Modèle linéaire de prédiction fonctionnelle sur données environnementales : choix de modélisation

Séverine Bayle, Pascal Monestiez, David Nerini

Journal de la Société Française de Statistique
Vol. 155 No. 2 (2014)

Modèle linéaire de prédiction fonctionnelle sur données environnementales : choix de modélisation

Title: Linear predictive functional model on environmental data: modeling choices

Séverine Bayle[1], Pascal Monestiez[1] et David Nerini[2]

Résumé : L'analyse de données fonctionnelles est devenue ces dernières années un champ d'étude important en statistiques, car de plus en plus de données observées dans différents domaines se trouvent sous forme de courbes (météorologie, économie, ...). Un des outils de l'analyse de données fonctionnelles est le modèle linéaire "pleinement" fonctionnel, qui est utilisé dans le cas où la variable à prédire et la variable prédictive sont toutes les deux des courbes. Ce modèle a fait l'objet de recherches théoriques approfondies, mais les applications l'utilisant restent peu nombreuses à ce jour. Nous proposons dans cet article une démarche méthodologique à travers un exemple d'application de ce modèle sur des profils océanographiques de lumière et de Chlorophylle *a*. Il est utilisé ici pour prédire des profils de Chlorophylle *a* à partir des dérivées des profils de luminosité. La démarche méthodologique permet de clarifier les choix de modélisation que nous avons eu à faire pour traiter les profils océanographiques. Les questionnements à travers notre étude de cas concernent entre autres le choix du type et du nombre de fonctions de base à utiliser, le choix de la valeur du paramètre de lissage, ainsi que le critère pour évaluer la qualité de l'ajustement. Nous montrons que l'utilisation du modèle linéaire fonctionnel permet d'obtenir une bonne qualité de reconstruction pour accéder aux variations hautes fréquences des profils de Chlorophylle *a* à fine échelle.

Abstract: Functional data analysis (FDA) has become in recent years an important field in statistics, because more data observed in different domains are in the shape of curves (meteorology, economics, linguistics, ...). One tool in FDA is the fully functional linear model, which is used in the particular case where the variable to be predicted and the predictor are both curves. This model has been the subject of extensive theoretical research, but applications using it are few in number to date. We propose in this paper a methodological approach through an application of this model on light and Chlorophyll *a* oceanographic profiles. It is used here to predict Chlorophyll *a* profiles from derivatives of light data. The methodological approach helps to clarify modeling choices necessary to treat oceanographic profiles. Questions through our case study include the choice of the type and the number of basis functions to use, the choice of the value of the smoothing parameter and the goodness of fit criterion. We show that the utilisation of the functional linear model provides a good quality of reconstruction to access high frequency variations of Chlorophyll *a* profiles at fine scale.

Mots-clés : Analyse de données fonctionnelle, modèle linéaire fonctionnel, splines, Chlorophylle a, luminosité

Keywords: Functional data analysis, functional linear model, splines, Chlorophyll a, light

Classification AMS 2000 : 62-07, 62P12, 65D07, 65D10, 86A05

[1] INRA, UR546 Biostatistique et Processus Spatiaux (BioSP), F-84914 AVIGNON, France.
E-mail : severine.bayle@avignon.inra.fr et E-mail : pascal.monestiez@avignon.inra.fr

[2] Institut Méditerranéen d'Océanographie (MIO), UMR 7294, Institut Pytheas (OSU), Aix-Marseille Université, Campus de Luminy, Case 901, 13288 MARSEILLE Cedex 09, France.
E-mail : david.nerini@univ-amu.fr

 ISSN: 2102-6238

1. Introduction

Depuis une trentaine d'années, l'analyse de données fonctionnelles est devenu un outil d'analyse statistique de plus en plus utilisé car les données fonctionnelles se retrouvent dans de nombreux domaines, comme le montrent les travaux de Brumback and Rice (1998) sur des données physiologiques , de Müller and Stadtmüller (2005) sur des données biologiques, ainsi que de Chiou and Müller (2009) sur des données démographiques.

Un des outils les plus prisés de l'analyse de données fonctionnelle est la régression fonctionnelle, qui peut notamment être appliquée dans le cas où la variable à prédire est fonctionnelle, et où l'ensemble des variables prédictives sont également des courbes. Cependant, malgré des recherches théoriques approfondies, ce type de modèle reste encore peu utilisé. En effet, le modèle de régression linéaire fonctionnel avec une réponse fonctionnelle ou continue a fait l'objet d'analyses théoriques par Ramsay and Dalzell (1991), Ramsay and Silverman (2005), Faraway (1997), Cardot et al. (1999), Cuevas et al. (2002), Fan and Zhang (2002) et Crambes and Mas (2013). Quelques applications ont été faites par Malfait and Ramsay (2003), Ramsay and Silverman (2005) et Yao et al. (2005).

L'adaptation de cette méthode de régression à des données fonctionnelles fait appel à l'estimation de deux coefficients de régression qui sont des fonctions univariée et bivariée. La modélisation des données brutes en courbes nécessite la définition de différentes fonctions de base qui implique des questionnements techniques, en particulier sur le type et le nombre de fonctions à choisir. De plus, la programmation de ce modèle (et de l'analyse de données fonctionnelles en général) reste complexe pour un non-statisticien.

Malgré ces aspects difficiles, Besse and Cardot (2003) ont identifié deux situations relativement fréquentes pour lesquelles la régression fonctionnelle est particulièrement bien adaptée et plus efficace qu'une approche vectorielle classique. Le premier cas concerne des courbes observées qui sont très régulières, et dont les variables issues de la discrétisation sont très corrélées deux à deux. Le second cas porte sur des données qui sont des observations bruitées d'un phénomène relativement régulier, et pour lesquelles il est important de faire intervenir une action de débruitage, et donc d'appliquer un lissage. Nous proposons dans cet article une application de la régression linéaire fonctionnelle avec réponse et prédicteur fonctionnels sur des données océanographiques.

Ces dernières années, un certain nombre de prédateurs marins capables de plonger à de grandes profondeurs ont été équipés de balises électroniques afin d'étudier leur comportement alimentaire. Ils sont de plus susceptibles de recueillir des données océanographiques sur de vastes secteurs de l'océan Austral (Charrassin et al., 2008; Boehlert et al., 2001; Fedak et al., 2002; Block et al., 2002; Biuw et al., 2007; McMahon et al., 2005). Des fluorimètres et des capteurs de lumière ont permis d'obtenir simultanément des mesures de Chlorophylle *a* (Chl *a*) et de luminosité, lors des nombreuses plongées de ces animaux le long de leur trajectoire géographique (Xing et al., 2012; Guinet et al., 2013). Cependant, en raison d'une forte consommation en énergie des balises, peu de profils de Chl *a* ont été échantillonnés, relativement aux autres variables échantillonnées.

Récemment, à l'aide d'éléphants de mer équipés de fluorimètres et de capteurs de lumière dans l'océan Antarctique, Jaud et al. (2012) ont montré que l'atténuation de la lumière est fortement corrélée avec la concentration en Chl *a* mesurée par le fluorimètre.

Dans la continuité de ce travail, nous montrons ici que le modèle linéaire fonctionnel peut être utilisé pour prédire avec une bonne précision les profils de concentration en Chl *a* à partir des

 ISSN: 2102-6238

données de lumière, échantillonnées fréquemment dans le temps et à différentes profondeurs. De plus, nous précisons certains choix méthodologiques afin de clarifier nos difficultés et questionnements rencontrés au cours de cette application : choix du type et du nombre de fonctions de base à utiliser, choix de la valeur du paramètre de lissage et justification d'un travail sur la monotonie et les dérivées des profils de lumière, ainsi que du critère pour évaluer la qualité de l'ajustement. Cette démarche est un support, généralisable à d'autres types de données fonctionnelles, pour les scientifiques susceptibles d'en analyser et se posant des questions méthodologiques, comme ce fut le cas ici.

