Clustering flou de séries temporelles discrètes pour la modélisation des trajectoires de soins de la douleur chronique

Armel Soubeiga, Violaine Antoine, Sylvain Moreno

Clustering flou de séries temporelles discrètes pour la modélisation des trajectoires de soins de la douleur chronique

Armel Soubeiga*, Violaine Antoine*, Sylvain Moreno**

*Université Clermont Auvergne, CNRS, LIMOS, ENSMSE LIMOS, F-63000 Clermont Ferrand, France
**Université Simon Fraser, Digital Health Hub, Vancouver, Canada

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 cette étude, 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 basé sur de multiples matrices de distance. Nous proposons une approche de clustering basée sur la distance qui intègre l'extraction de séquences de traitement discrètes, le calcul de relations de dissimilarité entre paires de séquences et le clustering relationnel collaboratif souple. Cette méthode a été appliquée à des données réelles issues d'une étude de suivi quotidien de patients souffrant de douleurs chroniques en France. 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.

1 Introduction

Les données de séries temporelles (TS) sont aujourd'hui omniprésentes. Alors que la majeure partie de la littérature sur ce sujet traite des séries temporelles à valeurs continues, les séries temporelles catégorielles et discrète ont reçu beaucoup moins d'attention, même si elles se présentent de manière naturelle dans de nombreuses applications. Des exemples incluent les décomptes hebdomadaires de nouvelles infections par une maladie spécifique (Weiss et Pollett (2014)), les enregistrements temporels des états de sommeil EEG (Stoffer et al. (1993)), la modélisation stochastique de données de séquence d'ADN (Fokianos et Kedem (2003)) ainsi que l'utilisation de processus de Markov cachés pour traiter les séquences de protéines (Krogh et al. (1994)), entre autres.

Les séries temporelles discrètes (DTS) que nous étudions ici sont constituées de trajectoires séquentielles comprenant des valeurs de scores barométriques sur une échelle de 0 à 10. Ces séries couvrent différents aspects de la douleur tels que la fatigue, le moral, le stress, le sommeil, le confort corporel, ainsi que l'activité sportive et non sportive. En vérité, la douleur chronique touche des millions de patients en France, soit environ 30% de la population générale. Les traitements disponibles sont anciens, ont une efficacité limitée et peuvent entraîner des effets indésirables importants (Kerckhove et al. (2022)). De plus, le parcours de santé des patients souffrant de douleur chronique est multiple ce qui entraîne des résultats médiocres en terme d'amélioration de leur santé. L'identification des typologies de parcours et les profils des patients permettrait aux corps médical d'améliorer les résultats de soins et de mieux soutenir ces patients. Pour aborder cette problématique, nous suggérons une approche de regroupement basée sur l'analyse séquentielle. C'est une technique de clustering basé sur la distance et dont le défi réside dans la recherche des mesures de dissimilarité optimales entre les paires de séquences qui sont extraites des données de DTS. Ces données de DTS sont complexes, sujettes à l'incertitude et au bruit. Il existe deux types d'approche pour le clustering de trajectoires DTS : la première se base sur des caractéristiques ou des modèles et la seconde sur la distance entre trajectoires (Nguefack et al. (2020)). L'approche basée sur la distance présente l'avantage de travailler directement avec les données brutes pour conserver le caractère longitudinal des séries temporelles, et surtout de bénéficier des performances des algorithmes de clustering relationnel. Dans ce travail, nous proposons une approche qui se situe à l'intersection des sciences sociales et de l'apprentissage automatique non supervisé. En effet, nous proposons l'utilisation de l'approche d'analyse séquentielle multichannel (MSA) (Robette (2021)) pour l'extraction des trajectoires séquentielles à partir des données DTS et le calcul des dissimilarités entre les pairs de trajectoires. Le states sequence (STS) processing est utilisé pour l'extraction des trajectoires et la Time Warp Edit Distance (TWED) est utilisé comme la fonction de mesure de distance entre les trajectoire. TWED est une mesure unidimensionnelle correspondant à l'élasticité temporelle et construit une matrice distance entre pairs de trajectoire par dimension. Une approche de clustering relationnelle floue basé sur les médoïdes avec de multiples matrices de dissimilarité (MFCMdd) est ensuite utilisé à la suite de MSA pour l'identification des typologies de trajectoires.

