Distance de Fréchet et Dynamic Time Warping pour la classification non supervisée de séries chronologiques d'observance dans le syndrome d'apnées du sommeil

Guillaume Bottaz-Bosson[1], Sébastien Bailly[2], Agnès Hamon[3] & Adeline Samson[4]

[1] *Laboratoires HP2 et LJK, Univ. Grenoble Alpes - guillaume.bottaz-bosson@univ-grenoble-alpes.fr*

[2] *Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, HP2, 38000 Grenoble - sbailly@chu-grenoble.fr*

[3] *Laboratoire LJK, Univ. Grenoble Alpes - agnes.hamon@univ-grenoble-alpes.fr*

[4] *Laboratoire LJK, Univ. Grenoble Alpes - adeline.leclercq-samson@univ-grenoble-alpes.fr*

Résumé. Nous nous intéressons à la classification non supervisée des séries chronologiques. Les algorithmes de classification reposent sur des fonctions de dissimilarité. Nous présentons la dissimilarité discrète sommée de Fréchet (sdF) qui est une variante de la distance de Fréchet. Une étude de simulation permet de comparer les performances de la sdF avec le dynamic time warping (DTW), la distance euclidienne et la distance de Manhattan. Ces mesures sont comparées en utilisant des algorithmes hiérarchiques et à nuées dynamiques. Les meilleures performances sont atteintes en classification hiérarchique avec les dissimilarités DTW et sdF. Enfin, la classification hiérarchique est appliquée avec ces deux mesures sur des données réelles d'observance thérapeutique à la ventilation en Pression Positive Continue (PPC).

Mots-clés. Classification non supervisée de séries chronologiques, Dynamic Time Warping, distance de Fréchet discrète, observance thérapeutique.

Abstract. This work treats on time series clustering. Clustering algorithms rely on dissimilarity measures. We present the sum discrete Fréchet distance (sdF) which is a variant of the Fréchet distance. A simulation study allows to compare the performances of this measure against dynamic time warping (DTW) dissimilarity, Euclidean distance and Manhattan distance. These measures are compared on hierarchical and partitioning algorithms. Best clusterings are provided by hierarchical clustering with DTW and sdF dissimilarities. Then hierarchical clustering is applied with these two measures on real dataset of compliance to Continuous Positive Airway Pressure (CPAP) therapy.

Keywords. Time series clustering, Dynamic Time Warping, discrete Fréchet distance, herapeutic compliance.

1 Contexte de l'étude

Le syndrome d'apnées obstructives du sommeil (SAOS) est une pathologie respiratoire chronique fréquente caractérisée par des obstructions partielles ou complètes des voies aériennes supérieures au cours du sommeil. Le SAOS est à l'origine d'une détérioration de la qualité de vie des patients, causant une fatigue excessive et une altération des capacités cognitives. De plus, il est fréquemment associé à une aggravation des comorbidités

métaboliques et cardiovasculaires. Le traitement efficace recommandé est la ventilation en pression positive continue (PPC). Les voies aériennes supérieures sont maintenues ouvertes par l'administration d'une pression constante dans le pharynx. Celle-ci est délivrée pendant le sommeil via un masque nasal ou naso-buccal raccordé à un appareil respiratoire bruyant. Cette thérapie est souvent perçue comme contraignante et intrusive pour le malade et son conjoint le cas échéant. 15% des individus refusent le traitement après la première nuit et le taux d'adhésion à long terme est compris entre 65% et 80% [Lévy et al., 2015]. L'observance à la PPC étant un paramètre essentiel pour l'efficacité du traitement, les cliniciens souhaitent comprendre les comportements individuels d'observance.

À partir des données journalières obtenues par le télésuivi des patients sous PPC, il est possible d'établir des trajectoires d'observance (Figure 1). L'objectif de cette étude est de classer ces comportements d'observance pour identifier des motifs typiques. On se limite au télésuivi des 91 premiers jours de thérapie pour identifier les principales trajectoires observées en début de traitement. En effet on considère que la plupart des abandons de PPC a lieu dans les 3 mois après l'instauration de la thérapie [Portier et al., 2010].

