ECOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION

STAGE DE FIN D'ÉTUDE

pour l'entreprise DataStorm

Estimation adaptative en analyse des données fonctionnelles

rédigé par
Hugo Brunet
Tuteur
Hassan Maissoro

Avril—Septembre 2023

Résumé

Les séries temporelles sont des données omniprésentes dans l'analyse et la prédiction de données. Elles concernent de nombreux secteurs critiques allant du secteur de l'énergie à la finance. Leur étude systématique depuis 1927 (Yule) est ainsi motivée par leur importance et utilité pour la mise en production.

Les données fonctionnelles quant à elles sont particulièrement présentes dans les données de capteurs ou à composante temporelle. Elles possèdent grâce au point de vue qu'elles offrent, d'obtenir notamment de meilleures estimation sur le long terme que le point de vue réel multivarié classique. Cependant, la littérature jusqu'alors ne prenait pas en compte les différences de régularité des données traitées, ce qui pose problème pour des données peu régulières pourtant fréquemment observées.

Ce stage porte sur l'estimation de la régularité locale des trajectoires des séries temporelles de données fonctionnelles afin d'obtenir une meilleure estimation de leur fonction moyenne et de l'opérateur d'auto-covariance.

contribution

si jamais vous apercevez des fautes méthodologiques ou orthographiques dans le rapport, merci de rédiger une *issue* sur Github à l'adresse :

correctif

ENSAI-stage-fin-etude-datastorm-fda-regularite/issues

contact

mail DEV : dev.allemandinstable@gmail.com

Table des matières

Chapitre 1

Motivations

Les données que l'on traite sont des données du secteur de l'énergie, et plus particulièrement des données de production électrique. On dispose ainsi de plusieurs éoliennes identifiées par le tag "id_(identifiant de l'éolienne)" dont l'énergie produite est mesurée toutes les demies heures, et ce pendant 4 ans (de de 2014 à 2017). Cette énergie produite est dénommée la courbe de charge (que l'on abbrégera par **CDC** par la suite). Il est cependant plus utile de s'intéresser au facteur de charge (ou **FDC**) qui est défini comme Facteur de Charge $= \frac{\text{Courbe de Charge}}{\text{Puissance Installée}}$. Ainsi **FDC** doit nécessairement être compris entre 0 et 1. c'est entre autre aussi une manière de détecter des anomalies et données atypiques comme la surproduction d'énergie par rapport à ce qui était attendu de la part d'une éolienne ou encore un défaut de capteur (tension / intenité, ...) pour mesurer la courbe de charge.

Ainsi, les données qui sont traitées dans le cadre de ce stage sont, entre autres, des courbes de charge éoliennes observées chaque 30 minutes. Le schémas d'observation est donc le 'common-design'. C'est-à-dire que les temps d'observations sont ici déterministes à intervalle de temps fixe.

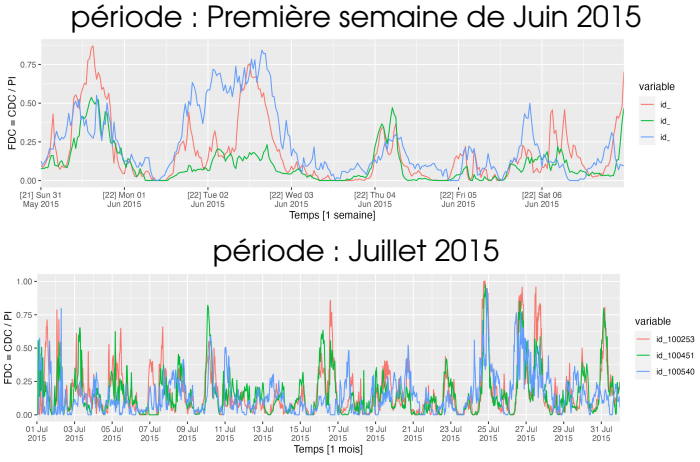

FIGURE 1.1 – Courbes de charges éoliennes sur 3 premiers parcs éoliens

Bien que la différenciation en analyse de séries temporelles soit une méthode efficace pour éliminer la tendance, qu'elle soit saisonnière ou non, permettant ainsi une bonne analyse des données ; ces modèles présentent des limites en termes de prédiction à long

terme, les rendant moins utiles lorsque l'objectif est de prédire à moyen ou long terme. De plus, ces modèles, ainsi que différents modèles de machine learning populaires, estiment les données courbe par courbe ce qui ne tire pas profit d'observations similaires entre les courbes.

