Classification de trajectoires d'observance de patients atteints d'un syndrome d'apnées obstructives du sommeil

Guillaume Bottaz-Bosson

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Classification de trajectoires d'observance de patients atteints d'un syndrome d'apnées obstructives du sommeil

Clustering of compliance trajectories for subjects with obstructive sleep apnea

Présentée par :

Guillaume BOTTAZ-BOSSON

Direction de thèse :

Adeline LECLERCQ-SAMSON PROFESSEURE DES UNIVERSITES, Université Grenoble Alpes	Directrice de thèse
Sébastien BAILLY CHARGE DE RECHERCHE, Université Grenoble Alpes	Co-directeur de thèse

Rapporteurs :

ANNE GEGOUT-PETIT
Professeur des Universités, UNIVERSITE DE LORRAINE
CAROLE PLANES
Professeur des Univ. - Praticien hosp., UNIVERSITE SORBONNE PARIS NORD

Thèse soutenue publiquement le **14 novembre 2022**, devant le jury composé de :

SEBASTIEN BAILLY Chargé de recherche HDR, INSERM DELEGATION AUVERGNE-RHONE-ALPES	Co-directeur de thèse
JEAN-FRANÇOIS COEURJOLLY Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Président
SYLVIE DEUFFIC-BURBAN Chargé de recherche HDR, INSERM PARIS ILE- DE-FRANCE CENTRE NORD	Examinatrice
ANNE GEGOUT-PETIT Professeur des Universités, UNIVERSITE DE LORRAINE	Rapporteure
ADELINE LECLERCQ-SAMSON Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Directrice de thèse
CAROLE PLANES Professeur des Univ. - Praticien hosp., UNIVERSITE SORBONNE PARIS NORD	Rapporteure
FABIEN SUBTIL Maître de conférences - Praticien hosp., UNIVERSITE LYON 1 - CLAUDE BERNARD	Examinateur

THÈSE

pour obtenir le grade de Docteur de

L' UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées

présentée par

Guillaume BOTTAZ-BOSSON

Classification de trajectoires d'observance de patients atteints d'un syndrome d'apnées obstructives du sommeil

Thèse codirigée par **Adeline LECLERCQ-SAMSON** et **Sébastien BAILLY**
préparée au sein des laboratoires **Jean Kuntzmann (LJK)**
et **Hypoxie et Physiopathologies cardiovasculaires et respiratoires (HP2)**
dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

soutenue publiquement le 14 Novembre 2022

Composition du Jury :

M. Sébastien BAILLY, *Co-directeur de thèse*
Chargé de recherche, INSERM

M. Jean-François COEURJOLLY, *Président*
Professeur des universités, Université Grenoble Alpes

Mme Sylvie DEUFFIC-BURBAN, *Examinatrice*
Chargé de recherche, INSERM

Mme Anne GÉGOUT-PETIT, *Rapporteure*
Professeur des universités, Université de Lorraine

Mme Adeline LECLERCQ-SAMSON, *Directrice de thèse*
Professeur des universités, Université Grenoble Alpes

Mme Carole PLANES, *Rapporteure*
Professeur des universités - Praticien hospitalier, Université Sorbonne Paris Nord

M. Fabien SUBTIL, *Examinateur*
Maître de conférences des universités - Praticien hospitalier, Université Claude Bernard Lyon 1

Remerciements

En premier lieu, je souhaite vivement remercier les membres de mon jury pour l'intérêt que vous avez porté à mes travaux de thèse, et pour les discussions intéressantes que vous avez initiées lors de ma soutenance.

Merci à Anne Gégout-Petit et Carole Planès d'avoir accepté de rapporter le présent mémoire, et pour vos commentaires enthousiastes.

Ensuite je remercie mes examinatrice et examinateurs pour l'honneur que vous m'avez fait de participer à mon jury de thèse : Jean-François Coeurjolly en sa qualité de président, Sylvie Deuffic-Burban et Fabien Subtil.

Je remercie également ce dernier pour avoir participé à mes comités de suivi individuel. Vos suggestions, tant scientifiques qu'organisationnelles, m'ont été précieuses pour mener au bout mon projet de thèse.

Je tiens à adresser mes plus sincères remerciements à ma directrice et à mon co-directeur de thèse, Adeline Samson et Sébastien Bailly. Votre complémentarité, savante et humaine, m'ont grandement poussé à puiser le meilleur de moi-même et à l'exploiter sur les deux fronts clinique/médical et statistique/méthodologique.

Adeline, tu as su faire preuve d'une grande exigence scientifique concernant mes travaux tout en te montrant bienveillante. Cela a été très enrichissant de bénéficier de ce subtil et rare équilibre entre ces deux qualités. Je te remercie également pour ta relecture attentive de chacun de mes chapitres.

Sébastien, outre le fait de m'avoir accordé un financement, tu as su te montrer réactif lorsque nécessaire. Je te remercie pour ta disponibilité à toute épreuve et ton honnêteté. L'énergie que tu as déployée a je pense joué un rôle moteur dans l'avancée de mes travaux de publication.

Merci pour ce que vous m'avez transmis, merci pour la confiance que vous avez placé en moi, et pour les responsabilités que vous m'avez laissé prendre par ma participation dans le choix des développements, à commencer par la construction du sujet de recherche à la suite de mon stage de M2.

Je souhaite ensuite remercier Agnès Hamon, qui a co-encadré ce stage préliminaire à ma thèse. Ton intérêt et ton implication dans les travaux qui sont présentés dans ce mémoire m'ont grandement permis d'en améliorer la qualité. Merci pour ton sens de l'intégrité et pour tout le temps que tu m'as consacré.

Je remercie également le Professeur Jean-Louis Pépin pour votre accueil au sein du laboratoire HP2, et pour m'avoir donné l'opportunité de travailler sur ce sujet porteur d'enjeux sociétaux, à la croisée entre recherche médicale et sciences des données.

Je remercie chaleureusement les personnes qui ont contribué à mes productions ainsi qu'au développements de mes connaissances :

Aux étudiants stagiaires avec qui j'ai eu l'occasion de travailler, particulièrement Théo S. et David F. pour vos contributions majeures sur l'application de visualisation.

À Alison Foote pour la relecture et l'optimisation de mes écrits en anglais.

À Alphanie pour nos échanges qualitatifs et nos travaux communs sur les données que nous avons analysés dans le cadre de nos thèses respectives.

À Gérard Grégoire et à Vincent Brault pour m'avoir partagé vos expertises statistiques sur la thématique du clustering et à Rémy Drouilhet pour la programmation R.

Aux personnels d'AGIR à dom. pour la fourniture des données de télésuivi et l'aide à leur interprétation. Particulièrement à l'endroit de Najeh, je garde un très bon souvenir de notre déplacement à Louvain pour le congrès de l'ISCB.

Au Professeur Renaud Tamisier pour votre intérêt à propos de mon sujet de thèse. Merci pour l'apport de vos connaissances sur le syndrome d'apnées du sommeil et de votre expérience dans l'accompagnement des patients traités par pression positive continue.

Je pense aussi aux personnels du laboratoire LJK ayant facilité les démarches administratives et l'accès aux ressources informatiques de l'université : Franck, Laurence, Frederic, Suzanne, Juana, Cathy, Emmanuel, Bruno, ainsi qu'à Audrey pour les aspects administratifs du côté HP2, et Clément pour la mise en place des outils de communication au niveau de l'équipe data, surtout dans les situations de confinements stricts.

Enfin, à Rosette et Éva pour la relecture de mon mémoire et de mon diaporama de soutenance. Et à mon "assistance", qui se reconnaitra, pour la correction des compte rendus de TPs des étudiants de licence.

Je remercie aussi les personnes que j'ai fréquemment côtoyées pendant ma thèse et que je n'ai pas déjà remerciées :

Aux membres du laboratoire LJK (et assimilés), et plus particulièrement aux permanents Vincent (oui, encore toi), Julien, Clovis, Jean-Charles, Annick, qui ont créé de bonnes conditions pour l'intégration des doctorants (Mario kart, loup garous, restaurants, running, ekiden, ...). Ensuite les étudiants, mes co-bureaux qui se sont succédés : Yassine et Simon c'était super de partager cette aventure avec vous depuis le début, et jusqu'au bout de l'ATER. Puis la relève : Ming Ming, Omayma, Weiss, Victor, Ieva ! Je peux partir l'esprit tranquille, le bureau 143 est passé entre de bonnes mains. Également les "jeunes" hors 143, la liste est longue... Je pense notamment à Chloé, pour nos footing à la bastille, dégustations de thés, et pour m'avoir donné la main sur nombre des déménagements encourus ces dernières années. Merci pour le soutien sincère que tu m'as apporté. À Modibo, je te remercie particulièrement pour ton assistance à l'utilisation des serveurs de calcul distants. À Anya, je garde aussi un bon souvenir de nos séjours communs à Aussois et aux JDS de Nancy. Je remercie également entre autres Jean-Baptiste, Léa, Philomène, Maria Belen, Margaux, Carlos, Mohamed, Nils, Hubert ou Flora pour les moments conviviaux à la Cafèt'.

Aux membres du laboratoire HP2, dont Meriem et Fabien, nous formions un chouette quatuor avec Najeh dans le bureau des « stats » lorsque j'ai intégré HP2. Puis les autres "jeunes" du labo : Rawaa, Jérémy, François, Guillemette, Sébastien.

Au groupe des « randonneurs indécis » pour entre autres nos sorties montagnardes et culturelles : merci Ariane mais aussi Zoé, Anna, Louise et Dany.

Je remercie les personnes que je connaissais avant d'entreprendre la thèse :

Ma famille de cœur, pour m'avoir tant apporté et avoir facilité mes possibilités de faire des études. Je vous suis reconnaissant d'avoir pu commencer ce doctorat. Merci également d'avoir été compréhensifs sur le peu de temps que j'avais pour venir vous voir.

Mes amis d'avant thèse qui se sont également montré compréhensifs, qui m'ont soutenu, qui ont été présents auprès de moi lorsque j'en avais besoin, et qui le sont toujours. J'espère être à la hauteur de ce que vous m'apportez.

Enfin, j'adresse mes tous derniers remerciements à toi, ma chère Ines. Quelle bonheur pour moi de t'avoir rencontrée en cours de thèse. Tu m'as tant apporté sur la fin de cette aventure qu'une partie du mérite te revient. Tu me rends heureux, tout simplement, et j'espère te le rendre.

Résumé

Le syndrome d'apnées obstructives du sommeil (SAOS) est une pathologie à forte prévalence en France. Le SAOS est associé à des comorbidités cardiovasculaires comme l'hypertension artérielle ou l'accident vasculaire cérébral. Le traitement de référence dans les cas les plus sévères est la ventilation en pression positive continue (PPC). La thérapie rencontre des problèmes d'observance et d'abandons, et certains effets potentiels sont à ce jour controversés. Les cliniciens ont besoin de comprendre et caractériser les comportements d'observance. Le télésuivi de la PPC rend disponibles les durées de traitement quotidiennes des patients.

Le travail de recherche présenté dans ce mémoire vise à identifier des comportements typiques d'observance à partir des données de télésuivi. Nous utilisons une approche de classification non supervisée de séries chronologiques.

Premièrement il convient de délimiter les séquences individuelles de données quotidiennes d'observance à partir des données du télésuivi. Les données sont produites par les appareils de PPC et peuvent transiter par les fabricants des machines et par les prestataires de soins à domicile (PSAD). Elles sont sujettes à contenir des biais lorsqu'elles parviennent aux cliniciens. Nous proposons des recommandations pour la production de séries chronologiques exhaustives et fidèles aux comportements d'observance des patients. Puis nous discutons de la pertinence et de la représentativité des données fournies par un PSAD pour prévenir de potentielles mauvaises utilisations.

Ensuite nous réalisons la classification non supervisée des séries temporelles d'observance. Elles sont caractérisées par des variabilités et des discontinuités dues aux jours de non utilisation de la PPC. Nous souhaitons que les clusters tiennent compte des niveaux d'observance, des variabilités et discontinuités des trajectoires, et que des individus abandonnant la thérapie soient réunis peu importe la date d'abandon. Nous mettons en avant l'intérêt de la dissimilarité du dynamic time warping (DTW) pour classer les trajectoires dans le contexte clinique. En particulier dans le cadre de la classification ascendante hiérarchique, le critère de Ward et l'indice de Dunn fournissent des résultats pertinents.

Enfin, il faut permettre aux cliniciens d'interpréter les clusters. La superposition et la juxtaposition des courbes d'un cluster sont inexploitables à cause du nombre de trajectoires et de leurs variabilités et discontinuités. Nous suggérons de diversifier les représentations graphiques afin de mettre en évidence les composantes comportementales communes d'un cluster ainsi que l'hétérogénéité des comportements. Les graphiques que nous proposons sont illustrés avec des clusters de données réelles. Une application web est en cours de développement pour permettre aux cliniciens de réaliser ces graphiques et de les personnaliser.

Mots-clés : Classification non supervisée de séries chronologiques, Dynamic time warping, Qualité des données, Visualisation de données, PPC, SAOS.

Abstract

Obstructive sleep apnea (OSA) is a highly prevalent disease in France. OSA is associated with cardiovascular comorbidities such as high blood pressure or stroke. Positive airway pressure (PAP) is the first line treatment of moderate to severe OSA. The therapy faces compliance issues and patient withdrawals, and some potential effects are still controversial. Clinicians need to understand and characterize compliance behaviors. PAP telemonitoring provides data regarding individual daily use.

The research work presented in this thesis aims to identify typical compliance behaviors from remote monitoring data. We employ a time series clustering approach.

Firstly, individual sequences of daily PAP usage duration must be built from telemonitoring data. Data are produced by PAP devices and can pass through devices manufacturers and homecare providers. Data may contain bias when they are made available to clinicians. We suggest recommendations for building complete and reliable individual time series. Then we discuss the relevance and representativeness of the data provided by a homecare provider to prevent potential misuses.

Next, we cluster the compliance time series characterized by high variabilities and discontinuities due to non uses of PAPs. We want the clusters to take into account of the level of compliance, variability and discontinuities within the trajectories, and to gather patients who dropped out of treatment regardless of the date they stopped. We highlight the benefits of the dynamic time warping dissimilarity to cluster these data in the clinical context. Particularly, in hierarchical ascending clustering, Ward's criterion and Dunn index provide relevant clusters.

Finally, clinicians must be able to interpret the clusters. The superposition or the juxtaposition of all the curves of a cluster is not informative because of the number of trajectories, their variability and discontinuities. We suggest diversifying the visualizations of the clusters to highlight both the common characteristics shared by the trajectories of a cluster and the heterogeneity of behaviors. The charts we propose are illustrated by real life clusters. A web application is under development to allow clinicians to create and customize these charts.

Key-words : Time series clustering, Dynamic time warping, Data quality, Data visualization, PAP, OSA.

Table des abréviations et acronymes

(r)AHI	:	*(residual) apnea hypopnea index*
ARI	:	*adjusted rand index*
BSS	:	*between-clusters sum of squares* (somme des carrés inter-clusters)
CAH	:	classification ascendante hiérarchique
CVI.s	:	*cluster validity (validation) index.indices*
DBA	:	*dynamic time warping barycenter averaging*
DTW	:	*dynamic time warping*
HAC	:	*hierachical agglomerative clustering*
IAH	:	index d'apnées hypopnées
iid	:	indépendantes et identiquement distribuées
MMG	:	modèle de mélange fini gaussien
OSA	:	*obstructive sleep apnea*
(A/C)PAP	:	*(automatic/continuous) positive airway pressure*
PPC	:	pression positive continue
PSAD	:	prestataire de soins à domicile
SAOS	:	syndrome d'apnées obstructive du sommeil
sdF	:	*generalized summed discrete Fréchet dissimilarity* (dissimilarité de Fréchet discrète sommée généralisée)
SS	:	*sum of squares* (somme des carrés)
VAS	:	voies aériennes supérieures
WSS	:	*within-clusters sum of squares* (somme des carrés intra-clusters)

Table des matières

Chapitre 1

Introduction

1.1 Contexte et problématique clinique

1.1.1 Le SAOS et la thérapie par PPC

Le syndrome d'apnées obstructives du sommeil (SAOS) est une pathologie respiratoire chronique fréquente. Elle est caractérisée par des arrêts (apnées) et/ou des réductions significatives (hypopnées) involontaires et anormalement récurrents du flux d'air respiratoire au cours du sommeil, principalement causés par des obstructions des voies aériennes supérieures (VAS). Elles sont provoquées par des affaissements des tissus mous de la gorge, résultant d'anomalies d'ordre anatomiques ou fonctionelles. Les principaux facteurs de risque de la pathologie sont l'obésité, un âge avancé, et le sexe masculin [1]. La Figure 1.1 montre les prédispositions aux SAOS induites par l'obésité.

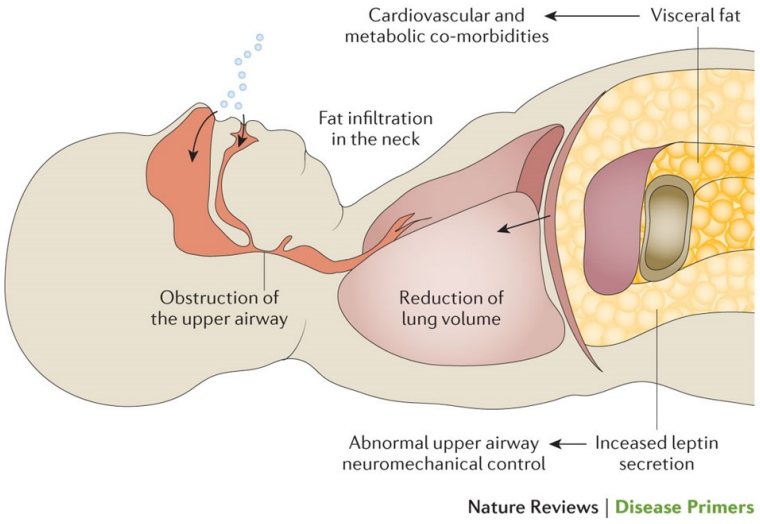


FIGURE 1.1 – Obésité et syndrome d'apnées obstructives du sommeil.
L'obésité prédispose au syndrome d'apnées obstructives du sommeil par l'infiltration de graisses dans le cou, conduisant à des affaissements des voies aériennes supérieures (VAS), et augmente la pression abdominale, réduisant le volume pulmonaire. L'accumulation de tissus adipeux pourrait aussi affecter le contrôle neuromécanique des VAS par les effets spécifiques de la leptine. La graisse viscérale favorise les comorbidités cardiométaboliques. Les "bulles" représentent les gaz inhalés et expirés.
Légende et figure tirées de [1], la figure est adaptée de Drager L, Togeiro S, Polotsky V, et al. Obstructive Sleep Apnea. J Am Coll Cardiol. 2013 Aug, 62 (7) 569–576, Elsevier.

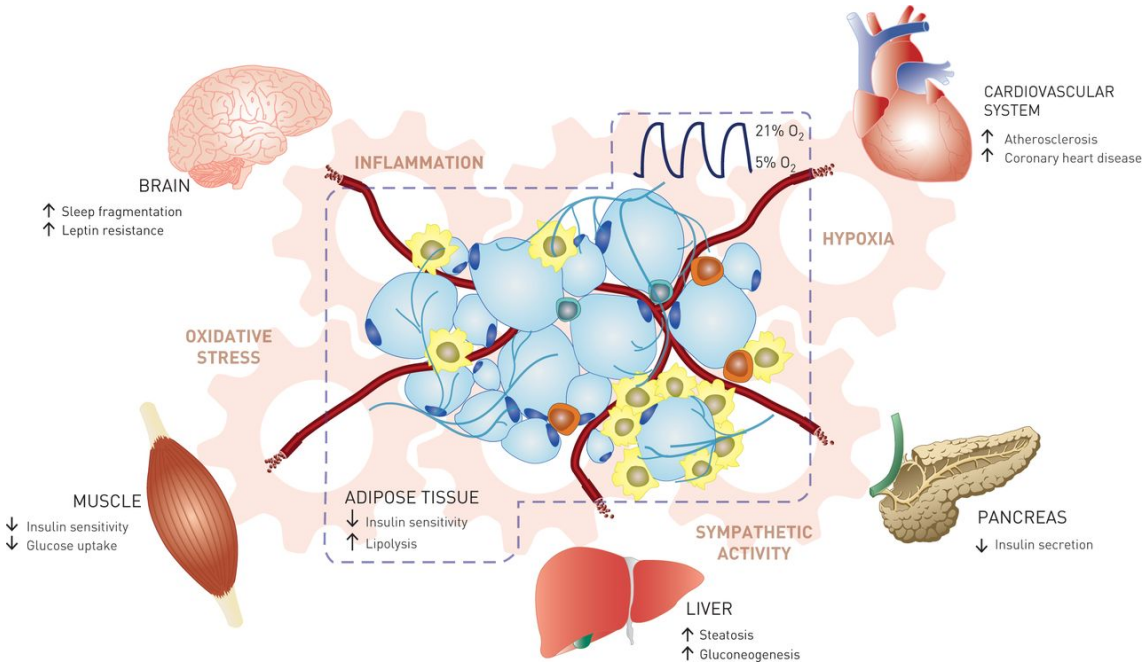


FIGURE 1.2 – Le tissu adipeux vu comme un acteur majeur dans les conséquences systémiques de l'hypoxie intermittente.
Source : [3].

Les conséquences directes du SAOS comprennent une augmentation de l'effort respiratoire avec des micro-éveils, des modifications des pressions intrathoraciques exerçant des contraintes mécaniques sur le cœur et les vaisseaux, et une exposition à l'hypoxie intermittente [1]. Il s'agit d'une alternance de situations où la disponibilité en oxygène dans l'organisme est réduite avant de retrouver un niveau normal. La qualité de vie est affectée notamment par des somnolences diurnes, une détérioration des capacités cognitives allant de difficultés de concentration à la depression et un risque accru d'accidents [1]. Une baisse de productivité ainsi qu'une majoration de l'absentéïsme professionels ont été montrés [2]. Le SAOS cause une perturbation systémique de l'organisme par de nombreux phénomènes interdépendants, illustrés en Figure 1.2 [3]. Cela induit une altération des fonctions métaboliques, hépatiques et cardiovasculaires. Le SAOS peut provoquer de l'hypertention artérielle et est associé à des comorbidités cardiométaboliques comme le diabète, l'insuffisance cardiaque où l'accident vasculaire cérébral [1, 4].

Le SAOS est une pathologie hétérogène où plusieurs phénotypes peuvent être distingués : 1) une multitude de facteurs de risques et de caractéristiques individuelles peuvent favoriser son apparition ; 2) il se manifeste de diverses façons lors du sommeil, à différents stades de sommeil et avec plusieurs niveaux de gravité, selon la fréquence des micro-éveils ainsi que l'ampleur et la fréquence des épisodes d'hypoxie ; 3) une variété de symptômes affecte les patients [1] ayant un SAOS ; et 4) ils ont des comorbidités différentes, entrainant des interactions spécifiques avec le SAOS et donc des conséquences à long terme différentes [5].

Le diagnostic et la gravité du SAOS sont évalués de plusieurs manières selon

1. Si nous utilisons principalement le genre masculin par soucis de simplification, "patient", "sujet", "individu", etc... doivent être compris au sens inclusif, et peuvent désigner toute patiente ou tout patient sans aucune distinction de genre.

les sociétés savantes [6, 7, 8]. La société de pneumologie de langue française tient compte, d'une part d'un examen clinique de la symptomatologie des sujets, et d'autre part de l'indice d'apnées hypopnées (IAH). L'IAH est un score objectif calculé selon la fréquence des évènements d'apnées et d'hypopnées observés lors d'un examen du sommeil [9]. Une apnée obstructive correspond à une réduction du flux respiratoire d'au moins 90% pendant plus de 10 secondes malgré la présence de mouvements respiratoires. Une hypopnée est définie par une réduction du flux respiratoire d'au moins 30% durant plus de 10 secondes, et accompagnée, soit d'une diminution de la teneur en oxygène du sang artériel (dé-saturation artérielle en oxygène) au delà de 3%, soit d'un micro-éveil. Au moins 5 évènements d'apnées ou d'hypopnées par heure de sommeil sont nécessaires pour diagnostiquer un SAOS, et le syndrome est considéré sévère lorsque l'IAH est supérieur à 30 évènements par heure.

Partant du constat que le SAOS est sous-diagnostiqué à travers le monde, Benjafield et al. [10] ont estimé au delà de 900 millions la prévalence mondiale d'individus ayant un IAH supérieur à 5 évènements par heure dans la tranche d'âge 30-69 ans. Plus de 400 millions d'entre eux auraient un IAH supérieur à 15 évènements par heure. Ces estimations tiennent compte de disparités dans les proportions d'individus touchés selon les pays ou régions du monde. La Figure 1.3 montre les 10 pays qui présenteraient les prévalences les plus élevées, dont la France avec environ 24 millions (respectivement 12 millions) d'individus estimés ayant un IAH supérieur à 5 (respectivement supérieur à 15) évènements par heure.

D'importants enjeux économiques gravitent autour du SAOS et de sa prise en charge. Le cabinet de conseil Frost & Sullivan a publié en 2016, sous le mandat de

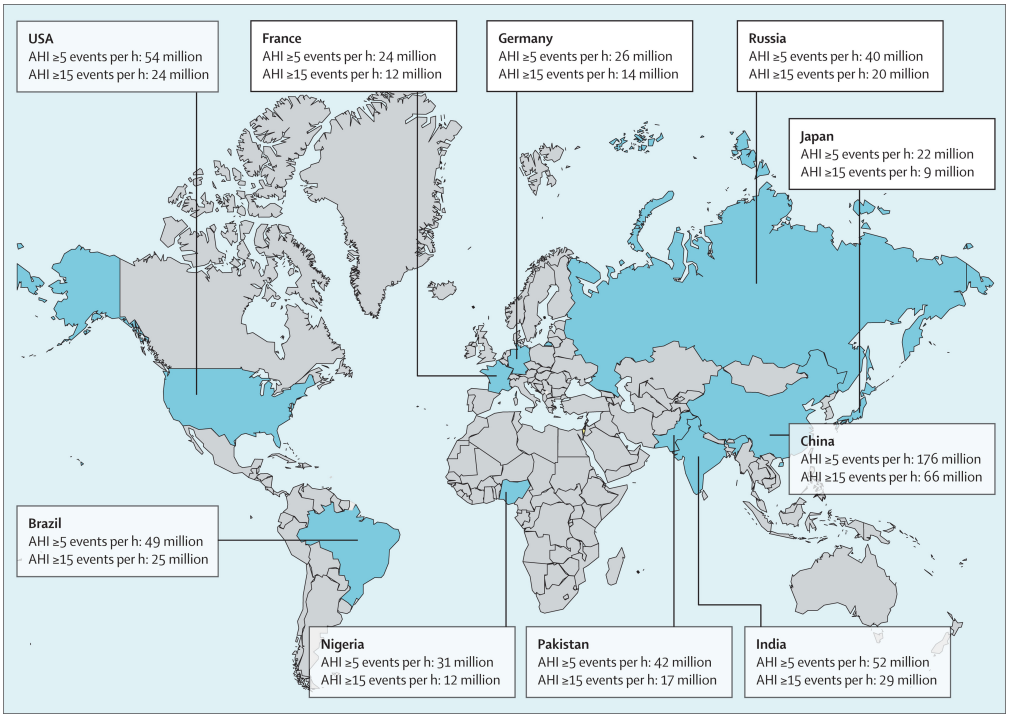


FIGURE 1.3 – Les 10 pays avec les plus grandes prévalences estimées du syndrome d'apnées obstructives du sommeil selon les critères 2012 de l'American Academy of Sleep Medicine [9].
AHI=apnoea-hypopnoea index, traduit par indice d'apnées hypopnées (IAH).
Source : [10].

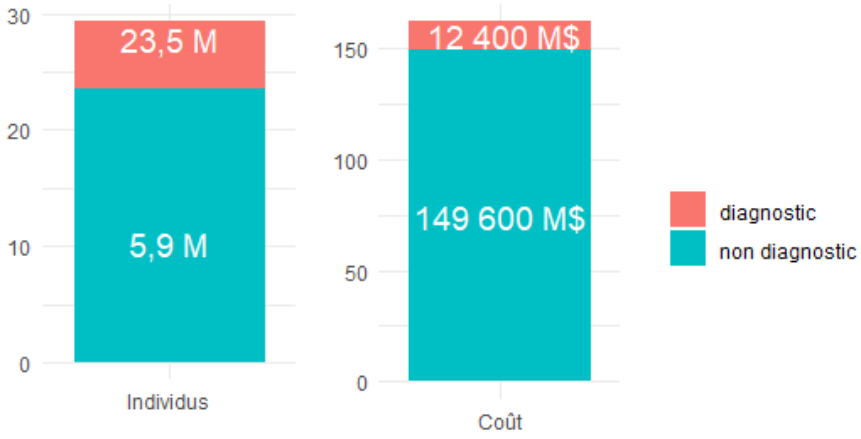
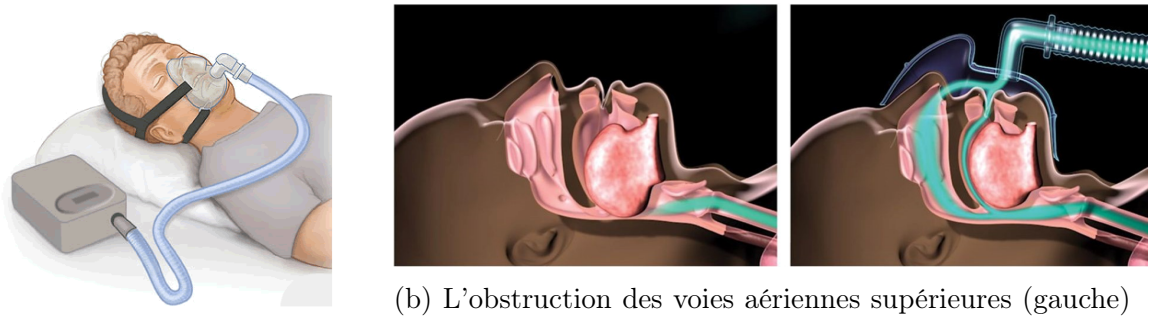


FIGURE 1.4 – Prévalences et coûts du diagnostic et du non diagnostic du SAOS aux États-Unis en 2015, d'après les estimations du cabinet de conseil Frost & Sullivan dans son rapport "Hidden Health Crisis Costing America Billions" (publié en 2016).

l'American Academy of Sleep Medicine (AASM), un rapport [2] présentant l'impact de l'évolution des diagnostics et prises en charge du SAOS sur le système de santé et le monde professionnel des États-Unis en 2015. Ce rapport mentionne d'un côté un coût de la prise en charge du SAOS estimé à 12,4 milliards de dollars américains pour 5,9 millions de patients diagnostiqués. Et d'un autre côté, un coût supporté du sous-diagnostic estimé à 149,6 milliards de dollars, pour 23,5 millions d'individus supposés non diagnostiqués (voir Figure 1.4). Les quatre postes de coûts considérés dans leur étude sont les coûts induits du fait des comorbidités et de l'impact sur la santé mentale (30 milliards de dollars), les coûts des accidents de véhicules motorisés (26,2 milliards), les coûts des accidents sur le lieu de travail non liés à l'utilisation de véhicules motorisés (6,5 milliards), et enfin les coûts liés à la perte de productivité (86,9 milliards). Ils recommandent de mieux diagnostiquer et traiter le SAOS. Wickwire et al. [11] ont publié en 2020 une étude médico-économique couvrant la période 2006-2013 et concernant les sujets de plus de 65 ans aux États-Unis. Ils estiment entre 13000 et 26000 dollars le surcoût annuel d'un SAOS non traité sur les 12 mois précédents le diagnostic ou la prescription d'un soin.

La ventilation en pression positive continue (PPC) est le traitement de première intention dans les cas modérés à sévères du SAOS. Elle concernait 1,2 millions de personnes en France début 2021, où la prescription médicale est accordée en fonction de l'IAH qui doit être supérieur à 15 évènements par heure, de la gravité des symptômes, du risque d'accidents et des comorbidités [12]. La thérapie consiste à normaliser le nombre d'événements respiratoires en maintenant les VAS ouvertes lors du sommeil par l'administration d'une pression d'air positive continue dans le pharynx. La figure 1.5 montre une représentation schématique de l'utilisation d'un appareil de PPC par l'intermédiaire d'un masque bucco-nasal, et de son action sur les obstructions des VAS.

2. En date du 12-06-2022, ce rapport intitulé "Hidden Health Crisis Costing America Billions" est accessible à l'url : `https://aasm.org/resources/pdf/sleep-apnea-economic-crisis.pdf`.



(a)

(b) L'obstruction des voies aériennes supérieures (gauche) est levée avec la pression d'air positive délivrée par la machine (droite).

FIGURE 1.5 – Représentations schématiques de l'utilisation d'un appareil de pression positive continue par l'intermédiaire d'un masque bucco-nasal (a) et de son effet sur les voies aériennes supérieures (b).
Sources : (a) `www.wikipedia.org`, (b) Cleveland Clinic. PAP therapy. `https://my.clevelandclinic.org/health/treatments/17320-pap-therapy` Accessed July 19, 2019.

Selon le degré d'utilisation du dispositif par les patients, la PPC atténue les symptômes diurnes, et améliore la qualité de vie [1], les fonctions neurocognitives [13] et la productivité au travail subjective [2]. Elle a également un impact bénéfique sur certains troubles cardiométaboliques, comme une diminution de l'hypertension artérielle [14, 15], ou une amélioration de la résistance à l'insuline [16]. L'étude observationelle de Marin et al. [17] va dans le sens d'un effet protecteur de la PPC sur l'incidence d'événements cardiovasculaires et sur la mortalité pour cause cardiovasculaire. En revanche l'étude randomisée SAVE [18] n'a pas confirmé cet effet. Cette étude présente cependant deux limites. Le temps de traitement quotidien moyen des sujets étudiés est seulement de 3,3 heures, et l'objectif principal de l'article compare les sujets traités avec et sans PPC, sans tenir compte du niveau effectif d'utilisation des machines par les patients. Pépin et al. [19] ont analysé les données du Système National des Données de Santé (SNDS) [3] des individus vivants en France de plus de 18 ans nouvellement traités par PPC entre 2015 et 2016. Les nombres de décès et d'évènements d'insuffisances cardiaques dans les trois ans après la première année de traitement sont significativement supérieurs chez les patients ayant arrêté la PPC (après exclusion des arrêts pour cause de décès ou de recours à un traitement alternatif) que chez ceux ayant continué. L'existence d'un effet bénéfique de la PPC reste alors à investiguer, et le cas échéant à préciser en tenant compte des phénotypes du SAOS, des causes de mortalité et du niveau d'observance des patients [19, 20, 21, 22].

Or, pour certains patients l'observance est souvent irrégulière et insuffisante [23]. De plus, l'acceptation du traitement ainsi que l'adhésion à long terme rencontrent des obstacles [24, 25]. Après exclusion des sujets décédés ou ayant eu recours à un traitement alternatif, 47,7% des prises en charge de la PPC initiées en France entre 2015 et 2016 ont été interrompues dans les trois premières années [12]. Ces problèmes d'acceptation, d'adhésion et d'observance à la PPC s'ajoutent à celui du sous-diagnostic du SAOS.

3. Le Système National des Données de Santé (SNDS) est un entrepôt de données médico-administratives pseudonymisées couvrant l'ensemble de la population française et contenant l'ensemble des soins présentés au remboursement (source au 12-06-2022 : `https://documentation-snds.health-data-hub.fr/`). Cette base de données permet notamment d'accéder aux données des décès et des prises en charge thérapeutiques par l'assurance maladie.

En France, où ce sont des prestataires de soins à domicile (PSAD) qui assurent la fourniture, le réglage et la maintenance des machines de PPC au domicile des patients, l'intérêt d'augmenter l'observance semble légitime à trois niveaux. Premièrement au niveau des patients pour améliorer la réponse au traitement. Deuxièmement au niveau de l'État, d'une part pour maintenir les bénéfices de l'investissement sur la prise en charge du SAOS, certains effets de la PPC étant réversibles dès la première nuit d'interruption [13, 26]. D'autre part pour atténuer les coûts de santé [11] et sociétaux [4] de la non prise en charge du SAOS. Enfin, troisièmement au niveau des PSAD pour obtenir un meilleur montant de remboursement de l'assurance maladie. Le remboursement est décliné selon plusieurs forfaits alloués en fonction l'observance quotidienne moyenne par périodes de 28 jours consécutifs [5]. Les PSAD doivent alors rendre compte pour chacun de leurs patients de l'utilisation des machines. Certaines peuvent enregistrer des données d'utilisation exploitables pour déterminer des indicateurs quotidiens et objectifs utiles au suivi de la thérapie : 1) la durée de port du masque lorsque la machine administre une pression thérapeutique, mesurant l'observance au traitement et que nous désignerons plus simplement par la durée de port du masque ; 2) les pressions et fuites d'air au niveau du masque permettant de contrôler l'usure des consommables et/ou les réglages de l'appareil ; et 3) l'IAH résiduel, qui est le nombre d'événements d'apnées ou d'hypopnées observés sous traitement et qui indique l'efficacité curative de la PPC. Depuis le 1er janvier 2018, les PSAD ont l'obligation légale de proposer aux patients le télésuivi de leur traitement [6]. Cela consiste à instaurer la transmission automatique de ces données au PSAD. Ce dernier les rend ensuite disponibles aux médecins et patients, et fournit les données d'observance à l'assurance maladie.

Dans la sous-section suivante nous présentons les opportunités médicales résultantes de la généralisation du télésuivi de la PPC.

1.1.2 Opportunités médicales offertes par la mise en place du télésuivi de la PPC en France

En France, la mise en place du télésuivi de la PPC d'un patient procure plusieurs avantages dans la prise en charge de son SAOS. Le PSAD peut déceler d'éventuelles pannes ou problèmes techniques entravant le bon déroulement du traitement dès leurs apparitions, et intervenir rapidement auprès du patient si besoin. Ensuite, médecin et patient peuvent facilement consulter jour après jour l'historique des indicateurs quotidiens. Le patient peut suivre objectivement son observance, évitant ainsi de sur-estimer son adhésion [27]. Enfin, selon l'observance, l'efficacité, la tolérance, les bénéfices ainsi que l'évolution de l'état de santé du patient, le médecin et le patient peuvent décider de la nécessité d'adapter le traitement, de l'interrompre

4. Voir le rapport "Hidden Health Crisis Costing America Billions" de Frost & Sullivan.

5. Source : Version au 09-05-2022 de la liste des produits et prestations remboursables (LPP) par l'assurance maladie. Document accessible en date du 12-06-2022 à l'url : `https://www.ameli.fr/professionnel-de-la-lpp/exercice-professionnel/facturation/liste-produits-prestations-lpp`.

6. Voir l'arrêté du 13-12-2017 du Journal officiel de la République française, modifiant la procédure d'inscription et les conditions de prise en charge du dispositif médical à pression positive continue pour traitement de l'apnée du sommeil et prestations associées au paragraphe 4 de la sous-section 2, section 1, chapitre 1er, titre Ier de la liste prévue à l'article L. 165-1 (LPPR) du code de la sécurité sociale.

ou de recourir à des thérapies alternatives.

Un nombre élevé de patients pourraient être concernés par le télésuivi : 1,2 millions d'individus sont actuellement traités par PPC en France [12], et jusqu'à 12 millions d'adultes seraient éligibles à la thérapie [10]. Le télésuivi, à l'échelle de l'ensemble des patients qui l'acceptent, facilite l'accès pour les cliniciens prescripteurs aux données générées quotidiennement du fait de l'utilisation des machines, dans un cadre légal. Elles couvrent l'ensemble des périodes de traitement après la mise en place du télésuivi, sont datées, et sont transmises automatiquement. D'une part, ces données peuvent permettre l'amélioration de la pratique médicale grâce à l'apport d'informations utilisables pour accompagner les patients au delà des consultations de suivi annuelles. D'autre part, elles peuvent s'agglomérer et être fusionnées avec les données de suivi médical. Cela ouvre la possibilité de réaliser des études cliniques sur le long terme et peut contribuer à une meilleure compréhension du SAOS.

Dans le cadre des travaux de cette thèse, nous avons souhaité favoriser l'exploitation des ces opportunités par les cliniciens. Notre apport consiste en la proposition de méthodes et d'outils permettant d'apporter des éléments de réponse à la problématique clinique présentée dans la sous-section suivante.

1.1.3 Problématique clinique motivant les travaux de la thèse

Après le diagnostic d'un SAOS, la prescription de la PPC pour un patient pose différents problèmes aux médecins. Premièrement, l'indication de la thérapie n'est à ce jour pas clairement établie en fonction des spécificités des SAOS pouvant affecter les patients. Par exemple le bénéfice reste controversé pour des sujets ayant un SAOS légér et fortement symptomatique [20]. Puis, au moment de la prescription, la PPC peut ne pas être acceptée par le patient. En cas d'acceptation il y a ensuite les risques d'observance insuffisante ou irrégulière qui pourraient amoindrir les bienfaits de la thérapie. Enfin il y a la possibilité que le patient abandonne le traitement à terme. Deux besoins des cliniciens peuvent alors être identifiés. D'une part, celui d'une meilleure connaissance des conséquences de la thérapie, notamment sur l'incidence d'évènements cardiovasculaires à long terme. Et d'autre part, celui de pouvoir accompagner les patients traités vers une "bonne" observance, le niveau d'observance optimal s'appréciant en fonction des effets de la PPC.