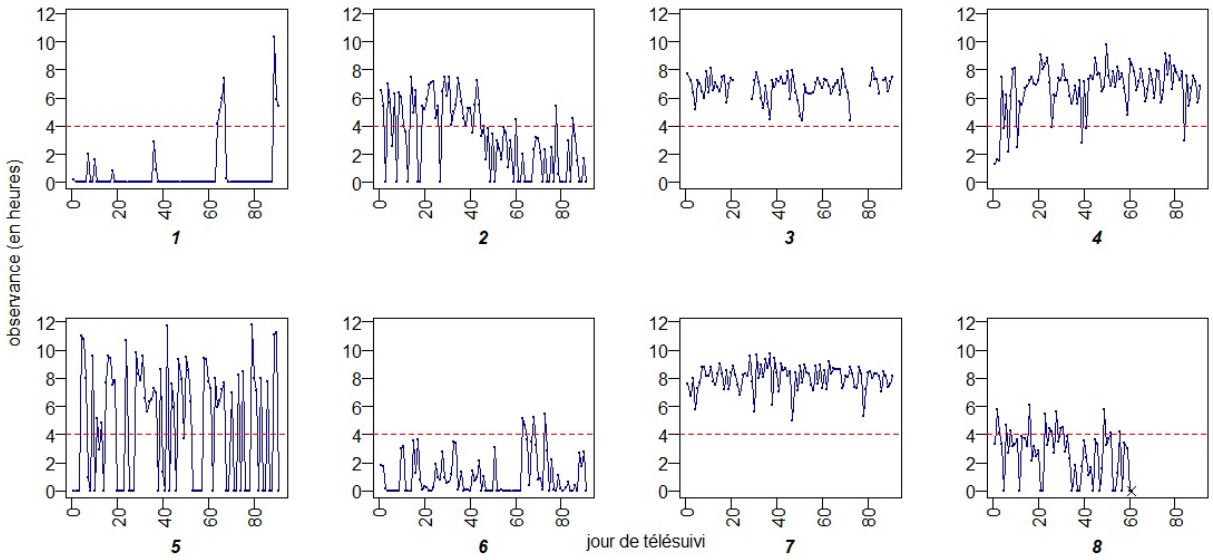

FIGURE 1 – Courbes d'observance à la PPC de 8 patients : heures d'utilisation de l'appareil pour chaque jour depuis le début du traitement (en jours de télésuivi).

La qualité des données brutes ne permet pas de distinguer avec certitude les données manquantes des inobservances à la PPC. En effet les absences de transmission de données peuvent être dues à des problèmes techniques mais aussi à l'inutilisation des machines. Une exploration de ces données a été réalisée pour permettre l'émission d'hypothèses sur les procédés de transmission propres aux différents modèles d'appareils ventilatoires. À titre d'exemple, la troisième courbe de la figure 1 comporte plusieurs périodes sans données

qui sont supposées être des périodes avec données manquantes.

2 Méthodologie

Considérant ces trajectoires comme des séries temporelles, la question est abordée par une approche en classification de séries chronologiques. Nous considérons d'une part des algorithmes hiérarchiques et d'autre part des classifications par nuées dynamiques (k-moyennes et k-medoïdes). Ces algorithmes nécessitent l'utilisation d'une fonction de distance ou dissimilarité. Les distances classiques de Manhattan et Euclidienne [Montero et Vilar, 2014] sont définies pour des données numériques quelconques et sont donc applicables aux séries chronologiques. Le Dynamic Time Warping (DTW) [Müller, 2007] est une technique d'alignement de séries temporelles fournissant une mesure de dissimilarité fréquemment utilisée en classification de séries temporelles. Cette mesure est dite "élastique" car des séries possédant la même forme avec décalages temporels sont considérées comme similaires.