Une première idée serait d'utiliser un modèle de série temporelle ARIMA afin de modéliser la dynamique des courbes de charge.

Un peu d'histoire sur les séries temporelles . . .

une grande partie des informations présentées dans cette section histoire provient de la référence (14)

Parmi les étapes importantes du développement des séries temporelles, on peut noter l'article *Time Series Analysis : Forecasting and Control* de Box et Jenkins (1970) qui introduit le modèle ARIMA et une approche aujourd'hui standarde d'évaluation du modèle à utiliser ainsi que son estimation. Ce développement est dû en grande partie à l'utilisation de telles données dans les secteurs économiques et des affaires afin de suivre l'évolution et la dynamique de différentes métriques

L'étude des séries temporelle a été divisée en l'étude du domaine fréquentiel, qui étudie le spectre des processus pour le décomposer en signaux principaux, et du domaine temporel, qui étudie les dépendances des indices temporels. L'utilisation de chacune des approches était sujet à débats mouvementés jusqu'aux alentours de l'an 2000.

Le développement des capacités de calcul a été une révolution notamment pour l'identification des modèles (le critère AIC, l'estimation par vraissemblance dans les années $1980, \ldots$).

À partir des années 1980, les modèles non linéaires émergent (ARCH par Engle, modèles à seuil . . .) et trouvent application en économie notamment. Enfin l'étude multivariée (modèle VAR) fait surface dans les années 1980 par Christopher Sims (16, lien de l'article)

Une large partie de la théorie s'appuie notamment sur l'étude des racines de l'unité, en considérant un polynôme d'opérateur $P(B) = (I + \sum_k a_k B^k)$ à partir duquel les relations d'autocorrélations peuvent se ré-écrire.

Toutefois, l'utilisation d'un modèle ARIMA ne permet de modéliser la dynamique du phénomène étudié. En effet, la sélection d'un modèle ARIMA sur le critère du BIC sélectionnait peut importe le parc éolien un modèle auto-régressif d'ordre 0. Ainsi le modèle sélectionné considérait les irrégularités comme étant du bruit. On en conclut que ces modèles peuvent ne pas capturer efficacement la structure complexe des données.

Afin de prédire sur le long terme, nous allons donc adopter une approche basée sur les données fonctionnelles pour capturer la structure de la consommation. Cette approche permettra de d'exploiter une information clé : la similarité entre les courbes observées.

Qu'est-ce qu'une donnée fonctionnelle ?

Une donnée est dite fonctionnelle lorsque la variable aléatoire qui nous intéresse n'est plus une variable aléatoire à valeur dans $\mathbb{R}^d$, comme le statisticien a l'habitude de manipuler, mais une variable aléatoire à valeur dans un espace de fonction. Concrètement, chaque réalisation n'est plus un nombre mais bien une courbe toute entière indexée (le plus souvent) sur un intervalle $\mathcal{T}$.

Si le statisticien est déjà à l'aise avec l'idée qu'une variable aléatoire réelle identiquement distribuéé puisse modéliser une expérience répétable provenant d'un même phénomène, il pourra se convaincre que les données fonctionnelles permettent elles aussi de modéliser des expériences en lien (fonctionnel) avec un certain paramètre. Et c'est le lien entre les deux valeurs, cette fois-ci, qui provient d'un même phénomène.

