Etude et conception d'un modèle mixte semiparamétrique stochastique pour l'analyse des données longitudinales environnementales.

Kairou Moumouni

N° d'ordre :

THÈSE

présentée à

L'UNIVERSITÉ DE RENNES II,
Haute Bretagne

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES II
Mention : Mathématiques appliquées
Spécialité : Statistique

par

Kairou MOUMOUNI

TITRE DE LA THÈSE

ETUDE ET CONCEPTION D'UN MODÈLE MIXTE SÉMIPARAMÉTRIQUE STOCHASTIQUE POUR L'ANALYSE DES DONNÉES LONGITUDINALES ENVIRONNEMENTALES

◇ ◇ ◇ ◇ ◇ ◇

Soutenue publiquement le 12 Décembre 2005 devant le jury composé de :

M.Michel Carbon	Directeur de recherche
M.Dominique Dehay	Président du jury
Mme Hélène Jacqmin-Gadda	Rapporteur
M Christian Lavergne	Rapporteur
Mme Séverine Deguen	Examinateur
M. Philippe Garat	Examinateur

Laboratoire EGERIES-LERES, ENSP

Remerciements

Il me plait en ce moment d'adresser mes remerciements à tous ceux qui, d'une manière ou d'une autre, ont contribué à l'aboutissement de ce travail et à rendre meilleures ces années de thèse.
Tout d'abord, je tiens à remercier le Professeur Michel Carbon d'avoir bien voulu diriger ce travail de thèse.
Merci au Professeur Dominique Dehay d'avoir bien voulu présider ce jury. Je tiens aussi à exprimer toute ma gratitude au Dr. Hélène Jacqmin-Gadda et au Professeur Christian Lavergne d'avoir accepté de rapporter ce travail de recherche.
Je tiens particulièrement à remercier Philippe Garat pour son encadrement et ses encouragements et Sèverine Deguen pour son aide précieuse et indispensable. Tous les deux ont été à l'origine de ce projet.
Pour mener à bien ce projet, j'ai été accueilli à l'École Nationale de Santé Publique dans les structures d'ÉGERIES-LÈRES, j'ai pu ainsi disposer des moyens nécessaires. Je tiens à remercier les différents responsables de cette institution pour m'avoir offert cette chance, qu'ils trouvent à travers ces mots le témoignage de toute ma reconnaissance.
Je ne saurais finir sans remercier Eric Maztner-Lober et tous les membres du Laboratoire de statistiques de Rennes 2 pour leur soutien.
La salle à café d'ÉGERIES est un endroit convivial et sans doute le plus convivial que j'ai connu en France, merci à ceux qui créent cette ambiance et qui sauront se reconnaître.
J'ai une pensée particulière pour tous les collègues-doctorants et jeunes docteurs rencontrés à l'ENSP et dans les associations, et merci pour bien de choses partagées.
Enfin à ma famille et surtout à Christelle pour avoir su me faire croire que c'est possible même dans les moments les plus critiques.

Table des matières

Chapitre 1

Introduction aux données environnementales longitudinales

1.1 Résumé et présentation

L'objet de ce chapitre est de présenter la problématique des données auxquelles nous nous intéressons et l'intérêt de leur modélisation.
Le cadre de collecte et les spécificités des données sont décrites. Pour fixer les idées, quelques exemples de données sont exposés ainsi que l'apport de leur modélisation à la compréhension du phénomène dont elles sont issues.
Ensuite deux approches d'analyse des données longitudinales environnementales sont décrites. Après la présentation de ces deux méthodes d'analyse, nous introduisons celle que nous développons dans cette thèse. C'est une extension de l'approche des données longitudinales que Diggle et *al* développe dans [46, 12]. Pour clore cette partie, nous donnons l'organisation du mémoire et fixons les points sur lesquels porte notre contribution.

1.2 Introduction

Dans les études statistiques, les données proviennent essentiellement de deux types de mécanismes de collecte : l'observation et l'expérimentation. Les données d'expérimentation sont les données dont le mécanisme de production est contrôlé par l'expérimentateur. Celui-ci s'assure de l'élaboration d'un schéma des conditions de production (plans d'expérience) de ces données permettant d'atteindre le but poursuivi. Lorsque ce mécanisme de production n'est pas contrôlé, les données sont désignées comme des données d'observation.
Dans le domaine environnemental, les conditions de collecte sont mises en place dans un cadre de suivi par rapport à la réglementation (seuils de référence à ne pas dépasser). De ce fait, la fréquence des observations n'est pas régulière. Les bases des données ainsi constituées sont plutôt des données d'observation.
Très souvent, ces données environnementales proviennent d'une configuration spatiale fixée par la mise en place de capteurs ou de stations d'observation. Les observations successives

provenant d'un même site (station de mesure) conduisent à des mesures répétées. Le facteur de répétition de ces mesures est, dans la plupart des cas, le temps. Cependant, il se peut que d'autres facteurs d'observation servent à référencer cette répétition.
Par la suite, nous entendrons par données longitudinales environnementales ce type de données. En général ce sont des observations faites *par routine*. C'est souvent le cas dans le domaine de la santé publique et de l'environnement.
Le contexte, dans lequel nous nous situons, correspond au cas pratique où la localisation spatiale est fixée (exactement connue), parfois mise en place suivant un a priori de l'expert. Les mesures faites à l'échelle des sites fournissent alors suffisamment d'information sur le phénomène étudié. Par exemple, dans un suivi de la qualité de l'eau de consommation, les mesures faites au niveau des stations de captage donnent une idée précise sur la qualité de celle-ci.
Les données observées correspondent exactement à des caractéristiques physiques, chimiques ou bactériologiques du milieu (par exemple, le niveau de pesticides (caractéristique chimique) dans l'eau (milieu) de consommation). C'est l'exposition du milieu désignée sous le nom d'exposition environnementale. De manière analogue, il existe l'exposition individuelle portant sur la dose de ces caractéristiques dans le corps humain avec des seuils de référence. Pour certaines subtances toxiques, le niveau de toxicité admissible par le corps humain fixé par l'OMS existe et est appelée valeur toxicologique de référence (VTR). La connaissance de l'exposition environnementale peut contribuer à la détermination de la dose individuelle d'exposition.
Dans un cadre de données recueillies par région, la station de mesure peut être incluse dans un groupe et ce groupe emboîté dans un autre groupe à un niveau supérieur. Cette configuration des stations de mesure correspondra alors à un modèle emboîté à plusieurs niveaux avec une hiérarchisation de ces niveaux. Il apparaît donc intéressant dans une telle configuration de pouvoir comparer les niveaux d'emboîtement. Ce type d'approche, très répandu est désigné sous le nom d'analyse multi-niveaux.
L'élaboration de telles bases de données d'exposition environnementale, et la tenue de différents registres (PMSI et autres) pour des maladies spécifiques comme le cancer dans les mêmes zones nécessitent le développement de méthodologies susceptibles de permettre la mise en valeur de ces données emmagasinées et de fournir des réponses aux besoins d'analyses sous-jacentes et autres requêtes. La modélisation de l'exposition environnementale apparaît alors comme une étape importante de l'évaluation efficace entre *facteurs de risque*, par exemple liés à la pollution et les *indicateurs de santé*.
Dans ce qui va suivre, nous allons exposer les caractéristiques des données et situer l'intérêt d'une étude de ces données. Il faut noter que la complexité de la structure de ces données est amplifiée par le schéma de collecte (échantillonnage spatio-temporel).

1.3 Présentation du type des données

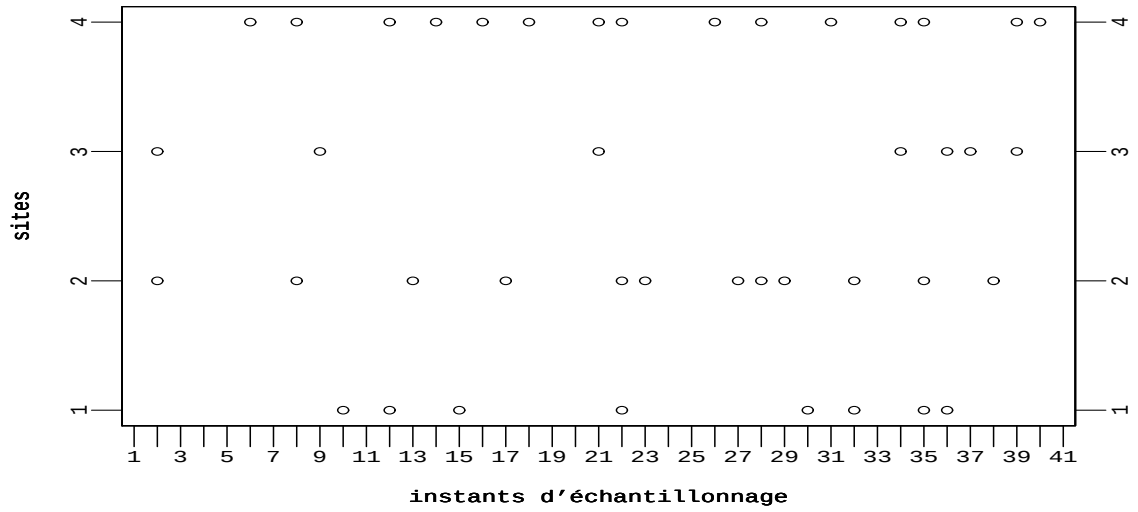
Cette étape permet de définir le type des données que nous conviendrons de désigner dans ce contexte comme *données longitudinales environnementales*. Le terme *données longitudinales*

suggère évidemment des observations faites au cours du temps. Ces observations peuvent provenir de plusieurs sites (stations de mesure, capteurs, . . . , etc). Ce sont des mesures répétées, dont la répétition vient du fait qu'il existe un ensemble d'observations provenant d'un individu (ici le site), et qui sont ordonnées suivant l'évolution du facteur de répétition au niveau de cet individu. La méthodologie développée dans cette thèse porte sur des données longitudinales provenant du champ de la santé-environnement. Le contexte relatif à l'effet de l'environnement, impose de chercher si la station joue un rôle sur le phénomène observé.
Les exemples suivants sont des données longitudinales environnementales :

- L'unité d'observation peut être une station de mesure pour le prélèvement de concentrations de pesticide ou de nitrate dans les eaux superficielles. Pour chaque station, les mesures de polluants sont effectuées de façon répétitive en plusieurs occasions référencées par le temps. Dans le cas idéal, la mesure des polluants est effectuée avec celle des variables météorologiques susceptibles d'expliquer le phénomène.
- L'unité d'observation peut être un arbre dans un système forestier. Les arbres peuvent être associés à des systèmes de culture différents et les mesures des diamètres de chaque arbre sont répétées pour chaque système de culture et à différents stades de vie de l'arbre.
- L'unité d'observation peut aussi être des points de contrôle de la qualité de l'eau de baignade sur les plages. Ces points sont choisis suivant un a priori de l'expert. Les mesures répétées de la présence bactérienne (entérocoques, *E.coli*) dans les prélèvements d'eau de mer sont des indicateurs de la qualité de l'eau de baignade.

Le terme de données longitudinales désignera donc tous types de données environnementales présentant les caractéristiques décrites et similaires par sa texture et sa structure aux exemples ci-dessus.
Il faut noter que, dans le cadre de la santé publique, les fréquences de mesures sont accélérées suivant le contexte socio-économique associé au phénomène étudié. Par exemple, un regain d'intérêt des médias sur le niveau élevé des teneurs des pesticides et des nitrates dans les eaux ou l'envahissement des plages par la marée verte entraînent une accélération des fréquences de ces mesures. Autre exemple, le suivi régulier de la qualité des eaux de baignades des plages n'est effectué que pendant l'été, saison de la fréquentation élevée.
Les données provenant de différentes stations présentent alors des données manquantes quand celles-ci sont regroupées pour le besoin d'analyse. Les différentes dates d'observation peuvent alors être distinctes d'un individu à l'autre. Le plan d'échantillonnage devient alors déséquilibré et asynchrone. Ce type de plan d'échantillonnage correspond globalement au plan d'échantillonnage décrit par la figure (cf.1.1).

FIG. 1.1 – Exemple de plan d'échantillonnage d'un monitoring environnemental : les petits cercles représentent la survenue d'une mesure sur l'un des 4 sites (fréquence de mesures au niveau de 4 stations). En ordonnées, sont représentés les 4 sites où les observations sont réalisées, et en abscisses les instants d'observation.



L'intérêt scientifique peut porter sur la courbe d'évolution ; c'est-à-dire la façon dont la variable étudiée varie au cours du temps avec une évaluation des effets de l'environnement sur cette évolution et ceux qui peuvent être imputés de façon spécifique aux variables météorologiques.
Les méthodes de régression sont applicables et la recherche des différentes sources de variation de nos données s'impose. Cet objectif, qui paraît relativement simple dans le cadre des données d'expérimentation, l'est beaucoup moins dans le cas des données d'observation. Aux instants précis de la collecte des mesures de la variable, les mesures des covariables associées ne sont pas effectuées, ce qui rend ces données complexes. Cela nécessite une approche tenant compte de cette complexité et des spécificités de ces données.

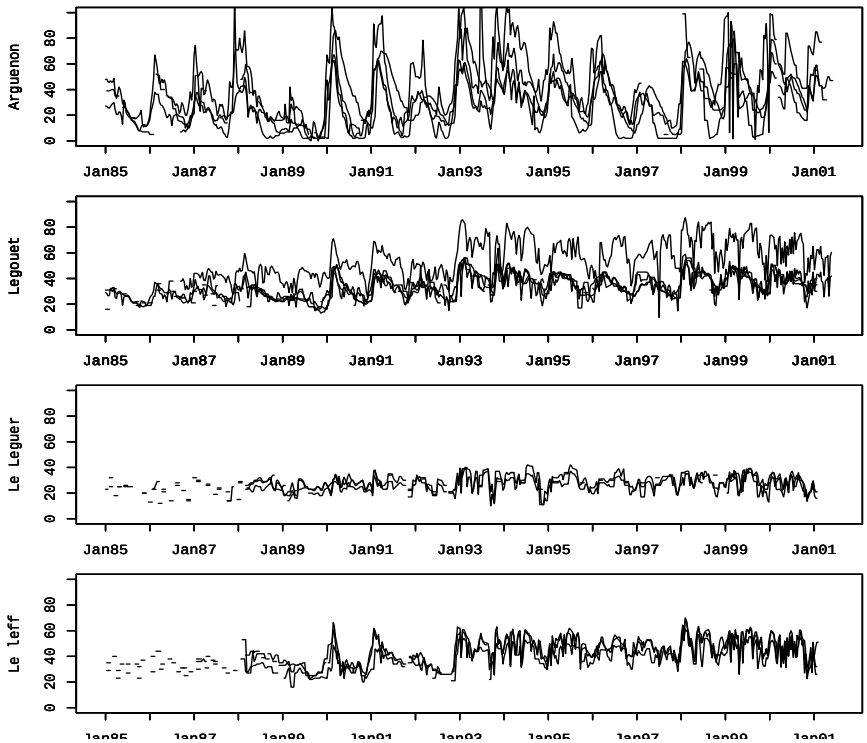
1.4 Quelques exemples de données

Les deux exemples suivant sont des jeux de données réelles qui vont permettre de fixer les idées sur les aspects décrits précédemment.

1.4.1 Les données de concentration de nitrates dans les bassins versants de Bretagne

Dans le cadre du suivi des niveaux de concentration de nitrates dans les eaux de surface en Bretagne, des stations de mesure de ces concentrations ont été mises en place, sur différents bassins versants tels que le gouet, l'Arguenon, en différents points correspondant parfois à des stations de traitements des eaux. Le nombre de stations de mesures par bassin versant varie entre trois et quatre. La période d'observation s'étend sur une période d'environ 16 ans. La fréquence des mesures change et varie de la mesure quotidienne à des mesures mensuelles suivant les stations mais aussi au cours de la période d'observation.
Les séries de mesures disponibles présentent des courbes d'évolution similaires au sein d'un même bassin versant. Sur l'ensemble de quatre stations, les observations sont faites avec une irrégularité remarquable conforme à la représentation graphique (1.1).
Les représentations graphiques de la figure ci-dessous ont été faites avec en abscisses les instants de mesures en décade (10 jours). Sur ces représentations graphiques, l'ensemble des courbes des concentrations de nitrates des 4 bassins versants semblent exhiber une composante commune (voir figure 1.2).

FIG. 1.2 – Exemple de données de concentration de nitrates (mg/l) des eaux superficielles issues de 4 bassins versants bretons. Chacune des 4 figures représente un bassin versant. Chaque bassin versant comporte 3 à 5 courbes représentant les mesures issues de ses stations d'observation. En abscisses le temps d'évolution est référencé par les mois.



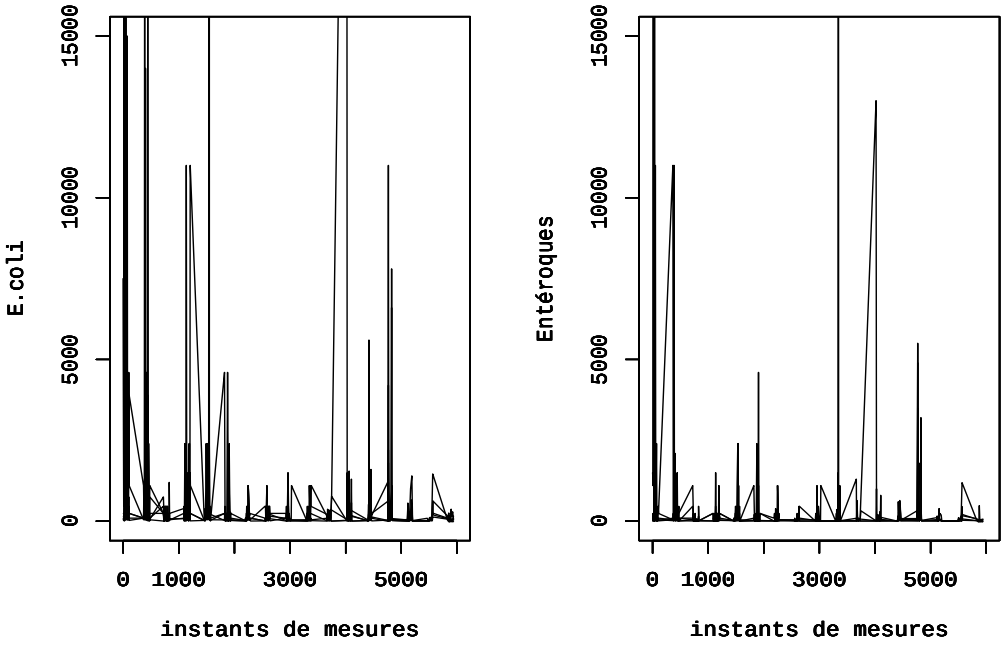
Quelques stations de mesure montrent un niveau assez élevé de concentration du polluant. L'intérêt d'une étude sur ces données consiste :
– à trouver l'allure de la courbe moyenne propre à l'ensemble des stations ;
– à tester si les différences entre bassins versants ou stations de mesure sont significatives.

1.4.2 Les données de suivi de la qualité des eaux de baignade

Pour juger de la qualité des eaux de baignade, les pratiques en santé publique consistent à effectuer des prélèvements des eaux de plage et à dénombrer le nombre de germes présents dans 100 ml d'eau. Les indicateurs microbiologiques retenus, dans le cas suivant, sont les entérocoques intestinaux et les *Escherichia coli*. Ces indicateurs retenus sont supposés avoir un comportement similaire et être influencés de la même façon par les facteurs environnementaux.
Les mesures sont effectuées sur différentes plages d'une station balnéaire donnée (Ille et Vilaine). Dans cet exemple, les plages constituent les stations de mesure et pour chaque plage, les mesures sont répétées pour les deux indicateurs microbiologiques, durant l'été, saison de fréquentation des plages. En dehors de cette période le suivi est interrompu.

La représentation graphique des données brutes (figure 1.3) n'exhibe aucune information particulière.

FIG. 1.3 – Exemple de données de suivi de la qualité des eaux de baignade. La planche de gauche représente les courbes de mesures brutes des bactéries *E.coli* contenues dans 100 ml d'eau. Ces mesures proviennent de différentes plages (5) représentées par les courbes. L'irrégularité des mesures rend les différentes courbes difficiles à observer. La planche de droite représente les courbes des mesures brutes des entéroques intestinaux.



L'intérêt d'une telle étude peut porter sur la courbe écologique des deux indicateurs afin de déceler si ces deux populations de germes réagissent effectivement de la même façon au cours du temps aux facteurs environnementaux à partir des donnés disponibles, et en cas de réponse positive, nous pouvons conclure qu'il était justifié de les considérer comme des indicateurs équivalents.

1.5 Méthodes d'analyse

Il est donc clair que l'intérêt d'une analyse statistique des données de ce type réside en partie dans la détermination de la courbe d'évolution associée au phénomène étudié sur l'ensemble des sites d'observation (composante commune à l'ensemle des sites) mais aussi dans l'évaluation du rôle particulier de chaque station sur les observations faites au niveau de celle-ci, (courbe spécifique à l'individu). La détermination de la courbe d'évolution passe

par celle de la relation linéaire ou non linéaire voire fonctionnelle entre la variable d'intérêt et les variables explicatives disponibles ou prédicteurs. La prise en compte de l'effet spécifique des stations sur un phénomène environnemental, est rarement pris en compte dans les méthodologies statistiques courantes. Très souvent les méthodes statistiques utilisées sont des méthodes de statistiques spatiales et ces méthodes ne sont pas tout à fait adaptées pour ce genre de données.
Enfin le plan d'échantillonnage irrégulier des données et la complexité de la dynamique des phénomènes environnementaux font que les méthodes de régression linéaires classiques peuvent paraître rigides. Au sens où, elles permettent certes une facile exploitation des covariables, une facilité d'interprétation et jouissent d'une immense popularité. Cependant elles sont souvent inadéquates et *trop restrictives* pour capturer les diverses formes des courbes d'évolution des variables d'intérêt. La recherche de méthode combinant la facilité d'interprétation avec des approches non linéaires et plus flexibles pour la prise en compte de cette complexité de la dynamique des phénomènes environnementaux, s'impose.

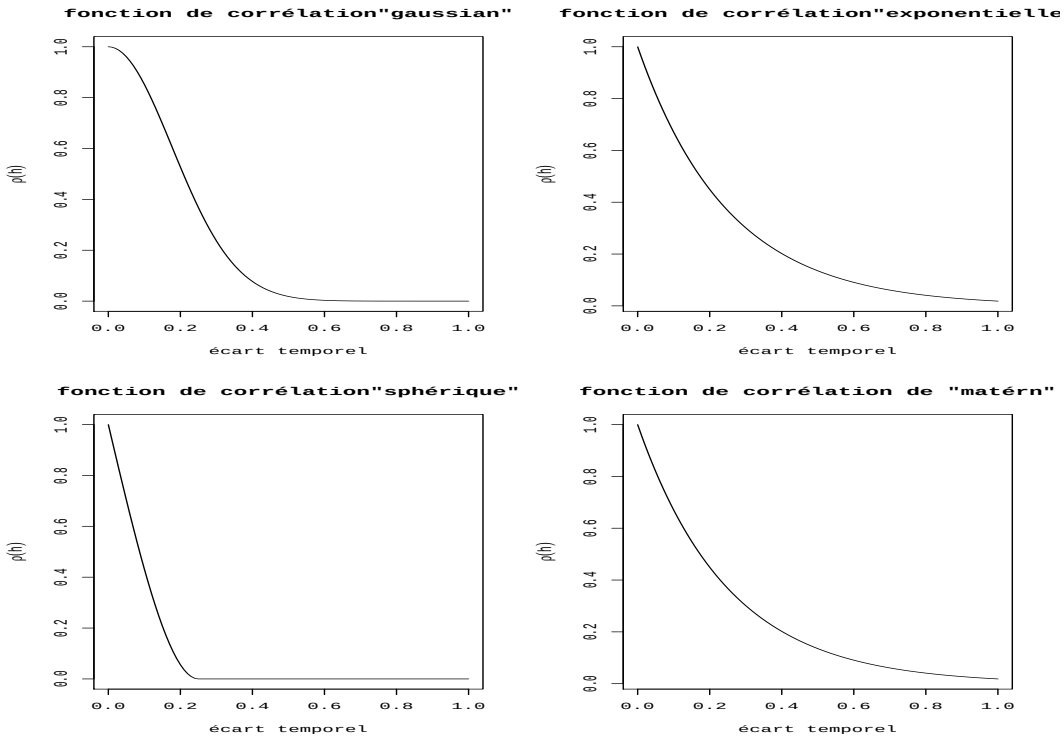
1.5.1 Approche spatio-temporelle des données environnementales et méthodes ARMAX

Compte tenu de la répartition spatiale des points d'observation et de la répétition des mesures, les méthodes des statistiques spatiales sont très souvent utilisées pour l'analyse des données environnementales. L'approche standard des statistiques spatiales est basée sur l'hypothèse que la dépendance spatiale entre deux points de l'espace est une fonction de la distance entre ces points. Les dépendances spatiales entre les stations de mesure d'un bassin versant sont prises en compte par les fonctions de corrélation. L'essentiel de l'approche réside dans le choix et/ou la formulation de la structure de covariance pour représenter l'aspect spatial des données. Les travaux de Matheron [29] et Cressie [9] ont servi de base pour le développement de la méthode.
L'approche générale est basée sur les structures de covariance des processus gaussiens stationnaires. Cependant, celles-ci dans la pratique peuvent présenter des restrictions car les données spatio-temporelles ne sont pas toujours des processus gaussiens stationnaires. De même une structure de covariance stationnaire peut entrainer dans l'estimation d'un processus spatial non stationnaire, suivant les zones géographiques, un lissage soit excessif soit insuffisant, bref dans tous les cas, inadapté au processus étudié. La modélisation des phénomènes spatio-temporels survenant dans le domaine de l'environnement nécessite en général l'évaluation d'une structure de covariance non stationnaire. Dans la représentation d'un tel phénomène, il est nécessaire de tenir compte des dépendances spatiales mais aussi de la façon dont celles-ci évoluent.
L'approche peut se résumer en deux étapes principales une fois les données spatiales disponibles :

- estimation du variogramme ou de la structure de covariance et sélection du modèle
- prédiction du variogramme ou de la structure de covariance *du nouveau point* et interpolation spatiale (kriging ou krigeage).

Les fonctions de covariance les plus répandues pour les processus gaussiens stationnaires sont : la fonction de corrélation exponentielle, la fonction de corrélation gaussienne, la fonction de corrélation sphérique ou encore de Matèrn. (cf. figure 1.4)

FIG. 1.4 – Exemples de fonctions de corrélation.



Les processus physiques étudiés dans l'environnement ne s'ajustent pas toujours à des processus gaussiens stationnaires. La prise en compte de la non stationnarité exige la recherche des méthodes alternatives. On peut citer la méthode de déformation spatiale bijective de Sampsom et Guttorp [36]. Dans cette méthode, la déformation est appliquée sur l'espace des indices $\mathbb{R}^n$. Il faut ajouter la récente méthode de Nott et Dunsmuir [30] illustrant les méthodes générales d'estimation prenant en compte les variabilités spatio-temporelles. Ils donnent une méthode d'estimation d'une structure de covariance non stationnaire par une décomposition de celle-ci en structures de covariance stationnaires.
Toutes les variables intégrées dans ces modèles sont considérées comme des variables spatiales, elles sont traduites en coordonnées spatiales. Les méthodes des statistiques spatiales semblent inadaptées pour l'analyse des données exposées. Le type des données auxquelles nous nous intéressons présentent peu de points spatiaux et un grand nombre d'observations répétées. Malgré la répartition spatiale des sites, ces données se présentent comme des séries chronologiques avec une tendance temporelle marquée, des fluctuations saisonnières et potentiellement une composante stochastique.
En plus, l'effet temporel des facteurs environnementaux (géologie locale, topographie, réseau

hydrogéologique, politiques locales, effets d'installations industrielles locales) ne pourra pas être intégré dans un tel modèle de type spatial.
Rares sont les approches spatio-temporelles (state space methods) qui prennent en compte conjointement les aspects spatial et longitudinal. Les méthodes recherchées doivent permettre une modélisation des données environnementales avec un ajustement de séries multiples (multisites), une décomposition en une tendance temporelle commune et des composantes spécifiques pour caractériser les différents sites.
Des modèles de kriging à coefficients de régression aléatoires de forme **ARMAX** peuvent offrir des approches alternatives. Parmi ces modèles, nous pouvons citer celui de Franke & Grunder [15] pour modéliser des données environnementales. Une particularité de ces données est qu'elles proviennent d'un faible nombre de points d'observation, mais avec une taille importante d'observations répétées au cours du temps. Par exemple, les données peuvent être des concentrations de gaz SO_2, associées à plusieurs paramètres météorologiques observés sur une période assez étendue. La méthode permet l'intégration des variables météorologiques dans le modèle. Leur modèle de base se présente sous la forme :

$$\zeta(t,x) = f(x)^T\beta(t) + \eta(t,x),$$

où l'opérateur de transposition est désigné par ^{T}, $t \in \mathbb{Z}$, $x \in \mathbb{R}^2$ et $f(x)$ est un vecteur de fonction de x.
Le coefficient $\beta(t)$ est un vecteur aléatoire dont les éléments sont des processus autorégressifs indépendants de la composante $\eta(t,x)$. La partie $f(x)^T\beta(t)$ représente la tendance globale temporelle des différents sites et $\eta(t,x)$ la composante spécifique à un site donné.
La méthode s'applique avec une procédure adaptative permettant de prendre en compte l'acquisition de nouvelles observations. Cependant la procédure adaptative dans l'estimation et le fait que le coefficient $\beta(t)$, soit un processus autorégressif nécessitant une fréquence régulière des observations, rendent sur un plan pratique la mise en oeuvre de la méthode délicate.

1.5.2 Analyse par composantes principales et méthodes dérivées

La disponibilité de plusieurs séries exhibant une dynamique commune conduit souvent à l'utilisation des méthodes dérivées de l'analyse en composantes principales. La méthode pour le type des données exposées se décline en deux approches.
Une approche *naïve* consiste à compléter les plages des données manquantes des variables dépendantes découlant de la différence entre les instants des mesures des différents individus (stations de mesure) et à développer une analyse en composantes principales sur les courbes des différentes stations. Cette approche permet d'extraire une composante commune propre à l'ensemble des stations susceptible de représenter la dynamique du phénomène étudié. La modélisation du phénomène tourne alors autour de cette composante commune. Bien évidemment, cette approche présente quelques insuffisances. La méthode d'analyse en composantes principales est développée sur l'ensemble des données en affectant le même poids

aussi bien aux données observées donc d'origine que celles complétant les données manquantes. Ces deux catégories des données ne fournissent pas la même qualité d'information, il apparaît donc convenable de leur attribuer des poids différents.
Une autre approche plus performante est celle proposée par James & Hastie [22] à travers leur modèle développé comme une extension du modèle mixte nonparamétrique des courbes des données fonctionnelles déséquilibrées de Rice & Wu [33]. Ces derniers approchent la matrice de covariance comme un produit tensoriel de splines. Les trajectoires individuelles sont obtenues alors comme des estimateurs BLUP, en combinant l'information individuelle complétée par l'information collective. Le modèle proposé par James & Hastie [22] se présente sous la forme :

$$Y_i = B_i\beta + B_i\gamma_i + \epsilon_i,$$

où B_i est une base de B-splines de dimension $(n_i \times q)$, avec n_i le nombre d'observations de l'individu i.
Cette base de fonctions permet de représenter une courbe fonctionnelle des observations de l'individu i. Le paramètre d'effets fixes de la fonction spline est β et le paramètre γ_i est associé aux effets aléatoires.
James & Hastie considèrent que l'approche de la matrice de covariance proposée par Rice & Wu [33] conduit à une matrice de grande dimension difficile à estimer. Pour éviter ce degré de complexité tout en conservant la performance et la flexibilité du modèle, ils proposent de procéder à l'extraction des composantes principales de la matrice de covariance du modèle. Cette réduction de la matrice de covariance s'obtient par régression sur des fonctions de courbes orthogonales entre elles. Ces fonctions de courbes orthogonales sont obtenues sous contraintes. Cette approche permet l'analyse des observations disponibles tout en évitant le problème relatif aux instants distincts de mesure des différents individus. Cette approche est basée sur la matrice de variance réduite contenant l'information retenue par les composantes principales de la matrice de variance.

1.5.3 Approche des données longitudinales environnementales proposée

Pour modéliser la variabilité inter-individu, les effets aléatoires du modèle linéaire mixte de Laird & Ware [26] ont été introduits pour l'analyse des données environnementales comme l'attestent les deux dernières approches de la section précédente. Le but de cette thèse est de proposer une approche permettant une modélisation des données environnementales avec une extraction d'une composante commune représentant la dynamique du phénomène observé pour l'ensemble des sites et aussi une reconstruction des courbes d'évolution de chaque site.
Les modèles mixtes permettent cette approche intégrée des observations provenant de différents individus et où chaque individu possède plusieurs observations avec une prise en compte des différentes sources de variation. Pour définir notre approche générale, le chapitre 2 est axé sur le modèle linéaire mixte étendu au modèle linéaire mixte stochastique. Les différentes méthodes d'estimation ainsi que certaines propriétés du modèle sont expo-

sées. Les corrélations intra-individus sont approchées à travers une modélisation sérielle prise en compte par la composante stochastique introduite. Les modèles classiques des séries chronologiques peuvent être utilisés pour représenter cette corrélation sérielle ou encore d'autres processus plus appropriés au phénomène étudié.
Ensuite, les techniques sémiparamétriques sont introduites. Le modèle permettant une estimation de la tendance temporelle par des méthodes nonparamétriques palliant ainsi à l'insuffisance de la régression sur les variables explicatives disponibles est le modèle sémiparamétrique qui s'écrit :

$$Y = X\beta + f(t) + Zb + \varepsilon. \tag{1.1}$$

Selon cette approche, la tendance temporelle des données pourra être estimée par des méthodes nonparamétriques (méthode des noyaux, fonctions splines cubiques, ondelettes,...).
Diggle et *al* [46, 12], dans leur approche, notent que le paramètre β peut être estimé par la méthode des moindres carrés généralisés (GLS) combinée de façon itérative à une estimation de la tendance temporelle par la méthode des noyaux. Le modèle de l'équation (1.1) est typiquement la forme du modèle mixte semiparamétrique de Diggle et *al* [46, 12]. Cependant le modèle proposé, dans cette thèse pour l'analyse des données longitudinales exposées, est plus proche du modèle mixte semiparamétrique stochastique de Zhang et *al* [47] où l'approximation de la fonction nonparamétrique est obtenue à l'aide des fonctions splines de lissage. Les fonctions splines d'approximation sont représentées comme un modèle linéaire mixte, s'intégrant naturellement au modèle. L'estimation de la fonction spline de lissage lorsque le nombre d'observations est élevé, est délicate. C'est pourquoi comme Jacqmin-Gadda et *al.* [21], nous proposons d'utiliser à la place de la fonction spline de lissage, une fonction spline de régression avec un nombre réduit de noeuds. Les outils utilisés pour définir cette approche semiparamétrique stochastique et les méthodes d'estimation proposées pour le modèle mixte semiparamétrique stochastique constituent l'objet du chapitre 3. Des résultats de simulation sont présentés afin d'illustrer les performances des méthodes d'estimation.
Dans le chapitre 4, nous introduisons la méthode d'influence locale afin de développer une analyse de sensibilité dans le modèle. En effet, l'influence locale présentée est une méthode permettant de détecter les effets sur le modèle, des perturbations de certaines de ces composantes. Nous adaptons la méthode d'influence locale de Cook [8] au modèle mixte semiparamétrique stochastique. Nous montrons certaines propriétés asymptotiques locales de la méthode et décomposons l'influence locale suivant les différents paramètres du modèle. En particulier, nous mettons en oeuvre une analyse de sensibilité locale sur le paramètre de lissage intégré comme paramètre de variance.
Dans le dernier chapitre, nous appliquons le modèle mixte semiparamétrique stochastique à deux jeux de données réelles. Nous exhibons ainsi l'intérêt de la méthodologie développée pour l'analyse de ces données dans le domaine de la santé publique.
La contribution de cette thèse se situe à différents niveaux :

– L'approximation de la fonction nonparamétrique se fait par une fonction spline formellement introduite comme un (sous) modèle à effets aléatoires sans modifier la composante des effets fixes du modèle global. Le paramètre de lissage est directement intégrée comme

une composante de variance des effets aléatoires de ce (sous) modèle.
- Ensuite, nous mettons l'accent sur le fait que la méthode de validation croisée généralisée (GCV), qui simplifie l'application de la validation croisée (CV) ne s'applique pas parfaitement lorsque l'intérêt de la modélisation porte sur la recherche d'une composante commune.
- Les propriétés asymptotiques démontrées pour la matrice d'influence permettent sa décomposition entre les paramètres d'effets fixes et les paramètres de variance du modèle. L'analyse de sensibilité basée sur la méthode d'influence locale ouvre des possibilités d'évaluation du comportement local de certains paramètres de variance dans le modèle mixte modifié.
- Le modèle mixte semiparamétrique stochastique adapté aux données environnementales permet l'évaluation de l'exposition environnementale dans certains domaines où les données sont plus ou moins parcellaires.

1.6 Notations

Dans cette section, nous présentons les différentes principales notations utilisées dans ce mémoire. Dans un souci de clarté, elles sont souvent rappelées au risque d'être redondantes.

Y : la variable aléatoire du phénomène étudié sur tous les sujets, Y_i la suite des mesures répétées au niveau du i^{eme} individu, y_{ij} la réalisation à l'instant noté t_{ij}. Y_i : $(Y_{i1}, Y_{i2}, ..., Y_{in_i})$ le vecteur aléatoire de taille n_i des mesures répétées au niveau de l'unité i avec $E(Y_i) = \mu_i$ et sa variance $Var(Y_i) = V_i$

Y : $(Y_1, ..., Y_m)$ l'ensemble des vecteurs aléatoires des mesures répétées des différentes unités. Var(Y)=V, $E(Y) = \mu$.

y_i : $(y_{i1}, y_{i2}, ..., y_{in_i})$ le vecteur des réalisations de taille n_i de Y_i

y : $(y_1, ..., y_m)$ l'ensemble des vecteurs des réalisations de Y.

β : un vecteur de p éléments du paramètre inconnu d'effets fixes.

b : un vecteur de q éléments d'effets aléatoires.

θ : le vecteur des paramètres contrôlant la structure de covariance des données, il contient tous les paramètres liés aux différentes composantes de la variance.

X : la matrice des p variables explicatives, mesures des variables explicatives reliées par le paramètre d'effets fixes β à la variable d'intérêt Y, $X_{ij} = (X_{ij1}, ..., X_{ijp})$ est un vecteur de p éléments donnant les valeurs prises par les différentes variables explicatives pour l'instant t_{ij} de la mesure y_{ij}.

Z : la matrice des q facteurs d'effets aléatoires, reliés par le paramètre d'effets aléatoires b à la variable d'intérêt, Z_{ij} est un vecteur de q éléments, contenant sauf mention contraire des 0 et des 1.

ε : un vecteur de N résidus de même distribution que Y sachant les effets aléatoires connus.

n_i : nombre d'observations par individu, l'individu peut être un site, une station de mesures dans le domaine environnemental ou une unité statistique en général.

N : le nombre d'observations disponibles sur l'ensemble des m individus soit $N = \sum_{i=1}^{m} n_i$.

G, R ,Γ : des matrices de variance-covariance respectivement de b,ε, et U.

I_{n_i},et I_N : les matrices identité d'ordre n_i et N.

1_{n_i} : vecteur de taille n_i contenant des 1.

$L(\beta, \theta; y)$: la logvraisemblance.

ϕ : l'ensemble des paramètres (β, θ).

$f(t)$: paramètre fonctionnel du modèle.

$U(t)$: processus stochastique du modèle.

B : matrice de B-splines B_j dont les coefficients sont a_j.

f : Vecteur dont les élementsf_j sont les valeurs l'approximation de la nonparamétrique.

a : les coefficints de regréssion reliant NB à f.

Ω : matrice de variance des paramètre a de la fonction $f = NBa$

Chapitre 2

Approche générale et modèles linéaires mixtes

2.1 Introduction

L'approche est motivée par la détermination des sources de variation potentielles des données. Le principe directeur de l'analyse, est basé sur l'hypothèse que ces sources de variation des données sont déterminées de façon opportune selon une variabilité naturelle entre des individus plus ou moins *similaires* (Searle, [37]). Cette variabilité devient mesurable grâce aux différents niveaux des facteurs du modèle. Les modèles à effets mixtes ou modèles mixtes sont des modèles associant des facteurs à effets fixes et des facteurs à effets aléatoires.
De manière conceptuelle, un facteur est à effets fixes lorsqu'il a un nombre fini de niveaux qui sont *entièrement* représentés. Les facteurs à effets aléatoires sont des facteurs dont le nombre des niveaux peuvent être infini et les données disponibles ne représentent alors qu'un échantillonnage des niveaux de chaque facteur.

Laird & Ware [26] sont parmi les premiers à présenter le modèle linéaire mixte sous une approche *globale* de modèle de régression et d'analyse des composantes de la variance avec les méthodes d'estimation (algorithme EM) à cette approche. Aujourd'hui le modèle linéaire mixte est d'un usage courant et trouve des applications dans de nombreuses sciences expérimentales.
Dans ce chapitre, nous allons tout d'abord rappeler le modèle linéaire mixte classique ainsi que les méthodes d'inférences propre à ce modèle. Ensuite l'extension (du modèle linéaire mixte classique) qu'est le modèle linéaire mixte stochastique sera présenté suivant une approche spécifique aux données longitudinales rencontrées dans le chapitre précédent.

2.2 Modèle linéaire mixte standard

Le modèle linéaire mixte peut être vu comme une extension du modèle de régression classique : comme lui, l'hypothèse admise est que les données Y suivent une loi normale du type :

$$Y \sim \mathcal{N}(X\beta, V_\theta). \tag{2.1}$$

Ce qui peut s'écrire de manière équivalente sous la forme :

$$Y = X\beta + \epsilon, \tag{2.2}$$

où le vecteur d'aléas ϵ est une variable gaussienne de variance $Var(\epsilon) = V_\theta$.
Dans les équations (2.1) et (2.2) :

- X est une matrice de dimension $(N \times p)$ de rang plein, il s'agit des variables explicatives (variables exogènes).

- Le paramètre $\beta \in \mathbb{R}^P$ est aussi un vecteur de p éléments désigné comme le paramètre de la moyenne du modèle, formalisant la relation linéaire entre l'espérance de la variable expliquée et les variables explicatives : X

$$E(Y) = X\beta.$$

- V_θ est la matrice de dimension $(N \times N)$ de covariance de ϵ donc de Y, c'est une fonction de θ désigné comme le paramètre de la variance.

Dans l'équation (2.1), la connaissance de la distribution de la variable Y est alors complètement déterminée par celle des paramètres de la variance θ et des paramètres de la moyenne β.
L'hypothèse de normalité faite sur la distribution de la variable Y est évidemment réductrice. Cependant son utilisation conduit à des expressions plus accessibles, simples à évaluer et les pratiques font d'elle la loi la plus intuitive. Le modèle linéaire mixte est aussi désigné sous le nom de modèle mixte gaussien.
Un cas particulier de la formulation de la structure de covariance V_θ du modèle linéaire mixte, s'effectue par l'adjonction d'effets aléatoires $Z_i b_i$ $i = 1, \ldots, m$, constituant m sources de variation supplémentaires aux résidus ε :

$$\epsilon = Z_1 b_1 + \ldots + Z_m b_m + \varepsilon.$$

Ce qui conduit à l'équation générale :

$$Y = X\beta + Zb + \varepsilon, \tag{2.3}$$

où $Z = [Z_1, \ldots, Z_m]$, chaque Z_i, $i = 1, \ldots, m$ étant une matrice de plans d'expériences de dimension $(N \times q_i)$. En posant $q = \sum_{i=1}^{m} q_i$ alors Z est de dimension $(N \times q)$.
Le vecteur b représente la concaténation des q paramètres aléatoires (gaussiens), reliant la matrice Z des facteurs et variables à effets aléatoires au vecteur Y des données.

Par définition, le vecteur ε est la composante résiduelle évoquée dans l'approche du modèle linéaire mixte qui sera désignée, par la suite, tout simplement comme le vecteur des résidus du modèle linéaire mixte. Les hypothèses de construction du modèle impliquent que $E(\varepsilon) = 0$ et $E(b_i) = 0$ et, au niveau des variances : $\text{Var}(b_i) = \sigma_i^2 G_i, \qquad i = 1, \ldots, m$, $\text{Var}(\varepsilon) = \sigma_0^2 R$ avec les matrices G_i et R supposées connues. En ajoutant l'hypothèse supplémentaire que les composantes b_i sont mutuellement indépendantes entre elles et indépendant du vecteur ε des résidus :

$$\forall \ \ i \ \ \text{et} \ \ \ j \ \ \text{, avec} \ \ i \neq j \ \ \text{Cov}(b_i, b_j) = 0 \ \ et \ \ \text{aussi} \ \ \text{Cov}(b, \varepsilon) = 0.$$

La variance V_θ devient :

$$V_\theta = \text{Var}(Y) = \text{Var}(X\beta + Zb + \varepsilon) = \text{Var}(Zb) + \text{Var}(\varepsilon) = \sum_{i=1}^{m} \sigma_i^2 Z_i G_i Z_i^T + \sigma_0^2 R.$$

On notera que V_θ ici est une fonction linéaire du paramètre de variance $\theta = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)$. Pour terminer cette présentation sur le modèle linéaire mixte, il faut noter que β est désigné comme le paramètre d'effets fixes et b comme le vecteur d'effets aléatoires.

Un cas particulier de modèle d'analyse des composantes de la variance, est le cas où chaque q_i vaut 1 et Z_i est un vecteur composé de 0 et de 1, variable indicatrice d'un bloc i de données. Chaque b_i représente alors la constante «random intercept» du bloc i correspondant. Dans ce cas, l'expression de la j^{eme} observation du bloc i se note :

$$Y_{ij} = x_j^T \beta + b_i + \varepsilon_{ij},$$

$j = 1, \ldots, n_i$.
Pour les données du bloc i, la variance de Y_i est donnée par :

$$\text{Var}(Y_i) = \sigma_0^2 R_i + \sigma_i^2 1_{n_i} 1_{n_i}^T.$$

2.3 Formulation d'un modèle linéaire mixte stochastique pour données longitudinales

Très souvent les données provenant d'un processus physique ou biologique telles que celles décrites au chapitre 1, possèdent une structure de covariance complexe. Cette dernière est difficile à représenter par un modèle simple. Lors de l'analyse des données longitudinales, il faut tenir compte de la dépendance au temps. Cette dépendance fait dorénavant apparaître les données Y comme des courbes d'évolution $Y(t)$. Les variables exogènes, peuvent aussi éventuellement dépendre du temps.
Mais dans le cadre d'une modélisation par modèle linéaire mixte, la dépendance par rapport au temps peut se présenter via des composantes de variance supplémentaires que nous détaillons ci-après.

