Détection d'anomalies dans des données fonctionnelles multivariées

Martial Amovin-Assagba, Julien Jacques, Irène Gannaz, Frédéric Fossi, Johann Mozul

Détection d'anomalies dans des données fonctionnelles multivariées

Martial AMOVIN-ASSAGBA[1,2], Julien JACQUES[2], Irène GANNAZ[3], Frédéric FOSSI[1] & Johann MOZUL[1]

[1] *Arpege Master K, Saint-Priest, France, martial.amovin@masterk.com, frederic.fossi@masterk.com & johann.mozul@masterk.com*
[2] *Univ Lyon, Lyon 2, ERIC EA3083, Lyon, France, julien.jacques@univ-lyon2.fr*
[3] *Univ Lyon, INSA de Lyon, CNRS UMR 5208, Institut Camille Jordan, F-69621 Villeurbanne, France, irene.gannaz@insa-lyon.fr*

Résumé. L'objectif de ce travail est de détecter des anomalies dans les données fonctionnelles multivariées provenant d'appareils de mesure, dans une optique de maintenance prédictive. Des méthodes statistiques comme le clustering fonctionnel et l'estimation linéaire par morceaux ont été testées. Nous montrons l'intérêt de ces méthodes ainsi que leurs insuffisances.

Mots-clés. données fonctionnelles, clustering, modélisation linéaire par morceaux

Abstract. This work aims to detect anomalies in multivariate functional data coming from measuring devices. Statistical methods such as functional clustering and piecewise linear estimation were tested. We show the interest of these methods as well as their lackness.

Keywords. functional data, clustering, piecewise linear modelling

1 Introduction

Nous disposons de mesures temporelles provenant simultanément de divers capteurs. Notre objectif est de détecter des anomalies dans ces données fonctionnelles multivariées. De nombreuses techniques d'apprentissage non supervisé pour la détection des anomalies existent (Chandola et al 2009), mais très peu sont adaptées aux données fonctionnelles (Ramsay et Silverman 2005). Certaines méthodes se basant sur les fonctions de profondeur sont proposées pour les données fonctionnelles univariées (López-Pintado et Romo 2009; Febrero et al 2008), d'autres dans le cadre multivarié (Hubert et al 2015; Dai et Genton 2018). Hubert et al (2015) utilisent le demi-espace de profondeur pour mesurer la "centralité" d'une courbe alors que Dai et Genton (2018) définissent une matrice de décalage en étendant le décalage directionnel.

On distingue plusieurs types d'anomalies quand il s'agit des données fonctionnelles: anomalie de forme, de localisation, etc. Dans l'application qui nous intéresse, nous sommes

confrontés principalement à des anomalies de forme. Si bon nombre d'approches se basant sur les fonctions de profondeur sont fiables dans la détection des anomalies de localisation, elles échouent souvent à identifier ce type de données aberrantes.

Une autre façon de détecter des anomalies pourrait être d'utiliser des techniques de clustering fonctionnelles multivariées (Schmutz et al 2020), qui pourraient permettre d'isoler des clusters de données atypiques.

Afin de détecter des anomalies, nous avons testé sur les données un algorithme de clustering fonctionnel, que nous avons comparé avec une autre approche paramétrique (estimation linéaire par morceaux) basée sur la forme des courbes. Nous présentons les résultats obtenus et leurs limites.

2 Les données

Nous considérons un jeu de données de taille 509, issu d'un matériel comportant 4 capteurs, sachant que nous avons aussi d'autres matériels qui ont plus de 4 capteurs. Une mesure est constituée de 4 courbes. Le nombre de points de mesure des trajectoires diffère d'une observation à l'autre. Il varie entre 2199 et 10675 avec une médiane égale à 5144. Puisque les données sont discrétisées sur des grilles fines, une approche de type données fonctionnelles est privilégiée (Ramsay and Silverman 2005). Nous représentons quelques données sur la Figure 1.