Le papier est organisé comme suit. La section 2 présente les données de l'étude et une analyse descriptive sur ces donnée. Les Sections 3 et 4 passent en revue MSA et l'algorithme MFCMdd qui introduisent l'approche de clustering proposée. Enfin la Section 5 présente les résultats du clustering et des analyses décisionnelles.

2 Données des trajectoires de soins

Les données ont été collectées à l'aide de l'application mHealth eDOL (Kerckhove et al. (2022)), permettant aux patients et à leurs médecins de compléter des questionnaires cliniques, personnels et barométriques concernant la douleur chronique. Les six attributs barométriques (douleur, fatigue, moral, stress, sommeil, confort corporel, activité sportive et non sportive) mesurés chaque semaine permettent d'évaluer l'intensité de la douleur et ses répercussions. Ces baromètres sont complétés par les patients via l'application mobile. Les patients attribuent un score de ressenti sur une échelle de 0 à 10 pour chaque baromètre. Ces évaluations fournissent

des informations sur la perception subjective de la douleur par les patients. Cependant, lors de la modélisation, il est essentiel de prendre en compte plusieurs aspects de ces données. Tout d'abord, il est important de noter que les séries temporelles dérivées de ces mesures sont souvent caractérisées par une irrégularité, car certains patients n'ont pas complété les baromètres chaque semaine. Cette irrégularité est traitée comme une non-observance du patient lors du calcul des dissimilarités entre les patients. En outre, le caractère discret est pris en compte en utilisant une mesure de distance séquentielle adaptée aux DTS. Enfin, il est essentiel de prendre en compte le caractère imprécis et incertain des données déclarées, compte tenu du contexte de collecte, notamment des informations subjectives sur le ressenti des patients. Cette dimension d'incertitude ajoute de la complexité à l'analyse, et est prise en compte par l'utilisation du clustering souple. 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 incomplètes. Parmi les 986 patients retenus, seules les données de 664 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.

3 Analyse de trajectoires séquentielles

En science médicale, l'analyse de trajectoires séquentielle concerne l'analyse d'ensemble de séquences catégorielles ou discrètes qui décrivent généralement des données longitudinale ou temporelle. Les séquences analysées sont des représentations codées, par exemple, de trajectoires de soins individuelles telles que les consultations, les transitions de soins de ville et hospitalières, les consommations pharmaceutique, mais elles peuvent également décrire l'utilisation du temps quotidien ou hebdomadaire ou représenter l'évolution de la santé observée ou auto-déclarée.

3.1 Clustering de trajectoires séquentielles

Le clustering séquentiel de trajectoire vise à décrire un ensemble de trajectoires en les classant dans un nombre fini de clusters, en fonction de leurs caractéristiques de séquences. Cependant, le séquençage des DTS est une tâche complexe dans le domaine de l'analyse séquentielle, en particulier lorsque de nombreuses séries temporelles sont prises en compte simultanément (Richter et al. (2015)). Bien qu'il existe de nombreuses études sur l'exploration de données séquentielles, le regroupement de séquences discrètes requiert encore plus d'attention dans la littérature, étant donné son utilité dans de nombreux domaines. Dans le sequence mining, le clustering séquentiel nécessite soit le calcul de la (dis)similarité entre les séquences, soit la découverte du modèle sous-jacent générant les séquences. Les approches basées des modèles, telles que les arbres suffixes probabilistes et les chaînes de Markov, ont été largement étudiées (Bouveyron et Brunet-Saumard (2014)). 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) (Nguefack et al. (2020)). Des études récentes évoquent également des approches basées sur les caractéristiques pour le clustering de trajectoires séquentielles. Les caractéristiques extraites peuvent inclure des mesures d'en-

tropie telles que la dispersion de Gini, Shannon, Chebycheff, et d'autres mesures telles que le coefficient d'incertitude, la mesure de Pearson, la mesure de Sakoda, la mesure de Φ^2, et bien d'autres encore. Ces caractéristiques sont ensuite utilisées dans un cadre de clustering transversal basé sur les variables (López-Oriona et Vilar (2023)).