La littérature clinique comporte de nombreux travaux visant à délimiter les effets de la thérapie, ou à identifier des facteurs de risque d'une mauvaise observance. Deux critères sont fréquemment employés pour qualifier la bonne adhérence des patients. Le premier est d'avoir une observance quotidienne moyenne supérieure à 4 heures [17, 28]. Le deuxième critère est d'avoir une observance supérieure à 4 heures pendant au moins 70% des jours [24, 29]. La Figure 1.6 montre l'évolution de l'observance quotidienne pour 5 patients satisfaisant le deuxième critère pendant leurs 3 premiers mois de traitement. Ces 5 courbes témoignent de l'incapacité du critère à discriminer qualitativement les comportements d'observance. Le troisième patient ne devrait pas être jugé adhérent sur la période de traitement représentée. De plus, considérer indistinctement ces comportements d'observance peut être un frein à l'identification de certains effets de la thérapie. Une hypothèse raisonnable est que la thérapie est davantage bénéfique pour des patients assidus à un haut niveau d'observance comme le $5^{\text{ème}}$ par rapport à des patients assidus à un plus bas niveau comme le $4^{\text{ème}}$ patient, inutilisant quelques fois la machine comme le 1^{er} patient, qui

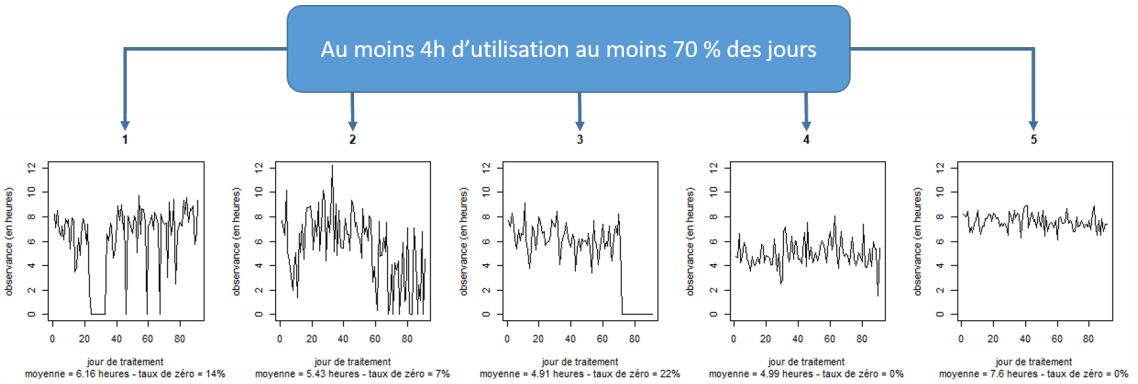


FIGURE 1.6 – Différents comportements d'observance qui sont ceux de sujets "adhérents" sur leurs trois premiers mois de traitement selon le critère "au moins 70% des jours avec une observance à la PPC supérieure à 4 heures". Chaque courbe représente l'évolution de l'observance quotidienne en heures. L'observance moyenne est calculée en incluant les jours d'inutilisation de la PPC, dont la proportion est donnée par le taux de zéro.

semblent décrocher comme le 2ème patient, voir abandonner (patient 3). Aussi on peut se demander s'il est préférable d'être régulièrement observant autour de 4h par jour comme le 4ème patient, ou être parfois inobservant mais avoir une plus haute observance les jours d'utilisation (patient 1).

Ainsi, pouvoir différencier les comportements d'observance en début de thérapie de manière plus fine en les considérant comme des "trajectoires" est crucial pour les cliniciens et chercheurs dans leur objectif d'optimiser la prise en charge du SAOS. Cela pourrait être une étape utile pour plusieurs questions cliniques. Premièrement pour approfondir l'étude de l'effet de la thérapie. Deuxièmement pour prolonger les travaux qui visent, soit à identifier des déterminants de l'adhésion, soit à prédire l'adhésion à terme en fonction de l'observance du début de traitement [30, 31, 32, 33]. Enfin, une approche complémentaire serait d'apprendre des motifs prédictifs d'une dégradation de l'observance et de les détecter dès leur apparition dans la trajectoire d'observance d'un patient au cours de son traitement. La finalité est de tendre vers une médecine de précision du SAOS qui consiste à soigner les patients de manière ciblée, selon leurs besoins, et en accompagnant prioritairement et au plus tôt ceux pour qui il y a un risque d'être insuffisamment observants ou d'abandon de la thérapie. Les differents phénotypes du SAOS devraient être intégrés à ces travaux. En effet ils pourraient être des déterminants de l'adhésion thérapeutique, et pourraient également impacter la réponse au traitement [5, 21].

Les cliniciens et chercheurs ont alors une forte motivation à s'intéresser aux données du télésuivi de la PPC et à investir des moyens pour se procurer des méthodes et des outils pour les intégrer dans leurs pratiques. Les travaux qui sont présentés dans cette thèse sont motivés par la volonté d'assister les cliniciens à résoudre la problématique clinique suivante :

"En intégrant la dimension temporelle des données issues du télésuivi de la PPC en France, identifier les comportements d'observance typiques des patients en début de thérapie."

Les critères communément utilisés par les cliniciens pour catégoriser les com-

portements d'observance sont inefficaces, motivant le recours à des méthodes statistiques.

1.2 Enjeux scientifiques abordés dans la thèse

Proposer une réponse à la problématique clinique soulevée dans la section précédente lève plusieurs enjeux scientifiques. Dans cette section nous présentons les enjeux que nous avons identifié lors de la production des données d'apprentissage, lors de l'analyse statistique de ces données, ainsi que lors de la restitution des résultats de cette analyse.

1.2.1 Production des données d'apprentissage

Outre leur important volume, les données générées par les machines de PPC présentent un intérêt par leur nature objective. En effet, elles ne sont exposées ni à des risques de mauvaise estimation par les patients, ni à des problématiques de mémorisation par ces derniers. Pour autant, cela ne dispense pas ces données d'être entachées de biais après leur fourniture par les PSAD. Certains biais peuvent être imputés au fait de considérer l'observance dans sa dimension longitudinale alors que d'autres biais sont plus génériques, comme dans toute situation impliquant une analyse statistique.

Les biais génériques sont de deux types. Le premier type de biais est lié à la sélection des données au regard de la question clinique pour laquelle nous cherchons à délimiter les comportements typiques d'observance. Nous dénombrons deux biais. Le premier vient de la pertinence de la variable choisie. Par exemple, la durée de port du masque est une variable pertinente mais incomplète si l'on s'intéresse à l'effet de la thérapie. Il faudrait aussi considérer l'IAH résiduel. Cette durée est en revanche moins pertinente pour l'étude des déterminants des comportements d'observance car des éléments externes (usure des consommables, nécessité de paramétrer la machine) peuvent réduire la durée de port du masque. Nous supposons que si une de ces situations se produisait fréquemment, le PSAD le détecterait et interviendrait sous quelques jours de sorte que le patient puisse être observant tel qu'il le désire. Ainsi nous considérons la durée de port du masque comme une variable pertinente pour étudier l'observance à la PPC, et nous nous limiterons à l'étude de cette variable. Le second biais dans la sélection des données peut être causé par la non-représentativité d'un échantillon, s'il ne couvre pas la population d'intérêt, ou si les données sont trop anciennes.

Le second type de biais générique est lié à la non-justesse de l'indicateur reporté chaque jour par rapport au phénomène mesuré. En particulier, certaines durées quotidiennes de port du masque peuvent être incohérentes ou erronées. Il peut y avoir plusieurs causes à cela comme une mauvaise estimation par les machines, des doublons, des problèmes d'agrégations inconsistantes, ou encore des retraitements incorrects de ces données.

Enfin, il reste les biais liés au caractère longitudinal de l'observance. Il ne s'agit pas d'extraire des observations ponctuelles indépendantes les unes des autres. L'unité statistique est un patient dont le comportement d'observance est descriptible par une séquence d'observations temporellement liées. Les biais se présentent lorsque les séquences définies ne reflètent pas les trajectoires d'observance effectives des

patients. Ils peuvent être causés par la non-justesse des indicateurs quotidiennement reportés, par la présence de données manquantes, ou par des incohérences dans les correspondances entre machines et patients dans les bases de données des PSAD.

Finalement plusieurs sources de biais peuvent dégrader la qualité des données, et il n'est pas forcément aisé de les identifier, ni de les dénouer. L'importance des questions cliniques sous-jacentes en termes de santé et économiques renforce la légitimité des préoccupations et de la rigueur à accorder sur la gestion de la qualité des données d'observance, même si cela doit être fait au détriment de leur quantité.

1.2.2 Analyse statistique

Considérant que les trajectoires individuelles d'observance des patients soient convenablement retranscrites dans un jeu de données, un objectif des travaux de cette thèse est d'en extraire les comportements typiques d'observance. L'observance au traitement par PPC est caractérisée par la régularité des patients à entreprendre chaque jour l'action de se traiter, avec des nuances selon les différentes durées d'utilisation quotidiennes. Il en découle une potentielle complexité de l'évolution des comportements individuels d'observance, avec des dynamiques pouvant plus ou moins progressivement varier au cours du temps, en des instants aléatoires. Cela peut être source d'une diversité des comportements d'observance entre différentes périodes pour un même individu, ou entre différents patients. Les cinq exemples de la Figure 1.6 (Section 1.1.3) illustrent ces diversités et l'inefficacité des indicateurs statistiques classiques utilisés avec des seuils. Outre le constat que plus de deux catégories semblent nécessaires, deux défauts sont imputables à ces seuils : 1) ils sont déterminés à priori, sans tenir compte des données ; et 2) ils négligent le caractère séquentiel de l'observance, et donc ne peuvent intégrer son caractère évolutif. Additionnellement à ces défauts, le choix même des indicateurs pour distinguer les comportements peut être remis en cause. Par exemple, il n'est pas forcément pertinent de différencier des individus, qui sur trois mois de suivi, ont eu une utilisation semblable de leur machine sauf à une journée près. De plus, certaines séquences sont caractérisées par de fortes variabilités et de nombreuses discontinuités liées aux non utilisations des dispositifs. À notre sens, il est primordial de considérer ces variabilités comme des composantes intégrantes des comportements individuels. Pour cela il est par exemple proscrit de considérer les séquences seulement à travers des valeurs moyennes.

Un des défis de l'analyse statistique de ces données est de tirer le plus pleinement parti de leur richesse pour distinguer les comportements individuels typiques d'observance. Une manière de procéder est de réaliser des groupes de séquences par comportements similaires afin que des séquences décrivant des comportements très différents soient dans des groupes distincts. Ainsi les caractéristiques qui émergeraient très majoritairement dans les comportements individuels de chacun des groupes pourraient décrire les comportements typiques d'observance. Aloia et al. [34] ont considéré les séquences d'une année d'observance de 71 patients et les ont classés en 7 groupes après inspection visuelle des courbes.

La classification non supervisée, couramment désignée par l'anglicisme "clustering", est une branche de l'apprentissage statistique qui conçoit des méthodes identifiant automatiquement des sous-groupes, ou "clusters", dans une collection d'observations. Ces méthodes sont fondées sur des approches mathématiques et sta-

tistiques, et sont exploratoires dans le sens où elles n'utilisent aucune connaissance experte qui aurait déjà catégorisée les observations. Cela permet de réaliser des regroupements en limitant les biais de confirmation. Le clustering longitudinal développe des méthodes pour intégrer le caractère séquentiel et/ou temporel des données dans le processus de classification. Ainsi ce type d'approches semble combler les deux défauts précédemment pointés de l'actuelle utilisation d'indicateurs statistiques pour décrire les comportements d'observance. Il conviendra alors de réaliser des regroupements en tenant compte de la variabilité de l'observance et des jours de non utilisation de la PPC, en distinguant différents niveaux d'observance, tout en reconnaissant la présence des motifs récurrents et ayant une signification clinique avérée. Le motif d'abandon présent sur la courbe du troisième patient de la Figure 1.6 (Section 1.1.3) est un exemple. Si certaines approches de clustering sont complémentaires, le panel des méthodes existantes dans la littérature méthodologique permet d'insuffler diverses philosophies de regroupement à un même jeu de données. Les principaux challenges auxquels il faudra se confronter sont d'identifier une méthode de clustering cliniquement pertinente, à travers ses manières 1) de considérer que des comportements individuels sont similaires, et 2) de définir les clusters et délimiter leur nombre.

1.2.3 Visualisation des résultats par le clinicien

Supposant que les séquences individuelles soient réparties en clusters de manière cliniquement pertinente, il reste à donner une signification à ces clusters en restituant les comportements individuels typiques d'observance associés. Les méthodes de clustering sont des méthodes statistiques exploratoires faisant se succéder plusieurs étapes entre les données qui leur sont fournies en entrée et les clusters qu'elles identifient en sortie. Elles impliquent des algorithmes pouvant nécessiter de régler certains paramètres et/ou de choisir un ou plusieurs critère(s) à optimiser. En conséquences, il n'est pas aisé d'accéder aux caractéristiques des données qui vont impacter la constitution des groupes à partir d'une méthode de clustering donnée, quand bien même celle-ci est connue et comprise sur ses aspects techniques et théoriques. Cette difficulté est davantage marquée avec des données longitudinales qui contraignent à utiliser des méthodes spécifiques ou des étapes additionnelles si l'on souhaite intégrer leur nature dans l'analyse.

Les clusters doivent être interprétables en termes humains par des chercheurs hospitaliers qui ne sont pas nécessairement spécialistes de la donnée. Ces derniers peuvent ensuite souhaiter reporter leurs résultats dans des communications scientifiques. Puis les cliniciens pourraient vouloir présenter les différents comportements types à leurs patients dans le cadre d'actions de prévention thérapeutiques. Les clusters ne sont finalement que des regroupements de données. La visualisation de données semble appropriée et plus efficace que la présentation d'un grand tableau de nombres pour chaque cluster.

Le tracé d'une courbe est un graphique facilement interprétable pour restituer une séquence d'observance individuelle. Cependant tracer toutes les courbes associées à un cluster n'assure pas l'extraction d'une information synthétique du fait de la variabilité présente au sein de certaines séquences ainsi que de leur nombre potentiellement élevé. Il est donc nécessaire de recourir à d'autres représentations graphiques.

Chaque cluster est supposé regrouper des comportements semblables sur certains de leurs aspects, avec une hétérogénéité qui peut être plus ou moins prononcée selon les clusters. Une difficulté est de précisément mettre en avant les spécificités comportementales propres à chaque cluster. Il peut y avoir par exemple un groupe d'individus ayant une observance régulière autour de 7 heures par nuit, avec peu de jours d'inutilisation de la PPC. Une courbe comme celle du patient 5 de la Figure 1.6 sera très efficace pour représenter ce comportement type. À l'inverse on peut imaginer un autre cluster rassemblant des individus abandonnant la thérapie dans les trois premiers mois. La courbe du troisième patient montre un tel abandon, mais sans information complémentaire, on ne peut savoir à quel point il faut extrapoler ce qui est représenté aux autres patients du cluster, notamment en ce qui concerne le jour de l'abandon, la variabilité de l'observance ou son niveau initial. Ces exemples montrent qu'un choix graphique admet une efficacité variable pour représenter différents clusters, et qu'il existe des risques de mauvaises interprétations.

Les méthodes de clustering identifient des structures cachées dans des données. Or les données des trajectoires d'observance sont susceptibles d'évoluer pour différentes raisons. Premièrement car les questions cliniques peuvent s'intéresser aux comportements d'observance sur des durées de thérapie diverses, et modifier les populations d'étude éligibles. Deuxièmement car la qualité des données est vouée à s'améliorer. Enfin car le sous-diagnostic du SAOS implique que le volume des données devrait augmenter. De plus, il n'existe pas une bonne et unique manière de procéder à l'identification des groupes. Toute méthode de clustering est sujette à optimisation ou à être remplacée par une méthode reposant sur une philosophie de groupement différente. Cela n'aurait alors pas beaucoup d'intérêt de représenter des clusters spécifiques obtenus dans le cadre d'une étude avec des graphiques adaptés à chacun de ces clusters. En revanche, il serait profitable de donner aux chercheurs les moyens de réaliser la représentation graphique des clusters avec un certain niveau de flexibilité, concernant par exemple le nombre de clusters ou la longueur des séquences.

La littérature médicale traitant de la problématique d'identification des comportements d'observance à la PPC fait principalement état de travaux relatifs à l'analyse statistique des séquences d'observance. Dans la section suivante, nous passons en revue cette littérature en contextualisant les méthodologies employées dans la littérature des méthodes de clustering longitudinal.

1.3 De la classification d'indicateurs statistiques à la classification des séquences d'observance

Outre les critères d'usage dans la pratique clinique pour qualifier l'adhérence, des premières études se sont intéressées à caractériser les comportements d'observance en discriminant les individus selon des seuils appliqués sur des indicateurs statistiques. En 1997 [35], il a été mis en évidence que les patients qui utilisent leur machine en moyenne au moins 6 jours par semaine ont, si l'on ne considère que les jours d'utilisation de la PPC, une utilisation moyenne quotidienne supérieure aux autres sujets. Cela a été confirmé dans une population d'enfants en 2011 [31]. En 2013 [36] des comportements ont été répartis en trois catégories tenant compte de

l'utilisation moyenne et du taux de jours d'utilisation, évalués en deux instants de la thérapie. Dans ces trois articles, les seuils sont trouvés de manière arbitraires et/ou définis a priori, et l'aspect évolutif des comportements d'observances n'est pas pris en compte.
Des travaux de classification des comportements d'observance à la PPC ont été publiés, avec parfois la recherche de prédicteurs. Ces travaux reposent sur diverses approches méthodologiques visant à fournir une partition d'un ensemble de comportements individuels. C'est à dire que tous les patients doivent individuellement être associés à un unique groupe. Nous distinguons d'une part les approches de classification à partir d'indicateurs statistiques (Sous-section 1.3.1), et d'autre part les approches réalisant la classification des séquences entières ou après modélisation de celles-ci (Sous-section 1.3.2).

Notations et définitions préliminaires

On considère x un ensemble fini de n séquences de même longueur t, $x = \{x_1, x_2, ..., x_n\}$.
On note $x_{i,j} \in \mathbb{R}$ la $j^{\text{ème}}$ cordonnée de la séquence x_i, de sorte que
$x_i = (x_{i,1},\ x_{i,2},\ ...,\ x_{i,t}) \in \mathbb{R}^t$.

Définition 1 *Soit x_i et $x_{i'}$ deux éléments de $\mathbb{R}^t$. Une dissimilarité (ou mesure ou fonction de dissimilarité) d sur $\mathbb{R}^t$ est une fonction définie sur $\mathbb{R}^t \times \mathbb{R}^t$ et à valeurs dans $\mathbb{R}^+$ vérifiant les deux conditions suivantes $\forall x_i, x_{i'} \in \mathbb{R}^t$: $d(x_i, x_i) = 0$ et $d(x_i, x_{i'}) = d(x_{i'}, x_i)$.*

Définition 2 *Une partition P de x en k groupes $P = \{C_1, C_2, ..., C_k\}$ est un ensemble de parties non vides de x, vérifiant $\bigcup_{C \in P} C = x$ et $\forall \ell, \ell',\ \ell \neq \ell' \Rightarrow C_\ell \cap C_{\ell'} = \emptyset$.*

Dans cette section nous introduisons des méthodes de clustering produisant des partitions de l'ensemble x. Le nombre k de groupes peut, selon les méthodes, être défini a priori ou déterminé par les algorithmes. Les k groupes $C_1, C_2, ..., C_k$ seront appelés des clusters (ou classes). On note n_C le cardinal d'un cluster C.

1.3.1 Classification de descripteurs (approches "feature based")

Dans chacune des approches présentées ci-après, chaque séquence x_i est décrite par un vecteur $y_i = (y_{i,1},\ y_{i,2},\ ...,\ y_{i,p}) \in \mathbb{R}^p$. La partition sur x est obtenue en réalisant la classification des éléments de y. Par abus de notation on note pareillement C_ℓ un cluster obtenu sur y ou celui correspondant dans x.
Nous recensons trois utilisations de ces approches "feature based" : avec un modèle de mélange gaussien (MMG), avec une classification ascendante hiérarchique (CAH) de Ward et avec l'algorithme des k-moyennes.

1.3.1.1 Modèle de mélange gaussien

La première classification basée sur des descripteurs a été réalisée en 2014 [37] avec un MMG ajusté selon les trois indicateurs statistiques suivants : taux de jours

d'utilisation de la PPC, taux d'utilisation supérieur à 4h et utilisation quotidienne moyenne.

L'application d'un modèle de mélange [38] pour la classification non supervisée suppose que $y = \{y_1, ..., y_n\}$ sont des réalisations de variables aléatoires $Y_1, ..., Y_n$ de $\mathbb{R}^p$ indépendantes et identiquement distribuées (iid). Les clusters sont constitués par identification de distributions qui leur sont propres. Dans le cas gaussien, chaque cluster C_ℓ est modélisé par une ellipsoïde dont la position est indiquée par sa moyenne $\boldsymbol{\mu_\ell} \in \mathbb{R}^p$, et dont la configuration est donnée selon la matrice de covariance $\boldsymbol{\Sigma_\ell} \in \mathcal{M}_p(\mathbb{R})$ associée. On note

$$\phi_\ell\left(y_i, \boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\right) = \frac{1}{\sqrt{(2\pi)^p \det\left(\boldsymbol{\Sigma_\ell}\right)}} e^{-\frac{1}{2}\,{}^t(y_i - \boldsymbol{\mu_\ell})\boldsymbol{\Sigma_\ell}^{-1}(y_i - \boldsymbol{\mu_\ell})}$$

la fonction de densité gaussienne du cluster ℓ où $\{\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\}$ sont les paramètres associés.

L'utilisation d'un modèle de mélange à k composantes (i.e. les clusters) sur y introduit n variables aléatoires iid $Z_1, ..., Z_n$ qui représentent l'appartenance de chaque observation aux différents clusters. Ces variables sont non observables et sont modélisées à l'aide d'une distribution multinomiale d'un seul tirage à k catégories. Les probabilités associées sont $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$. Elle vérifient $\forall \ell, \pi_\ell > 0$ et $\sum_{\ell=1}^{k} \pi_\ell = 1$. La fonction de densité ϕ du modèle de mélange gaussien est :

$$\phi\left(y_i, \boldsymbol{\Psi}\right) = \sum_{\ell=1}^{k} \pi_\ell \phi_\ell\left(y_i, \boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\right) \tag{1.1}$$

où $\boldsymbol{\Psi} = \boldsymbol{\pi} \cup \{\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}; \ell = 1,\ ..., k\}$ désigne l'ensemble des paramètres du modèle de mélange gaussien fini à k composantes.
L'ajustement d'un modèle de mélange gaussien à k composantes estime conjointement les paramètres $\boldsymbol{\Psi}$ ainsi que les probabilités des réalisations des classes cachées $\{z_1,\ ..., z_n\}$ où $z_i = (z_{i,1}, ..., z_{i,k})$. La classification de y est ensuite déduite selon la composante estimée la plus probable pour chaque y_i.

Le nombre de paramètres à estimer augmente avec les nombres de descripteurs utilisés et de clusters à identifier. Le nombre d'individus doit être suffisamment grand. En contrepartie, les critères de validation de modèle permettent le choix du nombre de clusters.

1.3.1.2 Classification ascendante hiérarchique

En 2021 [39], des comportements d'observance ont été classifiés par CAH avec le critère de Ward [40] à partir de deux descripteurs : taux de jours d'utilisation de la PPC et utilisation quotidienne moyenne.

La CAH est un algorithme qui repose sur l'utilisation d'une mesure de dissimilarité d entre les observations et d'un critère d'agrégation définissant la manière d'étendre l'utilisation de cette dissimilarité pour la formation des clusters. La CAH construit itérativement une suite de partitions $(P_1, P_2,\ ..., P_n)$ sur y vérifiant $P_1 = \{\{y_1\}, \{y_2\},\ ..., \{y_n\}\}$, $P_n = y$ et telle que $\forall m, \forall (C, C') \in P_m \times P_{m-1}, C \cap C' \in \{C', \emptyset\}$.
En pratique, l'application d'une CAH sur y peut se décomposer selon les étapes suivantes :

1. Calcul de toutes les dissimilarités entres les $\frac{n(n-1)}{2}$ paires d'éléments de y.
2. Construction de la suite de partitions :
   (a) Initialisation : Formation de la partition $P_1 = \{\{y_1\}, \{y_2\}, ..., \{y_n\}\}$ où toutes les observations de y sont réparties dans des clusters individuels et avec des dissimilarités initialisées selon celles calculées en étape (1).
   (b) Itérations, pour m allant de 2 à n :
      i. Formation de la partition $P_m = \{C_1^{(m)}, C_2^{(m)}, ..., C_{n-m+1}^{(m)}\}$ en réunissant les deux clusters les plus similaires de la partition $P_{m-1} = \{C_1^{(m-1)}, C_2^{(m-1)}, ..., C_{n-m+2}^{(m-1)}\}$, où $C_\ell^{(m)}$ désigne le cluster ℓ obtenu après l'itération m.
      ii. Mise à jour des dissimilarités entre les paires de clusters obtenus selon le critère d'agrégation choisi. Seules les dissimilarités impliquant le cluster dernièrement créé doivent être ré-évaluées.

La suite de partitions obtenue peut être représentée avec un dendrogramme. Choisir une partition est équivalent à "couper" ce dendrogramme et récolter les clusters par "grappes".

D'un point de vue géométrique, choisir une mesure de dissimilarité revient à choisir une manière de placer les observations dans un espace, et choisir un critère d'agrégation revient à donner des contraintes de formes géométriques aux clusters. Avant de présenter le critère de Ward défini pour le cas euclidien, nous rappelons quelques notations, définitions et résultats dans $\mathbb{R}^p$.

Définition 3 *Soit y_i et $y_{i'}$ deux éléments de $\mathbb{R}^p$. La distance euclidienne δ_E entre y_i et $y_{i'}$ est définie par :*

$$\delta_E(y_i, y_{i'}) = \|y_i - y_{i'}\|_2 = \sqrt{\sum_{j=1}^{p} (y_{i,j} - y_{i',j})^2}.$$

Remarque : une distance est une dissimilarité.

Définition 4 *Soit C un cluster quelconque de y. L'isobarycentre de C, noté $\overline{C}$, est l'élément de $\mathbb{R}^p$ défini par :*

$$\overline{C} = \underset{\xi \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{y_i \in C} ||\xi - y_i||_2^2 = \frac{1}{n_C} \sum_{y_i \in C} y_i.$$

Il s'agit de la moyenne arithmétique des éléments de C.

Définition 5 *Soit C un cluster de y. La somme des carrés de C, notée $SS(C)$ (sum of squares), est égale à la quantité :*

$$SS(C) = \sum_{y_i \in C} \left\|y_i - \overline{C}\right\|_2^2.$$

Remarque : La somme des carrés d'un cluster C est un indicateur de sa compacité. Plus les éléments de l'ensemble sont concentrés autour de leur moyenne, plus $SS(C)$ est petite.

Définition 6 *Soit C un cluster de y et d une mesure de dissimilarité sur $\mathbb{R}^p$. On définit $I(C)$ l'inertie de C pour la dissimilarité d par :*

$$I(C) = \frac{\sum_{y_i \in C} \sum_{y_{i'} \in C} d^2(y_i, y_{i'})}{2n_C}.$$

Remarque : Si d est la distance euclidienne, alors $SS(C) = I(C)$.

Il est possible de séparer la somme des carrés d'un ensemble partitionné en somme des carrés intra-clusters et somme des carrés inter-clusters. Ces deux indicateurs rendent respectivement compte de la compacité et de la séparabilité des clusters.

Définition 7 *On considère une partition $P = \{C_1, \ldots, C_k\}$ en k clusters de y. La somme des carrés intra-clusters de la partition P, notée $WSS(P)$ (within sum of squares), est définie par :*

$$WSS(P) = \sum_{C \in P} SS(C).$$

La somme des carrés inter-clusters, notée $BSS(P)$ (between sum of squares), est définie par :

$$BSS(P) = \sum_{C \in P} n_C ||\overline{C} - \overline{y}||_2^2.$$

Remarque : selon le théorème de Huygens, pour toute partition P de y, on a :

$$SS(y) = WSS(P) + BSS(P).$$

Nous présentons maintenant le critère de Ward qui est un des critères d'agrégation les plus utilisés en CAH. Ce dernier consiste lors de chaque itération à unir les deux clusters qui génèrent la croissance minimale de la somme des carrés intra-clusters (WSS). Pour simplifier les notations, on considère une partition $P_m = \{C_1, C_2, \ ..., C_k\}$ de y en k clusters construite après l'étape (i) de l'itération m. C_ℓ désigne le cluster de P_m qui est la réunion des deux clusters C_{ℓ_1} et C_{ℓ_2} de P_{m-1}, et $C_{\ell'}$ désigne un autre cluster de P_m distinct de C_ℓ et donc également existant comme tel dans P_{m-1}. On étend la notation $d(C_\ell, C_{\ell'})$ pour désigner la dissimilarité calculée lors de l'étape (ii) de l'itération m entre les clusters C_ℓ et $C_{\ell'}$ selon un critère d'agrégation. La variation positive de la somme des carrés intra-cluster $\Delta_{WSS}(m)$ à l'issue de l'itération m vaut :

$$\begin{aligned}\Delta_{WSS}(m) &= WSS(P_m) - WSS(P_{m-1}) = \sum_{C \in P_m} SS(C) - \sum_{C \in P_{m-1}} SS(C) \\ &= SS(C_\ell) - SS(C_{\ell_1}) - SS(C_{\ell_2}).\end{aligned}$$

Déterminer cette quantité pour tous les groupements deux à deux potentiels de clusters lors de chaque itération serait calculatoirement coûteux. L'implémentation de la CAH de Ward dans le logiciel R est réalisée en fusionnant à chaque itération m les 2 clusters minimisant la dissimilarité $d(C_\ell, C_{\ell'}) = \sqrt{2\Delta_{WSS}(m)}$ [41]. Cette quantité peut être calculée de manière récursive via la formule de Lance-Williams suivante, qui définit la dissimilarité entre $C_{\ell'}$ et $C_\ell = C_{\ell_1} \cup C_{\ell_2}$:

$$
\begin{aligned}
d^2(C_{\ell_1} \cup C_{\ell_2}, C_{\ell'}) = & \frac{n_{C_{\ell_1}} + n_{C_{\ell'}}}{n_{C_{\ell_1}} + n_{C_{\ell_2}} + n_{C_{\ell'}}} d^2(C_{\ell_1}, C_{\ell'}) \\
& + \frac{n_{C_{\ell_2}} + n_{C_{\ell'}}}{n_{C_{\ell_1}} + n_{C_{\ell_2}} + n_{C_{\ell'}}} d^2(C_{\ell_2}, C_{\ell'}) \\
& - \frac{n_{C_{\ell'}}}{n_{C_{\ell_1}} + n_{C_{\ell_2}} + n_{C_{\ell'}}} d^2(C_{\ell_1}, C_{\ell_2}).
\end{aligned} \tag{1.2}
$$

Remarque : L'utilisation de la formule 1.2 est possible grâce à l'égalité $SS(C) = I(C)$.

L'algorithme de classification de Ward favorise la constitution de clusters sphériques et de volumes équilibrés [42]. La CAH offre toutefois la possibilité d'explorer différentes philosophies de regroupement. Premièrement par la possibilité d'utiliser toute mesure de dissimilarité d, au prix d'un éventuel surcoût calculatoire. La formule de Lance-Williams (Équation 1.2) permet l'extension de l'algorithme de Ward avec d quelconque. Des travaux visent à proposer des fondements théoriques d'une méthode de Ward généralisée [43]. Secondairement, une fois les dissimilarités calculées, il est possible de comparer plusieurs critères d'agrégation avec des coûts calculatoires raisonnables. Le critère du saut minimal ("single linkage") construit par exemple des clusters rassemblant des observations contiguës. Le critère du diamètre ("complete linkage") construit des clusters bien séparés. Le critère de la dissimilarité moyenne ("average linkage") est intermédiaire entre les deux précédents. D'autres critères d'agrégation sont utilisables.
La représentation des partitions sous-forme de dendrogramme aide à l'interprétation des clusters. La hauteur des nœuds de fusion permet de matérialiser les dissemblances entre les clusters et peut être exploitée pour le choix du nombre de clusters. Il est de plus aisé d'extraire des partitions à différents nombres de clusters et de les comparer. Enfin, la contrainte hiérarchique imposée lors de la formation des groupes entraine de la rigidité dans l'optimisation d'un critère donné car une seule partition candidate est évaluée à nombre de clusters k fixé. Par exemple dans le cas euclidien où l'on souhaiterait optimiser le critère WSS, l'algorithme des k-moyennes est une bonne alternative.

1.3.1.3 k-moyennes

La dernière classification proposée à partir de descripteurs a été réalisée en 2021 [44] à l'aide de l'algorithme de k-moyennes [45]. Les descripteurs analysés sont la moyenne de la durée d'utilisation quotidienne ainsi que l'écart-type. L'approche proposée par Baddam et al. se distingue des précédentes car la classification est exécutée mois par mois de thérapie, en identifiant des clusters fortement similaires d'un mois sur l'autre. Ainsi leur approche définit des états mensuels d'utilisation de la PPC, et s'intéresse aux probabilités de passage d'un état à l'autre, mais ne définit pas des comportements individuels sur l'ensemble de la durée de thérapie considérée. Avec k clusters et τ mois, cela constituerait k^τ comportements d'observance possibles.

L'algorithme des k-moyennes, aussi dénommé par l'anglicisme "k-means", est une déclinaison des méthodes de partitionnement dits de "centres mobiles" également basée sur la minimisation de la WSS. L'algorithme répète alternativement deux phases où pour chaque itération m sont d'abord alloués dans des clusters les

éléments de y en fonction de leur proximité avec k centres déjà déterminés, appelés "centroïdes". Puis les centroïdes sont redéfinis par les barycentres des clusters issus de la dernière allocation. L'algorithme requiert à priori les définitions du nombre de clusters k souhaité et d'un nombre maximal d'itérations. En pratique, l'utilisation de l'algorithme des k-moyennes peut se décomposer selon les étapes suivantes :

1. Initialisation : Choix d'un ensemble de k centroïdes notés $\{\overline{C_1^{(0)}}, \overline{C_2^{(0)}}, \ ..., \overline{C_k^{(0)}}\}$.
2. Itération m, tant que les centroïdes évoluent :
   (a) Allocation des éléments de y dans des clusters $\{C_1^{(m)}, C_2^{(m)}, \ ..., C_k^{(m)}\}$ où $C_\ell^{(m)} = \{y_i \in y, \left\| y_i - \overline{C_\ell^{(m-1)}} \right\|_2 \leq \left\| y_i - \overline{C_{\ell'}^{(m-1)}} \right\|_2 \forall \ell' = 1, \ ..., k\}$.
   (b) Calcul des k moyennes $\{\overline{C_1^{(m)}}, \overline{C_2^{(m)}}, \ ..., \overline{C_k^{(m)}}\}$.

Cet algorithme assure la décroissance de la WSS après chaque itération vers un minimum local, qui dépend de l'initialisation. Dans la pratique, l'algorithme peut être exécuté plusieurs fois avec des initialisations aléatoires pour espérer atteindre le minimum global. Il faut d'autant augmenter le nombre d'exécutions que l'on souhaite considérer des partitions avec différents nombres de clusters. De même que la CAH de Ward, l'algorithme des k-moyennes classique permet la constitution de clusters sphériques et de volumes similaires dans un espace euclidien [46]. Si l'on souhaite modifier l'algorithme en utilisant une autre dissimilarité lors de l'allocation des individus à l'étape (a) de chaque itération, le calcul des centroïdes de l'étape (b) doit être adapté en conséquences. Cette adaptation n'est pas toujours aisée, rendant difficile l'extension de l'algorithme avec des mesures de dissimilarité quelconques.

L'ensemble des méthodes vues précédemment étendent facilement l'utilisation d'algorithmes de classifications classiques à des séquences en les positionnant toutes dans un même espace Euclidien. Cela présente l'avantage de regrouper automatiquement des séquences de longueurs variables mais contracte la dimension temporelle de chaque courbe en un seul vecteur, avec la conséquence d'être très réducteur de l'information portée par chaque séquence. Deux comportements d'observance différents, l'un croissant et l'autre décroissant peuvent par exemple être résumés de manière identique par la moyenne, le taux de jours d'utilisation ou l'écart-type. Ces méthodes ne peuvent suffire à prendre en compte l'aspect dynamique des comportements d'observance et leurs variations dans le temps. L'approche de Baddam et al. [44] réalisant indépendamment la classification de descripteurs calculés sur des intervalles de temps successifs n'est pas plus efficace pour cela.

1.3.2 Classification de séries chronologiques

Les valeurs des séquences étant prélevées quotidiennement, une alternative à laquelle nous adhérons est de considérer ces séquences individuelles comme des séries temporelles. Si l'on interprète les courbes de la Figure 1.6 (voir Section 1.1.3) via le prisme des séries chronologiques, on peut voir des séries avec des tendances croissantes (la première) ou décroissantes (la deuxième), une série caractérisée par la présence d'une rupture (la troisième), ou encore une série qui ressemble à du bruit blanc (la cinquième). Ceci motive d'utiliser la classification non supervisée de séries chronologiques [47]. Parmi les autres travaux de classification des comportements d'observance à la PPC publiés, nous distinguons deux grandes catégories de

méthodes de clustering : les méthodes basées sur des modélisation des séries chronologiques et les méthodes reposant sur l'utilisation d'une mesure de dissimilarité entre séries chronologiques.

1.3.2.1 Classification basée sur une modélisation des séries chronologiques (approches "model based")

Les approches listées dans cette section reposent sur l'utilisation de modèles de mélange finis où les densités peuvent prendre différentes formes pour modéliser les séquences d'observance en fonction du temps. La fonction de densité f d'un modèle de mélange dans le cadre général s'écrit de manière analogue à l'équation 1.1, en remplaçant les densités gaussiennes ϕ_ℓ par des densités f_ℓ et les paramètres spécifiques $\{\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\}$ par des paramètres génériques $\boldsymbol{\theta_\ell}$:

$$f\left(x_i, \boldsymbol{\Theta}\right) = \sum_{\ell=1}^{k} \pi_\ell f_\ell\left(x_i, \boldsymbol{\theta_\ell}\right)$$

où $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_k}\}$ désigne les paramètres du modèle de mélange fini à k groupes.

Deux publications ont ainsi réalisé la classification des comportements d'observance à la PPC. En 2013 [48], un modèle de mélange de chaînes de Markov homogènes a été appliqué. Cette approche requiert de discrétiser la variable d'observance selon des états via des seuils prédéterminés et indépendants de sa distribution. Cette méthode présente l'avantage de pouvoir traiter des séquences de longueurs variables, contrairement à la seconde méthode publiée en 2021 [49]. Dans cette approche, la modélisation de l'observance dans les clusters est réalisée avec des modèles additifs généralisés fonction du temps. Les modèles sont des lois normales tronquées avec effet "hurdle". Les modèles "hurdle" sont utilisés avec des variables présentant un nombre important de zéros, pour modéliser un "obstacle" à obtenir des valeurs non nulles. Ils intègrent une probabilité pour les valeurs nulles, et la probabilité complémentaire est associée à une une loi de probabilité tronquée en zéro. Les paramètres de chaque composante sont pour chaque temps la probabilité d'avoir une observance nulle, et en cas d'observance positive, la moyenne et la variance de l'observance. L'évolution de ces paramètres est prise en compte à travers des fonctions quadratiques du temps. Un paramètre supplémentaire est la variance du décalage individuel de l'ordonnée à l'origine de chaque courbe d'observance. Cette approche est intéressante car elle permet de modéliser d'une part l'apparition des zéros dans les séquences, ainsi que l'évolution du comportement d'observance dans le temps.

1.3.2.2 Classification reposant sur l'utilisation d'une dissimilarité entre séries chronologiques (approches "shape based")

Ces approches consistent à adapter des algorithmes de clustering "classiques" par l'utilisation de mesures de dissimilarité compatibles avec des séries chronologiques. Nous dénombrons trois publications concernant la PPC, publiées entre 2015 et 2020, et toutes considèrent les séquences dans un espace euclidien en minimisant la WSS. Ceci suppose de travailler avec un ensemble de séquences de même longueur.

Deux publications [50, 51] utilisent les k-moyennes. Dans [52] une CAH de Ward est utilisée après exclusion des séries comportant trop de données manquantes pour permettre à une stratégie d'imputation de fonctionner. Les méthodes détaillées en

Section 1.3.1 sont appliquées telles quelles sur l'ensemble x à la place de l'ensemble y.

À notre sens, ces choix méthodologiques sont intéressants car ils permettent de traiter les séquences entières sans les réduire via des indicateurs statistiques "ponctuels", ni introduire de biais de modélisation. Cependant l'utilisation de la distance euclidienne est discutable entre des séries chronologiques. Premièrement elle compare les comportements d'observance jour par jour. Des courbes admettant des formes identiques sauf à un décalage temporel près pourraient être jugées fortement dissimilaires bien qu'elles décrivent un même comportement d'observance. Secondairement la distance euclidienne est d'une certaine manière insensible au caractère longitudinal des séries chronologiques en les considérant comme des vecteurs avec des coordonnées temporellement indépendantes.

Il existe des fonctions de dissimilarité plus aptes à considérer le caractère temporel des séquences d'observance. Dans la section suivante nous présentons les contributions apportées par cette thèse pour le clustering des séries chronologiques d'observance à la PPC. Elles concernent tant les questions liées à l'analyse statistique des séries temporelles que les questions liées aux problématiques de production des séries chronologiques et de restitution des résultats du clustering.

1.4 Contribution de la thèse

Hypothèses fixées pour une réponse à la problématique clinique

La durée quotidienne de port du masque lorsque la machine administre une pression thérapeutique est supposée être un indicateur pertinent de l'observance thérapeutique à la PPC.

Les séquences individuelles d'observance sont considérées comme des séries chronologiques.

