Clustering multi-relationnel flou des trajectoires de la douleur chronique

Armel Soubeiga[1], Violaine Antoine[1], Sylvain Moreno[2]
[1]Université Clermont Auvergne, Clermont Auvergne INP, ENSM St Etienne, UMR 6158 CNRS, LIMOS, Clermont-Ferrand, France
[2]École d'arts interactifs et de technologie, Université Simon Fraser, Vancouver, Canada
{armel.soubeiga, violaine.antoine}@uca.fr, sylvain_moreno@sfu.ca

Résumé :
L'analyse des trajectoires a récemment fait l'objet d'une attention croissante dans le domaine de la santé en raison d'une progression importante du volume de données de suivi individuel des patients. L'identification des typologies de trajectoires de soins devient ainsi un défi majeur dans la perspective d'une médecine personnalisée. Cependant, cette tâche devient plus difficile lorsque les données des trajectoires sont complexes, imprécises et subjectives. Elle est d'autant plus ardue lorsque les informations médicales sont représentées par une série temporelle discrète auto-déclarée. Dans ce travail, nous étendons l'analyse séquentielle multicanal à l'extraction de trajectoires séquentielles décrites par des séries temporelles discrètes, couvrant différents aspects de la douleur chronique. De plus, nous exploitons les avantages du clustering relationnel flou pondéré basé sur plusieurs matrices de distance. Les résultats indiquent que cette approche améliore l'interprétabilité des typologies de trajectoires identifiées pour les professionnels de la santé et permet de considérer simultanément plusieurs dimensions intervenant dans une trajectoire de soins, facilitant ainsi la mise à l'échelle.

Mots-clés :
Clustering Relationnel Flou, Séries Temporelles Discrètes, Données Imprécises, Trajectoires de Soins, Douleur Chronique.

Abstract:
Trajectory analysis has recently received increasing attention in the healthcare domain, due to a significant increase in the volume of individual patient follow-up data. The identification of care trajectory patterns is thus becoming a major challenge in the perspective of personalized medicine. However, this task becomes more difficult when trajectory data is complex, imprecise and subjective. It is all the more difficult when medical information is represented by a self-reported discrete time series. In this work, we extend multichannel sequence analysis to the extraction of sequence trajectories described by discrete time series, covering different aspects of chronic pain. In addition, we exploit the advantages of weighted fuzzy relational clustering based on multiple distance matrices. The results show that this approach improves the interpretability of the trajectory typologies identified for medical professionals, and makes it possible to simultaneously consider several dimensions intervening in a care trajectory, thus facilitating scalability.

Keywords:
Fuzzy Relationnal Clustering, Discrete Time Series, Imprecise Data, Health Trajectories, Chronic Pain.

1 Introduction

Le clustering pour l'analyse des trajectoires a suscité un intérêt croissant dans divers domaines de recherche ces dernières années. Les trajectoires les plus courantes sont les trajectoires séquentielles, les trajectoires temporelles et les trajectoires spatio-temporelles. Les trajectoires séquentielles encore appelées séquences temporelles, caractérisées par des séries temporelles discrètes (STD) ou catégorielles (STC), sont des types de trajectoires fréquemment rencontrés dans des domaines d'applications variés tels que l'analyse de logs informatiques, l'analyse de parcours de vie en science sociale [7] ou encore l'analyse de parcours de soins [2].

Selon la littérature, diverses approches ont été développées pour le clustering des trajectoires séquentielles [3, 2]. Les approches basées sur les caractéristiques consistent en l'extraction de nouvelles variables à partir des trajectoires. Les caractéristiques extraites peuvent inclure des mesures d'entropie telles que la dispersion de Gini, Shannon, Chebycheff, et d'autres mesures telles que la mesure de Pearson, la mesure de Sakoda, la mesure de Φ^2 [5]. Cependant, si elle est la plus répandue, elle ne répond pas à la définition de la notion de trajectoire. En effet, ces caractéristiques sont ensuite utilisées dans un cadre de clustering transversal [10]. Les approches basées sur des modèles, telles que les arbres suffixes probabilistes et les chaînes de Markov, ont également été largement étudiées [11]. Dans la recherche épidémiologique, par exemple, les techniques de modélisation des trajectoires couramment utilisées comprennent les

approches de modélisation des classes latentes, c'est-à-dire la modélisation des mélanges de croissance (GMM), la modélisation des trajectoires basée sur les groupes (GBTM), l'analyse des classes latentes (LCA) et l'analyse des transitions latentes (LTA) [2].