2. Etude de cas : profils océanographiques de Chlorophylle *a* et de lumière

Dans cet article, nous présentons une application sur des données de Chl *a* et de luminosité qui ont été récoltées entre octobre 2009 et janvier 2010 à l'aide de balises fixées sur la tête de trois femelles éléphants de mer, dans la zone de l'Océan Austral s'étendant des îles Kerguelen jusqu'au plateau Antarctique. Ces balises ont permis d'échantillonner des données de fluorescence sur une profondeur allant de -180 mètres à la surface sur la phase ascendante de la plongée. D'autres capteurs permettant de mesurer des données de lumière toutes les deux secondes ont été placés sur le dos des éléphants de mer. Nous nous intéressons ici particulièrement aux données de Chl *a* (fluorimétrie) contenue dans les organismes photosynthétiques, car ceux-ci jouent un rôle essentiel de pompe à carbone. Chaque profil de Chl *a*, transmis via le système ARGOS, consiste en un maximum de dix-huit sections de dix mètres. La valeur moyenne de fluorescence est associée avec la profondeur médiane de chaque segment (-5 à -175 mètres). Cependant, en raison d'une forte consommation en énergie des balises, seulement deux profils de Chl *a* ont été échantillonnés quotidiennement. Nous nous intéressons alors dans cette étude à la relation entre les profils de Chl *a* et ceux de la lumière. Notre objectif est d'utiliser la lumière, variable océanographique échantillonnée fréquemment, pour prédire les profils de concentration en Chl *a* le long de la colonne d'eau traversée par les éléphants de mer.

Dans cette étude, seuls les profils de Chl *a* complets (c'est-à-dire avec dix-huit observations) ont été analysés afin d'avoir une meilleure estimation et un meilleur lissage des profils. Au total, sur les 436 profils échantillonnés, 407 ont été gardés et parmi eux, nous avons sélectionné ceux récoltés de jour, avec un angle solaire supérieur à 20° degrés au-dessus de l'horizon, pour les faire correspondre aux profils de lumière récoltés au même moment. Les profils récoltés de nuit ont été mis de côté puisque ceux de lumière sont composés de valeurs nulles. Finalement, 208 paires de profils de Chl *a* et de lumière ont été analysés (Tableau 1). Un modèle linéaire fonctionnel a été appliqué dans un premier temps sur ces 208 paires pour évaluer la qualité de la prédiction, puis dans un deuxième temps pour prédire la concentration en Chl *a* aux endroits où les profils de lumière n'ont pas de profils de Chl *a* correspondant. Les paramètres du modèle sont estimés séparément pour chaque éléphant car globalement, les mesures dépendent des zones dans lesquelles les animaux se déplacent. En effet, les éléphants 1 et 2 se dirigent vers le sud et le sud-est des îles Kerguelen, tandis que l'éléphant 3 se déplace vers le nord-ouest. Il est connu que les caractéristiques de l'océan austral dans ces différentes zones varient fortement. Il est donc problable que les paramètres (dans notre cas la lumière, mais aussi les types phytoplanctoniques, la température...) soient différents d'une zone à l'autre. Ces connaissances a priori des caractéristiques océaniques nous ont poussé à utiliser un modèle de prédiction par

 ISSN: 2102-6238

TABLE 1. *Tableau des différents éléphants de mer avec le nombre de profils enregistrés sur leur trajectoire respective.*

Numéro de l'éléphant de mer	Nombre total de profils de Chl *a* enregistrés avec 18 observations	Nombre de profils de Chl *a* considérés
1	142	73
2	148	73
3	117	62

éléphant (et donc par zone), et non pas un modèle global.

3. Méthodologie

3.1. Le modèle linéaire "pleinement" fonctionnel

Le modèle utilisé ici est le modèle "pleinement" fonctionnel (Ramsay and Silverman, 2005), qui est de la forme

$$y_i(t) = \alpha(t) + \int_0^T \beta(s,t)x_i(s)ds + \varepsilon_i(t)$$

où le paramètre $\beta(s,t)$ détermine l'effet de $x_i(s)$ (la lumière) au temps s (dans notre cas la profondeur) sur $y_i(t)$ (la Chl a) au moment t. Ce modèle donne la prédiction de $y_i(t)$ à partir de toutes les valeurs de $x_i(s)$ pour toutes valeurs de s. Il peut également être utilisé sur un sous-intervalle de temps de $[0,T]$ ou localement. Puisqu'il autorise une prédiction pour n'importe quel t, le modèle "pleinement" fonctionnel permet une resconstruction précise des courbes par lissage. De plus, l'utilisation de ce modèle apporte techniquement certains avantages. En effet, en fonction de la régularité des données échantillonnées, il permet de résumer chaque courbe à l'aide d'un nombre de fonctions de base restreint, ce qui permet de simplifier les calculs et d'éviter des problèmes de mauvais conditionnement par surparamétrisation. Nous proposons donc de travailler avec le modèle linéaire "pleinement" fonctionnel sur les 208 profils de lumière sélectionnés.

3.2. Estimation des paramètres

Supposons que l'on dispose d'un échantillon $\{(X_i,Y_i), i = 1,...,n\}$ défini par un ensemble de n paires de courbes échantillonnées à partir des variables fonctionnelles X et Y. Nous supposons que les fonctions appartiennent à $L^2[0;T]$, l'espace des fonctions de carré intégrable définies sur le segment $[0;T]$. Cet espace fonctionnel est muni du produit scalaire $\langle\cdot,\cdot\rangle$ et de la norme $\|\cdot\|$. Nous considérons un modèle de régression fonctionnelle dans lequel la fonction $X(s), s \in [0,T]$ est une variable utilisée pour expliquer les variations de la courbe réponse $Y(t), t \in [0,T]$. Entre autres, Besse and Cardot (1996), Bosq (2000), Cardot et al. (1999), James (2002) et Ramsay and Silverman (2005) considèrent le cas le plus général

$$Y = \alpha + B(X) + \varepsilon$$

où la fonction $\alpha(t)$ est un paramètre fonctionnel qui fait office d'ordonnée à l'origine et B est un opérateur linéaire tel que

$$B(X)(t) = \int_0^T \beta(s,t)X(s)\,ds.$$

 ISSN: 2102-6238

La fonction bivariée $\beta(s,t)$, noyau de l'opérateur B, agit comme un coefficient de régression qui donne l'influence de $X(s)$ sur $Y(t)$ en toute valeur de t. La fonction résiduelle ε donne l'erreur entre le modèle supposé et la variable réponse Y.

Les estimateurs $\hat{\alpha}$ et $\hat{B}$ sont déterminés par minimisation des moindres carrés de l'espérance

$$SSE(\alpha,B) = \mathbb{E}\|\varepsilon_i\|^2.$$

La solution du problème de minimisation amène alors aux équations normales pour le modèle linéaire fonctionnel dont l'allure générale rappelle celle du cas classique multivarié :

$$\left\{\begin{array}{rcl} V_X B^\star & = & V_{XY} \\ \alpha^\star & = & \mu_Y - B^\star(\mu_X) \end{array}\right. \tag{1}$$

où les fonctions μ_X et μ_Y sont respectivement les espérances des variables X et Y, V_X est l'opérateur de variance-covariance pour la variable X et V_{XY} est l'opérateur de covariance croisée entre les variables X and Y. Les estimateurs empiriques pour les opérateurs de covariance et pour les fonctions moyennes sont exprimés à partir de l'échantillon de la manière suivante :

$$\left\{\begin{array}{l} \hat{V}_X = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \hat{\mu}_X)\otimes(X_i - \hat{\mu}_X) \\ \hat{V}_{XY} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \hat{\mu}_X)\otimes(Y_i - \hat{\mu}_Y) \\ \hat{\mu}_X = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i \\ \hat{\mu}_Y = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}Y_i \end{array}\right.$$

où le produit tensoriel de deux élements A et B est un opérateur tel que

$$A\otimes B(F) = \langle A,F\rangle B$$

pour tout $F \in \mathbb{L}^2[0;T]$.