Nous fixons l'horizon des trois premiers mois (91 jours) de thérapie pour l'étude de l'observance, en supposant qu'il y a stabilisation du comportement dans ce délai. Cela ne signifie pas que l'observance est supposée être peu variable à l'issue des trois mois pour tous les patients, mais que ces derniers ont pu s'adapter au traitement et que leur comportement d'observance à l'issue de cette période est celui qu'ils pourront maintenir à plus long terme. Ce choix du délai de trois mois provient initiallement d'une publication [53] indiquant que la plupart des abandons de thérapie a lieu dans ce délai. Cela a récemment été contredit dans la littérature clinique où un nombre non négligeable d'abandons se produit après la première année. En 2020 [25], une étude sur le suivi de 181 patients au Japon pendant 10 ans révèle que près de la moitié des 56 patients ayant abandonné l'on fait entre le début de la deuxième année et avant le 76 $^{\text{ème}}$ mois de thérapie. En 2021 [12], l'analyse des données du SNDS dévoile que près de 25% des patients ayant commencé la thérapie en France entre 2015 et 2016 ont abandonné lors de la deuxième ou de la troisième année de traitement.

1.4.1 Production des données

Les données à disposition pour les travaux présentés dans cette thèse proviennent d'un PSAD implanté dans les alentours de Grenoble. La production des séries chronologiques d'observance individuelles requiert de faire attention à ce que les séquences produites reflètent fidèlement les comportements d'observance des patients. Or les données fournies peuvent présenter des biais inhérents à leurs modalités de récolte et de circulation qui implique plusieurs acteurs. Pour identifier les biais et sources d'erreur, il convient alors de s'approprier le processus de collecte qui intègre leur enregistrement par les machines, leurs flux, ainsi que les pratiques et retraitements opérés par les différents acteurs par lesquels elles transitent.

Nous avons entrepris cette démarche au sein du laboratoire HP2 avec de nombreux échanges avec le PSAD fournisseur des données. Considérant ces échanges, la littérature ainsi que notre expérience dans l'analyse de ces données, nous avons soumis une publication adressée à la communauté scientifique médicale dont l'objectif est de rendre plus intelligibles les données issues du télésuivi de la PPC. Nous exposons des biais et proposons des recommandations en termes de datamanagement. Le Chapitre 2 comporte cette prépublication et la relie à la problématique clinique adressée par les travaux de la thèse. La finalité est de permettre la production de séries temporelles fiables, de même longueur et sans données manquantes.

1.4.2 Analyse statistique

Afin de traiter les séries chronologiques avec leurs variabilités, discontinuités ou ruptures caractéristiques, nous optons pour une approche de classification basée sur une dissimilarité applicable entre séries chronologiques, sans recourir à aucun lissage de ces dernières. Nous choisissons d'utiliser une mesure de dissimilarité prenant en compte la forme des courbes et « annulant » l'effet des décalages temporels. Une motivation est qu'un abandon de traitement après deux semaines est d'un point de vue clinique identique à un abandon après douze semaines de PPC. Nous avons identifié deux mesures de dissimilarité compatibles dans la littérature : la dissimilarité produite par l'algorithme du dynamic time warping (DTW) [54], fréquemment utilisée pour réaliser la classification des séries chronologiques [47], et une variante sommée de la distance de Fréchet discrète [55] désignée sdF dans ce mémoire. Ces deux dissimilarités sont dites "élastiques" car elles procèdent à des dilatations simultanées des axes temporels lors de la comparaison de deux séries temporelles. Ces dilatations sont réalisées de manière à aligner d'éventuels motifs communs. La dissimilarité sdF peut être vue comme une généralisation de DTW dont le calcul tient compte de l'amplitude des dilatations effectuées.

Dans le Chapitre 3, nous formalisons la définition de la dissimilarité sdF de sorte qu'elle généralise la dissimilarité DTW. Nous évaluons si ces deux mesures de dissimilarité peuvent, selon différentes méthodologies de clustering, être pertinentes dans le contexte clinique. Nous nous plaçons dans le cadre de la CAH, permettant de comparer différentes manières de procéder au regroupement des séries. Nous avons dans un premier temps réalisé une étude de simulation dans laquelle nous avons intégré la question du choix du nombre de groupes d'une manière objective en comparant différents indices de validation de classification internes. Puis nous avons réalisé le clustering de données réelles.
Ce travail a fait l'objet d'une publication dans la revue "Statistics in Medicine" en

2021. Il montre que la dissmilarité DTW semble adaptée pour réaliser la classification des comportements d'observance à la PPC lorsque les clusters sont construits avec le critère d'agrégation de Ward, et que le nombre de clusters est choisi avec l'indice de Dunn [56]. Cette méthodologie de classification est la plus performante d'après l'étude de simulation, et elle permet d'identifier une partition des comportements d'observance réels en 6 clusters ayant un sens clinique. Les bons résultats obtenus par le critère de Ward semblent indiquer que les comportements typiques peuvent être décrits par des clusters organisés autour de centres pour la dissimilarité DTW. Les bonnes performances de l'indice de Dunn pour le choix du nombre de clusters semblent indiquer que les clusters sont compacts et bien séparés. Nous définissons ci-après la dissimilarité du DTW et l'indice de Dunn.

Définition 8 *Un chemin de déformation ("warping path") W de longueurs q et q' est une séquence $\big((a_1, b_1)\ , (a_2, b_2),\ ..., (a_{\ell_W}, b_{\ell_W})\big)$ de paires distinctes de $[\![1, q]\!] \times [\![1, q']\!]$ telle que : ℓ_W soit un entier supérieur ou égal à $\max(q, q')$, $a_1 = b_1 = 1$, $a_{\ell_W} = q$, $b_{\ell_W} = q'$ et $\forall \ell \in \{1,\ ...,\ \ell_W - 1\}$, $(a_{\ell+1} - a_\ell,\ b_{\ell+1} - b_\ell) \in \{(0, 1), (1, 0), (1, 1)\}$.*

Définition 9 *Soient $E = (e_1,\ ..., e_q)$ et $F = (f_1,\ ..., f_{q'})$ deux familles ordonnées d'éléments de cardinaux respectifs q et q'. Soit $W = \big((a_1, b_1)\ ,\ ..., (a_{\ell_W}, b_{\ell_W})\big)$ un chemin de déformation de longueurs q et q'.*
La séquence $L_W = \big((e_{a_1}, f_{b_1}),\ ..., (e_{a_{\ell_W}}, f_{b_{\ell_W}})\big)$ est appelée couplage ("coupling") entre E et F.

Définition 10 *Soient x_i et $x_{i'}$ deux séries chronologiques et $\mathcal{L}_{i,i'}$ l'ensemble de tous les couplages possibles entre x_i et $x_{i'}$. La dissimilarité du dynamic time warping d_{DTW} entre x_i et $x_{i'}$ est définie par :*

$$d_{DTW}\left(x_i, x_{i'}\right) = \min_{L_W \in \mathcal{L}_{i,i'}} \sum_{\ell=1}^{\ell_W} \left\| x_{i,a_\ell} - x_{i',b_\ell} \right\|_2 .$$

Nous avons introduit la version de base de la dissimilarité sur laquelle se sont appuyés nos travaux. Des variantes existent et nous renvoyons par exemple vers l'ouvrage de Mueller [54] pour plus de détails sur les possibles paramétrages de la dissimilarité. Cette dernière ne remplit pas les conditions de séparabilité et ne respecte pas l'inégalité triangulaire. Grâce à la Définition 9 du couplage entre deux séries chronologiques, il est possible d'appliquer DTW entre des séries de tailles différentes et possédant des données manquantes. Un algorithme de programmation dynamique peut être utilisé pour procéder au calcul de cette dissimilarité et éviter l'évaluation de tous les couplages possibles.

Définition 11 *Soient d une mesure de dissimilarité définie sur $\mathbb{R}^t$ et P une partition de x en k clusters. L'indice de Dunn évalué sur P avec d est :*

$$Dunn(P) = \frac{\min_{1 \le \ell < \ell' \le k}\ dmin(C_\ell, C_{\ell'})}{\max_{1 \le \ell \le k}\ dmax(C_\ell)}$$

où $dmin(C_\ell, C_{\ell'}) = \min_{(x_i, x_{i'}) \in C_\ell \times C_{\ell'}}\ d(x_i, x_{i'})$ et $dmax(C_\ell) = \max_{(x_i, x_{i'}) \in C_\ell}\ d(x_i, x_{i'})$.

Cet indice est le rapport entre la dissimilarité minimum qui sépare deux éléments de clusters distincts et la dissimilarité maximum qui sépare deux éléments d'un même cluster. En pratique on cherche à le maximiser car plus il est grand, plus les clusters sont compacts et séparés.

1.4.3 Visualisation des résultats

Pour rendre compte des différents comportements regroupés dans un même cluster, en mettant en avant leurs caractéristiques homogènes mais également leur diversité, nous proposons de diversifier les représentations graphiques des clusters. L'objectif est de permettre l'interprétation des clusters par croisement des informations portées par chacun des graphiques.

Les cliniciens et chercheurs du domaine médical n'ont pas nécessairement les compétences en programmation statistique pour réaliser des graphiques adaptés à la restitution des résultats du clustering. Le Chapitre 4 présente une application web en cours de développement et permettant de réaliser de multiples représentations graphiques. Nous les illustrons sur les 6 clusters obtenus sur données réelles dans le Chapitre 3. L'application est conçue via le langage de programmation R avec le package Rshiny pour être facilement utilisable par des chercheurs en milieu hospitalier. L'outil proposé est dédié à la description des clusters et intègre une part de fonctionnalités exploratoires par la possibilité de croiser les clusters avec des covariables pour affiner l'analyse des clusters. Les graphiques productibles permettent quelques interactions comme la modification des échelles, le choix des clusters à représenter ou la personnalisation des couleurs. Enfin les figures sont exportables en format image pour permettre de les utiliser en dehors de l'application web.

Chapitre 2

Les données, depuis le domicile des patients aux séquences exploitables pour l'apprentissage des comportements typiques d'observance

2.1 Introduction

Dans ce chapitre, il est question de la production des séquences décrivant les trajectoires d'observance individuelles en début de traitement à la PPC. La constitution des séquences est facilitée par la généralisation du télésuivi de la PPC et la mise à disposition des données par les PSAD. Des précautions sont cependant requises pour réduire la présence des biais qu'elles comportent. Notre contribution est, à partir de données fournies par un PSAD, de favoriser l'extraction des séquences de mêmes longueurs, sans données manquantes, et les plus justement représentatives des trajectoires d'observance des patients télésuivis. Il s'agit de privilégier la qualité des données, au détriment de leur quantité. Les critères de qualité considérés sont 1) l'exactitude des valeurs quotidiennes, 2) la complétude des séquences, 3) leur cohérence par rapport aux trajectoires réelles, et 4) la conformité du traitement administré, sur sa nature ainsi que son caractère initial.

La première section est une prépublication destinée à la communauté des soignants et des scientifiques susceptibles de manipuler les données machines produites dans le cadre du télésuivi. Nous exposons des biais et proposons des recommandations pour s'en affranchir. Cet article est co-écrit en premier auteur avec Alphanie Midelet, doctorante CIFRE au laboratoire HP2. Le cadre traité est plus large que la variable étudiée dans cette thèse car nous considérons l'IAH résiduel ainsi que des fuites qui sont également des indicateurs estimés par les machines. La deuxième section aborde les aspects spécifiques liés à l'étude de la variable d'observance pour la problématique médicale. Enfin des perspectives sont adressées en vue de l'amélioration de la qualité des données. Nous abordons des éléments relatifs à la pertinence de la variable étudiée, et des éléments concernant la représentativité d'un échantillon de séquences extractible des données d'un PSAD.

2.2 Considérations préalables à l'analyse des données produites par les appareils de PPC dans le cadre du télésuivi

Title: Remote monitoring of positive airway pressure data: Challenges and Pitfalls to consider for optimal data science applications

Authors:

Guillaume Bottaz-Bosson[1,2]*, guillaume.bottaz-bosson@univ-grenoble-alpes.fr

Alphanie Midelet[1,3]*, alphanie.midelet@probayes.com

Monique Mendelson[1], mmendelson@chu-grenoble.fr

Jean-Christian Borel[1,4], j.borel@agiradom.com

JB Martinot[5,6], martinot.j@respisom.be

Ronan Le Hy[3], ronan.lehy@probayes.com

Marie-Caroline Schaeffer[3], marie-caroline.schaeffer@probayes.com

Adeline Samson[2], adeline.leclercq-samson@univ-grenoble-alpes.fr

Agnès Hamon[2], agnes.hamon@univ-grenoble-alpes.fr

Renaud Tamisier[1], rtamisier@chu-grenoble.fr

Atul Malhotra[7], atulandkaren@gmail.com

Jean-Louis Pépin[1#], jpepin@chu-grenoble.fr

Sébastien Bailly[1#], sbailly@chu-grenoble.fr

*co-first authors

#co senior authors

1 Univ. Grenoble Alpes, Laboratoire HP2, U1300 Inserm, CHU Grenoble Alpes, Grenoble, France

2 Laboratoire Jean Kuntzmann, Univ. Grenoble Alpes, CNRS, Grenoble, France

3 Probayes, Montbonnot-Saint-Martin, France

4 AGIR à dom. HomeCare Charity, 38240 Meylan, France

5 Sleep Laboratory, CHU UCL Namur Site Sainte-Elisabeth, Namur, Belgium

6 Institute of Experimental and Clinical Research, UCL, Bruxelles Woluwe, Belgium

7 University of California San Diego, 8784, Division of Pulmonary, Critical Care and Sleep Medicine, La Jolla, California, United States.

Corresponding author: Dr Sébastien Bailly - Laboratoire EFCR, CHU de Grenoble, Rond point de la Chantourne, CS10217, 38043 Grenoble Cedex 9, France. Email: sbailly@chu-grenoble.fr

Funding

Funding GBB, JLP, SB, and RT are supported by the “e-health and integrated care and trajectories medicine and MIAI artificial intelligence” Chairs of excellence from the Grenoble Alpes University Foundation (ANR-19- P3IA-0003) and the French National Research Agency in the framework of the "Investissements d’avenir” program (ANR-15-IDEX-02). AM is supported by Probayes and MIAI (ANR-19- P3IA-0003) in the framework of a “Convention Industrielle de Formation par la Recherche” (CIFRE) PhD. Her PhD is also supported by the French National Research Agency (grant 2020/0007). Dr. Malhotra is funded by National Institutes of Health.

Conflict of interest: none of the authors have any competing interest in the manuscript

ABSTRACT

Remote monitoring has increased over the past years for obstructive sleep apnea and generates a huge quantity of data regarding positive airway pressure (PAP) device usage and airflow. Collected PAP data provide the opportunity to access valuable and objective information regarding patient treatment adherence and efficiency. However, the majority of studies based on longitudinal PAP remote monitoring data summarize the data trajectories in static and simplistic metrics for PAP adherence by using mean or median values. The aims of the present manuscript are to suggest directions for improving data processing and cleaning and to address major concerns for data science applications using daily remote monitoring data of PAP devices including: 1) the absence of metric standardization between reports and data provided by different PAP brands, 2) specific data management from aggregated data, 3) missing values and 4) consideration of PAP treatment interruptions.

To allow fair comparison between studies and to avoid biases in computation, PAP data processing and management should be carefully considered recognizing these points.

The potential of using PAP remote monitoring data is important and is currently underused in the field of sleep research. Improving the quality of data issued from PAP remote monitoring can

contribute to the development and sharing of data worldwide to advance toward artificial intelligence approaches adapted for big data.

Keywords: Positive Airway Pressure, Datamanagement, Time series, Obstructive sleep apnea

Introduction

Obstructive sleep apnea (OSA) is a highly prevalent chronic disease with nearly one billion adults aged 30-69 years affected worldwide [1]. OSA is a systemic disease which presents multiple clinical phenotypes [2] and is independently associated with cardiovascular comorbidities, decreased quality of life, alteration of neurocognitive function and depression [3, 4]**.** Positive airway pressure (PAP), the first-line therapy for moderate to severe OSA, used by up to 450 million people worldwide [1] improves symptoms and quality of life and can reduce cardiovascular risk [4-6]. The effectiveness of PAP treatment relies on adherence [7] and it is admitted that a minimum of four hours of PAP/night is required in order to improve blood pressure control in minimally symptomatic patients [8].

Over the past 10 years, the development of communicating PAP devices and willingness of manufacturer and home care providers to reshape follow-up care have enabled the emergence of remote monitoring platforms for visualization of nightly data generated by hundreds of millions of patients worldwide [9]. The technology has transformed the way patients are monitored by providing an opportunity for early interventions in order to improve disease management [10, 11], PAP adherence [12, 13] and complete telemedicine activities in sleep disordered breathing disorders [14]. In some countries, such as France, remote monitoring of PAP adherence is a condition for healthcare insurance reimbursement of PAP devices and the rates of reimbursement are related to levels of adherence.

PAP data can be used to improve our understanding of OSA and to personalize treatment by using innovative statistical approaches alongside artificial intelligence in the identification and

prediction of OSA patient trajectories under treatment. This issue constitutes a major challenge now identified in the treatment and management of OSA [15]. PAP monitoring data include daily aggregated measurements of adherence [16] and treatment efficacy (residual apnea-hypopnea index (rAHI), leaks) (Figure 1).

The majority of studies based on longitudinal PAP remote monitoring data summarize the data trajectories in static and simplistic metrics for PAP adherence by using mean or median values [10, 17-19] .Actually, the value of these data lies in the complexity of their evolution and variability over time which is highly relevant for patients care and management. As data are collected daily, they can be analyzed as time series, and modeled with multiple components, both deterministic and stochastic. Indeed, these data may reveal, for example, trends, cyclic components or disruptions in the observed behaviors [20-22].

At this step of knowledge, the reliability of data remains questionable, especially the management of missing data, the absence of standardization between PAP brands of PAP generated indicators [23] and the involvement of different intermediaries (i.e. home care providers or digital health private companies) processing original data with some approximations or dimensional reductions. Clinicians and scientists accessing these data are generally blinded regarding all the processing pipeline that might modify their interpretation of the data. Thus, there is a need to develop a common language for processing, reporting and interpreting these data.

The aim of the present manuscript is to suggest directions for improving the data processing and cleaning and address major concerns for data science applications using daily remote monitoring data of PAP devices (including fixed PAP, and automated PAP, APAP).

In the following sections, the objective is to provide a panoramic view of the landscape of PAP remote monitoring data, including concerns regarding: 1) conditions for rAHI reliability, 2) the absence of indicators standardization between data provided by different PAP brands, 3) missing

values and 4) consideration of treatment interruptions. These issues, impact and potential solutions are summarized Table 1.

Data origin

Raw data, including airflow and pressure, are collected at a high frequency (i.e. for example 25 Hz for the airflow signal) by the PAP devices. Real-time respiratory events detection is performed during the night by proprietary softwares embedded in the device. At the end of the night, raw records and summary statistics of the data are sent to the equipment provider and stored in the SD card of the PAP device. Then, summarized data are made available to healthcare providers, and these data include, whatever the PAP brand, time and duration of PAP usage, estimated residual AHI, mean or median, 95th percentile and/or maximum leaks.

Different technical ways exist to obtain data from device manufacturers and/or specific stakeholders involved in the follow-up pathway (Figure 2). This can vary between countries and health policies. Data flow can be obtained: 1) via an external modem, outside of the PAP device, which can be used to transmit PAP data or 2) via a connected PAP device which uses Wi-fi, Bluetooth or cellular service to automatically transmit PAP data.

Advances in technology have made the connected PAP device the most common remote monitoring system, sending data on a daily basis, and keeping data history over several weeks to send then later in case of transmission failure (e.g. because of interrupted internet connection). Some limitations have to be considered due to transmission systems or due to specific device limitations [24]. Respiratory events detection and summary statistics computation are performed according to individual manufacturer specifications. Data can be displayed on the manufacturer's platform or sent to different healthcare provider's platforms via Application Programming

Interfaces (API) as raw or aggregated data, which is one measure aggregated over 24 hours or from noon to noon for a given patient (Figure 2).

Reliability of the residual AHI

Independently to the data transmission procedure, there are two concerns to know about the reliability of the reported rAHI. First, under 2 hours of consecutive PAP use, the metric may not be reliable or may be biased as unstable sleep structure might affect AHI and its central component [25]. Secondly, above a maximum leak the machine is not able to maintain the pressure setting threshold, the respiratory event detection is unreliable and the rAHI can be distorted as illustrated in Figure 3 and should probably not be considered. Indeed, all masks are intentionally designed to leak in order to prevent re-breathing of CO_2. Thus, every mask is characterized by its intentional leak rate. The non-intentional leak is the leak due to a bad fit of the mask, an overused mask, mouth breathing, facial relaxation in deep sleep or movements of the mask while rolling over in bed. It is the volume of leaks above the intentional leak rate.

Device shift and absence of standardization of reported indicators

PAP devices shift can generate some transmission bugs during a periods of time leading to missing values, i.e. the absence of a record, even though the subject uses the device. Thus, the missing values cannot be interpreted as non-usage of the PAP therapy.

Moreover, device brand specificities are associated with significant variations in reported indicators as there is no standardization in computation nor in reporting methods between PAP brands. Indeed, parameters included in rAHI and summary statistics for leaks are different regarding PAP manufacturers (non-intentional or total leaks, median or mean and/or 95^{th} percentile and/or maximum value over the night) [26].

Thus, we can observe a sudden change in the behavior of the rAHI, due to both an adaptation period with the new device and/or a difference in the levels of reported rAHI (Figure 4) [23].

As a consequence, for a given patient, transition periods between two distinct devices require specific data pre-processing. To assess the impact of device changes at individual level, we suggest to systematically identify device changes and compare the distribution of the indicators and the rates of missing and null usage values over a period before and after the change. In the analysis of large databases for scientific purposes, a solution can be to consider only one device for a given patient and to censor the second device. Another possibility is to consider only devices from a single manufacturer to ensure that analyzed values are obtained by a common methodology.

Multiple records of the same device for a patient

This task aims to verify that records do not include duplicates or errors in device attribution. After considering device changes, this point can be considered in three different scenarios.

First, when multiple sleep bouts are transmitted owing to several PAP re-start during the night or during naps, several records can be sent across the same day. We suggest to aggregate them by summing the usage values and computing the residual AHI weighted means, keeping in mind that rAHI is probably reliable only for uses superior to 2 hours.

The second potential concern occurs when there is a mismatch between the machine identification and the patient identification, which can lead to assigning the records of two distinct devices to the same patient. The consequence is that a patient can have several rows for different devices although they use only one PAP device (Figure 5A). In practice, this situation can occur when a device has some technical issues and the patient gives it back to the device provider. Mismatch can also occur when patients are changing their home care provider. The problem can

be solved by asking the patient to provide the device ID (Figure 5B and 5C) or by removing the period with duplicate devices if it is impossible to identify which is the right one (Figure 5D).
The third situation is characterized by a dataflow issue, resulting in duplicate rows for a patient with the same device and collected values. This situation can be solved by removing duplicate rows where all the values are identical. This removal may be sufficient to only keep one row per patient per day.

Missing values

It is important to check the distribution of rates of missing values and null usage times, i.e. record with a zero on the usage time to reveal inconsistencies. Importantly, a possible source of missing data is related to PAP devices transmitting data by using an external modem, which can be disconnected or misconnected to the PAP device. These devices should be identified in the database and preferentially excluded before analysis to avoid error in interpretation of null or missing values. Moreover, according to countries' regulations and personalization of the follow-up, additional annotations may be collected to explain the non-usage of the devices and reduce the rate of missing values: hospitalizations, vacations or discontinuation of PAP treatment.

Imputation of missing values

In recent literature, different methods have been used to consider missing values in PAP adherence remote monitoring: 1) replacing missing values by zero [27-29], 2) using an imputation process when it is due to transmission failure [20, 22], 3) excluding subjects with excessive missing data [22] 4) not replacing data [30] or 5) not reporting what is done [31].

The relevance of each method depends on the situation: 1) when missing values are observed between two transmissions and 2) when missing values are observed at the end of the patient's time series.

a. Imputation of PAP adherence values in case of missing records between two transmissions

Let us consider a subject with several consecutive missing records between two transmissions (Figure 6A and 6B). The storage capacity of their device is ***p*** days. Then the period with missing records can be imputed with zero values if the length of this period is inferior to ***p*** (Figure 6A) and if the patient was not using another device, during a hospitalization for example.

In the case where information on periods such as vacations is known, it is possible to consider these as specific cases in which missing values can be replaced by zero, whatever the storage period of the device if the patient identifies that they did not use their device over this time period (Figure 6A).

Otherwise, missing values cannot be imputed with zero values and have to remain as missing values or a specific imputation method should be considered if the patient effectively used a PAP device, which requires the understanding of the adherence behaviors (Figure 6B).

b. Imputation of PAP adherence values in case of missing records at the end of a patient's sequence

Some patients can present missing values during the interval between the last available record and the date when the data are exported for analysis (Figure 6C). This can be due to PAP termination e.g. institution of alternative therapies, change in PAP provider or death. If it is known that the subject withdrew their therapy, then their missing sequence should be imputed with zero values. In the case of device restitutions due to a change in PAP provider, then the subjects have

continued their therapy so the sequences should not be filled with zero values. A sequence interrupted because of death should be censored before being included in any analysis.

c. Impact of missing values imputation on PAP adherence

From a sample of 407 patients with daily PAP adherence values collected during 3 months after PAP initiation, without PAP termination observed at 3 months, 33 patients (8%) had missing values if PAP adherence is considered over the first month of PAP use.

This proportion increased to 15% at month 2 and 19% at month 3. In the total sample of patients, no change in the mean of the individual average adherence values was observed whether individual averaging was done with or without imputing (Figure 7). The largest difference is observed if patients with missing values are removed from the sample (Table 2). If we consider only patients with missing values, the imputation strategy can lead to more important changes due to the limited sample size (N=33, 60, 78). The changes observed are more important in this sub-sample, especially if missing values are replaced by zeros (Table 2 and Figure 7).

In summary, if missing values occur between two transmissions: it is suggested to replace by zeros if the period is shorter than the PAP storage capacity and there was no other PAP device used (e.g. at the hospital), and to keep the missing or use an imputation strategy otherwise. If missing values are at the end of the exported sequence: it is suggested to censor in case of lost follow-up or death, to keep the missing values or use an imputation strategy in case of provider switch, and to replace by zeros in case of PAP termination.

To impute missing data, one solution consists in considering each indicator as an independent time series and comparing the performance of several methods: mean/mode imputation, last observation carried forward (LOCF), next observation carried backward (NOCB), mean of the last and the next observation, interpolation (linear, spline), and time series modeling and forecast [32]. Another solution consists in imputing the multiple incomplete variables with an imputation method

suited to multivariate longitudinal data: either using joint modelling (multivariate normal imputation) or with fully conditional specification (sequential regression and multiple imputation by chained equations) [33].

Treatment interruption

In the case when daily aggregated data are considered, there is no distinction whether the patient did three 2-hour naps or slept 6 hours consecutively at night. In a sample of 15.7 million of night from 2,607 patients using PAP from the same manufacturer between June 2017 and April 2021, the majority (81.3%) of the records correspond to a single or two-mask removals, and the mask was removed between 3 and 5 times in 15.9% of the night. Nevertheless, 2.8% of the records are the aggregation of 5 to 10 uses. If some data about the number of mask removals and the distinct periods of use over the 24 hours are available, the nights that are broken up in many parts should be identified.

Indeed the same aggregated value of PAP adherence might reflect completely different PAP usages potentially related to PAP termination and prognosis: 1) In patients with comorbid insomnia and sleep apnea (COMISA) [34, 35], one of the most frequent OSA phenotype number of sessions during the night and naps reflect the severity of insomnia, sleep deprivation and changes induced in the first months after PAP initiation. 2) It has been well established that AHI in REM sleep is related to hypertension and diabetes [36, 37]. REM sleep predominantly occur at the end of the night. Thus, it is preferable to use raw data which provide the exact time of PAP exposures during night and day.

Finally, in the case where the analysis focuses on the indicators measured while the patients use their PAP (residual AHI, leaks etc.), only nights where the patient removed their mask less than

5 times/night (i.e. the night was divided into 5 parts or less), and usages longer than 2 hours may be analyzed.

DISCUSSION

The remote monitoring of PAP data provides the opportunity to access valuable and objective information regarding patient treatment use and efficiency. It provides access to continuous monitoring of this information as well as its variability over time. Numerous studies were based on summary statistics of daily PAP uses and indicators and there is growing interest for considering the sequential structure of the daily aggregated data trajectories, that can be considered in time series analysis framework.

To improve the possibility to compare studies and to avoid biases in computation, PAP data processing and management should be carefully considered. According to the objectives, several points have to be considered. Firstly, excessive leaks, with definition varying according to devices manufacturers result in inaccurate rAHI estimations. Second, the duration of the PAP use has to be considered. In order to analyze PAP reported rAHI, device usage shorter than 2 hours should be carefully considered (they can be meaningful but certainly do not reflect the actual mean number of residual events of a patient). But these short usages should be kept for reporting detailed PAP adherence and sleep bouts across the 24 hour span [34]. Then, missing values should be investigated and imputation strategies should be developed. Future studies should compare imputation methods on PAP data in order to set guidelines. Additionally, errors in PAP device attribution or data transmission should be explored to avoid different situations including multiple devices for one patient in the case of a device change or not, or multiple data for one patient in the same day. Different granularity levels of aggregation for the data could be provided, such as details about mask removal during night, which can provide several records for a single night of PAP use. This could be very informative because these data provide contextual annotations on possible patient phenotypes for PAP device use, including possible nocturia, or

other sleep disorders such as insomnia or sequential sleep period including naps. To our knowledge, such data are not currently considered in PAP data studies. Finally, in the case of a device change, if the new device comes from a different manufacturer, the non-standardization of indicator computations (leaks or rAHI) leads to important biases in data analyses. To avoid such biases, it would be preferable to consider records for a patient with a single device model. A possibility is to use an external device which can ensure the PAP standardization regardless of the PAP manufacturer [38] but precautions must be taken to avoid the problems encountered with external modem transmissions. Alternatively, implementing international guidelines for standardization can help to solve this concern.

The development of PAP remote monitoring worldwide increases data availability for an important population of patients. This is a unique opportunity to develop and improve adequate data analyses by considering not only computed features at trajectories' level, but whole time series of daily records, and even raw data with a higher frequency such as the raw airflow signal recorded by PAP devices. This approach could contribute to the development of personalized tools to improve PAP management and to ensure data quality control for health insurance reimbursement. By developing standardized processes for data management of PAP records and by working on a standardization of metrics produced by PAP devices, it could be possible to compare different management patterns, develop standardized tools for patient monitoring in order to improve PAP adherence or treatment efficiency.

The potential of using PAP remote monitoring data is important and is currently underused in the field of sleep research. Improving data quality from PAP remote monitoring would help to develop and share PAP data worldwide to move toward artificial intelligence approaches adapted for big data. This could be an opportunity to develop personalized medicine by various applications including identification of patient trajectories of PAP use or variability of treatment efficiency based on rAHI measures [21, 22, 29, 39]. The different clusters of PAP trajectories are potentially linked to patients reported outcomes and long-term incident cardiovascular events and mortality.

However, some limitations have to been acknowledged when considering only the PAP measure. There are a number of relevant information, such as the number of hours of sleep [40], patient-reported outcomes measures [41], technical alerts or measures which can help to better characterize patient trajectories. Thus, the aggregation of clinical, technical and sociologic data can help to improve personalized medicine in sleep apnea by considering individual variability [27].

REFERENCES

1. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med. 2019;7 8:687-98. doi:10.1016/S2213-2600(19)30198-5.
2. Bailly S, Grote L, Hedner J, Schiza S, McNicholas WT, Basoglu OK, et al. Clusters of sleep apnoea phenotypes: A large pan-European study from the European Sleep Apnoea Database (ESADA). Respirology. 2021;26 4:378-87. doi:10.1111/resp.13969.
3. Javaheri S, Barbe F, Campos-Rodriguez F, Dempsey JA, Khayat R, Javaheri S, et al. Sleep Apnea: Types, Mechanisms, and Clinical Cardiovascular Consequences. J Am Coll Cardiol. 2017;69 7:841-58. doi:10.1016/j.jacc.2016.11.069.
4. Levy P, Kohler M, McNicholas WT, Barbe F, McEvoy RD, Somers VK, et al. Obstructive sleep apnoea syndrome. Nat Rev Dis Primers. 2015;1:15015. doi:10.1038/nrdp.2015.15.
5. Marin JM, Carrizo SJ, Vicente E and Agusti AG. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet. 2005;365 9464:1046-53. doi:10.1016/S0140-6736(05)71141-7.
6. Pepin JL, Bailly S, Rinder P, Adler D, Benjafield AV, Lavergne F, et al. Relationship Between CPAP Termination and All-Cause Mortality: A French Nationwide Database Analysis. Chest. 2022; doi:10.1016/j.chest.2022.02.013.
7. Patil SP, Ayappa IA, Caples SM, Kimoff RJ, Patel SR and Harrod CG. Treatment of Adult Obstructive Sleep Apnea with Positive Airway Pressure: An American Academy of Sleep Medicine Clinical Practice Guideline. J Clin Sleep Med. 2019;15 2:335-43. doi:10.5664/jcsm.7640.
8. Bratton DJ, Stradling JR, Barbe F and Kohler M. Effect of CPAP on blood pressure in patients with minimally symptomatic obstructive sleep apnoea: a meta-analysis using individual patient data from four randomised controlled trials. Thorax. 2014;69 12:1128-35. doi:10.1136/thoraxjnl-2013-204993.
9. Dusart C, Andre S, Mettay T and Bruyneel M. Telemonitoring for the Follow-Up of Obstructive Sleep Apnea Patients Treated with CPAP: Accuracy and Impact on Therapy. Sensors (Basel). 2022;22 7 doi:10.3390/s22072782.

10. Pepin JL, Jullian-Desayes I, Sapene M, Treptow E, Joyeux-Faure M, Benmerad M, et al. Multimodal Remote Monitoring of High Cardiovascular Risk Patients With OSA Initiating CPAP: A Randomized Trial. Chest. 2019;155 4:730-9. doi:10.1016/j.chest.2018.11.007.
11. Tamisier R, Treptow E, Joyeux-Faure M, Levy P, Sapene M, Benmerad M, et al. Impact of a Multimodal Telemonitoring Intervention on CPAP Adherence in Symptomatic OSA and Low Cardiovascular Risk: A Randomized Controlled Trial. Chest. 2020;158 5:2136-45. doi:10.1016/j.chest.2020.05.613.
12. Fox N, Hirsch-Allen AJ, Goodfellow E, Wenner J, Fleetham J, Ryan CF, et al. The impact of a telemedicine monitoring system on positive airway pressure adherence in patients with obstructive sleep apnea: a randomized controlled trial. Sleep. 2012;35 4:477-81. doi:10.5665/sleep.1728.
13. Malhotra A, Crocker ME, Willes L, Kelly C, Lynch S and Benjafield AV. Patient Engagement Using New Technology to Improve Adherence to Positive Airway Pressure Therapy: A Retrospective Analysis. Chest. 2018;153 4:843-50. doi:10.1016/j.chest.2017.11.005.
14. Verbraecken J. Telemedicine in Sleep-Disordered Breathing: Expanding the Horizons. Sleep Med Clin. 2021;16 3:417-45. doi:10.1016/j.jsmc.2021.05.009.
15. McNicholas WT, Bassetti CL, Ferini-Strambi L, Pepin JL, Pevernagie D, Verbraecken J, et al. Challenges in obstructive sleep apnoea. Lancet Respir Med. 2018;6 3:170-2. doi:10.1016/S2213-2600(18)30059-6.
16. Pepin JL, Bailly S and Tamisier R. Big Data in sleep apnoea: Opportunities and challenges. Respirology. 2020;25 5:486-94. doi:10.1111/resp.13669.
17. Giampa SQC, Furlan SF, Freitas LS, Macedo TA, Lebkuchen A, Cardozo KHM, et al. Effects of CPAP on Metabolic Syndrome in Patients With OSA: A Randomized Trial. Chest. 2022; doi:10.1016/j.chest.2021.12.669.
18. Wohlgemuth WK, Chirinos DA, Domingo S and Wallace DM. Attempters, adherers, and non-adherers: latent profile analysis of CPAP use with correlates. Sleep Med. 2015;16 3:336-42. doi:10.1016/j.sleep.2014.08.013.
19. Aardoom JJ, Loheide-Niesmann L, Ossebaard HC and Riper H. Effectiveness of eHealth Interventions in Improving Treatment Adherence for Adults With Obstructive Sleep Apnea: Meta-Analytic Review. J Med Internet Res. 2020;22 2:e16972. doi:10.2196/16972.
20. Aloia MS, Goodwin MS, Velicer WF, Arnedt JT, Zimmerman M, Skrekas J, et al. Time series analysis of treatment adherence patterns in individuals with obstructive sleep apnea. Ann Behav Med. 2008;36 1:44-53. doi:10.1007/s12160-008-9052-9.
21. Bottaz-Bosson G, Hamon A, Pepin JL, Bailly S and Samson A. Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Frechet and dynamic time warping dissimilarities. Stat Med. 2021;40 24:5373-96. doi:10.1002/sim.9130.
22. Babbin SF, Velicer WF, Aloia MS and Kushida CA. Identifying Longitudinal Patterns for Individuals and Subgroups: An Example with Adherence to Treatment for Obstructive Sleep Apnea. Multivariate Behav Res. 2015;50 1:91-108. doi:10.1080/00273171.2014.958211.
23. Midelet A, Borel JC, Tamisier R, Le Hy R, Schaeffer MC, Daabek N, et al. Apnea-hypopnea index supplied by CPAP devices: time for standardization? Sleep Med. 2021;81:120-2. doi:10.1016/j.sleep.2021.02.019.
24. Vidigal TA, Brasil EL, Ferreira MN, Mello-Fujita LL, Moreira GA, Drager LF, et al. Proposed management model for the use of telemonitoring of adherence to positive airway pressure equipment - position paper of the Brazilian Association of Sleep Medicine - ABMS. Sleep Sci. 2021;14 Spec 1:31-40. doi:10.5935/1984-0063.20200086.

25. Eiseman NA, Westover MB, Ellenbogen JM and Bianchi MT. The impact of body posture and sleep stages on sleep apnea severity in adults. J Clin Sleep Med. 2012;8 6:655-66A. doi:10.5664/jcsm.2258.
26. Schwab RJ, Badr SM, Epstein LJ, Gay PC, Gozal D, Kohler M, et al. An official American Thoracic Society statement: continuous positive airway pressure adherence tracking systems. The optimal monitoring strategies and outcome measures in adults. Am J Respir Crit Care Med. 2013;188 5:613-20. doi:10.1164/rccm.201307-1282ST.
27. Patel SR, Bakker JP, Stitt CJ, Aloia MS and Nouraie SM. Age and Sex Disparities in Adherence to CPAP. Chest. 2021;159 1:382-9. doi:10.1016/j.chest.2020.07.017.
28. Borker PV, Carmona E, Essien UR, Saeed GJ, Nouraie SM, Bakker JP, et al. Neighborhoods with Greater Prevalence of Minority Residents Have Lower Continuous Positive Airway Pressure Adherence. Am J Respir Crit Care Med. 2021;204 3:339-46. doi:10.1164/rccm.202009-3685OC.
29. NG PDT, van den Heuvel ER, Aloia MS and Pauws SC. A latent-class heteroskedastic hurdle trajectory model: patterns of adherence in obstructive sleep apnea patients on CPAP therapy. BMC Med Res Methodol. 2021;21 1:269. doi:10.1186/s12874-021-01407-6.
30. Contal O, Poncin W, Vaudan S, De Lys A, Takahashi H, Bochet S, et al. One-Year Adherence to Continuous Positive Airway Pressure With Telemonitoring in Sleep Apnea Hypopnea Syndrome: A Randomized Controlled Trial. Front Med (Lausanne). 2021;8:626361. doi:10.3389/fmed.2021.626361.
31. Bertelli F, Suehs CM, Mallet JP, Court-Fortune I, Gagnadoux F, Borel JC, et al. Did COVID-19 impact Positive Airway Pressure adherence in 2020? A cross-sectional study of 8477 patients with sleep apnea. Respir Res. 2022;23 1:46. doi:10.1186/s12931-022-01969-z.
32. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M and Stork J. Comparison of different Methods for Univariate Time Series Imputation in R. arXiv:151003924 [cs, stat]. 2015.
33. Huque MH, Carlin JB, Simpson JA and Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol. 2018;18 1:168. doi:10.1186/s12874-018-0615-6.
34. Ong JC, Crawford MR and Wallace DM. Sleep Apnea and Insomnia: Emerging Evidence for Effective Clinical Management. Chest. 2021;159 5:2020-8. doi:10.1016/j.chest.2020.12.002.
35. Pieh C, Bach M, Popp R, Jara C, Cronlein T, Hajak G, et al. Insomnia symptoms influence CPAP compliance. Sleep Breath. 2013;17 1:99-104. doi:10.1007/s11325-012-0655-9.
36. Mokhlesi B, Finn LA, Hagen EW, Young T, Hla KM, Van Cauter E, et al. Obstructive sleep apnea during REM sleep and hypertension. results of the Wisconsin Sleep Cohort. Am J Respir Crit Care Med. 2014;190 10:1158-67. doi:10.1164/rccm.201406-1136OC.
37. Grimaldi D, Beccuti G, Touma C, Van Cauter E and Mokhlesi B. Association of obstructive sleep apnea in rapid eye movement sleep with reduced glycemic control in type 2 diabetes: therapeutic implications. Diabetes Care. 2014;37 2:355-63. doi:10.2337/dc13-0933.
38. Leger D, Elbaz M, Piednoir B, Carron A and Texereau J. Evaluation of the add-on NOWAPI(R) medical device for remote monitoring of compliance to Continuous Positive Airway Pressure and treatment efficacy in obstructive sleep apnea. Biomed Eng Online. 2016;15:26. doi:10.1186/s12938-016-0139-4.
39. Midelet A, Bailly S, Tamisier R, Borel JC, Baillieul S, Le Hy R, et al. Hidden Markov model segmentation to demarcate trajectories of residual apnoea-hypopnoea index in CPAP-treated sleep apnoea patients to personalize follow-up and prevent treatment failure. EPMA J. 2021;12 4:535-44. doi:10.1007/s13167-021-00264-z.