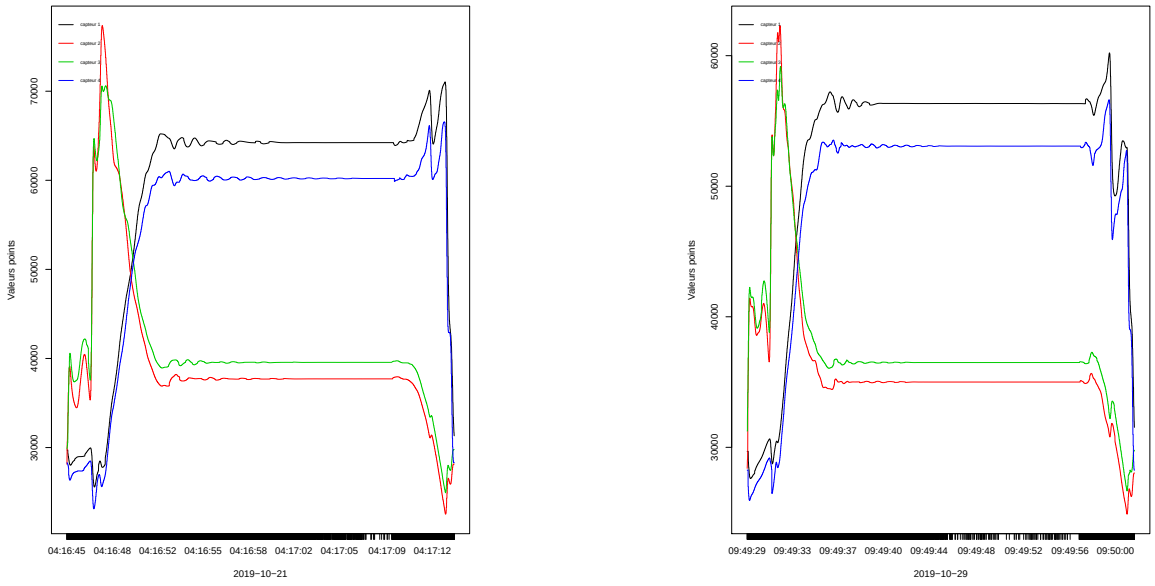

Figure 1: Exemples de courbes des instruments de mesure

D'une donnée normale à l'autre, il y a souvent une différence assez importante en

amplitude et en temps. Cette différence pourrait biaiser les résultats. Par conséquent, nous normalisons au préalable les données en amplitude en les divisant par une valeur moyenne relative aux courbes, et en temps en les ramenant dans l'intervalle $[0, 1]$.

Des analyses d'experts ont permis d'identifier deux mesures comme anormales. La première (notée "Anormale1") provient de la défaillance du Capteur 2 alors que la seconde (notée "Anormale2") est liée à un effet extérieur. Cette dernière n'est pas tout à fait considérée comme aberrante du point de vue de l'expert métier. Notre objectif est de proposer une procédure automatisée permettant sur cette base de données de retrouver ces deux données.

3 Clustering

Afin de détecter un ensemble de données atypiques liées à des défaillances des appareils de mesure, une caractérisation pourra se faire à partir d'une analyse non supervisée des historiques de mesures, de type clustering de données fonctionnelles multivariées. Notre objectif est d'identifier des clusters qui caractérisent les groupes de données atypiques.

Récemment Schmutz et al, ont proposé une nouvelle méthode de clustering fonctionnel multivarié (`funHDDC`) afin de permettre d'identifier des groupes d'individus homogènes. S'agissant des données fonctionnelles, la principale source de difficulté réside dans le fait que les courbes sont censées appartenir à un espace dimensionnel infini, alors qu'en pratique nous disposons d'échantillons observés en un ensemble de points finis. Les auteurs reconstituent la forme fonctionnelle des données en lissant les observations dans une base de fonctions de dimension finie. Leur méthode s'appuie ensuite sur un modèle de mélange latent fonctionnel.

Nous appliquons dans un premier temps cet algorithme dans le cas multivarié où nous prenons en compte toutes les courbes quel que soit l'appareil de mesure. Dans un second temps nous testons le cas univarié où nous ne considérons que les courbes d'un même capteur. Le package `funHDDC` de Schmutz et al propose plusieurs modèles parcimonieux. Il propose également 5 moyens d'initialisation de l'algorithme E-M. Nous avons testé tous les modèles, tout en variant le nombre maximum d'itérations et les initialisations de l'algorithme E-M.