Diggle et *al* donnent un exposé exhaustif sur le sujet dans [12]. Il paraît raisonnable d'étendre le modèle linéaire mixte standard de la façon suivante. Pour modéliser les variations locales propres à chaque station de collecte i des données environnementales, un terme $Z_i(t)b_i$ représentera cet effet du site i ; un processus stochastique $U_i(t)$ représentera, quant à lui, les corrélations sérielles de ce site. Ainsi la décomposition de la partie aléatoire ϵ de (2.2), pour le site i, devient :

$$\epsilon_i(t) = Z_i(t)b_i + U_i(t) + \varepsilon_i(t).$$

L'ensemble des composantes et des processus aléatoires du modèle (à savoir b_i, $U_i(t)$, $\varepsilon_i(t)$ $i = 1, \ldots, m$,) sont mutuellement indépendantes.
$U_i(t)$ est un processus continu (même si nous nous intéresserons surtout à sa restriction, aux instants de mesure pour sa prédiction). Un choix de structure de corrélation, le plus fréquent pour les données longitudinales, se porte sur la fonction de corrélation exponentielle : $U_i(t)$ peut être alors un processus stationnaire de type Ornstein-Uhlenbeck (O.U.) homogène de structure de covariance :

$$\Gamma_i(\gamma_i, \alpha_i, t, s) = \mathrm{cov}(U_i(t), U_i(s)) = \gamma_i^2 \exp(-\alpha_i|t - s|).$$

Ce choix de processus gaussiens stationnaires n'est pas unique. Cependant pour ce travail, nous nous limitons au processus stochastique exponentiel.
Très souvent les données longitudinales exhibent des structures non stationnaires, dans ce cas des spécifications adaptées peuvent être trouvées. Un choix, peut se porter sur le processus de type Ornstein-Uhlenbeck (O.U.) non homogène comme celui proposé par Zhang & *al* [47].
Sur l'ensemble des instants de mesure t_{ij} du site $i, j = 1, \ldots, n_i$, le modèle linéaire mixte stochastique s'écrit :

$$Y_i(t_{ij}) = X(t_{ij})\beta + Z_i(t_{ij})b_i + U_i(t_{ij}) + \varepsilon_{ij}, \tag{2.4}$$

avec i variant de $1, \ldots, m$ et j variant de $1, \ldots, n_i$.
Pour l'ensemble des données de la station i la variance devient alors :

$$V_{\theta i} = \sigma_0^2 R_i + \sigma_i^2 Z_i G_i Z_i^T + \Gamma_i(\gamma_i^2, \alpha_i),$$

où $\Gamma_i(\gamma_i^2, \alpha_i)$ est la matrice de dimension $(n_i \times n_i)$ des termes
$\{\Gamma_i(\gamma_i^2, \alpha_i, t_{ij_1}, t_{ij_2})\}_{j_1=1,\ldots,n_i,\ j_2=1,\ldots,n_i}$.

Pour une écriture globale du modèle pour l'ensemble des sites, nous poserons :

$$Y = X\beta + Zb + U + \varepsilon, \tag{2.5}$$

avec

$$\begin{aligned} Y &= \left(Y_1^T, \ldots, Y_m^T\right)^T, \\ Z &= diag\left(Z_1, \ldots, Z_m\right), \\ U &= \left(U_1^T, \ldots, U_m^T\right)^T. \end{aligned}$$

La variance V_θ s'écrit alors :

$$V_\theta = \sigma_0^2 R + ZG(\sigma_*^2)Z^T + \Gamma(\gamma^2, \alpha), \tag{2.6}$$

avec :

$$\begin{aligned}
\theta &= \left\{\sigma_0^2, \sigma_*^2, \gamma^2, \alpha\right\}, \\
\sigma_*^2 &= \left(\sigma_1^2, \ldots, \sigma_m^2\right), \\
\Gamma(\gamma^2, \alpha) &= diag(\Gamma_1(\gamma_1^2, \alpha_1), \ldots, \Gamma_m(\gamma_m^2, \alpha_m)), \\
\alpha &= (\alpha_1, \ldots, \alpha_m), \\
\gamma^2 &= \left(\gamma_1^2, \ldots, \gamma_m^2\right), \\
G(\sigma_*^2) &= diag\left(\sigma_1^2 G_1, \ldots, \sigma_m^2 G_m\right), \\
\phi &= (\beta, \theta).
\end{aligned}$$

2.4 Méthodes d'estimation

Les composantes de la variance, les paramètres des effets fixes et les réalisations des effets aléatoires et du processus stochastique sont les différents paramètres du modèle linéaire mixte stochastique et sa mise en œuvre passe par l'estimation de ces différents paramètres. Il arrive souvent que l'intérêt de la modélisation porte particulièrement sur l'évaluation des réalisations des effets aléatoires $Z_i b_i$ et $U_i(t)$ du modèle, c'est le cas des modèles désignés sous le nom de «multi-level random effect models».
La méthode du maximum de vraisemblance et la méthode du maximum de vraisemblance restreinte sont les méthodes d'estimation du modèle mixte. En tant que modèle paramétrique, la connaissance du modèle est complètement déterminée par celle du paramètre ϕ : le paramètre d'intérêt de la logvraisembance, (dans sa forme la plus générale on a Φ est l'ensemble des paramètres ϕ. Celle-ci s'effectue à travers la maximisation de la fonction objectif qu'est la vraisemblance $L(\phi|y)$.
Il s'agit de trouver $\hat{\phi}$ telle que :

$$\hat{\phi} = \arg\max_{\phi \in \Phi} L(\phi|y),$$

L'obtention de l'estimateur du maximum de logvraisemblance (ML) passe par la résolution des *équations normales* d'Euler-Lagrange associées :

$$grad_\phi(L(\phi|y))| = \left.\frac{\partial L(\phi|y)}{\partial \phi}\right|_{=} 0.$$

Ces équations sont aussi connues sous le nom d'*équations du maximum de logvraisemblance*. L'estimation ML des paramètres de variance induit un biais sur ces derniers. Cela est dû à l'absence de la prise en compte par la méthode ML, de la perte des degrés de liberté lors de l'estimation des effets fixes. La méthode du maximum de vraisemblance restreinte, dérivée

du maximum de vraisemblance, donne des estimateurs permettant de remédier à cela. Cette méthode est connue aussi sous de le nom de maximum de vraisemblance marginale en ce sens que la vraisemblance restreinte est obtenue après intégration par rapport au paramètre d'effets fixes β.
Pour finir sur les méthodes d'estimation du modèle linéaire mixte, et du fait que le modèle linéaire mixte est à l'interface de la statistique classique *fréquentiste* et de la statistique bayésienne, nous esquissons en fin de chapitre, l'approche *bayésienne* du modèle mixte stochastique après avoir présenté les *équations du modèle linéaire mixte* adaptées au contexte du modèle mixte stochastique.
L'ajustement du modèle linéaire mixte stochastique se résume succinctement aux deux points du schéma ci-dessous :

$$\beta,\ \mathrm{G},\ \mathrm{R},\ \Gamma \quad \Longleftarrow \boxed{(RE)ML}\ ou\ \boxed{BAYES}$$

$$\mathrm{b},\ \mathrm{U} \quad \Longleftarrow \boxed{BLUP}$$

où $BLUP$ = Best Linear Unbiased Predictor.

2.4.1 La méthode du maximum de vraisemblance

Les hypothèses de la méthode du maximum de vraisemblance notée ML (Maximum Likelihood) sont relativement fortes, mais il est connu qu'elles conduisent à des estimateurs avec des propriétés asymptotiques optimales (*Rao*, 1973) [32]. Ce qui constitue la justification essentielle de l'utilisation systématique de la méthode.

Estimation ML du paramètre β d'effets fixes à variance (supposée) connue

Sous les hypothèses de normalité et de la connaissance de la matrice de variance V, la logvraisemblance du modèle à effets mixtes s'écrit sous la forme :

$$L(\beta; y) = -\frac{1}{2}\{\log|V| + (y - X\beta)^T V^{-1}(y - X\beta) + N\log(2\pi)\}, \tag{2.7}$$

avec $|V| = det(V)$.
Cette logvraisemblance vérifie toutes les conditions d'application de la méthode du maximum de logvraisemblance. Selon la méthode ML, il s'agit de maximiser cette fonction objectif par rapport aux paramètres β.
Les équations ML deviennent alors :

$$L_\beta = \frac{\partial L(\beta; y)}{\partial \beta} = X^T V^{-1}(y - X\beta).$$

Finalement, la maximisation de la logvraisemblance revient à trouver la solution qui annule les équations ML obtenues ci-dessus tout simplement, d'où on obtient :

$$(X^T V^{-1} X)\hat{\beta}_{ML} = X^T V^{-1} y. \tag{2.8}$$

La solution de cette dernière est la même que celle obtenue si l'estimation avait été faite par les moindres carrés généralisés (GLS). L'estimateur par maximum de vraisemblance (MLE) à variance connue est :

$$\hat{\beta}_{ML} = (X^T V^{-1} X)^- X^T V^{-1} y.$$

Cet estimateur $\hat{\beta}_{ML}$ ainsi obtenu est un estimateur BLUE (Best Linear Unbiased Estimator). Dans l'expression de $\hat{\beta}_{ML}$, la notation $(X^T V^{-1} X)^-$ désigne une inverse généralisée de $(X^T V^{-1} X)$.

Estimation ML à paramètres de variance inconnus

Les équations ML pour le paramètre β ont été résolues en supposant la matrice de variance-covariance connue. Ce n'est pas le cas en pratique et il faut aussi l'estimer. Cette estimation passe par l'évaluation des équations ML pour les paramètres de la variance θ :

$$L_{\theta_l} = \frac{\partial L(\beta, \theta; y)}{\partial \theta_l} = -\frac{1}{2} tr(V_\theta{}^{-1} \dot{V}_{\theta_l}) + \frac{1}{2}(y - X\beta)^T V_\theta{}^{-1} \dot{V}_{\theta_l} V_\theta{}^{-1} (y - X\beta) \tag{2.9}$$

$l = 1, \ldots, s$, décrivant les différents éléments de θ et où $\dot{V}_{\theta_l} = \dfrac{\partial V(\theta)}{\partial \theta_l}$.
Les équations (2.9) sont aussi des fonctions du paramètre d'effets fixes β. La maximisation de la logvraisemblance, dans ce cas aussi, revient à trouver les solutions qui annulent (2.9). Les solutions ML des équations (2.8) et (2.9) vérifient bien évidemment le système ci-dessous :

$$\begin{aligned} (X^T V_\theta{}^{-1} X)\beta &= X^T V_\theta{}^{-1} y & (2.10)\\ \mathrm{tr}(V_\theta{}^{-1} \dot{V}_{\theta_l}) &= (y - X\beta)^T V_\theta{}^{-1} \dot{V}_{\theta_l} V_\theta{}^{-1} (y - X\beta) & (2.11) \end{aligned}$$

où $l = 1, \ldots, s$.
Il s'agit de trouver $\hat{\beta}_{ML}$ et $\hat{\theta}_{ML}$, les solutions ML vérifiant les équations (2.10) et (2.11).
En posant :

$$P_\theta = V_\theta^{-1} - V_\theta^{-1} X (X^T V_\theta^{-1} X)^- X^T V_\theta^{-1}, \tag{2.12}$$

nous obtenons l'égalité simple suivante $V_\theta^{-1}(y - X\hat{\beta}_{ML}) = P_\theta y$. (2.10) et (2.11) deviennent alors :

$$\begin{aligned} (X^T V_\theta^{-1} X)\dot{\beta} &= X^T V_\theta^{-1} y, & (2.13)\\ \mathrm{tr}(V_\theta^{-1} \dot{V}_{\theta_l}) &= y^T P_\theta \dot{V}_{\theta_l} P_\theta y, & (2.14) \end{aligned}$$

$l = 1, \ldots, s$.
Enfin, il faut noter que les équations (2.13) et (2.14) sont non linéaires en θ. La résolution de l'équation (2.14), à ce titre, ne donne pas d'expressions explicites de la solution.

Il existe des méthodes numériques itératives adaptées pour l'approche de ces solutions $[Newton - Raphson, Fisher - Scoring, EM]$.

Matrice de dispersion asymptotique des estimateurs ML

La méthode ML trouve tout son intérêt dans les propriétés des paramètres estimés. Il est établi que les estimateurs ML convergent presque sûrement, ($i.e$ lorsque $N \to \infty$) vers les vraies valeurs des paramètres (il faut noter que N représente le nombre des observations.) et vérifient la propriété de normalité asymptotique :

$$\forall \phi \in \Phi, \sqrt{N}\left(\hat{\phi}_N - \phi\right) \xrightarrow{L} \mathcal{N}\left(0, \{I(\phi)\}^{-1}\right), \; quand \; N \to \infty,$$

avec :

$$\begin{aligned} I(\phi) &= E\left(\frac{\partial L}{\partial \phi}\frac{\partial L}{\partial \phi^T}\right), \\ &= E\left\{\frac{\partial L}{\partial \phi_i}\frac{\partial L}{\partial \phi_j}\right\}_{i,j=1,\ldots,s+p}, \end{aligned}$$

la matrice d'information.
Il est facile de montrer que sous certaines conditions de régularité, cette matrice peut se mettre sous une forme plus simple à évaluer :

$$I(\phi) = -E\left(\frac{\partial^2 L}{\partial \phi \partial \phi^T}\right), \tag{2.15}$$

$$= -E\left\{\frac{\partial^2 L}{\partial \phi_i \partial \phi_j^T}\right\}_{i,j}. \tag{2.16}$$

Utilisant une formulation en blocs, la matrice de variance asymptotique des paramètres β et θ est alors :

$$I(\beta,\, \theta) = -E\begin{bmatrix} L_{\beta\beta} & L_{\beta\theta} \\ L_{\beta\theta} & L_{\theta\theta} \end{bmatrix},$$

où $L_{\beta\beta} = \dfrac{\partial^2 L}{\partial \beta \partial \beta^T} = -X^T V_\theta^{-1} X$, de même $L_{\beta\theta} = \dfrac{\partial^2 L}{\partial \beta \partial \theta^T} = -X^T V_\theta^{-1} \dot{V}_{\theta_l} V_\theta^{-1}(y - X\beta)$ et enfin :

$$L_{\theta\theta} = \frac{\partial^2 L}{\partial \theta \partial \theta^T} = \frac{1}{2}\mathrm{tr}(V_\theta^{-1}\dot{V}_\theta V^{-1}\dot{V}_\theta - V_\theta^{-1}\ddot{V}_\theta)$$
$$+\frac{1}{2}\left\{(y - X\beta)^T\left(V_\theta^{-1}\ddot{V}_{\theta_l\theta_k}V_\theta^{-1} - V_\theta^{-1}\dot{V}_{\theta_l}V_\theta^{-1}\dot{V}_{\theta_k}V_\theta^{-1}\right)(y - X\beta)\right\}_{lk}.$$

Le calcul direct des espérances de ces différentes expressions donne :

$$\begin{aligned} E(L_{\beta\beta}) &= -X^T V_\theta^{-1} X, \\ E(L_{\beta\theta}) &= 0, \\ E(L_{\theta\theta}) &= -\frac{1}{2}\mathrm{tr}(V_\theta^{-1}\dot{V_{\theta_l}} V_\theta^{-1}\dot{V_{\theta_k}})_{lk}. \end{aligned}$$

La matrice de variance asymptotique estimée se met alors sous la forme simple :

$$I(\hat{\beta}_{ML}, \hat{\theta}_{ML})^{-1} = \begin{bmatrix} (X^T V_{\hat{\theta}}^{-1} X)^{-1} & 0 \\ 0 & \{\frac{1}{2}\mathrm{tr}(V_{\hat{\theta}}^{-1}\dot{V}_{\hat{\theta}_l} V_{\hat{\theta}}^{-1}\dot{V}_{\hat{\theta}_k})\}_{lk}^{-1} \end{bmatrix},$$

l, $k = 1, \ldots, s$, décrivant les différents éléments de θ.
Pour finir l'expression $\dot{V_{\theta_l}}$, pour les différents éléments de θ est :

$$\dot{V_{\sigma_0^2}} = R, \tag{2.17}$$

$$\dot{V_{\sigma_i^2}} = Z_i G_i Z_i^T \qquad i = 1, \ldots, m, \tag{2.18}$$

$$\dot{V_{\gamma_i^2}} = \frac{\partial \Gamma}{\partial \gamma_i^2} \qquad i = 1, \ldots, m, \tag{2.19}$$

$$\dot{V_{\alpha_i}} = \frac{\partial \Gamma}{\partial \alpha_i} \qquad i = 1, \ldots, m, \tag{2.20}$$

$$\tag{2.21}$$

2.4.2 La méthode du maximum de vraisemblance restreinte

L'estimation des paramètres de variance par la méthode du maximum de vraisemblance (ML) conduit à des estimateurs biaisés. L'estimation (ML) des paramètres de variance ne tient pas compte de la perte des degrés de liberté induite par l'estimation des effets fixes. Le maximum de vraisemblance restreinte (REML) est une méthode dérivée du maximum de vraisemblance, permettant de prendre en compte cette perte de degré de liberté causée par l'estimation du paramètre β.

L'approche et les équations REML

L'exposé de la méthode nécessite l'introduction du rang de la matrice X noté : r_x.
La méthode est basée sur une transformation linéaire des données par une matrice notée A, cette matrice A est de dimension $(N \times (N - r_x))$ avec les $(N - r_x)$ vecteurs colonnes formés par le nombre maximal des $(N - r_x)$ vecteurs linéairement indépendants et orthogonaux aux vecteurs colonnes de la matrice X.
La transformation ainsi construite permet de vérifier aisément les hypothèses requises pour la transformée $A^T y$, en l'occurrence le fait qu'elle ne dépend pas de β : $E(A^T Y) = 0 \Rightarrow A^T X\beta = 0\,; \forall\beta$ car $A^T X = 0$. Une forme immédiate de A est la matrice associée à la base des vecteurs propres, (de valeurs propres non nulles) du projecteur $I - X(X^T V_\theta^{-1} X)^- X^T V_\theta^{-1}$.

La logvraisemblance associée au vecteur A^TY est ainsi dénommée logvraisemblance restreinte.
Comme $y \sim N(X\beta, V_\theta)$ alors $A^Ty \sim N(0, A^TV_\theta A)$. La logvraisemblance restreinte s'écrit :

$$L_r(\theta; y) = -\frac{1}{2}\{\log|A^TV_\theta A| + y'A(A^TV_\theta A)^{-1}A^Ty) + (N - r_x)\log(2\pi)\}. \tag{2.22}$$

On sait que la matrice P donnée par (2.12) peut se mettre sous la forme [37] :

$$P_\theta = A(A^TV_\theta A)^{-1}A^T.$$

Pour la dérivation du premier terme de (2.22), on aura :

$$\begin{aligned}
\frac{\partial \log|A^TV_\theta A|}{\partial\theta_l} &= |A^TV_\theta A|^{-1}\frac{\partial|A^TV_\theta A|}{\partial\theta_l}, \\
&= |A^TV_\theta A|^{-1}|A^TV_\theta A|\mathrm{tr}\left\{(A^TV_\theta A)^{-1}\frac{\partial(A^TV_\theta A)}{\partial\theta_l}\right\}, \\
&= \mathrm{tr}\left\{(A^TV_\theta A)^{-1}A^T\frac{\partial V_\theta}{\partial\theta_l}A\right\}, \\
&= \mathrm{tr}\left\{A(A^TV_\theta A)^{-1}A^T\frac{\partial V_\theta}{\partial\theta_l}\right\}, \\
&= \mathrm{tr}(P_\theta\frac{\partial V_\theta}{\partial\theta_l}) = \mathrm{tr}(P_\theta\dot{V}_{\theta_l}),
\end{aligned}$$

le second terme de (2.22) se dérive essentiellement à partir de l'expression de P :

$$\begin{aligned}
\frac{\partial P_\theta}{\partial\theta_l} = & \frac{\partial A(A^TV_\theta A)^{-1}A'}{\partial\theta_l}, \\
= & -A(A^TV_\theta A)^{-1}\frac{\partial(A^TV_\theta A)}{\partial\theta_l}(A^TV_\theta A)^{-1}A^T, \\
= & -A(A^TV_\theta A)^{-1}A^T\frac{\partial V_\theta}{\partial\theta_l}A(A^TV_\theta A)^{-1}A^T, \\
= & -P_\theta\frac{\partial V_\theta}{\partial\theta_l}P_\theta = -P_\theta\dot{V}_{\theta_l}P_\theta.
\end{aligned}$$

D'où les dérivées de la logvraisemblance restreinte (2.22) par rapport au paramètre θ deviennent :

$$\frac{\partial L_r}{\partial\theta_l} = -\tfrac{1}{2}\mathrm{tr}(P_\theta\dot{V}_{\theta_l}) + \tfrac{1}{2}y^TP_\theta\dot{V}_{\theta_l}P_\theta y, \tag{2.23}$$

$l = 1, \ldots, s$.
Les solutions des équations REML du modèle mixte stochastique vérifient alors :

$$\mathrm{tr}(P_\theta\dot{V}_{\theta_l}) = y^TP_\theta\dot{V}_{\theta_l}P_\theta y. \tag{2.24}$$

Il faut remarquer que les équations (2.24) peuvent être obtenues au niveau de (2.9), en appliquant la transformation linéaire A aux composantes du modèle y, Z, X, V de la manière suivante :

$$\begin{aligned} y &\rightarrow A^T y, \\ Z &\rightarrow A^T Z, \\ X &\rightarrow A^T X, \\ V &\rightarrow A^T V A. \end{aligned}$$

Tout comme pour la résolution des équations de la logvraisemblance classique, la résolution des équations REML requiert, en pratique, de recourir à des algorithmes itératifs pour approcher les solutions de ces équations non linéaires par rapport au paramètre θ.
Enfin comme, déjà noté la logvraisemblance restreinte est aussi appelée aussi *logvraisemblance marginale*. Il s'agit de la logvraisemblance obtenue après intégration par rapport au paramètre β qi se met sous la forme :

$$L_r(y;\theta) = -\frac{1}{2}\{\log|V(\theta)| + \log|X^T V(\theta) X| + (y - X\hat{\beta})^T V(\theta)^{-1}(y - X\hat{\beta}) + (N - r_x)\log(2\pi)\}. \tag{2.25}$$

Il faut noter que la méthode du REML ne concerne pas directement l'estimation des paramètres des effets fixes β. Cependant, si les paramètres de la variance intervenant dans l'estimation de (2.8), sont des estimations REML $\hat{\theta}_R$, alors on parle d'estimateurs REML pour les effets fixes β aussi :

$$\hat{\beta}_{REML} = (X^T V_{\hat{\theta}_R}^{-1} X)^- X^T V_{\hat{\theta}_R}^{-1} y. \tag{2.26}$$

Matrice de variance asymptotique des estimateurs REML

En appliquant la formule (2.15) de calcul de la matrice de variance asymptotique, celle des estimateurs des composantes de la variance obtenus par la méthode du REML est donnée par :

$$(-E(\frac{\partial^2 L_r}{\partial\theta_l \partial\theta_k^T}))^{-1} = \{\frac{1}{2}\mathrm{tr}(P_\theta \dot{V}_{\theta_l} P_\theta \dot{V}_{\theta_k})\}_{lk=1,\ldots,s}^{-1}, \tag{2.27}$$

Les différents calculs se font selon le même le cheminement que dans le cas ML.

2.4.3 Prédiction des effets aléatoires du modèle mixte stochastique

Dans l'introduction sur l'estimation dans le modèle mixte stochastique, nous avons rappelé que l'intérêt du modèle peut porter sur les valeurs que peuvent prendre les effets aléatoires et le processus stochastique $U(t)$. La procédure permettant cette inférence est désignée sous le nom de prédiction. Le terme estimation concerne les paramètres d'effets fixes qui sont comme l'indique leur nom fixes (non aléatoires) mais inconnus. Tandis que le terme prédiction est réservé aux composantes aléatoires ayant une distribution gaussienne dans

le modèle mixte stochastique. Il faut dire que les prédicteurs sont en fait les espérances conditionnelles sachant les données Y. Le développement théorique sur les méthodes de prédictions des composantes aléatoires est essentiellement basé sur les travaux de Harville [19].
Les prédicteurs possèdent des propriétés analogues aux estimateurs. On s'intéresse aux prédicteurs linéaires sans biais et de variance minimale désignés comme BLUP's (Best Linear Unbiaised Predictors).

Prédiction des effets aléatoires b

Il s'agit de trouver l'espérance conditionnelle de b sachant Y, i.e. $E(b|Y)$. La notion d'espérance conditionnelle permet de donner un sens à celle de "résultat attendu de b connaissant Y". Il est bien établi que l'espérance conditionnelle d'une variable aléatoire gaussienne est de forme générale :

$$E(b|Y) = E(b) + \mathrm{Cov}(b, Y)\mathrm{Var}(Y)^{-1}\left(Y - E(Y)\right). \tag{2.28}$$

Soit ici :

$$E(b|Y) = E(b) + GZ^T V^{-1}(Y - X\beta). \tag{2.29}$$

Cela nous permet d'introduire le résultat suivant. Sous les hypothèses du modèle (2.3), le prédicteur de b ayant la variance minimale est l'espérance conditionnelle de b sachant $Y = y$ notée : $E(b|Y)$. La preuve de ce résultat peut se trouver dans [19, 37] ou d'autres ouvrages traitant du sujet. D'après ce résultat, la prédiction des effets aléatoires b est donnée par :

$$\hat{b} = \hat{G}Z^T\hat{V}^{-1}(Y - E(Y)) = \hat{G}Z^T\hat{V}^{-1}(Y - X\hat{\beta}). \tag{2.30}$$

(Le paramètre d'effets fixes est remplacé, par exemple, par son estimateur REML de β noté par $\hat{\beta}_{REML} = (X^T V^{-1}_{\hat{\theta}_R} X)^- X^T V^{-1}_{\hat{\theta}_R} Y$ et les paramètres de variance par leurs estimateurs REML.)

Prédiction du processus continu du temps $U(t)$

Le processus est continu, cependant l'intérêt des prédictions porte sur des instants discrets quelconques. Une fois, cette grille d'instants t_0 fixée, l'application de la proposition du paragraphe précédent donne les prédicteurs de U.

$$\hat{U}_i(t) = \hat{\Gamma}_i(t_0, t)\hat{V}_i^{-1}(Y_i - E(y)) = \hat{\Gamma}_i(t_0, t)\hat{V}_i^{-1}(Y_i - X\hat{\beta}), \tag{2.31}$$

C'est un prédicteur BLUP.

Reconstruction des profils individuels $Y_i(t)$

Globalement l'intérêt de la modélisation, dans ce contexte, porte aussi sur la reconstruction des profils individuels $Y_i(t) = X_i(t)\beta + Z_i(t)b_i + U_i(t) + \varepsilon_i$ au niveau de chaque station. Ces

profils individuels combinent des composantes aléatoires (effets aléatoires b et processus U) et des effets fixes β. On peut se demander si ces profils individuels vérifient les propriétés de variance minimale et de biais nul en tant que fonctions continues du temps.
Sous les hypothèses du modèle (2.5), les $\hat{Y}_i(t)$ seront des prédicteurs et auront des propriétés BLUP :

$$\begin{aligned}\hat{Y}_i(t) = & \quad X_i(t)\hat{\beta}_r + Z(t)\hat{b}_i + \hat{U}_i(t), \\ = & \quad X_i(t)\text{BLUE}(\beta) + Z(t)\text{BLUP}(b_i) + \text{BLUP}(U_i(t)).\end{aligned}$$

2.4.4 Les équations du modèle mixte stochastique d'Henderson

Après ce rappel, sur les procédures permettant d'obtenir les différents paramètres, l'approche par estimation conjointe de l'ensemble de ces paramètres constitue l'objet de cette partie.
Sous l'hypothèse des paramètres de variance connus, il existe des algorithmes d'estimation conjointe des paramètres (β, b, U).
L'approche unifiée d'Henderson, très répandue, permet cette estimation simultanée des paramètres β, b *et* U. Cette approche s'effectue à travers l'optimisation de la logvraisemblance conjointe de b, *et* U, ε. Ces trois processus gaussiens sont indépendants, la logvraisemblance conjointe est alors la somme des trois logvraisemblances marginales (Robinson,[34]).
La maximisation de cette logvraisemblance conduit aux *équations du modèle mixte stochastique*, une extension des équations du modèle mixte ou *MME=Mixed Model Equations* ou encore les *équations d'Henderson*.
La fonction de logvraisemblance ainsi décrite se présente sous la forme :

$$(2\pi)^{-\frac{1}{2}n-\frac{1}{2}q}\det\left[\text{diag}\left(G(\sigma_*^2), \Gamma(\gamma^2,\alpha),\, R(\sigma_0^2)\right)\right]^{-\frac{1}{2}} \times \exp$$
$$\left\{-\frac{1}{2}\begin{pmatrix} b \\ U \\ Y - X\beta - Zb - U \end{pmatrix}^T \left\{\text{diag}\left(G(\sigma_*^2), \Gamma(\gamma^2,\alpha),\, R(\sigma_0^2)\right)\right\}^{-1} \begin{pmatrix} b \\ U \\ Y - X\beta - Zb - U \end{pmatrix}\right\}$$

La maximisation de cette fonction par rapport à β, b et U revient à maximiser :

$$\begin{pmatrix} b \\ U \\ Y - X\beta - Zb - U \end{pmatrix}^T \left\{\text{diag}\left(G(\sigma_*^2), \Gamma(\gamma^2,\alpha),\, R(\sigma_0^2)\right)\right\}^{-1} \begin{pmatrix} b \\ U \\ Y - X\beta - Zb - U \end{pmatrix}$$
$$= b^T G(\sigma_*^2)^{-1} b + U^T \Gamma(\gamma^2,\alpha)^{-1} U + (Y - X\beta - Zb - U)^T R(\sigma_0^2)^{-1}(Y - X\beta - Zb - U)$$

En dérivant cette expression toujours par rapport à β , b *et* U à paramètres de variance connus ou déjà estimés, on obtient les équations d'Henderson :

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} & X^T R^{-1} Z \\ R^{-1} X & R^{-1} + \Gamma^{-1} & R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ U \\ b \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}, \qquad (2.32)$$

Ce système d'équations est une extension au modèle mixte stochastique de l'équation du modèle mixte ; donc la résolution de ce système s'effectue après l'estimation des paramètres de variance. Les solutions de ce système sont des estimateurs BLUE de β et prédicteur BLUP de U et b. En désignant par $\bar{\beta}$, $\bar{U}$ et $\bar{b}$, les solutions de ces équations, on aura alors :

$$\begin{aligned} \bar{U} = & \quad (R^{-1} + \Gamma^{-1})^{-1} R^{-1} (Y - X\bar{\beta} - Z\bar{b}), \\ \bar{b} = & \quad (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} (Y - X\bar{\beta} - \bar{U}). \end{aligned}$$

Il faut noter, dans les différentes expressions de $\bar{U}$, $\bar{b}$ et $\bar{\beta}$, que les matrices Γ,G et R sont des fonctions des paramètres (γ^2, α), σ_*^2 et σ_0^2. Comme ces paramètres ne sont pas connus, ils sont estimés par l'utilisation de l'une des méthodes précédemment décrites. Les estimations de $\hat{\Gamma}$,$\hat{G}$ et $\hat{R}$ obtenues sont alors directement remplacées dans les équations du modèle mixte stochastique pour la résolution. Dans le cas de l'utilisation des estimations $\hat{\Gamma}$,$\hat{G}$ et $\hat{R}$, les estimateurs $\bar{b}$ et $\bar{U}$ mais aussi $\hat{\beta}$ ne sont plus de variance minimale. Pour clore, cette partie, il faut noter qu'il existe des méthodes alternatives de résolution de (2.32).
Les équations d'Henderson réduites pour β et b sont obtenues :

$$\begin{pmatrix} X^T\Xi^{-1}X & X^T\Xi^{-1}Z \\ Z^T\Xi^{-1}X & Z^T\Xi^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X^T\Xi^{-1}y \\ Z^T\Xi^{-1}y \end{pmatrix}, \tag{2.33}$$

avec $\Xi = R + \Gamma$.

2.5 Approche bayésienne du modèle linéaire mixte stochastique

2.5.1 Le modèle linéaire mixte stochastique comme modèle bayésien

Approche bayésienne générale

Le modèle linéaire mixte et ses extensions constituent un interface entre les statistiques classiques fréquentistes et les méthodes bayésiennes. A ce titre, il nous paraît important de rappeler les méthodes d'estimations dans l'approche bayésienne complétant ainsi l'approche fréquentiste du modèle. Cette approche bayésienne contribue à la compréhension de l'approche globale de ce modèle.
Nous exposons l'inférence bayésienne développée sur les familles conjuguées. (Les distributions a priori et a posteriori sont dites conjuguées en ce sens qu'elles appartiennent à la même famille de loi de probabilité.) Le principe de base de la méthode bayésienne, dans ce cas, repose sur l'estimation de cette distribution *a posteriori* qui s'obtient par la mise à jour de celle dite *a priori* à l'aide de l'information empirique issue des observations. La distribution *a priori* est choisie suivant une connaissance plus ou moins experte des données. L'approche bayésienne du modèle mixte stochastique est simple à exposer à travers le mécanisme d'un modèle linéaire hiérarchique.
Ce mécanisme consiste à caractériser les distributions conditionnelles des observations et

des paramètres aux différents niveaux du modèle linéaire hiérarchique. Ainsi le premier niveau consiste à spécifier une distribution conditionnelle des observations en supposant les paramètres $\phi = (\beta, \theta)$ connus. Puis la spécification à un second niveau des distributions a priori des paramètres ϕ sachant les hyperparamètres connus, ainsi de suite ....
Soit, suivant le schéma simplifié ci-dessous :

$$Y|\phi \sim f(y, \phi),$$
$$\phi \sim \pi(\phi).$$

Dans un modèle hiérarchique, le premier niveau concerne toujours la distribution des observations disponibles.
L'application du théorème de Bayes permet alors le calcul des distributions a posteriori des paramètres ainsi :

$$\pi(\phi|y) = \frac{f(y|\phi)\pi(\phi)}{\int f(y|\phi)\pi(\phi)d\phi},$$

où $\pi(\phi)$ est la distribution a priori et $f(y|\phi)\pi(\phi)$ la distribution conjointe de Y et de ϕ, tandis que $\int f(y|\phi)\pi(\phi)d\phi$ est la distribution marginale de Y. L'estimation du paramètre ϕ se fait alors par sa moyenne a posteriori qui s'obtient par l'espérance mathématique de cette dernière, une fois sa distribution a posteriori bien spécifiée :

$$\phi = \int \phi \frac{f(y|\phi)\pi(\phi)}{\int f(y|\phi)\pi(\phi)d\phi} d\phi. \tag{2.34}$$

L'approche bayésienne du modèle mixte stochastique

Après ce rappel succinct sur la méthode bayésienne en général, l'application de l'estimation bayésienne au modèle mixte stochastique se décline suivant la procédure suivante. En reprenant le modèle de l'équation (2.5), les données observées sont les variables y, X et Z, les paramètres inconnus sont γ^2, α, σ_*^2 et σ_0^2 dont dépendent les matrices de variance Γ, G et R. Enfin les paramètres d'effets β, et b, ainsi que le processus stochastique U sont aussi non observés. D'après cette description, l'approche du modèle mixte stochastique peut s'effectuer en trois niveaux.

- Le premier niveau concerne la distribution des résidus ε qui est aussi celle des données observées. Cette distribution est conditionnelle à la connaissance des paramètres que sont β, b mais aussi le processus U.
- Au deuxième niveau,les distributions a priori seront alors portées, sur les paramètres inconnus d'effets mixtes (β, b) et le processus stochastique U.
- Enfin, les hyperparamètres que sont σ_*^2, γ^2 et α dont dépendent les matrices G, Γ et R les matrices de variance covariance forment le troisième niveau.

Ce troisième niveau ajouté concerne essentiellement l'estimation des paramètres de variance des matrices G, Γ et R. Ces différentes étapes peuvent se résumer en :

1. Sachant U, b, β, R,
$$y \sim \mathcal{N}(X\beta + Zb + U, R),$$

2. Sachant β_0, B, G, Γ,

$$\beta \sim \mathcal{N}(\beta_0, B)\ ; b \sim \mathcal{N}(0, G)\ ; U \sim \mathcal{N}(0, \Gamma);$$

3.

$$(B, G, \Gamma, R) \sim \Pi_{B,G,\Gamma,R}(\ ,\ ,\ ,\ ).$$

Dans ce troisième niveau (B, G, Γ, R), la loi a priori et a posteriori conjuguées pour les matrices de covariance sont des lois de type de Wishart. Pour plus de détails sur cette loi, on peut se référer à Rao [32]. Cette approche, dans le cadre de l'estimation des composantes de la variance, conduit à des évaluations numériques complexes et délicates à mettre en oeuvre.
L'approche alternative souvent utilisée est la méthode d'estimation empirique de Bayes basée sur la densité conditionnelle induite pour les paramètres du troisième niveau de la procédure.

2.5.2 Les méthodes du maximum de vraisemblance et l'approche bayésienne : estimation empirique de Bayes

Ainsi, en définissant le modèle hiérarchique suivant :

$$\begin{aligned} y &\sim f(y|\beta, b, U, R), \\ \beta \sim f_\beta(\beta|\beta_0, B),\ b &\sim f_b(b|G),\ U \sim f_U(U|\Gamma), \end{aligned}$$

Nous attribuons un a priori non informatif à β, on obtient :

$$\begin{aligned} y|\beta, B, U &\sim \mathcal{N}(X\beta + Zb + U, R), \\ b \sim \mathcal{N}(0, G),\ U &\sim \mathcal{N}(0, \Gamma). \end{aligned}$$

La formulation de la densité de loi conditionnelle pour (B, G, Γ, R) connaissant les deux premiers niveaux de spécification du modèle linéaire hiérarchique et l'intégration par rapport à ces différents paramètres, permet d'obtenir la forme de vraisemblance ci-dessous :

$$\begin{aligned} l_r(G, \Gamma, R|y) = & \int\int\int L(\beta, G, \Gamma, R|y) dU d\beta db, \\ = & \ (2\pi)^{-\frac{N-r_x}{2}} |P^T V P|^{-\frac{1}{2}} \exp\left\{ y^T P \left\{ P^T V P \right\}^{-1} P^T y \right\}. \end{aligned}$$

On peut montrer que $l_r(G, \Gamma, R|y) = |X^T V X|^{-\frac{1}{2}} l(\beta, G, \Gamma, R|y)$ (pour la démonstration voir [37]).
Cette forme de vraisemblance peut se noter comme une fonction des paramètres θ soit :

$$l_r(\theta) = (2\pi)^{-\frac{N-r_x}{2}} |P_\theta^T V_\theta P_\theta|^{-\frac{1}{2}} \exp\left\{ y^T P_\theta \left\{ P_\theta^T V_\theta P_\theta \right\}^{-1} P_\theta^T y \right\}.$$

L'approche par la maximisation de la quantité ci-dessus est la méthode du bayésien empirique. C'est une méthode équivalente à la méthode d'estimation REML pour les paramètres

de variance du modèle à condition que les distributions a priori des paramètres soient non informatives.
Cette approche du modèle mixte stochastique nous permet de voir que des paramètres du modèle, découlant d'une formulation bayésienne peuvent être estimés par la méthode du REML. Les différentes approches du modèle mixte stochastique développées dans ce chapitre établissent les bases des procédures nécessaires pour la suite de notre travail.

Chapitre 3

Modèles mixtes semipramétriques stochastiques

3.1 Introduction aux méthodes semi-paramétriques

Les techniques semi-paramétriques viennent en complément des méthodes classiques de régression, elles constituent une approche beaucoup plus exploratoire d'analyse des données que leur contrepartie paramétrique. Formellement, elles consistent à remplacer (ou à compléter) les fonctions paramétriques usuelles de modélisation (polynômes, exponentielles, ...) par des opérateurs de lissage.
La régression semiparamétrique sur données indépendantes est largement traitée dans la littérature selon des approches fréquentiste ou bien bayésienne ; une illustration parfaite est la popularité dont jouit aujourd'hui la classe des modèles additifs généraisés GAM (Hastie & Tibshirani, [20]) (Generalized Additive Models). Les techniques les plus courantes s'appuient sur l'utilisation du lissage par noyau [46, 12], par les fonctions splines [12, 47, 21], ou bien récemment par les ondelettes [14].
Dans le chapitre précédent, nous avons présenté le modèle linéaire mixte stochastique : ce dernier est un modèle mixte présentant des composantes paramétriques fixes $X\beta$, aléatoires Zb et stochastiques $U(t)$. Le but de ces dernières est de modéliser, sur chaque site de mesure, les corrélations sérielles issues de la répétition des mesures. Etape supplémentaire dans la conception de notre modèle, nous souhaitons, dans ce chapitre 3, modifier et enrichir le modèle mixte stochastique en incluant une composante fonctionnelle non-paramétrique $f(t)$ dite *composante commune* ou (*baseline* en anglais). En théorie de nature "infini-dimensionnelle", une telle composante $f(t)$ sera bien sûr approchée à l'aide d'un estimateur de "dimension finie" choisi dans un espace d'approximation adapté.
Le choix des splines de régression pénalisées est motivé par le fait que ces dernières, possédant l'aspect «gaussian model shift », sont structurellement "compatibles" avec les modèles mixtes gaussiens et l'inférence à leur sujet pourra être réalisée à la base par les mêmes techniques déjà décrites au chapitre précédent (REML, équations d'Henderson, prédicteurs linéaires). Ainsi une seule procédure globale permettra l'estimation/prédiction de toutes les composantes du modèle, y compris la composante fonctionnelle $f(t)$. L'inclusion d'une

composante $f(t)$ au sein du modèle mixte stochastique peut se justifier pour de multiples raisons :
(1) cela peut servir à pallier le manque de variables explicatives exogènes du modèle construit : en effet, dans de nombreuses situations d'analyse de données environnementales, le praticien ne peut disposer de toutes les variables explicatives $X(t)$ dont il aurait besoin pour un modèle complet. L'approche semiparamétrique consiste alors à inclure une base $B(t)$ de variables endogènes (de type fonctions splines par exemple). Ces variables endogènes pourront "capturer" une part de la variabilité des données et ainsi éviter le report de cette variabilité sur les processus spécifiques $U_i(t)$ ou sur les résidus $\epsilon(t)$.
(2) la dimension de l'espace de représentation de la composante de la moyenne est augmentée, et cet apport de degrés de liberté permet d'obtenir des estimateurs plus adaptatifs pour les composantes de la variance.
Un ensemble de simulations sont présentées en deuxième partie de ce chapitre, elles visent à éclairer l'utilisateur sur les différents aspects évoqués ci-dessus, mais aussi dans la mise en œuvre pratique du modèle mixte : en particulier sur le compromis biais-variance. En particulier, la question du choix du paramètre de lissage sera abordé dans le contexte du type de modèle proposé c'est-à-dire les splines de régression pénalisées. La mise en œuvre pratique d'un modèle mixte semiparamétrique stochastique demande une économie, autant que possible, du nombre de nœuds servant à la construction de la base de représentation car, outre les problèmes numériques éventuels (inversion de matrices,...), il faut éviter le surajustement (*overfitting*) du modèle sur les données.

3.2 Outils semi-paramétriques usuels

3.2.1 Opérateurs de lissage : définition générale

Les opérateurs de lissage sont les outils incontournables dans les techniques de modélisation semiparamétriques. Ces techniques utilisent une fonction lisse pour l'approximation de la composante nonparamétrique (pour plus de détails sur l'utilisation du lissage dans les méthodes semiparamétriques, voir Hastie et Tibshirani [20]). Un opérateur de lissage se définit formellement comme une fonction permettant de passer d'une représentation "discrète" d'une fonction (un nuage de points $\{X_i, y_i\}_{i=1,\cdots,N}$ dans $\mathbb{R}^p \times \mathbb{R}$) à une représentation "continue" de celle-ci : $X \mapsto y = f(X)$ où f est une fonction continue de $\mathbb{R}^p$ à valeurs dans $\mathbb{R}$.
Soit $Y = (y_1, \cdots, y_N)$ un vecteur de N observations réelles, et soit $X = (x_1, \cdots, x_N)^T$ une matrice $N \times p$ de rang plein constituée de N vecteurs ligne $x_i \in \mathbb{R}^p \quad i = 1, \cdots, N$.
On appelle *opérateur de lissage* toute fonction $\mathcal{S}$ de $\mathbb{R}^{N\times(p+1)}$ dans un espace fonctionnel $\mathcal{H}(\mathbb{R}^p)$: $\quad (X, Y) \longmapsto \mathcal{S}(Y|X)$.
Ainsi $\mathcal{S}(Y|X)$ est une fonction de la variable $x \in \mathbb{R}^p$, à valeur dans $\mathbb{R}$, possédant à la fois des propriétés d'approximation, d'interpolation et de lissage :
(a) approximation : en tout point x_i de l'échantillon, l'évaluation $\mathcal{S}(Y|X)(x_i)$ doit être proche de la valeur y_i au point x_i ;

(b) interpolation : en tout nouveau point $x_0 \in \mathbb{R}^p$ non inclus dans l'échantillon, l'évaluation $\mathcal{S}(Y|X)(x_0)$ existe et doit tenir compte des valeurs y_i aux points voisins ;
(c) lissage : la fonction $x \longmapsto \mathcal{S}(Y|X)(x)$ doit posséder un degré de régularité suffisant. Dans la pratique, un *paramètre de lissage* devra permettre à l'utilisateur de choisir ou contrôler le niveau de régularité souhaité. On notera dorénavant $\mathcal{S}_\lambda(Y|X)$ la forme générale de l'opérateur de lissage, l'indice λ représentant le paramètre de contrôle du niveau de régularité.

Par ailleurs, placé dans un contexte de statistique inférentielle, nous considérerons que les observations y_i $i = 1, \cdots, N$ sont en fait issues d'un modèle de la forme :

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \cdots, N,$$

avec $(\epsilon_1, \cdots, \epsilon_N)$ un vecteur de variables aléatoires réelles i.i.d. d'espérance nulle, et f une fonction (inconnue) réelle continue et suffisamment régulière de la variable x. Dans ces conditions, l'opérateur de lissage $\mathcal{S}$ pourra être vu comme un estimateur de f :

$$\hat{f}_\lambda(x_0) = \mathcal{S}_\lambda(Y|X)(x_0), \qquad x_0 \in \mathbb{R}^p.$$

Nous exposons à présent les opérateurs usuels de lissage ; par souci de clarté de l'exposé, nous présentons les choses dans le cas unidimensionnel $p = 1$, car en dimension supérieure ($p > 1$) l'écriture est beaucoup plus complexe. La variable X étant unidimensionnelle, elle sera notée t dans la suite.