40. Bakker JP, Weaver TE, Parthasarathy S and Aloia MS. Adherence to CPAP: What Should We Be Aiming For, and How Can We Get There? Chest. 2019;155 6:1272-87. doi:10.1016/j.chest.2019.01.012.
41. Mehta N, Mandavia R, Patel A, Zhang H, Liu ZW, Kotecha B, et al. Patient-reported outcome measure for obstructive sleep apnea: Symptoms, Tiredness, Alertness, Mood and Psychosocial questionnaire: Preliminary results. J Sleep Res. 2020;29 2:e12960. doi:10.1111/jsr.12960.

Figure legends:

Figure 1:

Title: Daily computed indicators from positive airway pressure device records

Legend: PAP adherence, residual apnea-hypopnea index (AHI) and leaks

Figure 2:

Title: From positive airway pressure (PAP) prescription to PAP data visualisation: PAP data processing

Legend: **1**: PAP prescription from clinician to the patient.
2: PAP device distribution either directly from the manufacturer **(2a)** or from a home care provider **(2b)**.
3: real-time respiratory events detection by proprietary software embedded in PAP device.
4: data transfer from the PAP device to the manufacturer.
5: PAP data are stored in a secured online platform either from the manufacturer (**5a**) or through the home care provider (by using an Application Programming Interface (API)) who can make data transformation (**5b**).
6: clinician can visualize and download PAP data for their patient by using a web platform.
7: patient can visualize data on the device's screen or by using a web platform or a mobile application.

Figure 3:

Title: Illustration of the unreliability of the reported residual apnea-hypopnea index (AHI) in the presence of excessive leaks

Figure 4:

Title: Impact of device change in reported residual apnea-hypopnea index (AHI)

Legend: Dashed lines: mean residual AHI over time

Figure 5:

Title: Example of a situation where two positive airway pressure (PAP) devices are simultaneously assigned to a single patient

Legend: (A) Plot of the PAP usage of a patient who was assigned the records of two distinct devices. (B) Usage records to keep for analysis if the Device 2 was used by the patient after Device 1. (C) Usage records to keep for analysis if the Device 1 was used all over the period. (D) Usage records to keep for analysis if no information can be retrieved about which device was actually used by the patient after Device 1.

Figure 6:

Title: Management of missing data

Legend: A: Missing values are observed within a patient sequence. Either the length of the missing value is lower than the data storage capacity of the positive airway pressure (PAP) device without usage of another device (for example, during a hospitalization) or the patient provides a reason for non-use of any PAP (for example, vacation). In these cases, missing values can be replaced by 0.
B: Missing values are observed within a patient sequence and the patient does not provide a reason for non-use of PAP or the length of missing values is higher than PAP storage capacity: missing values cannot be replaced by 0. An imputation method may be considered.
C: Missing values are observed at the end of a patient sequence and three possibilities can be considered according to the information about missing values.

Figure 7:

Title: Change in mean positive airway pressure (PAP) individual average adherence over different number of months of therapy according to sample and missing value imputation strategy

Table 1: summary of problems, impacts and solutions

Problem to address	Potential impact	Proposed solution
Reliability of the residual AHI		
Unstable sleep structure might affect residual AHI and its central component	Under 2 hours of consecutive PAP use, the residual AHI may not be reliable or may be biased	Consider measures for more than 2 hours for analysis of residual AHI data
Drift of the flow curve due to important level of leaks	Respiratory events cannot be detected and the reported residual AHI can be distorted	Do not interpret residual AHI in case of excessive leaks
Device shift and absence of standardization		
Transmission bugs during device shift	Data may not be transmitted even though the subject uses the device	Missing values cannot be replaced by zero values
Data reported and summary statistics for leaks and residual AHI are different among PAP manufacturers	Impossible to compare directly measures from different devices	Transition periods between two distinct devices require specific data pre-processing
Multiple records of the same device for a patient		
Several records can be sent on the same day	Multiple rows for a patient with different values	Aggregate multiple rows by summing the usage values and computing the residual AHI weighted means
Mismatch between the machine ID and the patient ID	Assigning the records of two distinct devices to the same patient	Provide the device ID or remove the period with duplicate devices
Dataflow issue	Duplicate rows for a patient	Remove duplicate rows
Missing values		
Transmission with external modem	Missing values due to connecting issues	Remove data recorded by PAP device with external modem
Between two transmissions	Bias related to missing values	Missing values can be replaced by zero
At the end of a patient's sequence	Bias related to missing values	Imputation strategy should be defined
Treatment interruption		
With daily aggregated data there is no distinction whether the patient did three 2-hour naps or slept 6 hours consecutively at night	The same aggregated value of CPAP adherence might reflect completely different CPAP usages potentially related to CPAP termination and prognosis (e.g. Insomnia)	Use raw data which provide the exact time of CPAP exposures during night and day

Table 2: Impact of different strategies for imputing adherence missing values on mean adherence value and rate of patients with mean PAP use ≥4h/night. Results are presented for two different populations: the entire sample (N=407) and the sub-sample of patients with missing values.

	Over 1 month	Over 2 months	Over 3 Months
Number of missing values (mean and SD)	0.5 (2.6)	1.6 (5.9)	3.5 (11.2)
Maximum number of missing values	29	58	86
All sample			
No strategy for missing values			
Individual Average PAP Adherence hour/night (mean and SD)	4.61 (2.4)	4.62 (2.4)	4.6 (2.4)
PAP adherent patients* (N %)	257 (63)	255 (63)	255 (63)
Replacing missing values by zeros			
Individual Average PAP Adherence hour/night (mean and SD)	4.56 (2.4)	4.55 (2.4)	4.5 (2.4)
PAP adherent patients (N %)	252 (62)	249 (61)	251 (62)
Removing patients with missing values			
Individual Average PAP Adherence hour/night (mean and SD)	4.69 (2.4)	4.85 (2.3)	4.89 (2.3)
PAP adherent patients (N %)	239 (64)	230 (66)	221 (67)
Patients with missing values (N %)	33 (8)	60 (15)	78 (19)
No strategy for missing values			
Individual Average PAP Adherence hour/night (mean and SD)	3.67 (2.4)	3.29 (2.4)	3.36 (2.4)
PAP adherent patients (N %)	18 (55)	25 (42)	34 (44)
Replacing missing values by zeros			
Individual Average PAP Adherence hour/night (mean and SD)	3.0 (2.1)	2.78 (2.2)	2.86 (2.3)
PAP adherent patients (N %)	13 (39)	19 (32)	30 (38)

PAP: positive airway pressure; SD: standard deviation

*PAP adherent patients: number and percentage of patients with mean PAP use ≥4h/night

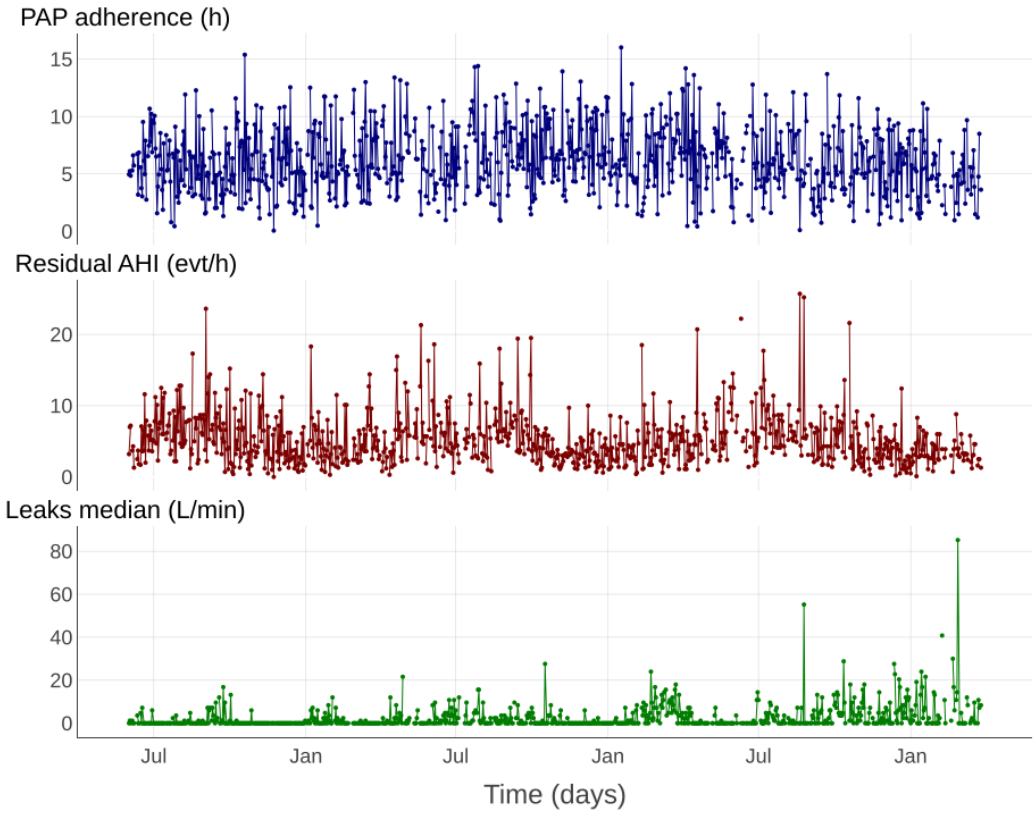


FIGURE 2.1 – Daily computed indicators from positive airway pressure device records.
PAP adherence, residual apnea-hypopnea index (AHI) and leaks.

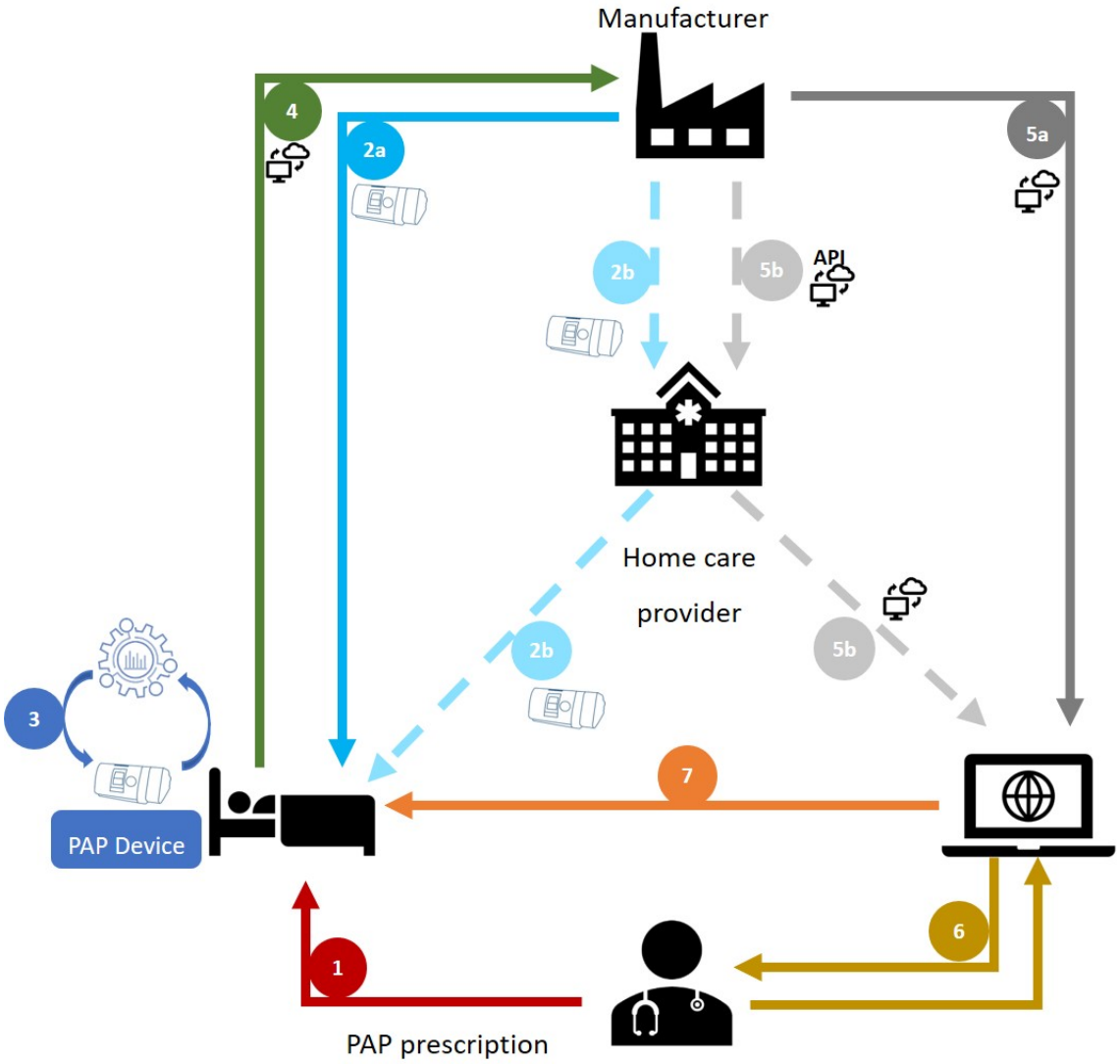


FIGURE 2.2 – From positive airway pressure (PAP) prescription to PAP data visualisation : PAP data processing.
1 : PAP prescription from clinician to the patient.
2 : PAP device distribution either directly from the manufacturer (2a) or from a home care provider (2b).
3 : real-time respiratory events detection by proprietary software embedded in PAP device.
4 : data transfer from the PAP device to the manufacturer.
5 : PAP data are stored in a secured online platform either from the manufacturer (5a) or through the home care provider (by using an Application Programming Interface (API)) who can make data transformation (5b).
6 : clinician can visualize and download PAP data for their patient by using a web platform.
7 : patient can visualize data on the device's screen or by using a web platform or a mobile application.

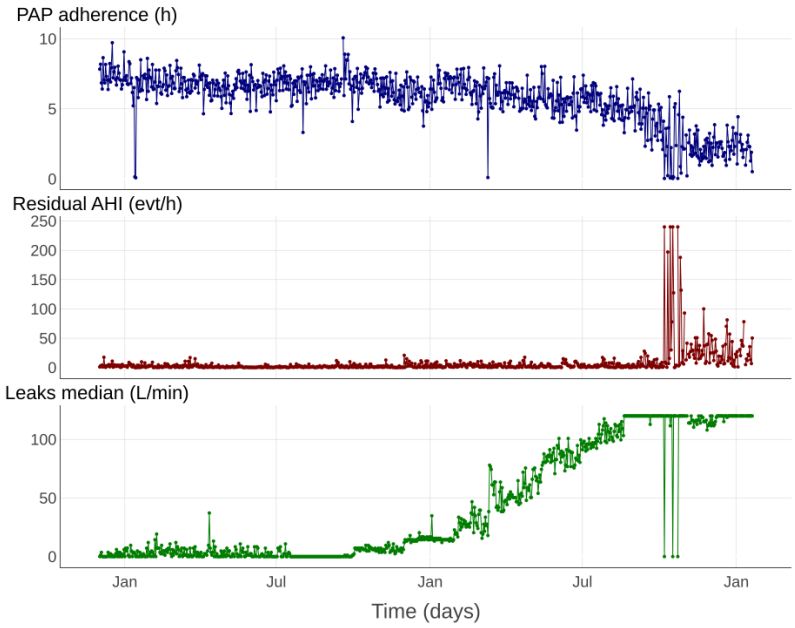


FIGURE 2.3 – Illustration of the unreliability of the reported residual apnea-hypopnea index (AHI) in the presence of excessive leaks.

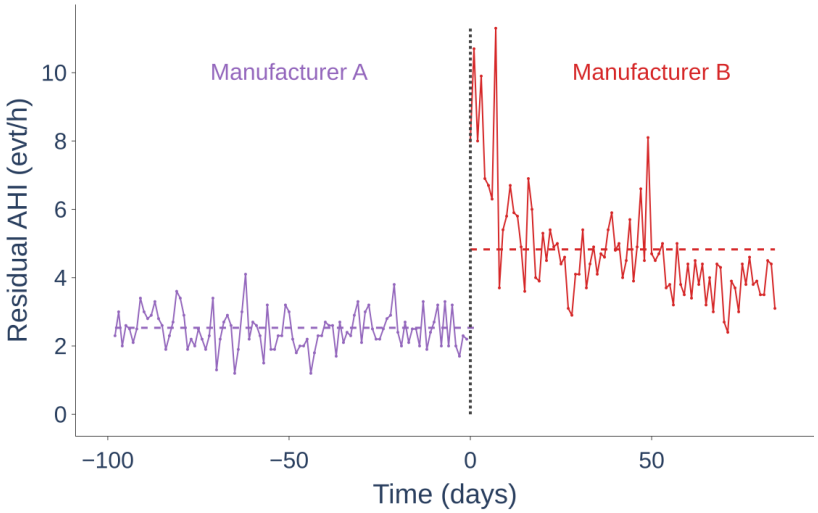


FIGURE 2.4 – Impact of device change in reported residual apnea-hypopnea index (AHI).
Dashed lines : mean residual AHI over time.

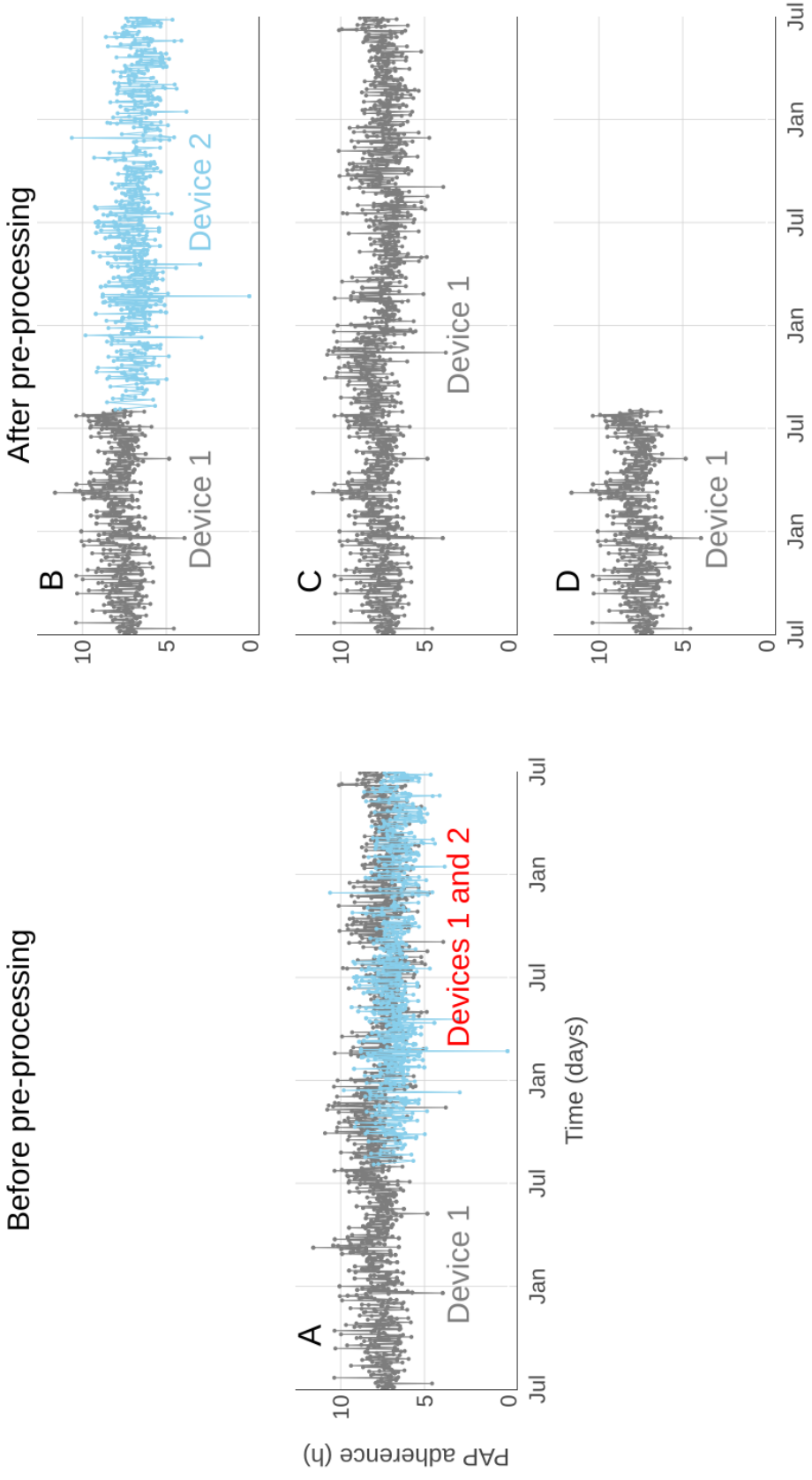


FIGURE 2.5 – Example of a situation where two positive airway pressure (PAP) devices are simultaneously assigned to a single patient.
(A) Plot of the PAP usage of a patient who was assigned the records of two distinct devices.
(B) Usage records to keep for analysis if the Device 2 was used by the patient after Device 1.
(C) Usage records to keep for analysis if the Device 1 was used all over the period.
(D) Usage records to keep for analysis if no information can be retrieved about which device was actually used by the patient after Device 1.

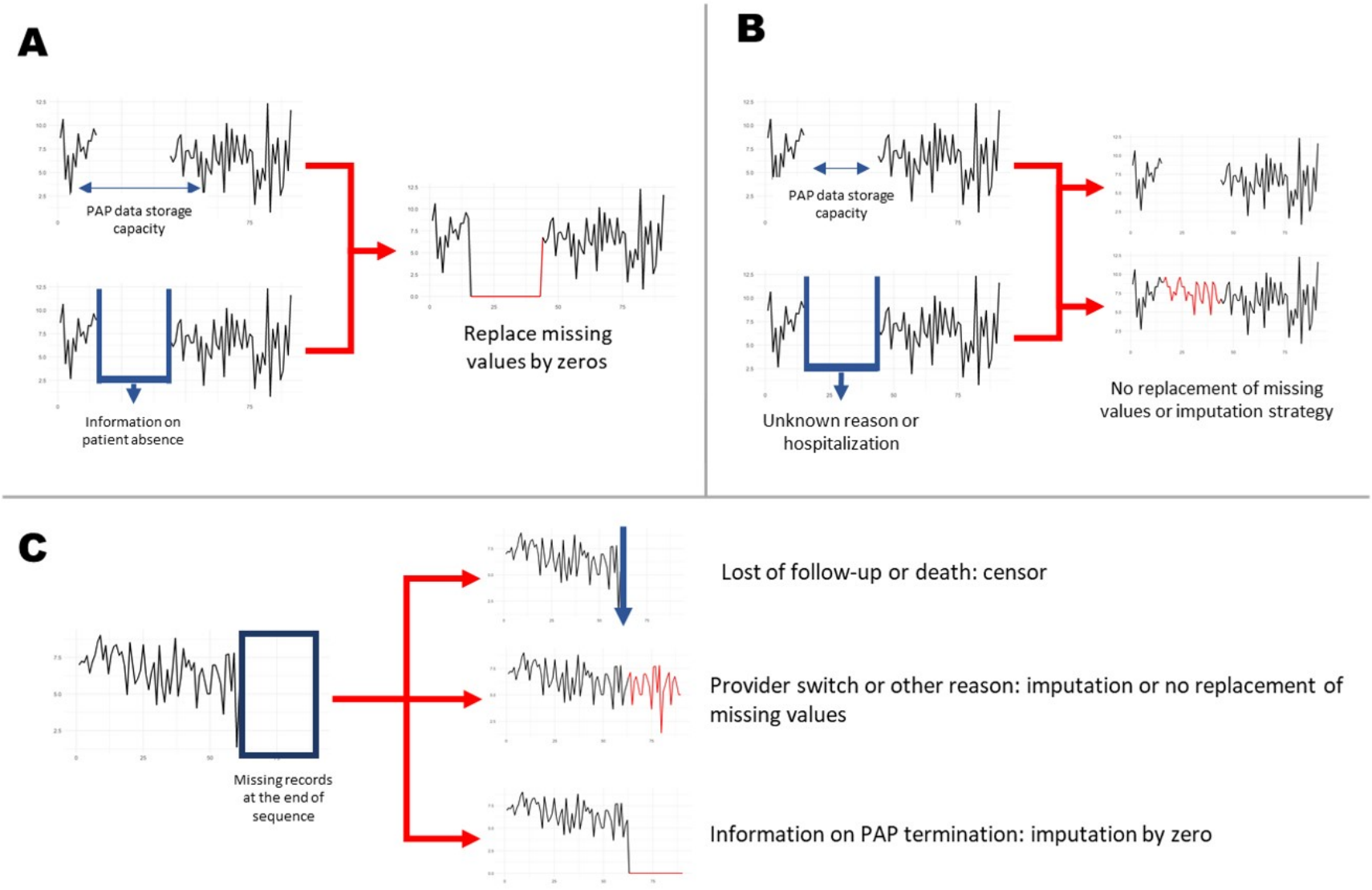


FIGURE 2.6 – Management of missing data.
A : Missing values are observed within a patient sequence. Either the length of the missing value is lower than the data storage capacity of the positive airway pressure (PAP) device without usage of another device (for example, during a hospitalization) or the patient provides a reason for non-use of any PAP (for example, vacation). In these cases, missing values can be replaced by 0.
B : Missing values are observed within a patient sequence and the patient does not provide a reason for non-use of PAP or the length of missing values is higher than PAP storage capacity : missing values cannot be replaced by 0. An imputation method may be considered.
C : Missing values are observed at the end of a patient sequence and three possibilities can be considered according to the information about missing values.

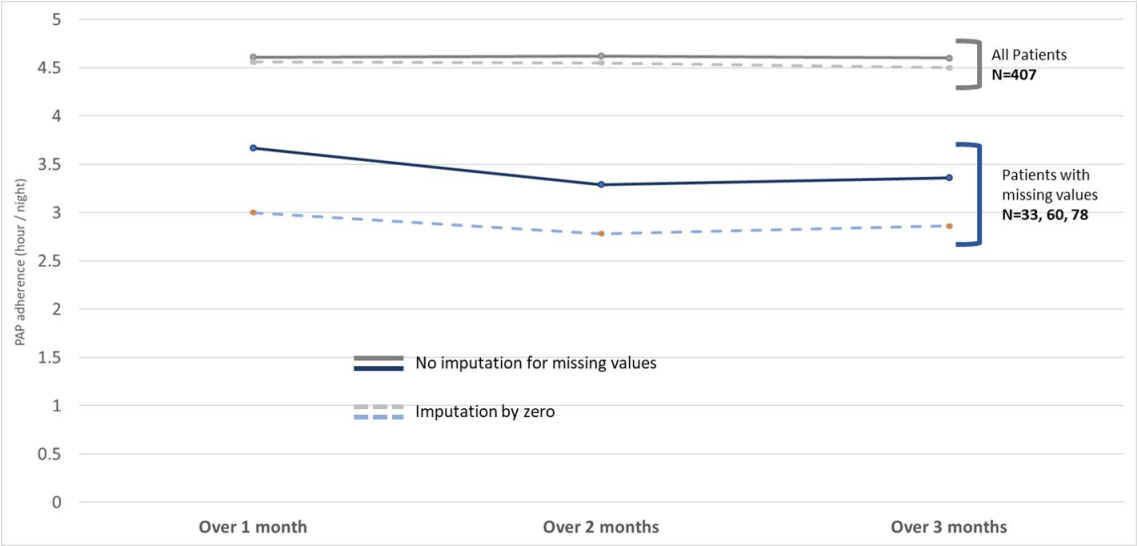


FIGURE 2.7 – Change in mean positive airway pressure (PAP) individual average adherence over different number of months of therapy according to sample and missing value imputation strategy.

2.3 Production des séquences d'observance de début de thérapie à la PPC

Au vu des éléments mentionnés dans l'article, nous identifions cinq points sur lesquels il convient d'être vigilant. La connaissance des différents appareils émettant les données pour chaque patient est nécessaire dans cette démarche.

1. Les données transmises par un modem externe sont liées à l'utilisation du modem et non pas à l'utilisation de la PPC. Il est nécessaire d'identifier les enregistrements qui sont produits directement par les appareils de PPC de ceux produits par des modems externes.
2. Il convient de distinguer une non utilisation de l'appareil pour un patient (un zéro) et une non transmission des données par la machine (créant une donnée manquante). Normalement une non utilisation de l'appareil entraine une non transmission de l'observance, et donc une non transmission est assimilée à une non utilisation de la machine. Or certaines valeurs sont déjà à zéro. De plus certains modèles d'appareils sont davantage affectés par des pannes de transmission et de recouvrement des données. Il est conseillé d'explorer les données pour détecter les modèles suspects de non transmission en dépit de l'utilisation de la PPC.
3. Il faut prêter attention aux patients ayant des données rattachées à plusieurs appareils de PPC. Certains patients peuvent être concernés par un ou plusieurs changement(s) de machine, nécessitant de prendre des précautions. Il peut aussi exister des incohérences de correspondance entre les patients et les machines dans la base de données du PSAD.
4. En dehors des situations où des patients sont individuellement associés à plusieurs appareils, de multiples enregistrements concernant une même date pour

un même patient peuvent être transmis. Si les doublons se résument en un seul enregistrement, les autres situations doivent être convenablement gérées car elles peuvent révéler des problèmes techniques et/ou des incohérences dans les données reçues.

5. Il reste à imputer les valeurs manquantes, dont une partie correspond aux inutilisations des appareils. Cette imputation doit se faire pour chaque patient. Elle doit tenir compte des informations à disposition pouvant expliquer les non transmissions des données (dates d'appareillage et de désapareillage le cas échéant, vacances, hospitalisation, utilisation d'une orthèse d'avancée mandibulaire en substitut de la PPC), de la position des non transmissions dans la trajectoire d'observance (entre d'autres transmissions ou en fin de trajectoire), et de la question clinique sous-jacente (évaluation des déterminants de l'observance ou des effets de la thérapie).

Réaliser le retraitement des données issues des machines n'est pas suffisant pour délimiter les séquences d'observance analysables. D'autres informations sont requises pour sélectionner un échantillon d'individus compatible avec les contraintes de la question clinique motivant leur production. En particulier nous souhaitons étudier les sujets au début de leur thérapie. Aussi il convient de faire attention aux traitements effectivement dispensés par les appareils de PPC. Certaines machines peuvent fournir une oxygénothérapie. Pour les patients concernés, les enjeux de leur adhésion vont au delà de ceux seulement liés au SAOS et d'autres comorbidités pourraient impacter leur pronostic. Ces patients devraient être écartés des études s'intéressant à l'observance thérapeutique pour un SAOS.

D'autres données peuvent être récoltées par les PSAD, à l'initiation du traitement, à l'issue de rendez-vous de suivi, ou auprès des fabricants. Elles peuvent porter sur le patient, sa/ses maladie(s), ainsi que sur les matériels thérapeutiques comme les réglages des machines ou les masques fournis. Ces métadonnées sont indispensables pour permettre la sélection des individus.

2.4 Discussion et perspectives

La production des séquences d'observance à partir des données générées par les machines de PPC est une étape critique pour tirer parti des opportunités offertes par le télésuivi. Les éléments abordés dans ce chapitre ne sont pas exhaustifs pour optimiser la qualité des données au regard des questions cliniques à adresser. Quelques pistes sont évoquées ci-après, d'une part concernant la fiabilité des séquences par rapport aux trajectoires individuelles à décrire, et d'autre part pour ouvrir sur la sélection des données pour la problématique et les questions cliniques.

Nous commençons par les remarques sur la fiabilité des séquences individuelles. Nous avons supposé que la variable "durée quotidienne de port du masque lorsque la machine administre une pression thérapeutique" est correctement estimée par les fabricants. Une manière de vérifier la fiabilité de l'indicateur reporté est de comparer l'homogénéité des distributions de la variable selon les constructeurs. En cas de différence, l'indicateur pourrait être uniformisé à partir des pressions théoriques et des signaux bruts de débit et pression mesurés à l'entrée des masques. Les données des patients changeant de machine seraient précieuses pour réaliser des comparaisons appariées, à condition d'identifier et d'isoler au maximum les variations de l'ob-

servance dues à d'autres causes explicatives (période d'adaptation aux nouveaux appareils et consommables, changements connus dans les habitudes de sommeil).

Cela conduit au deuxième point pour améliorer la qualité des séquences, qui est d'utiliser un maximum d'informations pour expliquer et contextualiser les trajectoires d'observance des patients. Certaines métadonnées collectées par les PSAD peuvent être stockées dans des champs de texte de compte-rendus de rendez-vous. Des approches basées sur de la fouille de texte pourraient permettre d'identifier des événements notoires pour la productions des séquences. En revanche, il peut être difficile à partir des seules données du PSAD, de définir les patients qui n'ont jamais été traités antérieurement avec un autre PSAD. Il peut être nécessaire de recourir à d'autres données.

Le dernier point d'amélioration des séquences individuelles concerne l'imputation des valeurs manquantes qui ne sont pas liées à des non-utilisations des machines. Cela permettrait de contribuer à la constitution d'un nombre plus grand de séquences complètes. Parmi les méthodes classiques telles que l'interpolation par des splines, l'imputation par la moyenne, ou encore le report de valeurs directement antérieures ou postérieures, aucune ne parait adaptée pour restituer la variabilité caractéristique des comportements d'observance. Une possibilité est de procéder à ces imputations avec des méthodes basées sur des modélisations. Ceci requiert d'avoir avancé sur l'analyse des séquences actuellement disponibles.

Additionnellement à ces éléments, les aspects de pertinence et de représentativité des données étudiées doivent être réfléchis selon la question clinique d'analyse des comportements d'observance. Concernant la pertinence de la variable "durée de port du masque", deux situations peuvent être distinguées selon les questions cliniques. Premièrement, lorsque l'objectif est d'évaluer les effets de la thérapie, cette variable est pertinente mais incomplète. L'IAH résiduel, qui indique l'efficacité du traitement en termes de réduction des évènements respiratoires, également reporté de manière quotidienne, devrait également être pris en compte. Cependant, comme détaillé dans la Section 2.1, l'indicateur est reconnu pour être soumis à plusieurs sources d'inexactitude dans son estimation. Analyser les séquences d'IAH résiduels quotidiens associées aux séquences d'observance pourrait drastiquement limiter le nombre de sujets dans les études si l'on souhaite conserver un niveau équivalent de qualité des données. Cela appuie l'intérêt de recourir à l'analyse des signaux bruts pour au moins contrecarrer le manque de standardisation entre les fabricants.

Secondairement, sur la pertinence de la durée de port du masque, lorsque l'objectif est d'identifier les déterminants de l'observance, il faut prendre en compte qu'elle est limitée par le temps de sommeil, qui est difficile d'accès de manière objective. De plus la durée de port du masque peut être diminuée en dépit de la volonté des patients. Un patient avec des congestions nasales aura plus de difficultés à se traiter s'il dispose d'un masque nasal. Le cas de la chute du masque au cours du sommeil est un autre exemple. Il faudrait s'assurer que le PSAD ait été réactif pour solutionner les problèmes auprès des patients. Si oui, les réductions accidentelles de l'observance peuvent être intégrées dans la variabilité de l'observance. Sinon, il faut analyser les données en conséquence, par exemple en écartant les sujets ou en imputant les valeurs concernées s'il est possible de les identifier.

Cela ouvre sur les questions de représentativité des données analysables car les PSAD peuvent avoir des pratiques différentes d'accompagnement de leurs patients vers une bonne observance. En conséquence, analyser les comportements d'obser-

vance sur un échantillon de patients traités par un même PSAD peut mettre en avant des comportements spécifiques. De plus, puisque les pratiques d'accompagnement des PSAD peuvent évoluer, la récence des données analysées peut impacter les comportements typiques identifiés. Dans tous les cas il conviendra d'avoir conscience de la qualité de la prise en charge associée aux comportements typiques identifiés si l'on souhaite personnaliser l'accompagnement thérapeutique d'un patient.
Ensuite, pour obtenir des résultats représentatifs de l'ensemble de la population des sujets traités par PPC, il faut intégrer des sujets de différentes localités. Les comportements d'observance individuels pourraient être influencés par des caractéristiques socio-démographiques et économiques sur-représentées dans une zone géographique donnée. De plus, des facteurs environnementaux tels que les niveaux de pollution pourraient interagir avec l'effet de la thérapie.
Enfin pour terminer sur les aspects liés à la représentativité de l'échantillon analysé, il faut prendre en compte que certains patients refusent le télésuivi. Si ce choix peut raisonnablement être considéré indépendant de l'effet de la thérapie, il peut en revanche être lié aux motivations psychologiques d'un patient à se traiter. La non présence de ces sujets dans l'échantillon analysé est adéquate pour les travaux visant à prédire l'adhésion à partir des séquences d'observance du début de traitement car les cliniciens n'ont pas le même accès aux données d'observance de cette sous-population. Cependant, si la question clinique est d'identifier des déterminants de l'observance antérieurs à la mise en place du traitement, cette sous-population devrait être étudiée pour déceler d'éventuelles spécificités.

Ces perspectives placent les PSAD comme des acteurs importants de la collecte des données, de leur qualité, mais aussi comme ayant une influence sur les comportements d'observance de leurs patients, ainsi que sur la qualité de la thérapie. Cela renforce l'importance de la relation entre cliniciens et PSAD tant dans la prise en charge du SAOS que dans la correcte valorisation scientifique des données du télésuivi de la PPC. Les contraintes de qualité sur la production des séquences limitent le nombre de sujets intégrables dans les études sur les comportements d'observance de début de thérapie. Cependant ce nombre de sujets est voué à augmenter au fur et à mesure des nouvelles prescriptions de PPC, avec des données de plus en plus fiables par exemple via la généralisation de l'utilisation des machines connectées. Les aspects sur la sécurité des données et leur accessibilité ont volontairement été mis de côté car ils impliquent surtout des considérations techniques et réglementaires, qui vont au delà du cadre statistique des travaux décrits dans cette thèse. Ils ne contribuent pas directement à rendre les données moins biaisées. Cependant, ils peuvent être des freins techniques à l'optimisation de l'actualité des données, ou restreindre légalement les possibilités d'accès et d'utilisation des données contextuelles et cliniques associées aux trajectoires d'observance.

Jusque là, dans le mémoire, nous avons distingué les termes "séquence" et "trajectoire" d'observance pour insister sur la différence entre un comportement individuel d'observance (i.e. la trajectoire d'observance) et les données qui le décrivent (i.e. la séquence d'observance). Les aspects de qualité des données étant traités dans ce chapitre, dans la suite nous emploierons "trajectoire d'observance" pour décrire autant un comportement individuel d'observance que sa séquence de données associée.

Chapitre 3

Classification des trajectoires individuelles

3.1 Introduction

Ce chapitre rapporte les premiers travaux réalisés dans le cadre de la présente thèse. L'objectif de ces travaux était d'identifier une méthodologie de clustering basée sur l'utilisation d'une mesure de dissimilarité applicable directement entre séries chronologiques. Nous avons considéré trois mesures de dissimilarité : la distance euclidienne, la dissimilarité DTW [54] et sdF une variante sommée de la distance de Fréchet discrète [55]. Nous avons implémenté ces dissimilarités en simulation dans le cadre de la CAH avec trois critères d'agrégation : le critère de Ward, le critère du diamètre et le critère de la dissimilarité moyenne. La question du choix du nombre de groupes est prise en compte avec la comparaison de six indices de validation de classification internes sur leur capacité à choisir des partitions proches des groupes simulés. Nous comparons les indices selon la méthodologie de Gurrutxaga et al. [57].

Ce travail a fait l'objet d'une publication dans la revue "Statistics in Medicine" qui constitue l'intégralité de ce chapitre. La publication présente l'étude de simulation ainsi que l'interprétation de clusters obtenus sur données réelles. Conformément aux recommandations des relecteurs, nous avons souhaité mettre en avant les apports de l'utilisation de DTW pour réaliser la classification de trajectoires médicales au delà du cadre de l'observance à la PPC.

Puisque l'article reporte le premier travail de cette thèse et qu'il est déjà publié, il y a des maladresses de rédaction et des corrections mineures à apporter, que nous précisons ci après par ordre d'apparition dans l'article.
L'utilisation du terme "partitionnal" à propos des méthodes de classifications réfère plus précisément aux méthodes dites de "centres mobiles".
Dans la section sur les données, nous indiquons que les machines ne gèrent pas identiquement les données manquantes et la génération des zéros d'observances. C'est incomplet car il y a également les retraitements opérés par les fabricants des machines et par le PSAD qui peuvent différer. Le terme "sample" serait plus adapté que le terme "population" dans le paragraphe "Population selection".
L'implémentation de l'indice de Calinski-Harabasz [58] proposée dans la version utilisée du package R "dtwclust" [59] est erronée. Le calcul correctement exécuté améliore très légèrement les performances de l'indice, mais ne permet pas à l'indice d'obtenir de meilleures performances que l'indice de Dunn [56].

La définition donnée du centroïde dans la section 3.4 ne correspond pas exactement à celle de la moyenne arithmétique puisqu'il faudrait prendre la distance au carré dans la formule.

Nous ne présentons pas les résultats du critère du saut minimal de la CAH, ni de l'algorithme des k-médoïdes. Ils ont été considérés dans le design de l'étude de simulation mais ont obtenu de mauvaises performances comparativement aux autres méthodes de classification.