En multivarié, un seul cluster est obtenu à chaque fois: le modèle de clustering n'arrive pas à former des groupes suffisamment distincts les uns des autres. Par contre dans le cas univarié, quel que soit le capteur considéré, le modèle retenu présente au moins 3 clusters. Nous représentons sur la Figure 2 les courbes moyennes des 5 clusters obtenus avec les données du capteur 2, puisque c'est le seul capteur ayant eu une défaillance. Bien qu'identifiant plusieurs clusters, l'algorithme n'arrive pas à distinguer un groupe spécifique de courbes anormales. Aucun cluster n'a de forme atypique d'un point de vue de l'expert métier. Les courbes identifiées auparavant comme anormales sont dans un même cluster que des courbes normales. Ceci est probablement dû au trop faible nombre de données

atypiques. La Figure 3 présente les courbes de ce cluster, avec une distinction en couleur des courbes supposées anormales.

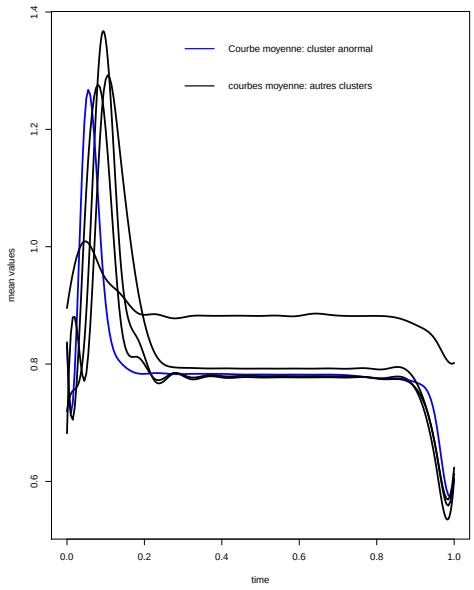

Figure 2: Courbes moyennes des 5 clusters formés par l'ensemble des courbes du capteur 2, cas univarié

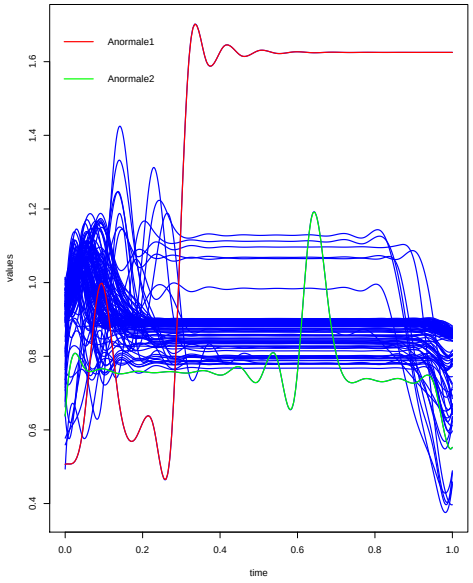

Figure 3: Ensemble des courbes du cluster contenant les données anormales du capteur 2, cas univarié

Avec cette méthode, il n'est donc pas possible de faire une caractérisation des données anormales. Nous ne pouvons donc pas détecter les courbes aberrantes du fait que l'algorithme les met dans un même cluster que les courbes normales. De plus cette approche présente des difficultés du fait que les données sont de tailles différentes. Nous pensons plus tard coupler cette méthode avec une étude de la distribution du temps et de l'amplitude des courbes.

4 Estimation linéaire par morceaux

Dans un second temps, nous avons décidé de nous appuyer sur la forme spécifique des courbes de notre application. Nous avons identifié une allure type pour chaque appareil de mesure. Nous définissons donc un patron pour chaque capteur. Ce patron forme une base de fonctions particulières (linéaires par morceaux, cf. Figure 4) dans laquelle nous lissons les données.

Chaque observation est approchée dans cette base linéaire par morceaux (Muggeo 2017) à l'aide du package R `segmented`. Muggeo définit des modèles de régressions avec

des relations segmentées entre la réponse et la variable tout en estimant les points de rupture. La Figure 4 présente une courbe du deuxième appareil de mesure avec sa reconstruction linéaire par morceaux (patron).