Les coefficients $\hat{B}$ et $\hat{\alpha}$ sont alors obtenus en substituant ces estimateurs empiriques dans (1) de telle manière que :

$$\left\{\begin{array}{rcl} \hat{V}_X\hat{B} & = & \hat{V}_{XY} \\ \hat{\alpha} & = & \hat{\mu}_Y - \hat{B}(\hat{\mu}_X) \end{array}\right. .$$

Une façon élégante d'effectuer les calculs ci-dessus est de décomposer les fonctions X et Y comme une combinaison linéaire de fonctions de base fixées à l'avance. Cependant, le problème majeur qui se pose est celui du calcul de $\hat{\beta}(s,t)$ c'est à dire celui du calcul d'un inverse pour $\hat{V}_X$. Ramsay and Silverman (2005) proposent une méthode qui permet de résoudre le problème d'estimation via une méthode de régularisation permettant de contrôler la forme de $\hat{\beta}(s,t)$ lors du processus d'estimation Malfait and Ramsay (2003). Une façon simple d'effectuer les calculs consiste à décomposer $\hat{\beta}(s,t)$ dans une base finie de fonctions bivariées, ce qui permet d'assurer sa régularité en controlant le nombre de fonctions de base. Le problème d'estimation est ensuite ramené à un problème classique multivarié par échantillonnage à différentes profondeurs des

 ISSN: 2102-6238

fonctions de l'échantillon de départ. C'est cette technique que nous employons et qui est proposée dans le package *fda* de R (fonction *linmod*). Cependant, de nouvelles méthodes, plus générales, sont proposées depuis peu par Crambes and Mas (2013), Lian (2012) et He et al. (2010). Ces méthodes sont basées sur le calcul d'un inverse pour $\hat{V}$ en régularisant de différentes façons et proposent différentes estimations de $\hat{B}$.

4. Application : inférence et choix de modélisation

4.1. Choix des fonctions de base pour l'ajustement fonctionnel des données de Chl a et de lumière

Afin de faire apparaître la forme fonctionnelle des données, il est nécessaire dans un premier temps d'effectuer un ajustement (ou lissage) sur ces dernières à l'aide de fonctions de base. Dans notre cas, nous avons choisi d'utiliser des fonctions splines qui conviennent tout particulièrement aux données non périodiques et aux fonctions. Elles sont définies par un intervalle de validité (dans notre cas l'ensemble des profondeurs sur lequel sont mesurées les données, de -5 à -175 mètres), des noeuds et un ordre. Il existe différentes sortes de splines, et nous avons choisi de travailler avec le système de base le plus courant pour contruire les fonctions splines, le système de base B-spline. Les autres possibilités sont les M-splines, les I-splines et les fonctions puissances par morceaux (Ramsay et al., 2009). Pour de plus amples discussions sur l'utilisation des splines, voir de Boor (2001) et Schumaker (2007). Le système B-spline a été l'objet d'études théoriques, notamment par Cardot et al. (2003), et a été mis en application par Abraham et al. (2003).

Le plus souvent, ce sont des B-splines d'ordre 4 qui sont utilisées pour la modélisation des courbes, comme dans Abraham et al. (2003), et qui consistent en des segments polynômiaux cubiques. Nous avons également fait le choix d'utiliser ici des B-splines d'ordre 4, car l'utilisation d'un plus grand ordre ou d'un trop grand nombre de fonctions conduit à un sur-ajustement des courbes sur les données.

De ce fait, le choix du nombre de B-splines à utiliser a été choisi par la méthode de validation croisée "leave-one-out" afin de minimiser l'erreur de prédiction entre un profil de Chl *a* ajusté $y_i(t)$ et un profil de Chl *a* prédit $\hat{y}_i(t)$.

Cette méthode consiste à diviser n fois l'échantillon (n étant le nombre total de profil considérés par éléphant), à sélectionner un des n éléments de l'échantillon comme ensemble de validation et les $(n-1)$ autres éléments constitueront l'ensemble d'apprentissage. On calcule alors la somme des erreurs quadratiques, telle que

$$SSE = \sum_{i=1}^{n-1} \int (y_i(t) - \hat{y}_i(t))^2 dt.$$

On répète l'opération en sélectionnant un autre élément de l'échantillon de validation parmi les $(n-1)$ éléments qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi n fois pour qu'en fin de compte chaque sous-échantillon de taille $n-1$ ait été utilisé exactement une fois comme ensemble de validation. La moyenne des n sommes d'erreurs quadratiques est alors calculée pour estimer l'erreur de prédiction, c'est-à-dire

$$SSE_{moy} = \frac{1}{n} \sum^{n} SSE.$$

 ISSN: 2102-6238

TABLE 2. *Tableau des différents nombres de fonctions de base nécessaires pour ajuster les profils de Chl a et lumière afin de minimiser l'erreur de prédiction entre un profil de Chl a prédit et un profil de Chl a mesuré, selon la méthode de validation croisée.*

Numéro de l'éléphant de mer	Nombre de B-splines pour ajuster les profils de lumière	Nombre de B-splines pour ajuster les profils de Chl *a*
1	5	5
2	6	4
3	4	4

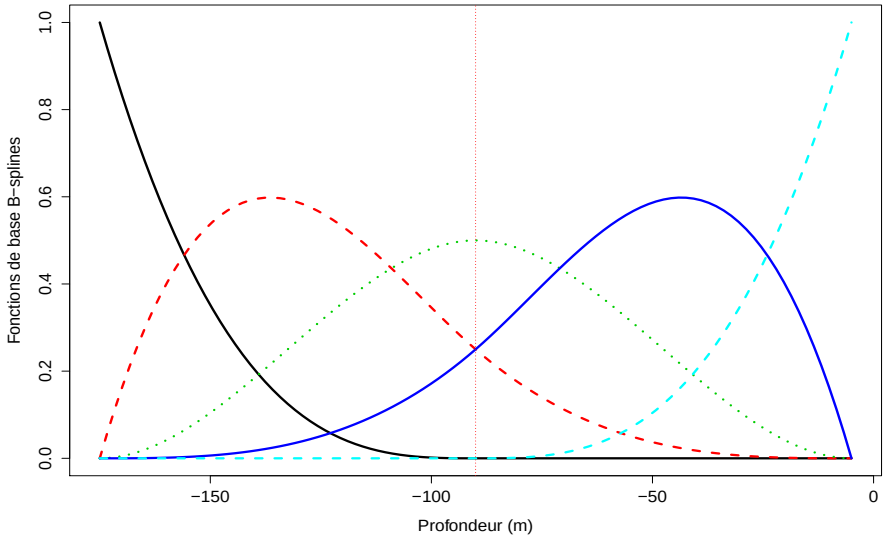


FIGURE 1. *Système de 5 B-splines d'ordre 4 défini sur l'intervalle [-175,0] et utilisé pour l'application. La ligne rouge en pointillés correspond au noeud. A cet endroit, les valeurs des polynômes et celles de leurs deux premieres dérivées doivent se correspondre.*

Les résultats détaillés de cette méthode sont regroupés dans le Tableau 2. Puisque le nombre de fonctions de base à utiliser diffère très peu d'un éléphant à un autre, et puisque nous avons constaté que pour les éléphants de mer 2 et 3, les erreurs de prédiction calculées en utilisant 5 fonctions de base pour ajuster la lumière et la Chl *a* sont très proches de celles obtenues pour les nombres de fonctions de base optimaux à utiliser (Tableau 2), nous avons choisi de généraliser les résultats obtenus sur l'éléphant 1 et d'utiliser également 5 B-splines sur les trajectoires des éléphants 2 et 3 pour modéliser les profils de Chl *a* et de lumière.