Dans ce papier, nous nous concentrons au clustering basé sur la distance (similarité ou dissimilarité) en utilisant l'analyse de séquences multiples.

3.2 Analyse séquentielle multicanal

L'analyse séquentielle multidimensionnelle s'intéresse à l'évolution dans le temps de plusieurs séquences donc de plusieurs variables catégorielles ou discrètes. Ces données 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.

3.2.1 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. Encore appelé sequence processing, il peut varier considérablement en fonction de la manière dont les données ont été collectées et de la façon dont l'information est organisée. C'est souvent une tâche laborieuse et intimidante en raison de la nature désordonnée des données de trajectoire. Par exemple, les taux d'échantillonnage et les longueurs peuvent varier d'une trajectoire à l'autre. De plus, la littérature offre peu de ressources pour le séquençage de DTS, (Giele et Elder (1998)) étant l'une des rares exceptions. 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.

Soit par exemple la trajectoire de soins d'un patient i atteint d'une pathologie grave et chronique, caractérisée par le suivi de son parcours de soins toutes les semaines pendant trois mois (douze semaines) à partir du diagnostic. La variable j contient l'état des soins du patient pour chaque semaine de suivi, avec un alphabet de quatre, représentant ainsi quatre états possibles de traitement. Les différentes représentations de sa séquence de traitement $S_j^{(i)}$ sont illustrées dans la Figure 1.

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

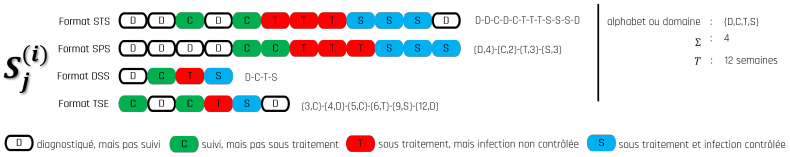

FIG. 1 – *Exemples de représentations d'une séquence de traitement à partir d'une série temporelle catégorielle ou discrète*

3.2.2 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, (Marteau (2009)). Par rapport à d'autres mesures de distance (par exemple DTW (Dynamic Time Warping) ou LCS (Longest Common Subsequence Problem)), 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 $O(n^2)$, mais peut être réduite de manière drastique dans certaines situations spécifiques en utilisant un couloir pour réduire l'espace de recherche (Halpin (2014)). La distance TWED entre deux séquences discrètes est mesurée comme le coût minimum des opérations d'édition nécessaires pour transformer une séquences en une autre. Pour définir les opérations d'édition, les auteurs utilisent le paradigme d'un processus d'édition graphique et aboutissent à un algorithme de programmation dynamique. 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, telles 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 le plaçant entre la distance euclidienne et DTW.

Soit un ensemble de séries temporelles finies $\{A_p/p \in \mathbb{N}\}$. A_1^p est une série temporelle avec des indices temporels discrets variant de 1 à p. A peut être transcrite en une séquence discrète finie, avec a' le $i^{\text{ème}}$ statut (mesure ou évènement) A. A_i^j avec $i \leq j$ est la sous-séquence composée des statuts $i^{\text{ème}}$ au $j^{\text{ème}}$ (inclus) de A. Une opération d'édition est une paire (a', b') représentant des échantillons de l'alphabet de A, notée $a' \rightarrow b'$. TWED recherche une séquence d'opérations d'édition qui peuvent simultanément transformer deux séquences avec un coût minimal. En adoptant un paradigme d'éditeur graphique, TWED permet à deux séries temporelles, A et B, d'être modifiées en utilisant trois opérations d'édition élémentaires : suppression-A, suppression-B et opérations de substitution (ou d'appariement), au lieu des opérations classiques de suppression, insertion et substitution généralement développées dans les distances d'édition. Pour cela, TWED utilise une fonction de coût $\delta_{\lambda,\nu}(A_1^p, B_1^q)$, où p et q sont des indices qui parcourent les échantillons des séries A et B. Cette fonction de coût représente la dissimilarité minimale entre deux sous-séquences des séries A et B jusqu'à l'échantillon p de A et l'échantillon q de B. Plus précisément, $\delta_{\lambda,\nu}(A_1^p, B_1^q)$ est définie comme suit :