3.2.2 Régression locale

Cette approche consiste à effectuer en chaque t_0 une régression polynomiale de degré d, avec une pondération des observations variable d'un t_0 à l'autre : pour un point $t_0 \in \mathbb{R}$ fixé, les poids $w_{0,i}$ utilisés donnent plus d'importance aux données voisines de t_0 par l'application d'un noyau $K(\,\cdot\,)$:

$$w_{0,i} = \frac{1}{h} K\left(\frac{|t_0 - t_i|}{h}\right), \qquad i = 1, \cdots, N.$$

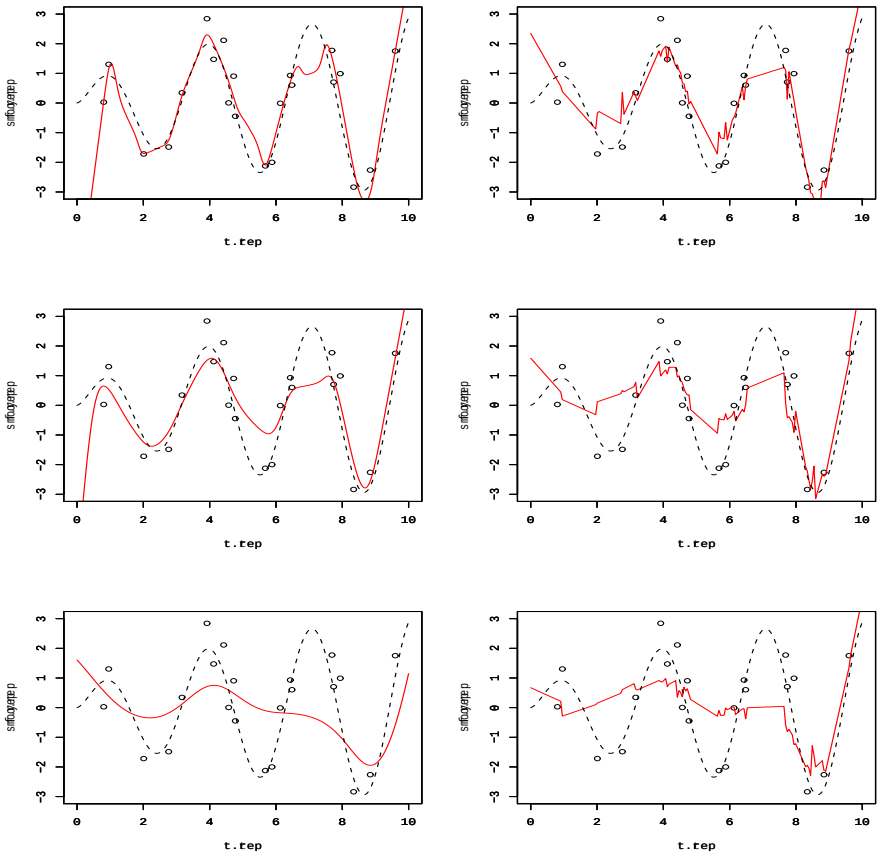
Les noyaux les plus populaires sont :

$$\begin{aligned} \text{Noyau Gaussien} &: \quad K_G(t) = exp(-t^2/2)/\sqrt{2\pi}, \\ \text{Nayau d'Epanechnikov} &: \quad K_E(t) = \tfrac{3}{4}(1-t^2)\ ; \ \ |t| < 1, \\ \text{Noyau de variance minimale} &: \quad K_{MV}(t) = \tfrac{3}{8}(3-5t^2)\ ; \ \ |t| < 1. \end{aligned}$$

Dans le cas d'une régression locale de degré d, en notant :

$$W_h(t_0) = \left(\sum_i w_{0,i}\right)^{-1} \times \text{diag}\{w_{0,1}, \cdots, w_{0,N}\},$$

FIG. 3.1 – A gauche : lissage par régression locale avec $h = 0.3$, 0.5 et 1 ; A droite : lissage par régression glissante avec $k = 2$, 3 et 4. En pointillé, la vraie fonction $f(t) = \sqrt{(t)} * sin(2t)$ sur $[0, 10]$.



$\mathcal{X} = (\mathbf{1}_N, T, T^2, \cdots, T^d)$ où $T^j = (t_1^j, \cdots, t_N^j)^T$, et $\mathcal{X}_0 = (1, t_0, t_0^2, \cdots, t_0^d)^T$, on écrira :

$$\mathcal{S}_h(Y|X)(t_0) = \mathcal{X}_0^T(\mathcal{X}^T W_h(t_0)\mathcal{X})^{-1}(\mathcal{X}^T W_h(t_0)Y).$$

Le cas de la régression locale de degré $d = 0$ correspond au célèbre opérateur de Nadaraya-Watson : $\mathcal{S}_h(Y|X)(x_0) = \sum w_{0i} y_i \,/\, \sum w_{0i}$;

3.2.3 Régression glissante sur les k-plus-proches-voisins

L'ensemble $\{t_1, \cdots, t_N\}$ étant ordonné par ordre croissant, définissons $N_k(t_0)$ le voisinage (symétrique) des k-plus proches voisins de t_0 de la manière suivante :

cas 1 : $t_0 \in \{t_1, \cdots, t_N\}$; par exemple : $t_0 = t_i$.

$$N_k(t_0) = \{max(i-k, 1), \cdots i-1, i, i+1, \cdots, min(i+k, N)\},$$

cas 2 : $t_0 \notin \{t_1, \cdots, t_n\}$; t_i et t_{i+1} sont les deux points de l'échantillon encadrant t_0.

$$N_k(t_0) = \{max(i-k+1, 1), \cdots i, i+1, \cdots, min(i+k, N)\}.$$

Notons $W_k(t_0) = \text{diag}\ \{\mathbf{1}_{N_k(t_0)}(1), \cdots, \mathbf{1}_{N_k(t_0)}(N)\}$ avec $\mathbf{1}_{N_k(t_0)}(i)$ égal à 0 ou 1 selon que la donnée i fait partie ou non du k-voisinage de t_0.
Alors $\mathcal{S}(Y|X)$ est évaluée au point t_0 par une régression polynomiale d'ordre d (moindres carrés ordinaires) sur les données sélectionnées par $W_k(t_0)$; avec les mêmes notations que précédemment, cela donne :

$$\mathcal{S}(Y|X)_k(t_0) = \mathcal{X}_0^T(\mathcal{X}^T W_k(x_0)\mathcal{X})^{-1}(\mathcal{X}^T W_k(x_0) Y).$$

3.2.4 Régression spline (polynomiale par morceaux)

La régression spline offre un bon compromis entre un ajustement global (tel que celui d'une régression polynomiale) et un ajustement purement local (tel que celui de la régression locale décrite ci-dessus). Dans son principe, la régression spline n'est pas différente d'une régression linéaire multiple classique (par moindres carrés ordinaires), elle utilise comme espace de représentation de $g(t)$ un espace (vectoriel) de dimension d de fonctions polynomiales par morceaux (encore appelées fonctions splines). Etant donnée une suite (ordonnée) de r nœuds *intérieurs* $(\kappa_j)_{j=1,\cdots,r}$, une fonction spline $g(t)$ d'ordre p est par définition une fonction polynomiale par morceaux, de degré $(p-1)$ entre deux nœuds successifs. Le domaine sur lequel la fonction $g(t)$ est non nulle, est délimité par deux nœuds *extérieurs* κ_0 et κ_{r+1}.
Le niveau de régularité de $g(t)$ est assuré en imposant un certain nombre de contraintes de continuité et/ou de nullité sur les dérivées successives aux points-nœuds (κ_j). Ces contraintes varient selon le *type* de fonctions splines construites (puissance tronquées, B-splines, N-splines, etc.) et seront précisées dans chaque cas. Les fonctions splines d'un type donné forment un espace vectoriel dont la dimension d est déterminée par le degré $(p-1)$ des polynômes, le nombre r de nœuds intérieurs et le nombre c de contraintes sur les dérivées successives.

Splines de type «puissances tronquées»

Une base relativement intuitive de fonctions splines est celle dite des «puissances tronquées», permettant de générer des fonctions splines sous la forme :

$$g(t) = \delta_0 + \delta_1 t + \delta_2 t^2 + \cdots + \delta_{p-1} t^{p-1} + \sum_{j=1}^{r} a_j^T (t - \kappa_j)_+^{p-1}, \tag{3.1}$$

avec :

$$\forall j = 1, \cdots, r \qquad (t-\kappa_j)_+ = \begin{cases} (t-\kappa_j) & \text{si} \quad (t-\kappa_j) \geq 0, \\ 0 & \text{sinon.} \end{cases}$$

Toute fonction spline de la forme (3.1) possède les propriétés de régularité suivantes :
– elle est continue sur tout le domaine de définition $(\kappa_0\,,\,\kappa_{r+1})$;

– ses dérivées successives sont également continues, jusqu'à l'ordre $p-1$.

Bien que simple dans sa construction, la représentation (3.1) présente l'inconvénient de ne pas être stable sur un point de vue numérique, rendant parfois délicate la mise en œuvre sur le plan pratique.

Splines de type B-splines

Une autre base beaucoup plus utilisée que les puissances tronquées est celle des "Basic"-splines (ou encore B-splines), appellation qui, d'après de Boor, provient de Schoenberg en 1946. Les fonctions B-splines sont souvent préférées pour leurs propriétés numériques, et aussi pour leur relative facilité de mise en œuvre algorithmique. Etant donné un ordre fixé p, la j-ème fonction B-spline d'ordre p sera notée $B_{j,p}(t)$, $j = 1-p, \cdots, r$. Elle se construit récursivement à partir des fonctions B-splines $B_{j,p-1}(t)$ d'ordre $p-1$ grâce à l'algorithme de Cox De Boor basé sur les différences divisées :

$$B_{j,p}(t) = \frac{t-\kappa_j}{\kappa_{j+p-1}-\kappa_j} B_{j,p-1}(t) + \frac{\kappa_{j+p}-t}{\kappa_{j+p}-\kappa_{j+1}} B_{j+1,p-1}(t)\ ; \qquad \forall j = 1-p, \cdots, r,$$

initialisé par $(p=0)$:

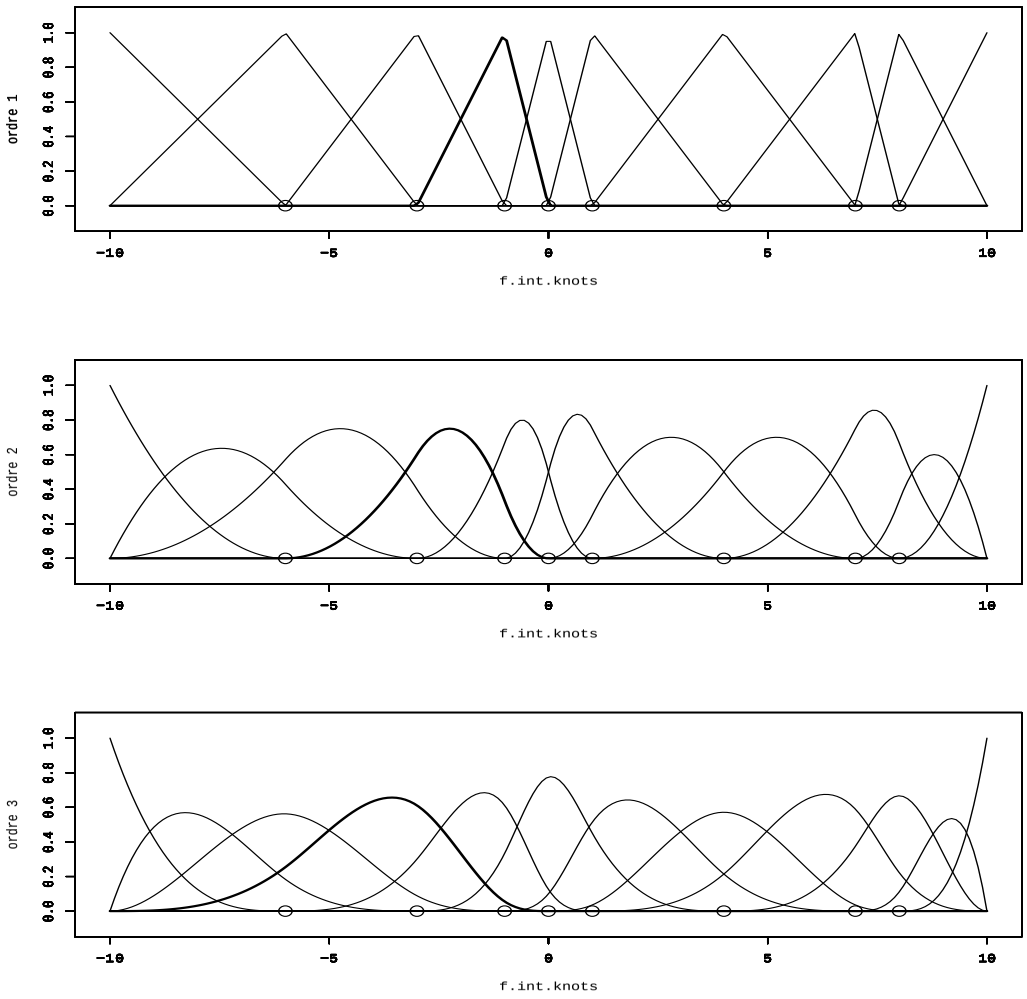
$$B_{j,1}(t) = \begin{cases} 1 & \text{si} \quad t \in (\kappa_j ; \kappa_{j+1}), \\ 0 & \text{sinon,} \end{cases} \quad \forall\, j = 0, \cdots, r.$$

Cette définition par récurrence permet de voir qu'effectivement :

– $B_{j,p}(t)$ est une fonction de type "polynôme par morceaux" de degré $(p-1)$ ayant pour nœuds $\kappa_j, \cdots, \kappa_{j+p}$;
– $B_{j,p}(t)$ est à support local ; elle est positive sur l'intervalle $(\kappa_j ; \kappa_{j+p})$ et nulle en dehors de cet intervalle ;
– A t fixé, la somme des $B-$splines d'ordre p vaut 1 : $\sum_{j=1-p}^{m} B_{j,p}(t) = 1$.

Ces propriétés sont bien établies aujourd'hui, le lecteur pourra consulter par exemple les travaux de de Boor [10, 11] sur le sujet.
A titre d'illustration, nous avons représenté en figure (3.2) les trois familles de B-splines, respectivement d'ordre $p=1$, $p=2$ puis $p=3$, et ceci pour un même jeu de huit nœuds intérieurs.

FIG. 3.2 – Fonctions B-splines de différent ordre p =1,2 et 3



Un théorème dû à Schoenberg établit que, à p fixé, la base que forment les B-splines $B_{j,p}(t)$ $j = 1-p, \cdots, r$; est équivalente à celle des puissances tronquées ; c'est-à-dire que toute fonction $g(t)$ écrite sous la forme (3.1) peut aussi s'écrire sur la base des B-splines : $g(t) = \sum_{j=1-p}^{r} a_j^B B_{j,p}(t)$, les coefficients a_j^B pouvant s'obtenir aisément à partir des relations de passage suivantes (dites relation de Marsden) :

$$\forall j = 1 \cdots, r \qquad (t-\kappa_j)_+^{p-1} = \sum_{j'>j} \psi_{j',p}(\kappa_j) B_{j',p}(t) \ ;$$

avec :

$$\psi_{j,p}(t) = (\kappa_{j+1} - t) \cdots (\kappa_{j+p-1} - t).$$

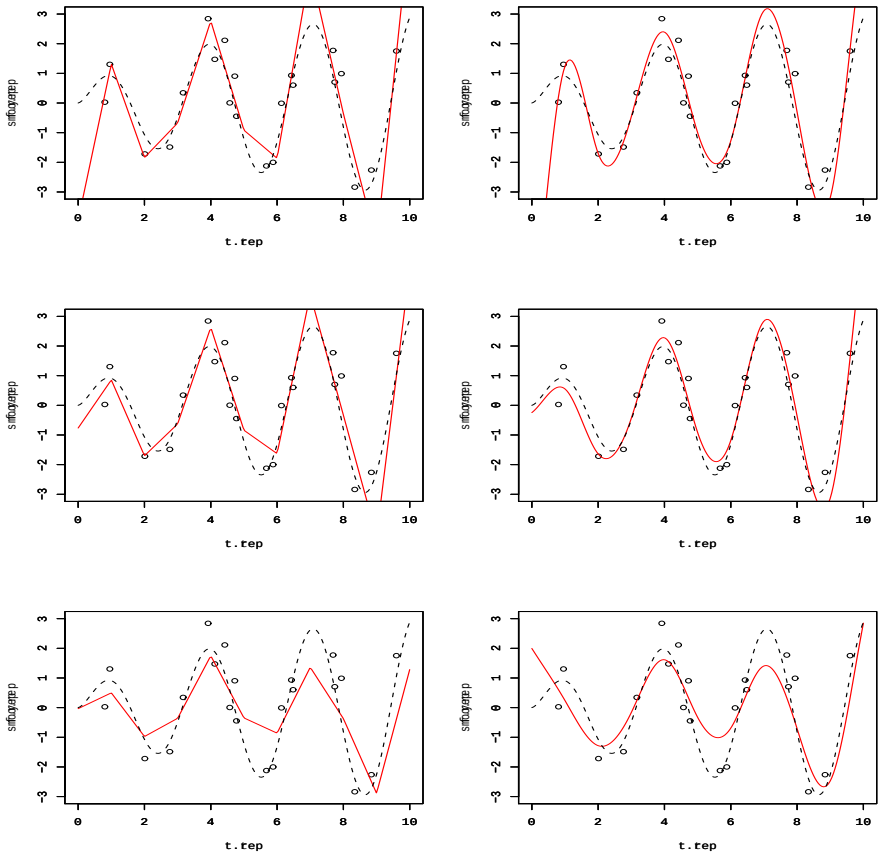
Lissage par régression spline pénalisée

Nous revenons sur la construction d'un opérateur de lissage par la technique de régression spline décrite ci-dessus ; nous souhaitons bien sûr pouvoir contrôler la régularité de la fonction $g(t)$ issue du processus de lissage ; il y a pour cela plusieurs moyens d'action possibles : soit agir sur le nombre de nœuds (c'est-à-dire la dimension de l'espace de représentation), soit agir sur leur position, soit encore agir sur les coefficients de régression eux-mêmes en appliquant une technique de pénalisation.
Ayant choisi une base de fonctions splines d'ordre p fixé, $\mathcal{B}_p = \{B_{1,p}(t), \cdots, B_{q,p}(t)\}$, notant par B la matrice $N \times q$ constituée des valeurs $B_{ij} = B_{j,p}(t_i) \quad i = 1, \cdots, N \quad j = 1, \cdots, q$, et notant $b_0 = (B_{1,p}(t_0), \cdots, B_{q,p}(t_0))^T$ le vecteur $q \times 1$ des valeurs des splines de base au point t_0, la régression spline pénalisée au point t_0 s'écrit :

$$\mathcal{S}_\lambda(Y|X)(t_0) = b_0^T (B^T B + \lambda\Omega)^{-1} B^T Y, \tag{3.2}$$

où Ω est une matrice de pénalisation fixée (symétrique définie positive de taille $q \times q$), Ω peut être obtenue par construction dans le cas des splines de lissage (voir Green & Silvermann [17]) ; Dans la suite,nous verrons d'autres de éthodes de choix de la matrice Ω. Le contrôle de régularité s'effectue en faisant varier le paramètre λ (une valeur élevée de λ correspond à un effet de lissage fort). Pour le cas des splines de régression pénalisées, nous reviendrons ultérieurement sur la question du choix de la matrice de pénalisation Ω par la suite. Un choix raisonné de la position des nœuds permet aussi de contrôler *localement* la régularité de l'opérateur de lissage.

FIG. 3.3 – Lissage par régression pénalisée de fonctions B-splines d'ordre 2 (à gauche) et d'ordre 3 (à droite) avec $\lambda = 0.01$,0.1 et 0.5 ; en pointillé, la vraie fonction $f(t) = \sqrt{(}t) * sin(2t)$ sur $[0, 10]$



3.2.5 Quelques propriétés générales des opérateurs de lissage linéaires dans le contexte de la régression nonparamétrique

Plaçons-nous dans le contexte du modèle de régression nonparamétrique :

$$y_i = f(t_i) + \varepsilon_i, \qquad i = 1, \quad , N,$$

avec $(\varepsilon_1, \cdots, \varepsilon_N)$ des aléas i.i.d. vérifiant $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{V}[\varepsilon_i] = \sigma^2$. Ainsi les opérateurs de lissage $\mathcal{S}_\lambda(Y|X)$ sont des estimateurs potentiels pour le paramètre fonctionnel $f \in \mathcal{H}$.

Propriété de linéarité

Tout d'abord, signalons que les opérateurs $\mathcal{S}_\lambda(Y|X)$ que nous venons de présenter dans la section précédente ont tous la propriété de *linéarité* par rapport aux observations Y, (voir

Hastie & Tibshirani, [20]) à savoir :

$$\mathcal{S}_\lambda(\alpha_1 Y_1 + \alpha_2 Y_2|X) = \alpha_1 \mathcal{S}_\lambda(Y_1|X) + \alpha_2 \mathcal{S}_\lambda(Y_2|X),$$

quels que soient $Y_1 \in \mathbb{R}^N$, $Y_1 \in \mathbb{R}^N$, $\alpha_1 \in \mathbb{R}$ et $\alpha_2 \in \mathbb{R}$.
Cela implique en particulier, que le vecteur $\hat{f}_\lambda = (\hat{f}_\lambda(t_1), \cdots, \hat{f}_\lambda(t_N))^T$ s'écrit comme transformation linéaire du vecteur des valeurs observées $y = (y_1, \cdots, y_N)^T$:

$$\hat{f}_\lambda = S_\lambda y.$$

La matrice S_λ de taille $N \times N$, dans le cas général, ne dépend que de la matrice d'expérience $\mathcal{X}$ et du paramètre de lissage λ. Tous les opérateurs de lissage ne sont pas linéaires en Y.
Dans le cas particulier du lissage par régression spline pénalisée (3.2), la matrice S_λ peut s'écrire :

$$\begin{aligned} S_\lambda &= B(B^T B + \lambda\Omega)^{-1} B^T \\ &= (I_N - \lambda K)^{-1}, \end{aligned}$$

avec

$$K = B^{-T} \Omega B^{-1}.$$

A ce titre, nous rappelons quelques propriétés et formules fondamentales, basées sur l'écriture de la matrice S_λ.

Mesures d'erreur d'estimation

De manière théorique, la performance d'un estimateur fonctionnel $\hat{f}_\lambda$ est usuellement évaluée, en un point t_0 fixé, par l'erreur quadratique moyenne MSE (en anglais *mean square error*) :

$$\mathrm{MSE}_{\hat{f}_\lambda}(t_0) := \mathbb{E}[\hat{f}_\lambda(t_0) - f(t_0)]^2,$$

qui, selon un calcul très classique, s'exprime comme la somme de la variance et du carré du biais :

$$\mathrm{MSE}_{\hat{f}_\lambda}(t_0) = \mathrm{var}[\hat{f}_\lambda(t_0)] + \{\mathbb{E}[f_\lambda(t_0)] - f(t_0)\}^2.$$

Pour l'ensemble des instants mesurés $\mathrm{t} = (t_1, \cdots, t_N)$ nous écrirons :

$$\begin{aligned} \mathrm{MSE}_{\hat{f}_\lambda(\mathrm{t})} &= \mathrm{var}\left[\hat{f}_\lambda\right] + \{E\left[\hat{f}_\lambda\right] - \mathrm{f}\}^2, \\ &= \sigma^2 \mathrm{diag}\left[S_\lambda S_\lambda^T\right] + (S_\lambda - I_N)\mathrm{ff}^{\mathrm{T}}(S_\lambda - I_N)^T. \end{aligned}$$

où $\hat{f}_\lambda = (\hat{f}_\lambda(t_1), \ldots, \hat{f}_\lambda(t_N))$ Lorsque le MSE est intégré sur le domaine de définition de f, cela donne une mesure de performance globale de l'estimateur $\hat{f}_\lambda$, appelée MISE (en anglais *mean integrated square error*) :

$$\mathrm{MISE}_\lambda = \int_{\mathcal{T}} \mathbb{E}[\hat{f}_\lambda(u) - f(u)]^2 du.$$

Lorsque le calcul du MISE est trop complexe, il est parfois remplacé par le ISB, (*Integrated Squared Bias*) dont la formule est :

$$\mathrm{ISB}_\lambda = \int_{\mathcal{T}} \left[\mathbb{E}\left[\hat{f}_\lambda(u)\right] - f(u)\right]^2 du.$$

ou sinon par la moyenne empirique du MSE [20] sur les points t_i observés :

$$\begin{aligned} \mathrm{MSSE}_\lambda &= N^{-1}\sum_{i=1}^{N} \mathrm{MSE}_{\hat{f}_\lambda}(t_i), \\ &= N^{-1}\sigma^2 \mathrm{tr}[S_\lambda S_\lambda^T] + N^{-1}(S_\lambda - I_N)^T \mathrm{f}^T \mathrm{f}(S_\lambda - I_N). \end{aligned}$$

Mesures d'erreur basées les résidus observés

$$\mathrm{MSR}_\lambda(t_i) = \mathbb{E}[f_\lambda(t_i)] - y_i\}^2.$$

En particulier sur l'ensemble des points observés $t_i\, i = 1 \cdots, N$, on calcule la moyenne empirique des $\mathrm{MSR}_\lambda(t_i)$ cela donne le MASR (en anglais *mean average square residual*) :

$$\mathrm{MASR}_\lambda = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[f_\lambda(t_i) - y_i\}^2.$$

Degrés de liberté et rang de S_λ

Comme dans le cadre du modèle paramétrique, des degrés de liberté sont attribués aussi dans le cadre d'une modélisation nonparamétrique, ou semipramétrique. Cette analogie avec les modèles paramétriques permet la comparaison des différents modèles nonpramétriques. Dans le cas de l'utilisation d'un opérateur de lissage S_λ, le degré de liberté est donné par : $tr(S_\lambda)$.
La matrice S_λ associée à l'opérateur de lissage est symétrique, donc la transformation associée appliquée à y peut s'écrire comme :

$$S_\lambda(y) = \sum_{j=1}^{J_0} \ell_j < v_j, y >,$$

où seules un certain nombre (J_0) des valeurs propres ℓ_j sont non nulles. Les valeurs propres ℓ_j associées aux deux premiers vecteurs propres sont égales à l'unité. (Ces deux vecteurs propres sont la constante et le monôme de degré 1 de la variable explicative t). Cela traduit le fait que le lissage par fonctions splines conserve les tendances linéaires. Toutes les autres valeurs propres (en dehors de celles associées au monôme et la constante) de la matrice de lissage sont inférieures à 1 en valeur absolue. Ce qui traduit ainsi l'effet «shrinkage» associé au lissage spline.

3.2.6 Régression splines à coefficients aléatoires

Pour exposer la régression splines à coefficients aléatoires, nous revenons sur la représentation dans la base de puissances tronquées. La représentation des splines d'après (3.1) considère les coefficients δ_i, i=1, ..., $(p-1)$ et a_j, j=1, ..., d, comme fixes, non aléatoires. Il est aisé de remarquer que l'équation (3.1) constitue une équation de régression linéaire dont les coefficients sont δ_i, i=1, ..., $(p-1)$ et a_j, j=1, ..., r. Ce modèle de régression peut se représenter comme un modèle linéaire mixte dont les facteurs à effets aléatoires sont $(t-\kappa_j)_+^{p-1}$,$j=1,\ldots,d$. Les coefficients a_j, $j=1,\ldots,r$, deviennent alors aléatoires donnant ainsi le modèle de *splines à coefficients aléatoires*.
Les conditions sur les dérivées de la fonction et le niveau de régularité sont réalisées à travers la relation entre la variance de ces coefficients et le paramètre de lissage. Cette approche est similaire à celle de Ruppert et *al* [35] et permet d'appliquer des méthodes d'inférence propres au cadre du modèle mixte aux fonctions nonparamétriques.

3.2.7 Régression semiparamétrique et logvraisemblance pénalisée

Présentons le problème général de la régression semiparamétrique au travers du modèle simple suivant : soit $y=(y_1,y_2,\cdots,y_N)$ un échantillon de taille N de la variable (aléatoire) réelle Y et soit $X=(x_1,\cdots,x_N)$ les covariables associées (formant une matrice $N\times p$). On particularise la k-ième covariable de X (supposée continue) et on la note t ; la relation linéaire entre X et y est alors supposée de la forme :

$$y_i = x_i^T\beta + f(t_i) + \varepsilon_i, \quad i=1, \quad ,N;$$

avec les hypothèses :
• $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_N)$ est un vecteur gaussien centré de loi $\mathcal{N}(0,V)$. La structure de variance des résidus est supposée connue à un scalaire σ^2 près : $V=\sigma^2 R$, R connue.
• la covariable t est une variable continue (à valeur dans un compact $\mathcal{T}$) et f est une fonction inconnue (a priori non paramétrique) de $\mathcal{T}\longrightarrow\mathbb{R}$, continûment dérivable, de la covariable réelle t. Pour des raisons d'identifiabilité, on impose par exemple la condition :

$$\int_{\mathcal{T}} f(t)dt = 0.$$

Le problème posé est l'estimation des paramètres inconnus $\beta\in R^p$, de la composante fonctionnelle $f(t)$ et du paramètre de variance $\sigma^2>0$. Ce problème n'est pas fondamentalement plus complexe que le problème initial de la régression nonparamétrique. La résolution de ce problème est menée selon une approche de vraisemblance pénalisée, que nous décrivons à présent.
Nous désignerons par $\mathcal{H}$, un espace fonctionnel dont les éléments approximent de manière satisfaisante la fonction $f(t)$ ($\mathcal{H}$ est un espace de hilbert). L'espace $\mathcal{H}$ peut se scinder en somme directe de deux sous espaces $\mathcal{H}_0\oplus\mathcal{H}_1$ selon un niveau de régularité. Ce niveau de régularité est déterminé par un critère précis, par exemple un ordre de différentiabilité. Le

problème peut se ramener à éviter que les fonctions approximant $f(t)$ ne dépassent pas le niveau de régularité fixé [18], une façon de contrôler la régularité de $\hat{f}$ s'effectue à travers le choix du paramètre de lissage. Traditionnellement les fonctions solutions sont obtenues par un problème de *moindres carrés pénalisés* :

$$(\hat{\beta}, \hat{f}) = \underset{(\beta,f)}{\arg\min}\Big\{ \sum (y_i - x_i^T\beta - f(t_i))^2 + \lambda \int (f''(t))^2 dt)\Big\}. \tag{3.3}$$

Résoudre (3.3) revient à trouver la solution d'un problème variationnel qui minimise la somme de deux termes ; le premier étant la somme des résidus au carré (RSS) et le deuxième est l'intégrale de la dérivée seconde au carré.
L'application d'un résultat célèbre dû à Schoenberg (Wahba, [41]) permet de dire que les fonctions $\hat{f}$ solutions de (3.3) sont des fonctions splines pénalisées d'ordre 4 avec pour noeuds la suite ordonnée des instants t_{ij}.
Posée sous sa forme la plus générale, l'estimation de (β, f) peut aussi être traitée par l'approche de la vraisemblance pénalisée qui consiste à rechercher $(\hat{\beta}, \hat{f}_\lambda)$ vérifiant :

$$(\hat{\beta}, \hat{f}_\lambda) = \underset{(\beta,f)}{\arg\min}\{L(\beta, f; y) + \lambda \int (f''(t))^2 dt)\}, \tag{3.4}$$

où λ est un paramètre réel fixé > 0. De plus, dans le cas gaussien :

$$L(\beta, f; y) = -\frac{1}{2}\{\log|V| + (y - X\beta - f)^T V^{-1}(y - X\beta - f) + N\log(2\pi)\}. \tag{3.5}$$

Le paramètre λ reste le même que précédemment et en faisant varier celui-ci, on peut contrôler le niveau de régularité de la fonction et obtenir le meilleur compromis entre le biais et la variance comme nous l'avons précisé précédemment. Cela introduit aussi le choix du paramètre de lissage dans la procédure d'estimation. Les méthodes d'approche de ce paramètre de lissage dans le contexte des données qui nous intéressent fera l'objet de la seconde partie de ce chapitre.
Il existe une formulation de (3.3) en problème sous contraintes (Green & Silvermann, [17]) permettant de l'approcher comme un problème avec un multiplicateur de Lagrange dont la solution est le point selle $(\hat{\beta}, \hat{f}, \hat{\lambda})$.
La méthode de logvraisemblance pénalisée présente l'avantage de permettre l'approche de certains paramètres ou spécifiquement le paramètre de lissage par les méthodes du maximum de vraisemblance (ML) et autres méthodes dérivées.

3.3 Présentation du modèle mixte semiparamétrique stochastique

Les techniques nonparamétriques ont été introduites pour répondre au besoin des méthodes avec des hypothèses plus générales pouvant résister aux «caprices» de nos données, c'est-à-dire des méthodes robustes. L'utilisation des fonctions nonparamétriques dans les méthodes

de régression est fortement liée à la robustesse de l'estimateur vis-à-vis des données.
Pour les méthodes nonparamétriques de régression, le modèle se construit en utilisant des fonctions d'approximation. Cette approche se base sur l'hypothèse qu'une ou plusieurs composantes du modèle est une fonction d'une variable explicative connue. Cette fonction étant inconnue, il s'agit d'utiliser des fonctions d'approximation convenables tout en exploitant certaines de ses propriétés comme les conditions de régularité pour l'approcher le plus près possible. Pour cela il existe plusieurs bases fonctionnelles d'approximation. Les éléments de ces bases d'approximation peuvent être de différents types de fonctions avec des propriétés spécifiques. Leur choix est souvent déterminé par la nature des données à traiter.
Cependant les modèles paramétriques présentent l'avantage d'être facilement interprétables grâce à la formulation théorique du modèle. Il paraît donc intéressant d'avoir un gain en robustesse dans nos estimations tout en conservant cette facilité d'interprétation du modèle mixte paramétrique. Il faut alors s'engager sur la voie qui consiste à retenir des procédures statistiques réalisant deux objectifs.
Tout d'abord concernant la facilité d'interprétation du modèle retenu, le modèle paramétrique mixte répond à cette condition. Ses différentes composantes de variance permettent une approche statistique adaptée à l'approche qualitative qui a permis leur formulation. De même la décomposition de la moyenne selon les différents individus (stations) permet une approche intégrée de données *sparses* avec une construction des profils individuels.
Enfin le second objectif est celui d'un modèle permettant une approche robuste, avec des propriétés statistiques restant stables par rapport aux différentes observations disponibles, avec des hypothèses plus larges (faciles à vérifier pour les données). Il faut rechercher un modèle construit à partir des données, autrement dit "laisser les données dicter le modèle".
Les modèles mixtes semipramétriques stochastiques permettent de concilier ces deux objectifs pour les modèles mixtes, cette composante nonparamétrique paraît naturellement comme une composante de la partie à effets fixes du modèle en particulier pour le cas des données longitudinales.

3.3.1 Présentation du modèle

De manière générale, considérons un échantillon longitudinal écrit formellement sous la forme : $\mathcal{X} = \{y_{ij}, t_{ij}\} \cup \{X_{ij}\} \quad i = 1 \cdots m \quad j = 1 \cdots n_i$. où les y_{ij} sont des mesures aux instants t_{ij} d'un processus physique continu $Y_i(t) \in \mathbb{R}^m$ observé sur m sites de mesure ($i = 1$ à m). Le nombre total d'observations est $N = \sum_i n_i$. Les instants t_{ij} sont quelconques dans un intervalle fixe $\mathcal{T} = (a, b)$.
Les mesures y_{ij} sont faites de manière asynchrone et déséquilibrée entre les sites de mesure, mais nous supposerons que les éventuelles covariables $X(t) \in \mathbb{R}^p$ sont mesurées (ou disponibles) sur tous les instants t_{ij} et qu'elles ne sont pas spécifiques aux sites de mesure. Le lecteur notera bien le rôle de chaque indice :

$$\begin{cases} i = & \text{indice spatial } (i = 1, \cdots, m), \\ j = & \text{indice temporel } (j = 1, \cdots, n_i). \end{cases}$$

Selon le formalisme usuel des modèles mixtes, nous décrivons le modèle par l'équation :

$$y_{ij} = f(t_{ij}) + X(t_{ij})^T \beta + Z_{ij}^T b_i + U_i(t_{ij}) + \varepsilon_{ij}, \tag{3.6}$$

avec les spécifications suivantes :

(i) $f(t)$ est une fonction du temps ;
(ii) $b = (b_i, \ldots, b_m) \sim \mathcal{N}(0, \sigma_*^2 Z G Z^T)$;
(iii) $\varepsilon = (\varepsilon_{11}, \cdots, \varepsilon_{mn_m})^T \sim \mathcal{N}(0, \sigma_0^2 R)$;

(iv) $U_i(t)$ est un processus gaussien centré de structure de covariance :

$$\Gamma_i(t, s) = \mathrm{cov}(U_i(t), U_i(s))$$

(v) Toutes les composantes aléatoires du modèles sont mutuellement indépendantes.
Comme déjà notifié dans le chapitre 2, plusieurs choix sont bien sûr possibles pour les fonctions $\Gamma_i(t, s)$; cependant nous privilégierons le cas exponentiel :

$$\Gamma_i(t, s) = \gamma_i^2 \exp(-\alpha_i |t - s|).$$

Du point de vue de l'interprétation, ce modèle revient à dire que l'espérance conditionnelle de $y_i(t)$ est de forme semiparamétrique $f(t) + X(t)^T \beta$ où $f(t)$ est une composante non-paramétrique et $X(t)^T \beta$ un prédicteur linéaire construit à partir des variables exogènes $X(t)$. Quant aux composantes de la variance $Z_{ij}^T b_i + U_i(t_{ij}) + \varepsilon_{ij}$, elles décrivent les corrélations spatio-temporelles des données : plus précisément, le choix a été fait de décrire les corrélations spatiales par un vecteur de paramètres aléatoires $b = (b_1^T, \cdots, b_m^T)^T \in \mathbb{R}^q$, et de décrire les corrélations temporelles par des paramètres fixes γ_i^2, α_i au travers d'un processus gaussien $U_i(t)$. Les composantes spécifiques $U_i(t)$ propres à chaque sujet contribueront à la prédiction des trajectoires individuelles. Dans certains cas, en particulier pour des données environnementales provenant de différents sites de mesure, ces composantes s'ajoutent au paramètres $b = (b_1^T, \cdots, b_m^T)^T$ pour donner des profils représentant des courbes de données réelles interprétables. Comme dans l chpitre 2, l'ensemble des paramètres de la variance est noté $\theta = (\sigma_0^2, \sigma_*^2, \gamma^{2T}, \alpha^T)^T$.

3.3.2 La matrice d'incidence

Ceci nous amène à l'introduction de la matrice d'incidence (voir Green & Silvermann, [17]) qui permet de ne faire l'estimation du paramètre fonctionnel qu'à partir des instants d' observation "valides" pour les différents sites .
Sa construction nécessite la définition du vecteur t_r^0 des instants distincts ordonnés de tous les points t_{ij} d'observation sur l'ensemble des stations. Alors la matrice d'incidence N_i de la i^{eme} station $(i = 1, \ldots, m)$ est la matrice ayant r colonnes référencées par les éléments du vecteur t_r^0 et en lignes les n_i instants d'observation propres à cette i^{eme} station avec le jk^{eme}

élement étant $I(t_{ij} = (t_r^0)_k)$,$j = 1, \ldots, n_i$ et $k = 1, \ldots, r$ et où I est la fonction indicatrice soit :

$$N_{ij,k} = \begin{cases} 1 & \text{si } t_{ij} = (t_r^0)_k, \\ 0 & \text{ailleurs.} \end{cases}$$

La matrice d'incidence permet d'établir cette relation entre les vecteurs des instants de mesure pour une station donnée et le vecteur t_r^0. Il faut noter que l'estimation de la fonction $f(t)$ est effectuée aux instants t_r^0, la matrice d'incidence N_i permet l'adéquation de $\hat{f}_i(t)$ aux instants de mesure de la station i. La matrice N sera l'empilement de ces m matrices d'incidence les unes sur les autres. Le modèle semiparamétrique pour le site i devient simplement $Y_i = X\beta + Z_i b + U_i + N_i f + \epsilon$.

3.3.3 Estimation des composantes du modèle mixte sochastique semiparamétrique

Estimation de la composante nonparamétrique $f(t)$ par régression spline pénalisée

Pour l'estimation de la composante nonparamétrique $f(t)$, diverses approches sont envisageables.

Zeger et Diggle [46] (1994) furent parmi les premiers à proposer un modèle mixte sémiparamétrique stochastique pour données longitudinales où la composante commune est approchée par une méthode nonparamétrique. En particulier, ils utilisèrent la méthode des noyaux. L'estimation du paramètre β se fait par GLS. La procédure générale est itérative et est une version de l'algorithme de Backfittinng de Hastie & Tibshirani.

Le modèle (3.6) est plus proche de celui proposé par Zhang et *al* [47] (1998) qui proposèrent une approche globale pour ce dernier. Dans sa procédure d'estimation, Zhang a approché la fonction nonparamétrique comme modèle à effets mixtes à part entière afin de l'intégrer dans le modèle (global).

Jacqmin-Gadda et *al* [21] (2002) utilisent le modèle (3.6) où la fonction nonparamétrique est approchée dans une base de M-splines cubiques avec un nombre réduit de noeuds et un choix du paramètre de lissage par des critères de sélection dont des formes modifiées du CV et d'AIC.

Pour notre part, nous nous plaçons d'emblée dans un espace de représentation de type splines cubiques avec un nombre réduit de noeuds, ce qui conduit à remplacer le paramètre fonctionnel infinidimensionnel f par une approximation en dimension finie $\tilde{f}(t) = \sum_j a_j B_{j,4}(t)$. La méthode décrite fait une synthèse des deux précédentes.

Dans notre procédure d'estimation, la fonction nonparamétrique est spécifiquement considérée comme un modèle à effets aléatoires n'entraînant pas de ce fait une modification de covariables X et du paramètre β. Ce qui n'est pas le cas de l'approche de Zhang [47]. Cette considération simplifie la procédure d'estimation.

La fonction de vraisemblance pénalisée que nous considérons est de la forme :

$$L(\beta, a, \theta, \lambda; y) = -\frac{1}{2}\log|V_\theta| - \frac{1}{2}\|y - X\beta - NBa\|^2_{V_\theta{}^{-1}} - \lambda a^T \Omega a, \tag{3.7}$$

qu'il s'agit de maximiser par rapport à l'ensemble des paramètres (β, θ) et par rapport aux coefficients splines a intervenant dans l'expression $\tilde{f} = Ba$ du vecteur des valeurs $f(t_{ij})$.
La structure de covariance vaut :

$$V_\theta = \sigma_0^2 R + ZG(\sigma_*^2)Z^T + \Gamma(\gamma^2, \alpha). \tag{3.8}$$

Mais en considérant les nouveaux paramètres a comme aléatoires $a \sim \mathcal{N}(0, (2\lambda)^{-1}\Omega^{-1})$ de sorte que le coefficient de lissage λ soit intégré comme paramètre de variance dans le modèle sous la forme $\tau^2 = 1/2\lambda$. La logvraisemblance pénalisée se réécrit :

$$L_*(\beta, a, \theta, \tau^2; y) = -\frac{1}{2}\log|V_*| - \frac{1}{2}\|y - X\beta\|^2_{V_*^{-1}}, \tag{3.9}$$

avec pour matrice de covariance théorique

$$V_* = \tau^2 NB\Omega^{-1}B^T N^T + \sigma_0^2 R + ZG(\sigma_*^2)Z^T + \Gamma(\gamma, \alpha), \tag{3.10}$$

correspondant formellement au modèle mixte :

$$y = X\beta + NBa + Zb + U + \varepsilon, \tag{3.11}$$

dont les effets aléatoires sont donnés par : $NBa + Zb + U + \varepsilon$.
Ces deux dernières équations sont la description exacte d'un modèle linéaire mixte étendu, appelé Linear Modified Mixed Model (L3M) dans la suite. L'ensemble des paramètres de variance sera noté $\theta_* = (\tau^2, \sigma_0^2, \sigma_*^{2T}, \gamma^{2T}, \alpha^T)^T$. Nous proposons à présent une résolution de ce modèle étendu, selon les trois étapes usuelles de la méthode d'estimation REML des modèles mixtes.

Estimation REML pour les paramètres de la variance θ_*
La marginalisation sur les paramètres de la moyenne de la log-vraisemblance (gaussienne) du modèle étendu (3.9) permet d'écrire un "profil" de vraisemblance pour les paramètres de variance (voir la section 2.4.2) :

$$L_{REML}(\theta_*; y, \hat{\beta}) = -\frac{1}{2}\log|V_*| - \frac{1}{2}\log|X^T V_*^{-1} X| - \frac{1}{2}\|y - X\hat{\beta}\|^2_{V_*^{-1}}.$$

La maximisation de cette dernière est réalisée par un algorithme Fisher-Scoring spécifique, fondé sur la résolution des équations REML découlant de la procédure en considérant tous les paramètres de la variance θ_*. En particulier avec le paramètre de lissage intégrée comme nouveau paramètre τ^2. Ces estimations REML de $\hat{\theta}_*$ sont obtenues en cherchant les valeurs qui annulent l'expression ci-dessous :

$$-\frac{1}{2}\mathrm{tr}(P_*\frac{\partial V_*}{\partial \theta_k}) + \frac{1}{2}y^T P_*^T \frac{\partial V_*}{\partial \theta_k} P_* y,$$

avec $P_* = V_*^{-1} - V_*^{-1}X(X^T V_*^{-1} X)^{-1} X^T V_*^{-1}$.
Equations de Henderson Réduites pour β et a
En pratique, il est trop coûteux, sur un plan de mise en oeuvre pratique, de maximiser (3.9) par rapport à β et $b_* = (a^T, b^T, U^T)^T$ à partir des équations normales standard, en raison de la dimension excessive de b_* et de la structure non diagonale de V_* (cf. [47]). A partir de l'égalité $V_*^{-1}(y - X\hat{\beta}) = V^{-1}(y - X\hat{\beta} - NB\hat{a})$ découlant des travaux sur les méthodes du maximum de vraisemblance de Harville, [19] il est aisé de montrer que, les paramètres de variance étant fixés, une estimation BLUE de β et a est obtenue par résolution d'un système réduit d'équations :

$$\begin{bmatrix} X^T V^{-1} X & X^T V^{-1} NB \\ B^T N^T V^{-1} X & B^T N^T V^{-1} BN + \tau^{-2}\Omega \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T V^{-1} y \\ B^T N^T V^{-1} y \end{bmatrix}. \tag{3.12}$$

.
Prédiction BLUP pour les effets aléatoires b et $U_i(t)$
Pour terminer sur les aspects inférentielles, nous donnons les prédictions BLUP des effets aléatoires b_i et des processus aléatoires $U_i(t)$ qui sont des extensions des applications de la partie (2.4.3) :

$$\hat{b} = \hat{G} Z^T \hat{V}_*^{-1}(y - X\hat{\beta}). \quad \hat{U}_i(t) = \hat{\Gamma}_i(t, t_i)\hat{V}_i^{-1}(Y_i - X_i\hat{\beta} - \hat{f}_i).$$

3.3.4 Biais et variance d'estimation conditionnellement à f

En premier lieu dans cette section, nous développons les calculs d'espérance "conditionnelle" des différents estimateurs et prédicteurs construits en fonction de la composante f inconnue. En effet, dans une situation expérimentale donnée (f inconnue mais fixée), on doit s'attendre à une valeur espérée d'estimation qui dépende de f mais aussi du plan d'échantillonnage dicté par la matrice d'incidence N.