On note $x_{i,t} \in \mathbb{R}$ la valeur d'observance du sujet i au t^{eme} jour de traitement et T_i son dernier jour de mesure. La série temporelle $x_i = (x_{i,1},\ x_{i,2},\ ...,\ x_{i,T_i}) \in \mathbb{R}^{T_i}$ constitue la trajectoire d'observance du sujet i. Pour chaque série chronologique x_i, on considère $\widetilde{x_i}$ la courbe polygonale déterminée par la suite de sommets consécutifs

$$(\widetilde{x_{i,1}},\ \widetilde{x_{i,2}},\ ...,\ \widetilde{x_{i,T_i}}) = \left(\begin{pmatrix} 1 \\ x_{i,1} \end{pmatrix}, \begin{pmatrix} 2 \\ x_{i,2} \end{pmatrix},\ ...\ , \begin{pmatrix} T_i \\ x_{i,T_i} \end{pmatrix} \right) \in \mathbb{R}^{2^{T_i}}$$

Définition : Un *chemin de déformation (warping path)* W de taille m et n est une séquence $\big((a_1, b_1)\ , (a_2, b_2),\ ..., (a_{l_W}, b_{l_W})\big)$ d'éléments distincts de $[\![1, m]\!] \times [\![1, n]\!]$ tel que l_W soit un entier supérieur ou égal à $\max(m, n)$, $a_1 = b_1 = 1$, $a_{l_W} = n$, $b_{l_W} = m$, et pour tout k dans $\{1,\ ...,\ l_W - 1\}$ nous avons $a_{k+1} - a_k$ et $b_{k+1} - b_k$ appartenant à $\{0, 1\}$.

Définition : Soient $E = (e_1, e_2, ..., e_m)$ et $F = (f_1, f_2, ..., f_n)$ deux ensembles d'éléments ordonnés de tailles respectives m et n. Soit $W = \big((a_1, b_1)\ , (a_2, b_2),\ ..., (a_{l_W}, b_{l_W})\big)$ un chemin de déformation de taille m et n. La séquence $L_W = \big((e_{a_1}, f_{b_1}), (e_{a_2}, f_{b_2}),\ ..., (e_{a_{l_W}}, f_{b_{l_W}})\big)$ constitue un *couplage (coupling)* entre E et F.

Définition : Soient x_i et x_j deux trajectoires et $\mathcal{L}_{i,j}$ l'ensemble de tous les couplages possibles entre x_i et x_j. La *dissimilarité du dynamic time warping* d_{DTW} entre x_i et x_j vaut

$$d_{DTW}\left(x_i, x_j\right) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{k=1}^{l_W} \left\| x_{i,a_k} - x_{j,b_k} \right\|_2$$

La distance de Fréchet est également "élastique" tout en pénalisant les décalages temporels. [Genolini et al., 2016] ont développé un package R nommé *kmlShape*. Il permet d'appliquer l'algorithme des kmoyennes sur des données longitudinales en utilisant une

variation discrète de la distance de Fréchet appelée distance de Fréchet généralisée. Nous présentons une autre variante : la dissimilarité discrète sommée de Fréchet (sdF).

Définitions : Soient $\widetilde{x_i}$ et $\widetilde{x_j}$ les courbes polygonales associées aux deux trajectoires x_i et x_j et soit $\mathcal{L}_{i,j}$ l'ensemble de tous les couplages possibles entre $\widetilde{x_i}$ et $\widetilde{x_j}$.

1. La *distance de Fréchet discrète* δ_{dF} entre $\widetilde{x_i}$ et $\widetilde{x_j}$ vaut

$$\delta_{dF}\left(\widetilde{x_i}, \widetilde{x_j}\right) = \min_{L_W \in \mathcal{L}_{i,j}} \max_{k=1}^{l_W} \|\widetilde{x_{i,a_k}} - \widetilde{x_{j,b_k}}\|_2$$