D'autres approches sont basées sur les données brutes, dont la mise en œuvre repose sur l'analyse séquentielle (AS) et le clustering relationnel [7]. L'AS permet la construction et l'appariement des séquences à partir des données de trajectoire, ainsi que la génération des mesures de similarité à vue unique entre les séquences, facilitant ainsi leur regroupement. Les techniques d'AS sont soit additives, les matrices de similarité sont additionnées, soit combinatoires, les séquences multivariées sont fusionnées en une seule séquence [7]. La principale limite de cette approche est l'interdépendance des dimensions, ignorant ainsi les liens potentiels entre ces dimensions. Néanmoins, il semble plus réaliste d'interpréter un parcours de soins comme des combinaisons de séquences de traitements.

Notre travail vise à étendre cette dernière approche dans un cadre collaboratif des dimensions, en attribuant des poids de pertinence à chaque dimension et en prenant en compte l'incertitude des données. Nous étudions des trajectoires de soins séquentielles issues de scores barométriques de 0 à 10, couvrant divers aspects tels que la fatigue, le moral, le stress, le sommeil, le confort corporel, et les activités sportives et non sportives. Nous utilisons l'AS pour extraire les trajectoires à partir des données STD et calculer les dissimilarités entre les paires de trajectoires. Pour cela, la technique d'extraction des séquences sous forme d'états (STates Sequence, STS) [7] est utilisée pour construire les trajectoires, tandis que la distance d'édition avec distorsion temporelle (Time Warp Edit Distance, TWED) [6] est employée comme fonction de mesure de distance entre les trajectoires. Enfin, nous appliquons l'approche de clustering flou basée sur les médoïdes avec plusieurs matrices de dissimilarité (MFCMdd) [8] pour identifier les typologies de trajectoires, et comparons MFCMdd avec d'autres algorithmes multi-matrices produisant des partitions dures.

Le papier est organisé comme suit. La section 2 présente les travaux connexes, permettant ainsi une contextualisation plus détaillée. Les Sections 3 et 4 passent en revue l'AS et l'algorithme MFCMdd. La section 5 présente l'application de l'approche à l'analyse des trajectoires de soins, commençant par les données de l'étude en 5.1, ensuite le contexte expérimental en 5.2 et les résultats en 5.3. Enfin, nous discutons de l'approche et des perspectives dans la section 6.

2 État de l'art

Le clustering des trajectoires séquentielles, notamment dans le domaine de la santé, est un sujet d'intérêt croissant en raison de son potentiel pour analyser les données longitudinales et les séries temporelles discrètes ou catégorielles [7]. Nguena Nguefack et al. [2] ont examiné diverses techniques de modélisation des trajectoires en épidémiologie, mettant en évidence l'importance de l'AS. L'AS permet de générer des mesures de similarité entre les trajectoires, facilitant ainsi leur regroupement. Diverses mesures de similarités entre les trajectoires séquentielles ont été proposées, parmi lesquelles les plus couramment utilisées sont l'alignement optimal ou les distances d'édition temporelle. Certaines mesures sont adaptées de celles des séries temporelles, comme DTW (Dynamic Time Warping) ou LCS (Longest Common Subsequence Problem). Une revue complète de ces mesures est disponible dans [14]. Cependant, la majorité de ces fonctions de similarité étant unidimensionnelles, lorsque des séquences multidimensionnelles sont considérées simultanément, l'analyse de séquences multicanaux (MSA) basée sur des techniques additives ou combinatoires a été proposée [15].

3 Analyse séquentielle multicanal

Les séquences multicanal peuvent être structurées en un tenseur tridimensionnel, c'est-à-dire un cube de données dont les dimensions sont définies par les individus (N), les variables (M) et le temps (T). Ainsi, les données sont définies sur $\mathbb{R}^{N\times M\times T}$ et la trajectoire $\tau^{(i)}$ d'un individu i peut être définie comme l'ensemble des séquences $S_j^{(i)}$, avec $j = 1, ...M$. La séquence S_j de la variable j est une liste d'états ou d'événements $E_t^{(k)}$, ordonnés par $t = 1, \ldots, T^{(i)}$ choisis dans un alphabet fini Σ, avec $k = 1, ..., \Sigma$. S_j peut être représenté comme une succession de paires (E_t, T_t), avec E_t représentant un état et T_t une date de mesure de l'état.