Les définitions des dissimilarité et critères d'agrégation utilisés sont données en annexe B de l'article. Les notations utilisées dans l'article diffèrent légèrement de celles employées en Section 1.3 Les principales divergences sont :

— l'utilisation de la lettre majuscule X à la place de x pour désigner l'ensemble des séries chronologiques à classer ;

— l'utilisation de la lettre majuscule N à la place de n pour désigner le nombre de séries de l'échantillon ;

— l'utilisation de la lettre majuscule T à la place de t pour désigner la longueur commune des séries chronologiques ;

— la lettre minuscule t devient la variable muette parcourant l'ensemble des indices temporels des séries ;

— et enfin la lettre j est employée à la place de la lettre i' pour désigner l'indice d'une série chronologique quelconque.

3.2 Article

Received: 15 June 2020 | Revised: 6 May 2021 | Accepted: 23 June 2021

DOI: 10.1002/sim.9130

RESEARCH ARTICLE

Statistics in Medicine WILEY

Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities

Guillaume Bottaz-Bosson[1,2] | **Agnès Hamon**[2] | **Jean-Louis Pépin**[1] | **Sébastien Bailly**[1] | **Adeline Samson**[2]

[1]Laboratoire HP2, Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, Grenoble, France

[2]LJK, Univ. Grenoble Alpes, CNRS, Grenoble, France

Correspondence
Guillaume Bottaz-Bosson, Laboratoire HP2, Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, 38000 Grenoble, France.
Email: guillaume.bottaz-bosson@univ-grenoble-alpes.fr

Background: Obstructive sleep apnea (OSA) is a chronic disease characterized by recurrent pharyngeal collapses during sleep. In most severe cases, continuous positive airway pressure (CPAP) consists in keeping the airways open by administering mild air pressure. This treatment faces adherence issues.

Objectives: Eight hundred and forty-eight subjects were equipped with CPAP prescribed at the Grenoble University Hospital between 2016 and 2018. Their daily CPAP uses have been recorded during the first 3 months. Our aim is to cluster these adherence time series. With hierarchical agglomerative clustering, we focused on the choices of the dissimilarity measure and the internal cluster validation index (CVI).

Methods: The Euclidean distance, the dynamic time warping (DTW) and the generalized summed discrete Fréchet dissimilarity were implemented with three linkage strategies ("average," "complete," and "Ward"). The performances of each method (dissimilarity and linkage) were evaluated on a simulation study through the adjusted Rand index (ARI). The Ward linkage with DTW dissimilarity provided the best ARI. Then six different internal CVIs (Silhouette, Calinski Harabasz, Davies Bouldin, Modified Davies Bouldin, Dunn, and COP) were compared on their ability to choose the best number of clusters. The Dunn index beat the others.

Results: CPAP data were clustered with the Ward linkage, the DTW dissimilarity and the Dunn index. It identified six clusters, from a cluster of patients (N = 29 subjects) whose stopped the therapy early on to a cluster (N = 105) with increasing adherence over time. Other clusters were extremely good users (N = 151), good users (N = 150), moderate users (N = 235), and poor adherers (N = 178).

KEYWORDS

cluster validation, CPAP adherence trajectories, discrete Fréchet distance, dynamic time warping, time series clustering

1 | INTRODUCTION

1.1 | Clinical context and rationale

Obstructive sleep apnea (OSA) is one of the most common chronic diseases affecting almost 1 billion people worldwide.[1] OSA is characterized by repeated episodes of complete (apneas) and/or incomplete (hypopneas) pharyngeal collapse during sleep producing intermittent hypoxia and sleep fragmentation which in turn generate disturbing symptoms including daytime sleepiness, impairment of daily functioning, deterioration of memory and cognition, and a higher risk of developing cardiovascular, metabolic, and cerebrovascular disease.[2]

Continuous positive airway pressure (CPAP) is the first line treatment of moderate to severe OSA. CPAP reopens and stabilizes the upper airway and allows the complete suppression of abnormal respiratory events during sleep. The therapy is highly effective in improving quality of life and suppressing symptoms, but adherence remains challenging. A period of 3 to 6 months after CPAP initiation is necessary for some patients to stabilize their adherence due to the need to get used to the mask interface, to the pressure, to using the CPAP device, and adapt to the side effects. The initial refusal rate by patients is close to 15% even when proposed by experienced teams, and long term treatment discontinuation is estimated at between 20% and 35%.[2] The majority of dropouts happens in the first 3 months.[3] While variable from one country to another, a minimal adherence to CPAP is required for CPAP treatment reimbursement (eg, in the United States more than 4 hours of usage per night for more than 70% of nights). There is a dose response relationship between CPAP adherence and the degree of improvement in symptoms and related quality of life.[4] The potential for cardio-metabolic risk reduction is also highly dependent on CPAP adherence levels.[2]

A unique specificity of OSA is that data are generated every night by telemonitoring of CPAP devices in millions of patients providing objective daily measurements of adherence.[5] Interventions to improve CPAP adherence have included educational, supportive, and behavioral strategies[6] or technical CPAP innovations to reduce device-related side effects. When implemented separately, these approaches have only had a limited impact on CPAP adherence[6] and recent strategies have aimed at combining information from remote home monitoring of CPAP use and patient coaching. In depth detailed analysis of CPAP telemonitoring data is a crucial step toward describing the different patterns of CPAP adherence. This is a prerequisite to identifying patients at risk of poor CPAP adherence and to proposing personalized follow-up adapted to these profiles. The main goal of the present work is to propose statistical tools to delineate the different trajectories of CPAP adherence during the first three months of follow-up.

1.2 | Methodological approaches

Every night CPAP adherence data available from telemonitoring allow us to construct chronological diagrams of individualized trajectories of adherence. Here we employ a time series clustering approach, to describe the typical patterns and reveal the most illustrative CPAP adherence trajectories. Clustering algorithms are generally designed to deal with static data and therefore are not completely appropriate to consider the temporal structure of time series. Transforming time series into static data (by estimating model parameters or feature extraction) is one method. Customizing the algorithms is another. Moreover there are various clustering algorithms that can produce different results from the same data. Furthermore, a key issue is to evaluate the goodness of the resulting partition as there is no prior information about the data. Consequently, it is difficult to both choose an appropriate clustering method and to select the number of clusters in which to split the data. A first distinction among time series clustering methods is between model-based and shape-based approaches. Model-based clustering assumes a specific model for each cluster, and suitable model distance measures and algorithms are applied. The main drawback is that the results rely on the model's assumptions. Conversely, shape-based methods do not require any assumptions and work on raw data by using an appropriate dissimilarity measure. In the present study, we preferred a shape-based approach because of the small amount of prior knowledge available on CPAP adherence trajectories. In these shape-based approaches, the two main unsupervised classification algorithms are partitional and hierarchical clustering. Both rely on the choice of a dissimilarity measure. The Euclidean distance is frequently used with these algorithms but is not specific to the context of time series clustering, contrary to the dynamic time warping (DTW) dissimilarity which considers shapes differing only in temporal shift as being similar. Along the same line, Genolini et al[7] introduced the generalized discrete Fréchet distance that comports a time scale parameter to take account of the time shifts in the dissimilarity value. We propose a variant of this dissimilarity called the "generalized summed discrete Fréchet dissimilarity" (sdF) which is a generalization of the DTW including a time scale parameter. In this context,

we consider both DTW and sdF dissimilarities because time shifts are not important from the medical point of view. For example, when we want to consider all patients who abandon their CPAP therapy as being in the same group, whatever the duration between CPAP initiation and treatment surrender. We are interested in how these dissimilarity measures identify shapes despite the particularities of CPAP trajectories, from recurrent discontinuities due to the random occurrence of zero use to high variability hiding central tendency. These particularities are of interest when describing adherence behaviors and they motivate the choice to cluster the data without applying any smoothing method. However, the relevance of the dissimilarities are dependent on the clustering algorithm used. Partitional clustering requires the choice of a centroid algorithm and setting the required number of clusters (k). The initialization step randomly spreads the data into k clusters and computes centroids, called cluster centers. Then the algorithm alternates between reallocating the objects to the clusters with the nearest cluster centers and computation of the new cluster centers (centroids) until a quite stable partition is obtained. Two notable variants of this algorithm are K-medoids and K-means. K-medoids work with arbitrary dissimilarities and use medoids as centroids. K-means are clustering using the Euclidean distance and the centroids are defined as the arithmetic mean between the points in a cluster. DTW and sdF dissimilarities can be used in partitional clustering with suitable centroids (see Section 3.4), but involve high computation cost, and a new calculation is required at each iteration, so the overall computational cost of the algorithm is strongly impacted. Furthermore, the resulting clusters are dependent on the initialization step and so in practice, for a chosen k, several clusterings are done and compared before validating the final partition. Considering the necessity to launch several executions, for both initialization reasons and the k-parameter input, we excluded partitioning algorithms from this study so as to concentrate on hierarchical approaches. Thus we focus on hierarchical agglomerative clustering (HAC): after the computation of a pairwise dissimilarity matrix, each object is placed in an individual cluster which are then gradually merged through a linkage strategy. Then a dendrogram presents the hierarchy of clusters and the resulting partition is obtained after cutting it at the desired height or with the adopted number of clusters. Optionally, centroids are only computed once during this final step.

In the case of time series clustering, as for classical clustering, the evaluation of the resulting clusters and the choice of cluster number remain open questions. Many cluster validity indices (CVIs) exist in the literature. They are split into two categories: external and internal indices. The first ones evaluate the similarity between two partitions of a same set. They judge the quality of a partition when, for example, the true partition is known. On the other hand, internal indices can measure the goodness of a partition, without external information. Each of them gives a different sense of what is a "good" partition. Gurrutxaga et al[8] proposed a methodology to compare internal CVIs. One novelty of our work is to transpose this approach to the case of time series clustering, which has not been done yet.

To recap, we focused on HAC where we needed first to select a dissimilarity measure and then to validate a CVI to choose the number of clusters. These two choices motivated the current study and this article is organized as follows. In the next section, we present the data (Section 2). Section 3 describes several alternatives we considered in the framework of the approach described above, and they were compared in a simulation study (Section 4). Then, in Section 5, we describe the clusters resulting after application of the selected clustering procedure to real data. Section 6 concludes the article with some discussion and potential applications of these method to time series produced in other fields of medicine.

2 | DATA

2.1 | Confusion between missing and null values

A problem with CPAP adherence data is the possible confusion between missing and null values. Missing data is the absence of an adherence value whereas a null value means the patient did not use her/his device. As a missing value can be due to a technical problem with the CPAP machine, or due to the fact a patient did not power her/his device, it is difficult to differentiate null and missing values. Different machines do not manage these situations in the same way so these values must be interpreted for each different CPAP model.

2.2 | Population selection

Our data included N = 1831 subjects who started a CPAP therapy between January 2016 and January 2018 prescribed by the Grenoble-Alpes University Hospital (France). The adherence follow-up was launched within 15 days after treatment

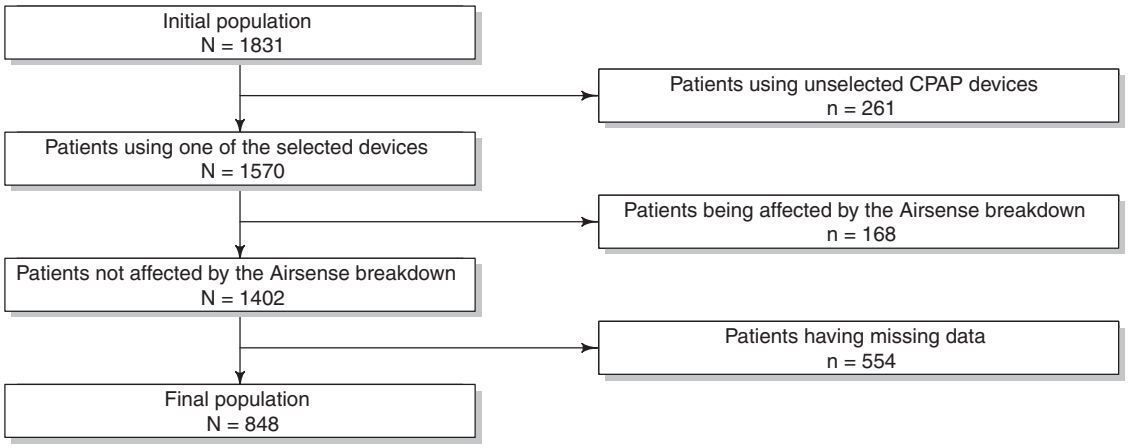


FIGURE 1 Real data: Subject inclusion flowchart

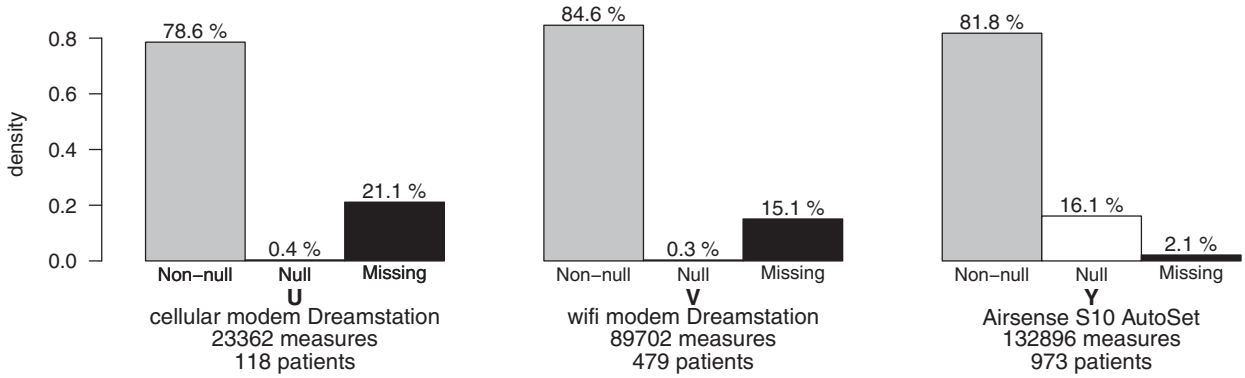


FIGURE 2 Real CPAP data: Distribution of adherence values (non-null, null, and missing) for the 3 selected CPAP devices

prescription. Figure 1 resumes the flowchart of patient inclusion in the study. These subjects did not change their CPAP device and do not have days with multiple measures during the period between 2016 and 2018.

A step preceding selection of the population to be analyzed was to distinguish missing from null values for each CPAP model. Therefore we considered only models of CPAP devices used by at least 100 subjects and we excluded CPAP devices that are unable to transmit data without being connected to an external modem. Indeed, for the latter, the data corresponds to modem utilization, which can be used without the CPAP device or inversely the device can be used without the modem. Three devices used by 1570 subjects were retained. We call these devices U, V, and Y where U and V are Dreamstation devices (Philips) with respectively internal cellular modems or dependent on WiFi, and Y is the Airsense S10 Autoset device (ResMed). Figure 2 shows the distribution of missing, null, and non-null adherence data, for the three selected devices, confirming that the meaning of "missing" and "null" differs according to the device. For machines U and V, and in view of the low zero rates, we considered that if a subject did not use her/his CPAP, the device remained off and so there was no data transmitted. Missing data for these devices were then replaced by null adherence. For the device Y, data were reliable as the modem sent a "zero" even if the patient did not use her/his device.

To simplify the statistical analysis, we considered trajectories without missing values and the length of the time series was fixed at 91 days. A breakdown occurred with the Y device between June 9, 2017 and June 18, 2017. This affected the transmission of adherence data of 168 subjects, who we excluded. Finally we excluded 554 more subjects with missing values, to obtain a final population of 848 patients with complete trajectories.

Table 3 briefly describes this population. They were middle aged (75% over 50 years), predominantly male (63%) and with a median body mass index of 30.4 kg/m^2. This table also summarizes individual adherence characteristics with individual means, standard deviations, rates of null values, rates of use for over 4 hours. The distributions of these characteristics are shown in the histograms in Appendix A (Figure A1). Figure 3 shows 8 examples of individual

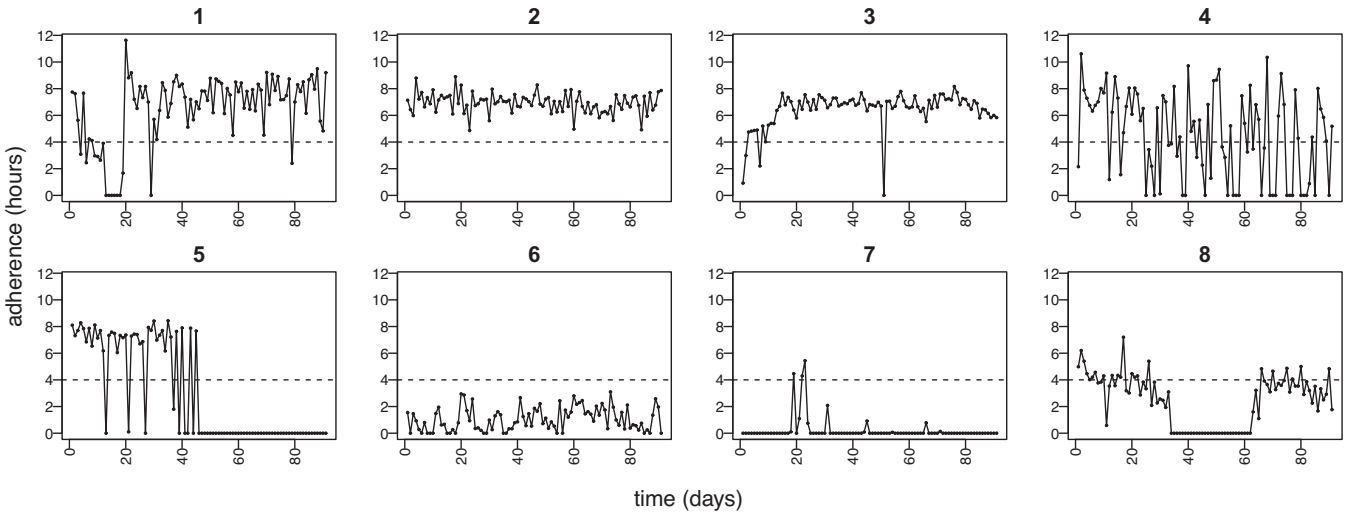


FIGURE 3 CPAP adherence time series during the first 3 months of therapy of 8 patients. The dashed line represents the efficiency threshold fixed at 4 hours of daily CPAP use

adherence trajectories, including subjects having good adherence trajectories (subject 2) patients with occasional or insufficient CPAP use (subjects 6 and 7) and patients who completely stopped CPAP (subject 5). High variability and random occurrences of zeros characterize these CPAP time series.

We remind readers that the clinical objective of this work was to find typical CPAP adherence profiles by clustering the data. This is a first step towards improving the personalized care of patients with sleep apnea. Several alternatives methods to cluster CPAP data in the framework of HAC are detailed in the next section.

3 | CLUSTERING PROCEDURE

Clustering time series using the HAC algorithm involves the computation of a pairwise dissimilarity matrix, a linkage strategy to define distances between clusters and the choice of the number of clusters. This choice can be made after inspecting the dendrogram node heights, but also thanks to internal CVIs. We focused on this second approach because it is more objective, although sometimes it can be dependent on the calculation of cluster centroids.

3.1 | Some dissimilarity measures between time series

Here, we describe the three dissimilarity measures considered in this study. Formal mathematical definitions are given in Appendix B.2.

The Euclidean distance computes the pointwise difference between the two time series. It is the square root of the summed daily squared differences (see Definition 3). Figure 4 illustrates the three dissimilarity measures on two fictive trajectories. The two trajectories have the same pattern with a time shift of one unit and differing from one unit on the y axis. This measure is fast to compute and compatible with the arithmetic mean used as a centroïd, making it easy to implement in all algorithms. However it is not specific to time series, contrary to DTW for example (see Definition 4).

This last measure is frequently used in time series clustering.[9] It is said to be "elastic" because the time series are warped and two patterns differing only in a time shift are thus considered as similar. The main drawbacks of this dissimilarity are the calculation cost, higher than for the Euclidean distance; and the choice of the centroid, as the arithmetic mean of several time series sometimes produces a series with a shape that is very different from all the individual series. We note that this measure considers patients stopping therapy as close and would identify a "dropout" cluster.

The Fréchet distance[10] applies to parametric curves. A discrete version exists for polygonal curves and can be applied to time series through the *generalized discrete Fréchet distance* and its variants.[7] We propose a formal definition of the generalized Fréchet distance and its variant, the *generalized summed discrete Fréchet dissimilarity* (Definition 5). These

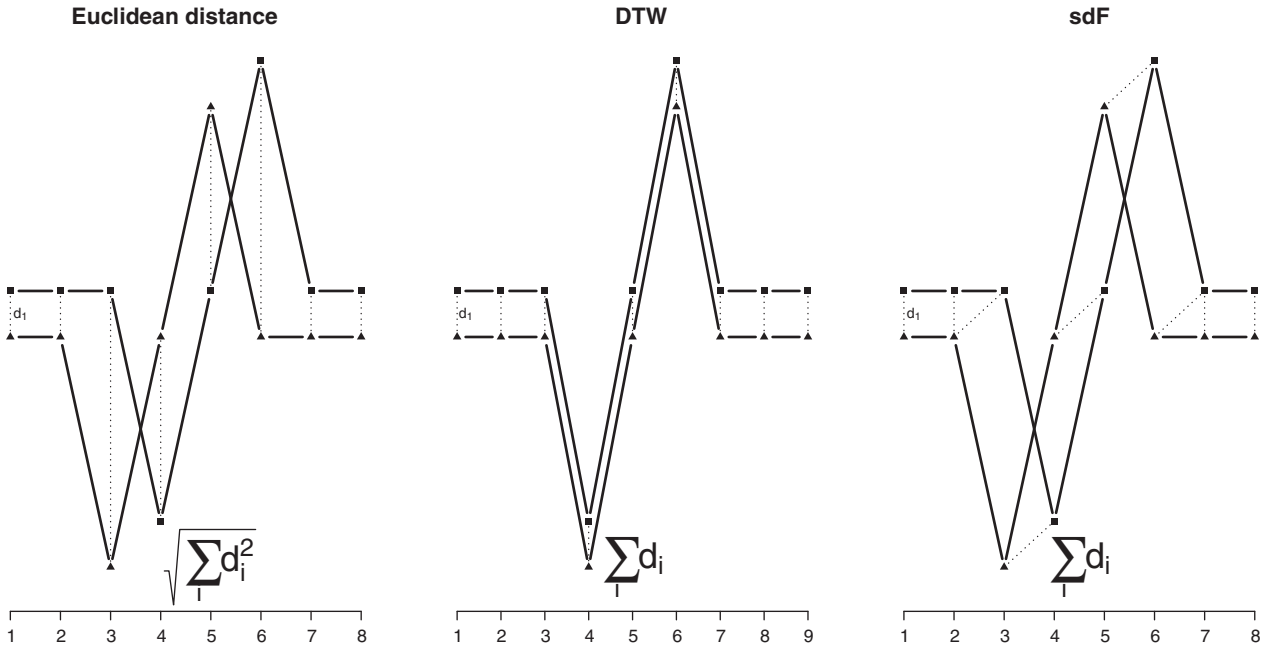


FIGURE 4 Schematic calculation principles for the 3 dissimilarity measures: Euclidean distance (left), dynamic time warping (middle), and generalized summed discrete Fréchet dissimilarity with $\lambda = 1$ (right). The Euclidean distance considers them pointwise. DTW first computes the optimal alignment before summing the pointwise differences on the *y* axis. sdF also computes an optimal alignment before summing the pointwise differences on the plane

measures are also "elastic" yet able to consider time shifts. As for DTW dissimilarity, they have a high calculation cost and also need a specific averaging process to provide cluster centroids. While the Euclidean distance requires two trajectories of the same length, an advantage of elastic dissimilarity such as DTW and sdF is that they can deal with trajectories having different numbers of points. Definitions 4 and 5 can be applied to time series of different lengths thanks to coupling, as recalled in Definition 2 (Appendix B.1).

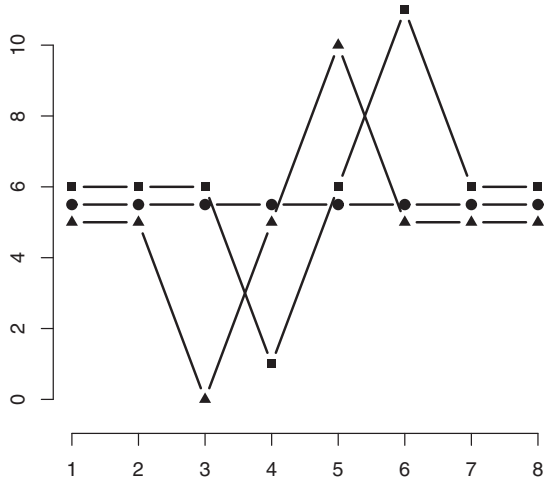
The parameter λ of the sdF dissimilarity can be understood as a time scale parameter. When $\lambda = 1$, the generalized discrete Fréchet distance becomes the classical discrete Fréchet distance and a difference of one unit on the variable of interest (the adhesion value here) is equivalent to a time lag of one time measure. A small value for λ gives more importance to differences in the variable of interest whereas a high value gives more importance to time lags (see Genolini et al[7] for more details). We note that the generalized summed Fréchet dissimilarity with $\lambda = 0$ matches the DTW measure. We have tested several values in a previous simulation experiment, and the results were not particularly sensitive to the choice of λ so we implemented a single value for this parameter. Within the context of our study, a difference of one hour between two trajectories was more clinically important than a time lag of one day, thus values inferior to 1 were preferred. We wanted to consider a constant difference of 4 hours between two trajectories as equivalent to a 15-day lag. Thus we set $\lambda = \frac{4}{15}$.

Figure 5 represents a third additional flat trajectory and the corresponding values for each dissimilarity and each pair of trajectories. The initial trajectories, marked with triangle and square are the farthest with the Euclidean distance, while they are the closest with DTW and sdF dissimilarities. This is explained because these trajectories share the same shape, with a time lag and only a difference of one unit on the *y* axis.

We were interested in how these dissimilarities manage the temporal shifts and find overall trends in the context of CPAP time series, which have high variability and many discontinuities due to the zeros. Especially, we asked if warping searches at any price to align the irregularities of the time series at the expense of grouping those with close trends.

3.2 | Linkage strategies

After the choice of a dissimilarity measure, the next step is to construct the hierarchy of the partition with a linkage strategy. This defines the dissimilarities between clusters from the individual pairwise dissimilarity matrix. The three strategies



Dissimilarity	Euclidean	DTW	sdF
$d(\blacktriangle, \bullet)$	7.21	13.00	13.00
$d(\blacksquare, \bullet)$	7.21	13.00	13.00
$d(\blacktriangle, \blacksquare)$	10.39	9.00	11.07

FIGURE 5 Examples of three fictive trajectories and the corresponding dissimilarities

we considered were "average," "complete," and "Ward," among the most used and are described below. Definitions and formula are given in Appendix B.3.

The "average" linkage strategy defines the dissimilarity between two clusters as the average dissimilarity between each point of the first cluster and each point of the second cluster.

The "complete" linkage strategy defines the dissimilarity between two clusters as the largest dissimilarity between two subjects in the two clusters. Hence this strategy avoids merging two clusters if they contain time series strongly dissimilars.

The "Ward" method[11] consists of successively merging the two clusters which results in a minimal increase of the total within-cluster sum of squares. It is based on distance between objects and cluster centroids and was initially designed to work with the Euclidean distance and the arithmetic mean as centroid. This method has been extended to be convenient when considering other dissimilarities[12] and moreover, the Ward algorithm can be run using only the dissimilarity matrix[13] without use of centroids.

3.3 | Choosing the number of clusters

After constructing the partition hierarchy, the following step is to define the partition by choosing the number of clusters. An internal CVI computes a score from a partition of the dataset in terms of a dissimilarity measure. In practice, the indices are computed for the partitions obtained with different numbers of clusters and compared, allowing one to choose the "best" number of clusters. The CVIs differ by their definition of what is a good partition. Some of them work with a subjective decision like a knee point. We focused on objective CVIs, such that the higher or smaller the CVI is, the better the partition. Most indices are based on two criteria: compactness and separation. They have different properties,[14] but as far as we know there is no study examining them in the context of time series clustering. Here we considered the six following indices: Calinski-Harabasz (CH),[15] COP,[16] Davies-Bouldin (DB),[17] Modified Davies-Bouldin (DB*),[18] Dunn (D),[19] and Silhouette (Sil).[20] Note that CH, DB, DB*, and COP indices require to compute cluster centroids and CH also needs a global centroid.

3.4 | Centroids

The centroid of a set of objects is an object at the center of the set, related to a dissimilarity measure. Determining a centroid is a way of summarizing the set in terms of one object. Given a dissimilarity measure d, the centroid R_C of the set of objects C is defined by $R_C = \text{argmin}_{\xi \in \Xi} \sum_{x \in C} d(x, \xi)$, where Ξ has to be chosen. When d is the Euclidean distance, the centroid corresponds to the arithmetic mean at each time point of the time series. However, the classical mean is not the best centroid for every dissimilarity, especially with elastic measures. With these kinds of dissimilarity, due to the warping preceding the computation of the dissimilarity, the centroid may be of different length and time points from the

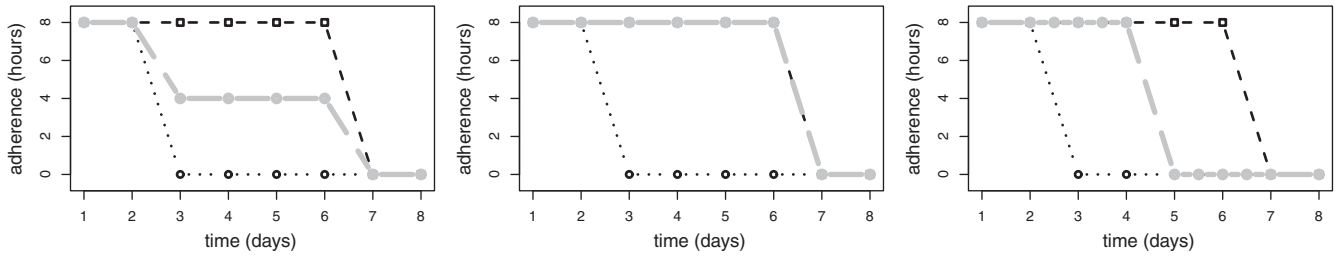


FIGURE 6 Examples (in gray) for the three centroids: The arithmetic mean (left), dynamic time warping barycenter averaging (DBA) (middle), and the Fréchet mean (right) on two fictive time series (dotted lines)

summarized time series. DTW barycenter averaging (DBA)[21] is a global averaging strategy providing centroids for groups of time series, well adapted to the DTW dissimilarity. The Fréchet mean[7] is an averaging strategy compatible with the sdF dissimilarity. These two strategies are heuristic and depend on an initialization step such that a distinction exists between the centroid algorithm and the centroid itself. Unlike the arithmetic mean, both the DBA and the Fréchet mean provide centroids with shapes similar to the individual time series, but at the price of a higher calculation cost. Figure 6 shows two fictive adherence time series over 8 days with application of the arithmetic mean, DBA, and Fréchet mean.

Another possibility is the medoid. The medoid of a group C is the centroid when $\Xi = C$. It is the object of C which minimizes the average dissimilarity to all other objects in C. The medoid is compatible with every dissimilarity measure.

3.5 | Software and packages

Within the R software environment (www.r-project.org), the package "dtwclust"[22] enables the clustering of time series with HAC among other methods. It includes the Euclidean distance and the DTW dissimilarity, with the DBA algorithm. It is possible to customize both the dissimilarity measure and the centroid process. The package "kmlShape"[7] includes the k-means algorithm with the sdF dissimilarity and the Fréchet mean. Combining these packages makes it possible to cluster time series with the sdF dissimilarity and using the HAC algorithm. "dtwclust" also includes the computation of some external CVIs such as the adjusted Rand index (ARI) and the 6 previously cited internal CVIs. We used this package for the simulation study.

4 | SIMULATION STUDY

In the simulation study, we compared the performance of the HAC algorithm using the three linkage strategies ("average," "complete," "Ward") and the three dissimilarity measures (Euclidean, DTW, sdF), that is, nine different clustering methods. The six internal CVIs mentioned in Section 3.3 were tested. R code is available upon request.

4.1 | Data generation process

Artificial data sets similar to real CPAP adherence time series were simulated following the results of Babbin et al,[23] in which clustering with HAC, the Euclidean and the Ward linkage strategy were applied to six month CPAP adherence series from 128 subjects. A partition with four clusters was obtained, called groups B, C, D, and E. As non-users and individuals stopping therapy early were excluded from their study, we added an extra group called "A" to our simulation study, corresponding to patients who dropped out, that is, completely stopped using CPAP. The length of the simulated time series was $T = 91$. Figure 7 shows the basic curves (see definition below) for each group. Simulated data were then obtained by adding white noise with variance σ^2.

Group A, dropout patients, was simulated with the model

$$X_{At} = \mathbb{1}_{t<\gamma} \times (5.5 + \epsilon_{At}) - \mathbb{1}_{\gamma-\tau\leq t<\gamma} \times \frac{5.5 \times (t - (\gamma - \tau - 1))}{\tau + 1},$$

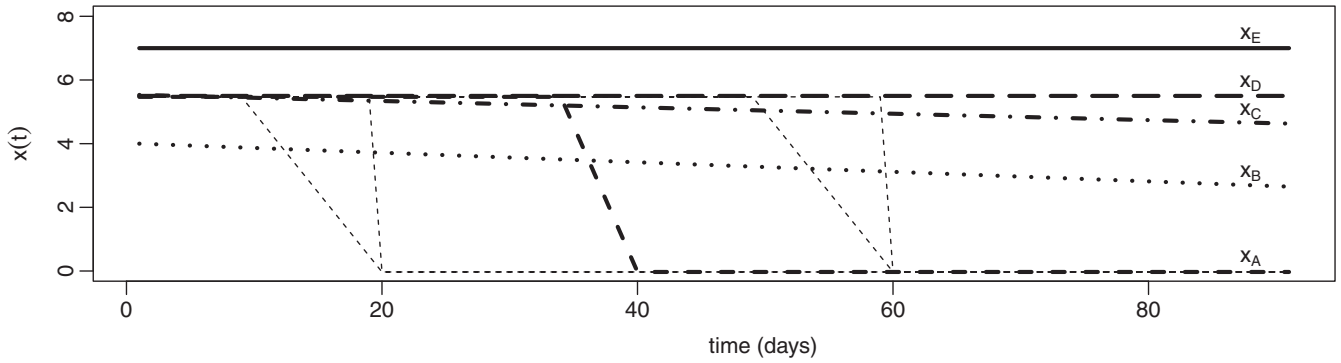


FIGURE 7 Non-noisy trends of groups A (dashed line), with 5 declensions varying the length of the phase of dwindling use and complete dropout, B (dotted), C (dots + dashes), D (long dashes), and E (solid) used for the simulation study

with an initial phase $[0, \gamma - \tau]$ with a constant trend, a phase of dwindling use $[\gamma - \tau, \gamma]$ and a "zero" or stopped phase $[\gamma, T]$, where γ and τ follow discrete uniform distributions $[20, 60]$ and $[0, 10]$ respectively. ϵ_{At} is white noise with variance σ^2. The four other groups were defined by:

$$X_{Bt} = 4 - 0.015t + \epsilon_{Bt},$$
$$X_{Ct} = 5.5 - 0.01t + \epsilon_{Ct},$$
$$X_{Dt} = 5.5 + \epsilon_{Dt},$$
$$X_{Et} = 7 + \epsilon_{Et},$$

where ϵ_{Bt}, ϵ_{Ct}, ϵ_{Dt}, ϵ_{Et} were white noise of variance σ^2.

To create data close to real-life data, four parameters were introduced: (1) group overlap, (2) presence of outliers, (3) variance of noise, and (4) null adherence values. Due to the proximity of groups C and D, including them or not in the datasets was a way to create group overlap. When the 5 groups were present in a sample (overlap), we generated 150 time series per group. For samples constituted only of groups A, B, and E (no overlap), each group contained 250 subjects. The second parameter was the presence or not of outliers. When included, they added 75 (10%) supplementary curves, each characterized by a first value and a slope. The model for an outlier trajectory was $X_t = \alpha + \beta t + \epsilon_t$, where α and β follow uniform distributions of $[2, 8]$ and $[-0.05, 0.05]$ respectively, and ϵ_t was white noise of variance σ^2. The variance of the noise σ^2 was either 2.5 or 6.5. These values corresponded approximately to the first and third quartiles of the individual variances estimated on real datasets (see Figure A1). The last parameter aimed at generating null adherence values. A null value was introduced when the simulated adherence value was below the censorship limit ρ, either fixed at 0 or at 1.5. In the real dataset, 4.1% of the 77 168 values were non-null and below 1.5 hours. Examples of simulated adherence trajectories from each group are shown in Figure 8.

These choices led to 16 parameter combinations. A hundred samples of 750 or 825 time series were simulated for each of them. Each of these datasets was clustered using the HAC algorithm, with either the "average," "complete," or "Ward" linkage strategies, and either the Euclidean distance, DTW dissimilarity or the generalized summed discrete Fréchet dissimilarity. The performances of these nine different clustering methods were compared using the methodology described below.

4.2 | Comparison of clustering methods

For a sample j, the true partition is p_j. In the case of a sample with outliers, each extra-group time series is considered as its own group. To evaluate the quality of the partition provided by a clustering algorithm when the true partition is known, an external CVI can be used returning a number in the range $[0, 1]$ (or $[-1, 1]$) such that the greater the similarity among partitions the nearer the value is to 1. Several external CVIs exist and we chose the ARI.[24] Let $h_{j,m}$ be the dendrogram obtained on sample j with the clustering method m, and $p^k_{j,m}$ the corresponding extracted partition with k clusters. A naive evaluation considers only the partition with the true known number of groups, and compares it with the real partition.

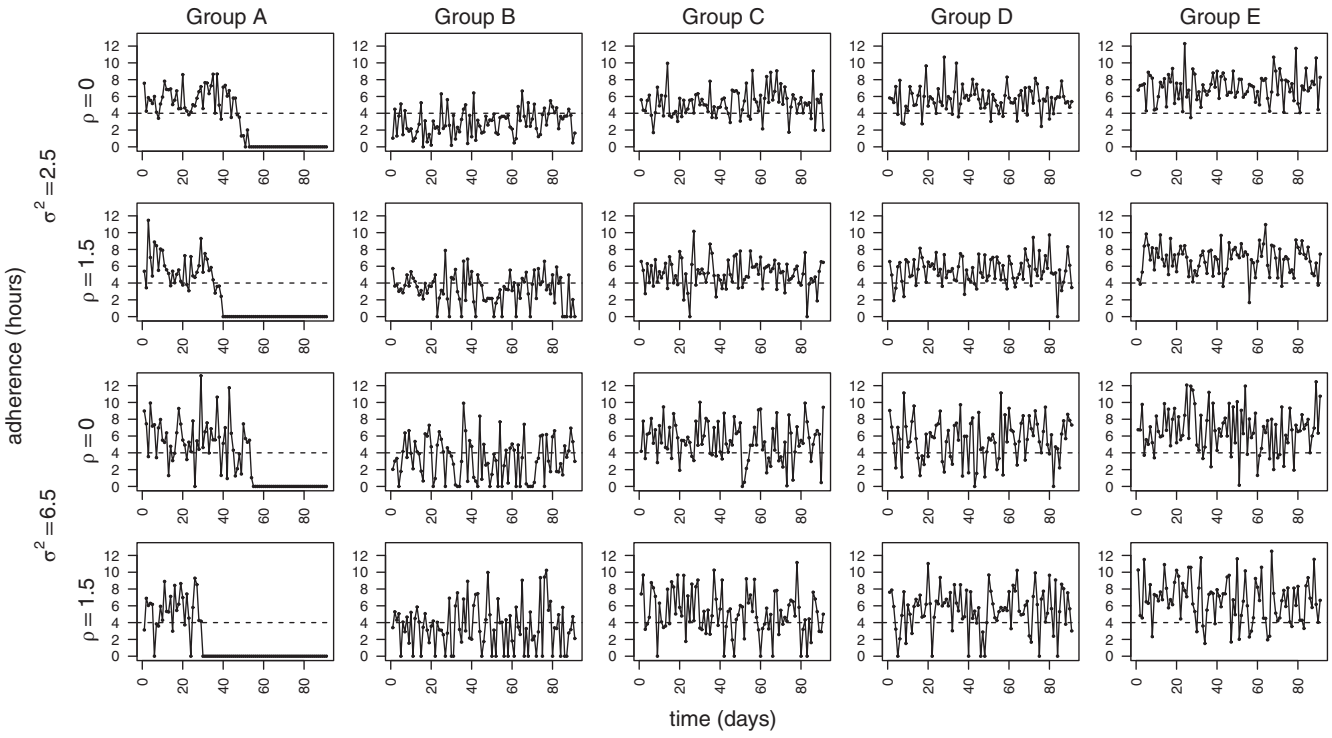


FIGURE 8 Examples of simulated trajectories. Each column stands for a group A, B, C, D, and E from left to right. Each combination of noise variance (σ^2) and censorship limit (ρ) are shown in rows

This approach assumes that among partitions resulting from a clustering method, the best one has the true number of clusters.[8] However this assumption does not hold, Figures C1 and C3 (in Appendix C) show an example where the best partitions do not have the true number of clusters. Instead of looking at only the partition with the true number of clusters, we computed the set of the best number of clusters with respect to ARI, with k ranging from 2 to 20. We note $\widehat{K_{j,m}} = \text{argmax}_{k\in[\![2;20]\!]}\ ARI(p^k_{j,m}, p_j)$ this set. However $\widehat{K_{j,m}}$ can be either unique (only one partition of the dendrogram maximizes the ARI) or a set number of clusters achieving the maximal ARI value. Let us denote $\widehat{P_{j,m}} = \{p^k_{j,m} \mid k \in \widehat{K_{j,m}}\}$ the set with the best partitions for the dendrogram $h_{j,m}$. The performance of the method m on sample j is evaluated through $s_{j,m} = ARI(p, p_j)$ where p is one element of $\widehat{P_{j,m}}$. The classification is perfect when its best partition exactly matches the original one (ie, $s_{j,m} = 1$). To compare methods we computed their perfect classification rates (*PCR*). We also looked at the $s_{j,m}$ distributions through their mean ($\overline{s_{.,m}}$) and standard deviation ($s_{s_{.,m}}$).