Le système de 5 B-splines d'ordre 4 utilisé dans notre application est illustré dans la Figure 1.

La Figure 2 montre quelques exemples d'ajustements de profils de Chl *a*. Les quatre premiers ajustements (Fig. 2.1 à Fig. 2.4) montrent des profils de Chl *a* bien ajustés. La Figure 2.5 désigne un profil avec un ajustement prenant des valeurs négatives en profondeur. Bien que la plupart des profils aient une forme relativement identique, les profils de Chl *a* à forme non régulière (Fig. 2.6) ont tendance à être mal ajustés. Ces différents types de problèmes peuvent être résolus en augmentant le nombre de fonctions de base afin d'ajuster au mieux les données, mais au

 ISSN: 2102-6238

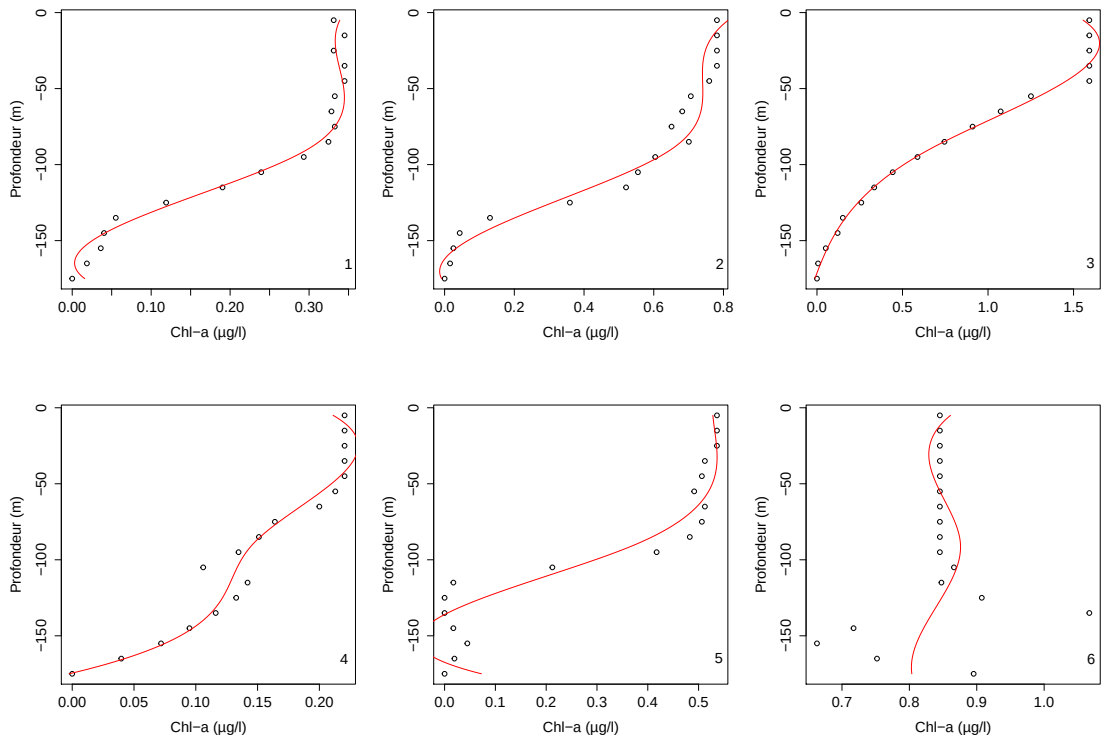


FIGURE 2. *Exemples d'ajustements de profils de Chl a avec 5 B-splines. Les ronds noirs désignent les données mesurées et la courbe rouge représente les données ajustées. Les figures 1, 2, 3 et 4 montrent des profils de Chl a bien ajustés. Les graphiques 5 et 6 présentent des profils avec différents problèmes de modélisation : une courbe dont l'ajustement admet des valeurs négatives et une courbe pour laquelle 5 B-splines sont insuffisantes pour bien ajuster.*

risque d'accroître l'erreur de prédiction. On peut également déplacer les noeuds par lesquels les ajustements doivent forcément passer.

4.2. Choix de la monotonie des profils de lumière et de la valeur du paramètre de lissage

Nous avons choisi de travailler avec les dérivées des courbes de lumière (voir paragraphe 4.3). Afin de rendre celles-ci positives, un lissage avec contrainte de monotonie a été imposé sur les données de lumière en même temps que l'ajustement par B-splines (Ramsay, 1988), car la quantité de luminosité croît lorsqu'un éléphant remonte vers la surface. Le fait de rendre la courbe monotone permet d'avoir un profil croissant, et en conséquence une dérivée à valeurs positives, en cohérence avec les lois de l'optique.

Dans le lissage des profils de lumière, un paramètre de lissage λ permet d'effectuer un arbitrage entre la fidélité aux données ($\lambda \to 0$) et la régularité de la solution ($\lambda \to +\infty$). Il intervient sur la dérivée seconde des profils fonctionnels de lumière et pénalise la courbe afin de la rendre plus lisse. Dans notre étude, λ a été choisi empiriquement en testant différentes valeurs possibles. Au vu de la forme des courbes de lumière déjà relativement monotone, prendre une valeur de