3.1 L'extraction des séquences

L'extraction des séquences est une étape cruciale dans l'analyse de séquences, qui consiste à préparer les données pour les organiser sous forme de séquences. Une ontologie définissant le format d'extraction des séquences concerne à la fois les états (événements ou statuts) ainsi que la temporalité. En se basant sur cette ontologie, plusieurs représentations séquentielles sont couramment utilisées. Il s'agit notamment du format *States Sequence (STS)*, qui énumère les états successifs d'un individu, du format *State Permanence Sequence (SPS)*, qui associe des états distincts à leur durée, du format *Distinct Successive State (DSS)*, qui fournit une représentation plus concise en mettant en évidence des états successifs uniques, et du format Vertical *Time-Stamped Event (TSE)*, qui enregistre des événements individuels avec leurs horodatages correspondants. Par exemple, considérons la trajectoire de soins d'un patient i atteint d'une maladie grave et chronique, suivie chaque semaine pendant douze semaines après le diagnostic. La variable j représente l'état des soins du patient pour chaque semaine, avec un alphabet de quatre états possibles de traitement. Les différentes représentations de sa séquence de traitement $S_j^{(i)}$ sont illustrées dans la Figure 1. A noter que le choix du format de séquençage dépend des objectifs d'analyse et de la fonction de similarité utilisée dans le cadre du clustering.

Semaine(s) depuis le diagnostic	0	1	2	3	4	5	6	7	8	9	10	11
Statut de soins	D	D	C	D	C	T	T	T	S	S	S	D

FIGURE 1 – Exemples de représentations d'une séquence de traitement à partir d'une série temporelle catégorielle ou discrète.

3.2 Mesures de Similarité de séquences discrètes

La Distance d'Edition Temporelle (TWED) est une mesure de distance pour l'appariement des séries temporelles discrètes avec élasticité temporelle [6]. Contrairement à d'autres mesures de distance temporelle, par exemple DTW ou LCS, TWED est une métrique. C'est une métrique élastique, qui exploite conjointement le décalage temporel et possède toutes les propriétés d'une distance, en particulier l'inégalité triangulaire. Sa complexité en temps de calcul est de $O(n^2)$. La distance TWED entre deux séquences discrètes est mesurée comme le coût minimum des opérations d'édition nécessaires pour transformer une séquence en une autre. Une opération d'édition entre deux séries temporelles discrètes A et B, est une séquence d'opérations qui permet leur appariement utilisant les opérations suppression-A, suppression-B et de substitution. TWED présente plusieurs avantages clés. Elle introduit une nouvelle métrique élastique (une pénalité d'écart temporelle) $\lambda > 0$, comblant l'écart entre les Lp-normes et les distances d'édition, telle que la distance d'édition avec pénalité réelle (ERP). TWED introduit également un paramètre $\nu \in [0, 1]$, appelé rigidité, permettant de contrôler son élasticité et la plaçant entre la distance euclidienne et la DTW.

4 Clustering multi-relationnel flou

Les données vectorielles et relationnelles sont couramment utilisées pour le clustering des objets. Les données relationnelles sont représentées par une matrice de dissimilarité, $\mathbf{D} = [d(e_i, e_j)]$ pour $i, j = 1, \ldots, n$, où chaque élément $d(e_i, e_j)$ indique la dissimilarité entre les objets e_i et e_j. Le clustering relationnel utilise cette matrice pour regrouper les objets, ce qui est particulièrement utile lorsque la mesure de la distance est complexe ou difficile à exprimer mathématiquement, comme avec la fonction TWED.