$$\delta_{\lambda,\nu}(A_1^p, B_1^q) = \min \begin{cases} \delta_{\lambda,\nu}(A_1^{p-1}, B_1^q) + d(a'_p, a'_{p-1}) + \nu + \lambda & \textit{delete-A} \\ \delta_{\lambda,\nu}(A_1^{p-1}, B_1^{q-1}) + d(a'_p, b'_q) + d(a'_{p-1}, b'_{q-1}) + 2\nu\,|p-q| & \textit{match} \\ \delta_{\lambda,\nu}(A_1^p, B_1^{q-1}) + d(b'_{q-1}, b'_q) + \nu + \lambda & \textit{delete-B} \end{cases}$$

Où $d(a'_p, b'_q)$ est le coût de substitution entre l'élément à l'indice p de la séquence A et l'élément à l'indice q de la séquence B. Le processus d'édition commence avec $p = q = 1$. Une opération de substitution incrémente à la fois p et q, une opération de suppression-A incrémente p et une opération de suppression-B incrémente q. La matrice de coût de substitution a ainsi une dimension $\Sigma \times \Sigma$, où l'élément (i, j) de la matrice est le coût de substitution du statut a'_p avec le statut b'_q. La littérature propose diverses méthodes empiriques pour générer des coûts de substitution et d'indel utilisés dans les analyses (Studer et Ritschard (2015)).

4 Fuzzy c-medoids basé sur multiple matrices

Les données de caractéristiques et les données relationnelles sont deux représentations courantes des objets sur lesquelles le clustering peut se baser. Lorsque chaque objet est décrit par un vecteur de valeurs quantitatives ou qualitatives, l'ensemble des vecteurs décrivant les objets est appelé données caractéristiques (feature data). Par ailleurs, lorsque chaque paire d'objets est représentée par une relation, on parle de données relationnelles. Le cas le plus courant de données relationnelles est celui où l'on dispose d'une matrice de dissimilarité, notée $\mathbf{D} = [d(e_i, e_j)]$ avec des éléments $d(e_i, e_j)$, qui représentent des dissimilarités, souvent des distances, entre les objets i et j. Le clustering relationnel permet de regrouper ou de partitionner les objets sur la base de la matrice de dissimilarité $\mathbf{D}$ résultante. Ce type de regroupement est particulièrement utile lorsque le calcul de la distance entre les objets est complexe ou lorsque la mesure de la distance ne peut être exprimée de manière mathématiquement simple, comme dans le cas de TWED. Le regroupement relationnel basé sur des matrices de dissimilarité multiples est très important pour le regroupement des trajectoires, car la plupart des mesures de similarité sont unidimensionnelles et génèrent une matrice par dimension. C'est également le cas de la mesure TWED adaptée aux trajectoires DTS.

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 (Krishnapuram et al. (1999)). Contrairement à FCM, FCMdd identifie des médoïdes pour chaque cluster, ce qui permet une représentation plus juste des clusters. Cette méthode attribue des degrés d'appartenance flous aux données, ce qui signifie que chaque point a une probabilité d'appartenance pour chaque cluster. La méthode FCMdd vise à minimiser la dissimilarité entre les points de données et les médoïdes. En supposant $\mathbf{E} = \{e_1, \ldots, e_n\}$ l'ensemble de n objets, $\mathbf{D} = [d(e_i, e_l)]$ la matrice de dissimilarité mesurant la dissimilarité entre les objets e_i et e_l et $\mathbf{G} = \{G_1, \ldots, G_k\}$, sous-ensembles de $\mathbf{E}$, représentant les médoïdes (centre) des k clusters. Ainsi FCMdd, consiste à minimiser la fonction objectif (J_{FCMdd}) définie comme suit :

$$J_{FCMdd}(\mathbf{U},\mathbf{G}) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m D(e_i,G_k) \tag{1}$$