En notant f le vecteur des valeurs $f(t_{ij})$ correspondant à l'échantillonnage de f aux instants $t_{ij}\, j = 1, \cdots, n_i\, i = 1\, cdots, m$, nous pouvons exprimer l'espérance "conditionnelle" de $\hat{\beta}$ et de $\hat{a}$ comme suit (les calculs sont analogues à ceux trouvés en [47]) :

$$\mathbb{E}[\hat{\beta}|f] = \beta + (X^T W_x X)^{-1} X^T W_x \mathrm{f}$$

avec $W_x = V^{-1} - V^{-1}NB(B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} B^T N^T V^{-1}$.

Il en va de même pour l'estimateur de a pour lequel nous obtenons :

$$\mathbb{E}[\hat{a}|f] = (B^T N^T W_f NB + \tau^{-2}\Omega)^{-1} B^T N^T W_f \mathrm{f}$$

avec $W_f = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}$.

A partir de $\mathbb{E}[\hat{a}|f]$ ci-dessus nous pouvons déduire l'expression de l'espérance "conditionnelle" de $\hat{f}(t) = \sum_j \hat{a}_j B_{j,4}(t)$, et l'écrire :

$$\mathbb{E}[\hat{f}(t)|f] = \sum_j \mathbb{E}[\hat{a}_j|f] B_{j,4}(t)$$

Cette expression sera utilisée dans le calcul du ISB "conditionnel" de $\hat{f}$:

$$\int_{\mathcal{T}} \Big[\mathbb{E}[\hat{f}(u)|f] - f(u)\Big]^2 du$$

En ce qui concerne les effets aléatoires prédits $\hat{b}$, l'expression retenue pour leur espérance "conditionnelle" est :

$$\mathbb{E}[\hat{b}|f] = GZ^T V^{-1}(\beta - \mathbb{E}[\hat{\beta}|f])$$

En deuxième lieu, il convient d'évoquer la question des variances d'estimation pour les différents paramètres fixes du modèle. Nous avons pour $\hat{\beta}$:

$$\mathrm{Var}[\hat{\beta}] = (X^T W_x X)^{-1}(X^T W_x V W_x X)(X^T W_x X)^{-1}$$

Quant aux différents paramètres de variance contenus dans θ_*, nous utiliserons les formules asymptotiques usuelles basées sur l'inverse l'information de Fisher, dont le terme général est :

$$I^*_{kl} = \frac{1}{2}\mathrm{tr}(P_* \frac{\partial V_*}{\partial \theta_{*k}} P_* \frac{\partial V_*}{\partial \theta_{*l}})$$

3.3.5 La validation croisée (CV) et la validation croisée généralisée (GCV) dans le modèle mixte semipramétrique stochastique

Le rôle joué par le paramètre de lissage λ dans la modélisation de la fonction nonparamétrique par les splines de régression pénalisées (ordre 4) a été noté dans l'exposé introduisant la logvraisemblance. Dans la procédure d'estimation exposée, nous tentons d'approcher le paramètre de lissage λ optimal par la méthode basée sur la logvaisemblance REML. Certains auteurs préfèrent estimer celui-ci par la méthode de la validation croisée (CV) ou la méthode de validation croisée généralisée (GCV). Ces deux méthodes sont les méthodes les plus répandues du choix du paramètre de lissage à partir des données. (voir Green & Silvermann [17], Hastie & Tibshirani [20], Chong Gu [18]). Dans cette partie, nous exposons ces deux approches adaptées aux données environnementales avec pour intérêt l'extraction d'une composante commune. Cela permettra de donner les raisons du choix d'estimer le paramètre de lissage par la méthode d'estimation REML.

D'autre part les méthodes (CV) et (GCV) sont des méthodes classiques basées sur la performance de l'estimateur à travers une fonction de perte. Le principe de la validation croisée repose sur l'erreur prédictive d'échantillonnage, c'est une forme particulière de l'échantillon test (échantillon réduit à une observation).
Pour faciliter l'exposé, nous notons $\tilde{Y}_i = Y_i - X\hat{\beta} - Z_i\hat{b} - \hat{U}_i = f(t) + \epsilon$. Le principe consiste à retenir le paramètre de lissage λ minimisant : $\frac{1}{N}\sum_i^m \sum_{j=1}^{n_i}(\hat{f}_\lambda(t_{ij}) - \tilde{y}_{ij})^2$. Mais comme très souvent la taille des données disponibles est insuffisante, la validation croisée consiste à choisir le paramètre λ qui minimise la quantité :

$$CV(\lambda) = \frac{1}{N}\sum_{i}^{m}\sum_{j=1}^{n_i}(\hat{f}_\lambda^{[-ij]}(t_{ij}) - \tilde{y}_{ij})^2. \tag{3.13}$$

où $\hat{f}_\lambda^{[-ij]}(t_{ij})$ est la prédiction de la valeur prise par $f(t)$ à l'instant t_{ij}, estimée sur l'ensemble des données sans l'observation $\tilde{y}_{ij}$. Dans ce cas précis, $\hat{f}_\lambda$ correspond à la solution du problème de minimisation ci-dessous :

$$\hat{f}_\lambda^{[-j]} = \underset{(f_\lambda)}{\arg\min}\left\{\frac{1}{N}\sum_{i}^{m}\sum_{k\neq j}(f_\lambda(t_{ik}) - \tilde{y}_{ik})^2 + \lambda f^T\Omega f\right\}. \tag{3.14}$$

Le lecteur notera que pour la prédiction croisée à un instant t_{ij}, (3.14) consiste à se passer de l'information issue des mesures de tous les sites consernés (où il y a eu une mesure à l'instant t_{ij}). Cette précaution n'est pas toujours prise en compte dans l'application de la méthode, pour ce type des données. Il faut ajouter que ces données proviennent de peu d'invidus (stations de mesure) avec un nombre important d'observations par conséquent il serait difficile de se passer des observations provenant d'un individu comme c'est souvent pratiqué dans le cas des données correlées.
Il est aisé alors d'établir que la solution de (3.14) [17, 18] est aussi solution de :

$$\underset{(f_\lambda)}{\arg\min}\left\{\frac{1}{N}\sum_{i}^{m}\sum_{j=1}^{n_i}(f_\lambda(t_{ij}) - \tilde{y}_{ij})^2 + \lambda f^T\Omega f\right\}.$$

L'expression de $CV(\lambda)$ fait évoluer celle de $\hat{f}_\lambda^{[-j]}$ dont la forme explicite n'est pas connue. La pratique ([35], [18]) consiste à utiliser une expression plus simple à évaluer, celle-ci découle de la remarque suivante. Elle consiste à utiliser la notation : $\hat{f}_\lambda = S_\lambda \tilde{Y}$. Cette dernière nous permet d'écrire :

$$\hat{f}_\lambda(t_{ij}) - \hat{f}_\lambda^{[-j]}(t_{ij}) = \sum_{k=1}^{n_i} S_{jk}\tilde{y}_{ik} - \hat{f}_\lambda^{[-j]}(t_{ij}) = \sum_{k=1}^{n_i} S_{jk}\tilde{y}_{ik} - \sum_{k\neq j} S_{jk}\tilde{y}_{ik} - S_{jj}\hat{f}_\lambda^{[-j]}(t_{ij}).$$

ce qui permet d'obtenir aisément que : $\hat{f}_\lambda(t_{ij}) - \hat{f}_\lambda^{[-j]}(t_{ij}) = S_{jj}(\tilde{y}_{ij} - f_\lambda^{[-j]}(t_{ij}))$.
A partir de cette égalité, il vient qu'une expression explicite de $f_\lambda^{[-j]}(t_{ij})$ est :

$$\hat{f}_\lambda^{[-j]}(t_{ij}) = \frac{\hat{f}_\lambda(t_{ij}) - S_{jj}\tilde{y}_{ij}}{1 - S_{jj}}. \tag{3.15}$$

Il s'en suit que :

$$\hat{f}_\lambda^{[-j]}(t_{ij}) - \tilde{y}_{ij} = \frac{\hat{f}_\lambda(t_{ij}) - \tilde{y}_{ij}}{1 - S_{jj}}.$$

D'où une expression équivalente à celle du (3.13)est alors :

$$CV(\lambda) = \frac{1}{N}\sum_{i}^{m}\sum_{j=1}^{n_i} \frac{(\hat{f}_\lambda(t_{ij}) - \tilde{y}_{ij})^2}{(1 - S_{jj})^2}. \tag{3.16}$$

Le choix du paramètre λ par validation croisée revient à choisir la valeur de λ qui minimise le critère (3.16). Suivant le plan d'échantillonnage, il n'est pas toujours possible que toutes les observations contribuent de manière uniforme à l'estimation de la fonction nonparamétrique. Cette situation se présente souvent dans le cas des mesures répétées. L'introduction d'une pondération adaptée dans le critère de la validation croisée est une méthode permettant d'uniformiser la contribution des différentes observations à l'estimation de la fonction nonparamétrique (Gu, [18]). Ce point devient considérable quand la fonction nonparamétrique est utilisée pour traduire une composante commune, caractérisant le phénomène étudié au niveau de tous les sites.
La forme pondérée de la validation croisée devient :

$$WCV(\lambda) = \frac{1}{N}\sum_{i}^{m}\sum_{j=1}^{n_i} w_{ij}\frac{(\hat{f}_\lambda(t_{ij}) - \tilde{y}_{ij})^2}{(1 - S_{jj})^2}. \tag{3.17}$$

La méthode du GCV consistant à choisir un poids $w_{ij} = \frac{(1-S_{jj})^2}{\left[\frac{1}{N}tr(I-S_\lambda)\right]^2}$ est souvent considérée comme une méthode alternative :

$$GCV(\lambda) = \frac{1}{N}\sum_{i}^{m}\sum_{j=1}^{n_i} \frac{(\hat{f}_\lambda(t_{ij}) - \tilde{y}_{ij})^2}{(1 - \frac{tr(S_\lambda)}{N})^2}. \tag{3.18}$$

En notant que la fonction $f(t)$ est estimée aux r instants distincts le poids aurait pu être alors de la forme $w_{ij} = \frac{(1-S_{jj})^2}{\left[\frac{1}{r}tr(I-S_\lambda)\right]^2}$. Il n'est pas alors assuré que la pondération retenue dans la formulation du (GCV) (3.18) conduise à l'uniformisation de la contribution des différentes observations à l'estimation de $f(t)$. Pour cette raison, il nous apparaît convenable d'utiliser la méthode du REML pour l'estimation du paramètre de lissage.

3.4 Simulations et performances comparées pour le modèle mixte stochastique semiparamétrique

3.4.1 Introduction

De nombreux résultats sont déjà établis dans la littérature concernant la comparaison des performances des méthodes d'estimation du paramètre de lissage dans le cas de données indépendantes. Ainsi Whaba & Wold [43]) puis Craven & Wahba [41] furent les pemiers à établir des résultats pour la méthode (GCV) dans le cas de données indépendantes asynchrones. La comparaison avec la méthode (ML) fut introduite par Ansley & Wecker ; cette comparaison fut ensuite étendue à l'approche (REML) par Wahba [42] et Ansley & Kohn [25].
Dans le domaine des données corrélées, les comparaisons des performances du (CV), (GCV) et (REML) sont peu nombreuses. Cependant il ressort des travaux de Lin & Zhang (1998) que les approches (CV) et (GCV) y sont peu performantes.
Pour ce qui nous concerne, nous avons souhaité évaluer les performances de l'approche (REML) dans le cadre du modèle mixte semiparamétrique stochastique, en considérant deux possibilités pour la matrice de pénalité Ω dans la fonction de vraisemblance pénalisée (3.7) :

1er cas : Ω=ID. La matrice Ω est choisie égale à l'identité (ID). Ce type de pénalisation peut être vu comme une version de la pénalisation partielle (appliquée sur une partie des coefficients de la régression) de Ruppert sur les coefficents de la régression spline dans une base de splines tronquées [35]. Nous étendons cette pénalisation sur tous les coefficients ce qui correspond à des contraintes de pénalisation plus fortes. Cette situation correspond à une résolution de type régression *Ridge* sur les coefficients spline a.

2eme cas : Ω=CISD. La pénalisation CISD (*Cross Integrated Second Derivative*) consiste à prendre pour Ω la matrice avec pour terme général (i, j) :

$$\Omega_{ij} = \int B_i''(t)B_j''(t)dt$$

Dans les deux cas, les performances de l'approche REML seront évaluées sur les critères usuels en estimation semi-paramétrique :
- biais et variance d'estimation des paramètres de variance $\hat{\theta}_*$
- biais et variance d'estimation des paramètres fixes $\hat{\beta}$
- MSE des composantes aléatoires prédites $\hat{b}$ et $\hat{U}$

– ISB et MISE de la composante fonctionnelle $\hat{f}(t)$

De plus, pour mesurer la capacité de l'approche REML à contrôler correctement la régularité de la composante fonctionnelle $f(t)$ estimée, nous calculerons lors de chaque essai Monte-Carlo le ISSD (*Integrated Square Second Derivative*) pour chaque fonction $\hat{f}(t)$ obtenue. Par définition le ISSD d'une fonction $f(t)$ vaut :

$$\lambda(f) = \int f''^2(t)dt$$

3.4.2 Description du modèle simulé

Les données de simulation que nous avons générées reproduisent un exemple virtuel (mais réaliste) à quatre sites de mesures ($m = 4$) avec la présence de deux covariables $X_1(t)$ et $X_2(t)$ égales respectivement à :

$$X_1(t) = 0.1 * t^2 \qquad \text{et} \qquad X_2(t) = \sin(1.2 * t - 2.0)$$

et évaluée sur l'intervalle de temps $\mathcal{T} = [-10\ ;\ 10]$. Ces deux fonctions ont ensuite été combinées linéairement pour former la composante "$X\beta$" du modèle :

$$0.1 * X_1(t) + 0.5 * X_2(t)$$

avec respectivement $\beta_1 = 0.1$ et $\beta_2 = 0.5$

La composante non-paramétrique $f(t)$ que nous avons choisie est construite sur une base fonctions splines cubiques avec $K = 17$ noeuds intérieurs :

$$(-8, -7, -6, -5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5, +6, +7, +8)$$

et $(-10, +10)$ pour noeuds extérieurs. Les 21 coefficients splines définissant exactement la fonction $f(t)$ ont été générés de manière aléatoire selon une loi $\mathcal{N}(0; \tau^2)$ avec $\tau^2 = 0.5$, sauf les deux coefficients extrêmes qui ont été fixés à zéro pour garantir une bonne régularité de $f(t)$ sur les bords du domaine d'étude. Pour information, les réalisations des coefficients spline retenus sont :

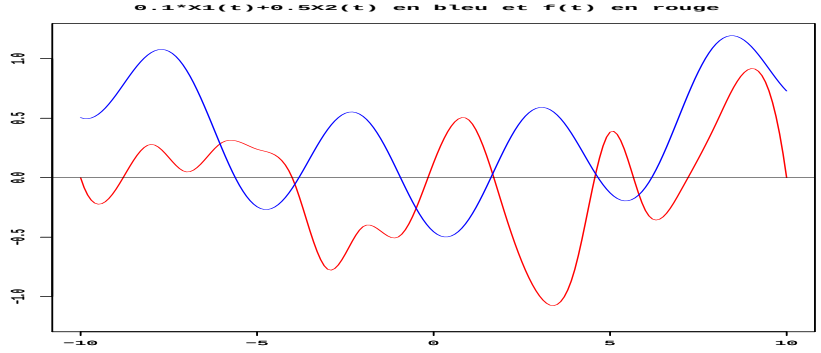
0.0	−0.6313	0.7275	−0.1628	0.4349	0.1883	0.2439
−1.1842	−0.1309	−0.7621	0.1873	0.7639	−0.3340	−1.1254
−1.1300	1.0184	−0.6412	−0.1097	0.5980	1.4222	0.0

La régularité de cette fonction $f(t)$ est mesurée par son ISSD que nous avons évalué par calcul à :

$$\lambda_0 = \lambda(f) = \int_{\mathcal{T}} f"^2(t)dt = 28.210$$

Le graphique ci-après nous donne l'allure comparée des deux composantes temporelles $\beta_1 X_1(t) + \beta_2 X_2(t)$ d'une part (en bleu) et $f(t)$ d'autre part (en rouge).

FIG. 3.4 – Composantes temporelles du modèle simulé : $X\beta$ en bleu et $f(t)$ en rouge



Quant aux composantes aléatoires b, $U_i(t)$ $i = 1, .., 4$ et ε du modèle, elles varient bien sûr lors de chaque simulation ; leur structure de covariance ont été prises égales à :

$$\text{var}\,(b) = G = \sigma_*^2 I_m = 4.0 * I_4$$

$$\text{cov}\,(U_i(t), U_i(t')) = \gamma^2 \exp[-\alpha * |t - t'|] = 1.0 \exp[-0.5 * |t - t'|]$$

et

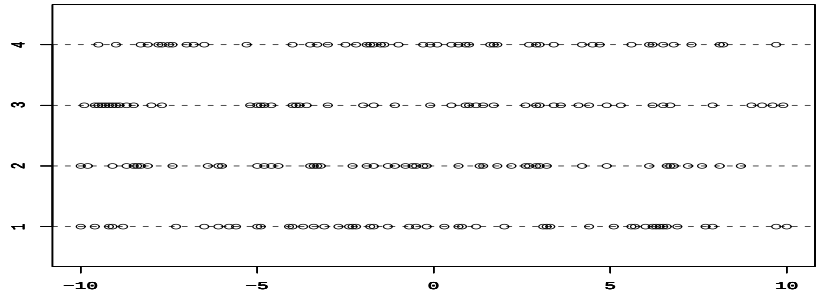
$$\text{var}\,(\varepsilon) = \sigma^2 I_n = 0.1 * I_{200}$$

En résumé, les divers paramètres fixés pour les simulations sont :

β_1	β_2	σ_*^2=var(b_i)	τ^2=var(a_k)	γ^2=var($U_i(t_{ij})$)	σ^2=var(ε_{ij})	α
0.1	0.5	4.0	0.5	1.0	0.1	0.5

Pour finir, nous devons décrire le plan d'échantillonnage choisi sur les quatre sites de mesure : pour chaque site $i = 1$ à 4, le plan d'échantillonnage correspond à une liste de 50 instants t_{ij} selon un tirage aléatoire (sans remise) de 50 parmi 200 valeurs possibles équiréparties sur l'intervalle $\mathcal{T}$. La situation se résume à l'aide de la représentation graphique suivante :

FIG. 3.5 – Plan d'échantillonnage sur les quatre sites



3.4.3 Performances théoriques attendues pour Ω=ID puis Ω= CISD

Nous calculons dans cette section les performances théoriques attendues sur les estimateurs REML $\hat{\theta}_* = (\hat{\tau}^2, \hat{\sigma}_0^2, \hat{\sigma}_*^2, \hat{\gamma}_2, \hat{\alpha})^T$ puis $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$; ensuite sur les prédictions $\hat{b} = (\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4)^T$ des effets aléatoires ; et enfin sur la composante commune estimée $\hat{f}(t) = \sum_k \hat{a}_k B_{4,k}(t)$.
Les calculs sont bien sûr effectués conditionnellement à la fonction $f(t)$ choisie, et au plan d'échantillonnage N choisi ; ce qui correspond aux calculs développés dans la section (3.3.4).
Les paramètres vrais sont fixés aux valeurs de simulation décrites dans la section précédente, seul le paramètre de lissage τ^2 est variable, allant d'une situation fortement pénalisée ($\tau^2 = 0.01$) à une situation très peu pénalisée ($\tau^2 = 100$).
Concernant le nombre et la position des nœuds pour la construction des splines, deux situations seront envisagées :

(a) v =20 nœuds intérieurs placés aux fractiles $k/21$ avec $k = 1, \cdots, 20$;
(b) v =100 nœuds intérieurs placés aux fractiles $k/101$ avec $k = 1, \cdots, 100$;

Les quatre tableaux qui suivent regroupent l'ensemble des calculs réalisés :
tableau (3.1) : matrice de pénalité Ω =ID et nombre de nœuds v =20
tableau (3.2) : matrice de pénalité Ω =ID et nombre de nœuds v =100
tableau (3.3) : matrice de pénalité Ω =CISD et nombre de nœuds v =20
tableau (3.4) : matrice de pénalité Ω =CISD et nombre de nœuds v =100

TAB. 3.1 – Performances théoriques attendues pour une matrice de pénaité Ω=ID et un choix de 20 noeuds

	$\tau^2 = 0.01$	$\tau^2 = 0.1$	$\tau^2 = 0.5$	$\tau^2 = 1$	$\tau^2 = 10$	$\tau^2 = 100$
$E[\hat{\beta}_1 \mid f]$	0.14061	0.14197	0.14363	0.14326	0.13076	0.10991
$E[\hat{\beta}_2 \mid f]$	0.29506	0.30174	0.31539	0.32082	0.28984	-0.10656
$E[\hat{b}_1 \mid f]$	-0.22882	-0.22738	-0.21993	-0.21050	-0.12050	-0.02273
$E[\hat{b}_2 \mid f]$	-0.16829	-0.17542	-0.18542	-0.18418	-0.10807	-0.01352
$E[\hat{b}_3 \mid f]$	-0.20668	-0.20760	-0.20436	-0.19714	-0.11267	-0.01672
$E[\hat{b}_4 \mid f]$	-0.19653	-0.19963	-0.20234	-0.19857	-0.12096	-0.02538
$V[\hat{\sigma}_b^2 \mid f]$	4.63139	4.68136	4.80236	4.90545	6.08337	7.01072
$V[\hat{\sigma}_a^2 \mid f]$	0.01436	0.02179	0.07049	0.16547	7.70663	172.84769
$V[\hat{\sigma}_U^2 \mid f]$	0.06740	0.06951	0.07407	0.07600	0.07741	0.07685
$V[\hat{\sigma}_\varepsilon^2 \mid f]$	0.00092	0.00092	0.00093	0.00093	0.00093	0.00093
$V[\hat{\alpha} \mid f]$	0.02624	0.02719	0.02932	0.03028	0.03108	0.03103
$V[\hat{\beta}_1 \mid f]$	0.00273	0.00274	0.00279	0.00290	0.00686	0.01792
$V[\hat{\beta}_2 \mid f]$	0.01765	0.01768	0.01797	0.01839	0.02412	0.31862
$MSE[\hat{b}_1 \mid f]$	0.21248	0.21260	0.21458	0.21929	0.41155	0.94859
$MSE[\hat{b}_2 \mid f]$	0.21207	0.21223	0.21445	0.21937	0.41279	0.95225
$MSE[\hat{b}_3 \mid f]$	0.21409	0.21421	0.21608	0.22061	0.41196	0.94863
$MSE[\hat{b}_4 \mid f]$	0.21180	0.21192	0.21377	0.21830	0.41022	0.94870
$ISB[\hat{f} \mid f]$	4.18678	3.14775	1.69656	1.24597	0.70115	3.65837

TAB. 3.2 – Performances théoriques attendues pour une matrice de pénaité Ω=ID et un choix de 100 noeuds

	$\tau^2 = 0.01$	$\tau^2 = 0.1$	$\tau^2 = 0.5$	$\tau^2 = 1$	$\tau^2 = 10$	$\tau^2 = 100$
$E[\hat{\beta}_1\|f]$	0.14065	0.14237	0.14577	0.14721	0.14434	0.12446
$E[\hat{\beta}_2\|f]$	0.29376	0.29167	0.28976	0.28985	0.29150	0.28907
$E[\hat{b}_1\|f]$	-0.22975	-0.23463	-0.24139	-0.24232	-0.21058	-0.08643
$E[\hat{b}_2\|f]$	-0.16832	-0.17674	-0.19801	-0.20931	-0.20318	-0.08540
$E[\hat{b}_3\|f]$	-0.20719	-0.21255	-0.22456	-0.22971	-0.20692	-0.08513
$E[b_4\|f]$	-0.19661	-0.20099	-0.21253	-0.21903	-0.20562	-0.08602
$V[\hat{\sigma}_b^2\|f]$	4.63436	4.67582	4.57675	4.44407	3.92597	4.20324
$V[\hat{\sigma}_a^2\|f]$	0.00270	0.00482	0.01868	0.04558	2.06956	162.97506
$V[\hat{\sigma}_U^2\|f]$	0.06445	0.06576	0.06705	0.06721	0.06266	0.05719
$V[\hat{\sigma}_\varepsilon^2\|f]$	0.00093	0.00104	0.00126	0.00139	0.00169	0.00175
$V[\hat{\alpha}\|f]$	0.02690	0.02816	0.02976	0.03016	0.02869	0.02636
$V[\hat{\beta}_1\|f]$	0.00273	0.00274	0.00279	0.00285	0.00375	0.01369
$V[\hat{\beta}_2\|f]$	0.01765	0.01771	0.01805	0.01834	0.01974	0.02278
$MSE[\hat{b}_1\|f]$	0.21248	0.21264	0.21356	0.21461	0.24188	0.61192
$MSE[\hat{b}_2\|f]$	0.21207	0.21222	0.21345	0.21492	0.24390	0.61396
$MSE[\hat{b}_3\|f]$	0.21409	0.21422	0.21520	0.21640	0.24600	0.61832
$MSE[\hat{b}_4\|f]$	0.21180	0.21192	0.21279	0.21382	0.24094	0.61101
$ISB[\hat{f}\|f]$	4.27645	3.72378	2.55175	1.97072	0.94343	0.52079

L'influence de τ^2 porte en premier lieu sur la variance de son propre estimateur $\hat{\tau}^2$ (voir ligne $V[\hat{\sigma}_a^2]$) : plus la valeur de τ^2 est grande, plus la variance de $\hat{\tau}^2$ est élevée. Par ailleurs, nous notons bien sûr que τ^2 agit dans le sens usuel de la pénalisation, à savoir : une faible pénalisation ($\tau^2 = 100$) conduit à des valeurs de biais plutôt faibles et des variances d'estimation plutôt élevées, une forte pénalisation ($\tau^2 = 0.01$) conduit à des valeurs de biais plutôt élevées et des variances d'estimation plutôt faibles.
Concernant la composante fonctionnelle f, en comparant les tableaux (3.1) et (3.2), puis (3.3) et (3.4), nous constatons que la situation à 100 nœuds dégrade sensiblement l'estimation de celle-ci (selon le critère de l'ISB) par rapport à la situation à 20 nœuds. C'est une chose qui s'explique par le fait que, plus on accorde de degrés de liberté pour la construction de $\hat{f}$ (100 nœuds au lieu de 20), plus l'ajustement spline a tendance à interpoler les points observés tout en produisant de grands écarts entre ces derniers. Heureusement, l'effet de pénalisation de τ^2 permet de limiter ce phénomène, et l'on constate en général une valeur optimale de τ^2 pour laquelle l'ISB atteint une valeur minimale.

TAB. 3.3 – Performances théoriques attendues pour une matrice de pénaité Ω=CISD et un choix de 20 noeuds

	$\tau^2 = 0.001$	$\tau^2 = 0.01$	$\tau^2 = 0.1$	$\tau^2 = 1$	$\tau^2 = 10$
$E[\hat{\beta}_1\|f]$	0.14119	0.12223	0.02589	-0.15020	-0.29756
$E[\hat{\beta}_2\|f]$	0.28530	0.27701	0.25865	0.24625	0.20665
$E[\hat{b}_1\|f]$	-0.03066	-0.03062	-0.02420	-0.00715	0.00119
$E[\hat{b}_2\|f]$	0.04335	0.04691	0.04292	0.01959	0.00493
$E[\hat{b}_3\|f]$	-0.01282	-0.01337	-0.00945	-0.00243	-0.00033
$E[b_4\|f]$	0.00013	-0.00292	-0.00927	-0.01001	-0.00578
$V[\hat{\sigma}_b^2\|f]$	4.61604	4.40152	3.78307	3.52567	3.73634
$V[\hat{\sigma}_a^2\|f]$	0.00009	0.00021	0.00328	0.15301	8.15089
$V[\hat{\sigma}_U^2\|f]$	0.06783	0.06809	0.06958	0.07039	0.07077
$V[\hat{\sigma}_e^2\|f]$	0.00089	0.00089	0.00090	0.00092	0.00093
$V[\hat{\alpha}\|f]$	0.02661	0.02664	0.02705	0.02776	0.02830
$V[\hat{\beta}_1\|f]$	0.00287	0.00330	0.00870	0.03462	0.21379
$V[\hat{\beta}_2\|f]$	0.01778	0.01783	0.01810	0.01923	0.02933
$MSE[\hat{b}_1\|f]$	1.12623	1.12623	1.12627	1.12644	1.12676
$MSE[\hat{b}_2\|f]$	1.13080	1.13084	1.13110	1.13156	1.13204
$MSE[\hat{b}_3\|f]$	1.12707	1.12708	1.12715	1.12734	1.12760
$MSE[\hat{b}_4\|f]$	1.12786	1.12787	1.12790	1.12799	1.12810
$ISB[\hat{f}\|f]$	4.71073	3.42636	3.60841	25.72823	64.98754

TAB. 3.4 – Performances théoriques attendues pour une matrice de pénaité Ω=CISD et un choix de 100 noeuds

	$\tau^2 = 0.001$	$\tau^2 = 0.01$	$\tau^2 = 0.1$	$\tau^2 = 1$	$\tau^2 = 10$
$E[\hat{\beta}_1\|f]$	0.14119	0.12224	0.02612	-0.14776	-0.23955
$E[\hat{\beta}_2\|f]$	0.28530	0.27701	0.25869	0.24970	0.26083
$E[\hat{b}_1\|f]$	-0.03063	-0.03063	-0.02430	-0.00735	0.00277
$E[\hat{b}_2\|f]$	0.04338	0.04691	0.04300	0.01982	0.00222
$E[\hat{b}_3\|f]$	-0.01279	-0.01337	-0.00950	-0.00299	-0.00244
$E[\hat{b}_4\|f]$	0.00016	-0.00291	-0.00921	-0.00947	-0.00255
$V[\hat{\sigma}_b^2\|f]$	4.60757	4.31848	3.29525	2.91128	2.86916
$V[\hat{\sigma}_a^2\|f]$	0.00008	0.00019	0.00270	0.12982	6.93640
$V[\hat{\sigma}_U^2\|f]$	0.06764	0.06844	0.07121	0.07286	0.07258
$V[\hat{\sigma}_e^2\|f]$	0.00089	0.00089	0.00090	0.00092	0.00096
$V[\hat{\alpha}\|f]$	0.02654	0.02678	0.02769	0.02892	0.03007
$V[\hat{\beta}_1\|f]$	0.00287	0.00330	0.00869	0.03451	0.17522
$V[\hat{\beta}_2\|f]$	0.01778	0.01783	0.01810	0.01916	0.02417
$MSE[\hat{b}_1\|f]$	1.12642	1.12625	1.12627	1.12644	1.12688
$MSE[\hat{b}_2\|f]$	1.13099	1.13086	1.13110	1.13155	1.13216
$MSE[\hat{b}_3\|f]$	1.12726	1.12710	1.12715	1.12734	1.12764
$MSE[\hat{b}_4\|f]$	1.12805	1.12789	1.12790	1.12799	1.12816
$ISB[\hat{f}\|f]$	4.71079	3.42524	3.58660	25.18434	47.35289

3.4.4 Essais Monte-Carlo à τ^2 fixé égal à 0.5 et Ω=ID

Lors de ces premiers essais Monte-Carlo, le paramètre τ^2 de contrôle de "lissage" est exclu de la procédure d'estimation REML, il est fixé à la vraie valeur 0.5 qui a été utilisée pour simuler la composante fonctionnelle $f(t)$. Pour la reconstruction $\hat{f}(t)$ sous forme de B-spline cubique, nous faisons varier la dimension de l'espace de reconstruction en augmentant progressivement le nombre de nœuds v utilisés, de 10 à 150. Les résultats Monte-Carlo suivants portent sur $R = 100$ répétitions :

	$\hat{\mathbb{E}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{E}}\,\hat{\alpha}$
v= 10	3.379023	0.5	1.027053	0.092060	0.614354
v= 15	3.484196	0.5	1.022018	0.096048	0.593247
v= 20	4.133774	0.5	1.013997	0.097472	0.545973
v= 25	3.645722	0.5	1.017246	0.100122	0.524762
v= 50	3.553642	0.5	1.101470	0.097913	0.488965
v= 100	3.336990	0.5	1.046654	0.075045	0.531928
v= 150	3.817476	0.5	1.071844	0.043925	0.551977

	$\hat{\mathbb{V}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{V}}\,\hat{\alpha}$
v= 10	11.345304	0	0.082181	0.001089	0.052945
v= 15	7.688250	0	0.083083	0.000990	0.063045
v= 20	8.288016	0	0.117973	0.000964	0.062710
v= 25	5.948421	0	0.131114	0.001145	0.045438
v= 50	8.250116	0	0.147055	0.000952	0.058766
v= 100	4.155969	0	0.112876	0.000980	0.054458
v= 150	9.487893	0	0.106311	0.001222	0.042317

	$\hat{\mathbb{E}}\hat{\beta}_1$	$\hat{\mathbb{E}}\hat{\beta}_2$	$\hat{\mathbb{V}}\hat{\beta}_1$	$\hat{\mathbb{V}}\hat{\beta}_1$
v= 10	0.1431	0.2521	0.0529	0.1513
v= 15	0.1389	0.3198	0.0570	0.1275
v= 20	0.1431	0.3193	0.0537	0.1406
v= 25	0.1489	0.3321	0.0501	0.1461
v= 50	0.1482	0.3005	0.0482	0.1417
v= 100	0.1336	0.2975	0.0532	0.1297
v= 150	0.1438	0.2951	0.0545	0.1322

	$\widehat{\mathrm{MSE}}\,\hat{b}$	$\widehat{\mathrm{MSE}}\,\hat{U}$	$\widehat{\mathbb{E}\,\lambda(\hat{f})}$	$\widehat{\mathrm{MISE}}\,\hat{f}$
v= 10	0.144440	0.223494	5.4265	3.5238
v= 15	0.127704	0.200240	14.3440	2.7624
v= 20	0.141987	0.202572	31.7269	2.5617
v= 25	0.136588	0.201787	55.8270	2.5381
v= 50	0.166776	0.222075	593.4784	2.8996
v= 100	0.126041	0.226290	8182.9818	3.5502
v= 150	0.131591	0.225003	41099.1194	3.9387

3.4.5 Essais Monte-Carlo à τ^2 estimé par REML et matrice Ω=ID

	$\hat{\mathbb{E}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{E}}\,\hat{\alpha}$
v= 10	3.554444	0.438723	1.080074	0.087925	0.642480
v= 15	3.609221	0.437830	1.021542	0.092092	0.615662
v= 20	4.031172	0.390282	1.088268	0.097499	0.545214
v= 25	3.683863	0.297961	1.044547	0.099739	0.570731
v= 50	3.738444	0.089499	1.205341	0.088925	0.501282
v= 100	3.234561	0.022861	1.256879	0.077496	0.543429
v= 150	3.961880	0.003881	1.249937	0.078226	0.583381

	$\hat{\mathbb{V}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{V}}\,\hat{\alpha}$
v= 10	10.517072	0.165428	0.116151	0.001101	0.058772
v= 15	7.541139	0.090269	0.117025	0.001222	0.076205
v= 20	7.359055	0.056383	0.199281	0.001021	0.054849
v= 25	7.200228	0.032007	0.178944	0.001224	0.058767
v= 50	6.597250	0.013155	0.146266	0.001046	0.032477
v= 100	5.204137	0.003737	0.134464	0.000790	0.030104
v= 150	10.370757	0.004248	0.128026	0.001063	0.034174

	$\hat{\mathbb{E}}\hat{\beta}_1$	$\hat{\mathbb{E}}\hat{\beta}_2$	$\hat{\mathbb{V}}\hat{\beta}_1$	$\hat{\mathbb{V}}\hat{\beta}_1$
v= 10	0.1456	0.2625	0.0588	0.1447
v= 15	0.1470	0.3116	0.0600	0.1233
v= 20	0.1443	0.3107	0.0472	0.1227
v= 25	0.1317	0.2900	0.0577	0.1343
v= 50	0.1486	0.3049	0.0554	0.1288
v= 100	0.1394	0.2870	0.0566	0.1463
v= 150	0.1328	0.2999	0.0508	0.1267

	$\widehat{\mathrm{MSE}}\,\hat{b}$	$\widehat{\mathrm{MSE}}\,\hat{U}$	$\mathbb{E}\,\lambda(\hat{f})$	$\widehat{\mathrm{MISE}}\,\hat{f}$
v= 10	0.157601	0.264551	22.9025	4.2771
v= 15	0.153120	0.221689	12.3161	3.1603
v= 20	0.199021	0.274651	29.4606	2.8205
v= 25	0.152412	0.249476	35.1455	2.8647
v= 50	0.148387	0.257994	127.5450	3.7871
v= 100	0.138325	0.270958	610.8833	4.3016
v= 150	0.119891	0.277760	2609.9939	4.4220

3.4.6 Essais Monte-Carlo à τ^2 estimé par REML et matrice Ω=CISD

	$\hat{\mathbb{E}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{E}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{E}}\,\hat{\alpha}$
v= 10	3.826241	0.042665	1.143775	0.078814	0.602241
v= 15	3.564028	0.156421	1.033181	0.085428	0.646430
v= 20	3.306657	0.118868	1.111692	0.083819	0.605141
v= 25	3.815514	0.154809	1.065304	0.090648	0.639991
v= 50	3.293482	0.171018	1.060109	0.084911	0.632482
v= 100	3.822364	0.156316	1.050547	0.087729	0.635165
v= 150	3.341284	0.071566	1.109268	0.083686	0.626425

	$\hat{\mathbb{V}}\,\hat{\sigma}_b^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_a^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_U^2$	$\hat{\mathbb{V}}\,\hat{\sigma}_\varepsilon^2$	$\hat{\mathbb{V}}\,\hat{\alpha}$
v= 10	7.716609	0.003834	0.103706	0.000831	0.046486
v= 15	9.805342	0.153155	0.090542	0.001278	0.075970
v= 20	5.636180	0.053799	0.125108	0.001058	0.057415
v= 25	8.676709	0.074586	0.100095	0.001049	0.101033
v= 50	7.660361	0.076976	0.105165	0.000985	0.049771
v= 100	11.364060	0.071065	0.137780	0.000917	0.052998
v= 150	7.261163	0.040321	0.148361	0.000958	0.060663

	$\hat{\mathbb{E}}\hat{\beta}_1$	$\hat{\mathbb{E}}\hat{\beta}_2$	$\hat{\mathbb{V}}\hat{\beta}_1$	$\hat{\mathbb{V}}\hat{\beta}_1$
v= 10	0.0839	0.2818	0.0948	0.1633
v= 15	0.0587	0.2872	0.1852	0.1398
v= 20	0.0597	0.2767	0.1348	0.1224
v= 25	0.0496	0.2827	0.1417	0.1303
v= 50	0.0186	0.2787	0.1821	0.1298
v= 100	0.0254	0.2655	0.1394	0.1479
v= 150	0.0922	0.2842	0.0993	0.1407

	$\widehat{\mathrm{MSE}}\,\hat{b}$	$\widehat{\mathrm{MSE}}\,\hat{U}$	$\mathbb{E}\,\lambda(\hat{f})$	$\widehat{\mathrm{MISE}}\,\hat{f}$
v= 10	1.290345	0.644037	0.3019	39.0879
v= 15	1.122389	0.331035	1.4276	41.1956
v= 20	0.889524	0.384683	0.9791	32.2671
v= 25	1.192514	0.302249	1.3385	35.5405
v= 50	1.025762	0.461010	1.5861	34.9511
v= 100	0.922464	0.297311	1.3810	32.8464
v= 150	1.014076	0.285783	0.7272	25.2109

3.4.7 Commentaires des essais Monte-Carlo

Le premier enseignement de ces essais Monte-Carlo est la différence notable de comportement de l'approche REML selon le choix de la matrice de pénalisation Ω :

Le cas Ω=ID produit une estimation globalement correcte de la composante $f(t)$ (d'après un MISE estimé proche de 4). L'augmentation du nombre de nœuds n'entraîne pas celle de la valeur du MISE. Par contre, cette augmentation détériore la régularité de la fonction $\hat{f}(t)$ (d'après un $\lambda(\hat{f})$ estimé passant de 12 à 2610) malgré une diminution du paramètre de contrôle τ^2 (qui passe approximativement de $0,4$ à $0,004$). On notera que la valeur théoriquement attendue $\lambda_0 \approx 28,2$ est atteinte en moyenne lorsque le nombre de nœuds est v

compris entre 15 et 20 ; ce qui correspond au nombre de nœuds effectivement utilisés pour simuler $f(t)$ ($K = 17$).
Le cas Ω=CISD donne des résultats dégradés sur les valeurs de MISE, et ce quel que soit le nombre de nœuds choisi. Par contre il faut bien sûr noter que la régularité de la fonction $\hat{f}(t)$ est beaucoup mieux contrôlée, puisque la régularité $\lambda(\hat{f})$ estimée reste inférieure à $1,5$ quel que soit le nombre de nœuds. Cette valeur $1,5$ est nettement inférieure à la valeur attendue $28,2$, ce qui nous laisse penser que la pénalisation Ω=CISD est trop sévère ici. Ce phénomène resterait à confirmer par de plus amples simulations encore ; et si tel est le cas, il serait bon d'envisager un pénalisation Ω qui serait un compromis entre le cas ID et le CISD (une combinaison linéaire des deux matrices Ω par exemple).
Le choix de Ω semble avoir aussi un effet sur la prédictions des composantes aléatoires du modèle. On remarque surtout une dégradation (d'un facteur dix) du MSE($\hat{b}$) dans le cas Ω=CISD par rapport au cas Ω=ID. ; la dégradation est moindre pour le MSE($\hat{U}$).

Chapitre 4

Analyse de sensibilité dans le modèle mixte semiparamétrique stochastique

4.1 Introduction

Dans les modèles paramétriques, il existe plusieurs méthodes et outils de diagnostics pour l'ajustement robuste du modèle par rapport aux particularités des données. Ces outils de diagnostics sont rares lorsque le modèle est semiparamétrique.
Dans le présent chapitre, nous exposons quelques outils de diagnostics et d'analyse de sensibilité *locale* pour le modèle décrit. Nous effectuons dans un premier temps un rappel sur la distance de Cook, aujourd'hui d'un usage courant en régression linéaire ou plus généralement pour les modèles paramétriques. Ensuite nous introduisons la méthode d'influence locale dans le modèle linéaire général.
Après ces rappels, nous développons une analyse de sensibilité pour le modèle mixte semiparamétrique stochastique afin d'identifier les composantes structurelles les plus influentes. En particulier, nous nous focalisons sur les composantes susceptibles de montrer des variations non contrôlées. Une analyse de sensibilité locale sur ces composantes permet d'évaluer les effets de telles perturbations sur le modèle mixte sémiparamétrique et de faire un choix basé sur la robustesse.
Une analyse de sensibilité locale de l'estimation de la fonction nonparamétrique est informative et peut être un indicateur à prendre en compte dans la recherche de la robustesse. L'extension de cette analyse de la sensibilité peut s'appliquer au cas de la sélection des hyperparamètres associées à l'estimation de la fonction nonparamétrique.

Cependant il faut noter que la méthode d'influence locale est à utiliser en combinaison avec les autres méthodes de mesure de l'adéquation du modèle telles que les intervalles de confiance, l'analyse des résidus avec les différents degrés de liberté des modèles.....
Dans le cadre des modèles semiparamétriques, des approches de cette nature ne sont pas très répandues bien que souvent l'hypothèse du modèle reste très proche du cadre paramétrique. Cependant le développement de tels éléments de diagnostics, est à nos yeux aussi nécessaire dans le cadre des modèles semiparamétriques voire nonparamétriques que paramètriques.

4.2 Rappel sur la distance de Cook

Dans la plupart des modèles statistiques, les estimateurs des paramètres ont des propriétés qui peuvent être qualifiées de globales en ce sens que les critères d'adéquation aux données permettant leur choix sont évalués sur la globalité des données. Cette approche globale conduit à des paramètres dont la robustesse peut être mise à rude épreuve par des perturbations locales sur les données. Ces perturbations peuvent être considérées comme des infimes changements dans nos données susceptibles de provoquer des effets majeurs sur les résultats et les interprétations. C'est pourquoi, nous pensons que pour tester la robustesse du modèle, une évaluation locale de ces propriétés est nécessaire.