 ISSN: 2102-6238

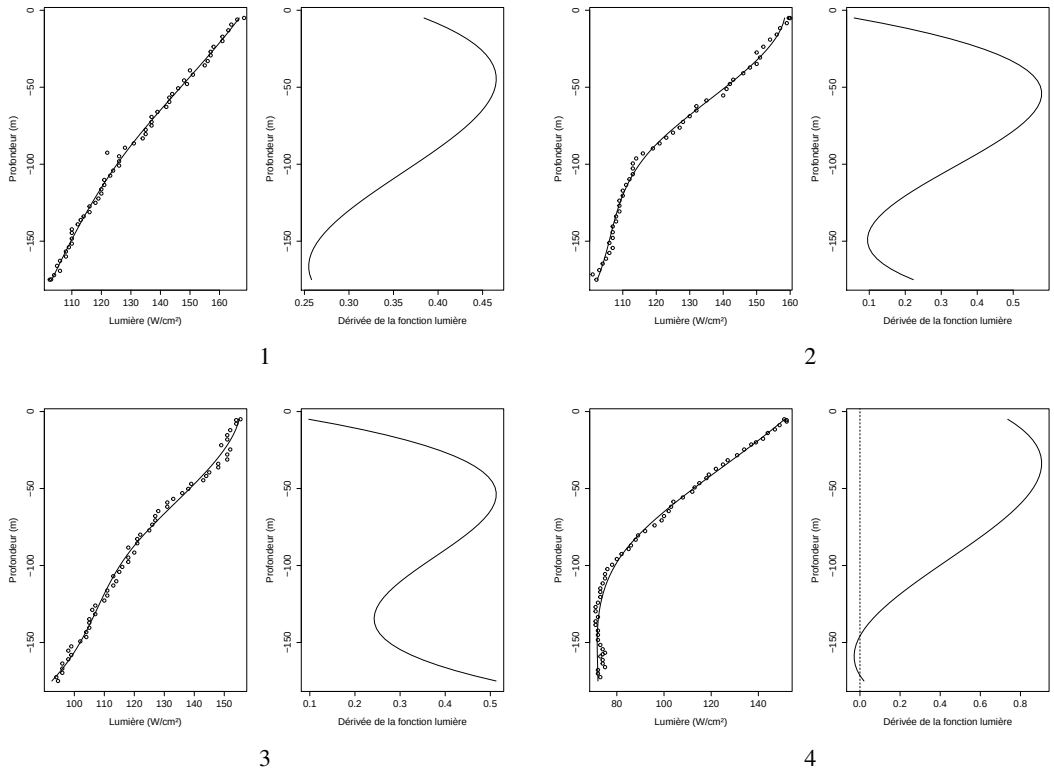


FIGURE 3. *Exemples de profils de lumière récoltés sur la trajectoire de l'éléphant 1, ajustés avec 5 B-splines et avec une contrainte de monotonie, avec leurs dérivées respectives. Pour chaque ensemble de graphiques, les figures de gauche se composent des données brutes (points ronds) et de l'ajustement, et les figures de droite représentent les dérivées associées. L'unité de la lumière est une caractéristique du capteur et s'étend de 0 à 250 en échelle logarithmique. Cette valeur est calibrée selon Jaud et al. (2012).*

$10^{-0.1}$ nous a paru raisonnable. Cependant, en régression non paramétrique, λ peut être choisi en minimisant le critère de validation croisée généralisée (Wahba, 1990). Cette méthode a déjà été utilisée par Hosseini-Nasab (2012) en régression linéaire fonctionnelle.

La Figure 3 montre des exemples de profils de lumière ajustés avec une contrainte de monotonie. Cependant, il reste des cas pour lesquels cette contrainte ne s'applique pas bien (Fig. 3.4) et apporte en profondeur des valeurs négatives sur les courbes de dérivées. En effet, il y a un problème de monotonie pour 13% des profils de lumière récoltés sur la trajectoire de l'éléphant 1, pour seulement 1% des profils enregistrés sur la trajectoire de l'éléphant 2 et pour environ 10% des profils du troisième éléphant. Les valeurs négatives peuvent s'étendre jusqu'à -0.3 pour l'éléphant 3 (Fig. 4). Cela est dû à un problème de reprojection de base. Une solution pour remédier à ce problème serait d'utiliser une base plus riche que celle des B-splines car celle-ci peut être trop simple pour ajuster la courbure de la fonction monotone. Par exemple, on peut envisager l'utilisation de la base des M-splines, qui sont des fonctions splines non-négatives.

 ISSN: 2102-6238

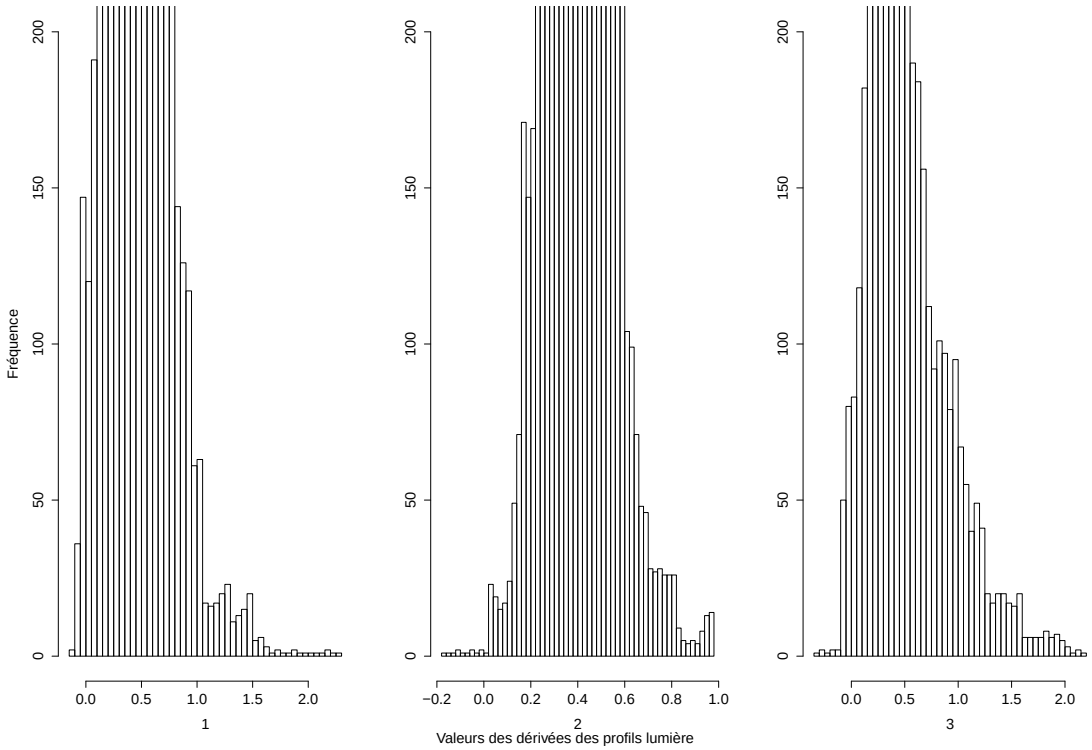


FIGURE 4. *Histogrammes des valeurs des dérivées des profils de lumière pour les éléphants 1, 2 et 3. Les graphiques mettent en évidence que les valeurs négatives sur les courbes des dérivées sont relativement peu nombreuses. Ils ont été volontairement coupés en haut, afin de mieux voir ces valeurs.*

4.3. Choix de travailler sur les dérivées

Le choix de travailler sur les dérivées des fonctions lumière repose sur une logique physique. En effet, depuis 1852, la luminosité dans un milieu liquide est définie grâce à la relation de Beer-Lambert (Bouguer, 1729) qui relie l'absorption de la lumière aux propriétés d'un environnement homogène. L'intensité d'un faisceau lumineux subit une décroissance exponentielle en fonction de la distance parcourue et de la densité des espèces absorbantes dans ce milieu selon la relation

$$L(z,\theta) = L_0(\theta)\exp(-kz)$$

où L désigne l'intensité lumineuse, z est la profondeur (en mètres), θ est la longueur d'onde du rayonnement (en mètres), L_0 représente l'intensité lumineuse en surface, et k est le coefficient d'absorption de la lumière dans le liquide considéré. Le niveau de lumière dans notre étude est exprimé dans une échelle logarithmique, donc

$$log(L(z,\theta)) = log(L_0(\theta)) - kz.$$

 ISSN: 2102-6238

Quand on dérive ce calcul par rapport à la profondeur z, il reste $-k$. Le lien entre ce coefficient et la Chl a est donc un coefficient de proportionalité ω. Soit C la Chl a. Nous obtenons alors

$$C(z) = \omega k.$$

Cependant, les prédictions issues de la régression linéaire sans les dérivées ne montrent pas de différence de résultat car nous travaillons avec le modèle "pleinement" fonctionnel (explications au paragraphe 3.1). Mais attention, avec un modèle concurrent, les résultats peuvent être différents. L'utilisation des dérivées peut s'avérer utile dans l'analyse de phénomènes similaires, comme par exemple des processus de diffusion.

4.4. Application du modèle linéaire fonctionnel

Nous travaillons avec le modèle linéaire "pleinement" fonctionnel sur les 208 profils de lumière sélectionnés au total sur les 3 éléphants. Nous rappelons que nous avons utilisé un modèle de prédiction par éléphant.