Avec, u_{ik} représentant le degré d'appartenance de l'objet e_i au cluster C_k et $D(e_i,G_k)$ mesure la dissimilarité entre l'objet e_i et le prototype du cluster G_k. $m \in [1,\infty)$ est un paramètre de pondération contrôlant le degré de flou de la partition finale. L'algorithme minimisant J_{FCMdd} commence par initialiser les centres, puis répète itérativement le calcul des prototypes, suivi du calcul de la partition floue $\mathbf{U}$ jusqu'à leurs convergence.

Le multiple relational fuzzy c-medoids (MRFCMdd) (de A.T. de Carvalho et al. (2013)), est une extension de l'algorithme FCMdd qui vise à partitionner des objets en prenant en compte leurs descriptions relationnelles fournies par plusieurs matrices de dissimilarité simultanément. L'objectif est d'obtenir une collaboration entre les différentes matrices de dissimilarité pour obtenir une partition finale consensuelle. MRFCMdd est conçu pour fournir une partition et un prototype pour chaque cluster flou, ainsi que pour apprendre un poids de pertinence pour chaque matrice de dissimilarité en optimisant un critère de correspondance qui mesure l'adéquation entre les clusters et leurs prototypes. Ces poids de pertinence changent à chaque itération de l'algorithme et peuvent être identiques pour tous les clusters (poids globaux) ou différents d'un cluster à l'autre (poids locaux).

4.1.1 MRFCMdd avec poids de pertinence estimés localement

L'algorithme MRFCMdd-RWL (MRFCMdd with local relevance weight) est conçu pour fournir une partition floue et un prototype pour chaque cluster. Il apprend également des poids de pertinence pour les matrices de dissimilarité, qui changent à chaque itération et diffèrent d'un groupe à l'autre. Son objectif est de partitionner de manière floue par $\mathbf{U}$ un ensemble $\mathbf{E}$ en K clusters, et de calculer un vecteur de poids de pertinence (un pour chaque cluster) $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K)$, tels que $\boldsymbol{\lambda}_k \in \mathbb{R}^p$ et $\boldsymbol{\mathcal{D}} = \{\mathbf{D}_1,\ldots,\mathbf{D}_p\}$ l'ensemble des matrices de dissimilarité en entrée du clustering. La fonction objectif mesurant l'adéquation entre les clusters et leurs prototypes $\mathbf{G}$ est définie par :

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

$(\lambda_{kj})^s D_j(e_i,G_k)$ représente la correspondance entre un exemple $e_i \in \mathbf{E}$ et le prototype de 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. $J_{MRFCMdd-RWL}$ est optimisé sous les contraintes suivantes :

$$(a): \quad s=1,\ \lambda_{kj}>0,\ \text{et}\ \prod_{j=1}^{p}\lambda_{kj}=1, \quad \text{ou}\ (b): \quad s\geq 1,\ \lambda_{kj}\in[0,1],\ \text{et}\ \sum_{j=1}^{p}\lambda_{kj}=1.$$

Ainsi, les auteurs proposent deux versions de l'algorithme, dont MRFCMdd-RWL-P pour l'optimisation sous (a) et MRFCMdd-RWL-S pour l'optimisation sous (b). L'algorithme suit les mêmes étapes que l'algorithme FCMdd en ajoutant une étapes de calcul des poids de pertinence.