4.1 Fuzzy c-means relationnel

Le Fuzzy c-means relationnel ou fuzzy c-medoids (FCMdd) est une variante de fuzzy c-means (FCM) conçue pour le regroupement de données floues [4]. En considérant $\mathbf{E} = \{e_1, \ldots, e_n\}$ l'ensemble de n objets, $\mathbf{D}$ la matrice de dissimilarité et $\mathbf{G} = \{G_1, \ldots, G_K\}$ un sous-ensembles de $\mathbf{E}$, représentant les médoïdes des K clusters, l'algorithme FCMdd cherche à minimiser la fonction objectif J_{FCMdd} définie comme suit :

$$\begin{aligned} J_{FCMdd}(\mathbf{U}, \mathbf{G}) &= \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m D(e_i, G_k), \\ \text{tel que } \sum_{k=1}^{K} u_{ik} &= 1, u_{ik} > 0 \quad \forall i = 1, \ldots, n. \end{aligned} \tag{1}$$

Avecu_{ik} représentant le degré d'appartenance de l'objet e_i au cluster C_k et $D(e_i, G_k)$ mesurant la dissimilarité entre l'objet e_i et le prototype (médoïde) G_k du cluster C_k. $m \in [1, \infty)$ est un paramètre de pondération contrôlant le degré de flou de la partition finale.

MFCMdd [8] est une extension de FCMdd qui partitionne les objets en utilisant plusieurs matrices de dissimilarité simultanément. Son objectif est de collaborer entre ces matrices pour obtenir une partition finale consensuelle. MFCMdd génère une partition floue tout en apprenant des poids de pertinence pour chaque matrice de dissimilarité. Ces poids sont ajustés à chaque itération de l'algorithme et peuvent être estimés soit localement pour chaque cluster, soit globalement pour l'ensemble des clusters.

4.2 MFCMdd avec poids de pertinence estimés localement

L'algorithme MFCMdd-RWL (MFCMdd with relevance weight estimated locally) [8] fournit une partition floue d'un ensemble $\mathbf{E}$ en K clusters et un prototype pour chaque cluster. Il apprend un vecteur de poids de pertinence, $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$, pour chaque matrice de dissimilarité, qui varient à chaque itération et diffèrent selon les clusters, avec $\boldsymbol{\lambda}_k \in \mathbb{R}^p$. La fonction objectif mesurant l'adéquation entre les clusters et leurs prototypes est définie par :

$$\begin{aligned} &J_{MFCMdd-RWL}(\mathbf{U}, \mathbf{G}, \mathbf{\Lambda}) = \\ &\sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (\lambda_{kj})^s D_j(e_i, G_k). \end{aligned} \tag{2}$$

$(\lambda_{kj})^s D_j(e_i, G_k)$ représente la relation entre un exemple $e_i \in \mathbf{E}$ et le prototype du cluster $G_k \in \mathbf{E}$, paramétrée par $s \geq 1$ et le vecteur de poids $\mathbf{\Lambda}$ des matrices de dissimilarité D_j dans le cluster C_k. $J_{MFCMdd-RWL}$ est optimisée sous les contraintes suivantes [8] :

$$\begin{aligned} \sum_{k=1}^{K} u_{ik} = 1, u_{ik} > 0 \quad \forall i = 1, ..., n, \\ (a) : s = 1,\ \lambda_{kj} > 0, \prod_{j=1}^{p} \lambda_{kj} = 1, \\ \text{ou } (b) : s \geq 1,\ \lambda_{kj} \in [0, 1], \sum_{j=1}^{p} \lambda_{kj} = 1. \end{aligned} \tag{3}$$

Ainsi, deux versions de l'algorithme sont proposées : MFCMdd-RWL-P pour l'optimisation sous (3-a) et MFCMdd-RWL-S pour l'optimisation sous (3-b). L'algorithme suit les mêmes étapes que l'algorithme FCMdd en ajoutant une étape de calcul des poids de pertinence.

4.3 MFCMdd avec poids de pertinence estimés globalement

L'algorithme MFCMdd-RWL peut présenter une instabilité numérique lorsqu'il génère des clusters uniques ou des clusters contenant des objets ayant une dissimilarité nulle entre eux, indiquant $\sum_{i=1}^{n}(u_{ik})^m D_j(e_i, G_k) \rightarrow 0$. Pour adresser cette limite, MFCMdd-RWG (MFCMdd with global relevance weight) [8] a été proposé. Il fournit ainsi une partition floue et un vecteur de poids $\boldsymbol{\lambda} \in \mathbb{R}^p$ estimé globalement. Sous les mêmes contraintes d'optimisation que Eq. (3), avec $\lambda_j \quad \forall k = 1, ..., K$, on peut également avoir les versions MFCMdd-RWG-P et MFCMdd-RWG-S [8].