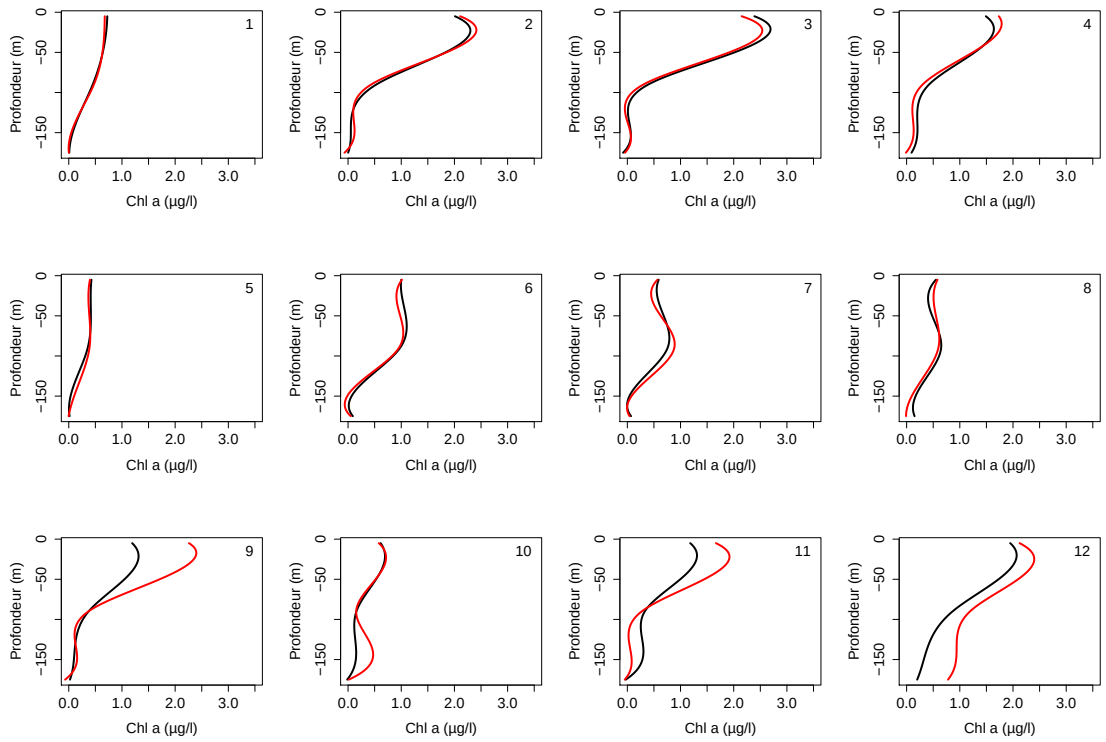


FIGURE 5. *Exemples de prédiction de profils de Chl a. Les courbes rouges montrent les profils modélisés à partir des mesures directes avec 5 B-splines d'ordre 4, et les courbes noires représentent les profils prédits avec le modèle linéaire "pleinement" fonctionnel. Les profils 1, 2, 3, 4, 9 et 12 ont été pris sur la trajectoire de l'éléphant 1 ; les profils 5, 6, 7, 10 et 11 ont été enregistrés par l'éléphant 2, et le profil 8 est pris sur l'itinéraire de l'éléphant 3.*

 ISSN: 2102-6238

Quelques exemples de résultats sont représentés dans la Figure 5. Les huit premiers graphiques (Fig. 5.1 à Fig. 5.8) montrent des profils de Chl *a* bien prédits. Les courbes rouges (profils modélisés) et les courbes noires (profils prédits) ont la même forme et les quantités de Chl *a* prédites par le modèle correspondent plutôt bien aux quantités modélisées.

Cependant, quatre types de difficultés peuvent être distingués. Dans un premier temps, il existe des problèmes de prédiction en surface et en profondeur (Fig. 5.9 et Fig. 5.10). Les erreurs de prédiction en surface sont fréquentes, elles se retrouvent dans 20% des cas sur la trajectoire de l'éléphant 1, 17% des cas pour l'éléphant 2, et sont quasi inexistantes sur le troisième itinéraire. Au contraire, les erreurs de prévision en profondeur ont une prévalence très faible (1,3% pour l'éléphant 2 et 3,2% pour l'éléphant 3). D'un point de vue océanographique, ce schéma est susceptible d'être la conséquence d'une plus grande atténuation de la lumière. En effet, les particules en suspension dans la colonne d'eau, autre que le phytoplancton, peuvent contribuer à l'atténuation de la lumière, telles que les particules organiques et le zooplancton. De plus, la fluorescence n'est qu'un proxy de la concentration en phytoplancton et elle peut varier selon les espèces de phytoplancton rencontrées et l'état physiologique du milieu (Xing et al., 2012). Par conséquent, pour une concentration en Chl *a* estimée à partir du fluorimètre, on peut s'attendre à une variation du coefficient d'atténuation de la lumière en fonction des espèces de phytoplancton. D'un point de vue statistique, ces erreurs sont liées à un problème d'adéquation du modèle à la réalité. Afin de réduire les erreurs de prédiction, il serait intéressant de prendre en compte d'autres variables se rapportant également au processus modélisé, ce qui dans notre cas pourrait correspondre à la température. L'ajout de variables supplémentaires dans le modèle peut s'avérer utile dans les cas où les erreurs de prédiction sont non négligeables, mais il faut faire attention à disposer de données de validation suffisantes pour ne pas risquer les sur-ajustements.

Dans un deuxième temps, les profils représentés dans les Figures 5.11 et 5.12 sont mal prédits par rapport à l'ensemble du profil modélisé. Nous pouvons voir que l'erreur de prédiction entre les profils modélisés et prédits est importante. Soit une intersection entre les courbes est observée (Fig. 5.11), mettant en avant un problème de forme, soit la totalité du profil prédit est décalée par rapport au profil modélisé (Fig. 5.12), et c'est un problème d'offset : quand le capteur ne peut pas mesurer la concentration en Chl *a*, il renvoie la valeur 0. Les croisements entre les courbes se retrouvent assez fréquemment sur les trois trajectoires (respectivement 15, 10 et 16% des cas). Le problème de décalage est plus limité. Il apparaît dans seulement 6% des cas sur la trajectoire de l'éléphant 1, 1% des cas sur l'itinéraire de l'éléphant 2 mais il représente l'erreur la plus fréquente sur la troisième trajectoire : 30% des paires de profils sont concernées par ce problème. Ces erreurs mettent en avant la question de la sensibilité de l'ajustement aux valeurs aberrantes. Deux solutions peuvent alors être envisagées. Premièrement, on peut simplement enlever de la régression des profils mal modélisés par les fonctions de base, mais ceci entraine une réduction de l'échantillon de courbes étudié. Deuxièmement, lors de l'ajustement des fonctions de base sur les données brutes, il peut être envisagé d'effectuer un lissage en excluant ces données aberrantes.

4.5. Choix du critère pour évaluer la qualité de l'ajustement

Il y a différentes façons d'évaluer la qualité de l'ajustement d'un modèle linéaire fonctionnel. Ramsay and Silverman (2005) proposent d'évaluer la qualité de l'ajustement à l'aide de la fonction

 ISSN: 2102-6238

de corrélation au carré

$$R^2(t) = \frac{\sum (y_i(t) - \bar{y}(t))^2 - \sum (\hat{y}_i(t) - y_i(t))^2}{\sum (y_i(t) - \bar{y}(t))^2}$$

où $y_i(t)$ est un profil de Chl *a* modélisé, $\bar{y}(t)$ est la valeur moyenne de tous ces profils et $\hat{y}_i(t)$ un profil de Chl *a* prédit. Cette formule correspond dans notre cas à un calcul de R^2 moyen par profondeur, mais n'a pas vraiment de sens dans notre application car elle juge la qualité de la forme de la courbe. Or, par exemple, la Figure 2.6. montre un mauvais ajustement de la courbe sur les données. Nous n'utiliserons donc pas cette formule ici, mais elle peut être utile dans d'autres cas si l'on souhaite calculer une mesure de l'ajustement en un point précis (par exemple un jour ou un mois si on analyse des données temporelles).

