

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



STAGE DE FIN D'ÉTUDE
pour l'entreprise DataStorm

Étude et application de l'estimation adaptative pour les séries temporelles 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 estimations 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

1	Méthodologie	2
1.1	Méthodes pour les séries temporelles	3
1.1.1	Présentation rapide de l'histoire des séries temporelles et de leurs applications (14)	3
1.2	Données fonctionnelles	3
1.2.1	Motivations de l'utilisation de données fonctionnelles	3
1.2.2	Séries temporelles de données fonctionnelles et prise en compte de la régularité des trajectoires	4
1.3	Séries temporelles de données fonctionnelles	5
2	Application des séries temporelles pour les données fonctionnelles	6
2.1	Le jeu de données	6
2.2	Étude avec des outils traditionnels	6
2.3	Étude en utilisant la théorie des données fonctionnelles	6
2.3.1	Ancienne méthodologie : sans prise en compte de la régularité des trajectoires	6
2.3.2	Nouvelle méthodologie : avec la prise en compte de la régularité des trajectoires	6
3	Chapter 3	7
3.1	G	7
3.2	H	7
3.3	I	7

Chapitre 1

Méthodologie

Contents

1.1	Méthodes pour les séries temporelles	3
1.1.1	Présentation rapide de l'histoire des séries temporelles et de leurs applications [14]	3
1.2	Données fonctionnelles	3
1.2.1	Motivations de l'utilisation de données fonctionnelles	3
1.2.2	Séries temporelles de données fonctionnelles et prise en compte de la régularité des trajectoires	4
1.3	Séries temporelles de données fonctionnelles	5



notes personnelles :

Refaire la structure en présentant plutôt sous cet angle :

faire juste un bloc données fonctionnelles et présenter d'entrée de jeu les données fonctionnelles

- Histoire des données fonctionnelles et motivations (avec graphiques à l'appui)

- mais dans les données fonctionnelles même si on suppose que c est souvent iid, il y a des phénomènes avec des corrélations temporelles : \Rightarrow considérer naturellement des séries temporelles fonctionnelles

- avantages et inconvénients de cette méthode, parallèle avec les données fonctionnelles avec le cas réel (\mathbb{R}) (histoire et méthode)

- présenter ce qui se fait actuellement dans les données fonctionnelles

- dire pourquoi il est important de considérer la régularité des trajectoires (avec graphique à l'appui)

- annoncer ce qui va être fait : application et comparaison des méthodes actuelles et la nouvelle (régularité) des données fonctionnelles à la fois sur des données simulées et réelles

1.1 Méthodes pour les séries temporelles



cette partie ne sera pas utilisée pour le rapport final est fait en ce moment office d'éléments où je peux aller piocher pour la rédaction.

1.1.1 Présentation rapide de l'histoire des séries temporelles et de leurs applications (14)



une grande partie des informations présentées dans cette sous-section 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 standard 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 vraisemblance dans les années 1980, modèles à espace d'états et le filtre de Kalman pour évaluer cette vraisemblance efficacement, MCMC, ...).

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.

1.2 Données fonctionnelles

1.2.1 Motivations de l'utilisation de 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). 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.

1.2.2 Séries temporelles de données fonctionnelles et prise en compte de la régularité des trajectoires

Une large partie de la théorie des données fonctionnelles suppose que l'on observe des courbes $X_i : \Omega \rightarrow \mathcal{C}^0(I, \mathbb{R})$ **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 nouveaux 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 auto-régressifs 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

$$X_n = \sum_{k=1}^{n-1} \varphi_k X_k + \varepsilon_n \text{ où } \varphi_k \in \mathbb{R} \text{ et } \varepsilon_n \begin{cases} \in \text{VA}(\mathbb{R}) \\ \perp \sigma(X_i)_{1:n-1} \end{cases}, \text{ dans le cadre fonctionnel on cap-}$$

ture la même idée en considérant $X_n = \sum_{k=1}^{n-1} \phi_k(X_k) + \varepsilon_n$ 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 $\phi : \mathbb{R} \rightarrow \mathbb{R}$ 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.

Toutefois, jusque maintenant, une large partie de la littérature sur les données fonctionnelles considère comme hypothèse que l'on observe des courbes entières dans la construction de leur estimateur alors que la réalité du monde physique dans lequel nous vivons est que nous pouvons observer avec nos capteurs qu'un nombre fini de points. Ainsi, nos observations sont fondamentalement de nature discontinue alors que les objets que l'on souhaite modéliser sont de nature continue. Un argument fréquemment utilisé est qu'il suffit d'effectuer un lissage, en utilisant notamment des splines cubiques naturelles et d'utiliser les courbes lissées en plug-in dans l'estimateur considéré  ( citation requise ). Cette approche ne tient pas compte de la régularité de la courbe considérée, qui même si continue peut s'avérer irrégulière (non dérivable par exemple), alors que les splines cubiques sont \mathcal{C}^2 . Il n'est d'ailleurs pas rare d'observer des processus fortement irréguliers dans le monde physique dans lequel on vit  (exemple concret requis). De plus la régularité du processus que l'on observe peut même varier sur la trajectoire de celui-ci (i.e. sur l'ensemble I où $X : \Omega \rightarrow \mathcal{C}^0(I, \mathbb{R})$). Il est ainsi raisonnable de penser qu'inclure la régularité du processus considéré dans son estimation fournira de meilleures estimations et prédictions de trajectoires.

Ce stage se concentrera sur l'estimation et l'utilisation de la régularité des trajectoires de séries temporelles de données fonctionnelles dans le cadre de courbes de charges énergétiques. Il s'agit de données de production dont la bonne estimation et la précision des prévisions constituent un enjeu stratégique. Nous comparons dans ce rapport les résultats de la méthode avec la prise en compte de la régularité avec les méthodes classiques utilisées jusqu'alors dans le domaine des données fonctionnelles.

1.3 Séries temporelles de données fonctionnelles

Chapitre 2

Application des séries temporelles pour les données fonctionnelles

Contents

2.1	Le jeu de données	6
2.2	Étude avec des outils traditionnels	6
2.3	Étude en utilisant la théorie des données fonctionnelles	6
2.3.1	Ancienne méthodologie : sans prise en compte de la régularité des trajectoires	6
2.3.2	Nouvelle méthodologie : avec la prise en compte de la régularité des trajectoires	6

2.1 Le jeu de données

2.2 Étude avec des outils traditionnels

2.3 Étude en utilisant la théorie des données fonctionnelles

2.3.1 Ancienne méthodologie : sans prise en compte de la régularité des trajectoires

2.3.2 Nouvelle méthodologie : avec la prise en compte de la régularité des trajectoires

Chapitre 3

Chapter 3

Contents

3.1	G	7
3.2	H	7
3.3	I	7

3.1 G

3.2 H

3.3 I

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 probabilités 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.