5 Application aux parcours de soins

L'objectif de cette étude est d'identifier des typologies de parcours des patients souffrant de douleur chronique. Nous commençons par décrire les données des trajectoires de soins des patients dans la section 5.1. Ensuite, nous présentons le cadre expérimental dans la section 5.2. Enfin, les résultats sont présentés dans la section 5.3.

5.1 Les trajectoires de soins de la douleur chronique

Les données ont été collectées à l'aide de l'application mHealth eDOL [1], permettant aux patients et à leurs médecins de compléter des questionnaires cliniques, personnels et barométriques concernant l'état de douleur des patients. Les huit attributs barométriques (douleur, fatigue, moral, stress, sommeil, confort corporel, activité sportive et non sportive) mesurés hebdomadairement permettent d'évaluer l'intensité de la douleur et ses répercussions. Les patients attribuent un score de 0 à 10 pour chaque baromètre via l'application mobile, fournissant des informations sur leur perception subjective de la douleur. En 2019, une étude de faisabilité a montré un taux d'adhésion initial de 61,9% à l'application eDOL. A ce jour, sur 1 590 patients inclus, ce taux s'est amélioré pour atteindre 67,3%. Environ 38% des patients ont été exclus pour des données trop incomplètes. Parmi les 986 patients retenus, seules les données de 636 ont été analysées, totalisant 14 090 séries de remplissage sur une moyenne de suivi de 5 mois, avec une durée totale d'environ 19 mois. Ce jeu de données présente plusieurs défis de modélisation que notre approche aborde à différentes étapes. Nous utilisons la fonction TWED pour gérer les séquences irrégulières dues aux durées de suivi variables des patients, en adaptant les alignements temporels pour réduire les écarts entre les séquences. Les données manquantes sont codées comme des statuts de non-observance (hors suivi). La fonction TWED prend également en compte la nature discrète des données. De plus, le caractère imprécis et incertain des informations, notamment les données subjectives sur le ressenti des patients, est adressé grâce au clustering flou, qui gère cette dimension d'incertitude.

5.2 Contexte expérimental

Sur ces jeux de données eDOL, les variables constituant les trajectoires ont toutes le même alphabet (domaine) $\Sigma = [0 - 10]$, c'est-à-dire que toutes les valeurs possibles de ces variables sont dans Σ. Les coûts de substitution entre deux statuts sont dérivés de la somme des coûts d'intel (suppression-A ou suppression-B) estimés pour chacun d'eux. Les coûts d'intel ont été calculés de façon empirique à travers les fréquences relatives, telles que définies dans [14], et exprimées par la formule exprimées par $indel_{a_p} = log\left[2/(1 + f_{a_p})\right]$, où f_{a_p} représente la fréquence observée du statut a_p. Les paramètres $\nu = 0.5$ et $\lambda = 0.5$ ont été fixés par défaut. Toutes les matrices de dissimilarité ont été normalisées en fonction de leur dispersion globale. Cela implique que chaque dissimilarité $d_{ii'} = \delta_{\lambda,\nu}\left(S_j^{(i)}, S_j^{(i')}\right)$ dans une matrice de dissimilarité $j = 1, \ldots, 8$, a été normalisée selon la formule $2 * d_{ii'}/(m + d_{ii'})$, où m représente la dissimilarité maximale possible de la matrice

j. Le flux de travail expérimental est résumé à travers la Figure 2.

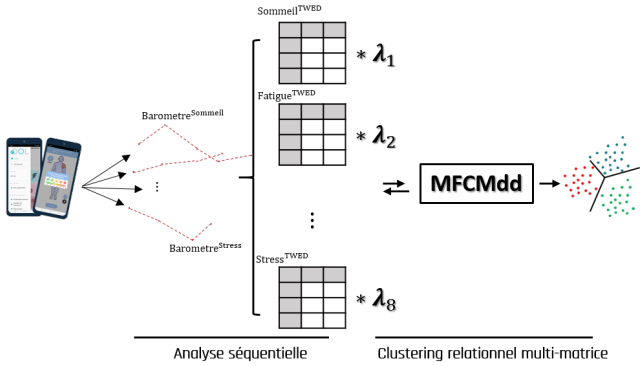

FIGURE 2 – Flux de travail expérimental.