After the choice of the clustering method, the next step was to determine the best number of clusters.

4.3 | Comparison of internal CVIs

Let us consider a dendrogram $h_{j,m}$ with its associated partitions $\{p^k_{j,m} \mid 2 \le k \le 20\}$, an internal CVI v and a centroid algorithm $\mathcal{R}$. We note d the dissimilarity measure that is related to the clustering method m.

Let $\widetilde{K^{v,\mathcal{R}}_{j,m}}$ be the set of value $k \in [\![2; 20]\!]$ which optimizes $v(p^k_{j,m}, d, \mathcal{R})$ and $\widetilde{P^{v,\mathcal{R}}_{j,m}} = \{p^k_{j,m} \mid k \in \widetilde{K^{v,\mathcal{R}}_{j,m}}\}$. The performance of v with the centroid algorithm $\mathcal{R}$ and the method m on sample j is evaluated through $s^{v,\mathcal{R}}_{j,m} = \frac{\sum_{k\in\widetilde{K^{v,\mathcal{R}}_{j,m}}} ARI(p^k_{j,m}, p_j)}{|\widetilde{K^{v,\mathcal{R}}_{j,m}}|}$, the mean ARI between the true partition p_j and each element of $\widetilde{P^{v,\mathcal{R}}_{j,m}}$. The CVI v was considered to be a success on the sample j with clustering method m and centroid algorithm $\mathcal{R}$ when there was a partition maximizing ARI among the partition optimizing v (ie, $\widehat{P^m_j} \cap \widetilde{P^{v,\mathcal{R}}_{j,m}} \neq \emptyset$). We compared the performance of the internal CVIs with their success rates (*SR*) and also by comparing their $s^{v,\mathcal{R}}_{j,m}$ scores through their mean ($\overline{s^{v,\mathcal{R}}_{.,m}}$) and standard deviation ($s_{s^{v,\mathcal{R}}_{.,m}}$).

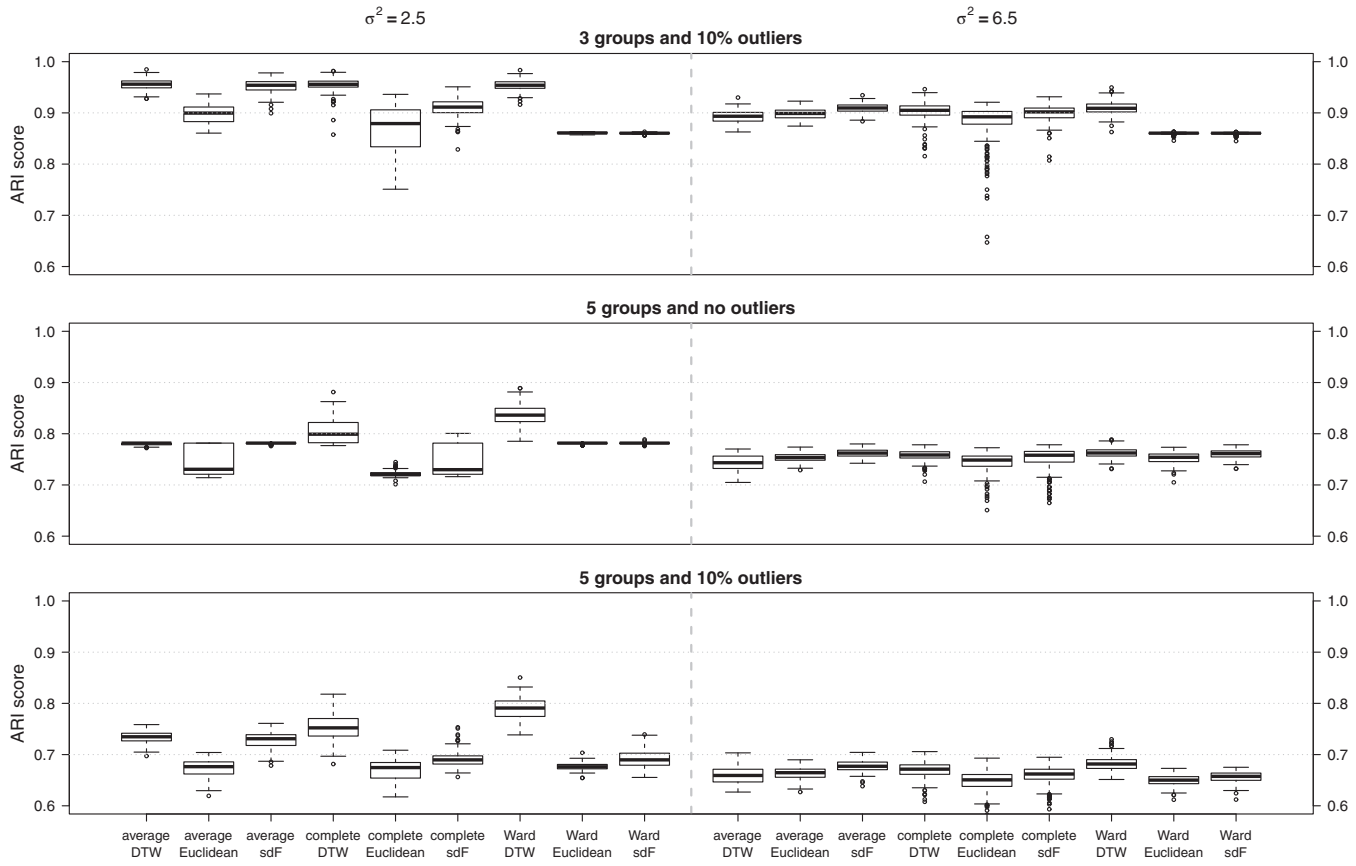


FIGURE 9 Simulation study: Boxplots of $s_{j,m}$ scores for each method in cases with 3 groups and 10% of outliers, with 5 groups without outliers, and with 5 groups and 10% of outliers. Boxplots are split according to the variance σ^2 making 6 different cases of 200 samples

4.4 | Simulation results

4.4.1 | Comparison of clustering methods

The simplest case is 3-group clustering. When there is no outlier, almost all methods reach near 100% perfect classification, except two of them. Average linkage with DTW dissimilarity and complete linkage with Euclidean distance exhibited rates of perfect classification of around 50% (see Table 1). With 10% of outliers, there was no perfect classification. Average and complete linkage strategies could produce a lot of clusters in the final partition (see Figure C2 in Appendix C), which, in practice, is a drawback. The distributions of $s_{j,m}$ scores with 3 groups and outliers, or with 5 groups without and with 10% of outliers are presented in Figure 9. With 5 groups and no outlier the Ward linkage strategy gave the best results in term of $\overline{s_{.,m}}$ (Table 1). In the case of low variance ($\sigma^2 = 2.5$), 5 theoretical groups without outliers (see middle plot in Figure 9), the DTW dissimilarity measure with the Ward linkage strategy clearly yielded the best scores even for 5 groups and 10% of outliers, whatever the variance (see bottom plot in Figure 9). In this configuration the DTW dissimilarity with the Ward linkage strategy gave the best mean $\overline{s_{.,m}}$ (see Table 1). We also noticed that there was a large range of values for the number of clusters in the final partition with average and complete linkage (see Figure C4 in Appendix C). The presence of outliers considerably influenced the mean $\overline{s_{.,m}}$ and the number of clusters in the resulting partitions (see Figures C1-C4 in Appendix C). A large variance σ^2 shrunk the global performance of the clustering methods and reduced the differences between methods. The censorship limit and thus a greater presence of null values had no impact on the clustering performances. This is important regarding its practical application to CPAP time series.

Finally, from the simulation study, the DTW dissimilarity used with the Ward linkage strategy provided the best results both in terms of $s_{j,m}$ score distributions and perfect classification rates. Each clustering method was applied to the same 1600 datasets. We ran 8 Wilcoxon tests comparing each combination to DTW-Ward. The highest P-value after Bonferroni correction was close to 4.99×10^{-73} showing that DTW-Ward was thus significantly better than the other combinations.

TABLE 1 Simulation study: Comparison of the methods according to group overlap (no overlap with 3 groups, overlap with 5 groups) and presence of outliers (400 samples per combination)

		3 groups					5 groups			
		No outliers			10% of outliers		No outliers		10% of outliers	
Linkage	**Dissimilarity**	**PCR**	$\overline{s_{.,m}}$	$s_{s_{.,m}}$	$\overline{s_{.,m}}$	$s_{s_{.,m}}$	$\overline{s_{.,m}}$	$s_{s_{.,m}}$	$\overline{s_{.,m}}$	$s_{s_{.,m}}$
Average	DTW	0.54	0.99	0.01	0.92	0.05	0.76	0.02	0.69	0.05
	Euclidean	0.97	1.00	0.00	0.90	0.02	0.75	0.02	0.67	0.01
	sdF	1.00	1.00	0.00	0.93	0.02	0.77	0.01	0.70	0.03
Complete	DTW	1.00	1.00	0.00	0.93	0.03	0.78	0.03	0.71	0.05
	Euclidean	0.53	0.94	0.06	0.88	0.04	0.73	0.02	0.66	0.02
	sdF	1.00	1.00	0.00	0.90	0.02	0.75	0.02	0.68	0.02
Ward	DTW	1.00	1.00	0.00	0.93	0.03	0.80	0.04	0.74	0.06
	Euclidean	1.00	1.00	0.00	0.86	0.00	0.77	0.02	0.66	0.02
	sdF	1.00	1.00	0.00	0.86	0.00	0.77	0.01	0.67	0.02

Note: Perfect classifications rates (PCR), mean $s_{j,m}$ scores ($\overline{s_{.,m}}$), and standard deviation ($s_{s_{.,m}}$).

TABLE 2 Simulation study: Success rates (SR), mean $s_{j,m}^{v,\mathcal{R}}$ scores ($\overline{s_{.,m}^{v,\mathcal{R}}}$), and standard deviation ($s_{s_{.,m}^{v,\mathcal{R}}}$) for each internal cluster validity index (CVI) depending on the number of groups and presence of outliers (400 samples for each combination)

	3 groups						5 groups					
	No outliers			10% of outliers			No outliers			10% of outliers		
CVI (centroid)	**SR**	$\overline{s_{.,m}^{v,\mathcal{R}}}$	$s_{s_{.,m}^{v,\mathcal{R}}}$	**SR**	$\overline{s_{.,m}^{v,\mathcal{R}}}$	$s_{s_{.,m}^{v,\mathcal{R}}}$	**SR**	$\overline{s_{.,m}^{v,\mathcal{R}}}$	$s_{s_{.,m}^{v,\mathcal{R}}}$	**SR**	$\overline{s_{.,m}^{v,\mathcal{R}}}$	$s_{s_{.,m}^{v,\mathcal{R}}}$
Silhouette	0.00	0.57	0.00	0.00	0.48	0.01	0.00	0.37	0.00	0.00	0.30	0.02
Calinski Harabasz (Medoid)	0.00	0.57	0.00	0.00	0.48	0.01	0.00	0.37	0.03	0.00	0.30	0.04
Calinski Harabasz (DBA)	0.01	0.57	0.04	0.00	0.48	0.01	0.00	0.37	0.01	0.00	0.30	0.02
Davies Bouldin (Medoid)	0.00	0.57	0.00	0.00	0.48	0.04	0.01	0.38	0.05	0.00	0.31	0.05
Davies Bouldin (DBA)	0.32	0.71	0.20	0.00	0.60	0.18	0.00	0.42	0.05	0.00	0.36	0.08
Modified Davies Bouldin (Medoid)	0.00	0.57	0.00	0.00	0.48	0.02	0.02	0.38	0.06	0.00	0.31	0.05
Modified Davies Bouldin (DBA)	0.11	0.62	0.13	0.00	0.52	0.12	0.00	0.42	0.05	0.00	0.36	0.08
Dunn	0.86	0.94	0.15	0.76	0.86	0.16	0.28	0.75	0.06	0.34	0.67	0.09
COP (Medoid)	0.00	0.32	0.02	0.00	0.39	0.03	0.00	0.38	0.02	0.00	0.43	0.05
COP (DBA)	0.00	0.33	0.02	0.00	0.40	0.03	0.00	0.38	0.02	0.00	0.43	0.05

Note: CVIs for which a centroid is needed are displayed on two lines. The first line corresponds to the CVI computed with the medoid and the second corresponds to the CVI computed with dynamic time warping barycenter averaging.

4.4.2 | Comparison of internal CVIs

The internal CVIs were evaluated with the clustering based on Ward linkage and DTW dissimilarity (see Table 2). Except for the Dunn index, and whatever the simulation parameters, all internal CVIs performed very poorly in identifying the number of clusters which maximizes the ARI index. The Dunn index was the most effective with high success rates with 3 groups but somewhat lower with 5 groups. This was confirmed by the mean $\overline{s_{j,m}^{v,\mathcal{R}}}$ scores (Table 2). Based on this simulation study with the clustering method previously selected, the best internal CVI to choose the number of clusters for CPAP adherence series is the Dunn index.

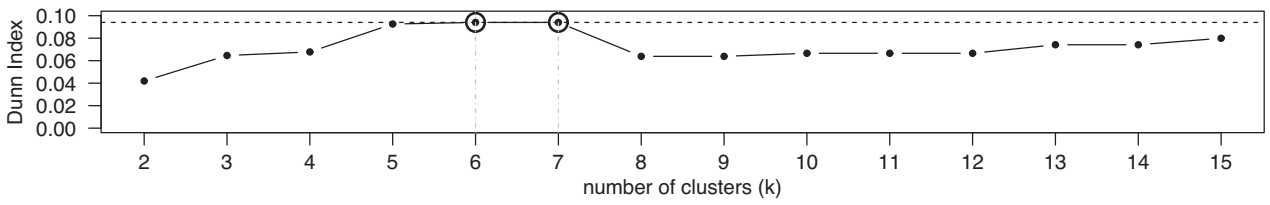


FIGURE 10 Real CPAP data: Dunn index values by number of clusters

5 | REAL-LIFE CPAP ADHERENCE CLUSTERING

Our aim was to explore individual CPAP adherence data using cluster analysis, to reveal typical patterns. According to the simulation results, the data presented in Section 2 were clustered with the Ward linkage strategy and the DTW dissimilarity, and the number of clusters was selected with the Dunn index.

Partitioning giving from 2 to 15 clusters were considered. Figure 10 represents the Dunn index values with the number of clusters. Partitions with 6 or 7 clusters both maximized the internal CVI. For sparsity reasons, the partition with the lowest number of clusters was chosen first. Figure 11 shows the median adherence trajectories for the 6 clusters. The first cluster contains 105 subjects with a median use of 4 hours per night at the beginning of therapy; gradually increasing nearly 8 hours per night at the end of 3 months. The variances in the time series are high, especially at the beginning of treatment, and patients have some zeros. The second cluster comprises 151 highly adherent users with a median trajectory stabilizing at around 8 hours use per night. Zero values are rare and the individual variances are low. The third cluster groups 29 patients who stopped CPAP. Individual variances are high and many zeros appear, especially towards the complete discontinuation of treatment. The fourth cluster contains 150 subjects with good adherence, a median trajectory of nearly 6 hours per night, individual variances are quite low and few zeros appear. Cluster 5 brings together 178 non-adherent users, a median trajectory beginning at 2 hours that decreases. Individual variances are low to moderate and several individual trajectories contain many zeros. The sixth and largest cluster contains 235 subjects, the median trajectory is about 4 hours, but individual variances are moderate and some zeros appear. Clusters 2, 4, and 6 represent subjects with predominately stable patterns but at different levels. One might consider that the difference between these clusters is due to sleep duration. However, for clusters 2 and 4, the difference in the median mean adherence is almost an hour and a half. This corresponds to a complete sleep cycle, which has clinical meaning. Cluster 6 presents smaller proportions of nights with usage at over 4 hours, with higher rates of zero use and greater variances, indicating that cluster 6 also differs from clusters 2 or 4 in usage regularity.

Clustering was performed using a HAC method. Thus the partition among 6 clusters is obtained from the partition into 7 clusters and then merging two clusters. The cluster 2 results from merging 2.1 and 2.2, containing respectively 40 and 111 subjects. Figure 12 shows their median trajectories. Really good users are thus spread between two clusters where median trajectories remain stable near 8 hours per night for both clusters. The difference between these two clusters is in the regularity of adherence. Individual medians are higher in cluster 2.1 than in cluster 2.2, but with higher variances and more occurrences of zero. Keeping in mind that the main motivation of this study was personalized support for subjects at risk of dropping out or being insufficiently adherent, this distinction is not relevant. The 6-cluster partition was finally retained. Table 3 also describes these clusters and compares them. The independence between these clusters and continuous variables was tested using the Kruskal-Wallis test, and the independence with the discrete variables was tested using the chi-square test. The significance level of the tests is fixed at 5%. We note that belonging to one cluster or another is not independent of the age of the patient, nor of other individual adherence characteristics presented on this table (mean, standard deviation, rate of zeros and proportion of nights with adherence over 4 hours).

The median trajectories are a way to represent the 6 clusters but do not properly show individual characteristics in terms of null values and dropouts for example. This is especially so for clusters 3 and 5 that group subjects with many zeros, the first because of dropouts and the second because of poor and irregular adherence. Boxplots in Figure 13 give more information, showing for each cluster the distribution of the day with the first zero and of the day of completely stopping CPAP, that is, dropout, when applicable. The distributions of these two characteristics for the overall population are also shown in the histograms in Appendix A (Figure A1). The distribution taking into account the day of the first zero showed that subjects in the "dropout" cluster (cluster 3) have zero use at later times than subjects in cluster 5 (poor adherers). The distribution that considers the time of dropout revealed two facts about the resulting partition. The first was that

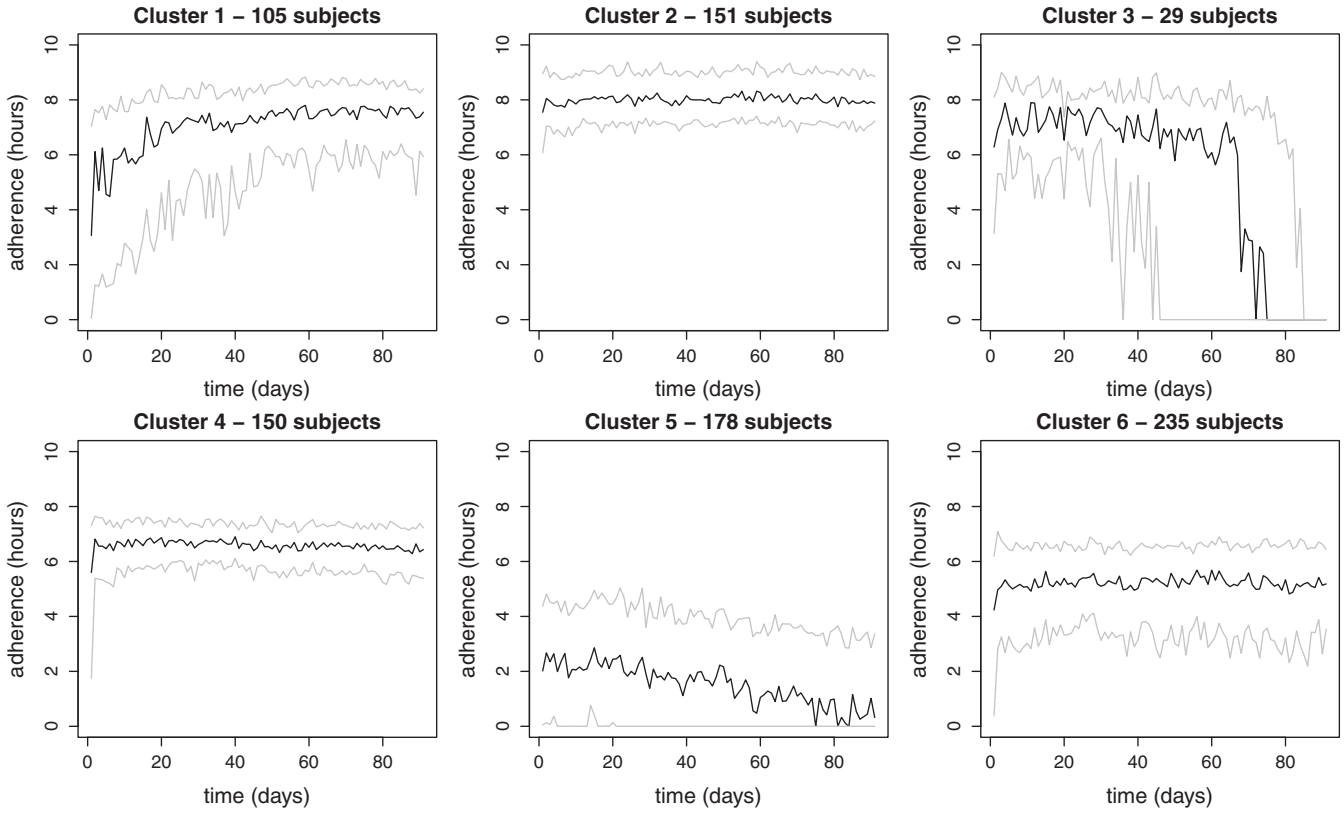


FIGURE 11 Real CPAP data: Representatives of the 6 clusters with median trajectories (black) and quartile trajectories (gray). Median and quartile trajectories are calculated pointwise

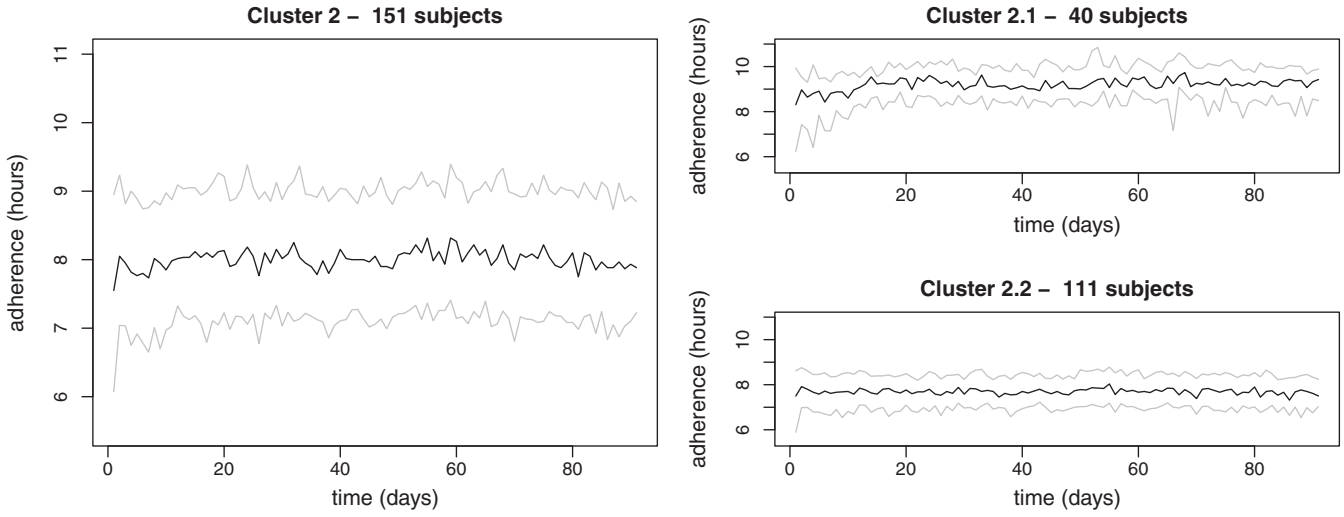


FIGURE 12 Real CPAP data: Median (black) and quartile (gray) trajectories for cluster 2 (left) from the partition with 6 clusters and for clusters 2.1 and 2.2 (right) from the partition with 7 clusters. Median and quartile trajectories are calculated pointwise

TABLE 3 Real CPAP data: Descriptive statistics for the selected population and comparison of the 6 clusters

	Male sex	Age (year)	Body mass index (kg/m^2)	Apnea hypopnea index (event/hour)	Mean adherence (hour)	Standard deviation of adherence (hour)	Rate of zeros (%)	Proportion of adherence >4 hours (%)
Whole population								
N = 848	533 (62.9%)	59 [50; 69]	30.4 [26.7; 34.6]	36.0 [30.0; 50.1]	5.5 [3.8; 6.7]	2.1 [1.5; 2.6]	6.6 [1.1; 19.8]	78.6 [51.6; 93.4]
Missing values	0 (0%)	2 (0.2%)	66 (7.8 %)	204 (24.1 %)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Cluster comparison, n (%)								
Cluster 1, 105 (12.4%)	68 (64.8%)	62 [45; 70]	28.0 [25.6; 33.1]	37.5 [30.0; 54.5]	6.3 [5.5; 7.0]	2.9 [2.4; 3.3]	8.8 [4.4; 14.3]	81.3 [68.1; 87.9]
Cluster 2, 151 (17.8%)	91 (60.3%)	60 [52; 70]	30.8 [27.1; 35.2]	39.0 [31.0; 58.0]	7.8 [7.3; 8.3]	1.4 [1.0; 1.8]	0.0 [0.0; 1.6]	97.8 [95.6; 98.9]
Cluster 3, 29 (3.4%)	15 (51.7%)	58 [45; 69]	30.9 [27.9; 36.0]	35.0 [30.0; 40.0]	5.5 [3.0; 6.4]	2.9 [2.6; 3.4]	22.0 [9.9; 53.8]	72.5 [40.7; 85.7]
Cluster 4, 150 (17.7%)	90 (60.0%)	63 [54; 72]	30.0 [27.0; 33.1]	35.0 [30.0; 48.0]	6.4 [6.0; 6.6]	1.7 [1.4; 2.0]	1.1 [0.0; 3.3]	92.3 [88.2; 95.6]
Cluster 5, 178 (21.0%)	120 (67.4%)	57 [48; 65]	30.9 [26.8; 36.4]	34.5 [29.0; 45.6]	2.2 [1.3; 3.2]	2.0 [1.5; 2.5]	33.5 [17.6; 53.8]	23.6 [8.8; 39.3]
Cluster 6, 235 (27.7%)	149 (63.4%)	58 [50; 67]	30.2 [26.2; 34.7]	36.5 [30.0; 51.0]	4.8 [4.2; 5.3]	2.3 [2.0; 2.8]	8.8 [3.3; 18.7]	70.3 [58.8; 79.1]
P-value	0.510	**0.003**	0.091	0.133	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Note: Continuous variables are summarized using median [interquartile range] while discrete variables are summarized using frequency (proportion). The independence between these variables and clusters was tested using the Kruskal-Wallis test (continuous variable) and the chi-square test (discrete variable). The p-values under the significance level fixed at 5% are given in bold font.

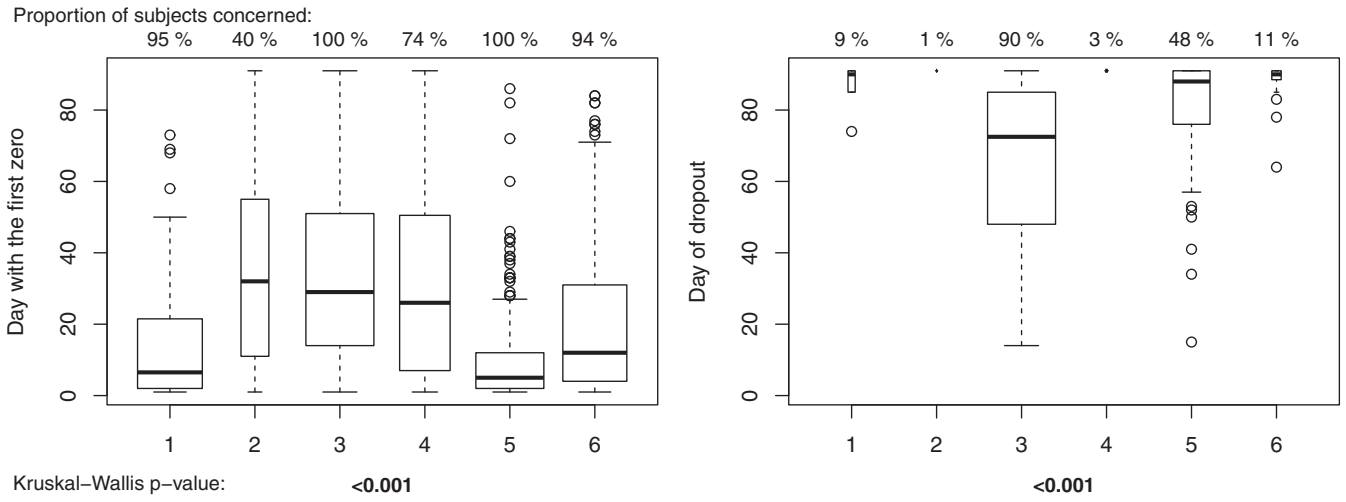


FIGURE 13 Real CPAP data: For each cluster, distribution of the day of the first zero (left plot) and distribution of the day of dropout (right plot). The proportions of subjects concerned are given above each boxplot and the widths of the box are proportional to these proportions. Clusters are compared using the Kruskal-Wallis test

more subjects in cluster 3 (90%) actually end their trajectory with a zero, than in cluster 5 (48%). The second was that the time when complete dropout occurred is earlier in the "dropout" cluster. Some patients who eventually stop CPAP, who begin with poor adherence are thus grouped in cluster 5. This is not visible using the median and quartile trajectories because these patients are grouped with subjects still continuing their CPAP therapy at the end of the third month. This highlights the difficulty in giving representatives of clusters based on time series clustering alone and advocates the need to use several tools. Figure D1 in Appendix D shows these clusters with their medoids.

Nevertheless, the comparison of the clusters gave significant differences for characteristics involving zero values (rate of zero [see Table 3], day of the first zero [Figure 13], and time of complete dropout [Figure 13]). This observation underlines that the distribution of zeros have been taken into account by the clustering algorithm, which gives reassurance in the resulting partition.

The two clustering approaches which obtained the nearest performances to that of the DTW-Ward combination in the simulation study were sdF-average and DTW-complete combinations. They were applied to the real-life dataset fixing the number of clusters at six for comparability reasons. Both approaches produced partitions with clusters of unbalanced sizes. SdF-average clustering identified four clusters with very few subjects (from one individual to seven) and DTW-complete produced one large cluster (385 subjects) and two small clusters (30 and 21 subjects). Clustering with the DTW measure and Ward linkage tended to identify more equilibrated clusters. Hence the resulting clusters are less susceptible to be dependent on a given dataset.

6 | DISCUSSION

In this work, we investigated complete clustering procedures primarily intended for use with CPAP adherence time series in an HAC framework. We considered the choice of dissimilarity measure taking into account several linkage strategies comparing an original dissimilarity measure, the generalized summed discrete Fréchet dissimilarity, with the classical Euclidean distance and the DTW dissimilarity that is widely used in time series clustering. We also examined the question of the selection of the number of clusters, using the six following internal clustering validation indices: Calinski-Harabasz, COP, Davies-Bouldin, Modified Davies-Bouldin, Dunn, and Silhouette.

In a simulation study we have shown that a combination of DTW dissimilarity, Ward linkage and the Dunn index provided the best results within the HCA context. Fictive datasets with parameters that included the presence of extra-group outliers, the variance of noise and the number of zeros were used. We found the presence of outliers to be the most influential parameter, reducing the performances of the clustering algorithm (using the Ward linkage and the DTW dissimilarity) and that of the Dunn index with this algorithm. Next, the variance of the noise decreased the performances of the internal CVI, with and without the presence of outliers and whatever the number of zeros, and also that of the

clustering algorithm, except in the case of 3 groups without outliers where performance was not affected. An increase in zeros did not change the performance of the clustering algorithm but substantially diminished the CVI. Nevertheless, the effects of these parameters were minimal so the whole clustering method appears stable enough to give confidence in the partition we would obtain with real data.

Some improvements in the design of the simulation could be made regarding the clinical question and the specificity of data to cluster. A first is to consider the timing of the phase of diminishing CPAP use and of complete dropout in group A as simulation parameters and to measure their impact on clustering performances. This question is related to the time scale parameter λ of the summed discrete Fréchet dissimilarity. Only a single value of λ was used, however the performances of the clustering methods based on sdF dissimilarity could vary with the supplementary simulation parameter previously mentioned and with other values of λ. We simulated only one group with a shape recognizable by elastic dissimilarity measures. An additional group with a zero phase, like subject 8 in Figure 3, could be simulated, also varying the number of days with such phases.

From a methodological point of view, it would be interesting to use partitional clustering with the sdF dissimilarity and Fréchet mean centroids, but the computational cost is too high. A further point concerns the choice of the number of clusters. Except for the Dunn index, all internal CVIs we implemented in the simulation study gave very poor performances in selecting partitions close to the theoretical ones. This was even true with the Euclidean distance or sdF dissimilarity. It appears that the CVIs we tested are not suitable for time series clustering. More research is needed to develop new internal CVIs that are appropriate for use with time series.

The application of HAC with Ward linkage, DTW measure and the Dunn index provided six clusters, ranging from a small cluster (N = 29 subjects) in which patients stopped the therapy early on, to a cluster (N = 105) in which patients increased their adherence over time. Other clusters included extremely good users (N = 151), good users (N = 150), moderate users (N = 235), and poor adherers (N = 178). Regarding the previous study on CPAP data conducted by Babbin et al[23] we found two supplementary and considerably different CPAP clusters, both characterized by their shape. As expected, one was users who discontinued the therapy and the other contained subjects who began with only moderate use and greatly increased CPAP use by the end of the 3 months. This was probably due to two main reasons. First because we considered non-users and zero values. It is obviously necessary to define clusters having extended periods of non-use. The second reason may be the dissimilarity we used that aligns time series to find similar shapes over time. Application of the Euclidean distance and Ward linkage strategy, as used by Babbin et al[23] along with the Dunn index, to our real dataset gave 8 clusters, but did not produce a "dropout" cluster. The subjects who dropped out were dispatched into several clusters depending on the date of dropout, but the date is not of clinical importance. More importantly, the six clusters identified using HAC with Ward linkage, DTW measure and the Dunn index made clinical sense and allow both clinicians and home care-providers to ensure a better allocation of resources by individualizing patient management during the stabilization phase.

These results were obtained considering a 3-month time window after treatment initiation for the CPAP trajectories. Applying the same clustering methodology to only the first 2 months of follow-up would not have recognized the dropouts. This underlines how the choice of time window impacts on the resulting clusters. An interesting research question would be to apply the clustering to the same trajectories varying the time window so as to identify the moment after which the clusters become quite similar over time, and giving an estimate of the time required before CPAP adherence stabilizes.

The description of CPAP adherence trajectories from individual trajectories has previously been attempted.[23] However, to our knowledge the work we report here is the first investigating clustering methods that consider the specificities of CPAP data, that is, high variability and many discontinuities, which render them different from classical time series.

In this study, we highlighted some advantages of using the DTW dissimilarity for clustering trajectories: (1) the possibility to apply it to trajectories of different lengths; (2) its ability for recognize shapes, illustrated by grouping together dropouts whatever the dates of dropout were. However, this was at the expense of an increase in computation time.

We also investigated the data emission process for the different models of CPAP device, so as to formulate hypothesis enabling us to distinguish zero use from missing values. One hypothesis was that for the devices U and V there were no missing values. This seems reasonable, but this topic merits complementary work to confirm the reliability of the data supplied by each CPAP model; especially null and missing values. A special focus on missing values could be a future research question, avoiding the patient exclusion seen in this study. It could use an imputation strategy or a dissimilarity measure that takes otherwise excluded patients into account. Another direction of study would be to look at the zero values because they could furnish information on individual adherence behaviors. First, an extension of this work would be to search for a smoothing strategy that preserves the zero values, without replacing them with positive averaged values; and a complementary approach would be to model the occurrence of zero values to better define poor adherence profiles.

A limitation to these results is that some clusters, such as extremely good users and good users, may differ due to their sleep duration rather than to adherence behaviors. Currently, sleep duration is mainly subjectively reported at treatment initiation and, to our knowledge, there is no dataset, which combines both daily sleep duration and daily CPAP adherence. For this study, we used the criteria of the French National Health Insurance System for CPAP reimbursement, which is based only on the number of hours of CPAP use. However, the method used here for hours of CPAP adherence could be used in the same way for the daily ratio of CPAP use to sleep duration. This type of data is currently unavailable, but the development of connected health applications might make such more complete data available in the future.

Beyond application to CPAP adherence data, the clustering algorithm combining the DTW dissimilarity and HAC with the Ward linkage strategy, could be extended to all time series containing discontinuities and high variability which are common in medical area. These methods can be applied in sleep apnea for other data generated from CPAP telemonitoring: the residual number of apnea events under CPAP (residual apnea hypopnea index) or air leaks during CPAP use. These data patterns are also observed during the initiation period of non-invasive ventilation in patients with chronic obstructive pulmonary disease. In other fields, daily data recorded by activity trackers, accelerometers, pedometers, smart-phone applications, or wearable medical devices can present characteristics of variability and discontinuities. The necessity for unsupervised clustering was highlighted in previous studies.[25,26] Another field of application of such methods is the monitoring of blood markers, if collected at regular time periods, such as the residual concentration of immunosuppressant drugs after an organ transplant. Indeed, for data presenting the same specific problematics, clustering them through the DTW dissimilarity and HAC with the Ward linkage strategy might help to identify trajectory patterns at the beginning of a treatment or an intervention.

To conclude, our initial clinical objective was to enhance CPAP adherence follow-up for patients newly diagnosed with OSA. The DTW dissimilarity and HAC with the Ward linkage strategy seems well adapted to this approach in the field of OSA patient management. The resulting clusters showed clinical relevance and provided a richer description of the adherence behaviors than mean adherence values do. Comparing CPAP adherence clusters could confirm the effectiveness of a therapy and help to define more precisely efficiency thresholds.[27] The application of such methods is in line with the development of "P4" medicine (personalized, predictive, preventative, and participatory) in the field of OSA identifying specific trajectories of patient treatment..[28] Moreover, this clustering method can be applied to a broad spectrum of regularly collected data in different medical fields so as to improve personalized medicine.

The next step will be to use individual data collected at CPAP prescription to predict whether a given patient belongs to one cluster or another and to characterize these clusters. Such a predictive approach would help doctors and home care providers to lighten the follow-up of patients who are predicted to be adherent and would allow them to concentrate their efforts on subjects at risk of dropping out or being poor adherers. Another clinical application of this work would be to evaluate the impact of belonging to a particular cluster on the improvement in daytime symptoms and on the risk of comorbidities.

ACKNOWLEDGEMENTS

GBB, SB, ALS, and JLP are supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) and the "e-health and integrated care and trajectories medicine and MIAI artificial intelligence" Chairs of excellence from the Grenoble Alpes University Foundation. This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

The authors also acknowledge the Grenoble Alpes Data Institute, which is supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

We thank Alison Foote for rereading of this article and improving our English.

DATA AVAILABILITY STATEMENT

Data availability is restricted.

ORCID

Guillaume Bottaz-Bosson https://orcid.org/0000-0001-6346-3510
Sébastien Bailly https://orcid.org/0000-0002-2179-4650

REFERENCES

1. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;7(8):687-698.

2. Lévy P, Kohler M, McNicholas WT, et al. Obstructive sleep apnoea syndrome. *Nat Rev Dis Primers*. 2015;1:15015.
3. Portier F, Frija EO, Chavaillon JM, et al. Traitement du SAHOS par ventilation en pression positive continue (PPC). *Rev Mal Respir*. 2010;27:S137-S145.
4. Siccoli MM, Pepperell JC, Kohler M, Craig SE, Davies RJ, Stradling JR. Effects of continuous positive airway pressure on quality of life in patients with moderate to severe obstructive sleep apnea: data from a randomized controlled trial. *Sleep*. 2008;31(11):1551-1558.
5. Pépin JL, Bailly S, Tamisier R. Big data in sleep apnoea: opportunities and challenges. *Respirology*. 2020;25(5):486-494.
6. Aardoom JJ, Loheide-Niesmann L, Ossebaard HC, Riper H. Effectiveness of electronic health interventions in improving treatment adherence for adults with obstructive sleep apnea: meta-analytic review. *J Med Internet Res*. 2020;22(2):e16972.
7. Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS One*. 2016;11(6):e0150738.
8. Gurrutxaga I, Muguerza J, Arbelaitz O, Pérez JM, Martín JI. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recogn Lett*. 2011;32(3):505-515.
9. Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering – a decade review. *Inf Syst*. 2015;53:16-38.
10. Fréchet MM. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*. 1906;22(1):1-72.
11. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236-244.
12. Batagelj V. Generalized ward and related clustering problems. *Classification and Related Methods of Data Analysis*. Amsterdam, Netherlands: North-Holland; 1988.
13. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif*. 2014;31(3):274-295.
14. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S. Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern*. 2013;43(3):982-994.
15. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theory Methods*. 1974;3(1):1-27.
16. Gurrutxaga I, Albisua I, Arbelaitz O, et al. SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recogn*. 2010;43(10):3364-3373.
17. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227.
18. Kim M, Ramakrishna R. New indices for cluster validity assessment. *Pattern Recogn Lett*. 2005;26(15):2353-2363.
19. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern*. 1973;3(3):32-57.
20. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
21. Petitjean F, Ketterlin A, Gançarski P. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recogn*. 2011;44(3):678-693.
22. Sardá-Espinosa A. Comparing time-series clustering algorithms in R using the dtwclust package. Technical report; 2018.
23. Babbin SF, Velicer WF, Aloia MS, Kushida CA. Identifying longitudinal patterns for individuals and subgroups: an example with adherence to treatment for obstructive sleep apnea. *Multivar Behav Res*. 2015;50(1):91-108.
24. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2(1):193-218.
25. Lee IM, Shiroma EJ. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *Br J Sports Med*. 2014;48(3):197-201.
26. Fatouhi DE, Delrieu L, Goetzinger C, et al. Associations of physical activity level and variability with 6-month weight change among 26,935 users of connected devices: observational real-life study. *JMIR mHealth uHealth*. 2021;9(4):e25385.
27. Randerath W, Bassetti CL, Bonsignore MR, et al. Challenges and perspectives in obstructive sleep apnoea. *Eur Respir J*. 2018;52(3):1702616.
28. Pack AI. Application of personalized, predictive, preventative, and participatory (P4) medicine to obstructive sleep apnea. a roadmap for improving care? *Ann Am Thorac Soc*. 2016;13(9):1456-1467.