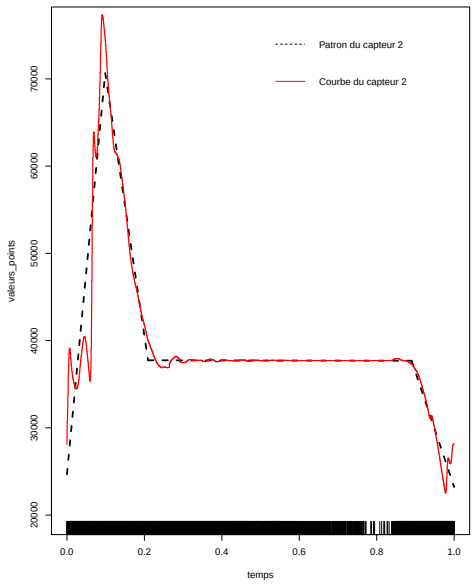

Figure 4: Exemple de patron : cas du capteur 2

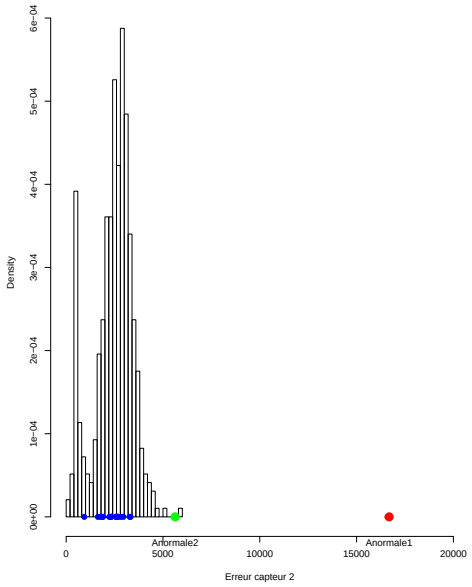

Figure 5: Histogramme des erreurs de reconstruction du capteur 2

Après avoir approché chaque mesure dans la base linéaire par morceaux définissant le patron d'une mesure normale, nous regardons la distribution empirique des erreurs d'approximations (écarts entre les vraies valeurs et les valeurs estimées). Il est alors possible de se baser sur cette distribution empirique pour proposer une évaluation de la probabilité qu'une nouvelle mesure soit atypique.

La Figure 5 présente l'histogramme des erreurs de reconstruction des courbes du capteur 2. Les points en rouge et vert représentent respectivement les erreurs de reconstruction des données défaillantes "Anormale1" et "Anormale2". Visiblement, le point rouge est très éloigné de l'ensemble des points de l'histogramme. La valeur réelle de cette erreur est 16681.9, soit environ 3 fois le maximum des erreurs de reconstruction des courbes normales provenant du capteur 2. Cette méthode nous permet de détecter la donnée atypique issue de la défaillance du capteur 2. Elle permet également de distinguer la seconde anomalie, provenant de la défaillance d'un capteur issue d'une variable exogène ou d'un effet extérieur.

Afin de voir s'il y a un signe précurseur à la défaillance du capteur 2, nous représentons les erreurs de reconstruction de quelques données avant la panne. Ces points (en bleu) sont tous considérés comme normaux suivant la distribution empirique des erreurs. Avec

cette méthode, il ne semble donc pas y avoir de signes précurseurs à la panne, pour cette défaillance.

5 Conclusion

Dans ce travail préliminaire, nous avons appliqué deux méthodes paramétriques pour détecter des anomalies, un clustering fonctionnel multivarié et une méthode basée sur la distribution des erreurs de reconstruction par une base de fonctions linéaires par morceaux. Seule la seconde méthode a permis de détecter la donnée atypique provenant d'un appareil de mesure défaillant. L'algorithme de clustering fonctionnel utilisé nous renvoie un cluster composé d'un mélange de données anormales et normales. Une difficulté en utilisant l'approche fonctionnelle est que les données ne sont pas de même taille. Nous pensons coupler l'approche de clustering fonctionnel avec une étude de la distribution du temps et de l'amplitude de la donnée. Un autre point important est de considérer des métriques basées sur les dérivées. Nous les utilisons déjà avec la méthode FIF : Functional Isolation Forest (Stearman et al 2019).

Bibliographie

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58.

Dai, W., & Genton, M. G. (2018). An outlyingness matrix for multivariate functional data classification. *Statistica Sinica*, 28(4), 2435-2454.

Febrero, M., Galeano, P., & Gonzàlez-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics: The official journal of the International Environmetrics Society*, 19(4), 331-345.

Hubert, M., Rousseeuw, P. J., & Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), 177-202.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486), 718-734.

Muggeo, V. M. (2017). Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. Australian & New Zealand *Journal of Statistics*, 59(3), 311-322.

Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis. Springer series in statistics.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 1-31.

Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019). Functional Isolation Forest. *arXiv preprint arXiv:1904.04573.*