Donnons en un exemple : observons la consommation électrique d'un foyer dans une journée. Lorsque l'on travaille sur $\mathbb{R}$, on s'intéresse à sa consommation électrique disons en l'instant $t = 12\text{h}$ $\left(\in \mathcal{T} \underset{\text{déf}}{\equiv} [0, 24[= 1 \text{ jour avec t en heure}\right)$. La consommation du foyer i à midi, notée y_i, suit la loi d'un phénomène général Y, comme un normale $\mathcal{N}(50\,kW\text{h}, 9)$ par exemple. Travailler sur des données fonctionnelles dans ce cadre c'est étudier non plus la consommation y_i à midi, mais regarder l'ensemble de sa consommation en même temps sur toute la journée $y_i(t) = x_i(t)$ avec $t \in \mathcal{T}$. Et de remarquer ainsi que toutes les consommations électriques le long de la journée d'un foyer à l'autre suivent la même tendance (on consomme plus le matin avant le travail et le soir, pendant la journée on consomme moins car on est au travail), ainsi c'est la fonction $x_i : \mathcal{T} \longrightarrow \mathbb{R}$ qui suit la loi d'un phénomène X général. Ce que l'on vient de dire c'est que la **relation** entre le temps $t \in \mathcal{T}$ et la consommation électrique $\underbrace{y_i(t)}_{=x_i(t)}$ est elle même sujet à une loi plus générale. Grossièrement, les courbes auront la même allure, mais chaque individu à sa consommation propre.

Plus formellement : comme on a défini une variable aléatoire réelle comme une application $X : \begin{array}{ccc} \Omega & \longrightarrow & \mathbb{R} \\ \omega & \longmapsto & x = X(\omega) \end{array}$, on définit de même une donnée fonctionnelle comme une application $X : \begin{array}{ccc} \Omega & \longrightarrow & \mathcal{C}^0(\mathcal{T}, \mathbb{R}) \\ \omega & \longmapsto & x = X(\omega) \end{array}$, et ce que l'on observe sont donc les valeurs des paramètres $t \in \mathcal{T}$ ainsi que l'image de t par x : $y = x(t)$. Les points que le statisticien observe sont donc les couples de la forme $(t_k^{(\text{individu } i)}, y_k^{(\text{individu } i)})_{i \in [\![1,m]\!]}$, générés par le processus aléatoire X dont la réalisation est la véritable courbe x_i de l'individu i que l'on souhaite estimer pour travailler avec.

certaines ressources sur l'analyse de données fonctionnelles définissent les données fonctionnelles de la manière suivante

$$X : \begin{array}{ccc} \Omega \times \mathcal{T} & \longrightarrow & \mathbb{R} \\ (\omega, t) & \longmapsto & X(\omega, t) = y \end{array}$$

qui selon mon humble avis, ne permet pas une interprétation clé en main du concept.

...et un peu d'histoire sur les données fonctionnelles

Pour une description plus complète de l'histoire du développement de l'analyse fonctionnelle, on pourra se référer à cet article de Wang, Chiou et Müller (15)

Bien que l'histoire du développement de l'Analyse de Données Fonctionnelles (FDA) puisse être retracée jusqu'aux travaux de Grenander et Karhunen (7) dans les années 1940 et 1950, où l'outil a été utilisé pour étudier les courbes de croissance en biométrie, ce sous-domaine de la statistique a été étudié de manière systématique à partir des années 1980.

En effet, c'est J.O. Ramsay qui a introduit l'appellation de "données fonctionnelles" en 1982 (11) et qui contribuera en partie à sa popularisation. La thèse de Dauxois et Pousse en 1976 sur l'analyse factorielle dans le cadre des données fonctionnelles(4) a ouvert la voie à l'analyse par composante principale fonctionnelle (FPCA), un outil clé pour l'étude des données fonctionnelles. La FPCA permet d'étudier des objets fonctionnels qui sont de dimension infinie, difficiles à manipuler et impossibles à observer empiriquement, en dimension finie.

Au cours des années 2000, de nombreux outils statistiques déjà développés pour des données à valeurs dans $\mathbb{R}^d$ depuis un siècle, tels que la régression linéaire (éventuellement généralisée), les séries temporelles ou encore les modèles additifs, ont été adaptés aux données fonctionnelles. Par exemple, les modèles de régression linéaire fonctionnelle ont été développés avec une réponse fonctionnelle (12) ou scalaire (2) en 1999. Les modèles linéaires généralisés ont également été étudiés (6, 10), avec l'estimation de la fonction de lien par méthode non paramétrique à direction révélatrice *(Single Index Model)* récemment étudiée en 2011 (3). Cette méthode avait déjà été utilisée en économétrie pour des données de $\mathbb{R}^d$ depuis 1963 (13), et leur estimation directe a été étudiée en 2001 par Hristache, Juditsky et Spokoiny (5). De même, les modèles additifs ont été étendus aux données fonctionnelles en 1999 par Lin et Zhang (9). Enfin, le livre de Bosq, *Linear Processes in Function Spaces : Theory and Applications* (1), publié en 2000, a contribué au développement des séries temporelles pour les données fonctionnelles.