Une approche complémentaire pour mesurer l'ajustement est de considérer une mesure de R^2 globale pour chaque individu fonctionnel (c'est-à-dire dans notre cas un profil de Chl *a*), défini par

$$R_i^2 = \frac{\int (y_i(t) - \bar{y}(t))^2 dt - \int (\hat{y}_i(t) - y_i(t))^2 dt}{\int (y_i(t) - \bar{y}(t))^2 dt}. \tag{2}$$

Nous proposons également d'utiliser la formule du R^2 classique

$$R_i^2 = \frac{\int (y_i(t) - \bar{y}_i)^2 dt - \int (\hat{y}_i(t) - y_i(t))^2 dt}{\int (y_i(t) - \bar{y}_i)^2 dt} \tag{3}$$

dans laquelle $\bar{y}_i$ est toujours la valeur moyenne de tous les profils de Chl *a* modélisés, mais ce profil moyen est maintenant constant. Il ne dépend plus de la profondeur.

Chacune des formules (2) et (3) mesure respectivement l'apport de la prédiction issue du modèle par rapport à l'utilisation d'un profil moyen et d'une constante. Cependant, les résultats obtenus dans les deux cas diffèrent.

Avec la formule (2), plus de 50% des 73 paires de profils ont un R^2 supérieur à 0.71 pour l'éléphant 1. Huit paires ont des valeurs négatives (-0.008, -0.07, -0.13, -0.23, -0.43, -1.13, -2.15, -21.77), ce qui signifie que pour ces courbes, le profil moyen $\bar{y}$ donne un meilleur ajustement de y_i que le prédicteur $\hat{y}_i$. Pour l'éléphant 2, 50% des 73 paires de profils ont un R^2 supérieur à 0.52. Il y a cependant vingt-six paires de profils avec un R^2 négatif et des valeurs comprises entre -0.09 et -14.96. De même, pour l'éléphant 3, 50% des 62 paires de profils ont un R^2 supérieur à 0.52. Treize paires ont aussi une valeur de R^2 négative (valeurs comprises entre -0.0001 et -1.07).

En ce qui concerne la formule (3), plus de 50% des 73 paires de profils ont un R^2 supérieur à 0.93 sur la trajectoire de l'éléphant 1. Pour l'éléphant 2, 50% des 73 paires de profils ont un R^2 supérieur à 0.87. Il y a sur cette trajectoire une paire de profils avec une valeur de R^2 négative (-0.94). Enfin, pour ce qui est du troisième éléphant, 50% des 62 paires ont un R^2 supérieur à 0.85. Ici, une petite proportion de paires de profils a aussi un R^2 négatif (-0.09, -0.47, -0.86, -3,26).

L'emploi de l'une ou l'autre des formules donne un ajustement globalement correct. Cependant, les R^2 obtenus avec la formule (3) sont meilleurs que ceux obtenus avec la formule (2). Ceci vient du fait qu'avec la formule (3), nous mesurons par rapport à une constante s'il y a de forts écarts entre les deux courbes. Or, si cet écart est important, la proportion de variation expliquée par la droite constante sera elle aussi importante, ce qui explique les valeurs de R^2 élevées.

 ISSN: 2102-6238

4.6. Prédiction de Chlorophylle a à fine échelle

Comme nous avons obtenu une erreur de reconstitution relativement bonne des profils de Chl *a* entre les profils modélisés (à partir des profils mesurés) et les profils prédits, nous avons alors essayé de prédire la concentration en Chl *a* à des endroits où aucune quantité de Chl *a* n'a été mesurée, à partir des dérivées des profils de lumière (ces derniers étant également modélisés avec 5 B-splines d'ordre 4). La Figure 6 montre des exemples de profils de Chl *a* prédits sur deux différentes journées. La Figure 6.1 représente la prédiction pour la journée du 27 octobre 2009, et la Figure 6.2 se rapporte au 9 novembre de la même année. Cette prédiction met en avant des variations à fine échelle.

Afin de prendre en compte la dimension spatiale des profils, et ainsi prédire la concentration en Chl *a* dans les zones traversées de nuit par les éléphants de mer, une interpolation par krigeage fonctionnel (Nerini et al., 2010) peut être utilisée.

5. Discussion

L'application traitée dans cet article nous incitait à prendre en compte le caractère fonctionnel des données observées. Ces données impliquaient une analyse méthodologique particulière, que nous avons développé ici. Cette méthodologie peut être étendue à n'importe quel type de données fonctionnelles (météorologiques, océanographiques...) et propose au lecteur une base de questionnements pour débuter une analyse de ses propres données.

Notre procédure est bien adaptée pour prédire les profils de Chl *a* à partir des données de lumière pour déduire les changements de concentration en Chl *a* à fine échelle. Cette prédiction est faite pour chaque plongée se déroulant de jour le long des trajectoires de trois éléphants de mer. La reconstruction des profils de Chl *a* à partir des mesures de lumière, et plus particulièrement la question de la qualité de telles reconstructions, a été posée par Jaud et al. (2012).

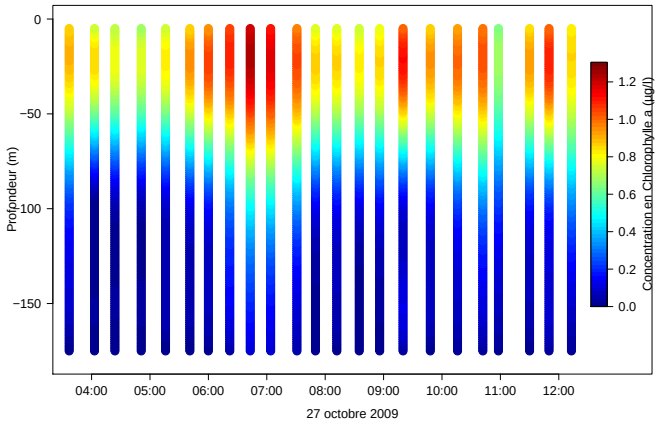
Le modèle linéaire fonctionnel que nous avons utilisé donne la prédiction de $y_i(t)$ à partir de toutes les valeurs de $x_i(s)$. Cependant, lorqu'on travaille sur des données fonctionnelles s'inscrivant dans le temps, il faut faire attention à l'utilisation de $x_i(s)$ pour prédire $y_i(t)$ quand $s > t$. En effet, cela implique des causalités avec le temps passé. Pour éviter cette contradiction, il faut considérer seulement les valeurs de x_i avant le temps t. Il est également possible de rajouter une restriction sur un paramètre de délai δ estimé à partir des données, tel que $s \in [t-\delta, t]$. Malfait and Ramsay (2003) accordent une attention particulière à ce cas.