2. La *dissimilarité sommée de Fréchet discrète* d_{sdF} entre $\widetilde{x_i}$ et $\widetilde{x_j}$ vaut

$$d_{sdF}\left(\widetilde{x_i}, \widetilde{x_j}\right) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{k=1}^{l_W} \|\widetilde{x_{i,a_k}} - \widetilde{x_{j,b_k}}\|_2$$

Afin d'illustrer ces différentes mesures de dissimilarités (Euclidienne, Manhattan, DTW, Fréchet discrète et sdF) voici 3 exemples fictifs de trajectoires d'observance de 8 jours (Figure 2). La première (resp. la troisième) correspond à un patient qui commencerait avec une observance à 10 heures par nuit avant d'arrêter la thérapie le septième (resp. le troisième) jour. Le second patient utiliserait son appareil 5 heures chacune des nuits. On souhaiterait utiliser une mesure où la dissimilarité la plus petite est réalisée entre les individus 1 et 3.

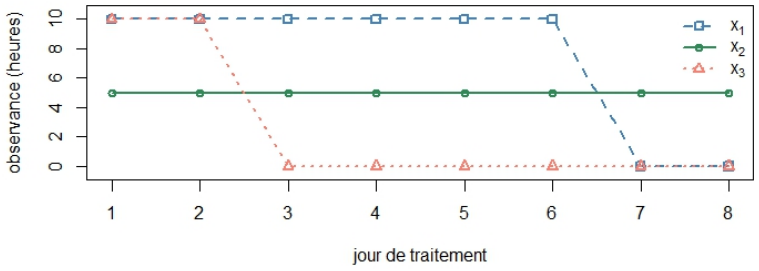

FIGURE 2 – Trois exemples fictifs de trajectoires

Le Tableau 1 présente les dissimilarités calculées entre les 3 paires d'individus pour chaque mesure. Pour chacune on indique en gras la (les) plus petite(s) valeur(s). Contrairement aux distances Euclidienne et de Manhattan, les mesures DTW, dF et sdF renvoient leurs plus petites valeurs entre x_1 et x_3. En ce sens, ces mesures semblent plus aptes à reconnaitre des comportements similaires.

dissimilarité :	Euclidienne	Manhattan	DTW	dF	sdF
$d(x_1, x_2)$	**14,14**	40	40	5	40
$d(x_2, x_3)$	**14,14**	40	40	5	40
$d(x_1, x_3)$	20,00	40	**0**	**4**	**20**

TABLE 1 – Dissimilarités entre les individus fictifs

3 Étude de simulation et application à des données réelles

Une étude de simulation permet de tester l'implémentation de ces distances dans différents algorithmes de classification. Sur des données artificielles dont nous connaissant les groupes, nous essayons de retrouver ces derniers via des classifications non supervisées. Sur la base des résultats de l'étude de [Babbin et al., 2015] nous avons simulé des jeux de données contenant 5 groupes distincts d'observance à la PPC (Figure 3). Trois paramètres varient : 1) la proportion des individus entre les 5 groupes, 2) la variance des résidus et 3) l'autocorrélation des résidus.

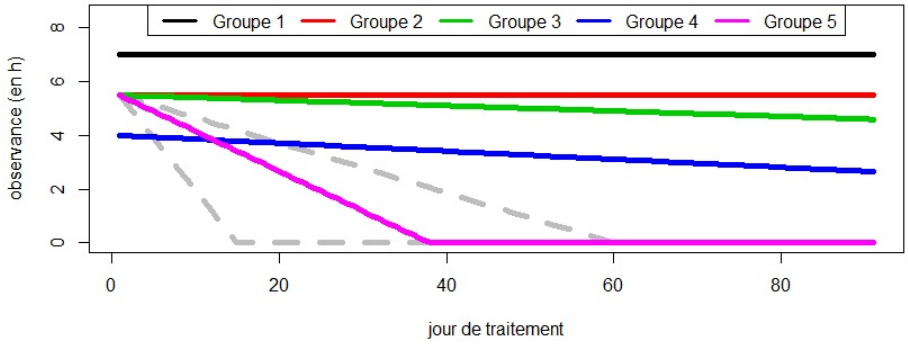

FIGURE 3 – Groupes de référence pour les simulations : les quatre premiers sont similaires aux groupes identifiés par Babbin et al., le cinquième représente des individus qui interrompent la PPC entre le 15^{eme} et le 60^{eme} jour de traitement.