4.1.2 MRFCM avec poids de pertinence estimés globalement

L'algorithme MRFCMdd-RWL peut rencontrer une instabilité numérique lorsqu'il génère des clusters uniques ou des clusters contenant des objets ayant une dissimilarité nulle entre eux, donnant $\sum_{i=1}^{n}(u_{ik})^m D_j(e_i, G_k) \to 0$. Pour pallier cette limitation, l'algorithme MRFCMdd-RWG (MRFCMdd with global relevance weight) est conçu pour fournir une partition floue et un prototype pour chaque cluster tout en apprenant des poids de pertinence pour l'ensemble des clusters dans chaque matrice de dissimilarité. Il minimise la fonction objectif suivant :

$$J_{MRFCM-RWG}(\mathbf{U}, \mathbf{G}, \mathbf{\Lambda}) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \sum_{j=1}^{p}(\lambda_j)^s D_j(e_i, G_k). \tag{3}$$

avec $(\lambda_j)^s D_j(e_i, G_k)$ représentant ici la correspondance entre un exemple $e_i \in \mathbf{E}$ et le prototype de 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. Cette minimisation prend en compte les contraintes (c) et (d) ci-dessous pour la mise à jour des poids $\boldsymbol{\lambda}$, ce qui donne lieu à deux variantes dont MRFCMdd-RWG-P (pour l'optimisation sous (c)) et MRFCMdd-RWG-S (pour l'optimisation sous (d)).

$$(c): \quad s = 1,\ \lambda_j > 0,\ \text{et} \prod_{j=1}^{p}\lambda_j = 1, \quad \text{ou } (d): \quad s \geq 1,\ \lambda_j \in [0,1],\ \text{et} \sum_{j=1}^{p}\lambda_j = 1.$$

4.1.3 Évaluation du clustering

La détermination du nombre optimal de clusters et l'évaluation de la qualité des résultats sont des étapes cruciales dans le clustering. Parmi les multiples critères d'évaluation du clustering flou existant (Wang et al. (2022)), dans ce travail, nous utilisons un indice de validité interne basé uniquement sur les degrés de partitionnement flou des individus : l'Entropie de Partition (PE). L'objectif est de minimiser PE comprise entre $[0, log_b(K)]$. L'indice atteint la borne supérieure lorsque l'incertitude est total. En revanche, la valeur minimale de 0 est obtenu lorsque la partition est dure.

$$PE = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik} log_b(u_{ik}). \tag{4}$$

5 Application et resultats

Nous avons expérimenté notre approche sur les jeux de données du projet eDOL, décrits dans la section 2. Les variables constituant les trajectoires ont toutes le même alphabet $\Sigma = [0-10]$, c'est à dire toutes les valeurs (statuts) 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, 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 (Yujian et Bo (2007)). 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. Les algorithmes de clustering multiple fuzzy c-medoïds sont implémentés à l'aide de R version 3.6.3. Nous avons considéré le nombre optimal de clusters utilisé dans l'article suivant (Soubeiga et al. (2023)), qui a travaillé sur les mêmes données, à savoir 3 clusters. Le critère de convergence et le nombre d'itérations maximal ont été respectivement fixés à 1e-9 et 1e+6. Nous avons ajusté l'hyperparamètre fuzziness en testant différentes valeurs et en sélectionnant celle qui minimise la mesure de l'entropie de partition, pour laquelle la valeur est m=1.5. L'algorithme MRFCMdd-RWL-S donne la meilleure performance de clustering pour l'identification des typologies de trajectoires de soins, avec une valeur du critère PE à 0,013 (voir la table 1). Par conséquent, les interprétations suivantes sont liées à la partition générée par l'algorithme MRFCMdd-RWL-S.