Depuis lors, des ressources telles que l'ouvrage de Kokoszka et Reimherr, *Introduction to Functional Data Analysis (2017)* (8), rendent la théorie et la mise en production des méthodes d'analyse et de prédiction de données fonctionnelles plus accessibles.

Rappel :
"Ainsi le modèle (arima) sélectionné considérait les irrégularités comme étant du bruit [...] Afin de prédire sur le long terme, nous allons donc adopter une approche basée sur les données fonctionnelles pour capturer la structure de la consommation [...]"

Pourquoi la régularité est-elle importante ? Et surtout, en quoi est ce que les données fonctionnelles vont nous permettre de mieux capturer la régularité ?

La production électrique est un phénomène très irrégulier (1.1), influencé par la consommation, la météo, etc. Par conséquent, la prévision de ces courbes de charge doit prendre en compte la nature fondamentalement irrégulière du phénomène afin de proprement le modéliser et, en définitive, mieux le prédire. Ce qui est notamment contraire à la plupart des méthodes qui utilisent des fonctions de classe $\mathcal{C}^2$ pour lisser les points observés en données fonctionnelles, ce qui limite la prédiction à des courbes de nature $\mathcal{C}^2$. Il est ainsi important pour des phénomènes de nature irrégulière de ne pas négliger des précautions lors du lissage afin de ne pas perdre l'information irrégulière. L'approche fonctionnelle est clé dans l'estimation de cette régularité, car c'est la réplication de courbes de même nature qui permet in-fine d'estimer la régularité du phénomène.

Chapitre 2

Méthodologie

Contents

toute la rédaction de ce chapitre est une ébauche grossière, destinée à former le squelette du rapport. Le processus de rédaction est itératif sur toute la durée du stage. avancement du stage : 2 / 6 mois

2.1 Données Fonctionnelles : l'essentiel

2.1.1 Cas indépendant : données fonctionnelles

2.1.2 Cas non indépendant : séries temporelles de données fonctionnelles

Une large partie de la théorie des données fonctionnelles suppose que l'on observe des courbes $X_i : \Omega \to \mathcal{C}^0(I, \mathbb{R})$ **indépendantes** et identiquement distribuées. Cependant une partie non négligeable des données que l'on observe ont des dépendances avec les valeurs passées. Par exemple, il est raisonnable de penser que la consommation électrique d'un foyer au cours d'une année croît avec l'ajout successif de nouveau appareils électroniques. L'hypothèse d'indépendance entre les données n'est donc plus pertinente pour les données que l'on traite et il devient important de considérer des processus autorégressifs adaptés aux données fonctionnelles. Si dans le cadre des données de $\mathbb{R}$ cette relation de *dépendance linéaire* avec le passé pouvait s'écrire sous la forme suivante

$X_n = \sum_{k=1}^{n-1} \varphi_k X_k + \varepsilon_n$ où $\varphi_k \in \mathbb{R}$ et $\varepsilon_n \begin{cases} \in \mathrm{VA}(\mathbb{R}) \\ \perp\!\!\!\perp \sigma\left(X_i\right)_{1:n-1} \end{cases}$, dans le cadre fonctionnel on capture la même idée en considérant $X_n = \sum_{k=1}^{n-1} \phi_k\left(X_k\right) + \varepsilon_n$ où ϕ_k est un *opérateur linéaire* de $\mathbb{L}^2(I, \mathbb{R})$, le plus souvent intégral.

Il s'agit d'une généralisation naturelle de la relation dans le cadre réel, puisqu'on peut démontrer que sur l'espace des nombres réels l'ensemble des fonctions linéaires $\phi : \mathbb{R} \to \mathbb{R}$ sont de la forme $x \mapsto ax$ avec $a \in \mathbb{R}$. La relation sur $\mathbb{R}$ que l'on a vue juste avant peut alors se ré-écrire de façon similaire à la version fonctionnelle.