De plus, Ramsay and Silverman (2005) distinguent un autre type de modèle linéaire fonctionnel, appelé modèle fonctionnel concurrent (ou modèle "point par point")

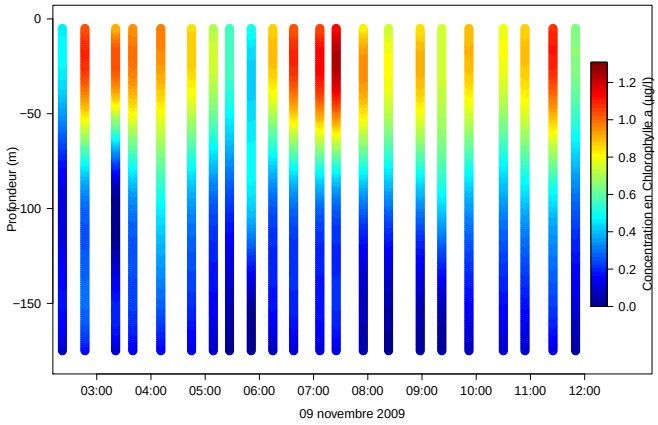
$$y_i(t) = \alpha(t) + x_i(t)\beta(t) + \varepsilon_i(t)$$

qui permet de faire la prédiction de la variable $y_i(t)$ au même moment $s = t$ (dans notre cas à la même profondeur). Ce modèle aurait pu être intéressant à utiliser si nous avions considéré que la variation de la concentration en Chl *a* dependait uniquement de la valeur de la lumière à la même profondeur. Or, dans le cadre d'une logique physique, nous considérons que cette variation peut dépendre de relations entre les différentes couches de profondeur, c'est pour cela que nous nous sommes tournés vers le modèle "pleinement" fonctionnel. De plus, à l'inverse de ce dernier, le

 ISSN: 2102-6238



1



2

FIGURE 6. *Deux exemples de prédiction de concentration en Chl a entre deux profils de Chl a enregistrés par les balises, montrant bien la dimension spatio-temporelle des profils. Plus la couleur est rouge, plus la concentration en Chl a est élevée.*

modèle concurrent s'avère moins précis puisqu'il nécessite une interpolation point par point pour modéliser un profil.

Le calcul du paramètre de lissage des splines par validation croisée généralisée en régression non paramétrique comporte certaines limites. En effet, une valeur correcte pour λ ne peut pas être attendue sur de petits échantillons, car il n'y a pas assez d'information dans les données pour séparer le signal du bruit (Wahba, 1990). De même, la méthode ne donnera pas un bon résultat si les erreurs sont corellées.

Enfin, les valeurs négatives obtenues dans les différents calculs des R^2 résultent de numérateurs négatifs quand la variance totale a une valeur plus petite que la somme des erreurs au carré. Cela peut arriver quand un profil de Chl *a* prédit est plat. Afin de réduire la fréquence de ce type de valeurs, on peut envisager de les traiter à part.

 ISSN: 2102-6238

Remerciements

Nous tenons à remercier la région Provence-Alpes-Côte-d'Azur pour le co-financement de la thèse de Séverine Bayle, l'ANR pour le programme IPSOS-SEAL ainsi que Christophe Guinet (CEBC-CNRS) pour le recueil et la mise à disposition des données.

Références

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30(3) :581–595.

Besse, P. and Cardot, H. (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24(4) :467–487.

Besse, P. and Cardot, H. (2003). *Modélisation statistique de données fonctionnelles*, chapter 6, pages 169–200. Traitement du Signal et de l'Image. Lavoisier.

Biuw, M., Boehme, L., Guinet, C., Hindell, M., Costa, M., Charrassin, J.-B., Roquet, F., Bailleul, F., Meredith, M., Thorpe, S., Tremblay, Y., McDonald, B., Park, Y.-H., Rintoul, S., Bindoff, N., Goebel, M., Crocker, D., Lovell, P., Nicholson, J., Monks, F., and Fedak, M. A. (2007). Variations in behavior and condition of a southern ocean top predator in relation to in situ oceanographic conditions. *Proceedings of the National Academy of Sciences*, 104(34) :13705–13710.

Block, B., Costa, D., Boehlert, G., and Kochevar, R. (2002). Revealing pelagic habitat use : the tagging of pacific pelagics program. *Oceanologica Acta*, 25(5) :255–266.

Boehlert, G., Costa, D., Crocker, D., Green, P., O'Brien, T., Levitus, S., and Le Boeuf, B. (2001). Autonomous pinniped environmental samplers : using instrumented animals as oceanographic data collectors. *Journal of atmospheric and oceanic technology*, 18(11) :1882–1893.

Bosq, D. (2000). *Linear processes in function spaces : theory and applications*, volume 149. Springer.

Bouguer, P. (1729). *Essai d'optique sur la gradation de la lumière*. Claude Jombert, Paris.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443) :961–976.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45 :11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13 :571–591.

Charrassin, J., Hindell, M., Rintoul, S., Roquet, F., Sokolov, S., Biuw, M., Costa, D., Boehme, L., Lovell, P., Coleman, R., Timmermann, R., Meijers, A., Meredith, M., Park, Y.-H., Bailleul, F., Goebel, M., Tremblay, Y., Bost, C.-A., McMahon, C., Field, I., Fedak, M., and Guinet, C. (2008). Southern ocean frontal structure and sea-ice formation rates revealed by elephant seals. *Proceedings of the National Academy of Sciences*, 105(33) :11634 –11639.

Chiou, J.-M. and Müller, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association*, 104(486) :572–585.

Crambes, C. and Mas, A. (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*. A paraître.

Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression : the case of fixed design and functional response. *Canadian Journal of Statistics*, 30(2) :285–300.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York.

Fan, J. and Zhang, J.-T. (2002). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 62(2) :303–322.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3) :254–261.

Fedak, M., Lovell, P., McConnell, B., and Hunter, C. (2002). Overcoming the constraints of long range radio telemetry from animals : getting more useful data from smaller packages. *Integrative and Comparative Biology*, 42(1) :3–10.

Guinet, C., Xing, X., Walker, E., Monestiez, P., Marchand, S., Picard, B., Jaud, T., Authier, M., Cotté, C., Dragon, A.-C., Diamond, E., Antoine, D., Lovell, P., Blain, S., D'Ortenzio, F., and Claustre, H. (2013). Calibration procedures and first data set of southern ocean chlorophyll-a profiles collected by elephant seal equipped with a newly developed ctd-fluorescence tags. *Earth System Science Data*, 5 :15–29.

He, G., Müller, H.-G., Wang, J.-L., and Yang, W. (2010). Functional linear regression via canonical analysis. *Bernoulli*, 16(3) :705–729.

 ISSN: 2102-6238

Hosseini-Nasab, M. (2012). Cross-validation approximation in functional linear regression. *Journal of Statistical Computation and Simulation*, pages 1–11.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :411–432.

Jaud, T., Dragon, A.-C., Garcia, J., and Guinet, C. (2012). Relationship between chlorophyll a concentration, light attenuation and diving depth of the southern elephant seal mirounga leonina. *PLoS one*, 7(10) :e47444.

Lian, H. (2012). Minimax prediction for functional linear regression with functional responses in reproducing kernel hilbert spaces. *arXiv preprint arXiv :1211.4080*.

Malfait, N. and Ramsay, J. (2003). The historical functional linear model. *The Canadian Journal of Statistics*, 31 :115–128.

McMahon, C. R., Autret, E., Houghton, J., Lovell, P., Myers, A., and Hays, G. (2005). Animal-borne sensors successfully capture the real-time thermal properties of ocean basins. *Limnology and Oceanography : Methods*, 3 :392–398.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33(2) :774–805.

Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101 :409–418.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, pages 425–441.

Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–572.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Schumaker, L. (2007). *Splines Functions : Basic Theory*. Cambridge University Press, New York.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Society for Industrial Mathematics.

Xing, X., Claustre, H., Blain, S., D'Ortenzio, F., Antoine, D., Ras, J., and Guinet, C. (2012). Quenching correction for in vivo chlorophyll fluorescence measured by instrumented elephant seals in the kerguelen region (southern ocean). *Limnology and Oceanography : Method*, 10 :483–495.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6) :2873–2903.

 ISSN: 2102-6238