Nous avons choisi de fixer le nombre de clusters à 3, en nous appuyant sur les travaux de [10] sur les mêmes données qui ont identifié, à travers une recherche minutieuse impliquant plusieurs itérations, que 3 clusters correspondait à une solution optimale. Nous avons également comparé notre approche de partitionnement flou à un partitionnement dur en utilisant l'algorithme de clustering relationnel dur dynamique basé sur des matrices de dissimilarité multiples (MRDCA, pour Multiple Relational Dynamic Clustering Algorithm) [13]. MRDCA fournit une partition dure avec des poids de pertinence pour chaque vue, estimés soit localement, soit globalement. Pour mesurer et comparer la performance des partitions, nous avons considéré deux principaux critères d'évaluation. Dans une première comparaison entre les variantes de la partition floue (MFCMdd), nous utilisons l'Entropie de Partition (PE) utilisant le degré de flou [9]. Cette étape a permis d'identifier l'algorithme flou le plus performant, compte tenu des contraintes de somme et de produit des poids égal à 1. Dans une seconde comparaison, l'algorithme MFCMdd avec le schéma d'optimisation le plus performant est comparé à MRDCA. Cette comparaison est basée sur le critère de l'indice de silhouette (SI) adapté aux données relationnelles p-matrices utilisant le poids de chaque matrice [12]. Les partitions floues sont transformées en partitions dures en utilisant la règle du principe du maximum, comme suit : $G_k = \{\mathbf{x_i} : u_{ik} \geq u_{im} \quad \forall m \in \{1, \ldots, K\}\}$. Tous les algorithmes ont été exécutés 10 fois, avec des initialisations différentes. Ensuite, nous sélectionnons les scores PE et SI correspondant au coût le plus bas de la fonction objectif de l'algorithme. Une valeur minimale de PE signifie une meilleure partition, tandis que SI est maximisé. Nous avons fixé le paramètre m gérant le degré de flou à 1.5.

5.3 Résultats

La Table 1 montre que les algorithmes MFCMdd-RWG-S et MFCMdd-RWL-S ont les coûts les plus bas (0.098). Cela signifie que la contrainte de somme sur les poids produit des minima locaux plus petits et est moins sensible aux initialisations. MFCMdd-RWL-S offre la meilleure qualité de partitionnement avec un indice de silhouette (SI) de 0.539 et une entropie de partition (PE) de 0.076. Les algorithmes MRDCA ont des coûts plus élevés et des indices de silhouette inférieurs, indiquant une performance moindre.

TABLEAU 1 – Performance des algorithmes de clustering en termes de PE et SI.

Algorithmes	J_{Cost}	PE	SI
MFCMdd-RWG-P	6.330	0.081	0.520
MFCMdd-RWG-S	**0.098**	0.079	0.528
MFCMdd-RWL-P	6.310	0.081	0.521
MFCMdd-RWL-S	**0.098**	**0.076**	**0.539**
MRDCA-RWG	8.280	-	0.481
MRDCA-RWL	8.240	-	0.509

Cependant une comparaison qualitative des partitions floues à l'aide d'un tableau de confusion, illustrée dans la Figure 3, révèle que les erreurs de classification sont relativement faibles, avec des résultats similaires dans 96% des cas.

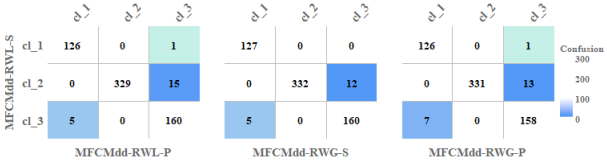

FIGURE 3 – Comparaison qualitatives entre les partitions floues.

La Figure 4 montre une forte appartenance des patients des clusters 1 et 2 à leurs clusters respectifs, indiquée par une médiane proche de 1. En revanche, le cluster 3 compte davantage d'objets imprécis, compte tenu du degré de flou de plusieurs patients proche de 1/3.

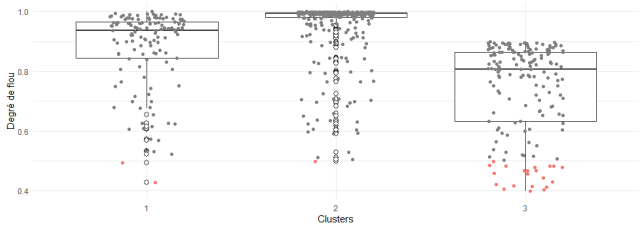

FIGURE 4 – Distribution des valeurs d'appartenance flou.