How to cite this article: Bottaz-Bosson G, Hamon A, Pépin J-L, Bailly S, Samson A. Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities. *Statistics in Medicine*. 2021;1–24. https://doi.org/10.1002/sim.9130

APPENDIX A. REAL CPAP DATA: DISTRIBUTION OF INDIVIDUAL ADHERENCE CHARACTERISTICS

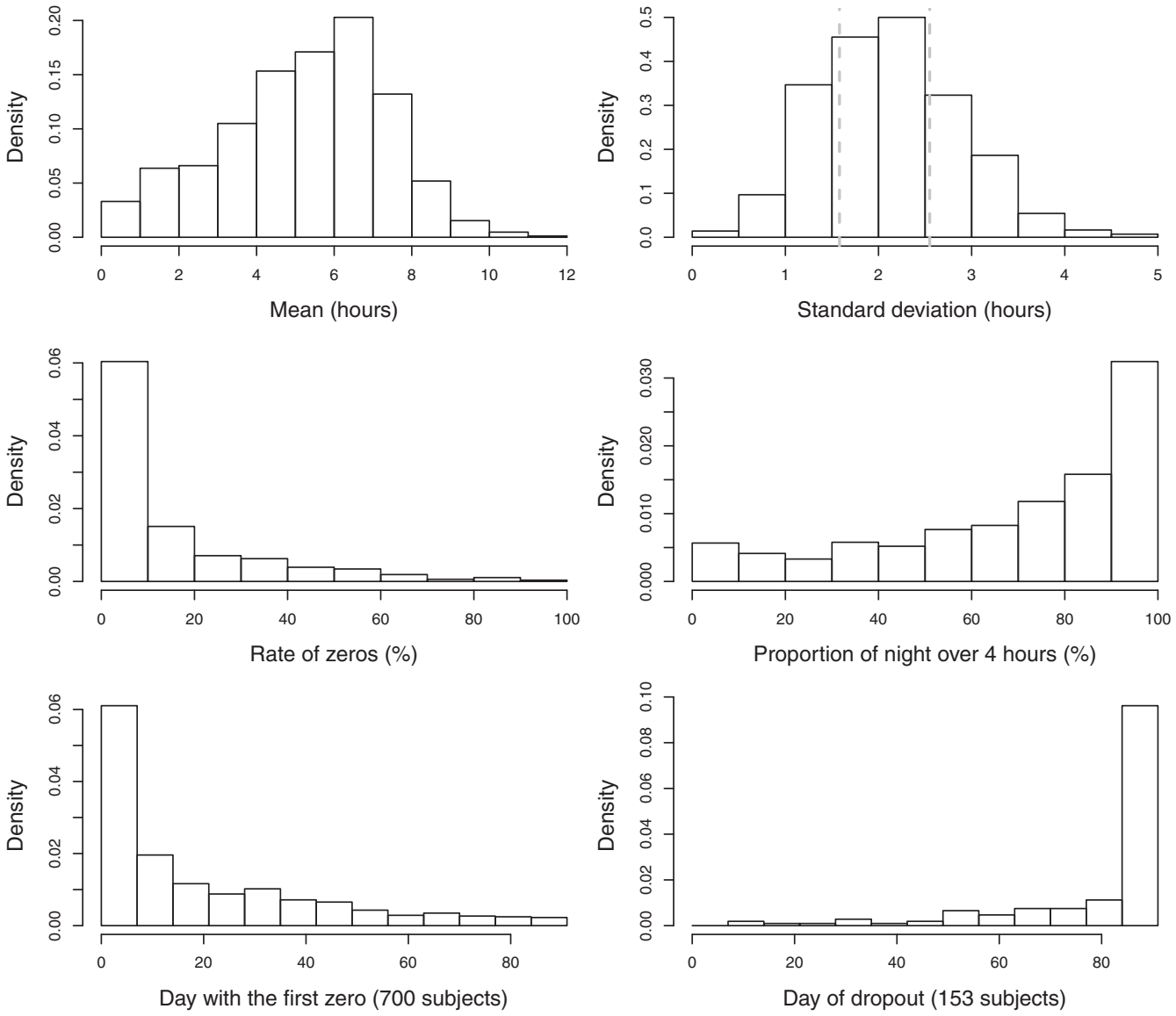


FIGURE A1 Distribution of individual adherence characteristics computed from the 91 days of observations: Mean (top-left), standard deviation (top-right), rate of null adherence (middle-left), proportion of nights with adherence over 4 hours (middle-right), day with the first zero (bottom-left), and day of dropout (bottom-right). The two dashed lines on the standard deviation plot represent the selected values used to generate the fictive datasets (with variance equal to 2.5 or 6.5)

APPENDIX B. MATHEMATICAL DEFINITIONS AND FORMULAS

B.1 Notations and preliminary definitions

Let $N \in \mathbb{N}^*$ the whole population size and $X := \{x_1, x_2, \dots, x_N\}$ be the whole dataset where $x_i, i \in [\![1, N]\!]$ is an individual's adherence time series. All time series are assumed to be of the same length $T \in \mathbb{N}^*$ and without missing data. For subject i, let $x_{i,t} \in \mathbb{R}$ represent her/his adherence value at the tth day of treatment, $t = 1, \dots, T$. So her/his adherence time series is $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,T}) \in \mathbb{R}^T$. For each time series x_i, let $\widetilde{x_i}$ be the polygonal curve specified by the ordered vertices sequence

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

A *partition* P with k clusters on X is a set of non-empty subsets of X, $P = \{C_1, C_2, \dots, C_k\}$ such that $\bigcup_{C_l \in P} C_l = X$ and $\forall l, l',\ l \neq l' \Rightarrow C_l \cap C_{l'} = \emptyset$.

For a chosen dissimilarity measure d, and two time series x and x', the dissimilarity between x and x' is expressed as $d(x, x')$ and the dissimilarity between two clusters C_l and $C_{l'}$ computed through a linkage strategy is $d(C_l, C_{l'})$.

Then, let us define a warping path and a coupling, that are both included in the definitions of DTW and sdF dissimilarities.

Definition 1. A *warping path* W of lengths m and n is a sequence $((a_1, b_1), (a_2, b_2), \dots, (a_{l_W}, b_{l_W}))$ of distinct pairs from $[\![1, m]\!] \times [\![1, n]\!]$ such that: l_W is an integer non inferior to $\max(m, n)$, $a_1 = b_1 = 1$, $a_{l_W} = m$, $b_{l_W} = n$ and $\forall l$ in $\{1, \dots, l_W - 1\}$, $(a_{l+1} - a_l, b_{l+1} - b_l) \in \{(0, 1), (1, 0), (1, 1)\}$.

Definition 2. Let $E = (e_1, e_2, \dots, e_m)$ and $F = (f_1, f_2, \dots, f_n)$ be two sets of ordered elements of size m and n respectively. Let $W = ((a_1, b_1), (a_2, b_2), \dots, (a_{l_W}, b_{l_W}))$ be a warping path of lengths m and n.

The sequence $L_W = \left((e_{a_1}, f_{b_1}), (e_{a_2}, f_{b_2}), \dots, (e_{a_{l_W}}, f_{b_{l_W}}) \right)$ is called a *coupling* between E and F.

B.2 Dissimilarity measures

Definition 3. Let x_i and x_j be two times series. The *Euclidean distance* δ_E between x_i and x_j is defined by:

$$\delta_E(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{\sum_{t=1}^{T} (x_{i,t} - x_{j,t})^2}.$$

Definition 4. Let x_i and x_j be two time series and $\mathcal{L}_{i,j}$ be the set of all possible couplings between x_i and x_j. The *dynamic time warping dissimilarity* d_{DTW} between x_i and x_j is defined by:

$$d_{DTW}(x_i, x_j) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{l=1}^{l_W} \|x_{i,a_l} - x_{j,b_l}\|_2 .$$

Definition 5. Let $\widetilde{x_i}$ and $\widetilde{x_j}$ be the polygonal curves associated with the time series x_i and x_j and $\mathcal{L}_{i,j}$ be the set of all possible couplings between $\widetilde{x_i}$ and $\widetilde{x_j}$. Let $\lambda \in \mathbb{R}^+$ and $\Lambda : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\Lambda(a, b) = (\lambda a, b)$. The *generalized discrete Fréchet distance of parameter* λ, d_{dF_λ} between $\widetilde{x_i}$ and $\widetilde{x_j}$ is defined by:

$$d_{dF_\lambda}(\widetilde{x_i}, \widetilde{x_j}) = \min_{L_W \in \mathcal{L}_{i,j}} \max_{l=1}^{l_W} \|\Lambda(a_l, x_{i,a_l}) - \Lambda(b_l, x_{j,b_l})\|_2 = \min_{L_W \in \mathcal{L}_{i,j}} \max_{l=1}^{l_W} \|\Lambda(\widetilde{x_{i,a_l}}) - \Lambda(\widetilde{x_{j,b_l}})\|_2 .$$

Thus the *generalized summed discrete Fréchet dissimilarity of parameter* λ, d_{sdF_λ} between $\widetilde{x_i}$ and $\widetilde{x_j}$ is defined by:

$$d_{sdF_\lambda}(\widetilde{x_i}, \widetilde{x_j}) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{l=1}^{l_W} \|\Lambda(\widetilde{x_{i,a_l}}) - \Lambda(\widetilde{x_{j,b_l}})\|_2 .$$

B.3 Linkage strategies

The "average" linkage strategy defines the dissimilarity between two clusters C_l and $C_{l'}$ as:

$$d(C_l, C_{l'}) = \frac{1}{|C_l| \times |C_{l'}|} \sum_{x \in C_l} \sum_{x' \in C_{l'}} d(x, x').$$

The "complete" linkage strategy defines the dissimilarity between two clusters C_l and $C_{l'}$ as:

$$d(C_l, C_{l'}) = \max_{(x,x') \in (C_l \times C_{l'})} d(x, x').$$

The "Ward" linkage strategy recursively defines the dissimilarity between a merged cluster $C_{l'} \cup C_{l''}$ and a third cluster C_l as:

$$d^2(C_{l'} \cup C_{l''}, C_l) = \frac{|C_{l'}| + |C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l'}, C_l) + \frac{|C_{l''}| + |C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l''}, C_l) - \frac{|C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l'}, C_{l''}),$$

where $|.|$ denotes the number of elements of a cluster.

If two clusters are singletons then the dissimilarity value between these two clusters is set with the dissimilarity value between the corresponding objects.

APPENDIX C. SIMULATION STUDY: DISTRIBUTION OF NUMBER OF CLUSTERS RETURNED FOR EACH METHOD

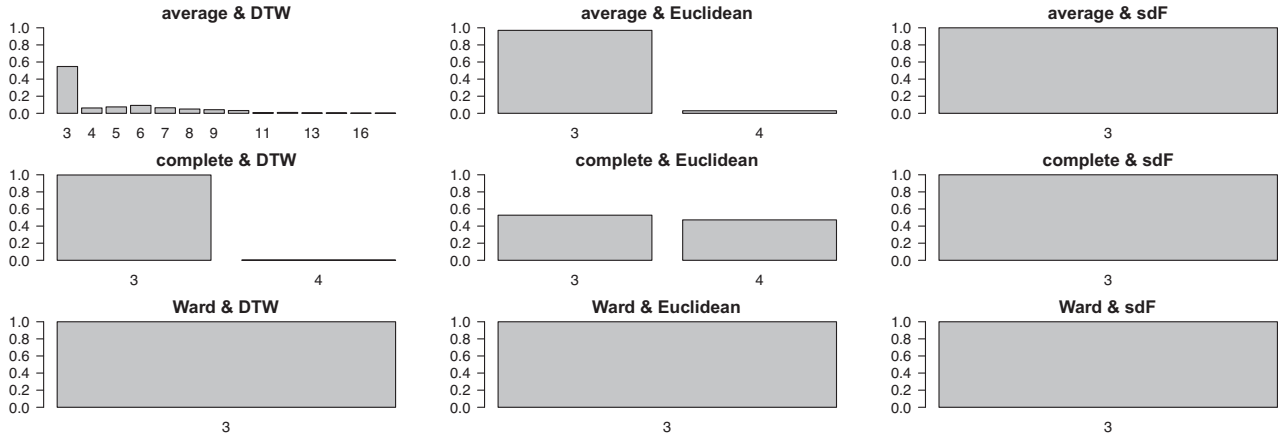


FIGURE C1 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 3 groups and without outliers

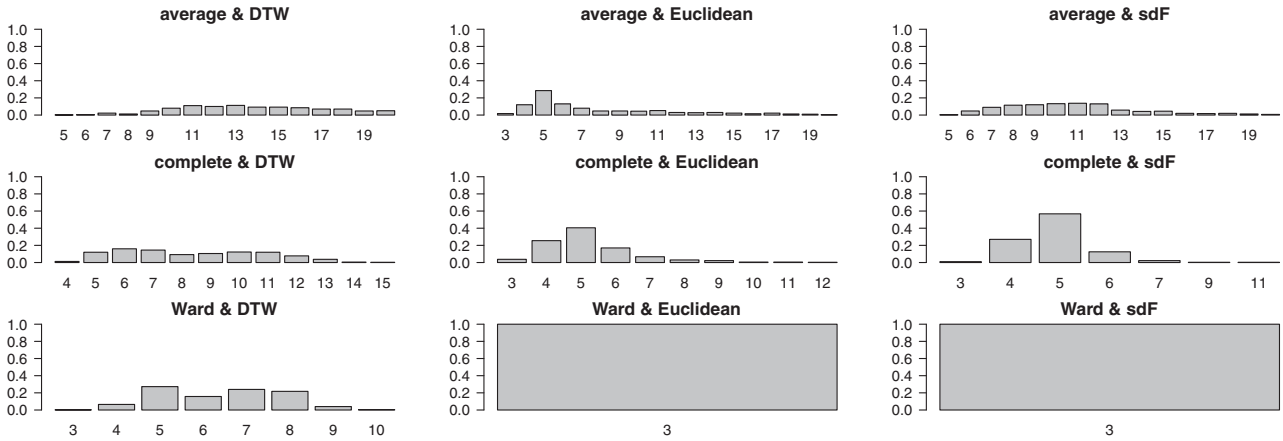


FIGURE C2 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 3 groups and 10% of outliers

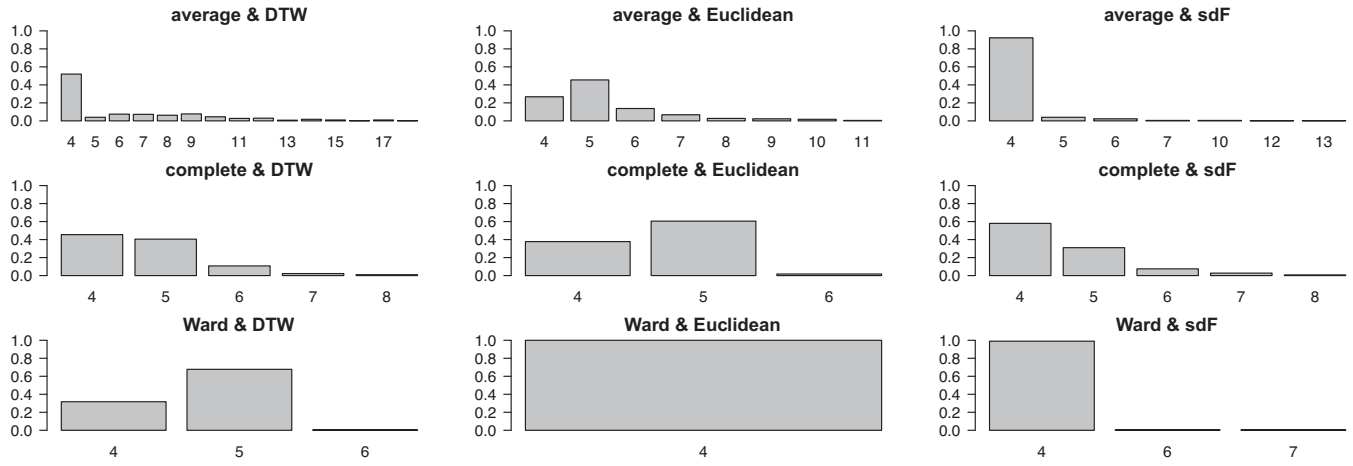


FIGURE C3 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 5 groups and without outliers

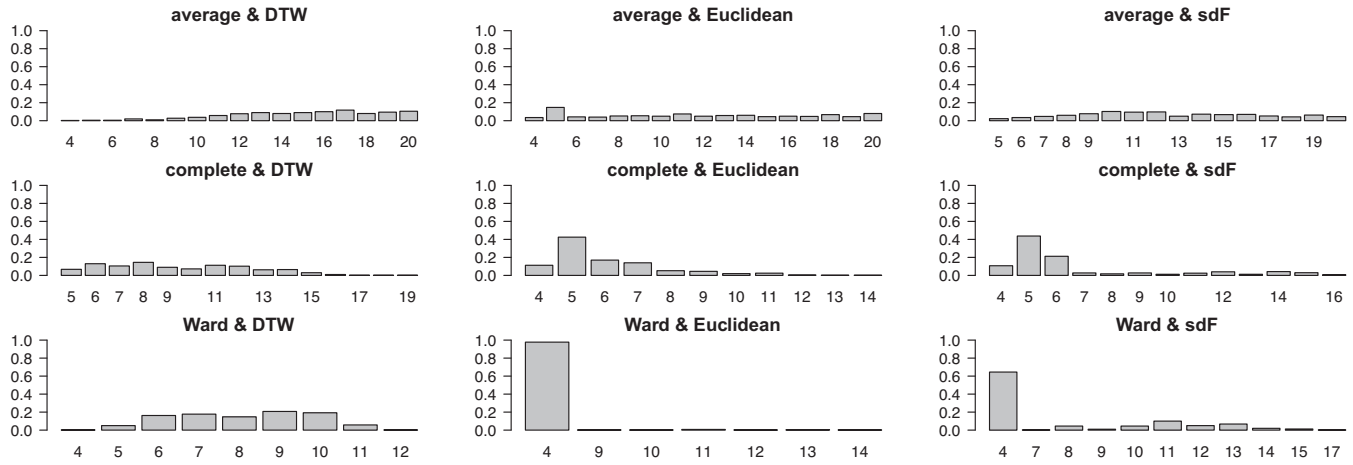


FIGURE C4 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 5 groups and 10% of outliers

APPENDIX D. REAL CPAP DATA: MEDOID REPRESENTATIVES OF THE 6 CLUSTERS

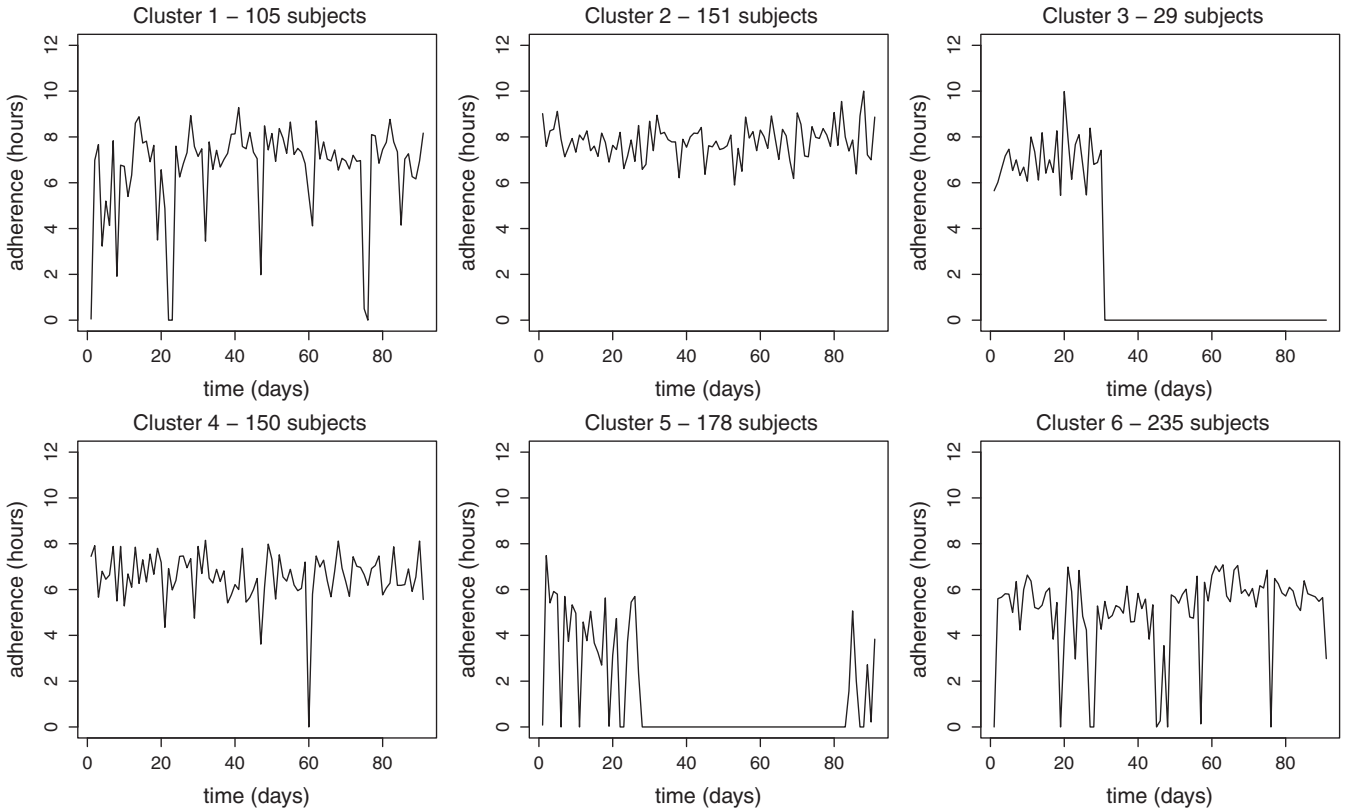


FIGURE D1 Real CPAP data: Medoid representatives of the 6 clusters

Chapitre 4

Développement d'une application web pour la visualisation de clusters de séries temporelles.

4.1 Introduction

On considère que les comportements typiques d'adhésion à la PPC sont décrits par des clusters de séries chronologiques. Leur représentation visuelle est une solution pour dévoiler leurs significations cliniques. Des méthodes classiques pour la visualisation de clusters (de séries temporelles ou non) incluent le dendrogramme dans le cas de la CAH ou le nuage de points d'une projection des observations dans un sous-espace de dimension réduite, comme après une analyse en composantes principales [60] pour des données quantitatives classiques. Ces graphiques aident à l'interprétation des clusters en les situant les uns par rapport aux autres mais ne donnent pas une image concrète des objets constituant chaque cluster.

Lorsque les observations sont des séries chronologiques, chacune admet des attributs propres : niveau de départ, tendance, cyclicité(s) périodiques(s) ou non, variabilité, présence de valeurs extrêmes, ou encore rupture(s) entraînant une ou des modification(s) de certains attributs dans le temps. La combinaison de différents attributs peut former des "motifs" visibles sur les courbes individuelles. À notre sens, un des enjeux de la représentation graphique d'un cluster de séries chronologiques est, en prenant en compte le caractère temporel des données, de faire ressortir les attributs et/ou motifs individuels partagés par les séries du cluster, sans masquer la disparité dans le cluster. La transparence sur cette disparité est primordiale pour réduire le risque de mauvaise interprétation. Deux manières fréquentes pour représenter visuellement un cluster de séries temporelles sont le tracé des courbes de ses séries, avec notamment un diagramme spaghetti qui les superpose sur un même graphique [61, 62, 63] ; et/ou la visualisation de la courbe d'un représentant du cluster. Dans [63] chaque cluster est graphiquement représenté par sa trajectoire moyenne en plus des diagrammes spaghetti.

Dans notre cas, nous traitons un nombre assez élevé de séries chronologiques. Ceci rend inconvenant la juxtaposition de toutes les courbes d'un cluster. De plus, certaines séries présentent de grandes variabilités, des discontinuités ou des ruptures. Le diagramme spaghetti est inexploitable. La courbe d'un représentant ne peut suffire à reconstituer fidèlement l'ensemble des attributs ou motifs individuels

caractéristiques d'un cluster, leurs temporalités, tout en restituant la diversité au sein du cluster. Pour tendre vers cela, il nous semble judicieux de recourir à plusieurs représentations graphiques, apportant chacune des informations spécifiques. Ceci permettrait de fournir des angles de vue complémentaires sur les clusters.

Des packages informatiques existent pour réaliser le clustering de séries chronologiques. Dans l'environnement Python (www.python.org), la libraire **Tslearn** [64] permet plus généralement l'application de méthodes d'apprentissage statistique aux séries temporelles, et permet de les traiter avec l'interface de programmation d'application **scikit-learn** [65]. Avec R (www.r-project.org) nous pouvons citer les packages **TSclust** [66], **TSdist** [67] ou encore **dtwclust** [59]. Dans les deux premiers sont implémentées de nombreuses mesures de dissimilarité et quelques fonctionnalités pour étendre l'utilisation d'algorithmes de classification classiques aux séries temporelles. Le package **dtwclust** propose une procédure couvrant plusieurs étapes du clustering en intégrant différents algorithmes dans un cadre uniformisé. Beaucoup de fonctions du package sont liées à l'algorithme DTW, mais il est possible de personnaliser la procédure et de la combiner avec d'autres packages. Les étapes incluses comprennent le pré-traitement des données ; la classification avec le calcul des dissimilarités et de représentants des clusters le cas échéant ; l'évaluation de la classification et la visualisation des clusters obtenus avec des diagrammes spaghetti et/ou les représentants des clusters. Les graphiques implémentés correspondent à ce qui est fréquemment proposé dans la littérature, et ne sont pas suffisants dans notre cas.

À notre connaissance, aucun outil informatique existant n'est dédié à la visualisation de clusters de séries chronologiques. Notre besoin de diversifier les représentations graphiques des clusters de trajectoires d'observance motive la conception d'une application web que nous présentons dans ce chapitre. Elle est créée pour les chercheurs du laboratoire HP2 comme un outil graphique d'aide pour étudier l'adhésion des patients démarrant une thérapie PPC. Les chercheurs en milieu hospitaliers ne disposent pas nécessairement de compétences en programmation statistique pour produire des graphiques. L'application est réalisée avec le package **Shiny** [68] de l'environnement de programmation R. Cela offre la possibilité de déployer un outil sur mesure, facilement utilisable, et permettant d'exploiter les possibilités graphiques du langage R, tant sur la diversité des représentations graphiques, que sur les possibilités de personnalisation des figures. La version actuelle de l'application web a été réalisée avec la contribution de 5 étudiants que j'ai encadrés lors de leur première année de Master, en cursus "Statistiques et Sciences des Données" à l'université Grenoble-Alpes. Dans le cadre d'un projet tutoré entre Novembre 2020 et Avril 2021, quatre étudiants Théo Silvestre, Liwa Fleury, Nabil Mehdaoui, et Fabien Jossaud ont produit les différents graphiques et les ont intégrés dans une première version de l'application. David Ferrero, stagiaire de mai à août 2021, a ensuite amélioré l'interface graphique et procédé à une restructuration du code de l'application pour une meilleure compatibilité de celle-ci avec un dépôt sur un serveur.

La Section 4.2 introduit la problématique de représentation pour des clusters de séries temporelles. La Section 4.3 est dédiée à la présentation de l'application et de ses fonctionnalités. Les deux sections suivantes sont consacrées aux visualisations proposées, où la Section 4.5 porte sur des fonctionnalités plus avancées de représentation d'indicateurs statistiques individuels. La Section 4.6 propose une synthèse avec un tableau récapitulatif et comparatif des différentes représentations graphiques.

4.2 Problématique de la représentation de clusters de séries chronologiques

Nous commençons cette section par une clarification de terminologie. La visualisation d'un cluster consiste pour nous à résumer et/ou décrire un cluster en utilisant des graphiques. Nous employons indifféremment les terminologies "représentation visuelle" ou "représentation graphique" d'un cluster pour désigner cette pratique. Cela ne doit pas être confondu avec ce que nous désignons plus simplement par la "représentation" d'un cluster, qui consiste à utiliser un représentant de cluster, aussi appelé "prototype" dans la littérature anglo-saxonne. Il s'agit d'un élément situé dans l'espace des observations ("a data object that is representative of the objects in the cluster" [69]), souvent "central" pour le cluster. Au delà de l'apport pour la synthétisation et la visualisation des clusters (en particulier avec des séries chronologiques), le choix des représentants est une étape fréquente du clustering. Ils sont au cœur de certains algorithmes privilégiant la formation de clusters convexes organisés autour de centres. Ils sont également impliqués dans le calcul d'indices de validation internes valorisant les clusters construits selon cette logique de regroupement.

Lorsque les observations $x_1, ..., x_n$ sont vues comme des éléments de $\mathbb{R}^t$ muni de la norme euclidienne, le centre d'un cluster C est l'isobarycentre $\overline{C}$ [1] des observations qui constituent ce cluster [45, 58, 70, 71, 72]. Ce représentant satisfait l'Équation 4.1 :

$$\overline{C} = \underset{\xi \in \mathbb{R}^t}{\operatorname{argmin}} \sum_{x_i \in C} ||\xi - x_i||_2^2. \tag{4.1}$$

Dans le cadre de l'application web, les observations sont des séries chronologiques dont certaines peuvent être caractérisées par des dépendances et/ou évolutions temporelles. Elles peuvent aussi présenter des motifs communs sans que ces motifs soient réalisés aux mêmes instants, comme les trajectoires d'abandon de traitement rencontrées dans le contexte clinique de nos travaux. Par l'utilisation d'une mesure de dissimilarité adéquate, que l'on note d, le regroupement des observations peut tenir compte de ces dépendances, et/ou s'intéresser à l'ordre dans lequel les motifs apparaissent plutôt qu'à leurs temps d'apparition. Si les séries chronologiques sont de même longueur alors il est possible de calculer le représentant défini en Équation 4.1. Cependant rien n'assure qu'il soit "central" pour d. Définir un représentant $\widetilde{C}$ satisfaisant une équation analogue à l'Équation 4.1 n'est pas trivial. Il peut ne pas être calculable temps par temps (contrairement au cas Euclidien). Il peut même, selon la dissimilarité utilisée, être autorisé à comporter un nombre de coordonnées différent de celui des séries classifiées. Il faut alors résoudre l'Équation 4.2 :

$$\widetilde{C} = \underset{\xi \in \mathbb{R}^q, q \in \mathbb{N}^*}{\operatorname{argmin}} \sum_{x_i \in C} d^2(\xi, x_i). \tag{4.2}$$

Résoudre cette équation est algorithmiquement complexe. Une alternative est de représenter un cluster avec son médoïde m_C, qui minimise la somme des dissimilarités aux autres observations du cluster, sans appliquer la fonction carré aux dissimilarités [73]. Mathématiquement, le médoïde satisfait l'Équation 4.3 :

$$m_C = \underset{x_{i'} \in C}{\operatorname{argmin}} \sum_{x_i \in C} d(x_{i'}, x_i). \tag{4.3}$$

1. On le note également $\overline{C}$ mais il ne doit pas être confondu avec celui introduit en Définition 4.

Dans tous les cas, utiliser des représentants "au centre" des clusters implique l'hypothèse de convexité des clusters pour la dissimilarité d, et admet un sens lorsque les clusters sont effectivement construits autour de centres. Pour ces raisons, et considérant, d'une part la difficulté de déterminer ces "centres" dans un cas non Euclidien, et d'autre part l'incapacité pour un représentant de capter la diversité des séries d'un cluster, nous ne nous concentrons pas uniquement sur la représentation visuelle des clusters à travers des éléments "centraux". Certaines représentations graphiques ont pour vocation de décrire les clusters, et non pas de les représenter à l'aide d'objets interprétables comme des séries chronologiques.

4.3 Présentation de l'application

La version actuelle de l'application est exécutable sur la version 4.0.4 de R, avec la version 1.6.0 du package **shiny**. Ses fonctionnalités que nous présentons ci-après sont illustrées en Figure 4.1. Après le téléversement d'un ensemble de trajectoires journalières individuelles regroupées en clusters ❶, l'application propose différentes représentations graphiques des clusters ❷. Les valeurs des trajectoires doivent être indicées par leurs temps de réalisation (i.e. la position de ces valeurs dans les trajectoires). Les trajectoires doivent de plus être de même longueur et comporter les mêmes temps de réalisation sans données manquantes. Pour plus de détails techniques, comme sur le format attendu des données à importer, se référer au guide d'utilisation fournit en Annexe A. Les graphiques sont personnalisables et exportables au format ".png". L'application permet d'explorer visuellement divers regroupements de trajectoires par la prise en compte de covariables modifiables (par exemple, en groupant des modalités d'une variable qualitative, ou en discrétisant une variable continue) ❸. Enfin il est possible de calculer des indicateurs statistiques par trajectoire et de représenter ces indicateurs par cluster ❹. Pour permettre l'utilisation de ces graphiques avec des clusters de trajectoires issus d'autres domaines d'étude, les codes de calcul de ces indicateurs sont à importer depuis un fichier ".Rdata".

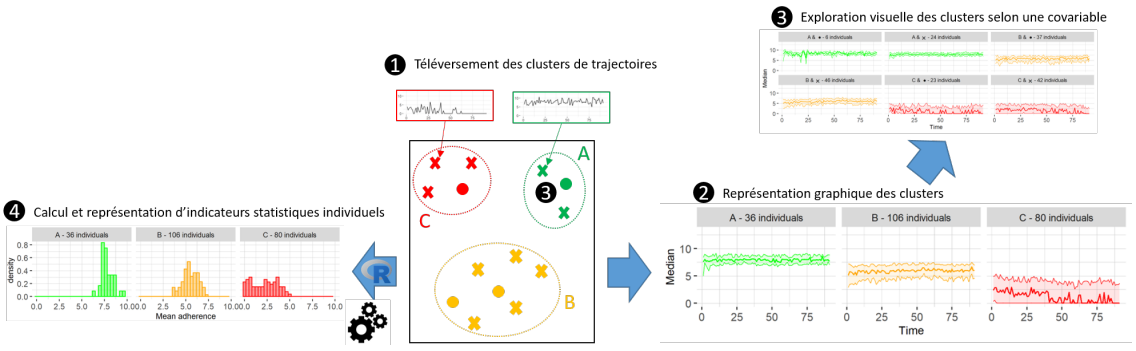


FIGURE 4.1 – Schéma récapitulatif des fonctionnalités offertes par l'application informatique. Le logo R indique lorsque des compétences en programmation R sont requises. L'engrenage matérialise lorsque l'application doit procéder à des calculs. Les chiffres en blanc sur fond rond noir sont des marques utilisées pour des renvois depuis le corps du texte de la section.

4.4 Représentations graphiques des clusters

Les différentes représentations graphiques proposées dans cette section se répartissent en 4 catégories, constituant des sous-sections propres. Elles comportent la visualisation de trajectoires individuelles, de trajectoires d'évolution d'un indicateur de tendance centrale quotidien, de diagrammes en boîtes et de heatmaps. La plupart des représentations graphiques proposées sont illustrées sur les 6 clusters identifiés dans la Section 3.2, en associant une couleur à chaque cluster et en la conservant lorsque cela est possible.

4.4.1 Trajectoires individuelles

Ces graphiques représentent des séries chronologiques sous forme de courbes, restituant leur dimension temporelle et rendant compte de détails observables au niveau individuel. Il est premièrement possible de visualiser une trajectoire individuelle spécifiée par son identifiant. Puis, plusieurs trajectoires individuelles tirées aléatoirement parmi une sélection de clusters peuvent être juxtaposées. Sous condition de sélectionner un échantillon suffisamment grand, ceci permet de disposer d'un aperçu de la diversité des trajectoires au sein d'un cluster. La Figure 4.2 montre 12 trajectoires prises au hasard dans le cluster 3. On peut voir un motif systématiquement présent : le passage brutal, et à une date quelconque, d'une bonne utilisation de la machine à un arrêt d'utilisation complet jusqu'à la fin des trois premiers mois. Les séries de ce cluster semblent caractérisées par la présence d'une rupture. On note de plus qu'excepté un bon niveau d'adhésion, il n'apparaît pas de particularité spécifique avant les arrêts. Certaines trajectoires sont plutôt stables (séries numéro 386 ou 587) quand d'autres sont très irrégulières, sans zéro (série numéro 719) ou avec (séries numéro 463 ou 585). D'autres encore semblent admettre une tendance croissante (séries 350 ou 39). La Figure 4.3 montre 12 autres trajectoires aléatoires issues du cluster 5. Il peut être plus difficile ici d'identifier une caractéristique ty-

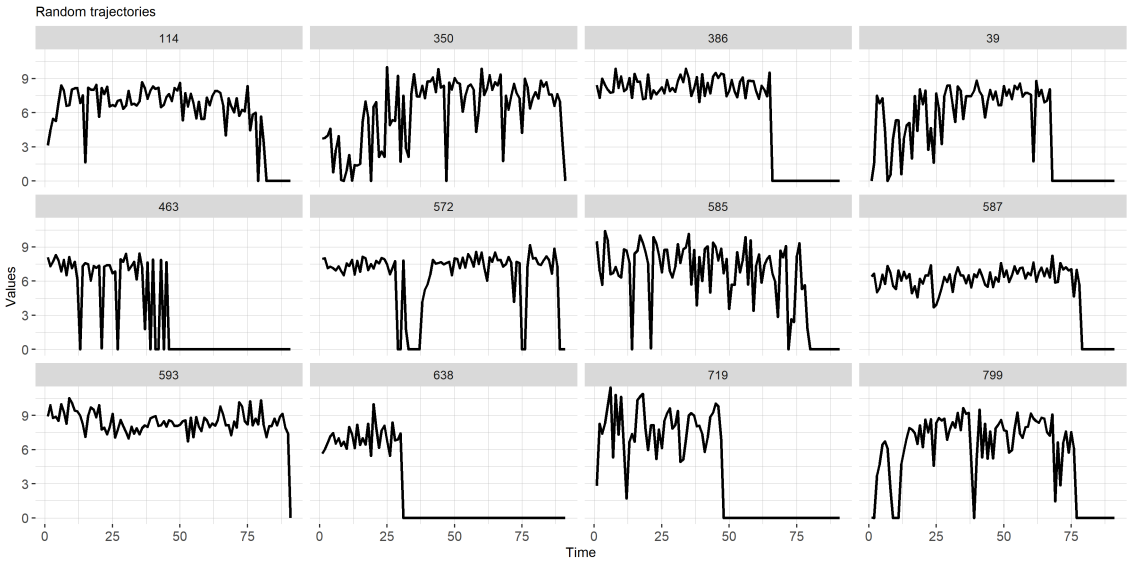


FIGURE 4.2 – Représentation graphique juxtaposée de 12 trajectoires aléatoires du troisième cluster d'adhésion. Figure exportée depuis l'application.

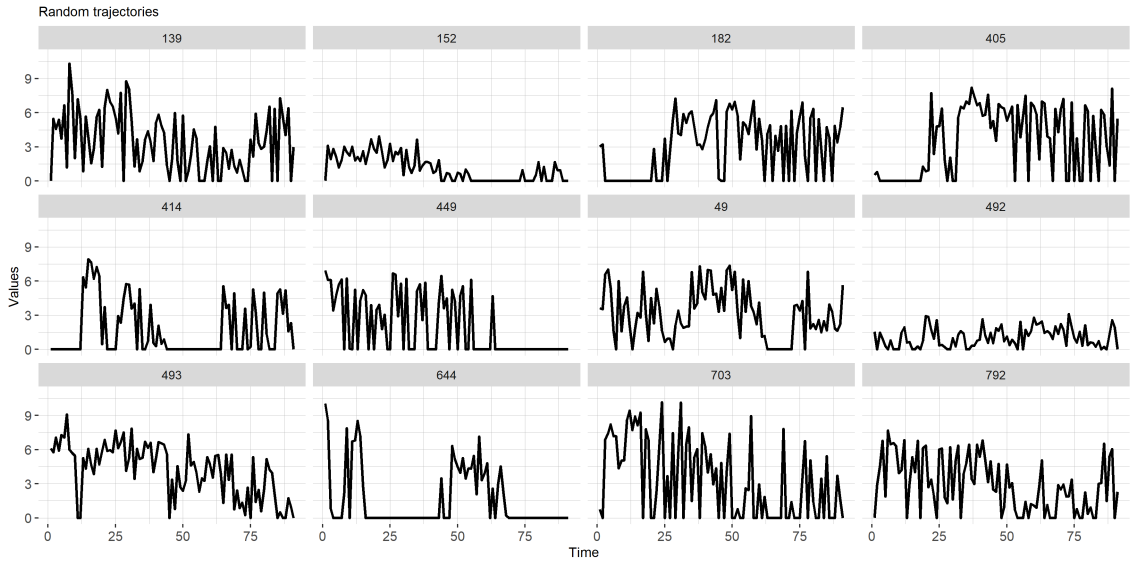


FIGURE 4.3 – Représentation graphique juxtaposée de 12 trajectoires aléatoires du cinquième cluster d'adhésion. Figure exportée depuis l'application.

pique. Le cluster regroupe des trajectoires erratiques ayant une grande variabilité et beaucoup de discontinuités, pouvant masquer d'autres caractéristiques individuelles, avec des trajectoires présentant une adhésion régulièrement faible (séries numéro 152 et 492). Ce qui semble caractériser ces trajectoires est leur nombre élevé de passages à zéros, ces derniers étant moins visibles lorsque les valeurs des séries sont basses. Cet exemple montre que représenter graphiquement ainsi un cluster n'assure pas l'extraction d'une information synthétique. De plus il est délicat d'avoir une vue d'ensemble et de comparer différents clusters si plusieurs courbes sont nécessaires pour robustement illustrer un même cluster.

