Décomposition et clustering de données fonctionnelles pour l'analyse de la consommation d'eau potable

Zineb Noumir*, Allou Samé*, Nicolas Cheifetz**, Anne-Claire Sandraz**, Cédric Féliers**

*Université Paris-Est, IFSTTAR, COSYS, GRETTIA, F-77447 Marne-la-Vallée, France
{Zineb.Noumir, Allou-Badara.Same}@ifsttar.fr
**Veolia Eau d'Ile-de-France, Le Vermont, 28 Boulevard de Pesaro, 92739 Nanterre Cedex
{Nicolas.Cheifetz, Anne-Claire.Sandraz, Cedric.Feliers}@veolia.com

1 Introduction

L'analyse de données fonctionnelles (Ramsay et Silverman, 2005) regroupe des techniques statistiques pour le traitement de données, considérées comme des fonctions (ou courbes). On s'intéresse ici au clustering de courbes de consommation d'eau ; l'approche proposée est notamment inspirée par Gaffney (2004). La suite du papier est organisée de la façon suivante : la section 2 présente la méthodologie proposée pour l'extraction de profil-type de consommation. Des expérimentations sur un jeu de signaux réels et leurs interprétations sont décrites dans la section 3. Enfin, quelques conclusions et perspectives sont listées dans la section 4.

2 Approche pour l'analyse des courbes de consommation

Cette section introduit un modèle de mélange dont chaque composante est un régresseur sous la forme d'une série de Fourier. Ce type d'approche permet notamment de s'affranchir d'un calcul de distances (L2, DTW,...) entre les séries temporelles comme dans l'algorithme des K-moyennes. Soit $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, un ensemble de n séries temporelles, où chaque série $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})$ correspond à une courbe de consommations avec $y_{it} \in \mathbb{R}$ et $t > 0$. Chaque série temporelle est d'abord exprimée par la décomposition additive suivante :

$$y_{it} = f_{it} + x_{it} + d_{it} + \epsilon_{it}\,, \tag{1}$$

Avec

- f_{it} est la tendance globale de la série estimée par un filtre à moyenne mobile,
- x_{it} est la composante saisonnière exprimée par une série de Fourier avec une saisonnalité journalière (période 24) et hebdomadaire (période 168) – pas de temps horaire :

$$x_{it} = \sum_{j=1}^{q_1} \left(\alpha_{1j} \cos\left(\frac{2\pi jt}{24}\right) + \alpha_{2j} \sin\left(\frac{2\pi jt}{24}\right)\right) + \sum_{j=1}^{q_2} \left(\alpha_{3j} \cos\left(\frac{2\pi jt}{168}\right) + \alpha_{4j} \sin\left(\frac{2\pi jt}{168}\right)\right)\,, \tag{2}$$

- d_{it} est une composante pour capturer l'effet des jours fériés, et ϵ_{it} est un bruit blanc.

Une stratégie est adoptée pour estimer les coefficients $(\alpha_{*j})_j$ et les hyperparamètres (q_1, q_2) mais elle n'est pas décrite ici par souci de brièveté. Chaque donnée fonctionnelle $\mathbf{y}_i$ est ainsi approximée par une courbe périodique de consommation : $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ avec $m \leq T_i$. On suppose que chaque série $\mathbf{x}_i$ est distribuée indépendamment selon un modèle de mélange :

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}) \,, \tag{3}$$

où le vecteur paramètre $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$ est estimé par maximisation de la vraisemblance $\sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\theta})$ via un algorithme EM (Dempster et al., 1977) dédié. Après l'estimation de $\boldsymbol{\theta}$, on obtient K prototypes $(\mathbf{U}\boldsymbol{\beta}_k)_k$ ainsi qu'une partition des courbes, où chaque série est associée au cluster ayant la plus grande probabilité a posteriori.

3 Expérimentations sur un jeu de données réelles

Le jeu de données représente les volumes horaires d'eau potable consommés sur 490 jours (de novembre 2013 à avril 2015) et télérelevés à partir de 10233 compteurs autour de Paris. Suivant la section précédente, on pose $n = 10233$, $m = 168$ et après sélection de modèle via l'étude d'un critère BIC, on choisit un nombre de clusters $K = 8$.

La figure 1 illustre chacun des huit clusters (lignes de la figure) en représentant des log-consommations dont la normalisation est décrite par Gaffney (2004). Selon notre interprétation, le cluster 1 représente une utilisation de bureaux ou industrielle (faible pendant le WE), les clusters 3, 4, 5 sont associés à une consommation résidentielle (pics matin et soir), les clusters 6, 7, 8 correspondent à des commerces (similaires tous les jours) et le cluster 2 est considéré comme un cluster de bruit (formes atypiques et large variance).

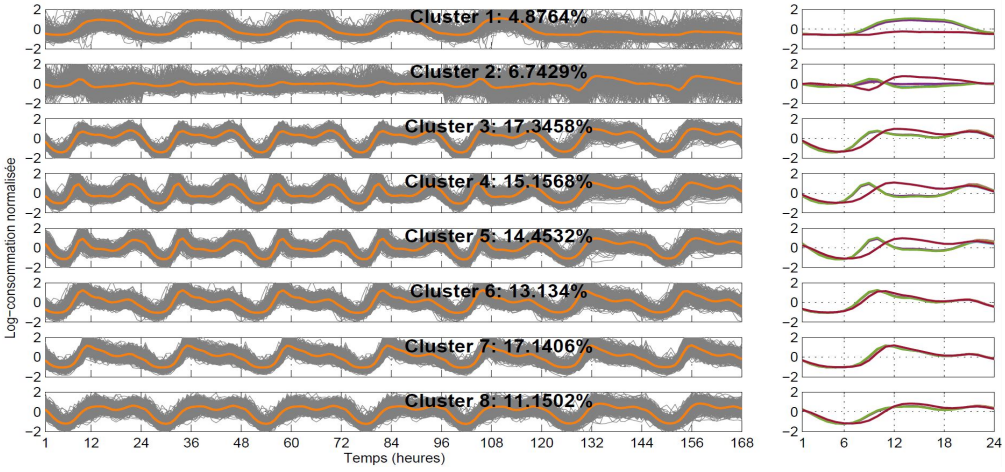

FIG. 1 – *Partitionnement des courbes de consommations en huit clusters avec prototypes hebdomadaires (gauche) en orange et journaliers (droite) de jaune (jours ouvrés) à rouge (WE).*

4 Conclusion

Dans cet article, une méthodologie générique est proposée pour l'analyse de données fonctionnelles et en particulier, le clustering de courbes et l'extraction de profils-type. Cette approche a été succinctement présentée et décrite en deux étapes : l'extraction de composantes saisonnières via une décomposition de Fourier puis l'apprentissage non supervisé des courbes via un algorithme EM. Huit clusters ont été identifiés à partir d'une base de données réelles et un catégorie réaliste a été attribuée à chaque cluster.

En perspectives, des données complémentaires de contexte seront explorées de manière à raffiner l'interprétation du clustering. Les résultats obtenus seront comparés à d'autres méthodes proposées pour répondre à une problématique similaire. On pense notamment à utiliser une décomposition à base d'ondelettes à la place de la transformée de Fourier afin de prendre en compte la localité temporelle des patterns.

Références

Dempster, A. P., N. M. Laird, et D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Gaffney, S. J. (2004). *Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models*. Ph. D. thesis, University of California, Irvine.

Ramsay, J. O. et B. W. Silverman (2005). *Functional Data Analysis* (Seconde ed.). Springer Series in Statistics. Springer.