Algo.	MRFCMdd	MRFCMdd-RWL-P	MRFCMdd-RWL-S	MRFCMdd-RWG-P	MRFCMdd-RWG-S
PE	0.170	0.082	0.013	0.129	0.078

TAB. 1 – *Valeurs de performance des algorithmes de clustering flous basés sur multiple matrices de dissimilarités*

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. Les valeurs représentent l'importance relative de chaque dimension dans la caractérisation de chaque cluster. Le Cluster 1 présente des valeurs plus élevées pour les dimensions de Fatigue, Douleur, Stress, Moral et Sommeil par rapport aux autres dimensions. Cela pourrait indiquer que ce cluster est associé à des trajectoires avec une forte incidence de ces dimensions, suggérant des profils de douleur chronique caractérisés par une fatigue, un stress et une douleur plus élevés, associés à des aspects émotionnels (Moral) et des perturbations du sommeil. Le Cluster 2 met en évidence des valeurs relativement plus équilibrées pour toutes les dimensions, mais avec une prédominance légère pour le Confort corporel. Le Cluster 3 affiche des valeurs plus élevées pour les dimensions de Douleur, Stress, et Moral, avec des valeurs relativement plus faibles pour les autres dimensions. Ce cluster pourrait caractériser des trajectoires où la douleur, le stress et les aspects émotionnels jouent un rôle plus prédominant par rapport aux autres dimensions. Même si les clusters 1 et 2 sont caractérisés tous les deux par la Douleur, le Stress et le Moral, on peut observer dans la Figure 2 que le temps moyens passe dans chaque statut de ces dimensions sont plus important dans le cluster 1 que dans le cluster 3. Par exemple, les patients appartenant au cluster 1 ont en moyenne vécu une période de trois semaines avec un niveau de stress évalué à 2, tandis que pour les individus du cluster 2, cette période était en moyenne d'une semaine.

	Douleur	Stress	Fatigue	Sommeil	Moral	Confort corporel	Activité non-sportive	Activité sportive
Cluster 1	0.144	0.143	0.161	0.141	0.144	0.104	0.103	0.059
Cluster 2	0.141	0.132	0.133	0.116	0.125	0.134	0.111	0.109
Cluster 3	0.139	0.144	0.144	0.124	0.143	0.084	0.116	0.107

TAB. 2 – *Valeurs des poids de pertinence de chaque dimension des trajectoires pour chaque clusters*

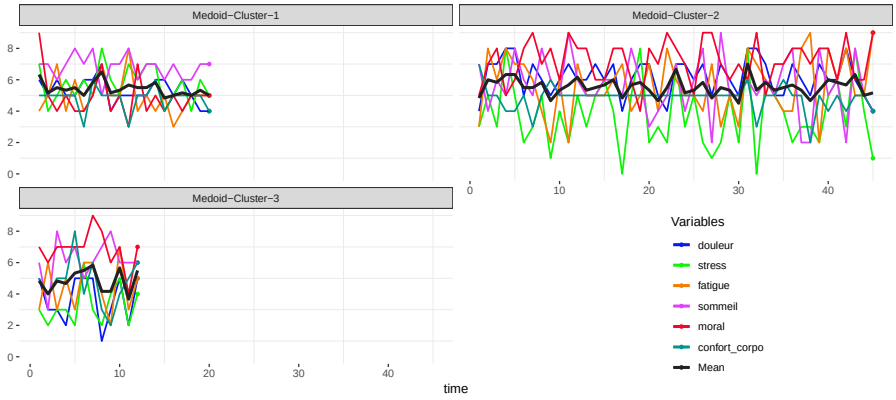

FIG. 2 – *Trajectoire des médoïde de chaque cluster*

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 sources 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. D'autre part, notre approche permet une collaboration entre les dimensions pour fournir une partition floue avec des poids de pertinence pour chaque dimension dans le partitionnement. Les résultats montrent qu'elle est plus simple à mettre en œuvre et pourrait traiter des dimensions plus grandes que les approches d'analyse séquentielle multicanal présentes dans la littérature (Robette (2021)), dont les objectifs visent la réduction de dimension en une seule matrice de similarité pour le clustering, en combinant les statuts, les coûts ou encore les distances elles-mêmes. 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, le choix de l'indice de qualité des clusters a influencé la sélection de la partition finale. Il est possible que d'autres indices, n'étant pas exclusivement basés sur les degrés de partitionne-

ment flou, puissent être plus adaptés. Ensuite, nous avons opté pour le clustering relationnel basé sur les médoïdes, bien que d'autres méthodes de clustering relationnel souple peuvent être étudier. Enfin, 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 basé sur plusieurs matrices de dissimilarité à d'autres techniques de clustering souple pouvant traiter l'incertitude. Nous explorerons aussi d'autres mesures de qualité , telles que l'indice de silhouette basé sur plusieurs matrices et prenant en compte les poids des matrices.