La table 2 montre les valeurs des poids de pertinence pour chaque dimension (Douleur, Stress, Fatigue, Sommeil, Moral, Confort corporel, Activité non-sportive, Activité sportive) dans chaque cluster selon la meilleure partition trouvée (MFCMdd-RWL-S). Les valeurs représentent l'importance relative de chaque dimension dans la caractérisation de chaque cluster. Le cluster 1 se caractérise par des valeurs élevées pour les dimensions de Douleur, Fatigue, Sommeil et Confort corporel, suggérant des profils associés à une douleur chronique avec fatigue, perturbation du sommeil et altération du confort corporel. Le cluster 2 présente des valeurs plus élevées pour les dimensions de Stress, Moral, Activité non-sportive et Activité sportive, indiquant des trajectoires où le stress, les aspects émotionnels et l'activité physique jouent un rôle plus prédominant. En revanche, le cluster 3 semble plus généraliste que les autres, sans caractéristiques spécifiques en termes de symptômes ou de comportements.

La Figure 5 la meilleure partition trouvée (MFCMdd-RWL-S) montre le nombre de patients par cluster, sur 20 semaines de suivi et en fonction de l'évolution de leur douleur caractérisée par l'évolution observée de chaque baromètres. Chaque tapis illustre en ligne la séquence des fréquences d'état transversales

TABLEAU 2 – Valeurs des poids de pertinence de chaque dimension des trajectoires pour chaque cluster.

	Cluster 1	Cluster 2	Cluster 3
Douleur	**0.147**	0.136	0.142
Stress	0.135	**0.137**	0.131
Fatigue	**0.148**	0.135	0.146
Sommeil	**0.143**	0.125	0.135
Moral	0.129	**0.133**	0.130
Confort corporel	**0.120**	0.095	0.115
Activité non-sportive	0.086	**0.124**	0.109
Activité sportive	0.093	**0.114**	0.091

du cluster concerné pour le baromètre correspondant. L'état Faible correspond à une valeur du baromètre inférieure à 4, et Modéré sinon. L'état hors de suivi correspond à la non-observation du patient pendant la période donnée. On observe que les trajectoires des douleurs des patients du cluster 2 deviennent irrégulières après dix semaine. Par exemple, tandis que la douleur et le stress diminuent, les patients présentent des niveaux modérés de fatigue et de moral.

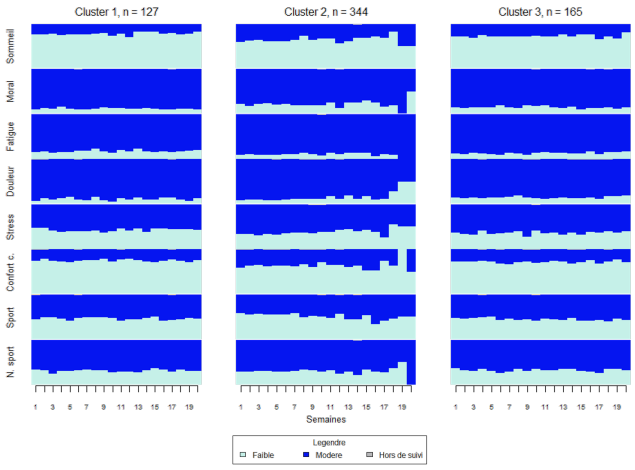

FIGURE 5 – Visualisation des clusters de trajectoires de la douleur chronique des patients eDOL sur une fenêtre de 20 semaines.

6 Discussions et conclusion

Dans cet article, nous avons proposé une approche qui combine l'analyse séquentielle et le clustering flou basé sur plusieurs matrices de dissimilarité. Contrairement aux autres