Dans le cadre d'un modèle paramétrique, il existe des éléments de diagnostics pour évaluer l'ajustement "*local*" des paramètres aux données ; la distance de Cook, aujourd'hui d'un usage «quasi» courant dans le cadre paramétrique est un outil de diagnostic qui est à la base des méthodes de l'évaluation locale des modèles. Nous allons présenter les approches d'influence locale obtenues comme extension de la distance de Cook. L'une des premières extensions de la distance de Cook a été proposée par Pregibon [31] pour le modèle logistique en tant qu'une extension de la distance de Cook introduite en 1977.
La distance de Cook est un critère facile à interpréter, introduite pour mesurer l'effet de chaque observation sur l'estimation des paramètres obtenus par moindres carrés généralisés (GLS). C'est une méthode combinant l'information conjointe des résidus studentisés, la variance des résidus ainsi que les valeurs prédites. Elle a été développée au sein du modèle linéaire général : $Y \sim \mathcal{N}(X\beta, V)$ où $V = \sigma^2 R$.
Le DFBETAS défini ci-dessous permet d'évaluer l'effet de l'amputation d'une observation sur le paramètre β estimé par moindres carrés généralisés :

Définition 1
En notant par $\hat{\beta}$ l'estimateur par maximum de vraisemblance du modèle $Y = X\beta + \varepsilon$ où $\varepsilon \sim \mathcal{N}(0, V)$ et par $\hat{\beta}_{(i)}$, l'estimateur par maximum de vraisemblance sur les données sans la i^{eme} observation.

$$DFBETAS = (\Delta_{(i)}\hat{\beta})^T (X^T \hat{V}^{-1} X)^{-1} (\Delta_{(i)}\hat{\beta}),$$

où $\Delta_{(i)}\hat{\beta} = \hat{\beta}_{(i)} - \hat{\beta}$.

Quand une observation ou des observations entraînent des valeurs plus élevées par rapport à l'ensemble des DFBETAS alors ces dernières sont considérées comme influentes sur l'estimation des paramètres de la régression.
En 1977, Cook [7] introduit une mesure basée sur l'ellipsoïde de confiance de β pour juger de la contribution de chaque observation à l'estimation du paramètre. Cette distance est connue depuis lors sous le nom de la distance de Cook et est définie par :

Définition 2
Pour déterminer le niveau d'influence du retrait des données de la i^eobservation sur l'estimation du paramètre d'intérêt β, Cook propose la distance notée D_i ci-dessous :
$D_i = \dfrac{\left\|\hat{Y} - \hat{Y}_{(i)}\right\|^2}{p\hat{\sigma}^2}$, *où* $p = rang(X)$ *et* $\hat{\sigma}^2 = \dfrac{\hat{\varepsilon}^T\hat{\varepsilon}}{N-p}$.
Avec $\hat{Y}$, le vecteur des $\hat{y}_i$, la prédiction de la i^eobservation à partir du modèle construit sur l'ensemble des observations ;
$\hat{Y}_{(i)}$ le vecteur des $\hat{y}_{i(i)}$, la prédiction de la i^eobservation à partir du modèle construit sans cette dernière.

Remarque 1
La distance de Cook D_i se met sous la forme suivante :

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T (\widehat{Var}(\hat{\beta}))^{-1} \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{p},$$

où $\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2(X^TR^{-1}X)^{-1}$ dans le cadre de l'estimateur par GLS.

La *distance de Cook* ainsi définie, mesure l'écart entre les valeurs prédites par le modèle construit sans la i^{ieme} observation de celles prédites par le modèle initial proportionnellement à la variance résiduelle. La distance de Cook offre une méthode de mesure de la sensibilité des paramètres estimés à la présence de certaines observations particulières désignées sous le nom de *outliers* en raison de leur comportement atypique dans la construction du modèle.
Dans son article [7] de [1979], Cook apporte une contribution appréciable dans l'interprétation de cette distance en accentuant la relation explicite entre l'ellipsoïde de confiance pour les paramètres estimés et la distance de Cook. Il détailla les différents rôles de la variance résiduelle, de la corrélation résiduelle et du plus petit convexe construit autour des points d'échantillonnage.

Définition 3
Le convexe de Hull IVH défini par Cook sous le nom de the Independent Variable Hull *est un ellipsoïde conçu autour des points d'apprentissage. Si nous désignons par $H = X\left(X^TV^{-1}X\right)^{-1}X^TV^{-1}$, alors pour tout point arbitraire x, le convexe de Hull est défini par :*

$$IVH = \left\{x \in \mathbb{R}^p, tel\ que\ x\left(X^TV^{-1}X\right)^{-1}x^T \leq \max_i(HV)_{ii}\right\}.$$

La distance de Cook peut s'établir comme une fonction monotone des termes du i^{e} résidu studentisé et aussi de la position de l'observation considérée dans le convexe IVH. Or ces deux quantités caractérisent respectivement le fait que l'observation i considérée est un *outlier* et sa position dans le convexe de Hull IVH. Pour finir, notons que la distance de Cook peut être reliée à la statistique du *likelihood-ratio* LR *test* pour l'hypothèse que cette

observation est justement un *outlier* [7].
La distance de Cook [7] est donc reliée à la position de cette $i^{éme}$ observation dans le convexe IVH tout en sachant que cette position détermine en partie le niveau d'influence de l'observation dans le processus (procédure) de construction du modèle. Ainsi les observations situées sur le bord du convexe tendent à devenir influentes. L'influence du point diminue lorsque la représentation de l'observation se situe dans un endroit du convexe IVH avec une forte densité de positions de points, c'est-à-dire avec d'autres observations similaires du moins assez proches vis-à-vis de leur rôle dans la construction du modèle. La distance de Cook permet ainsi d'évaluer l'influence de la $i^{éme}$ observation enlevée sur le paramètre estimé et cette notion d'observation influente permet d'introduire celle d'influence locale.
En 1986, (Cook [8]) a donné une généralisation de la méthode à différentes composantes d'un modèle linéaire général en passant de l'estimation en l'absence de la $i^{ème}$ observation à une estimation plus douce : l'observation n'est plus enlevée, mais perturbée par la présence w_i. Ainsi les schémas précédents correspondent à des perturbations extrêmes avec la pondération nulle pour la $i^{ème}$ observation.
Les perturbations peuvent porter sur deux aspects du modèle selon l'approche adoptée ; une perturbation des données telle que nous l'avons abordée ici, ensuite une perturbation du modèle qui portent sur les paramètres d'intérêt, c'est celle adoptée par Magnus & Vasnev [28].
A travers cette approche, il apparaît naturellement que le concept d'influence locale peut être étendu à l'ensemble des modèles paramétriques. Cette approche a été mise en oeuvre au sein du modèle linéaire mixte généralisé pour différents schémas de perturbations sur les observations par des auteurs comme (Zhu & Lee [49]). Il nous paraît donc intéressant dans ce contexte d'étendre cette notion au modèle semiparamétrique et au modèle mixte modifié. L'approche de l'influence locale dans le modèle mixte à effets aléatoires ayant déjà été exposé par Lesaffre & Verbeke [27].

4.3 L'analyse de sensibilité dans le cadre du modèle linéaire général

Soit un modèle de régression à matrice de variance connue avec un paramètre $\theta \in \mathbb{R}^d$ à déterminer et soit $l(\theta)$ (notation spécifique à cette partie) la fonction de logvraisemblance associée.
Un schéma de perturbation est introduit, pour le modèle à travers un vecteur w élément de W un ouvert. Ainsi ce vecteur w induit des perturbations infinitésimales sur la composante du modèle choisi. Nous désignerons souvent cette composante comme étant structurelle pour notifier qu'il s'agit des données servant à construire le modèle que nous perturbons.
On introduit la notion de logvraisemblance perturbée :
$l(\theta|w)$: la logvraisemblance perturbée du modèle ;
$l(\theta) = l(\theta|w_0)$ la logvraisemblance non perturbée du modèle donc correspondant à $w = w_0$.

Nous faisons les hypothèses suivantes :

Hypothèses 1

1. *L'espace des paramètres Θ est un compact de $\mathbb{R}^d$.*
2. *La valeur vraie θ_0 du paramètre θ est à l'interieur du compact Θ.*
3. *$l(\theta)$ est continue sur Θ*
4. *$l(\theta)$ est deux fois continûment différentiable.*

Avec une extension naturelle de ces hypothèses à $l(\theta|w)$

Ces conditions assez courantes dans le cadre des modèles paramétriques assurent que l'estimateur du maximum de vraisemblance $\hat{\theta}_{ML}$ existe et qu'il est consistent. L'approche de modèle mixte modifié permet d'assurer que le modèle mixte semiparamétrique vérifie toutes ces conditions. Donc $\hat{\theta}_{ML}$ est asymptotiquement efficace :

$$\forall \theta \in \Theta, \sqrt{N}\left(\hat{\theta}^N_{ML} - \theta_0\right) \xrightarrow{L} \mathcal{N}(0, \{I_{\theta_0\theta_0}\}^{-1}), \ quand\ N \to \infty,$$

où θ_0 est la vraie valeur du paramètre.
Les hypothèses faites sont sans doute fortes, Magnus & Vasnev [28] donnnent des conditions relativement plus faibles pour garantir la consistance de l'estimateur ML dans le cadre paramétrique.

Définition 4
Le déplacement de la logvraisemblance est alors défini par :

$$LD(w) = 2\left[l(\hat{\theta}) - l(\hat{\theta}_w)\right],$$

avec

$$\hat{\theta} = arg \underset{\theta}{max}\, l(\theta),$$

et

$$\hat{\theta}_w = arg \underset{\theta}{max}\, l(\theta|w).$$

Le déplacement de logvraisemblance introduit par Cook permet d'évaluer l'effet donc l'influence de la variation introduite par w. Les directions d'évaluation de ces influences sont données par les différents éléments de la base de W à travers lesquels $LD(w)$ montre des variations importantes au point $w = w_0$.

Remarque 2
Il faut noter que : $l(\hat{\theta}) = l(\hat{\theta}_{w_0})$ où w_0 correspond au cas du modèle non perturbé. Comme le maximum de la logvraisemblance est atteint au point $\hat{\theta}$ alors par définition $LD(w) \geq 0$ et on obtient $LD(w_0) = 0$.

Dans le cadre du modèle linéaire généralisé à matrice de variance connue, la fonction de logvraisemblance peut s'écrire à une constante près :

$$l(\beta|w_0) = -\frac{1}{2}\left(Y - X\beta\right)^T V^{-1}\left(Y - X\beta\right),$$

A titre d'exemples, voici trois schémas usuels de perturbation :

1. Les perturbations portent sur les individus, suivant un schéma multiplicatif, alors la matrice diagonale W des perturbations s'introduit comme une matrice de pondération pour donner la forme :

   $$l(\beta|w) = -\frac{1}{2}\left(Y - X\beta\right)^T V^{-\frac{1}{2}}\tilde{W}V^{-\frac{1}{2}}\left(Y - X\beta\right),$$

   Ce cas de perturbation est souvent désigné sous le nom de *logvraisemblance locale*.
   Les perturbations portent sur les individus, sous forme d'une matrice diagonale $\tilde{W} = diag(w)$ distincte de la matrice identité. Le cas sans perturbation correspond à $w_0 = (1, \ldots, 1) \in \mathbb{R}^N$
2. Les perturbations portent sur la variable d'intérêt Y sous forme additive $Y_w = Y + \tilde{W}$, nous obtenons alors la logvraisemblance perturbée par :
   $l(\beta|w) = -\frac{1}{2}\left(\left(Y + \tilde{W}\right) - X\beta\right)^T V^{-1}\left(\left(Y + \tilde{W}\right) - X\beta\right)$. $\tilde{W}$ est un vecteur de même taille que Y. Dans ce cas, la logvraisemblance est obtenue pour $w_0 = (0, \ldots, 0)$.
3. Les perturbations portent sur les variables explicatives, nous obtenons alors la logvraisemblance perturbée définie par :
   $l(\beta|w) = -\frac{1}{2}\left(Y - \left(X + \tilde{W}\right)\beta\right)^T V^{-1}\left(Y - \left(X + \tilde{W}\right)\beta\right)$ La perturbation, comme dans le cas précédent, est aussi de forme additive et $\tilde{W}$ est un vecteur de même taille que X.

Le déplacement de la logvraisemblance est le critère permettant d'observer la sensibilité des paramètres estimés aux perturbations w relativement petites, c'est sur ce critère qu'est basée la méthode d'influence locale.

4.3.1 L'approche géométrique de Cook de l'influence locale

La notion de la courbure d'un graphe utilisée en géométrie différentielle a été à la base de l'approche intuitive proposée par Cook. Cette notion permet la présentation simple des outils de l'évaluation du déplacement de logvraisemblance et donc de l'influence locale.
Ainsi par exemple pour un cercle, la courbure est le rayon de ce cercle, et pour une courbe quelconque, en un point donné de cette courbe c'est le rayon du cercle qui s'ajuste le mieux à cette courbe en ce point précis(dans le voisinage du point).
Alors, la fonction de la logvraisemblance $l(\theta)$ peut être vue comme une courbe, suivant une paramétrisation convenable, de la composante à perturber du modèle ; cette composante pouvant être $Y, X, \ldots$.

Au point $\hat{\theta}$, la courbe notée $\mathfrak{C}(w) = \begin{pmatrix} w \\ LD(w) \end{pmatrix}$ de la logvraisemblance atteint son maximum ; ce type de courbe $\mathfrak{C}(w)$ est connue, en géométrie différentielle, sous le nom de *Monge Path*. Ce point $\hat{\theta}$ est aussi l'intersection de la courbe avec son plan tangent, plan orthogonal à la normale à la courbe en ce point. Une perturbation w suivant une direction donnée permet d'observer comment la courbe quitte le plan tangent à ce point $\hat{\theta}$. L'allure selon laquelle la courbe de la logvraisemblance quitte son plan tangent suivant la direction de cette perturbation dénote de la sensibilité du paramètre à la perturbation w. Cette élévation s'évalue aisément suivant la direction de la normale. La coordonnée associée dans cette direction est la courbure, par analogie, c'est la vitesse selon laquelle la courbe quitte son plan tangent dans la direction w.
La formule de la courbure pour $\mathfrak{C}(w)$ d'après Cook [8] est donnée par :

$$\mathcal{C} = \left|\dot{\mathfrak{C}}_1\ddot{\mathfrak{C}}_2 - \ddot{\mathfrak{C}}_1\dot{\mathfrak{C}}_2\right| / \left(\dot{\mathfrak{C}}_1^2 + \dot{\mathfrak{C}}_2^2\right)^{3/2}. \tag{4.1}$$

La définition de la matrice d'influence qui reste reliée à la perturbation w des modèles est utile pour la suite.

Définition 5
La matrice d'influence est la quantité notée par $\ddot{F}$ définie ci-dessous :

$$\ddot{F} = \left.\frac{\partial^2 l(\hat{\theta}_w)}{\partial w\, \partial w^T}\right|_{w=w_0}.$$

L'application de la formule (4.1) à $\mathfrak{C}(w)$, donne la courbure de la fonction $LD(w)$, au voisinage d'un point noté w_* dans la direction $\vec{l}$ avec $\left\|\vec{l}\right\| = 1$ et $w = w_* + \vec{l}$.

$$\mathcal{C}(w_*, \vec{l}) = \frac{-2\vec{l}^T\ddot{F}\vec{l}}{\left(1 + \left\|\dot{F}\right\|^2\right)^{-\frac{1}{2}}\vec{l}^T\left(I + \dot{F}\dot{F}^T\right)\vec{l}},$$

où $\dot{F} = \left.\dfrac{\partial l(\hat{\theta}_w))}{\partial w}\right|_{w=w_0}$.
L'évaluation de cette quantité au point w_0 où $LD(w_0) = 0$ se ramène tout simplement à évaluer le numérateur car $\dot{F} = 0$ en w_0
La courbure $\mathcal{C}$ est donnée par $\ddot{\mathfrak{C}}_2 = \ddot{\overbrace{LD}}(w_0)$.
Le calcul direct de $\ddot{\overbrace{LD}}(w_0)$ donne :

$$\ddot{\overbrace{LD}}(w_0) = -2\left.\frac{\partial^2 l(\hat{\theta}_w)}{\partial w\, \partial w^T}\right|_{w=w_0}, \tag{4.2}$$

Avec ces notations, nous introduisons la proposition suivante due à Cook :

Proposition 1 [*Cook*, 1986]

$$\ddot{F} = \Delta^T \ddot{L}^{-1} \Delta,$$

où $\Delta = \left.\dfrac{\partial^2 l(\theta|w)}{\partial\theta\partial w^T}\right|_{\substack{\theta=\hat{\theta}\\ w=w_0}}$ *la dérivée seconde de la logvraisemblance par rapport à* θ *et* w.

Preuve
En utilisant les règles classiques de la dérivation et d'autre part en supposant θ comme une fonction suffisamment régulière en w, on obtient les égalités :

$$\ddot{F} = \left.\frac{\partial\hat{\theta}_w}{\partial w^T}\right|_{w=w_0} \left.\frac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta^T}\right|_{\theta=\hat{\theta}} \left.\frac{\partial\hat{\theta}_w}{\partial w^T}\right|^T_{w=w_0} = J^T \ddot{L} J, \tag{4.3}$$

où $J^T = \left.\dfrac{\partial\hat{\theta}_w}{\partial w^T}\right|_{w=w_0}$.
Pour évaluer la matrice J, il faut remarquer que :

$$\left.\frac{\partial l(\theta|w)}{\partial\theta}\right|_{\theta=\hat{\theta}_w} = 0. \tag{4.4}$$

En dérivant l'équation (4.4) par rapport aux différents éléments de w et θ et en évaluant cela à $w = w_0$ et à θ_w, on obtient l'égalité :

$$\left.\frac{\partial^2 l(\theta|w)}{\partial\theta\partial\theta^T}\right|_{\substack{w=w_0\\ \theta=\hat{\theta}}} \left.\frac{\partial\hat{\theta}_w}{\partial w^T}\right|_{w=w_0} + \left.\frac{\partial^2 l(\theta|w)}{\partial\theta\partial w^T}\right|_{\substack{w=w_0\\ \theta=\hat{\theta}}} = 0.$$

Cette dernière permet de noter que :

$$J = -\left(\ddot{L}\right)^{-1}\Delta. \tag{4.5}$$

En remplaçant l'expression (4.5) dans (4.3), en ajoutant cette dernière égalité dans (4.2), on obtient :

$$\ddot{F} = \Delta^T \ddot{L}^{-1} \Delta. \tag{4.6}$$

□

Définition 6
La mesure de l'influence locale suivant une direction notée $\vec{l}$ *est donnée par la courbure notée par :*

$$\forall\ \vec{l}\ avec\ \left\|\vec{l}\right\| = 1,\ C_{\vec{l}} = -2\vec{l}^T\Delta^T\ddot{L}^{-1}\Delta\vec{l}.$$

Une quantité assez informative pour l'interprétation des résultas de l'influence locale est le C_{max} défini par :

Définition 7 *La valeur propre associée au vecteur propre de $\ddot{F}$ entraînant le maximum de courbure est notée par :*

$$C_{max} = \underset{\|\vec{l}\|=1}{arg\,max} \left(-2\vec{l}^T \Delta^T \ddot{L}^{-1} \Delta \vec{l}\right). \tag{4.7}$$

Dans le cas il y a une unique direction de perturbation, la courbure est alors égale à C_{max}. Ainsi les courbures associées aux vecteurs colonnes de $\ddot{F}$ sont les éléments de la diagonale de la matrice d'influence $\ddot{F}$ losrque celle-ci est remplacée par la matrice réduite correspondante. Les grandes valeurs de ces éléments diagonaux dénotent d'une grande influence locale du vecteur colonne associé tandis que les petites valeurs dénotent d'une influence moindre. L'objectif est d'utiliser ces résultats de façon heuristique pour mettre en place un choix des paramètres avec une évaluation de la sensibilité locale aux perturbations des données

4.3.2 Approche analytique de l'influence locale

Nous élargissons les résultats de la proposition (1) dans un cadre analytique, suivant une autre approche qui nous paraît plus commode et justifierait pleinement le cadre dans lequel nous effectuons les perturbations des données. Sous les hypothèses (1), nous avons la proposition suivante :

Proposition 2
Soient $\hat{\theta}$ et $\hat{\theta}_w$, les estimateurs indiqués précédemment et en supposant que la logvraisemblance est deux fois continûment différentiable par rapport à w, alors une approximation du déplacement de la logvraisemblance suite à la perturbation par $w \in W$ est donnée par :

$$LD(w) = 2[l(\hat{\theta}) - l(\hat{\theta}_w)] \approx -(w - w_0)^T \Delta^T (\ddot{L})^{-1} \Delta (w - w_0).$$

Preuve :
Sous les hypothèses de la proposition, le développement de Taylor à l'ordre 2 de $l(\theta)$ au voisinage de $\hat{\theta}$ donne :

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta}) \frac{\partial l(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + o(\theta - \hat{\theta})^2.$$

En particulier pour $\theta = \hat{\theta}_w$, on obtient :

$$l(\hat{\theta}_w) = l(\hat{\theta}) + (\hat{\theta}_w - \hat{\theta}) \frac{\partial l(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} + \frac{1}{2}(\hat{\theta}_w - \hat{\theta})^T \frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T}\bigg|_{\theta=\hat{\theta}} (\hat{\theta}_w - \hat{\theta}) + o(\hat{\theta}_w - \hat{\theta})^2.$$

Comme $l(\theta)$ atteint son maximun au point $\hat{\theta}$, sa fonction tangente en ce point est nulle : $\dfrac{\partial l(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} = 0$, en tenant compte de cette dernière égalité, l'approximation suivante est obtenue pour le déplacement de la logvraisemblance :

$$2[l(\hat{\theta}_w) - l(\hat{\theta})] \approx (\hat{\theta}_w - \hat{\theta})^T \left.\frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T}\right|_{\theta=\hat{\theta}} (\hat{\theta}_w - \hat{\theta}). \tag{4.8}$$

D'autre part, le développement à l'ordre 1 de $\hat{\theta}_w$ au voisinage de $\hat{\theta}$ permet de noter que :

$$\hat{\theta}_w = \hat{\theta} + \left.\frac{\partial \hat{\theta}_w}{\partial w^T}\right|_{w=w_0} (w - w_0) + o(w - w_0), \tag{4.9}$$

d'où $\hat{\theta}_w - \hat{\theta} \approx \left.\dfrac{\partial \hat{\theta}_w}{\partial w^T}\right|_{w=w_0} (w - w_0)$.
Cette approximation permet d'écrire l'équation 4.8, sous la forme :

$$2[l(\hat{\theta}_w) - l(\hat{\theta})] \approx (w - w_0)^T \left.\frac{\partial \hat{\theta}_w}{\partial w^T}\right|_{w=w_0}^{T} \left.\frac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta^T}\right|_{\theta=\hat{\theta}} \left.\frac{\partial \hat{\theta}_w}{\partial w^T}\right|_{w=w_0} (w - w_0). \tag{4.10}$$

Considérons la fonction $\mathfrak{M}$ définie par : $\mathfrak{M}(\theta, w) = \dfrac{\partial l(\theta|w)}{\partial\theta}$.
Cette fonction est définie et continue au voisinage $\Theta_0 \times W$ de $(\hat{\theta}, w_0)$ et admet une dérivée partielle par rapport à θ $\left.\dfrac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta^T}\right|_{\theta=\hat{\theta}} \neq 0$ non nulle et continue. Donc on peut appliquer le théorème des fonctions implicites au point $(\hat{\theta}, w_0)$. La fonction implicite :

$$\begin{array}{rccl} \hat{\theta}_w & : & W & \to \Theta_0, \\ & & w & \to \hat{\theta}_w, \end{array}$$

où $\hat{\theta}_w$ solution de $\mathfrak{M}(\theta, w) = 0$. Ainsi d'après le théorème des fonctions implicites, on obtient finalement :

$$\left.\frac{\partial \hat{\theta}_w}{\partial w^T}\right|_{w=w_0} = - \left.\frac{\partial^2 l(\theta|w)}{\partial\theta\,\partial\theta^T}\right|^{-1}_{\substack{w=w_0 \\ \theta=\hat{\theta}}} \left.\frac{\partial^2 l(\theta|w)}{\partial\theta\,\partial w^T}\right|_{\substack{w=w_0 \\ \theta=\hat{\theta}}}, \tag{4.11}$$

En remplaçant l'expression obtenue en (4.11) dans (4.10), l'ègalité de la proposition en découle. □
Ainsi les conditions d'application du théorème des fonctions implicites sont réunies, permettant ainsi d'associer à chaque $w \to \hat{\theta}_w$ et du coup de définir le cadre de développement de nos perturbations.

Remarque 3
D'abord une conséquence immédiate est que :

$$\ddot{F} = \frac{1}{2}\overbrace{LD}^{\ddot{}}(w_0).$$

Sachant que le maximun de $l(\theta)$ est atteint au point $\hat{\theta}_{w_0}$, alors la matrice d'influence est définie négative.C'est une conséquence directe du fait $LD(w)$ est est une fonction positive atteignant som minimum en w_0.

Remarque 4
Une observation de la construction de la matrice d'influence $\ddot{F} = \Delta^T \ddot{L}^{-1} \Delta$ *permet d'identifier des rôles équivalents entre la matrice* Δ *et* $\Delta_{(i)}\hat{\beta}$ *de l'expression du DFBETAS et de noter que la matrice d'influence* $\ddot{F}$ *généralise ainsi le DFBETAS.*

4.3.3 Décomposition de la matrice d'influence

La matrice d'influence permet d'évaluer l'effet des perturbations sur les différents paramètres dont dépend la fonction de la logvraisemblance. Cette matrice d'influence est à la base de la méthode de l'influence locale d'où l'intérêt de sa décomposition suivant les différents paramètres du modèle (paramètres de moyenne et paramètres de variance). Nous espérons ainsi développer l'analyse de sensibilité sur les différents paramètres du modèle, en particulier, ceux du modèle mixte modifié.
Cook a établi que la courbure à travers la matrice d'influence $\ddot{F}$ se décompose suivant les différentes composantes du paramètre d'intérêt d'un modèle. Ainsi si le paramètre θ peut se scinder en deux composantes $\theta^T = (\theta_1^T, \theta_2^T)$; Cook définit :

$$LD_{\theta_1} = 2\left[l(\hat{\theta}_1, \hat{\theta}_2) - l(\hat{\theta}_{1w}, g_2(\hat{\theta}_{1w}))\right],$$

avec :$(\hat{\theta}_1^T, \hat{\theta}_2^T) = \underset{\theta_1,\theta_2}{\arg\max}\, l(\theta_1, \theta_2)$ et $(\hat{\theta}_{1w}^T, \hat{\theta}_{2w}^T) = \underset{\theta_1,\theta_2}{\arg\max}\, l(\theta_1, \theta_2|w)$. La fonction g_2 est définie comme suit :

$$\begin{array}{rccl} g_2: & \Theta_1 & \longrightarrow & \Theta_2, \\ & \theta_1 & \longmapsto & g_2(\theta_1) = \underset{\theta_2}{\arg\max}\, l(\theta_1, \theta_2) \end{array}$$
.

A partir du déplacement LD_{θ_1}, Cook introduit la matrice $\ddot{F}_{\theta_1}$ que nous désignerons comme matrice d'influence partielle. Selon la méthode de décomposition de Cook de l'influence locale, l'expression de la matrice d'influence partielle du paramètre θ_1 peut se mettre au final sous la forme :

$$\ddot{F}_{\theta_1} = \Delta^T \left(\ddot{L}^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & L_{\theta_2\theta_2}^{-1} \end{pmatrix} \right) \Delta. \tag{4.12}$$

Dans le cas du modèle mixte en général, les paramètres d'intérêt sont (β, θ) où β est le paramètre de la moyenne et θ constitue les paramètres de variance. Ces deux paramètres peuvent s'identifier aux paramètres de θ_1 et θ_2 notés précédemment.
En notant le hessien de la logvraisemblance par :

$$\ddot{L} = \begin{pmatrix} L_{\beta\beta} & L_{\beta\theta} \\ L_{\theta\beta} & L_{\theta\theta} \end{pmatrix},$$

l'inverse du hessien s'écrit alors :

$$\ddot{L}^{-1} = \begin{pmatrix} \left[L_{\beta\beta} - L_{\beta\theta} L_{\theta\theta}^{-1} L_{\theta\beta}\right]^{-1} & -L_{\beta\beta}^{-1} L_{\beta\theta} \left\{L^{\theta\theta}\right\} \\ -L_{\theta\theta}^{-1} L_{\theta\beta} \left\{L^{\beta\beta}\right\} & \left[L_{\theta\theta} - L_{\theta\beta} L_{\beta\beta}^{-1} L_{\beta\theta}\right]^{-1} \end{pmatrix},$$

où $L^{\theta\theta} = \left[L_{\theta\theta} - L_{\theta\beta}L_{\beta\beta}^{-1}L_{\beta\theta}\right]^{-1}$ et de même pour $L^{\beta\beta} = L_{\beta\beta}^{-1} + L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}L_{\theta\beta}L_{\beta\beta}^{-1}$.
Avec ces notations sur le hessien de la logvraisemblance et son inverse, la matrice d'influence partielle du paramètre β ou celle de θ peut être évaluée.

Un choix pertinent sur le paramètre θ permettrait en particulier de développer l'influence locale pour les hyperparamètres de la fonction nonparamétrique introduite dans le modèle mixte semiparamétrique stochastique comme composante de variance. Plus spécifiquement l'analyse de sensibilité sur ces hyperparamètres fournit des éléments de diagnostics pour affiner le choix de la fonction nonparamétrique. Par exemple, ces derniers peuvent être exploités dans le choix du nombre et de l'emplacement des nœuds.
Dans l'expression de la matrice d'influence $\ddot{F}$, l'inverse du hessien de la fonction de logvraisemblance notée par $(\ddot{L})^{-1}$ intervenant peut se décomposer asymptotiquement en une matrice diagonale en blocs. Nous allons montrer que de même la matrice d'influence peut aussi se décomposer asymptotiquement en blocs suivant la matrice d'influence partielle des paramètres de moyenne et celle des paramètres de la variance.
Pour cela, nous avons besoin d'établir les résultats des deux lemmes suivants. β_0 et θ_0 désignent les valeurs vraies des paramètres du modèle, les trois expressions ci-dessous utiles par la suite sont ainsi introduites :

$$\begin{aligned}
I_{\beta_0\beta_0} &= -\lim_{N\to\infty}\frac{1}{N}\frac{\partial^2 l(\beta|w)}{\partial\beta\,\partial\beta^T}\bigg|_{\beta=\beta_0}, \\
I_{\theta_0\theta_0} &= -\lim_{N\to\infty}\frac{1}{N}\frac{\partial^2 l(\beta|w)}{\partial\theta\,\partial\theta^T}\bigg|_{\theta=\theta_0}, \\
I_{\theta_0\beta_0} &= -\lim_{N\to\infty}\frac{1}{N}\frac{\partial^2 l(\beta|w)}{\partial\theta\,\partial\beta^T}\bigg|_{\substack{\beta=\beta_0\\ \theta=\theta_0}}, \\
K_{\beta_0 w_0} &= \lim_{N\to\infty}\frac{1}{N}\frac{\partial^2 l(\beta|w)}{\partial\beta\,\partial w^T}\bigg|_{\substack{\beta=\beta_0\\ w=w_0}}, \\
K_{\theta_0 w_0} &= \lim_{N\to\infty}\frac{1}{N}\frac{\partial^2 l(\theta|w)}{\partial\theta\,\partial w^T}\bigg|_{\substack{\theta=\theta_0\\ w=w_0}}.
\end{aligned}$$

Lemme 1
$\forall\beta_0, \sqrt{N}\left(\hat{\beta}_w - \beta_0\right) \xrightarrow{L} \mathcal{N}(I_{\beta_0\beta_0}^{-1}K_{\beta_0 w_0}(w - w_0), \{I_{\beta_0\beta_0}\}^{-1})$ *quand* $N \to \infty$,
si en plus $K_{\beta_0 w_0} = 0$
alors on a : $\forall\beta, \sqrt{N}\left(\hat{\beta}_w - \beta_0\right) \xrightarrow{L} \mathcal{N}(0, \{I_{\beta_0\beta_0}\}^{-1})$ *quand* $N \to \infty$.
De même on a pour le paramètre de variance θ :
$\forall\theta_0, \sqrt{N}\left(\hat{\theta}_w - \theta_0\right) \xrightarrow{L} \mathcal{N}(I_{\theta_0\theta_0}^{-1}K_{\theta_0 w_0}(w - w_0), \{I_{\theta_0\theta_0}\}^{-1})$ *quand* $N \to \infty$,
si on a $K_{\theta_0 w_0} = 0$,
on aura : $\forall\theta, \sqrt{N}\left(\hat{\theta}_w - \theta_0\right) \xrightarrow{L} \mathcal{N}(0, \{I_{\theta_0\theta_0}\}^{-1})$ *quand* $N \to \infty$.

Preuve

Nous allons démontrer le lemme pour β. On sait que : $\hat{\beta}_w = \hat{\beta} + \frac{\partial \hat{\beta}_w}{\partial w}\Big|_{w=w_0} (w - w_0) = \hat{\beta} - \ddot{L}_{\beta\beta}^{-1} \Delta_{\beta w}(w - w_0)$, d'ou $E(\hat{\beta}_w) = \beta_0 + I_{\beta_0\beta_0}^{-1} K_{\beta_0 w_0} (w - w_0)$ d'où quand $K_{\beta_0 w_0} = 0$ le résultat du lemme en découle. □

Lemme 2

Sous les hypothèses de la proposition (2), si $\left(\hat{\beta}, \hat{\theta}\right)$ *et* $\left(\hat{\beta}_w, \hat{\theta}_w\right)$ *sont les estimateurs par maximum de vraisemblance pour* $l(\beta, \theta | w_0)$ *et* $l(\beta, \theta | w)$ *et qu'en plus on a les deux égalités suivantes :* $K_{\theta_0 w_0} = 0$, $K_{\beta_0 w_0} = 0$, *alors* $\hat{\beta}$ *et* $\hat{\theta}_w$ *sont asymptotiquement indépendantes. Il en est de même pour* $\hat{\theta}$ *et* $\hat{\beta}_w$.

Preuve

Pour la démonstration du lemme, nous considérons les suites $\left(\hat{\beta}_N, \hat{\theta}_N\right)$ et $\left(\hat{\beta}_{wN}, \hat{\theta}_{wN}\right)$ d'estimateurs du maximum de vraisemblance de (β, θ), on sait par leurs propriétés qu'elles sont convergentes, comme $K_{\theta_0 w_0} = 0$, $K_{\beta_0 w_0} = 0$ en raison du lemme [1] et qu'elles vérifient les convergences en loi ci-dessous :

$$\begin{aligned}
\forall \beta, \sqrt{N}\left(\hat{\beta}_N - \beta_0\right) &\xrightarrow{L} \mathcal{N}(0, \{I_{\beta_0\beta_0}\}^{-1}) \quad quand\ N \to \infty, \\
\forall \beta, \sqrt{N}\left(\hat{\beta}_{wN} - \beta_0\right) &\xrightarrow{L} \mathcal{N}(0, \{I_{\beta_0\beta_0}\}^{-1}) \quad quand\ N \to \infty, \\
\forall \theta \in \Theta, \sqrt{N}\left(\hat{\theta}_N - \theta_0\right) &\xrightarrow{L} \mathcal{N}(0, \{I_{\theta_0\theta_0}\}^{-1}) \quad quand\ N \to \infty, \\
\forall \theta \in \Theta, \sqrt{N}\left(\hat{\theta}_{wN} - \theta_0\right) &\xrightarrow{L} \mathcal{N}(0, \{I_{\theta_0\theta_0}\}^{-1}) \quad quand\ N \to \infty.
\end{aligned}$$

Il apparaît donc que $\hat{\beta}$ est asymptotiquement équivalent à $\hat{\beta}_w$ et il en est de même pour $\hat{\theta}$ et $\hat{\theta}_w$.

Pour démontrer le lemme, il suffit d'établir le résultat ci-dessous :

$$E\left(Q_{\hat{\beta}_w} Q_{\hat{\theta}}^T\right) = 0,$$

où

$$\begin{aligned}
Q_{\hat{\beta}_w} &= \frac{\partial l(\beta, \theta)}{\partial \beta}\Big|_{\beta = \hat{\beta}_w}, \\
Q_{\hat{\beta}} &= \frac{\partial l(\beta, \theta)}{\partial \beta}\Big|_{\beta = \hat{\beta}}, \\
Q_{\hat{\theta}} &= \frac{\partial l(\beta, \theta)}{\partial \theta}\Big|_{\theta = \hat{\theta}}.
\end{aligned}$$

Or $Q_{\hat{\beta}_w} \approx Q_{\hat{\beta}} + L_{\beta\beta}\left(Q_{\hat{\beta}_w} - Q_{\hat{\beta}}\right) + L_{\beta\theta}\left(Q_{\hat{\theta}_w} - Q_{\hat{\theta}}\right)$.

Par la suite, nous ne considérons que le terme du premier ordre, faisant évoluer $Q_{\hat{\theta}_w}$ et $Q_{\hat{\theta}}$ d'où on obtient :

$$Q_{\hat{\beta}_w} \approx Q_{\hat{\beta}} + L_{\beta\theta}\left(Q_{\hat{\theta}_w} - Q_{\hat{\theta}}\right).$$

L'équivalence asymptotique entre $\hat{\theta}_w$ et $\hat{\theta}$ permet d'écrire que :

$$\left(Q_{\hat{\theta}_w} - Q_{\hat{\theta}}\right) \approx L_{\theta\theta}^{-1} Q_{\hat{\theta}}.$$

Cette dernière peut être vue comme une étape de l'algorithme de Fisher en notant que $Q_{\hat{\theta}} = Q_{\hat{\theta}_w} - L_{\theta\theta}^{-1} Q_{\hat{\theta}}$ et considérant $Q_{\hat{\theta}_w}$ comme la fonction score de l'étape du départ

$$Q_{\hat{\beta}_w} \approx Q_{\hat{\beta}} + L_{\beta\theta} L_{\theta\theta}^{-1} Q_{\hat{\theta}},$$

alors,

$$\begin{aligned} E\left(Q_{\hat{\beta}_w} Q_{\hat{\theta}}^T\right) &\approx E\left(\left\{Q_{\hat{\beta}} + L_{\beta\theta} L_{\theta\theta}^{-1} Q_{\hat{\theta}}\right\} Q_{\hat{\theta}}^T\right), \\ &\approx E\left(Q_{\hat{\beta}}\, Q_{\hat{\theta}}^T\right) - L_{\beta\theta} L_{\theta\theta}^{-1} E\left(Q_{\hat{\theta}}\, Q_{\hat{\theta}}^T\right). \end{aligned}$$

Cette dernière montre que quand N devient grand $E\left(Q_{\hat{\beta}_w} Q_{\hat{\theta}}^T\right) = 0$.

Ce lemme donne les conditions sous lesquelles le paramètre $\hat{\beta}_w$ et $\hat{\theta}$ peuvent être considérés comme asymptotiquement indépendants comme c'est souvent le cas en pratique pour $\hat{\beta}$ et $\hat{\theta}$. Pour finir, il permet ainsi d'introduire la proposition suivante qui permet d'avoir une décomposition asymptotique de la matrice d'influence suivant les paramètres d'effets et ceux de la variance du modèle.

Proposition 3

Sous les hypothèses du lemme [2]*, si* $\left(\hat{\beta}, \hat{\theta}\right)$ *et* $\left(\hat{\beta}_w, \hat{\theta}_w\right)$ *sont les estimateurs par maximum de vraisemblance pour* $l(\beta, \theta|w_0)$ *et* $l(\beta, \theta|w)$ *alors, le déplacement de la logvraisemblance suite à la perturbation par* $w = w_0 + \rho\vec{\imath}$ *se met sous la forme :*

$$\begin{aligned} LD(w) = 2\left[l(\beta, \theta|w) - l(\beta, \theta|w_0)\right] = \\ -(w - w_0)^T \ddot{F}_{11}(w - w_0) - (w - w_0)^T \ddot{F}_{22}(w - w_0). \end{aligned}$$

Preuve

Sous les hypothèses du lemme [2], la démonstration du théorème reste similaire à celle du théorème de la proposition (2). Nous obtenons l'approximation suivante suivant une démarche similaire à celle de la proposition (2).

$$2\left[l\begin{pmatrix}\hat{\beta}_w \\ \hat{\theta}_w\end{pmatrix} - l\begin{pmatrix}\hat{\beta} \\ \hat{\theta}\end{pmatrix}\right] \approx$$

$$(w - w_0)^T \left[\frac{\partial\hat{\beta}_w}{\partial w} : \frac{\partial\hat{\theta}_w}{\partial w}\right] \begin{bmatrix} \frac{\partial^2 l(\beta,\theta)}{\partial\beta\partial\beta^T} & \frac{\partial^2 l(\beta,\theta)}{\partial\beta,\partial\theta^T} \\ \frac{\partial^2 l(\beta,\theta)}{\partial\theta\partial\beta^T} & \frac{\partial^2 l(\beta,\theta)}{\partial\theta\partial\theta^T} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial\hat{\beta}_w}{\partial w} \\ \frac{\partial\hat{\theta}_w}{\partial w} \end{bmatrix} (w - w_0).$$

L'utilisation du théorème des fonctions implicites pour l'obtention des expressions de $\begin{bmatrix} \frac{\partial\hat{\beta}(w)}{\partial w} \\ \frac{\partial\hat{\theta}(w)}{\partial w} \end{bmatrix}$, et d'autre part en remarquant que la matrice Δ est bloc-diagonal par construction permet

d'avoir l'approximation suivante :

$$2\left[l\begin{pmatrix}\hat{\beta}_w\\ \hat{\theta}_w\end{pmatrix}-l\begin{pmatrix}\hat{\beta}\\ \hat{\theta}\end{pmatrix}\right]\approx$$
$$(w-w_0)^T\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}^T\begin{bmatrix}\frac{\partial^2 l(\beta,\theta)}{\partial\beta\partial\beta^T} & \frac{\partial^2 l(\beta,\theta)}{\partial\beta\partial\theta^T}\\ \frac{\partial^2 l(\beta,\theta)}{\partial\theta\partial\beta^T} & \frac{\partial^2 l(\beta,\theta)}{\partial\theta\partial\theta^T}\end{bmatrix}^{-1}\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}(w-w_0).$$

Lorsque $N\to\infty$ et en tenant compte de l'indépendance asymptotique des paramètres de variance avec β, nous obtenons :

$$2\left[l\begin{pmatrix}\hat{\beta}_w\\ \hat{\theta}_w\end{pmatrix}-l\begin{pmatrix}\hat{\beta}\\ \hat{\theta}\end{pmatrix}\right]\approx$$
$$(w-w_0)^T\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}^T\begin{bmatrix}\left[\frac{\partial^2 l(\beta,\theta)}{\partial\beta\partial\beta^T}\right]^{-1} & 0\\ 0 & \left[\frac{\partial^2 l(\beta,\theta)}{\partial\theta\partial\theta^T}\right]^{-1}\end{bmatrix}\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}(w-w_0),$$

d'où le résultat de la proposition à savoir que : $LD(w)=-(w-w_0)^T\ddot{F}_{11}(w-w_0)-(w-w_0)^T\ddot{F}_{22}(w-w_0)$. □

Comme dans Cook [1986] [8], nous définissons les quantités $C_{\vec{\mathfrak{l}}}$ comme des coefficients de mesure de l'analyse de sensibilité dans la direction $\vec{\mathfrak{l}}$ par :

$$C_{\vec{\mathfrak{l}}}=-2\mathfrak{l}^T\ddot{F}\mathfrak{l}=-2\mathfrak{l}^T\Delta^T(\ddot{L})^{-1}\Delta\mathfrak{l}. \tag{4.13}$$

D'après la proposition (3) quand N devient grand, le coefficient $C_{\vec{\mathfrak{l}}}$ se décompose en :

$$C_{\vec{\mathfrak{l}}}=C_{\vec{\mathfrak{l}}}(\beta)+C_{\vec{\mathfrak{l}}}(\theta). \tag{4.14}$$

Remarque 5 *Dans le cadre du modèle linéaire général, le hessien de la logvraisemblance s'écrit :*

$$\ddot{L}^{-1}=\begin{bmatrix}\left[L_{\beta\beta}-L_{\beta\theta}L_{\theta\theta}^{-1}L_{\theta\beta}\right]^{-1} & -L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}\\ -L_{\theta\theta}^{-1}L_{\theta\beta}\left\{L^{\beta\beta}\right\} & \left[L_{\theta\theta}-L_{\theta\beta}L_{\beta\beta}^{-1}L_{\beta\theta}\right]^{-1}\end{bmatrix}.$$

En notant que la matrice $\Delta=\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}$, *la matrice d'influence peut se mettre sous la forme :*

$$\ddot{F}=\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}^T\begin{bmatrix}L_{\beta\beta}^{-1}+L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}L_{\theta\beta}L_{\beta\beta}^{-1} & -L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}\\ -L_{\theta\theta}^{-1}L_{\theta\beta}\left\{L^{\beta\beta}\right\} & \left[L_{\theta\theta}-L_{\theta\beta}L_{\beta\beta}^{-1}L_{\beta\theta}\right]^{-1}\end{bmatrix}\begin{bmatrix}\Delta_{\beta w}\\ \Delta_{\theta w}\end{bmatrix}.$$

Soit la matrice d'influence :

$$\begin{aligned}\ddot{F}=\ & \Delta_{\beta w}^T\left(L_{\beta\beta}^{-1}+L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}L_{\theta\beta}L_{\beta\beta}^{-1}\right)\Delta_{\beta w}-\Delta_{\beta w}^T\left(L_{\beta\beta}^{-1}L_{\beta\theta}\left\{L^{\theta\theta}\right\}\right)\Delta_{\theta w}\\ & -\Delta_{\theta w}^T\left(L_{\theta\theta}^{-1}L_{\theta\beta}\left\{L^{\beta\beta}\right\}\right)\Delta_{\beta w}+\Delta_{\theta w}^T\left[L_{\theta\theta}-L_{\theta\beta}L_{\beta\beta}^{-1}L_{\beta\theta}\right]^{-1}\Delta_{\theta w},\end{aligned}$$

dans le cas gaussien, asymptotiquement la quantité $L_{\theta\beta}=L_{\beta\theta}=0$ *; cela conduit à la simplification :* $\ddot{F}=\Delta_{\beta w}^TL_{\beta\beta}^{-1}\Delta_{\beta w}+\Delta_{\theta w}^TL_{\theta\theta}^{-1}\Delta_{\theta w}$. *Par ailleurs, conformément la décomposition de Cook, on sait que :* $\ddot{F}_\theta=\ddot{F}-\Delta_{\beta w}^TL_{\beta\beta}^{-1}\Delta_{\beta w}=\Delta_{\theta w}^TL_{\theta\theta}^{-1}\Delta_{\theta w}$ *d'où par identification, nous obtenons :* $\ddot{F}_\beta=\Delta_{\beta w}^TL_{\beta\beta}^{-1}\Delta_{\beta w}$. *Asymptotiquement, on a bien la décomposition :* $\ddot{F}=\ddot{F}_\theta+\ddot{F}_\beta$.