Remerciements

Les auteurs remercient l'Agence nationale de la recherche française pour son soutien dans le cadre du financement CAP 20-25.

Références

Bouveyron, C. et C. Brunet-Saumard (2014). Model-based clustering of high-dimensional data : A review. *Computational Statistics & Data Analysis 71*, 52–78.

de A.T. de Carvalho, F., Y. Lechevallier, et F. M. de Melo (2013). Relational partitioning fuzzy clustering algorithms based on multiple dissimilarity matrices. *Fuzzy Sets and Systems 215*, 1–28. Theme : Clustering.

Fokianos, K. et B. Kedem (2003). Regression theory for categorical time series. *Statistical science 18*(3), 357–376.

Giele, J. et G. Elder (1998). *Methods of life course research : qualitative and quantitative approaches*. SAGE Publications, Inc.

Halpin, B. (2014). Three narratives of sequence analysis. In *Advances in sequence analysis : Theory, method, applications*, pp. 75–103. Springer.

Kerckhove, N. et al. (2022). eDOL mhealth app and web platform for self-monitoring and medical follow-up of patients with chronic pain : Observational feasibility study. *JMIR Form Res 6*(3), e30052.

Krishnapuram, R., A. Joshi, et L. Yi (1999). A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315)*, Volume 3, pp. 1281–1286. IEEE.

Krogh, A., M. Brown, I. S. Mian, K. Sjölander, et D. Haussler (1994). Hidden markov models in computational biology : Applications to protein modeling. *Journal of molecular biology 235*(5), 1501–1531.

López-Oriona, Á. et J. A. Vilar (2023). Ordinal time series analysis with the r package otsfeatures. *Mathematics 11*(11), 2565.

Marteau, P. F. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 306–318.

Nguefack, H. L. N., M. G. Pagé, J. Katz, M. Choinière, A. Vanasse, M. Dorais, O. M. Samb, et A. Lacasse (2020). Trajectory modelling techniques useful to epidemiological research : A comparative narrative review of approaches. *Clinical Epidemiology 12*, 1205–1222.

Richter, C., M. Luboschik, M. Röhlig, et H. Schumann (2015). Sequencing of categorical time series. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 213–214. IEEE.

Robette, N. (2021). *L'analyse statistique des trajectoires : Typologies de séquences et autres approches*. Méthodes et savoirs. Ined Éditions.

Soubeiga, A., J. Ettaghouti, V. Antoine, A. Corteval, N. Kerckhove, et S. Moreno (2023). Classification automatique de séries chronologiques de patients souffrant de douleurs chroniques. *Revue des Nouvelles Technologies de l'Information Extraction et Gestion des Connaissances, RNTI-E-39*, 651–652.

Stoffer, D. S., D. E. Tyler, et A. J. McDougall (1993). Spectral analysis for categorical time series : Scaling and the spectral envelope. *Biometrika 80*(3), 611–622.

Studer, M. et G. Ritschard (2015). 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), 481–511.

Wang, H.-Y., J.-S. Wang, et G. Wang (2022). A survey of fuzzy clustering validity evaluation methods. *Information Sciences 618*, 270–297.

Weiss, C. H. et P. K. Pollett (2014). Binomial autoregressive processes with density-dependent thinning. *Journal of Time Series Analysis 35*(2), 115–132.

Yujian, L. et L. Bo (2007). A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*(6), 1091–1095.

Summary

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 fuzzy relational clustering based on multiple distance matrices. We propose a distance-based clustering approach that integrates the extraction of discrete treatment sequences, the computation of dissimilarity relations between pairs of sequences and soft collaborative relational clustering. This method was applied to real data from a daily follow-up study of chronic pain patients in France. 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.