méthodes de clustering des trajectoires, elle présente l'avantage, d'une part, de travailler avec les données brutes sans aucune transformation des séries temporelles, et aussi de pouvoir utiliser des fonctions de distance complexes adaptées à chaque dimension et à la nature des données. Notre approche permet de pondérer les dimensions pour obtenir une partition floue avec des poids de pertinence pour chaque dimension. Elle est plus facile à mettre en œuvre et peut gérer des dimensions plus importantes que les méthodes d'analyse séquentielle multicanales existantes, dont les objectifs sont additifs [7]. L'étude de cas sur le projet eDOL pour l'identification des typologies de trajectoires de soins pour la douleur chronique illustre l'efficacité de l'approche pour identifier des trajectoires types. Ces résultats prometteurs ouvrent de nouvelles perspectives en analyse de parcours de soins. Notre travail peut cependant présenter certaines limites. Tout d'abord, nous avons opté pour le clustering relationnel basé sur les médoïdes, bien que d'autres méthodes de clustering relationnel flou peuvent être étudiées. Le choix des paramètres par défaut de la fonction de dissimilarité TWED aurait pu être optimisé. Pour les travaux futurs, nous prévoyons d'étudier le profil des patients de chaque cluster à travers l'utilisation des données socio-démographiques et cliniques des patients associées à l'étude. Nous envisageons également d'étendre cette approche de clustering basée sur plusieurs matrices de dissimilarité à d'autres techniques de clustering pouvant traiter l'incertitude.

Remerciements :

Les auteurs remercient l'Agence nationale de la recherche française pour son soutien dans le cadre du programme Investissements d'avenir (16-IDEX-0001 CAP 20-25). Les auteurs souhaitent remercier également l'Institut Analgesia pour la mise à disposition de l'ensemble des données du projet eDOL.

Références

[1] Kerckhove, Nicolas et al. eDOL mHealth App and Web Platform for Self-monitoring and Medical Follow-up of Patients With Chronic Pain : Observational Feasibility Study. *JMIR Form Res*, 2022, 6(3).

[2] Nguena Nguefack, Hermine Lore, Pagé, M. Gabrielle, Katz, Joel, Choinière, Manon, Vanasse, Alain, Dorais, Marc, Samb, Oumar Mallé, Lacasse, Anaïs. Trajectory modelling techniques useful to epidemiological research : A comparative narrative review of approaches. *Clinical Epidemiology*, 2020, 12, pp. 1205-1222. Dove Medical Press Ltd.

[3] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, Teh Ying Wah. Time-series clustering – A decade review. *Information Systems*, 2015, 53, pp. 16-38.

[4] Krishnapuram, Raghu, Joshi, Anupam, Yi, Liyu. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315)*, 1999, 3, pp. 1281-1286. IEEE.

[5] López-Oriona, Ángel, Vilar, José A. Ordinal Time Series Analysis with the R Package otsfeatures. *Mathematics*, 2023, 11(11), 2565. MDPI.

[6] Marteau, Pierre François. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(2), pp. 306-318.

[7] Robette, Nicolas. L'analyse statistique des trajectoires : Typologies de séquences et autres approches. *Ined Éditions*, 2021.

[8] de Carvalho, Francisco de A.T., Lechevallier, Yves, de Melo, Filipe M. Relational partitioning fuzzy clustering algorithms based on multiple dissimilarity matrices. *Fuzzy Sets and Systems*, 2013, 215, pp. 1-28.

[9] Wang, Hong-Yu, Wang, Jie-Sheng, Wang, Guan. A survey of fuzzy clustering validity evaluation methods. *Information Sciences*, 2022, 618, pp. 270-297.

[10] Soubeiga, Armel, Ettaghouti, Jessem, Antoine, Violaine, Corteval, Alice, Kerckhove, Nicolas, Moreno, Sylvain. Classification automatique de séries chronologiques de patients souffrant de douleurs chroniques. *Revue des Nouvelles Technologies de l'Information*, 2023, Extraction et Gestion des Connaissances, RNTI-E-39, pp. 651-652.

[11] Bouveyron, Charles, Brunet-Saumard, Camille. Model-based clustering of high-dimensional data : A review. *Computational Statistics & Data Analysis*, 2014, 71, pp. 52-78.

[12] Renê Pereira de Gusmão and Francisco de A.T. de Carvalho. Clustering of multi-view relational data based on particle swarm optimization. *Expert Systems with Applications*, vol. 123, 2019, pp. 34-53.

[13] Francisco de A.T. de Carvalho, Yves Lechevallier, Filipe M. de Melo. Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, 2012, 45(1), pp. 447-464.

[14] Studer, Matthias, et Gilbert Ritschard. What Matters in Differences Between Life Trajectories : A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society Series A : Statistics in Society* 179.2 (2015) : 481-511.

[15] Gilbert Ritschard, Tim F. Liao, et Emanuela Struffolino. "Strategies for Multidomain Sequence Analysis in Social Research." *Sociological Methodology* 53.2 (2023) : 288-322.