4.4 L'analyse de sensibilité dans le modèle mixte semiparamétrique stochastique

Nous appliquons la méthode de l'influence locale exposée au modèle mixte semiparamétrique en considérant logvraisemblance pénalisée. Cette dernière $l(\beta, a, \theta; y)$ s'écrit :

$$-\frac{n}{2}\log|V(\theta)| - \frac{1}{2}(Y - X\beta - NBa)^T V(\theta)^{-1}(Y - X\beta - NBa) - \frac{1}{2\tau^2}a^T\Omega a, \qquad (4.15)$$

où toutes les notations correspondent à celles définies dans les chapitres précédents.
Nous mettons la logvraisemblance pénalisée sous la forme :

$$-\frac{n}{2}log|V(\theta)| - \frac{1}{2}(Y - \mathfrak{X}\nabla)^T V(\theta)^{-1}(Y - \mathfrak{X}\nabla) - \nabla^T\mathfrak{J}^{-1}\nabla, \qquad (4.16)$$

où $\mathfrak{X} = (X; NB)$,$\nabla = (\beta, a)$ et $\mathfrak{J}$ est une matrice choisie de telle sorte que l'égalité $\frac{1}{2\tau^2}a^T\Omega a = \nabla^T\mathfrak{J}^{-1}\nabla$ soit vérifiée. Ces dernières notations introduites permettent d'interpréter facilement les résultats de l'application de l'influence locale et d'effectuer l'analyse de sensibilité locale.
Dans cette approche, nous considérons la courbe de la logvraisemblance pénalisée dans un espace dont la dimension est la taille des observations de la composante perturbée du modèle.
Dans le modèle mixte sémiparamétique, l'analyse de sensibilité peut s'effectuer à travers la méthode d'influence locale. Autrement dit à travers une perturbation des données ayant servi à mettre en place le modèle. Ce type d'influence est différent de la perturbation du modèle ou des paramètres d'intérêt du modèle.
Un vecteur directeur peut être associé à chaque élément w de l'espace W en vue de l'évaluation de l'effet de la perturbation. Cet effet est pratiquement évalué à travers les coefficients de la définition (6) de ce chapitre. Ces quantités une fois calculées peuvent être utilisées dans la pratique de différentes façons. L'évaluation s'effectue à travers une décomposition des vecteurs colonnes de la matrice $\ddot{F}$ dans cet espace. La décomposition des variables que constituent les vecteurs colonnes de la matrice $\ddot{F}$ en valeurs et vecteurs propres est l'étape principale dans cette évaluation. Ensuite pour la sensibilité du modèle, il faut considérer les vecteurs propres associés aux plus grandes valeurs propres. Ces vecteurs propres sont les directions des grandes perturbations et méritent une attention particulière. Un élément de la composante est influent si sa contribution suivant les directions données par ces vecteurs propres est importante. Les coordonnées de cet élément sur les axes associés aux principaux vecteurs propres sont les mesures de l'importance de l'influence.
La direction de perturbation importante est bien évidemment celle associée au vecteur propre de la plus grande valeur propre ; c'est la direction dont la perturbation entraîne un déplacement notable de la forme de logvraisemblance associée.

4.4.1 La matrice d'information de Fisher pour le modèle mixte semiparamétrique stochastique

L'application de la méthode d'influence locale décrite, dans la section précédente, au sein du modèle mixte semiparamétrique stochastique passe tout d'abord par l'évaluation du hessien $\ddot{L}$ du modèle non perturbé. La logvraisemblance pénalisée de ce dernier est donnée par (4.15). La mise en oeuvre de cette application nécessite la connaissance de l'inverse du hessien $\ddot{L}$ de la logvraisemblance pénalisée. Il faut noter que dans l'expression de la matrice d'influence $\ddot{F}$, celle-ci est présente de façon indépendante à la composante structurelle perturbée. Il est donc possible de calculer cette matrice une fois pour toutes et quelque soit le schéma de perturbaton.

$$\ddot{L} = \begin{pmatrix} X^T V^{-1} X & X^T V^{-1} NB \\ B^T N^T V^{-1} X & B^T N^T V^{-1} NB + \tau^{-2}\Omega \end{pmatrix} = \left(\mathfrak{X}^T V^{-1} \mathfrak{X} - \mathfrak{J}^{-1}\right).$$

Dans cette expression du hessien, nous avons supposé les paramètres de variance connus. Utilisant les règles d'inversion des matrices partitionnées en blocs et les notations de l'expression de $\ddot{L}$, l'inverse de cette matrice se calcule aisément. Nous introduisons les inverses par blocs d'abord, ensuite nous finissons par donner l'inverse complète de la matrice $\ddot{L}$:

$$\begin{aligned} [L_{22}]^{-1} = \quad & (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} + (B^T N^T V^{-1} NB + \\ & \tau^{-2}\Omega)^{-1} B^T N^T V^{-1} X \\ & \left\{X^T V^{-1} X - X^T V^{-1} NB (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1}\right. \\ & \left. B^T N^T V^{-1} X\right\}^{-1} X^T V^{-1} NB (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} \end{aligned}$$

$$\begin{aligned} = \quad & (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} + (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} \\ & B^T N^T V^{-1} X (X^T W_X X)^{-1} X^T V^{-1} NB (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1}, \end{aligned}$$

et

$$[L_{12}]^{-1} = \ -(X^T W_X X)^{-1} X^T V^{-1} NB (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1},$$

où comme dans Zhang [47], [1998] nous posons :

$$W_X = V^{-1} - V^{-1} NB (B^T N^T V^{-1} NB + \tau^{-2}\Omega)^{-1} B^T N^T V^{-1}.$$

L'inverse du hessien se met sous la forme :

$$\ddot{L}^{-1} = \begin{pmatrix} (X^T W_X X)^{-1} & [L_{12}]^{-1} \\ [L_{21}]^{-1} & [L_{22}]^{-1} \end{pmatrix} = \left(\mathfrak{X}^T V^{-1} \mathfrak{X} - \mathfrak{J}^{-1}\right)^{-1}.$$

4.4.2 Perturbation de la variable réponse

Il s'agit de remplacer dans cette partie, la réponse Y par $(Y + W)$ dans l'expression de la logvraisemblance pénalisée.

$$l(\theta; (Y + W)) = -\frac{1}{2}\log|V| - \frac{1}{2}\left((Y + \tilde{W}) - \mathfrak{X}\nabla\right)^T V^{-1}\left((Y + \tilde{W}) - \mathfrak{X}\nabla\right) - \nabla^T\mathfrak{J}^{-1}\nabla,$$

La matrice Δ s'obtient facilement et se présente sous la forme

$$\Delta = \mathfrak{X}^T V^{-1}.$$

Le calcul de $\ddot{F}$ conduit à l'expression :

$$\ddot{F} = \quad V^{-1}\mathfrak{X}\left(\mathfrak{X}^T V^{-1}\mathfrak{X} - \mathfrak{J}^{-1}\right)^{-1}\mathfrak{X}^T V^{-1} \tag{4.17}$$

L'expression (4.17) de $\ddot{F}$ est le produit de l'inverse de la matrice de variance V^{-1} et de la matrice H définie à la section (4.2) du modèle mixte modifié. Nous remarquerons qu'une perturbation sur la variable réponse correspond à un schéma de perturbations ne tenant pas compte des résidus (source de variation). Ceux-ci n'interviennent pas dans l'expression de la matrice d'influence (4.17). La décomposition de cette dernière dans l'espace de perturbation nous renseigne alors essentiellement sur la répartition (position) des différents individus dans le convexe de Hull défini par Cook. L'information apportée par ce schéma de perturbation est paradoxalement liée seulement à la variation introduite par la matrice d'expérience.

4.4.3 Perturbation des variables explicatives

Dans une régression classique, les variables explicatives sont considérées connues donc fixes. De ce fait, au sein du modèle statistique, la légitimité d'opérer des perturbations sur les composantes que sont les variables explicatives peut paraître ambiguë. Cependant le modèle mixte semiparamétrique est appliqué sur des données environnementales, domaine dans lequel les variables en particulier météorologiques présentent souvent des variations légitimant une telle approche. Il faut ajouter que les paramètres estimés par des méthodes proches des moindres carrés sont très sensibles à la présence de colinéarité au niveau des variables explicatives

Pour les variables à effets aléatoires, cette approche donnerait la sensibilité du modèle à des variations plus ou moins importantes susceptibles d'intervenir par des mécanismes non contrôlés. Entre autres, il serait intéressant d'évaluer la sensibilité du modèle à la fonction nonparamétrique faisant partie des composantes aléatoires du modèle. Nous avons noté que cette fonction nonparamétrique permettrait ainsi d'intégrer tous les mécanismes non contrôlées. Auparavant, nous aurions procédé à l'évaluation de l'influence du choix des noeuds pour le modèle. Cette question des nœuds influents sera traitée en considérant que l'introduction d'un nouveau noeud est équivalente à celle d'une nouvelle variable. Cette approche paraît plus abordable.

La logvraisemblance perturbée s'écrit ici :

$$\begin{aligned} l(\theta; Y, (\mathfrak{X} + \tilde{W})) = & -\tfrac{1}{2}\log|V| - \tfrac{1}{2}\left(Y - (\mathfrak{X} + \tilde{W}\Upsilon)\nabla\right)^T V^{-1}\left(Y - (\mathfrak{X} + \tilde{W}\Upsilon)\nabla\right) \\ & -\nabla^T\mathfrak{J}^{-1}\nabla, \\ = & -\tfrac{1}{2}\log|V| - \tfrac{1}{2}\left\{Y - (X + \tilde{W}_1\Upsilon_1)\beta - (NB + \tilde{W}_2\Upsilon_2)a\right\}^T \\ & V^{-1}\left\{Y - (X + \tilde{W}_1\Upsilon_1)\beta - (NB + \tilde{W}_2\Upsilon_2)a\right\} - \tfrac{1}{\tau^2}a^T\Omega a. \end{aligned}$$

La méthode que nous proposons est similaire sur le principe aux schémas de perturbation de Cook sur variable explicatives dans un modèle linéaire classique. Ainsi la matrice d'expérience des variables perturbées est donnée par $X_w = X + \tilde{W}_1\Upsilon_1$ (respectivement par $(NB + \tilde{W}_2\Upsilon_2)$ où $\tilde{W}_1$ (respectivement $\tilde{W}_2$) est une matrice $(n \times p)$ (respectivement $(n \times d)$ des poids w_1^{ij} et $\Upsilon_1 = diag(\upsilon_1^1, \ldots, \upsilon_1^p)$ (respectivement $\Upsilon_2 = diag(\upsilon_2^1, \ldots, \upsilon_2^d)$). L'élément diagonal υ_1^j de Υ_1 permet d'adapter les w_1^{ij} en taille et unité appropriées de telle sorte que $w_1^{ij}\upsilon_1^j$ (respectivement $w_2^{ij}\upsilon_2^j$) correspond au ij^{eme} élement de X (respectivement NB). Il faut noter que cette matrice d'expérience peut se noter globalement par :

$$\mathfrak{X}_w = (\mathfrak{X} + \tilde{W}) = (X + \tilde{W}_1\Upsilon_1, NB + \tilde{W}_2\Upsilon_2).$$

Pour une analyse de sensibilité des noeuds de la fonction spline, un schéma simplifié se ramenant à la perturbation d'une seule variable explicative est suffisant. Nous allons limité donc le schéma à la perturbation d'une variable.
Le premier cas, nous perturbons la j^{eme} variable explicative, alors Υ_1 se réduit au scalaire υ_1^j, la matrice Δ s'obtient en procédant comme dans le paragraphe précédent par

$$\Delta = \begin{pmatrix} \upsilon_1^j d_j\hat{\varepsilon}^T V^{-1} - \hat{\beta}_j\upsilon_1^j X^T V^{-1} \\ -\hat{\beta}_j\upsilon_1^j B^T N^T V^{-1} \end{pmatrix}, \tag{4.18}$$

où Δ est une matrice de dimension $((p+d) \times n)$, d_j un $(p+d)$-vecteur de 0 avec un 1 à la j^{eme} position et $\hat{\beta}_j$ le coefficient associé à cette j^{eme} variable.
Le deuxième cas, la perturbation porte sur le j^{eme} noeud parmi les (d) nœuds de la fonction spline une matrice Δ de la forme :

$$\Delta = \begin{pmatrix} -\hat{a}_j\upsilon_2^j X^T V^{-1} \\ \upsilon_2^j d_j\hat{\varepsilon}^T V^{-1} - \hat{a}_j\upsilon_2^j B^T N^T V^{-1} \end{pmatrix}, \tag{4.19}$$

où Δ est une matrice de la même dimension $((p+d) \times n)$, d_j un $(p+d)$-vecteur de 0 avec un 1 à la j^{eme} position du noeud perturbé et $\hat{a}_j$, le coefficient associé à ce j^{eme} noeud.

Pour simplifier prenons $\upsilon_1^j = \upsilon_2^j = 1$, d'où dans ce cas, la matrice d'influence de la perturbation de la j^{eme} variable explicative est :

$$\begin{aligned} \ddot{F} = & V^{-1}(d_j\hat{\varepsilon}^T)^T\left(\mathfrak{X}^T V^{-1}\mathfrak{X} - \mathfrak{J}^{-1}\right)^{-1}(d_j\hat{\varepsilon}^T)V^{-1} \\ & +\nabla_j^2 V^{-1}\mathfrak{X}\left(\mathfrak{X}^T V^{-1}\mathfrak{X} - \mathfrak{J}^{-1}\right)^{-1}\mathfrak{X}^T V^{-1}. \end{aligned}$$

Pour la perturbation du nœud le calcul de la matrice $\ddot{F}$ est similaire.
La matrice d'influence dépend uniquement des données et la somme de deux quantités : une première quantité $V^{-1}(d_j\hat{\varepsilon}^T)^T\left(\mathfrak{X}^TV^{-1}\mathfrak{X}-\mathfrak{J}^{-1}\right)^{-1}(d_j\hat{\varepsilon}^T)V^{-1}$ très proche des résidus studentisées et l'autre donnée par $\nabla_j^2V^{-1}\mathfrak{X}\left(\mathfrak{X}^TV^{-1}\mathfrak{X}-\mathfrak{J}^{-1}\right)^{-1}\mathfrak{X}^TV^{-1}$ nous rappelle une expression déja rencontrée : le produit de la hat matrice H du modèle mixte semiparamétrique par la matrice de variance. Il faut remarquer que la matrice d'influence est la somme de ces deux quantités par conséquent elle donne la sensibilité du modèle à chacune. La présence des résidus studentisés dans l'expression de la matrice d'influence permet d'affirmer que la perturbation sur les variables permet aussi de détecter les *outliers*. Cette approche par les perturbations des variables explicatives paraît donc plus étendue que celle de la variable réponse.

4.5 Analyse de sensibilité sur les paramètres de variance

Après ces deux schémas de perturbations effectués en considérant les paramètres de variance connus. Nous allons mettre en oeuvre cette analyse de sensibilité locale dans un cadre plus général, en considérant que les paramètres de variance sont aussi à estimer. Cela nous amènera à développer un schéma de perturbation permettant d'évaluer l'effet des perturbations des composantes structurelles du modèle mixte semiparamétrique sur les différents paramètres de variance. Dans ce cas, la logvraisemblace que nous allons considérer sera aussi une fonction des paramètres de variance correspondant (4.15).
La dérivation deux fois par rapport aux paramètres d'effets fixes ∇ et de variance θ donne un hessien de la forme :

$$\mathfrak{L}=\begin{pmatrix}\mathfrak{X}^TV_\theta^{-1}\mathfrak{X}-\mathfrak{J} & \mathfrak{X}^TV_\theta^{-1}\dot{V}_\theta V_\theta^{-1}(Y-\mathfrak{X}\nabla)\\ (Y-\mathfrak{X}\nabla)^TV_\theta^{-1}\dot{V}_\theta V_\theta^{-1}\mathfrak{X} & l_{\theta\theta}\end{pmatrix}, \tag{4.20}$$

où $l_{\theta\theta}=\frac{1}{2}tr(V_\theta^{-1}\dot{V}_\theta V_\theta^{-1}\dot{V}_\theta)-(Y-\mathfrak{X}\nabla)^TV_\theta^{-1}\dot{V}_\theta V_\theta^{-1}\dot{V}_\theta V_\theta^{-1}(Y-\mathfrak{X}\nabla)$.

$\mathfrak{L}$ va nous permettre d'appliquer la formule de la décomposition de la matrice d'influence afin d'évaluer la matrice d'influence partielle aux paramètres de variance.
Nous allons utilisé la méthode de décomposition présentée dans la section (4.3.3) pour le calcul de l'influence partielle par rapport aux paramètres de variance.

$$\ddot{F}_\theta=\Delta^T\left(\mathfrak{L}^{-1}-\begin{pmatrix}\ddot{L}^{-1} & 0\\ 0 & 0\end{pmatrix}\right)\Delta, \tag{4.21}$$

où $\ddot{L}^{-1}$ est l'inverse du hessien de la section (4.3.3).
Par ailleurs, dans le cas des schémas de perturbation de la variable Y, la matrice Δ sera de la forme :

$$\Delta=\begin{pmatrix}\mathfrak{X}^TV_\theta^{-1}\\ (Y-\mathfrak{X}\nabla)^TV_\theta^{-1}\dot{V}_\theta V_\theta^{-1}\end{pmatrix}. \tag{4.22}$$

Tandis que, le cas où la perturbation porte sur les variables explicatives équivaut à une matrice Δ de la forme :

$$\Delta = \begin{pmatrix} (d_j\hat{\varepsilon}^T) - \nabla_j \mathfrak{X} \\ \nabla^T V_\theta^{-1} \dot{V}_\theta V_\theta^{-1} (Y - \mathfrak{X}\nabla) \end{pmatrix}. \tag{4.23}$$

Ainsi décrit ce schéma de perturbation permet la mise en oeuvre d'une évaluation des effets des perturbations sur les paramètres de variance. Nous avons intégré le paramètre de lissage de la fonction nonparamétrique comme une composante de variance, ce schéma convient pour une analyse de sensibilité sur cette fonction nomparamétrique. En effet, l'approche du modèle mixte semiparamétrique stochastique comme modèle mixte modifié permet d'appliquer cette analyse de sensibilité au paramètre de lissage en le considérant comme un des paramètres de variance du modèle.
L'effet des perturbations sur le paramètre de lissage sera évalué alors dans ce cadre paramétrique ainsi défini.

Nous souhaitons rapprocher la méthode décrite à celle de Lesaffre & Verbeke [27]. En effet, ces derniers font une décomposition de l'influence locale du modèle linéaire mixte en différents blocs (paramètres de variance et paramètres de moyenne) suivant les individus de l'étude tandis que la méthode décrite ici consiste à faire cette décomposition suivant les composantes structurelles (fonction nonparamétrique, variables exogènes) du modèle mixte modifié.
Dans leur approche, Verbeke et Lesaffre évaluent les effets des perturbations suivant les individus. Ils utilisent la formule de décomposition de la matrice d'influence pour avoir les effets de perturbations sur les paramètres de la moyenne mais aussi des paramètres de variance du modèle mixte.
Ainsi selon leur approche, la matrice d'influence partielle propre à un paramètre quelconque est obtenue en remplaçant dans l'expression générale $\Delta^T A^{-1} \Delta$, la matrice A par la matrice convenable. L'expression contenant la matrice A est similaire à $\ddot{F}$.
Ils évaluent l'influence des différents paramètres par la norme de la matrice d'influence partielle celle-ci est déterminée principalement par celle de la matrice Δ associée. Ainsi cette norme dans le cas du paramètre du lissage de la fonction nonparamétrique intégré comme paramètre de variance du modèle mixte modifié se note :

$$\|B^T N^T V^{-1} N B - B^T N^T V^{-1} \hat{\varepsilon}\hat{\varepsilon}^T V^{-1} N B\|^2.$$

Par (4.21), nous n'obtenons certes pas la forme analytique finale de la matrice d'influence, mais cependant comme dans le cas de Lesaffre, cette dernière nous permet de représenter des graphiques de diagnostics utiles pour l'analyse de sensibilité locale.

Chapitre 5

Application sur des données réelles

5.1 Introduction

Après l'exposé du modèle sémiparamétrique mixte stochastique (chap. 3) et l'analyse de sensibilité développée dans ce modèle (chap.4), nous appliquons la méthode à des données réelles. La première application portera sur l'analyse des concentrations des nitrates provenant du bassin vesant du Gouet (Côtes d'Armor). Une deuxième application traitera les données de qualité des eaux de baignade des plages de la commune de Dinard (Ille et Vilaine).

5.2 Les données de concentration de nitrates dans les eaux superficielles : exemple du bassin versant du Gouet

Le suivi des concentrations des nitrates dans les eaux superficielles montre, qu'en France comme dans la plupart des pays d'Europe occidentale, ces teneurs augmentent. Cette évolution résulte de l'intensification de l'agriculture qui s'est accompagnée d'un usage massif des fertilisants d'origine animale ou de synthèse. De plus en Bretagne, l'épandage des déjections animales provenant des élevages porcins est aussi à l'origine des excès de matières azotées. En effet, cette région qui représente 7,7 % de la surface agricole française, concentre environ 22,5 % des productions animales dont 53 % de la production des porcs, 41 % de celle des volailles et 20 % de celle de lait (Cann, [1]).
Fournies par le Conseil Général et la DDASS des Côtes d'Armor, les données utilisées pour ce travail ont été instaurées dans le cadre d'un contrôle officiel de la qualité des eaux d'alimentation du département. Les données de concentration de nitrates dans les eaux superficielles présentent différentes spéficités. Ces données présentent un aspect longitudinal mais aussi spatial, résultat des différents points de mesure.

5.2.1 Problématique des nitrates

Problématiques générales

Avec un accroissement des teneurs de l'ordre de 1 à 3 mg/l/an , les valeurs en nitrates observées sur la quasi-totalité des prises d'eaux utilisées pour la production d'eau potable en Bretagne, dépassent aujourd'hui, sur une période plus ou moins longue dans l'année, la valeur limite de 50 mg/l fixée pour les eaux destinées à la consommation humaine.
Certes les risques pour la santé des nitrates et en particulier de ceux présents dans les eaux ont fait l'objet de réévaluations qui exonèrent presque totalement ces composés des risques de méthémoglobinémie chez le nourrisson ([24]) ou pour lesquels les risques de cancer ou d'effets sur la reproduction et le développement sont peu probables, l'obligation de respect de la valeur limite dans les eaux de distribution contraint les collectivités à importer des eaux de meilleures qualité ou à mettre en oeuvre des traitements de dénitratation onéreux.
Les valeurs élevées en nitrates sont l'un des facteurs de l'eutrophisation des eaux douces et de celles du littoral. L'eutrophisation des eaux gêne le traitement de potabilisation des eaux et dans le milieu marin est responsable du développement des marées vertes.
Le cadre réglementaire défini dans la directive n°91/676/CEE appelée «directive nitrate », a contraint les collectivités à engager des programmes d'actions sur les bassins versants afin de reconquérir la qualité des eaux. Si ces actions se sont accompagnées parfois de la stabilisation des teneurs en nitrates, les résultats sont dans l'ensemble décevants (Avery[3]) (Knobeloch, [24]). Cette situation peut s'expliquer par la seule participation aux programmes de reconquête des agriculteurs volontaires, par la fixation de délais trop brefs pour l'obtention de résultats (5 ans) et par les stocks de matière organique accumulée dans les sols.

Problématiques de l'analyse statistique des courbes des concentrations des nitrates

L'étude de ces données de concentration des nitrates vise à comprendre et analyser la variabilité des données puis à dégager la courbe des tendances, ses différentes saisonnalités, *etc.* Cela implique l'intégration des données disparates (non synchrones) avec des données manquantes. Pour ce type de phénomène dont la collecte des données se répartit sur plusieurs stations, plusieurs problématiques sont envisageables selon la dialectique choisie :

(a) : *Local/régional* Lorsque l'observation d'un phénomène environnemental se fait sur différents sites dont la source peut varier d'une localité à l'autre, l'un des premiers objectifs est de pouvoir comparer les niveaux des différents sites. Le lissage spatial sert à dégager une information «régionale » servant de référence pour positionner l'information «locale ». Cette comparaison peut se faire en regroupant les sites les plus similaires :les plus proches voisins, par exemple, ou d'autres critères de dissimilarités plus pertinents, emboîtements de vallées dans un réseau hydrographique.

(b) : *Inter-annuel/intra-annuel* Il est essentiel de noter que cette approche reste liée à l'aspect longitudinal des concentrations des nitrates. Ce sont des courbes d'évolution. Une étude du comportement de ces courbes sur un an, avec une périodicité des mesures adaptée, est une échelle permettant de déceler des effets de périodicités très courtes tandis

que d'autres ne sont perceptibles qu'à long terme.

(c) : *Méthode d'analyse intégrée* L'approche intégrée permet une analyse intégrant l'ensemble des observations. Il existe des approches utilisant des méthodes dérivant de l'analyse en composantes principales. Ces deux approches ne sont pas forcément à comparer mais peuvent parfois être perçues comme complémentaires. L'analyse des données est alors considérée comme une analyse exploratoire des données.

(d) : *Reconstruction des profils des stations de mesure* Enfin, pour des stations avec très peu de mesures, l'intérêt de l'étude peut se porter sur la reconstruction de leur profil de courbe d'évolution, en exploitant de l'information provenant des autres sites.

L'un des points essentiels est de pouvoir évaluer l'effet des variables météorologiques sur le phénomène étudié. Dans les études comparées décrites ci-dessus, la part expliquée par les variables ne peut pas toujours être correctement évaluée. Nous pensons que l'un des apports du modèle mixte sémiparamétrique stochastique est de permettre cette évaluation sans aucune ambiguïté vis vis des autres composantes du modèle. Dans un premier temps, nous allons développer une approche que nous qualifierons de *naïve*. Elle s'appuie sur un ensemble d'outils classiques d'analyse des données dont l'utilisation dans le contexte peut masquer certains aspects des données. En deuxième approche, nous appliquerons le modèle mixte sémiparamétrique stochastique.

5.2.2 Approche *naïve*

Les prélèvements des nitrates se présentent sous forme de données dont la variabilité peut être liée à diverses raisons. Le pas de temps qui sépare deux prélèvements ainsi que la méthode de mesure utilisée (mesure manuelle ou automatique), sont très variables d'une part, au sein d'une même station et d'autre part, d'une station à une autre. Ainsi, les sources d'erreurs des données peuvent être multiples : soit liées aux corrélations des mesures entre elles, soit liées aux différentes méthodes de mesure des données ou encore d'autres sources. L'analyse exploratoire des séries temporelles a été essentiellement basée sur la représentation graphique de ces séries. Elle permet de visualiser l'évolution temporelle du phénomène étudié. Les représentations graphiques des séries de nitrates mettent en évidence l'irrégularité à la fois dans la fréquence et le pas de temps des relevés. Cette irrégularité que nous avons évoquée comme l'une des spécificités des données. En guise de préambule à l'approche naïve, nous allons procéder à un recalage des données.

Algorithme de recalage temporel

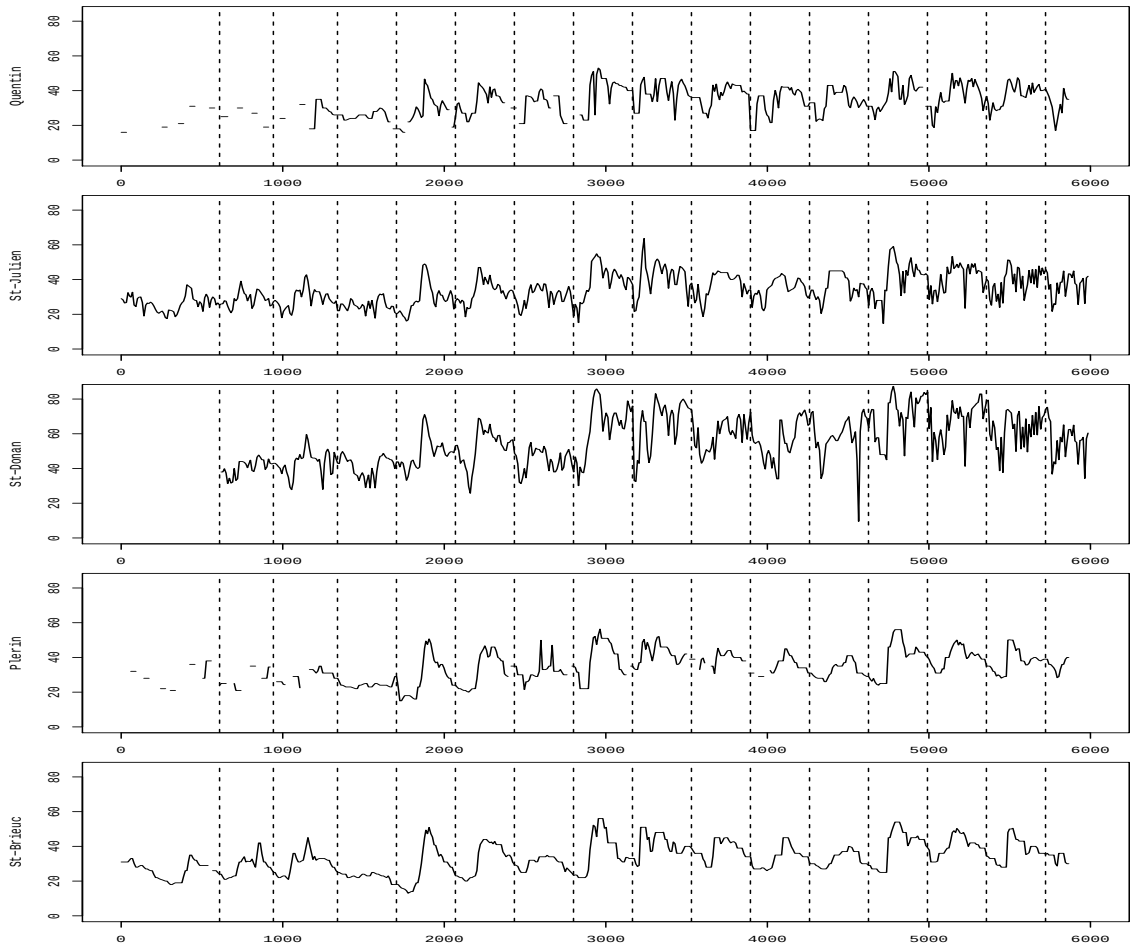
L'objectif de ce recalage est d'interpoler les données brutes et de produire une estimation des valeurs de nitrate sur une grille de temps régulière identique à tous les sites et, de permettre le suivi au cours du temps du phénomène étudié pour détecter les éventuels accidents par rapport au comportement général de la série. La partie essentielle du recalage est le choix du *pas de temps* (intervalle de temps entre deux dates de prélèvement). Afin d'utiliser les données météorologiques à un niveau susceptible d'influencer le phénomène étudié, en l'occurence la modélisation des nitrates, le pas de temps retenu est de 10 jours. Un algorithme de construction de la fonction de recalage est élaboré. Pour cela, il faut tout d'abord définir une date de début unique pour tous les cours d'eau d'un même bassin versant. Puis, un calendrier théorique est construit, constitué d'une suite de dates séparées de 10 jours (approximativement). Les dates réelles de prélèvement peuvent être comparer avec celles du calendrier construit. A chaque date du calendrier théorique, la valeur de prélèvement la plus proche dans le temps ou une interpolation de la valeur de ce prélèvement lui est associée.

Découpage en années hydrologiques

Pour étudier le comportement (tendance, cycle ou saisonnalité) des séries temporelles, les données ont été «découpées» en années hydrologiques. Une année hydrologique commence approximativement au début du mois de Septembre. Les graphiques (5.1) mettent en évidence à la fois l'existence d'une saisonnalité pour chaque série temporelle et la similitude du comportement entre chaque série temporelle. (La date de référence notée comme date 0 est la première date de mesure. Cette date correspond au 10 Mai 1983). Le phénomène saisonnier, équivalent à une année hydrologique, peut être décrit de la façon suivante : en début d'année hydrologique et jusqu'au mois de décembre, le niveau des nitrates est faible et atteint sur cette période le minimum de l'année. Vers la fin du mois de Janvier, le taux des nitrates augmente progressivement et atteint le maximum en Février - Mars. Ce niveau

très élevé est interrompu au mois d'Avril par une légère baisse suivie d'une période de nouveau élevée des nitrates qui se maintient jusqu'au mois de Juin -Juillet. Cette interruption confère à la série l'aspect de deux zones de pics. Les données manquantes, sont importantes pour Quintin et Plérin (10%) et visibles surtout au début de la période d'étude. Le taux de données manquantes est d'environ 2% pour Saint Brieuc et nul pour Saint Donan et Saint Julien.

FIG. 5.1 – Représentation graphique de 5 courbes des données recalées de concentration du bassin versant du Gouet



Corrélation linéaire entre les différentes courbes

De forts coefficients de corrélation (cf. tableau : 5.1) laissent à penser qu'il existe une dynamique commune aux 5 stations traduisant mieux le comportement à l'échelle du bassin versant. La date de début a été réajustée à la première date de mesure de Saint Donan.

	Saint-Donan	Saint-Julien	Saint-Brieuc	Quentin	Plerin
Saint-Donan	1				
Saint-Julien	0,82	1			
Saint-Brieuc	0,68	0,78	1		
Quentin	0,69	0.80	0,75	1	
Plérin	0,68	0,74	0,93	0,74	1

TAB. 5.1 – Tableau de corrélations linéaires entre les sites du Gouet

Le graphique (5.1) met en évidence le comportement similaire des différentes courbes. Les pics et les creux coïncident assez remarquablement. Le niveau des nitrates dans les eaux de Saint Donan est systématiquement supérieur au niveau des autres stations de mesure : cette différence peut s'expliquer par une dynamique locale propre à la station. Les analogies relevées entre les différentes stations de ce même bassin versant sont représentées par la suite par «composante commune» est considérée comme la dynamique propre aux différentes stations du bassin. La suite de cette approche *naive* consistera à identifier cette composante commune et d'en caractériser l'allure de la courbe.

Recherche de la composante commune

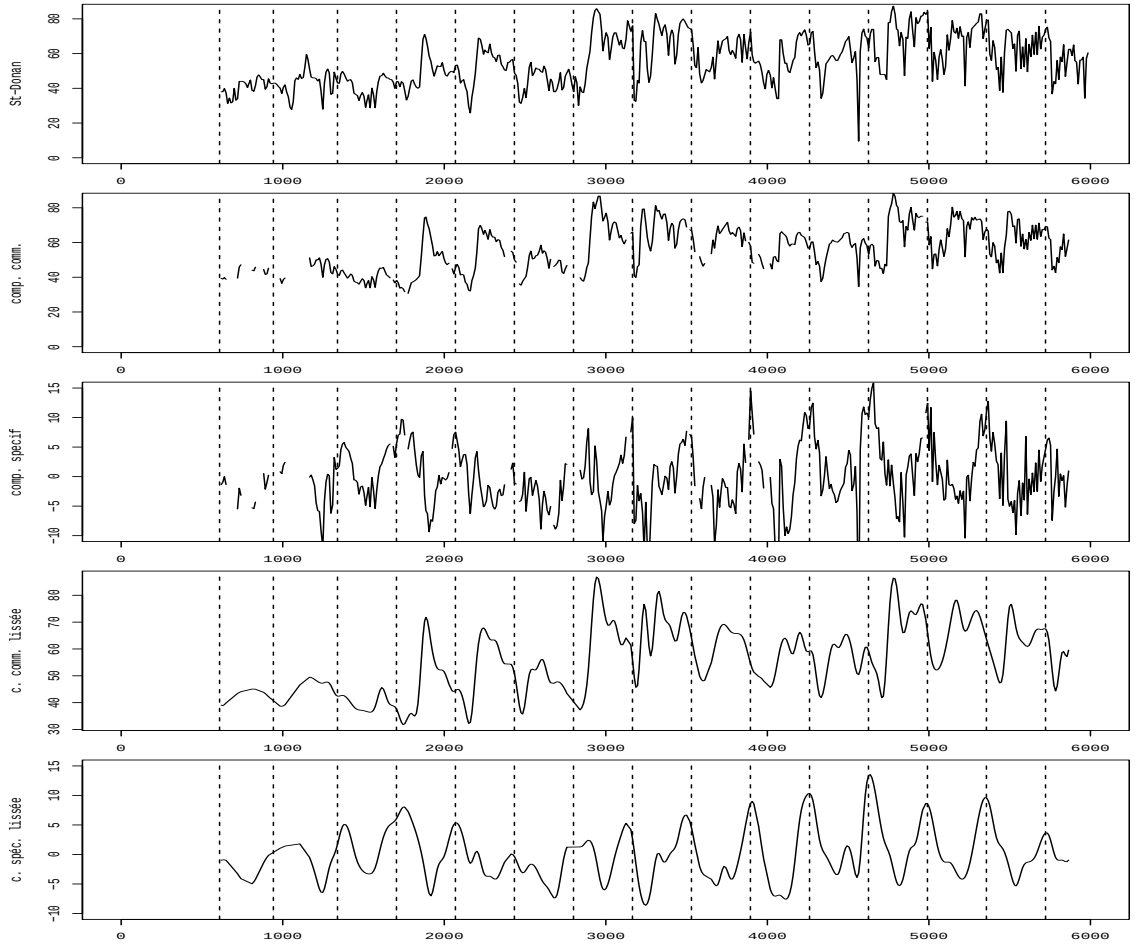
L'existence d'une dépendance linéaire forte entre les différentes stations peut s'expliquer à travers une dynamique commune. La recherche de variables latentes, en particulier d'une composante susceptible de caractériser cette dynamique commune s'impose. Un outil possible pour prendre en compte ce type d'analyse est l'analyse factorielle linéaire «*The Normal Linear Factor Model*». Les cinq courbes obtenues après le recalage aux instants t_j, $j = 1, \ldots, N$ sont considérées comme les variables de départ sur lesquelles va être développée la méthode. Cette dernière revient à appliquer un modèle factoriel *à un facteur* de la forme :

$$Y_i(t_j) = \beta_0 + \lambda_i f(t_j) + \varepsilon_i(t_j) \qquad i = 1, \cdots, 5 \quad j = 1, \cdots, N. \tag{5.1}$$

où $Y_i(t)$ représente le processus observé à la station i et $\varepsilon_i(t)$ un composante spécifique aléatoire (bruit blanc gaussien stationnaire) à la station i, de variance ψ_i^2 (appelée *variance spécifique*). Les termes $\lambda_i \;\; i = 1, \cdots, 5$ constituent les *contributions factorielles* ("loadings" en anglais) sur chacun des sites de mesure. Ils expriment le fait que la composante commune $f(t)$ "se réalise" selon une dynamique (ou amplitude) différente d'un site à l'autre. (λ_i est facteur d'échelle).
La composant $f(t)$ est tout simplement estimée par une ACP réalisée sur le tableau $N \times 5$ des mesures recalées : $\hat{f}(t)$ correspond à la première composante principale, celle associée à la plus grande valeur propre de la matrice de covariance empirique des données recalées. La composante $\hat{f}(t)$ ainsi construite reccueille la plus grande part possible de variance de la diffusion des nitrates mesurée sur l'ensemble des sites disponibles. A ce titre, elle constitue un *lissage spatial* et une *standardisation* des données prélevées sur le bassin versant du Gouet.

Les mesures de concentration de nitrates des cinq courbes constituent les différents points du nuage dont les principaux axes d'inertie vont être déterminés.

FIG. 5.2 – Représentation comparative de le composante commune et du profil individuel de Saint Donan



Le graphique (5.2) donne une représentation de la courbe des observations de la station de Saint Donan, sa composante spécifique (comp. spécif.) ainsi que la composante commune (comp.com.) et, aussi la courbe de la différence entre la série de nitrate de Saint Donan et la composante commune.
Cette composante commune repésente la dynamique commune de la diffusion des nitrates à l'échelle du bassin versant, par conséquent sa modélisation constitue la suite de cette approche.
Pour mieux appréhender l'allure de la composante commune et afin d'obtenir une courbe plus simple et des mouvements cycliques facilement perceptibles, nous avons appliqué un lissage sur la composante commune (cf. 5.2).

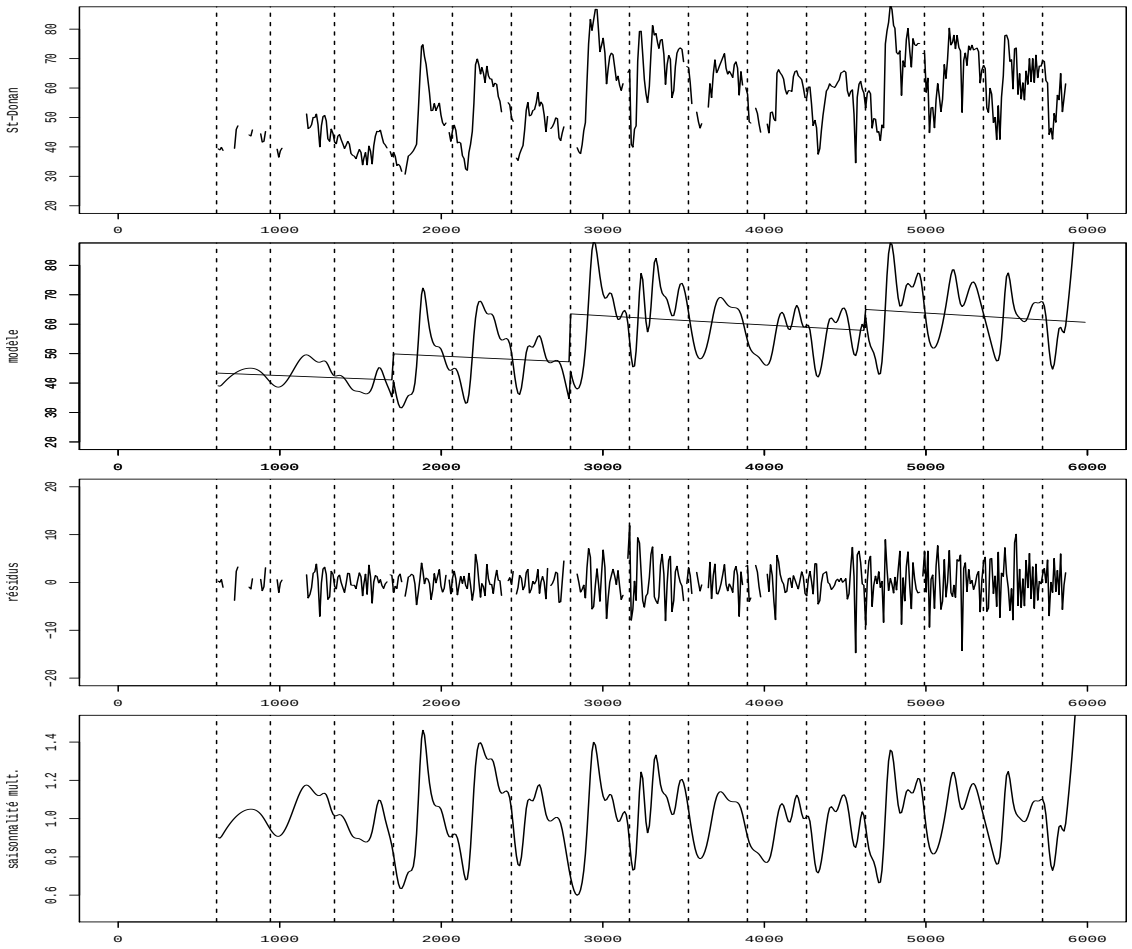
Estimation de la tendance

Le phénomène de diffusion des nitrates est très marqué à travers l'allure de sa courbe par les différentes ruptures dans la tendance coïncidant avec les années de sécheresse. Afin d'extraire la tendance de la série, une estimation précise de ces ruptures va être réalisée en identifiant tout d'abord les différentes périodes inter sécheresses :

Du 12 Septembre 1986 au 22 Août 1989 pour la première période
Du 2 Septembre 1989 au 22 Août 1992 pour la seconde période
Du 2 Septembre 1992 au 22 Août 1997 pour la troisième période
Du 2 Septembre 1997 au 22 mai 2001 pour la quatrième période

A chacune de ces périodes équivaut une tendance en dents de scie qui décroît au fil du temps. Cette diminution progressive est interrompue lorsqu'une autre année de sécheresse est observée provoquant une rupture de tendance. Cette rupture de tendance se caractérise par une augmentation brutale du niveau des nitrates dans l'eau suivie de nouveau par une diminution linéaire et progressive (et ainsi de suite se répète ce phénomène rythmé par les années de sécheresse). Une tendance unique a été estimée sous l'hypothèse d'une droite de même coefficient directeur pour les 4 périodes identifiées. La tendance se compose des tendances en dents de scie liées aux différentes ruptures auxquelles s'ajoutent la moyenne de la variation saisonnière de la série. La tendance paraît être gouvernée par la dynamique locale propre à la station et aux perturbations météorologiques exceptionnelles. A partir de l'estimation de la tendance, par soustraction à la transformée logarithmique de la composante commune lissée, on obtient une bonne estimation de la composante saisonnière. L'évolution temporelle de cette composante saisonnière (cf. figure 5.3) suit une trajectoire approximativement identique chaque année qui traduit le cycle régulier de la diffusion des nitrates en dehors de toute perturbation ou accident dus à des conditions météorologiques particulières.

FIG. 5.3 – Décomposition de la composante commune en Tendance + Saisonnalité + Résidus



Application du modèle GAM et utilisation des variables météorologiques

La composante saisonnière ainsi obtenue malgré le caractère cyclique de la courbe (cf. 5.3) ne s'ajuste pas suffisamment sur un modèle de type SARIMA. Ces variations peuvent être dues à des phénomènes climatiques qui ne peuvent pas, par définition, être pris en compte à travers une modélisation de type autorégressif. Des variables météorologiques susceptibles d'améliorer le modèle vont être introduites dans un modèle de régression GAM pour expliquer le phénomène étudié.

Les covariables environnementales La pluviométrie est une variable connue, susceptible d'influencer le phénomène de diffusion des nitrates. Des méthodes de calcul (Méthode de Tisson) permettent d'estimer de façon plus précise la pluie tombée sur le bassin versant à partir de données de pluviométrie provenant de stations encadrant celui-ci. Concernant le bassin du Gouet, il n'existe pas de données suffisamment complètes pour appliquer ce type de méthode. La régression a donc été réalisée sur les données de pluviométrie d'une seule station celles de Saint Donan : ce sont des mesures journalières de Janvier 1989 à Août 2001. Ces données ont été cumulées sur 10 jours correspondant au pas retenu. Les données d'ETP (évaporation transpiration potentielle) ont été recueillies sur la station de Quintenic disponibles sans valeur manquante à partir de Mars 1994. L'analyse par régression (GAM) a donc été restreinte à la période de Mars 1994 à août 2001.
A partir des connaissances hydrogéologiques et des deux covariables pluviométrie et ETP, d'autres covariables peuvent être construites. A l'issue de ces calculs, trois covariables supplémentaires (l'évapo-transpiration réelle couramment notée ETR, la réserve du sol et l'écoulement) pourront être introduites dans le modèle comme variables explicatives. Les séries de nitrates utilisées dans le calcul des corrélations sont la composante commune du bassin versant du Gouet ainsi que sa composante saisonnière afin de sélectionner la série à expliquer la plus adaptée pour cette régression.