Nous avons appliqué la Classification Ascendante Hiérarchique (CAH) avec les stratégies d'agglomération *average*, *complete* et *Ward*. Nous avons aussi utilisé les algorithmes des k-medoïdes et des k-moyennes. Avec chacun de ces algorithmes nous avons utilisé les distances euclidienne et de Manhattan et les dissimilarités DTW et sdF. Cependant le DTW n'a pas été appliqué aux k-moyennes car cela nécessite la définition d'un représentant autre que la moyenne classique. La qualité de chaque classification est évaluée selon l'indice de similarité [Montero et Vilar, 2014] donnant un score entre 0 et 1. Plus ce score est élevé plus les groupes obtenus sont proches des groupes théoriques.

Définition : Soient $\mathcal{G} = G_1, ..., G_K$ la partition théorique connue des individus et $\mathcal{C} = C_1, ..., C_K$ la partition obtenue après la réalisation d'une classification.
On note $sim(G_{j_1}, C_{j_2}) = \frac{2|G_{j_1} \cap C_{j_2}|}{|G_{j_1}|+|C_{j_2}|}$. L'indice de similarité $\mathcal{SIM}$ entre $\mathcal{G}$ et $\mathcal{C}$ est :

$$\mathcal{SIM}(\mathcal{G}, \mathcal{C}) = \frac{1}{K} \sum_{j_1=1}^{K} \max_{1 \leq j_2 \leq K} sim(G_{j_1}, C_{j_2})$$

Sur les simulations, le DTW fournit généralement les meilleures performances. Cependant, lorsque la variance et l'autocorrélation des trajectoires sont plus élevées, les performances de la sdF restent stables et peuvent fournir de meilleurs résultats. Dans la majorité des cas, la CAH avec le critère de Ward est la plus performante.

Sur des données réelles fournies par un prestataire de soins à domicile, nous avons appliqué la CAH avec le critère agglomératif de Ward. Le nombre de groupe est choisit à partir des hauteurs des branches des dendogrammes. Les dissimilarités DTW et sdF fournissent respectivement 6 et 7 groupes.

4 Conclusion

La validation d'une classification non supervisée, ainsi que le choix du nombre de groupe restent toujours des questions ouvertes. Sur les données réelles, des critères reconnus tels que l'indice de Calinski Harabasz ou l'indice de silhouette sont optimaux avec deux groupes seulement, limitant l'apport d'une telle analyse. Un nouveau plan de simulation intégrera et comparera différents indices de validation internes afin de déterminer une méthode permettant de choisir objectivement le nombre de profils d'observance retenus.

Bibliographie

Aghabozorgi, S. & Shirkhorshidi, A.S. & Wah T.Y. (2015). Time-series clustering – A decade review, *Information Systems*, 53, pp. 16-38.
Babbin, S.F. et al. (2015). Identifying Longitudinal Patterns for Individuals and Subgroups : An Example with Adherence to Treatment for Obstructive Sleep Apnea, *Multivariate Behavioral Research*, 50, 1, pp. 91-108.
Genolini, C. et al. (2016). kmlShape : An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes, *Plos One*, 11, 6.
Levy, P. et al. (2015). Obstructive sleep apnoea syndrome, *Nature Reviews Disease Primers*, 1.
Montero, P. et Vilar, J.A. (2014). TSclust : An R Package for Time Series Clustering, *Journal of Statistical Software*, 62, 1.
Müller, M. (2007). Dynamic Time Warping dans *Information Retrieval for Music and Motion*, Springer, pp. 69-84.
Portier et al. (2010). Traitement du SAHOS par ventilation en pression positive continue (PPC), *Revue des Maladies Respiratoires*, 27, pp. 137-145.