ij} = \rho\varepsilon_{i,j-1} + \alpha_{ij}$

f et g représentent les relations non linéaires entre DME et TB, ainsi qu'entre DME et le temps, respectivement. Coull et coll. (2001) ont exprimé ces relations non linéaires à l'aide de fonctions lisses, au moyen de *splines pénalisées* (par exemple, Eilers et Marx, 1996), c'est-à-dire une application plus complexe des splines que celle utilisée dans le modèle (2) de l'exemple du plomb chez les enfants.

Aussi complexe que puisse paraître le modèle (4) susmentionné, il est possible d'utiliser pour les modèles mixtes additifs, y compris les modèles tels que celui présenté en (4) pour les données respiratoires, un logiciel standard pour modèle linéaire à effets mixtes, tel que *lme* dans R et S-Plus, et *Proc MIXED* dans SAS. Pour les modèles additifs plus complexes, un logiciel spécialisé est nécessaire. Le progiciel SemiPar dans R pourrait être une option, et les modèles additifs mixtes complexes sont décrits dans Ruppert et coll. (2003). Pour les modèles additifs mixtes généralisés (par exemple, Lin et Zhang, 1999; Wood, 2006), il est possible d'utiliser le progiciel mgcv dans R.

3. Discussion

Nous avons présenté dans cet article de synthèse plusieurs exemples informatifs en mettant l'accent sur les données longitudinales pour lesquelles une modélisation complexe est nécessaire. Ces exemples ne représentent que quelques approches parmi le grand nombre proposé dans la littérature statistique et biostatistique, quoique plusieurs de ces approches souples ne soient pas encore devenues d'usage courant dans la littérature sur la santé des populations et dans la littérature biomédicale. Certaines autres pourraient être classées dans des catégories de *modèles à coefficients variables*, où les paramètres β proprement dits sont une fonction du temps (par exemple, Chiang et coll., 2001; Lin et Ying, 2001; Fan et Li, 2004), de modèles entièrement non paramétriques fondés sur des courbes (par exemple, Rice et Silverman, 1991; Dubin et Müller, 2005), et de modèles analytiques à données fonctionnelles (par exemple, Ramsay et Silverman, 2005; Yao et coll., 2005), les deux dernières approches utilisant des courbes (représentant la trajectoire des réponses au cours du temps) comme fondement de l'analyse. Heureusement, on a observé récemment chez de nombreux chercheurs élaborant ces nouvelles méthodes un effort concerté en vue de créer des logiciels qui permettent d'utiliser ces méthodes longitudinales souples plus fréquemment dans les études appliquées sur la santé. Les études décrites dans le présent article sont des exemples de cet effort.

Références

Breslow, N.E., et Clayton, D.G. (1993), "Approximate inference in generalized linear mixed models", *Journal of the American Statistical Association*, 88, 125-134.

Chiang, C.T., Rice, J.A., Wu, C.O. (2001), "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables", *Journal of the American Statistical Association*, 96, 605-619.

Coull, B.A., Schwartz, J., et Wand, M.P. (2001), "Respiratory health and air pollution: Additive mixed model analyses", *Biostatistics*, 2: 337-349.

Diggle, P.J., Heagerty, P., Liang, K.-Y., et Zeger, S.L. (2002), *Analysis of Longitudinal Data*, 2nd ed., Oxford: Oxford.

Dubin, J.A., et Müller, H.G. (2005), "Dynamical correlation of multivariate longitudinal data", *Journal of the American Statistical Association*, 100, 872-881.

Eilers, P.H.C. et Marx, B.D., (1996), "Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11: 89-121.

Fan, J.Q., Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data", *Journal of the American Statistical Association*, 99, 710-723.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H. (2004), *Applied Longitudinal Analysis*, Hoboken: Wiley.

Laird, N.M., et Ware, J.H. (1982), "Random-effects models for longitudinal data", *Biometrics*, 38, 963-974.

Liang, K.-Y., et Zeger, S.L. (1986), "Longitudinal data analysis using generalized linear models", *Biometrika*, 73, 13-22.

Lin, D.Y., Ying, Z. (2001), "Semiparametric and nonparametric regression analysis of longitudinal data", *Journal of the American Statistical Association*, 96, 103-113.

Lin, X., et Zhang, D. (1999), "Inference in generalized additive mixed models by using smoothing splines", *Journal of the Royal Statistical Society, Series B*, 61: 381-400.

McCulloch, C.E. et Searle, S.R. (2000), *Generalized, Linear, and Mixed Models*, New York: Wiley.

Pinheiro, J.C., et Bates, D.M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag.

Pope, C.A., Dockery, D.W., Spengler, J.D., et Raizenne, M.E. (1991), "Respiratory health and PM_{10} pollution: A daily time series analysis", *American Review of Respiratory Disease*, 144, 668-674.

Ramsay, J.O. et Silverman, B.W. (2005), *Functional Data Analysis, 2nd ed.*, New York: Springer-Verlag.

Rice, J.A. et Silverman, B.W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves", *Journal of the Royal Statistical Society, Series B*, 53: 233-243.

Ruppert, D., Wand, M.P., et Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge: Cambridge.

Verbeke, G. et Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Wood, S. (2006), *Generalized Additive Models: An Introduction with R*, London: Chapman & Hall.

Yao, F., Müller, H.G., Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data", *Journal of the American Statistical Association*, 100, 577-590.