2.2 Estimation de la régularité locale des trajectoires

2.2.1 Ce qu'on entend par régularité locale

Longtemps, il était cru que les fonctions continues étaient dérivables presque partout. C'est notamment Weierstrass qui a démontré qu'il existe des fonctions continues partout mais dérivable nulle part. Poincaré notamment disait de tels objets qu'ils n'existaient que pour contredire le travail des pères. Cependant, des objets manipulés tous les jours comme le monde de la finance notamment traitent des processus qui sont fondamentalement irréguliers (au point de vue de l'analyse, où l'on traite souvent des fonctions au moins dérivables). Il est donc important de pouvoir quantifier la régularité d'une fonction de façon plus fine que le nombre de dérivées qu'elle possède.

Fonction Continue :

$$(\forall \varepsilon > 0)\,(\forall x)(\exists \delta_x > 0)(\forall y)\,|x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

uniforme :

$$(\forall \varepsilon > 0)\,(\exists \delta > 0)(\forall x, y)\,|x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

Fonction Lipschitzienne (+ régulier) :

$$(\forall x, y) \quad |f(x) - f(y)| < L|x-y|$$

Fonction Hölderienne :

$$\begin{cases} (\forall x, y) \quad |f(x) - f(y)| < L_\alpha |x-y|^\alpha \\ 0 < \alpha \leq 1 \end{cases}$$

une fonction lipschitz est une fonction Holderienne avec $\alpha = 1$

Fonction Dérivable (+ régulier) :

Régularité locale :

$$\forall x_0 \quad \begin{cases} (\forall x) \quad |f(x) - f(x_0)| < L_{\alpha(x_0)} |x - x_0|^{\alpha(x_0)} \\ 0 < \alpha(x_0) \leq 1 \end{cases}$$

2.2.2 Prélissage : lissage à ondelettes

Une brève introduction aux ondelettes

Motivation dans le cadre de l'analyse de données fonctionnelles

Effets du lissage à ondelettes sur la régularité locale

2.2.3 Estimation des paramètres régularité locale des trajectoire

2.3 Estimation adaptative

2.3.1 Estimation adaptative de la fonction moyenne

2.3.2 Estimation adaptative de l'opérateur de covariance

2.3.3 Estimation adaptative de l'auto-covariance des séries temporelles fonctionnelles

Chapitre 3

Applications et comparaison des différentes méthodologies

Contents

toute la rédaction de ce chapitre est une ébauche grossière, destinée à former le squelette du rapport. Le processus de rédaction est itératif sur toute la durée du stage. avancement du stage : 2 / 6 mois

3.1 Données simulées

3.1.1 Simulation d'un processus Brownien (multi)-Fractionnaire

3.2 Données Réelles

3.2.1 Courbes de charge éolienne

3.2.2 Données Hydrauliques

3.3 Conclusion sur l'efficacité des différentes méthodologies

Bibliographie

(1) Denis Bosq. *Linear processes in function spaces : theory and applications*, volume 149. Springer Science & Business Media, 2000.

(2) Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1) :11–22, 1999.

(3) Dong Chen, Peter Hall, and Hans-Georg Müller. Single and multiple index functional regression models with nonparametric link. 2011.

(4) Jacques Dauxois and Alain Pousse. *Les analyses factorielles en calcul des probabiblités et en statistique : Essai d'étude synthétique*. PhD thesis, Éditeur inconnu, 1976.

(5) Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

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

(7) Kari Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34, 1946.

(8) Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC press, 2017.

(9) Xihong Lin and Daowen Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 61(2) :381–400, 1999.

(10) Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. 2005.

(11) James O Ramsay. When the data are functions. *Psychometrika*, 47 :379–396, 1982.

(12) James O Ramsay and CJ1125714 Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society : Series B (Methodological)*, 53(3) :539–561, 1991.

(13) William F Sharpe. A simplified model for portfolio analysis. *Management science*, 9(2) :277–293, 1963.

(14) Ruey S. Tsay. Time series and forecasting : Brief history and future research. *Journal of the American Statistical Association*, 95(450), 2000. DOI : https ://doi.org/10.2307/2669408.

(15) Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3 :257–295, 2016. télécharger.

(16) James H. Stock & Mark W. Watson. Vector autoregressions. *Journal of the American Statistical Association / Journal of Economic Perspective*, 15(4), 2001. page 101 - DOI : https ://doi.org/10.1257/jep.15.4.101 - télécharger.