	Comp com	saisonière	Pluvio	ETP	ETR	Réssol	Ecoul
Comp com	1						
saisonière	0,77	1					
Pluvio	-0,10	-0,27	1				
ETP	0,35	0,56	-0,16	1			
ETR	0,34	0,56	-0,16	0,999	1		
Réser	-0,11	-0,24	0,30	-0,26	-0,35	1	
Ecoul	-0,13	-0,30	0,996	-0,22	-0,22	0,3	1

TAB. 5.2 – Tableau des corrélations entre variables environnementales avec ETP=Evapo-transpiration potentielle, Pluvio= pluviométrie, ETR= Evapo-Transpiration Réelle, Réserv= réserve du sol et enfin Ecoul = Ecoulement

Modèles additifs (GAM) L'application d'un modèle de régression linéaire sur la composante commune ne permet pas un meilleur ajustement des données. Les modèles additifs et en général les modèles additifs généralisés (GAM), beaucoup plus souples que le modèle

linéaire classique, vont être développés sur les données de nitrates. Tout en conservant les hypothèses usuelles du modèle de régression linéaire, le modèle additif généralisé permet l'introduction des transformations nonparamétriques des covariables du modèle. Cette intrusion permet au modèle additif de découvrir certains aspects non linéaires ou nonparamétriques des observations. L'observation des représentations graphiques des transformations des covariables apporte beaucoup à l'interprétation de la relation traduite par le modèle GAM. Comme dans la plupart des modèles de régression, l'introduction des variables pas à pas (stepwise) permet de choisir le modèle «optimal ». Le tableau (5.3) résume ce choix pas à pas.
Le modèle retenu est le modèle $M\,4$ qui est une régression sur les transformées de la pluviométrie et de l'ETP. En considérant comme acceptable un nivau de significativité de $0,11$, le modèle M4 est retenu comme le meilleur (score GCV le plus faible). Plus de 50% de la variabilité est expliquée par ce modèle ($R^2 = 50\%$) dans lequel sont inclues les variables ETP et pluviométrie.

Modèles	Variables transfomées	RSS.	ddl	R^2	GCV	seuil de significativité
M1	ETP	2,88	11	47%	0,013	10^{-16}
M2	Pluviométrie	4,60	11	15,4%	0,02	10^{-8}
M3	Réserve du sol	5,23	11	4%	0,023	0,035
M4	ETP & pluvio	2,55	23	50,7%	0,012	0,11
M5	ETP & Pluvio & réserv du sol	2,47	34	49,7%	0,013	0,81

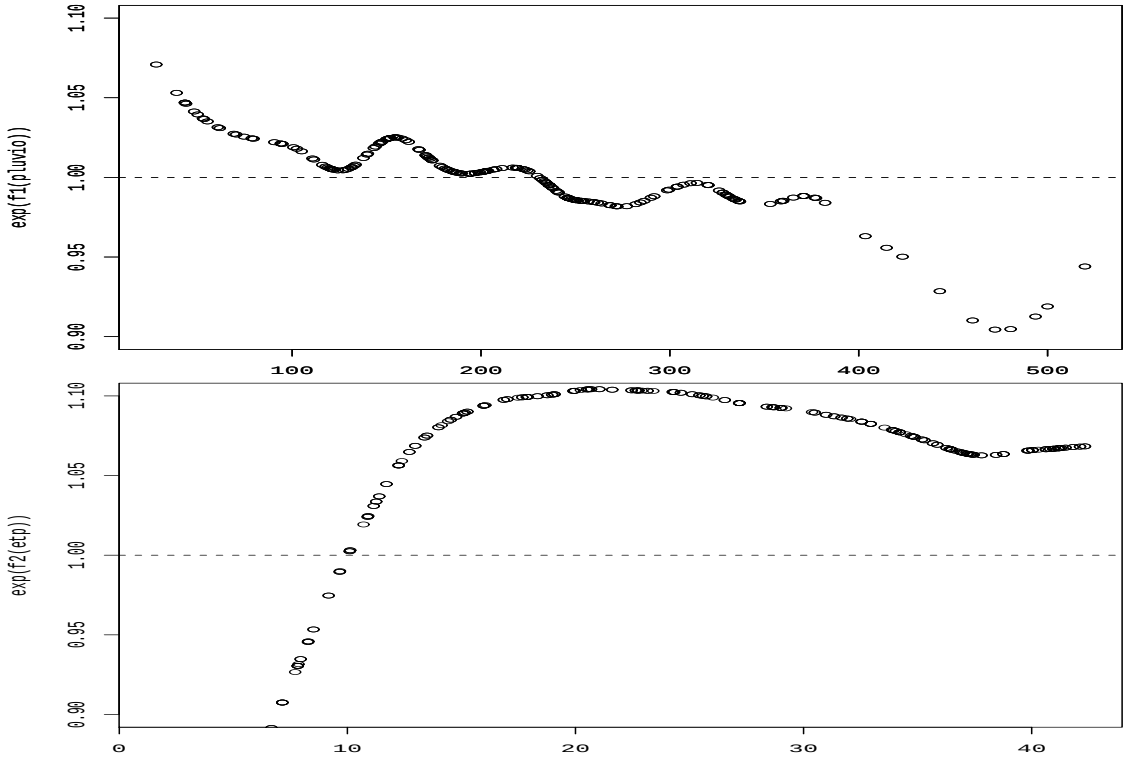
TAB. 5.3 – Tableau des critères de choix du modèle GAM avec ETP=Evapo-transpiration potentielle, Pluvio= pluviométrie, ETR= Evapo-Transpiration Réelle, Réserv= réserve du sol et enfin Ecoul = Ecoulement, ddl=dégré de liberté donné par le package mgcv de *R*.

Pour interpréter les différents résultats du modèle, nous allons faire les représentations graphiques des transformées de l'ETP et de la pluviométrie d'autre part :

★ Pour la forme prise par la variable «pluviométrie», la courbe comporte une alternance de pics et de creux. Cela pourrait corrrespondre aux différents niveaux selon lesquels, les effets de la pluviométrie sont perceptibles sur les courbes des concentrations des nitrates. Le premier pic «précoce » peut être dû aux effets des ruissellements. Cet effet s'estompe quand la pluie s'intensifie. Les charges de polluants en surface sont balayées par le premier lessivage du sol effectué par le ruisselement.
Les deux pics observés qui suivent sont attribuables aux deux niveaux suivants : Subsurface et nappe phréatique qui agissent à travers l'infiltration de la pluie à travers les sols. Ces deux niveaux peuvent agir conjointement ou de façon dissocié. A une certaine quantité de pluie tombée, l'effet de la pluie dimunie les concentrations, on note un effet de dilution dans l'allure globale de cette transformée de la pluviométrie.

★ La variable ETP (corrélation de $0,34$ avec la composante saisonnière des nitrates)(cf.5.4) donne une courbe à palier (à seuil) qui croit pour les faibles valeurs de l'ETP jusqu'à 20. Pour des valeurs supérieures à 20, on remarque une certaine saturation de l'effet de l'ETP sur nitrates. A partir de ce seuil, l'augmentation de l'ETP n'a plus d'effet sur la concentration des nitrates dans les eaux superficielles.

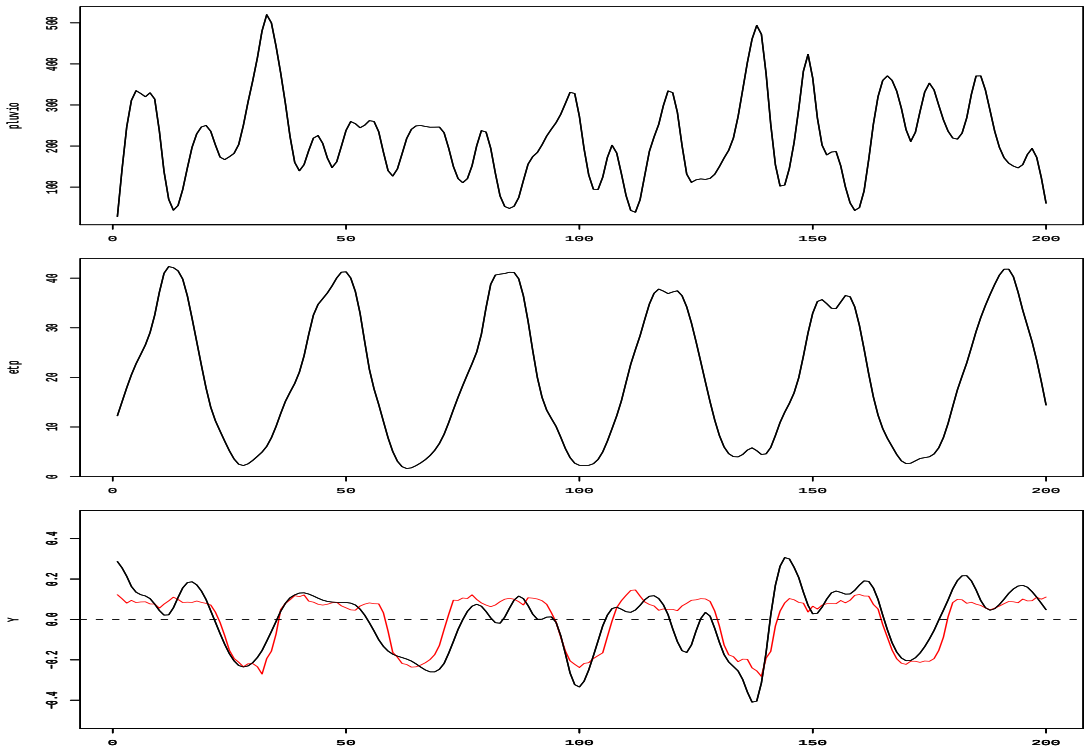
FIG. 5.4 – Représentation graphique des transformées des variables



Discussion

Malgré cet apport certain de la modélisation GAM, toute la dynamique de la diffusion des nitrates n'est pas expliquée par cette modèlisation comme le montre la représentation graphique de l'ajustement des prédictions (en rouge, figure. 5.5) de ce modèle aux données réelles. D'autre part, il faut noter que les pics ainsi que les zones de creux ne sont pas complètement appréhendés par le modèle.
L'utilisation de la fonction de recalage donne le même poids aux données complétant les observations, obtenues par recalage qu'aux observations initiales porteuses de toute l'information. Intuitivement, une méthode accordant un poids supérieur à ces observations initiales paraîtrait meilleure que l'approche par recalage des données. C'est pourquoi, nous allons appliquer de façon contextuelle le modèle sémiparamétrique mixte stochastique aux données des concentrations des nitrates.

FIG. 5.5 – Courbe estimée par GAM (rouge) ajustée à la courbe de la composante saisonnière



5.2.3 Application du modèle sémiparamétrique mixte stochastique aux données des concentrations des nitrates

La première partie a permis de découvrir les aspects essentiels des données des concentrations des nitrates. L'application du modèle mixte sémiparamétrique stochastique s'appuiera sur cette première analyse.

L'application du modèle mixte sémiparamétrique stochastique dans un premier temps se fera dans une approche similaire à celle décrite dans l'approche dite «naïve» qui peut être considérée comme une analyse exploratoire des données en vue de l'application de cette seconde partie.

Dans la partie précédente, nous avons noté que la composante commune des concentrations des nitrates du Gouet présente une dynamique avec des ruptures remarquables sur la période d'observation. Ces ruptures déterminent les courbes d'évolution des concentrations des nitrates et laissent apparaître que raisonnablement il existerait un modèle mixte sémiparamétrique stochastique pour chaque partie de la période d'observation délimitée par ces ruptures. Cette approche est justifiée par le choix des processus stochastiques stationnaires $U_i(t)$ pour modéliser la corrélation sérielle. Les ruptures remarquées introduisent éventuellement des structures de corrélation non stationnaires. Nous allons donc scinder la période d'observation en autant de parties déterminées par les ruptures notées dans l'approche dite naïve. On note ainsi quatre zones. Pour chaque zone, la composante commune

sera déterminée ainsi que les composantes spécifiques propres à chaque station de mesures.

Détermination des paramètres du modèle mixte sémiparamétrique stochastique

La mise en oeuvre du modèle mixte sémiparamétrique stochastique nécessite le choix de certains paramètres. Les critères de choix de ces paramètres ont été exposés dans le chapitre 3. Le type de modèle mixte sémiparamétrique stochastique retenu pour être appliqué aux données est de la forme :

$$Y_i(t_{ij}) = \beta_0 + f(t_{ij}) + b_i + U_i(t_{ij}) + \epsilon_{ij}. \tag{5.2}$$

C'est un modèle ne comportant pas de variable explicative, le coefficient β_0 représente la constante (*intercept*) du modèle. La fonction $f(t)$, dans ce cas précis, permet de captuer les fluctuations des données exhibées par la composante désignée comme étant une composante saisonnière dans l'approche «naïve ».
Les différents modèles candidats se distinguent les uns des autres par un nombre de noeuds différent de la fonction $\tilde{f}(t)$ approximation de $f(t)$ dans la base de fonctions splines. Pour le choix du modèle, la méthode retenue est le critère de la logvraisembance REML. Nous considérons que le meilleur modèle est celui pour lequel la logvraisemblance se «stabilise » tout en conservant un degré de liberté (Df-fit) faible. C'est une méthode ad-hoc qui nous paraît ici convenable car la structure des effets fixes est rigoureusement identique d'un modèle à l'autre.

nb de noeuds de $f(t)$	10	20	30	50	100	200	260
log. REML	-224,9	-249,2	-252,3	-260,9	-254,3	-260,1	-261,6
Df-fit	7,0459	18,3877	24,7676	38,9735	62,2777	103,8990	125,9798

TAB. 5.4 – Tableau des performances des différents modèles candidats

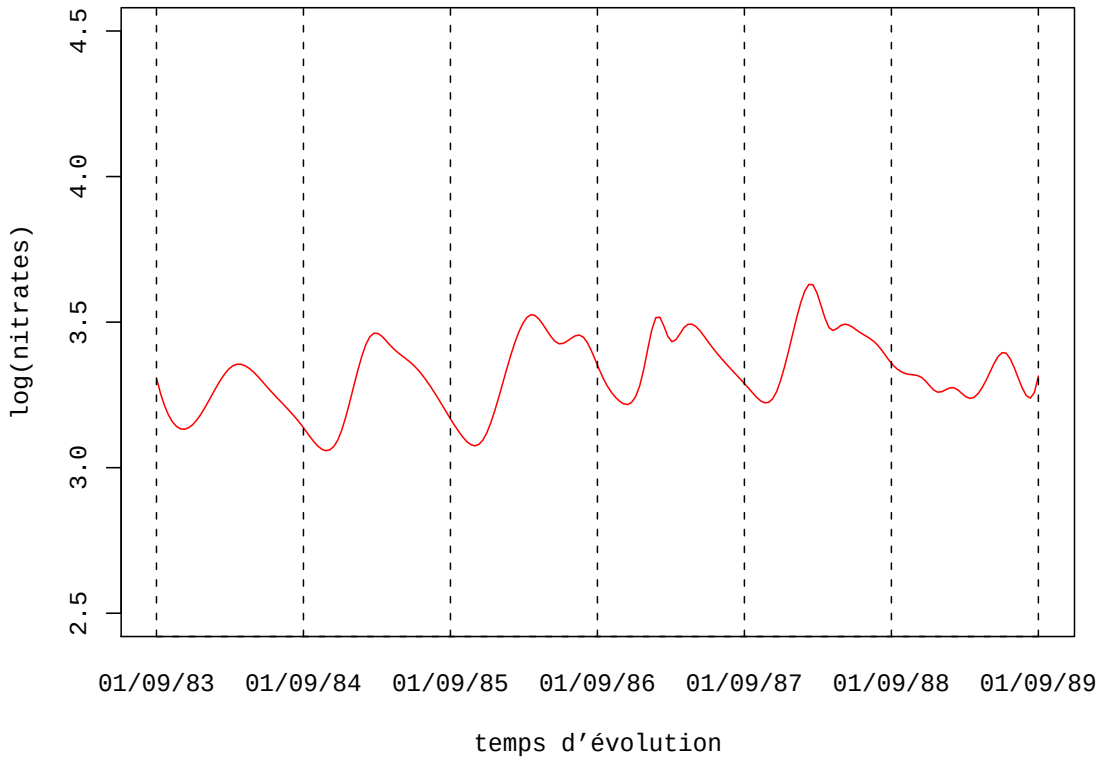
En examinant le tableau (5.4), on constate que la logvraisemblance REML n'augmente pas de façon monotone lorsque le nombre de noeuds croît, cela est sans doute dû au fait que les modèles candidats ne sont pas rigoureusement emboîtés.
Il aurait sans doute été plus facile et plus naturel, comme le suggèrent certains auteurs, de construire le tableau ci-dessus sur la base de la logvraisemblance et non la logvraisemblance REML et faire la sélection du modèle par des critères AIC et BIC.
Au regard des résultats du tableau (5.4), les modèles à 30 et 50 noeuds sont les plus convenables. Par souci de parcimonie, le modèle à 30 noeuds a été mis en oeuvre sur les données. Ce choix coïncide quasiment avec le nombre de noeuds recommandé (Ruppert et *al*, [35]) pour des donnnées de cette taille.

Période de Mai 1983 **à Septembre** 1989

Cest le début de la période d'observation. Cette première zone est sutout marquée par l'absence d'observations sur certains sites. L'allure générale au sein du bassin versant pour cette période est donnée par la composante commune de la figure (5.6). Le modèle étant

appliqué aux données transformées (logarithmiques), la composante restituée est représentée à une échelle logarithmique. Sur les deux années de Septembre 1985 à Septembre 1987, ayant suffisamment d'observations, on notera que la courbe d'évolution des nitrates présente deux pics sur une année hydrologique sur la dynamique commune à l'ensemble des cinq stations. Cette composante commune ainsi,aide à la détermination de la courbe spécifique de l'évolution des concentrations des nitrates pour un site donné. Les différents paramètres estimés sont reportés dans le tableau (5.5)

FIG. 5.6 – Composante commmune de la première période : du 01/Septembre/1983 au 02 Septembre 1989



Type de paramètre	Notation	Estimation	Erreur-type	P.valeur
Intercept	$\hat{\beta}_0$	3,332	0,09	0,00
variance de b	$\hat{\sigma}_*^2$	0,04	0,025	$5,64.10^{-2}$
variance de a	$\hat{\tau}^2$	0,0353	0,01	$6,62.10^{-4}$
paramètre γ_1^2 de $U_1(t)$	$\hat{\gamma}_1^2$	0,0293	0,01	$1,55.10^{-3}$
paramètre γ_2^2 de $U_2(t)$	$\hat{\gamma}_2^2$	0,0269	0,004	0,00
paramètre γ_3^2 de $U_3(t)$	$\hat{\gamma}_3^2$	0,0284	0,0054	0,00
paramètre γ_4^2 de $U_4(t)$	$\hat{\gamma}_4^2$	0,0134	0,0054	$7,33.10^{-3}$
paramètre γ_5^2 de $U_5(t)$	$\hat{\gamma}_5^2$	0,0188	0,0053	$2,56.10^{-4}$
Variance des résidus	$\hat{\sigma}_0^2$	0,0001291	0,0012	$4,58.10^{-1}$
paramètre α_1 de $U_1(t)$	$\hat{\alpha}_1$	3,36	2,146	$5,92.10^{-2}$
paramètre α_2 de $U_2(t)$	$\hat{\alpha}_2$	13,8	4,785	$2,09.10^{-3}$
paramètre α_3 de $U_3(t)$	$\hat{\alpha}_3$	14,34	7,17	$2,31.10^{-2}$
paramètre α_4 de $U_4(t)$	$\hat{\alpha}_4$	1,72	1,02	$4,55.10^{-2}$
paramètre α_5 de $U_5(t)$	$\hat{\alpha}_5$	2,62	1,112	$5,76.10^{-3}$

TAB. 5.5 – Les paramètres estimés du modèle ajusté à la période de Mai 1983 à Septembre 1989

Nous constatons que pour cette première période tous les paramètres sont significatifs. En particulier le paramètre τ^2 du modèle estimé est significatif, montrant ainsi l'utilisation d'une fonction nonparamétrique convenable. Sachant que la matrice G est choisie égale à l'identité, les prédictions du paramètre b sont :

	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
valeurs	-0,136	-0,085	0,34	-0,041	-0,082

TAB. 5.6 – Les prédictions de b

Les composantes spécifiques estimées que sont les processus stochastiques $U_i(t)$ s'ajoutent à la composante commune pour donner les courbes spécifiques. En l'absence, d'observations sur un site considéré comme c'est le cas sur cette période où les mesures des concentrations des nitrates n'ont pas démarré en même temps sur l'ensemble des 5 sites qui nous intéressent. La composante spécifique est quasiment nulle (cf. figure 5.7).
En plus, de la constante aléatoire (random intercept) b_i estimée par site, ces composantes spécifiques s'ajoutent à la composante commune pour donner les courbes d'évolution des nitrates par station de mesures. (voir figure 5.8)

FIG. 5.7 – Composantes spécifiques $U_i(t)$ des 5 sites

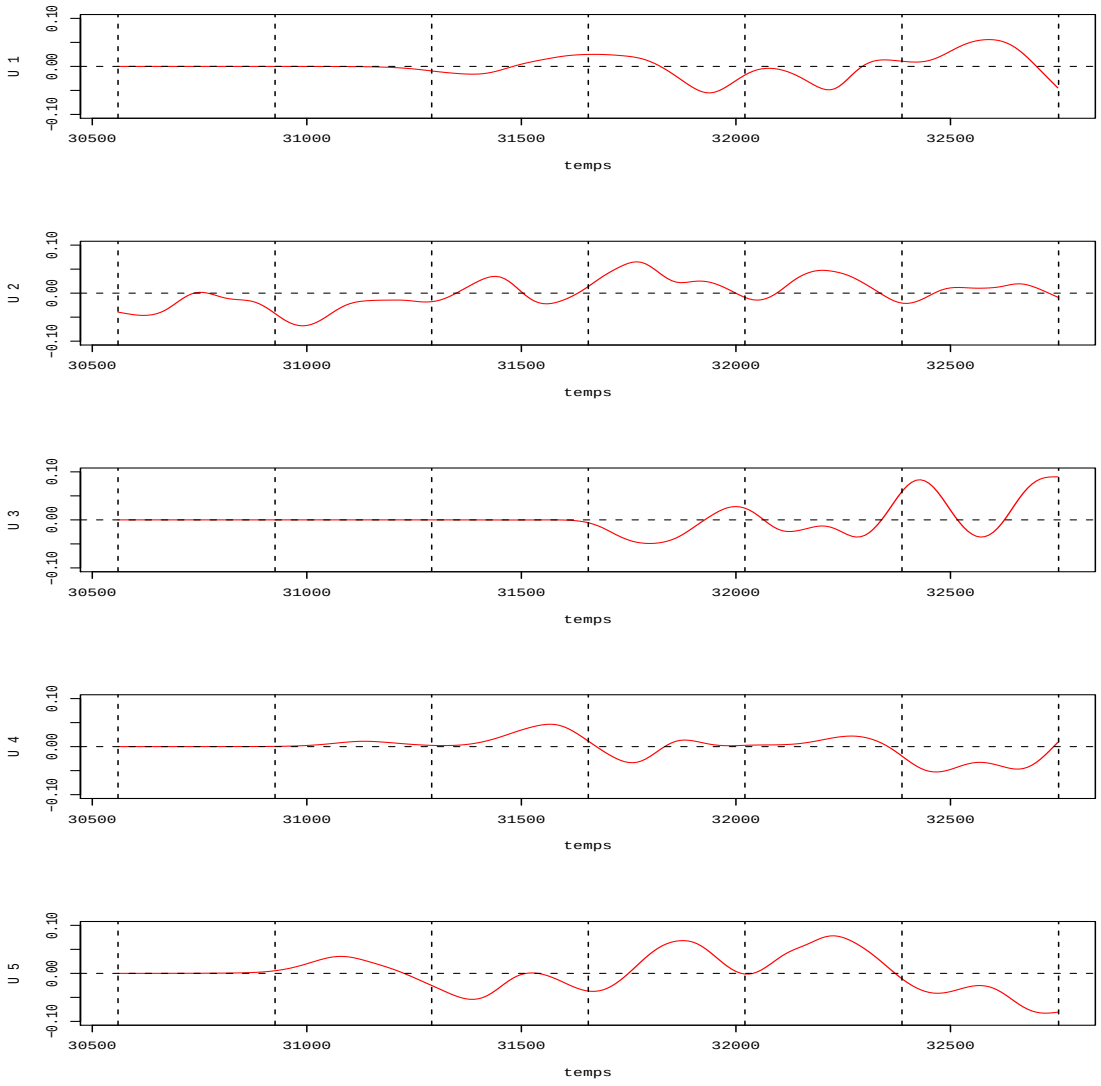
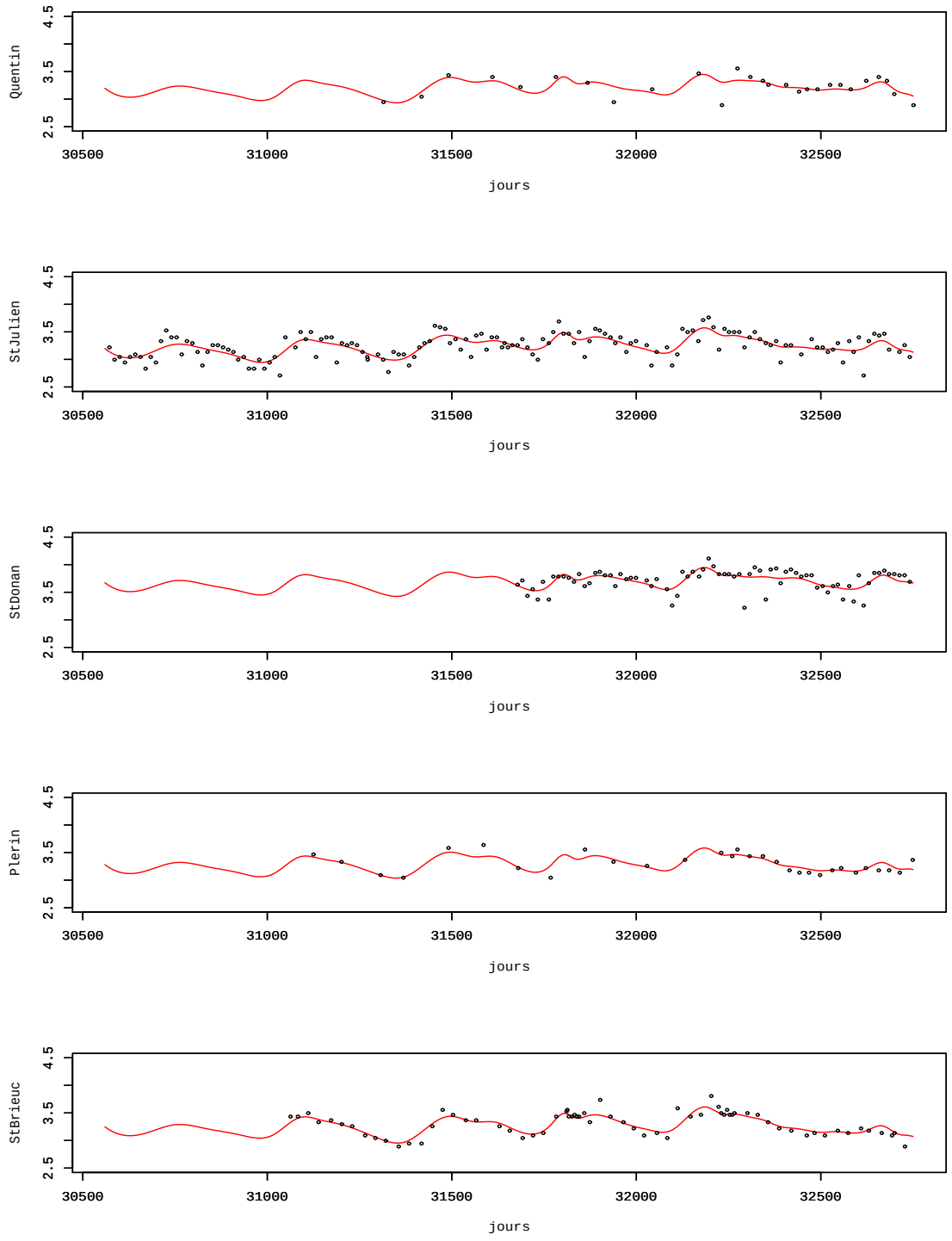


FIG. 5.8 – Courbes d'évolution des 5 stations de mesure avec les observations d'origine (points noirs)



Période de Septembre 1989 **à Septembre** 1992

Aprés application du modèle mixte sémiparamétrique stochastique, les résultats obtenus sont proches à ceux du pargraphe précédent. La différence entre les deux modèles n'est pas significative au niveau des paramètres estimés. Ces paramètres sont reportés dans le tableau ci-dessous :

Type de paramètre	Notation	Estimation	Erreur-type	P.valeur
Intercept	$\hat{\beta}_0$	3,4643	0,105	0,00
variance de b	$\hat{\sigma}_*^2$	$5,173.10^{-2}$	$3,05.10^{-2}$	$4,55.10^{-2}$
variance de a	$\hat{\tau}^2$	$6,74.10^{-2}$	$6,82.10^{-2}$	$1,30.10^{-4}$
paramètre γ_1^2 de $U_1(t)$	$\hat{\gamma}_1^2$	$2,84.10^{-2}$	$8,86.10^{-3}$	$7,50.10^{-4}$
paramètre γ_2^2 de $U_2(t)$	$\hat{\gamma}_2^2$	$2,34.10^{-2}$	$4,79.10^{-3}$	0,00
paramètre γ_3^2 de $U_3(t)$	$\hat{\gamma}_3^2$	$2,17.10^{-2}$	$5,64.10^{-3}$	$7.20.10^{-5}$
paramètre γ_4^2 de $U_4(t)$	$\hat{\gamma}_4^2$	$3,93.10^{-2}$	$2,1.10^{-2}$	$3,10.10^{-2}$
paramètre γ_5^2 de $U_5(t)$	$\hat{\gamma}_5^2$	$4,89.10^{-2}$	$2,16.10^{-2}$	$1,22.10^{-2}$
Variance des résidus	$\hat{\sigma}_0^2$	$1,22.10^{-4}$	$5,19.10^{-4}$	$4,06.10^{-1}$
paramètre α_1 de $U_1(t)$	$\hat{\alpha}_1$	7,64512	3,532	$1,56.10^{-2}$
paramètre α_2 de $U_2(t)$	$\hat{\alpha}_2$	18,795	12,63	$6,90.10^{-2}$
paramètre α_3 de $U_3(t)$	$\hat{\alpha}_3$	5,85	2,11	$3,00.10^{-3}$
paramètre α_4 de $U_4(t)$	$\hat{\alpha}_4$	0,818	0,498	$5,00.10^{-2}$
paramètre α_5 de $U_5(t)$	$\hat{\alpha}_5$	1,183	$5,899.10^{-1}$	$2,29.10^{-2}$

TAB. 5.7 – Les paramètres estimés du modèle ajusté à la période de Mai 1989 à Septembre 1992

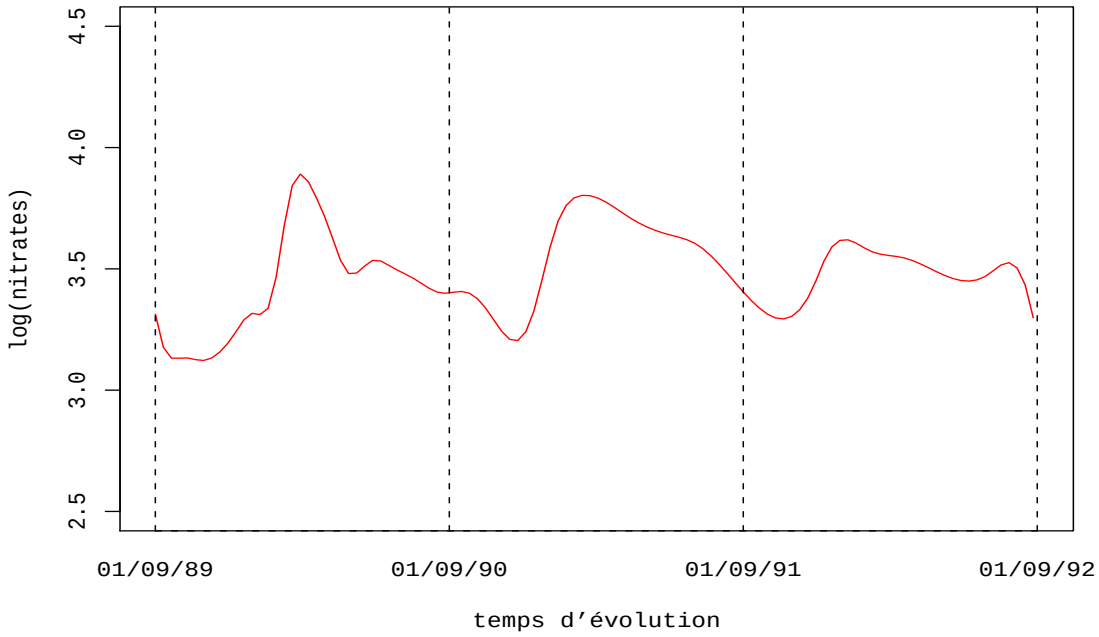
Sachant comme précédemment que la matrice G est choisie égale à l'identité, les réalsations du paramètre b sont :

	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
valeurs	-0,078	-0,079	0,38	-0,113	-0,0110

TAB. 5.8 – Les réalisations de b

On remarque, pour cette composante commune,(voir figure 5.9) que le niveau des concentrations de nitrates est globalement plus élevé par rapport au niveau avant la rupture de l'année 1989. Sur cette période de trois ans, une dimunition des concentrations des nitrates s'observe.

FIG. 5.9 – Composante commune de la période de Septembre 1989 à Septembre 1992



Période de Septembre 1992 **à Septembre** 1997

Ainsi, sur cette période, la représentation graphique de la composante commune (figure 5.10) issue de l'analyse met en évidence un phénomène d'amortissement du niveau des concentrations des nitrates. Cet amortissement s'observe vers la fin de cette troisième période.

Les deux tableaux suivants donnent un résumé des différents paramètres estimés :

Type de paramètre	Notation	Estimation	Erreur-type	P.valeur
Intercept	$\hat{\beta}_0$	3,644	0,37	0,00
variance de b	$\hat{\sigma}^2_*$	0,7	0,41	$4,83.10^{-2}$
variance de a	$\hat{\tau}^2$	0,24	0,065	10^{-4}
paramètre γ^2_1 de $U_1(t)$	$\hat{\gamma}^2_1$	$3,519.10^{-2}$	0,0337	$1,50.10^{-1}$
paramètre γ^2_2 de $U_2(t)$	$\hat{\gamma}^2_2$	$3,68.10^{-2}$	$2,54.10^{-2}$	$7,38.10^{-2}$
paramètre γ^2_3 de $U_3(t)$	$\hat{\gamma}^2_3$	$3,9.10^{-2}$	$2,2.10^{-2}$	$3,72.10^{-2}$
paramètre γ^2_4 de $U_4(t)$	$\hat{\gamma}^2_4$	$4,46.10^{-2}$	1,33	$4,80.10^{-1}$
paramètre γ^2_5 de $U_5(t)$	$\hat{\gamma}^2_5$	$4,45.10^{-2}$	0,41	$4,50.10^{-1}$
Variance des résidus	$\hat{\sigma}^2_0$	$3,8.10^{-2}$	$5,13.10^{-3}$	0,00
paramètre α_1 de $U_1(t)$	$\hat{\alpha}_1$	8,12	81,84	$4,60.10^{-1}$
paramètre α_2 de $U_2(t)$	$\hat{\alpha}_2$	0,57	0,59	$1,69.10^{-1}$
paramètre α_3 de $U_3(t)$	$\hat{\alpha}_3$	0,94870207	0,78	$1.13.10^{-1}$
paramètre α_4 de $U_4(t)$	$\hat{\alpha}_4$	$1,05.10^{-2}$	0,34	$4,87.10^{-1}$
paramètre α_5 de $U_5(t)$	$\hat{\alpha}_5$	$3,87.10^{-2}$	0,448	$4,65.10^{-1}$

TAB. 5.9 – Les paramètres estimés du modèle ajusté à la période de Mai 1992 à Septembre 1997

Sachant comme pour les deux premières périodes que la matrice G est choisie égale à l'identité, les prédictions des effets aléatoires b sont :

	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
valeurs	-0,091	-0,0794	0,4091	-0,131	-0,106

TAB. 5.10 – Les prédictions de b

Les composantes spécifiques par site sont données aussi, ces dernières s'ajoutent à la composante commune de la figure (5.10) pour donner, à une constante près, les profils de l'évolution des concentrations des nitrates des 5 sites sur cette période (cf. figure 5.11).

FIG. 5.10 – Composante commune de la période de Septembre 1992 à Septembre 1997

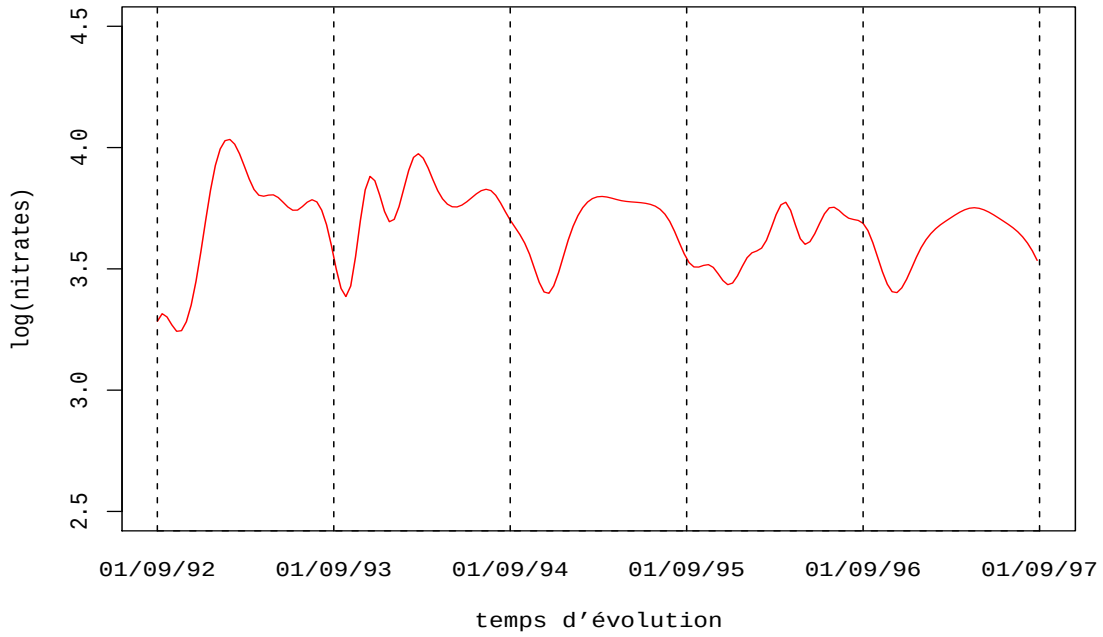
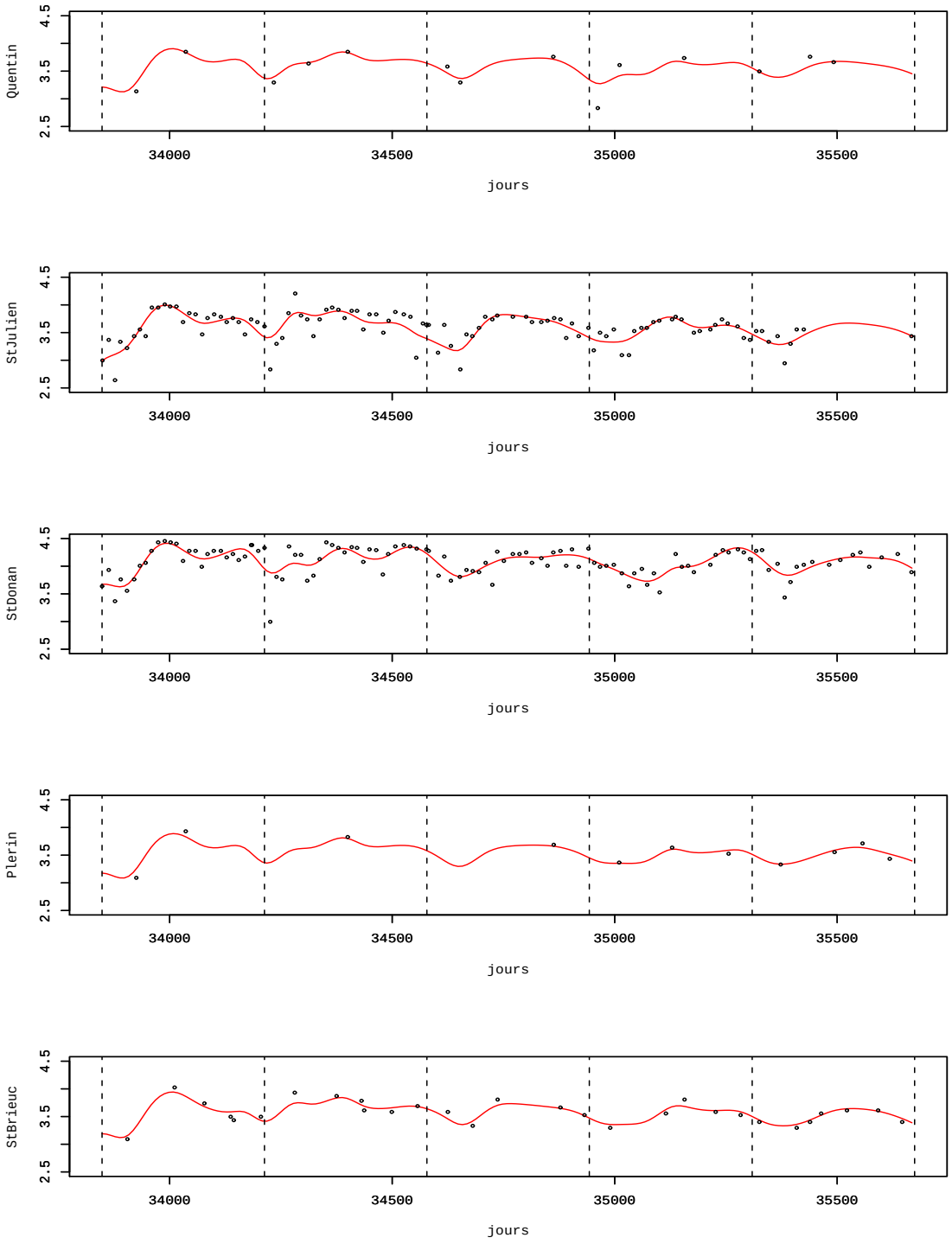


FIG. 5.11 – Courbes d'évolution propres à chaque site pour la période de Septembre 1992 à Septembre 1997



Période de Septembre 1997 **à Mars** 2001

Pour cette dernière partie de la période d'observation, l'application d'un modèle de type (5.2), donne la composante commune pour cette dernière partie. L'intercept pour cette courbe des transformées logarithmique est égal à $(3,73)$ tandis que celui de la période de Septembre 1992 à Septembre 1997, qui est $(3,65)$.

Type de paramètre	Notation	Estimation	Erreur-type	P.valeur
Intercept	$\hat{\beta}_0$	3,73	0,124	0,00
variance de b	$\hat{\sigma}_*^2$	$7,62.10^{-2}$	$4,88.10^{-2}$	$6,02.10^{-2}$
variance de a	$\hat{\tau}^2$	$1,3.10^{-1}$	$2,75.10^{-2}$	10^{-6}
paramètre γ_1^2 de $U_1(t)$	$\hat{\gamma}_1^2$	$2,34.10^{-2}$	$5,58.10^{-3}$	10^{-5}
paramètre γ_2^2 de $U_2(t)$	$\hat{\gamma}_2^2$	$1,85.10^{-2}$	$8,96.10^{-3}$	$1,98.10^{-2}$
paramètre γ_3^2 de $U_3(t)$	$\hat{\gamma}_3^2$	$4,25.10^{-2}$	$1,11.10^{-2}$	$9,15.10^{-5}$
paramètre γ_4^2 de $U_4(t)$	$\hat{\gamma}_4^2$	10^{-6}	10^{-6}	10^{-6}
paramètre γ_5^2 de $U_5(t)$	$\hat{\gamma}_5^2$	$1,043.10^{-3}$	$2,20.10^{-3}$	$3,10.10^{-1}$
Variance des résidus	$\hat{\sigma}_0^2$	$3,432.10^{-3}$	$1,46.10^{-3}$	0,00
paramètre α_1 de $U_1(t)$	$\hat{\alpha}_1$	15,49	7,243164	$1,68.10^{-2}$
paramètre α_2 de $U_2(t)$	$\hat{\alpha}_2$	18,5477	17,1475	$4,56.10^{-1}$
paramètre α_3 de $U_3(t)$	$\hat{\alpha}_3$	4,95	2,429	$2,14.10^{-2}$
paramètre α_4 de $U_4(t)$	$\hat{\alpha}_4$	130,71	$6,47933.10^3$	$4,91.10^{-1}$
paramètre α_5 de $U_5(t)$	$\hat{\alpha}_5$	1,78	5,3366	$3,69.10^{-1}$

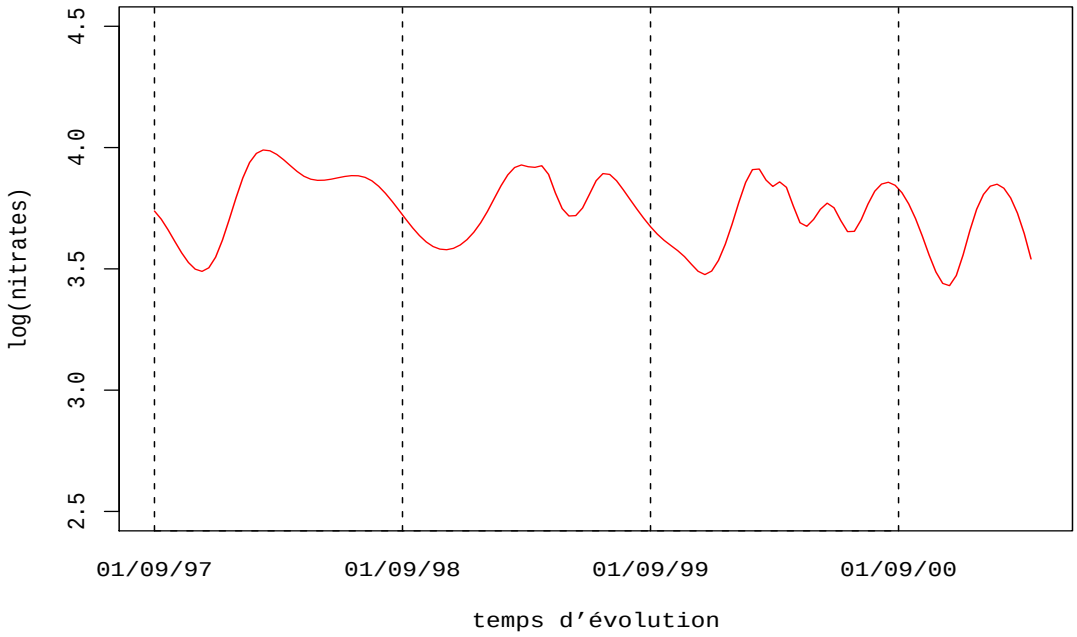
TAB. 5.11 – Les paramètres estimés du modèle ajusté à la période de Mai 1992 à Septembre 1997

En considérant toujours que la matrice G est choisie égale à l'identité, les prédictions des effets aléatoires b sont :

	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
valeurs	-0,140	-0,085	0,381	-0,0812	-0,0746

TAB. 5.12 – Les prédictions de b

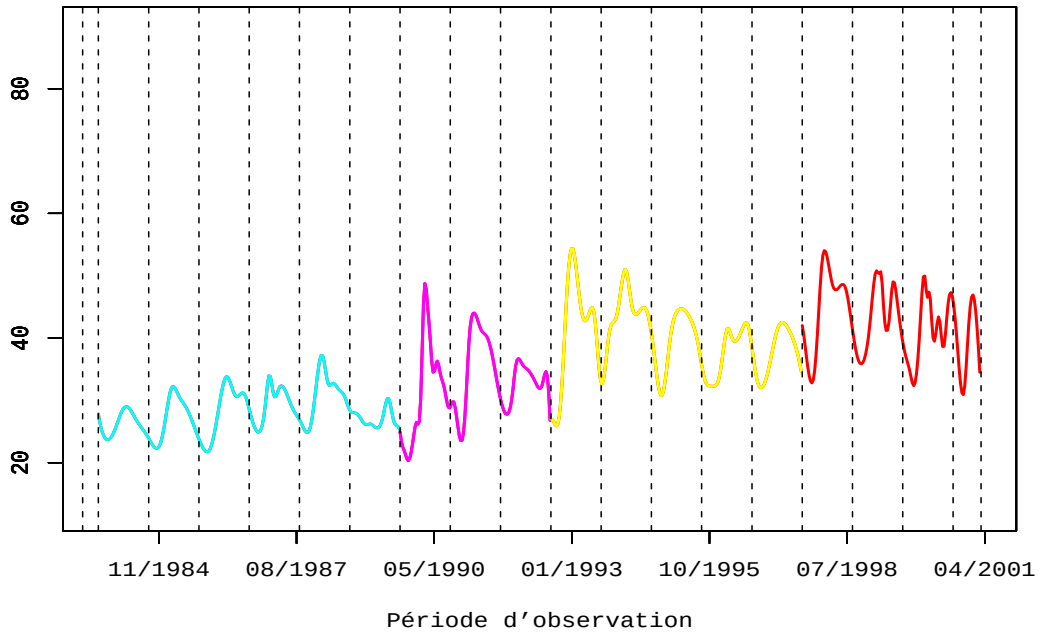
FIG. 5.12 – Composante commune de la période de Septembre 1997 à Mars 2001



Conclusion

L'application du modèle mixte sémiparamétrique stochastique a permis l'estimation des composantes communes des quatre périodes détectées par les ruptures des modèles. Il permet surtout la reconstruction des profils de l'évolution des concentrations des nitrates sur certains sites où les observations ne sont pas nombreuses via l'exploitation des observations provenant d'autres sites. Pour la période globale de l'étude, la reconstruction de certaines courbes nécessite des contraintes de continuité, pour le modèle (5.2), comme le montre la figure (5.13) pour la composante commune sur la période globale d'observation au point de rupture de Septembre 1997. En dehors de ce point, la composante commune sur les zones données est correctement restituée.
L'utilisation d'autres variables explicatives pour rendre le modèle (5.2) plus explicatif nécessiterait une contrainte d'orthogonalité entre ces dernières et la fonction nonparamétrique $f(t)$. C'est la raison pour laquelle malgré l'apport noté par le modèle GAM pour des variables comme L'ETP et la pluviométrie, ces variables n'ont pas été introduites dans le modèle. L'introduction de ces variables nécessiterait des développements futurs.
Cette application montre aussi que le modèle mixte sémiparamétrique améliore les résultats obtenus à partir de l'approche par variables latentes.

FIG. 5.13 – Composante commune pour la période totale d'observation de l'étude



5.3 Les données de qualité d'eau de baignade en mer de Dinard

Les côtes de la ville de Dinard font l'objet d'une surveillance pour avertir les baigneurs, de la qualité de l'eau de mer. En effet, le risque de gastroentérites et autres affections (maladies respiratoires) associées à la baignade est directement mesurable par la présence des agents pathogènes dans les eaux de baignade. Des organismes microbiologiques comme les Escherichia Coli (EC) et les Entérocoques Intestinaux (IC) sont des excellents indicateurs de la présence des micro-organismes pathogènes (bactéries, virus, parasites) présents dans l'eau. En effet, ces bactéries et dans une moindre mesure les Coliformes Totaux (CT) sont des indicateurs de la charge en organismes pathogènes présents dans l'eau de baignade. Des études épidémiologiques essentiellement basées sur les relations dose-effet ont établi les risques de contracter une gastroentérite pour les différents niveaux de présence de ces bactéries dans l'eau de baignade.
Le dénombrement de ces bactéries contenues dans 100 ml d'échantillon d'eau permet le classement des différents sites de baignades. Les plages sont classées dans différentes catégories en fonction de leur niveau de qualité (et conformément à la réglementation en vigueur). Ce classement s'effectue à la fin de chaque saison, avec les observations issues de la saison écoulée. Les percentiles de la distribution ajustée aux mesures microbiologiques de l'échantillon de la saison écoulée sont comparées à des valeurs de référence (seuils). La dernière directive en application est celle approuvée par la commission européenne depuis Juin 2004. Cette directive établit les références ci-dessous :

Indicateurs	Exellente qualité	Bonne qualité	Qualité satisfaisante
IE en UFC/100ml	100★	200★	200✠
EC en UFC/100ml	250★	500★	500✠

TAB. 5.13 – Pour qu'une eau de baignade soit classée dans une catégorie de qualité donnée, il faut que les percentiles des concentrations sur les deux indicateurs microbiologiques soient inferieurs aux valeurs seuils de la classe de qualitée considérée ; ★ valeurs seuils à comparer aux percentiles 0,95 des mesures microbiologiques, ✠ valeurs seuils à comparer aux percentiles 0,90 des mesures microbiologiques

.

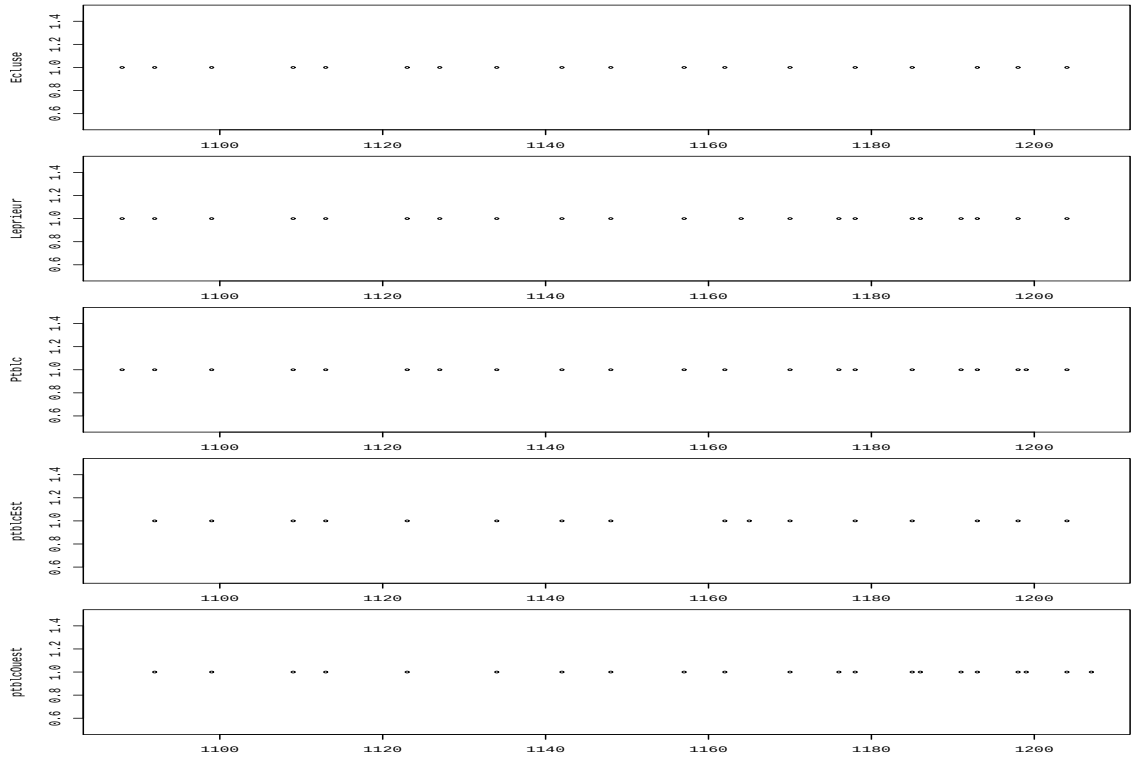
L'intérêt porte sur l'analyse des courbes des données de concentrations microbiologiques recueillies sur plusieurs saisons successives. Nous essayons à l'aide de l'analyse de ces courbes de voir si effectivement les coubes ont des comportements inter-annuels proches. Le classement des plages se fait sur les observations de la saison précédente ou des quatre dernières saisons. L'analyse des comportements inter-annuels permet d'évaluer la pertinence de cette méthode de classement.
Le modèle mixte sémiparamétrique stochastique proposé permet la construction des courbes en temps continu et par la mise en commun d'observations de différents sites (plages). L'observation de la carte de la qualité des plages laisse apparaître une homogénéité de la qualité des plages à l'échelle d'une ville. Nous utilisons cette caractéristique des plages de la ville de Dinard pour extraire leur composante commune sur différentes années. Celle-ci fournit le

comportement des paramètres considérés, corrigés des différentes erreurs de mesures et des spécificités locales aux différentes plages. Sur une année, l'analyse des courbes permettra de comparer les différentes plages de Dinard entre elles et si le lieu où sont situées ces plages influencent sur les différences.
A l'échelle de l'année, une comparaison des courbes de concentrations des deux indicateurs, permet de voir si les indicateurs concordent sur la qualité des plages.

5.3.1 Description des données

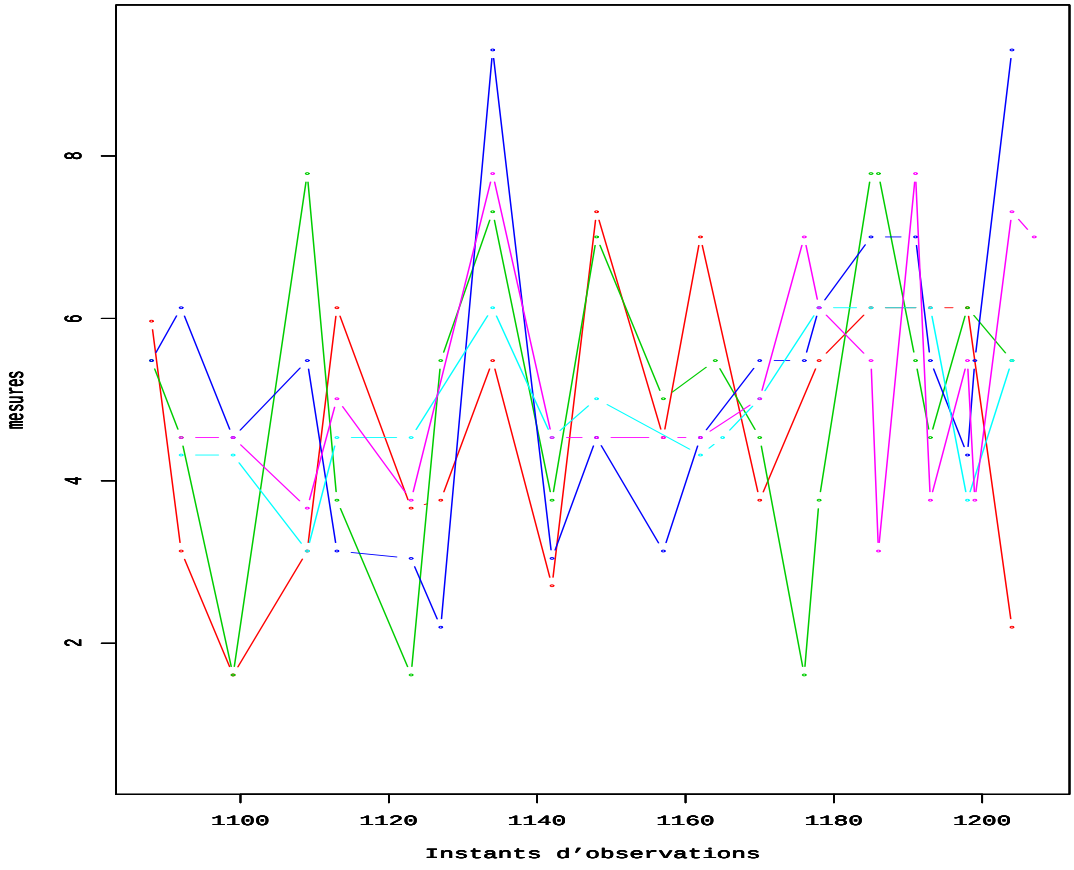
Les données ont été recueillies dans le cadre de suivi de la qualité des plages pendant les saisons de baignades de 1979 à 1994. Les sites surveillés sont les plages suivantes de la ville de Dinard : L'Ecluse, Le Prieure, Port Blanc, Port Riou et Saint Enogat. La station de Port Blanc, aux premières années de la période de surveillance était scindée en deux stations : Port Blanc Est et Port Blanc Ouest.
Sur une année, la période d'observation s'étend sur toute la saison d'été commençant parfois au début du mois de Juin jusqu'à fin Septembre. Le nombre d'observation par site sur une saison varie approximativement entre 7 et 20. Cette variation du nombre d'observation se remarque d'une année sur l'autre.
Ainsi sur la période d'observation de l'année 1982, le plan d'échantillonnage de collecte des données se présente sous la forme suivante :

FIG. 5.14 – Plan d'échantillonnage de collecte des données pour la saison de l'année 1982



On notera la différence entre la collecte des observations sur la plage «L'Ecluse» et la denière plage de «Saint Enogat» où certaines observations sont effectuées de façon régulière. Les courbes d'évolution, des concentrations microbiologiques recueillies sur la saison sont données par la figure (5.15).

FIG. 5.15 – Représentation des courbes d'évolution représentées (E.C.) à partir des observation des 5 plages : L'ecluse (rouge), le prieure (vert), Port Blanc (bleu), Port Riou (cyan), St Enogat (mauve)



Il faut noter que ces courbes présentent une dynamique commune qui peut représenter la relative homogéneïté à l'échelle des 5 plages de la ville.
L'approche consistant à analyser les courbes sur une saison est effectuée en raison du nombre important des jours en continu et sans observation en dehors des saisons. Cette longue période sans observations compromet la représentation des courbes sur plusieurs années, d'où la nécessité de faire l'analyse des courbes par saison.

5.3.2 Représentation des deux paramètres à l'échelle de la ville

La représentation graphique (5.15) permet d'observer l'allure des courbes d'évolution des concentrations de l'indicateur E.C. recensées dans les pélèvements d'eau de mer autour de la ville de Dinard. Cette similarité entre les courbes issues des différentes plages s'observe aussi pour l'indicateur E.I. L'une des possibilités offertes par l'application du modèle mixte sémiparamétrique stochastique est d'estimer une composante commune représentant la dynamique d'évolution du paramètre mesuré (E.C. ou E.I.) commune à l'ensemble des cinq plages, à partir de l'ensemble des observations disponibles. La courbe ainsi obtenue représente l'allure des populations de bactéries présentes dans l'eau de mer de Dinard indifféremment de la localisation de la plage. Cette localisation de la plage peut influencer le niveau des agents pathogènes présents dans l'eau.
La composante commune ainsi extraite est affranchie des différentes spécificités dues à la localisation de la plage en question. La composante spécifique d'une plage i donne la différence entre le comportement de cette plage et leur composante commune. Le modèle appliqué aux données de la qualité des plages de la ville de Dinard, pour le profil de la station est représenté par :

$$Y_i(t_{ij}) = \beta_0 + f(t_{ij}) + b_i + U_i(t_{ij}) + \epsilon_{ij}. \tag{5.3}$$

La composante commune du modèle est représentée par la partie $\hat{\beta}_0 + \hat{f}$ de l'équation (5.3). Cette partie de l'équation est présente dans toutes les constructions des profils des courbes d'évolution des agents pathogènes indépendamment de la plage i. Elle permet de présenter le caractère d'homogénéïté observé sur les plages de la ville. Sur la figure (5.16), les composantes communes obtenues pour les années allant de 1980 à 1995 sont ainsi représentées. Les méthodes d'estimation n'ont pas convergé pour les années 1988 (notée sur la représentation graphique : année 10) à 1990 (année 12) ainsi que la dernière année (année 17). Ce sont des années où le nombre d'observations recueillies n'est pas suffisamment élevé. Cela se réflète sur les graphiques. La courbe représentée, dans ces ces cas-là est très lisse.
La planche (5.17) comprend seulement les graphiques des années où la convergence de la méthode est observée. Pour ces années, l'observation des différentes courbes laisse apparaitre que les niveaux des concentrations des bactéries pathogènes dans l'eau sont assez différents d'une année sur l'autre. Ors la méthode de classement à partir des observations issues de quatre saisons est faite sur la base que d'une saison à une autre la courbe des populations de bactéries évoluent de façon identique. L'observation de ces graphiques (voir figure 5.17) montre des différences entre les courbes des différentes saisons et ne conforte pas cette hypothèse.

FIG. 5.16 – Représentation graphique des composantes communes de 16 différentes années (de 1980 à 1994)

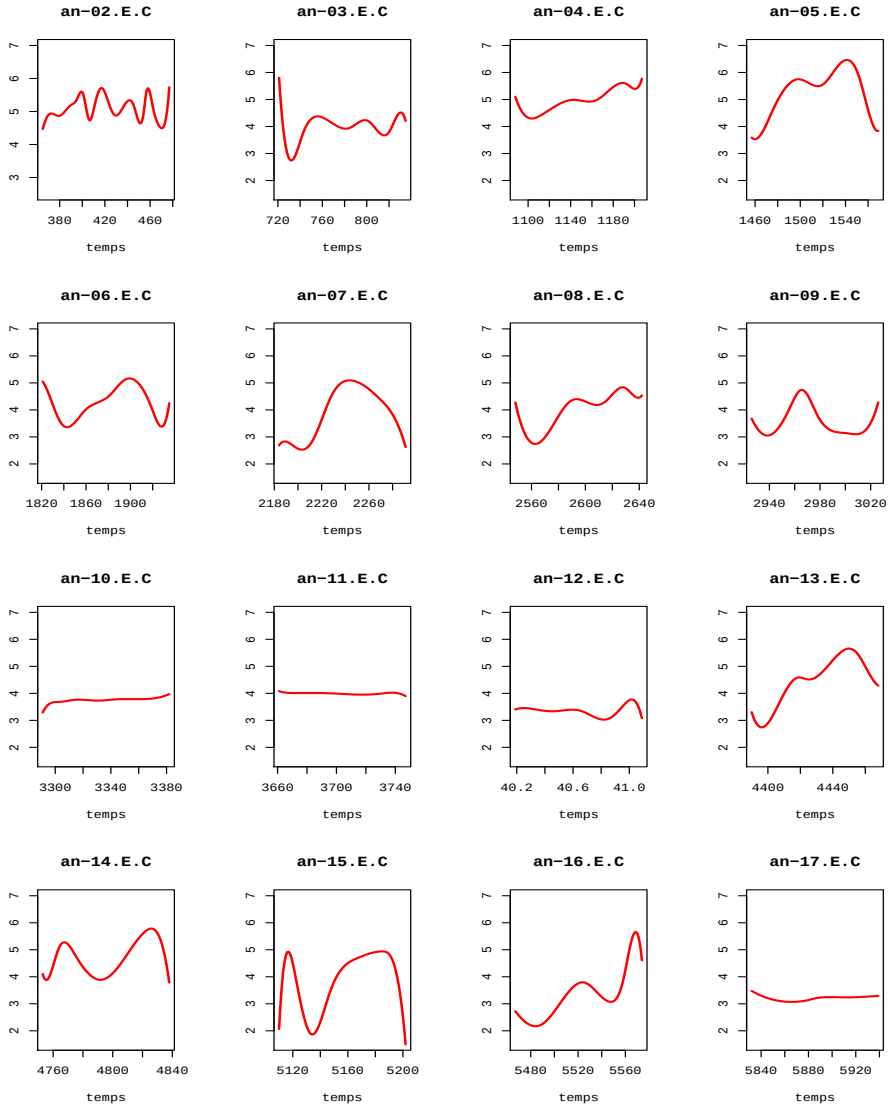
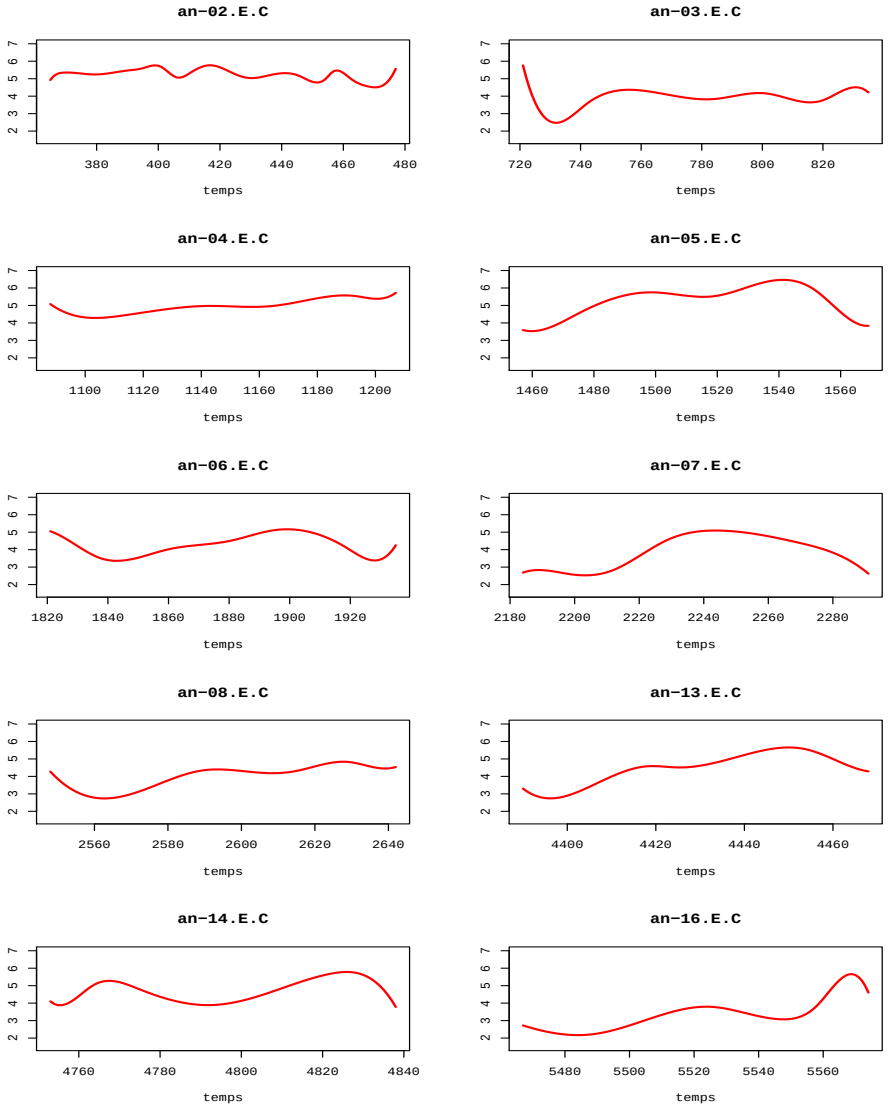


FIG. 5.17 – Représentation graphique des composantes communes (E.C.) pour les années où il y a eu convergence de la logvraisemblance



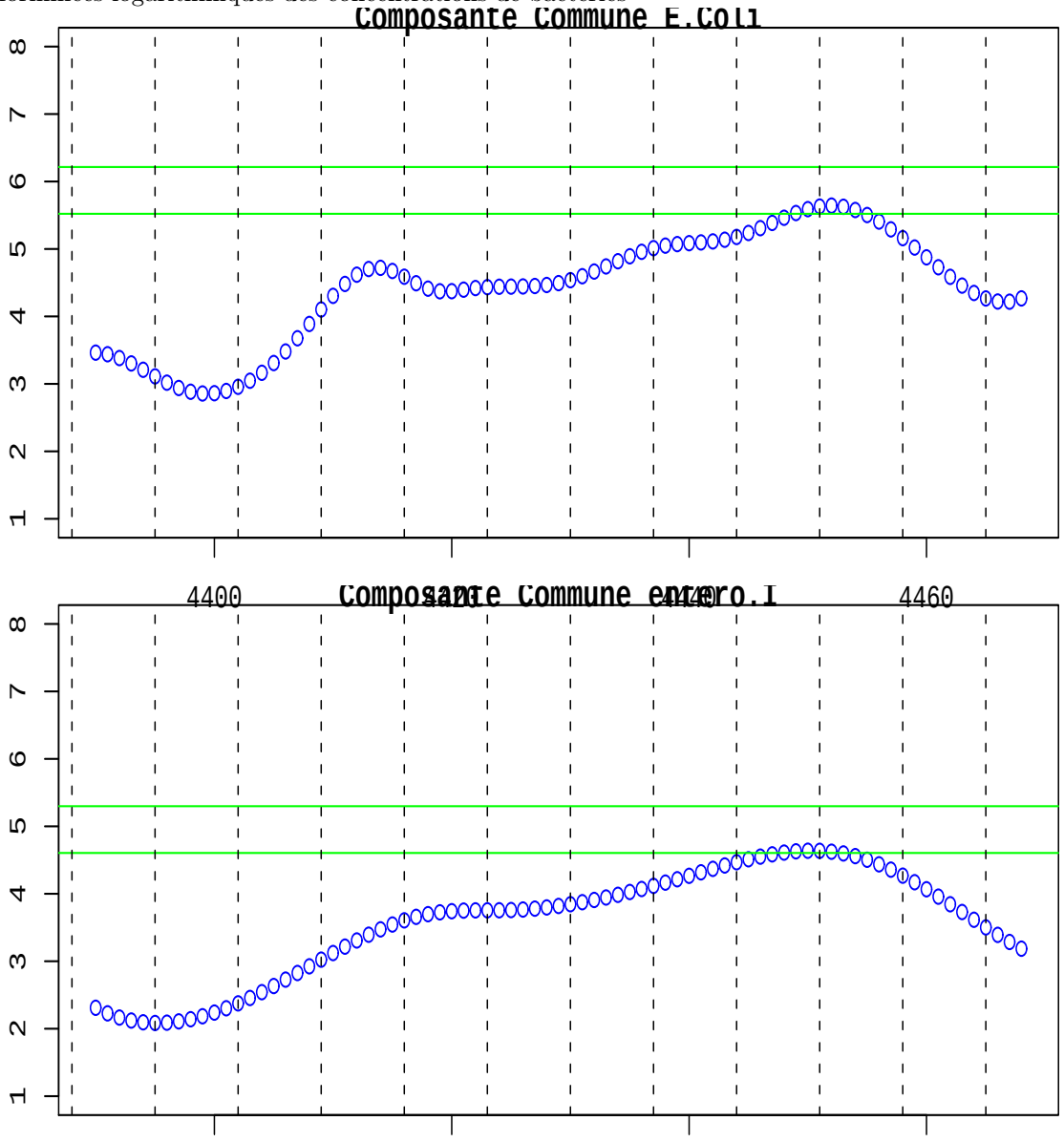
5.3.3 Comportement sur une saison des différentes plages

Dans cette partie, l'intérêt de l'étude consiste à observer l'évolution des agents pathogènes utilisés comme indicateurs à l'échelle d'une saison. L'étude du phénomène sur une année présente deux intérêts :

– Une comparaison des niveaux de qualité d'eau de baignade indiqués par les deux populations de bactéries, chaque semaine et ceci tout au long de la saison. Cela permettra de noter les écarts d'indications entre ces deux populations de bactéries.
– La comparaison des courbes d'évolution de la population des bactéries des différentes plages permet de noter si la localisation des plages a des effets sur la qualité des eaux de

baignade.

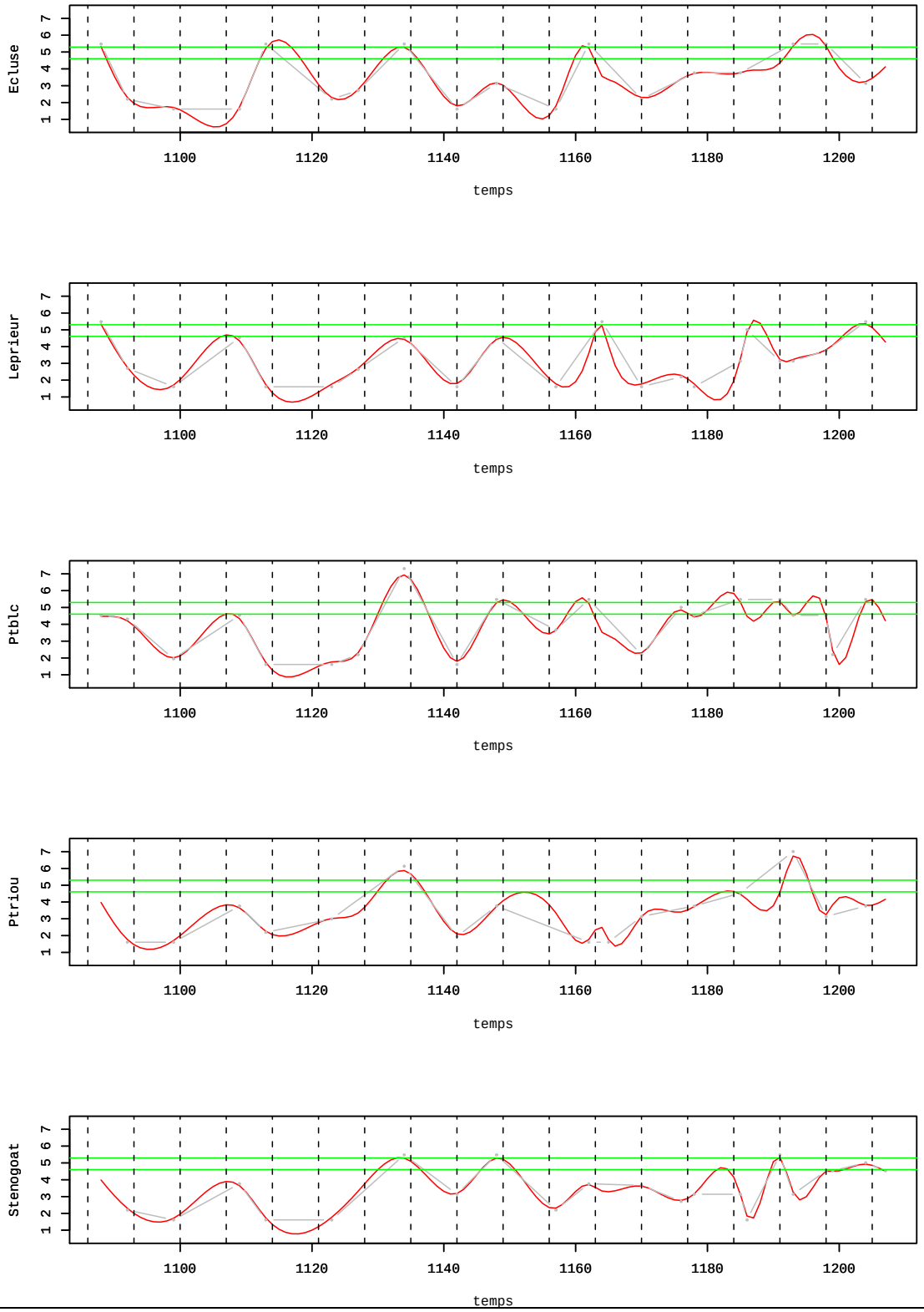
FIG. 5.18 – Représentation graphique des composantes communes (en bleu) (E.C.) et (E.I) pour l'année 1991 : Comparaison d'indication des 2 populations avec les seuils de référence en vert et les semaines d'observations délimités par les traits pointillés avec la première semaine d'observation à gauche. On notera la concordance des 2 courbes reconstruites par rapport aux traits verts. En ordonnées : ce sont les transformmées logarithmiques des concentrations de bactéries



Pour le premier cas, nous avons surtout utilisé les composantes communes des deux populations de bactéries, ce choix permet de considérer seulement l'indication de la population des bactéries sans considérer les effets spécifiques du lieu où la plage est située. Ainsi, pour

l'année 1991, on notera que les E.I. indiquent dans l'ensemble les eaux des cinq plages de la ville de Dinard sont d'excellente qualité sur 10 des 12 semaines de la saison d'observation. Alors que pour la neuvième et dixième semaine, pour ces deux semaines successives la qualité des eaux se dégrade. (voir figure (5.18)). La même indication est obtenue pour la population de l'indicateur d'E.coli. Il faut noter que cette constation coïncide avec l'hypothèse retenue dans la littérature à savoir que pour la dernière réglementation ces deux populations de bactéries avec les seuils de références notés indiquent la même qualité d'eau de baignade. Cette coïncidence conforte donc cette hypothèse.
Après cette comparaison des courbes des indicateurs microbiologiques par ajustement de la composante commune, le résultat coïncide avec les hypothèses faites à savoir que les deux indicateurs donnent les mêmes indications sur les qualités des eaux de baignades. Ce résultat valide la méthode de reconstruction des courbes d'évolution des populations de bactéries. Nous pouvons donc retenir indifféremment l'un des indicateurs pour la reconstruction des courbes d'évolutions pour les profils des cinq sites d'observation. Les entérocoques intestinaux (E.I.) ont été retenus pour cette deuxième partie, les composantes spécifiques des 5 sites ont été représentées pour l'année 1982. Sur l'ensemble des 5 sites, la composante spécifique de la plage de Port Blanc se détache assez nettement des autres. Par rapport aux 4 autres plages, c'est celle dont le niveau de la qualité de l'eau demeure dans la catégorie inférieure un nombre de jours relativement très élevé sur la saison considérée (cf. la gigure de reconstruction 5.19). . Le suivi de la qualité des eaux de baignade de cette plage devrait attirer l'attention afin de déceler la cause de cette détérioration. Les plages de Port Riou et de Le Prieur sont celles dont les profils individuels présentent peu de jours de présence dans la catégorie de niveau inférieur (voir figure (5.19)).

FIG. 5.19 – Représentation graphique des composantes spécifiques (E.I.) pour l'année 1982 :gris (courbe d'observations d'origines), rouge (courbe d'observations reconstruites)



5.3.4 Proposition d'une méthode de classement des plages dans les différentes catégories de qualité d'eau de baignade

Dans la partie introductive de cette seconde application, la méthode de classement des plages dans les différentes catégories de qualité d'eau de baignade a été décrite.
La méthode se base sur l'hypothèse que les observations provenant de la saison précédente (ou les quatre saisons précédentes selon le cas) sont issus d'un même échantillon et que la loi de cet échantillon est une loi lognormale. Les paramètres empiriques obtenus à partir de l'échantillon considéré permettent d'ajuster la loi empirique et de calculer la distribution de ces différents quantiles. La méthode consiste à considérer les quantiles 95% pour les deux indicateurs microbiologiques de les comparer à 100 (pour les E.I.), à 250 (pour les E.C) pour la qualité excellente et à 200 (pour les E.I.), à 500 (pour les E.C) pour la bonne qualité. Sinon le percentile 90% est comparé aux valeurs 200 (pour les E.I.), à 500 (pour les E.C) pour la qualité satisfaisante. Une plage est classée dans une catégorie de qualité lorsque les percentiles des deux indicateurs microbiologiques sont inférieurs aux valeurs seuils de la classe considérée. Autrement dit, cette méthode revient à classer :

- Dans la catégorie «*excellente qualité*», si 5% de la distribution de la loi lognormale ajustée à l'échantillon (observations de la saison précédente ou des quatre saisons précédentes) ne dépasse pas la valeur seuil de 100.
- Dans la catégorie «*bonne qualité* », si 5% de la distribution de la loi lognormale ajustée à l'échantillon (observations de la saison précédente ou des quatre saisons prédédentes) ne dépasse pas la valeur seuil de 200
- Dans la catégorie «*qualité satisfaisante* », si 10% de la distribution de la loi lognormale ajustée à l'échantillon (observations de la saison précédente ou des quatre saisons précédentes) ne dépasse pas la valeur seuil de 200

C'est l'approche paramétrique de classement des plages suivant la qualité de leur eau de baignade, méthode la plus souvent appliquée en raison de l'insuffisance du nombre d'observation par site.
Cependant cette approche souffre de quelques hypothèses assez fortes : la première est que la loi de l'échantillon retenue ne s'ajuste pas correctement à une loi lognormale ; la seconde est que cette façon de classer ne renseigne aucunement sur les périodes de la saison où les seuils sont dépassés.
La méthode que nous proposons pour classer consiste simplement à reconstruire les courbes d'évolution des concentrations microbiologiques des plages de la saison passée à l'aide du modèle mixte sémiparamétrique stochastique en exploitant l'ensemble des observations des autres plages de la ville et noter le nombre de dépassement des seuils de la saison écoulée pour classer la plage. Nous avons ainsi comparé les méthodes de classement à partir des courbes d'évolution construites (saison de l'année 1982) dans le paragraphe précédent.
L'observation des courbes reconstruites laisse apparaître que la fin des saisons présente plus des risques de dépassement de seuils que les débuts, une tendance à la détérioration des eaux au cours des saisons. Il apparaît donc intéressant de tenir compte de ce fait dans le classement des eaux de baignade des plages.

	"I'Ecluse"	"Prieure"	"PrtBlanc"	"PrtRiou"	"StEnogat"
(seuil 100)Méthode Classique	23,96	16,98	35,56	20,56	19,40
(seuil 200)Méthode Classique	12,74	7,73	21,60	10,50	8,40
(seuil 100)Méthode Proposée	20	13,33	32,5	11,66	17.5
(seuil 200) Méthode Proposée	9,17	4,17	16,67	6,67	1,67

TAB. 5.14 – Les pourcentages de dépassement des valeurs seuils de 100 et 200 sont représentés pour les cinq plages à l'issue des observations de la saison 1982, pour les deux méthodes. On notera que la méthode classique effectue un classement plus sévère. Cela sans doute est dû au fait que les dépassements de seuils sont calculés à l'aveugle, sans aucune information sur les dates où ces seuils sont dépassés alors que la reconstruction des courbes d'évolution fournit ces informations.

Cet exemple traduit tout l'intérêt des approches sémiparamétriques. Dans le cadre de l'exposition environnementale où l'information n'est pas toujours complète, la méthode permet la reconstruction des courbes d'évolution de l'exposition aux différentes bactéries présentes dans les eaux de baignade. La connaissance de cette exposition quotidienne peut permettre la détermination des effets négatifs résultant et les conséquences en terme de santé publique.

Bibliographie

[1] *Proceedings of the First Interceltic Colloquium on Hydrology and water management*, Rennes, 1996. Cann, C.

[2] R Development Core Team. *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. 3-900051-07-0.

[3] A.A. Avery. Infantile methemoglobinemia : reexamining the role of drinking water nitrates. *Environmental Health Perspectives*, 107 :583–586, 1999.

[4] R.J. Beckman, C.J. Nachtsheim, and R.D. Cook. Diagnostics for mixed-models analysis of variance. *Technometrics*, 29, 1987.

[5] D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics*. J. Wiley & Sons, New Jersey, 2004.

[6] R. D. Cook. Detection of influtiential observations in linear regression. *Technometrics*, 19 :15–18, 1977.

[7] R. D. Cook. Influtiential observations in linear regression. *Journal of the American Statistical Association*, 74 :169–174, 1979.

[8] R. D. Cook. Assessment of local influence. *Journal of the Royal Statistical Society*, 48 :133–169, 1986.

[9] N.A. Cressie. *Statistics for spatial data*. Wiley Series in Probability and mathematical statistics : Applied Probability and statistics. Wiley, New York, revised ed edition, 1993.

[10] C. de Boor. Splines as linear combination of b.splines. a survey. In C.K. Chui G.G. Lorentz and L.L. Schumaker, editors, *Approximation Theory II*, pages 1–47. 1976.

[11] C. de Boor. B.(asic) spline basics. In Les Piegl, editor, *Fundamental Developments of Computer-Aided Geometric Modeling*, pages 27–49. London, 1993.

[12] P. J. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford University Press, Oxford, seconde edition, 2002.

[13] C. R. DiMatteo, I. andGenovese and R. E. Kass. Bayesian curve-fitting with free-fnots splines. *Biometrika*, 88 :1055–1071, 2001.

[14] D. L. Donoho and I.M. JOHNSTONE. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 11 :425–455, 1994.

[15] J. Franke and B. Grunder. *Athens Conference on Applied Probability and Time Series*, volume 2, chapter General kriging for spatial-temporal processes with random ARX-regression parameters, pages 177–190. Springer-Verlag, 1996.

[16] P.J. Green. Reversible jump markov chain monte carlo computation and bayesian models determination. *Biometrika*, 82 :711–732, 1995.

[17] P.J. Green and P.J. B.W.Silverman. *Nonparametric Regressions and Genralized Linear Models*. Chapman & Hall, London, 2nd edition, 1994.

[18] C. Gu. *Smooothing spline ANOVA models*. Springer-verlag, New Yok, 2002.

[19] D.A. Harville. Maximum likelihood approaches to variance component estimation and related problems. *Journal of the American Statistical Association*, 72 :320–339, 1977.

[20] T.J. Hastie and R.J. Tibshirani. *Genralized Additive Models*. Chapman & Hall, London, 4th edition, 1990.

[21] H. Jacqmin-Gadda, P. Joly, D. Commenges, and C. Binquet. Penalized likehood approach to estimate a smooth curve on longitudinal data. *Statistics in Medecine*, 21 :2391–2402, 2002.

[22] G. James and T. Hastie. Principal component models for sparse functional data. *Biometrika*, 82 :711–732, 2000.

[23] Ath. Kehagias. A hidden markov model segmentation of hydrological and environemental time series. http ://citeseer.nj.nec.com/564055.html, 2002. Documentation on the web.

[24] B. Knobeloch, L and.Saina, A. Hogan, J. Postle, and H. Anderson. Blue babies and nitrate-contaminated well water. *Environmental Health Perspectives*, 108 :675–678, 2000.

[25] R. Kohn, C.F. Ansley, and D. Tharm. The performance of cross-validation and maximum likelihood estimator of spline smoothing parameters. *Journal of the american statistical association*, 86 :1042–1050, 1991.

[26] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38 :963–974, 1982.

[27] E Lesaffre and G. Verbeke. Local influence in linear mixed models. *Biometrics*, 54 :570–582, 1998.

[28] J. R. Magnus and A. L. Vasnev. Local sensitvity and diagnostics tests. Technical report, Tilburg University, Center and Department of Econometrics and Operations Research, October 2004.

[29] G. Matheron. *Traité de géostatistique appliquée*. Bur. Rech. Géol. Minières, Paris, 1962.

[30] D.J. Nott and T.M. Dunsmuir. Estimation of nonstationary spatial covariance stucture. *Biometrika*, 89 :819–829, 2002.

[31] D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9, 1981.

[32] C. R. Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New York, second edition, 1973.

[33] J.A. Rice and C.O. Wu. Nonparametric mixed models for unequally sampled noisy curves. *Biometrics*, 57 :253–259, 2001.

[34] G.K. Robinson. That BLUP is a good thing : the estimation of random effects. *Statistical science*, 6 :15–51, 1991.

[35] D. Ruppert, M.P. Wand, and R. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, 2003.

[36] P.D. Sampson and P. Guttorp. Nonparametric estimation of non-stationary spatial covariance stucture. *Journal of American Statistical Association*, 87 :108–119, 1992.

[37] S. R. Searle, G. Casella, and C.E. McCulloch. *Variance Components*. John Wiley & Sons, New York, first edition, 1992.

[38] B. W. Silverman. Some aspects of the smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47 :1–52, 1985.

[39] M.L. Stien. A comparaison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *The Annals of Statistics*, 18 :1139–1159, 1990.

[40] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer-Verlag, New Yok, 2000.

[41] G. Wahba. Improper priors, spline smoothing and the problem of guarding against models errors in regression. *Journal of the Royal Statistical Society*, 40 :364–372, 1978.

[42] G. Wahba. A compparaison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13 :1378–1402, 1985.

[43] G. Wahba and S. Wold. A completely automatic french curve. *Commun. Statistics*, 4 :1–17, 1975.

[44] M.P. Wand. Vector differential calculus in statistics. *American statistician*, 56 :55–62, 2002.

[45] Y. Wang. Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society*, 60 :159–174, 1998.

[46] S.L. Zeger and P.J. Diggle. Semiparametric models for longitudinal data with applications to cd4 numbers in hiv seroconverters. *Biometrics*, 50 :689–699, 1994.

[47] D. Zhang, X. Lin, J. Raz, and M. Sowers. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93 :710–719, 1998.

[48] X. Zhang and M. L. King. Influence diagnostics in garch processes. Technical Report 19/2002, Monash University, Department of Econometrics and Business Statistics, December 2002. available at http ://ideas.repec.org/p/msh/ebswps/2002-19.html.

[49] H.T. Zhu and S.Y. Lee. Local influence in generalized linear models. *La revue canadienne de statistisque*, 31 :293–309, 2003.