Médoïde

Du fait de sa position la plus proche de toutes les observations d'un cluster, le médoïde est lié à l'utilisation d'une mesure de dissimilarité et est pertinent pour représenter un cluster convexe construit autour d'un centre. Nous subordonnons leur représentation graphique au dépôt par l'utilisateur [2] d'une matrice de dissimilarité. Celle utilisée pour réaliser la classification est le cas échéant attendue. L'import de la matrice plutôt que son calcul dans l'application procure deux avantages. Le premier est que cela assure de sélectionner les séries chronologiques effectivement situées les plus au centre des clusters importés. En effet, l'utilisateur de l'application peut modifier la sélection des séries ainsi que leur longueur. Retirer des séries d'un cluster peut avoir l'effet de déplacer le centre du cluster, et faire varier leur longueur risque surtout de bouleverser leurs dissimilarités relatives. Le deuxième avantage à ne pas calculer la matrice de dissimilarité dans l'application est de permettre une visualisation des médoïdes quels que soient l'algorithme de classification et la mesure de dissimilarité employés.

2. La formulation "l'utilisateur" peut désigner toute utilisatrice ou tout utilisateur potentiel.le de l'application, sans aucune distinction de genre.

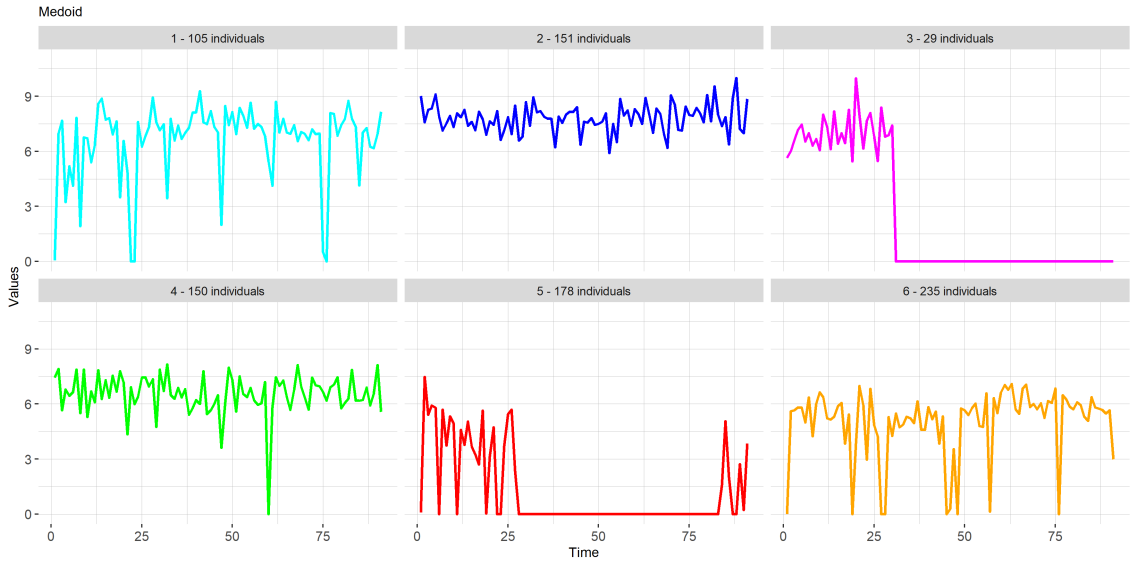


FIGURE 4.4 – Représentation graphique des médoïdes juxtaposés des 6 clusters d'adhésion. Figure exportée depuis l'application.

S'agissant de trajectoires réelles, elles offrent l'avantage de leur simplicité d'interprétation et permettent une comparaison facile des clusters par juxtaposition ou superposition des courbes. Cependant, il serait réducteur de résumer un cluster avec une seule série chronologique. La Figure 4.4 représente graphiquement les médoïdes des 6 clusters d'adhésion de manière juxtaposée. Les représentants des clusters 4 et 6 (en vert et orange) semblent fidèles aux indicateurs présentés dans la Table 3 de l'article inséré dans la Section 3.2 : ils se distinguent par leurs niveaux, leurs variabilités, et sur leurs nombres de zéros. Attention toutefois à ne pas formuler de conclusions erronées de la comparaison des médoïdes. Celui du cluster 3 (en magenta) est conforme au motif d'arrêt de traitement caractéristique du cluster. Cependant, en termes de variabilité intra-trajectoire ou de discontinuités, la partie de la trajectoire antérieure à l'arrêt n'est pas représentative de toutes les trajectoires de la Figure 4.2. Aussi, le médoïde du cluster 5 est caractérisé par une longue séquence de zéros. Cette période est excessivement longue relativement aux autres trajectoires de la Figure 4.3. Ceci peut s'expliquer par l'algorithme DTW standard utilisé pour la classification : avec les dilatations des axes temporels, il ne fait pas la différence entre un arrêt d'un jour et un arrêt de plusieurs jours consécutifs. Il est donc recommandé d'avoir une bonne intuition des spécificités de la mesure de dissimilarité utilisée lors de l'interprétation des médoïdes.

Diagramme Spaghetti

La dernière représentation visuelle utilisant les trajectoires individuelles est le diagramme spaghetti. Nous proposons ce graphique car il peut être très efficace dans certaines situations [61, 62, 63]. Son interprétabilité est dépendante des caractéristiques des séries chronologiques, de leur nombre et de leur philosophie de regroupement. Pour une meilleure lisibilité, les couleurs propres aux différents clusters ne sont pas reprises : de nouvelles couleurs sont utilisées pour distinguer les différentes trajectoires superposées. La Figure 4.5 représente les diagrammes spa-

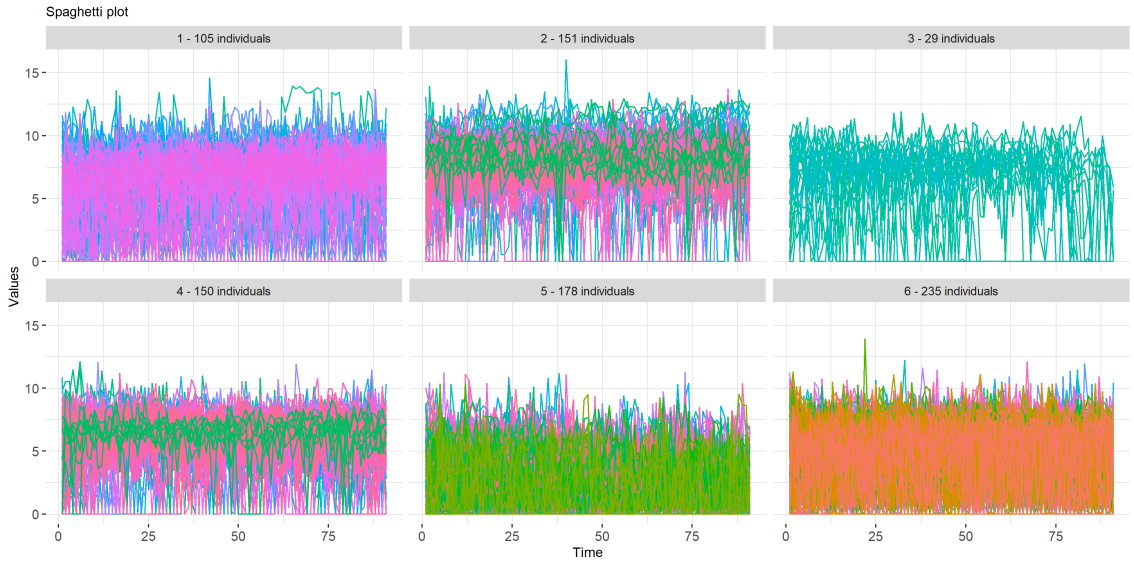


FIGURE 4.5 – Représentation visuelle des 6 clusters d'adhésion avec des diagrammes spaghetti. Figure exportée depuis l'application.

ghetti des 6 clusters d'adhésion. Il est difficile d'en déduire des informations sur les trajectoires individuelles qui les composent.

4.4.2 Trajectoires d'évolution d'un indicateur de tendance centrale quotidien

Les graphiques de cette sous-section tracent la courbe d'un indicateur de tendance centrale calculé temps par temps sur l'ensemble des séries d'un cluster. Nous appelons ces séquences des trajectoires bien qu'elles ne reflètent pas les caractéristiques individuelles des séries chronologiques qu'elles représentent. En particulier, elles gomment les variabilités intra-individuelles et des comportements différents peuvent être résumés par une courbe n'incarnant aucun de ces comportements. En revanche, ces trajectoires sont représentables avec des indicateurs de dispersion tenant compte de l'évolution des variabilités inter-trajectoires au sein d'un cluster. Les courbes peuvent être juxtaposées, ou superposées sur un même graphique. Les indicateurs proposés sont la moyenne avec son intervalle de confiance, et la médiane et l'écart inter-quartile.

Trajectoire moyenne

La trajectoire moyenne est le représentant $\overline{C}$, satisfaisant l'Équation 4.1 dans le cas d'utilisation de la distance Euclidienne. Dans cette situation, cette trajectoire fictive est interprétable comme la série chronologique située au "centre" du cluster. Cependant l'utilisation de cette séquence ne se cantonne pas au cas Euclidien. La Figure 4.6 représente graphiquement les trajectoires moyennes des 6 clusters d'adhésion. On remarque premièrement que tous les clusters admettent une adhésion moyenne plus faible au temps 1 qu'au temps 2, de manière plus ou moins prononcée selon les clusters. La première journée de traitement semble être un temps d'adaptation quels que soient les profils d'adhésion des patients. Dans la suite, nous

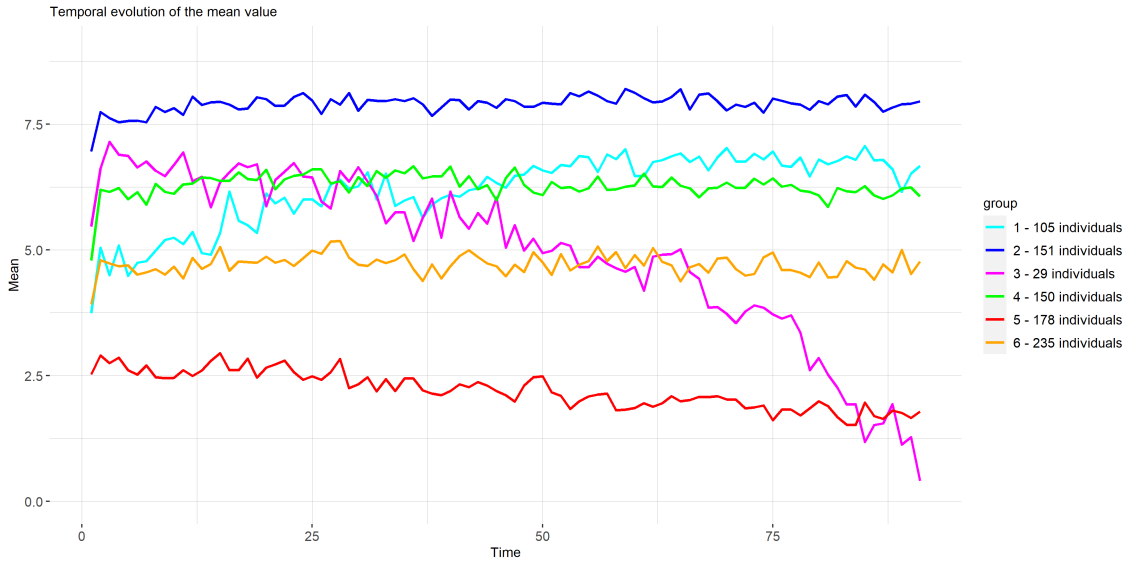


FIGURE 4.6 – Représentation visuelle des 6 clusters d'adhésion avec les trajectoires moyennes superposées. Figure exportée depuis l'application.

regardons donc l'évolution de la moyenne à partir du deuxième jour de traitement. Trois clusters ont une moyenne stable, du plus haut niveau au plus bas : le cluster 2 (en bleu), le cluster 4 (en vert) et le cluster 6 (en orange). Le cluster 1 (en cyan) admet une moyenne croissante se stabilisant à un bon niveau. Enfin les deux clusters restants admettent des moyennes décroissantes, celle du cluster 3 (en magenta) plus fortement car commençant plus haut et finissant plus bas que celle du cluster 5 (en rouge). L'affichage superposé montre que l'adhésion moyenne des individus du cluster 1 (en cyan) semble commencer comme celle du cluster 6 (en orange) pour finir au niveau de celle des sujets regroupés dans le cluster 4 (en vert).

Afin de fiabiliser la comparaison des moyennes entre les clusters, il est pratique de matérialiser l'évolution de l'intervalle de confiance ponctuel associé au niveau 95%. Ceci a aussi l'effet de relativiser les variations des moyennes d'un jour sur l'autre au sein d'un cluster. L'intervalle de confiance calculé est asymptotique pour ne nécessiter aucune hypothèse sur les distributions quotidiennes des valeurs. Sa fiabilité requiert néanmoins un nombre de séries temporelles suffisamment élevé. Il convient de borner les intervalles de confiance pour que l'estimation de la moyenne reste sur le support de la variable aléatoire étudiée, ici entre 0 et 24 heures pour le temps d'utilisation quotidien de l'appareil de PPC. Sur la Figure 4.7 sont représentées les trajectoires moyennes avec intervalles de confiance des clusters 1, 4 et 6. Ce graphique confirme les observations faites précédemment sur l'évolution de la moyenne du cluster 1 (en cyan).

Interpréter seulement les trajectoires moyennes ne permet pas de déduire beaucoup d'informations sur les clusters. Une moyenne de cluster stable n'étant pas par exemple une condition suffisante pour que toutes les séries chronologiques du cluster aient une tendance stable. Ceci justifie l'utilisation de trajectoires d'autres indicateurs statistiques.

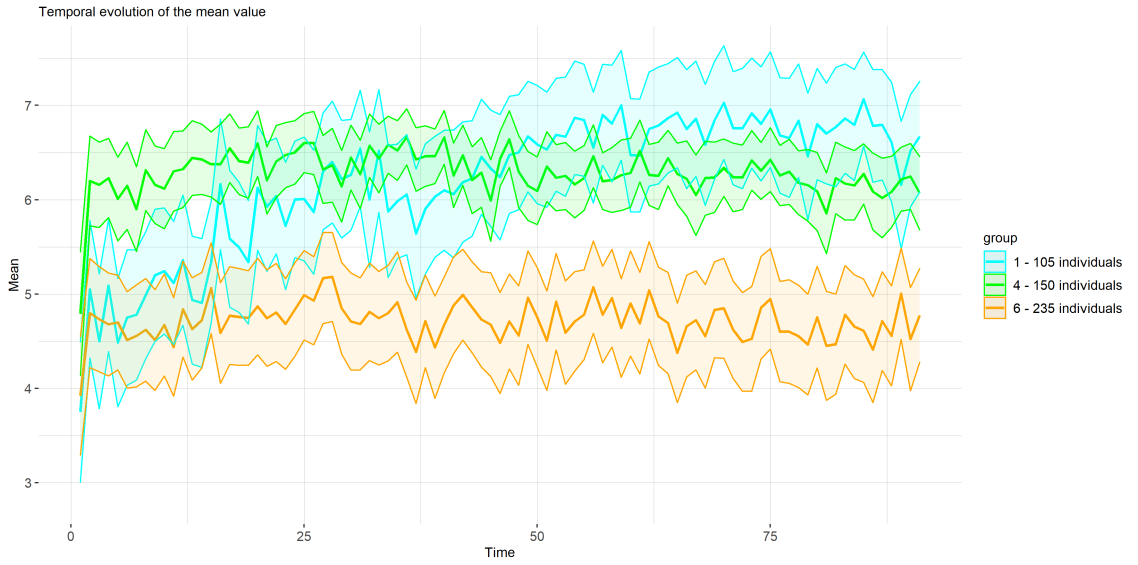


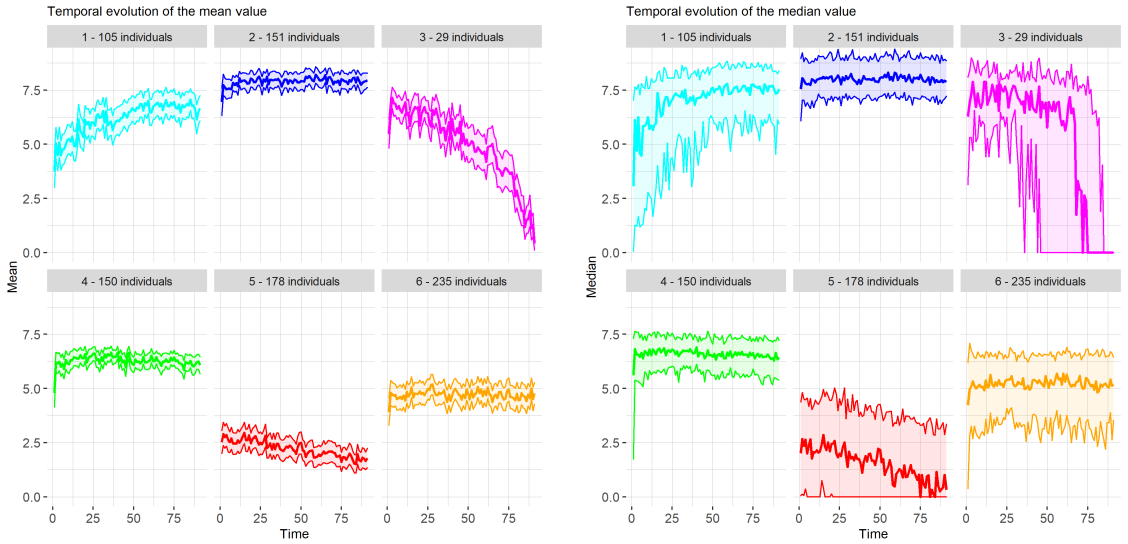
FIGURE 4.7 – Représentation graphique superposée des trajectoires moyennes avec intervalles de confiance ponctuels asymptotiques de niveau 95%, pour les clusters 1, 4 et 6. Figure exportée depuis l'application.

Trajectoire médiane

Avec la trajectoire médiane, la dispersion des valeurs temps par temps au sein d'un cluster peut être représentée en ajoutant la bande inter-quartiles. Ces bandes sont plus instructives sur les variabilités inter-trajectoires au sein des clusters que ne le sont les intervalles de confiance ponctuels des moyennes. Elles permettent de mettre en relief des différences de variabilité inter-trajectoires entre les clusters, de montrer d'éventuelles asymétries des distributions autour des tendances centrales, et de comparer les évolutions relatives des médianes et des quartiles. Leur représentation visuelle est pertinente pour un cluster si l'ensemble des valeurs possibles est majoré et/ou minoré et que l'atteinte d'une borne est informative. Par exemple, la médiane est égale à cette borne les jours où au moins 50% des valeurs l'atteignent.

Aussi, sous certaines conditions, cette représentation graphique est plus apte à illustrer des ruptures caractéristiques d'un cluster. Prenons par exemple le cas d'une rupture entre deux phases stables, se produisant de manière identique à différents instants sur un ensemble de séries chronologiques. La trajectoire moyenne va évoluer de manière continue entre le niveau d'avant la rupture et celui d'après, selon la proportion de séries qui ont atteint cette rupture à chaque temps. En revanche, il est attendu que la trajectoire des quantiles d'ordre q restitue cette rupture au temps où la proportion de séries ayant passé la rupture atteint q.

La Figure 4.8 compare les trajectoires moyennes avec intervalles de confiance ponctuels (a) et les trajectoires médianes avec bandes inter-quartiles (b) des 6 clusters d'adhésion de manière juxtaposée. Premièrement, les trajectoires médianes et les bandes inter-quartiles marquent également le temps d'adaptation correspondant au premier jour de traitement, dont nous ne tiendrons toujours pas compte pour analyse les courbes. Nous regardons ensuite ce que cette figure indique sur les dynamiques d'évolution individuelles existantes dans chaque cluster. Excepté pour le troisième cluster (en magenta), l'information apportée par les trajectoires médianes semble la



(a) (b)

FIGURE 4.8 – Représentation visuelle juxtaposée des 6 clusters d'adhésion selon les trajectoires moyennes avec intervalles de confiance asymptotiques ponctuels de niveau 95% (a) et selon les trajectoires médianes avec bandes inter-quartiles (b).

même que celle fournie par les trajectoires moyennes. Cependant, la prise en compte de la bande inter-quartiles, et les contrastes entre les trajectoires moyenne, médiane et des premiers et troisièmes quartiles permettent de faire ressortir des éléments plus précis sur les séries des clusters.

Les trajectoires moyennes, médianes et quartiles des clusters 2, 4 et 6 (en bleu, vert et orange) suivent des tendances stables. Ceci va dans le sens qu'il ne semble pas y avoir de tendance croissante ou décroissante dans ces clusters, et donc que la plupart de leurs trajectoires suivent des tendances stables. Grâce aux bandes inter-quartiles, on constate que relativement aux deux autres clusters, les valeurs à chaque temps dans le clusters 6 semblent plus varier d'une trajectoire à l'autre. On remarque aussi une asymétrie autour des médianes : les valeurs inférieures en sont plus éloignées que ne le sont les valeurs supérieures, et semblent également plus variables au vu des oscillations d'un jour à l'autre de la trajectoire des premiers quartiles. Pour les clusters 2 et 4, la tendance stable qui caractérise les séries temporelles, et la faible variabilité inter-individuelle à chaque temps impliquent que les variabilités intra-individuelles sont également petites. Il n'est pas possible de conclure concernant les variabilités intra-individuelles au sein du cluster 6.

Pour le cluster 3 (en magenta), les motifs de l'évolution de la médiane et de la moyenne sont différents. La trajectoire moyenne est une courbe à tendance décroissante, ce qui ne représente aucune des trajectoires de la Figure 4.2. La trajectoire médiane est plus fidèle à la rupture caractéristique de ces séries. Les bords des bandes de confiance présentent aussi des motifs de rupture vers zéro. Cette similarité des formes entre la trajectoire médiane et les trajectoires des premiers et troisièmes quartiles indiquerait que le phénomène de rupture est commun à la plupart des séries temporelles du cluster, et qu'il se produirait à différents instants selon les séries. Cependant il n'est pas possible de conclure sur une tendance stable des séries chronologiques avant l'abandon car la trajectoire des premiers quartiles semble présenter

une rupture antérieure à celle qui conduit à la descente vers zéro. Aussi, si la trajectoire médiane et la bande inter-quartiles se terminent en restant stables à zéro, cette visualisation ne permet pas de montrer que les séries temporelles du cluster finissent également par des séquences de zéros.
Concernant les deux clusters restants, il n'est pas possible d'identifier un motif d'évolution typique pour chaque cluster. Pour le cluster 5 (en rouge), la trajectoire médiane suit un tendance qui décroît progressivement vers zéro, en l'atteignant parfois, et la trajectoire moyenne suit une tendance décroissante, mais plus lente. Si ceci peut s'expliquer par une augmentation du nombre de valeurs nulles et que seules décroissent les valeurs strictement positives, rien ne nous certifie ni la croissance du taux de zéro, ni qu'aucune autre cause ne soit impliquée. Cela peut aussi masquer une diversité dans les motifs de décroissance. En réalité, au vu de la Figure 4.3, ce cluster ne comporte pas que des séries ayant un comportement décroissant. Certaines semblent présenter une tendance de décroissance progressive vers zéro (séries numéro 152 ou 493), d'autres ressemblent à des trajectoires d'arrêts brutaux (séries numéro 449 ou 644), mais on observe aussi des séries sans tendance évidente (séries 182 ou 492). Cela peut contribuer à la différence dans l'évolution des moyennes et des médianes.
Pour le cluster 1 (en cyan), la trajectoire médiane ainsi que la bande inter-quartiles décrivent une croissance globale des séries chronologiques du cluster, avec un bon niveau d'adhésion à la fin des trois mois et une diminution en fonction du temps de la variabilité inter-trajectoires quotidienne. Cependant, la trajectoire médiane ne partage pas de motif commun avec les bords de la bande inter-quartiles, et leurs croissances suivent des dynamiques différentes. Tout ceci ne permet de distinguer ni une tendance, ni une rupture qui concerneraient la plupart des séries du cluster.

Pour la problématique clinique, les trajectoires médianes avec les bandes inter-quartiles associées semblent plus appropriées que les trajectoires moyennes. Elles exhibent les inutilisations des appareils, et le motif de rupture qui caractérise le cluster 3 est bien restitué. Il est toutefois fructueux de disposer des deux représentations graphiques pour conclure sur l'unité d'une dynamique d'évolution au sein d'un cluster. De plus la médiane et les quartiles à chaque temps ne correspondent qu'à trois valeurs réelles, alors que la moyenne tient compte de l'ensemble des observations d'un cluster. Pour le cluster 3 des abandons de traitement, le fait que la trajectoire moyenne ne termine pas à zéro reflète que c'est également le cas de certaines trajectoires du cluster. Si cela ne concerne qu'une minorité des séries chronologiques, la trajectoire moyenne a le mérite de manifester une information exploitable pour l'étape de validation de la classification.

4.4.3 Diagrammes en boîtes

Nous proposons quatre représentations graphiques utilisant des diagrammes en boîte, aussi désignées par l'anglicisme "boxplot". Elles diffèrent par les regroupements de valeurs dont sont affichées les distributions : selon les journées (ou de manière équivalente selon les positions des valeurs dans leurs trajectoires respectives), selon des intervalles de journées réglables, selon le jour de la semaine, ou selon le mois de réalisation. Ces diagrammes neutralisent les caractéristiques individuelles des séries temporelles pour permettre la comparaison des distributions selon les critères de regroupement des valeurs. Tous ces graphiques ne sont proposés qu'en

version juxtaposée.

Diagrammes en boîtes conservant l'ordre temporel des valeurs

Les diagrammes en boîtes par journée restituent l'information visible sur les trajectoires médianes avec bandes inter-quartiles. Ils donnent plus de détails car on peut également distinguer l'évolution quotidienne des valeurs extrêmes, ces dernières pouvant être issues de trajectoires distinctes. Le fait de ne pas présenter ces informations sous formes de trajectoires diminue le risque d'interprétation à tort sur les dynamiques d'évolution individuelles.

Représenter les distributions temps par temps distingue la variabilité inter-individuelle à chaque temps de la variabilité des valeurs d'un jour sur l'autre. Ce niveau de détail n'est pas toujours nécessaire et peut avoir pour conséquence de surcharger les graphiques. Dans notre cas par exemple, où le clustering regroupe les séries temporelles en procédant à des dilatations temporelles, la variabilité des temps d'utilisation d'un jour sur l'autre à l'échelle des clusters n'est pas très informative. Ceci justifie l'utilisation des diagrammes en boîtes par intervalle de journées.

Diagrammes en boîtes "calendaires"

Les deux derniers diagrammes en boîtes tiennent compte des jours calendaires. Les valeurs sont regroupées selon le jour de la semaine où elles sont réalisées ou selon leurs mois de réalisations. Le caractère séquentiel des données est perdu pour mettre en évidence d'éventuelles périodicités marquées à l'échelle des clusters. La figure 4.9 représente graphiquement les 6 clusters d'adhésion avec les diagrammes en boîtes selon les jours de la semaine. On remarque par exemple que les quartiles des temps d'utilisation quotidiens de la PPC varient peu pour les clusters 2 (en bleu) et 3 (en magenta), contrairement au cluster 5 (en rouge) où les médianes et troisièmes quartiles sont moins élevés les soirs de weekend (vendredi et samedi).

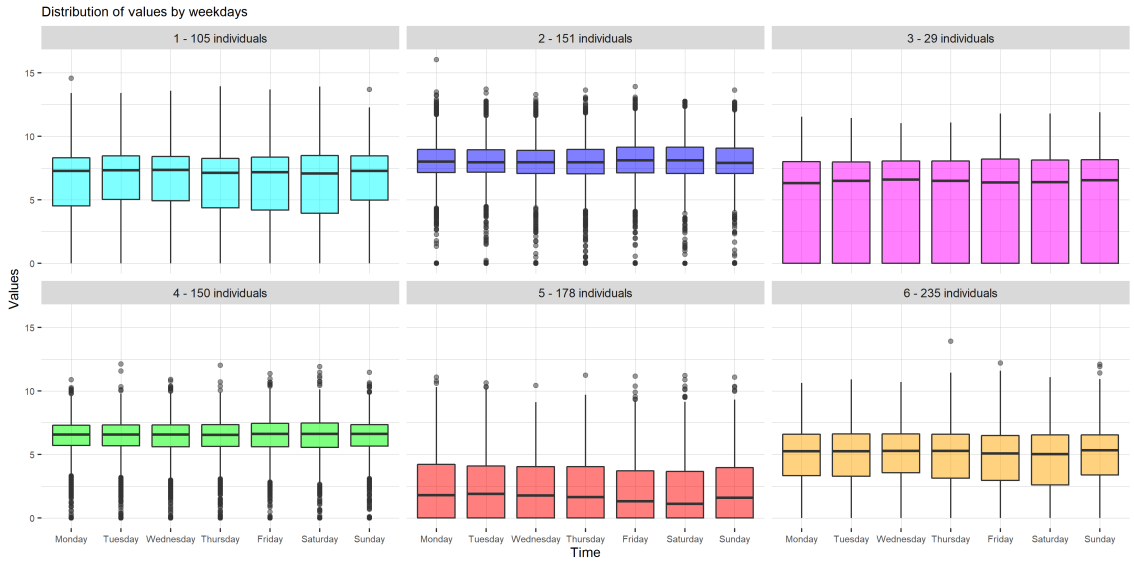


FIGURE 4.9 – Représentation visuelle des 6 clusters d'adhésion avec les diagrammes en boîtes selon les jours de la semaine. Figure exportée depuis l'application.

4.4.4 Heatmaps

Ces représentations graphiques sont désignées par un anglicisme traduisible par "carte de chaleur". Une heatmap représente, dans un plan, des données quantitatives par des couleurs dont l'intensité est liée à l'intensité des valeurs de la grandeur à représenter. Deux types de heatmaps sont proposés. Les couleurs spécifiques des clusters ne sont pas conservées. Une palette de couleurs commune est utilisée pour une meilleure comparaison des variations de teinte d'un cluster à un autre. Pour ces graphiques, seule la juxtaposition des clusters est possible.

Heatmap exhaustive

Le premier type de heatmap représente graphiquement un cluster avec en lignes l'ensemble des séries chronologiques et en colonnes les temps de réalisation des valeurs. Chaque case correspond à une valeur d'une série temporelle et l'intensité de la couleur de la case est proportionnelle à la valeur maximale observée sur l'ensemble des séries représentées. Si il ne montre pas directement les courbes des séries chronologiques d'un cluster, ce graphique a l'avantage de toutes les représenter en conservant leur caractère séquentiel. Il permet également de comparer visuellement les tailles des clusters. Le niveau d'un cluster est lisible par l'intensité globale des couleurs et la variabilité globale des valeurs est visible grâce à la variabilité des intensités des couleurs. Il est difficile de distinguer précisément la variabilité intra-trajectoire de la variabilité inter-trajectoires, mais plus les couleurs sont homogènes sur les lignes de la heatmap, moins la part de variabilité intra-individuelle est importante. Les discontinuités sont marquées par des contrastes élevés de couleurs entre des cases voisines en lignes. Enfin, l'évolution globale des valeurs au cours du temps est visible lorsqu'elle est suffisamment marquée face aux variabilités et aux discontinuités.

La Figure 4.10 représente graphiquement les clusters 2, 3, 4 et 6 avec des heatmaps exhaustives en nuances de couleurs du blanc au violet. Ces heatmaps n'apportent pas beaucoup d'informations à celles déduites après analyse de la Figure 4.8. Pour les trois clusters à tendances stables, elles confirment que les clusters 2 et 4 se différencient principalement par le niveau d'adhésion des patients alors que le cluster 6 diffère aussi par une variabilité globale plus élevée. La heatmap révèle aussi plus de discontinuités pour ce dernier cluster. La heatmap du cluster 3 certifie le bon niveau d'adhésion avant les abandons (sauf pour deux séries).

Les clusters 1 et 5 qui étaient difficilement interprétables avec les trajectoires moyennes et médianes sont représentés en Figure 4.11 pour une meilleure lisibilité. Concernant le cluster 1, on voit une diminution de la variabilité globale ainsi qu'une augmentation des valeurs au cours du temps, pour atteindre un haut niveau d'adhésion à la fin des trois mois. La heatmap pour ce cluster montre plus précisément la diversité des comportements d'adhésion en début du traitement, avec des séries qui commencent à un bas niveau (zéro pour certaines) et d'autres qui sont passées par une phase creuse après un bon début. Il n'est par contre pas possible d'identifier le caractère progressif ou brutal des changements d'adhésion. Pour le cluster 5, la heatmap indique la faible adhésion des sujets, et montre une forte variabilité, ainsi que de nombreuses discontinuités. Contrairement à la Figure 4.8, il n'apparaît pas de diminution de l'adhésion dans le temps.

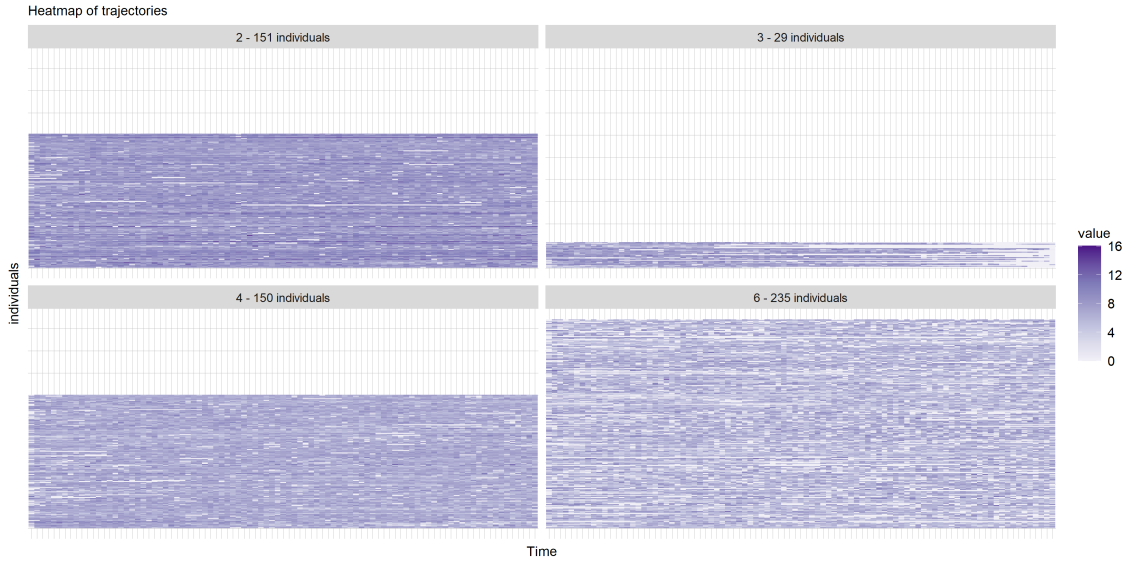


FIGURE 4.10 – Représentation visuelle des clusters d'adhésion 2, 3, 4 et 6 avec des heatmaps exhaustives. Figure exportée depuis l'application.

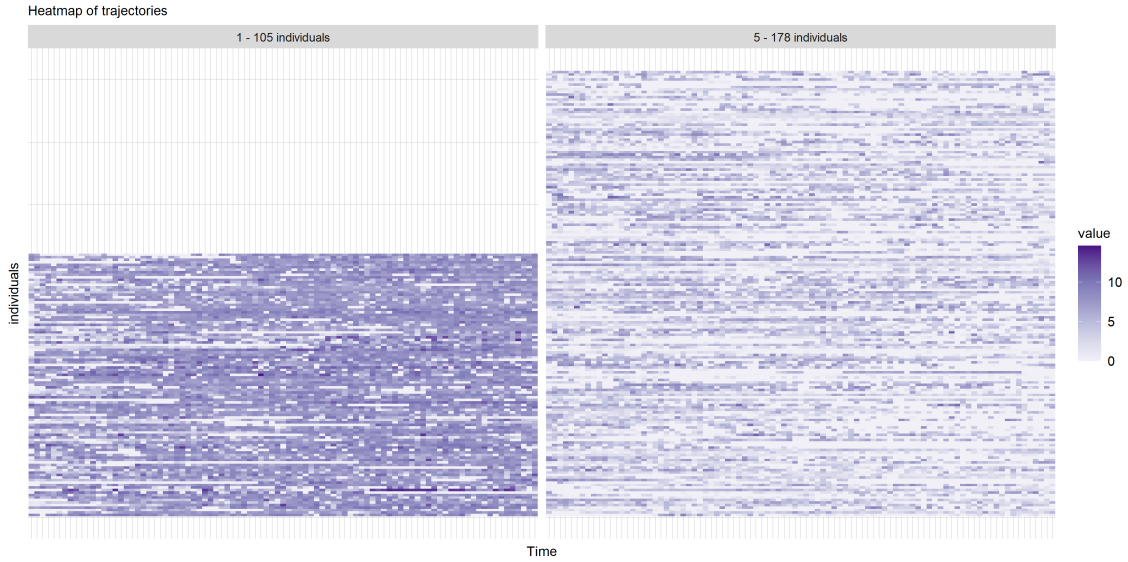


FIGURE 4.11 – Représentation visuelle des clusters d'adhésion 1 et 5 avec des heatmaps exhaustives. Figure exportée depuis l'application.

Heatmap de distributions quotidiennes selon des classes

Le second type de heatmap représente la distribution des valeurs selon des classes disjointes et personnalisables, conditionnellement aux journées. Un cluster est premièrement décrit dans un tableau contenant une ligne par classe et une colonne par journée. Le croisement de la ligne correspondant à une classe avec la colonne du $j^{\text{ème}}$ jour des trajectoires contient la fréquence de valeurs qui sont dans cette classe parmi celles réalisées la journée j, représentée par gradient de couleurs. Ce graphique montre l'évolution des proportions quotidiennes de valeurs dans des intervalles cibles. Ces proportions sont calculées temps par temps au niveau du cluster et ne permettent donc pas de restituer des comportements individuels. Sous réserve de définir des classes pertinentes, outre l'illustration des niveaux des clusters et de la variabilité inter-individuelle considérée temps à temps, cette représentation peut souligner l'apparition de valeurs extrêmes ou atypiques.

La Figure 4.12 représente graphiquement les 6 clusters d'adhésion avec des heatmaps d'appartenance à des classes quotidiennes, en nuances de couleurs du blanc au violet. Les classes choisies permettent de distinguer les nuits d'inutilisation, les nuits de très faible utilisation, (en dessous de deux heures), les utilisations insuffisantes (entre 2h et 4h), les utilisations suffisantes (entre 4h et 7h) ainsi que les très bonnes utilisations (au dessus de 7h). L'apport de ces heatmaps sur ce que nous avons déjà pu observer avec les graphiques précédents concerne principalement les clusters 5 et 6. Par rapport aux clusters 2 et 4, le cluster 6 se démarque aussi par un nombre plus élevé de nuits d'inutilisation des appareils. Le cluster 5 se caractérise par une augmentation du taux d'inutilisation des appareils au cours du temps, ce que la trajectoire médiane (Figure 4.8.b) ne peut certifier. Comme pour la heatmap exhaustive (Figure 4.11), cette figure ne permet pas de conclure d'une diminution globale des valeurs au cours du temps dans le cluster 5.

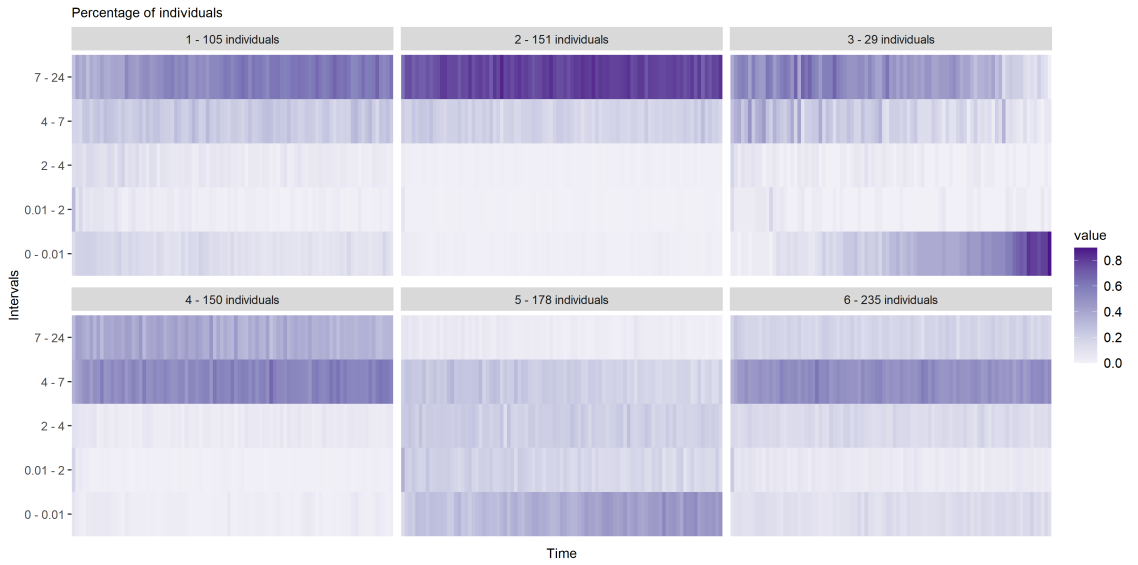


FIGURE 4.12 – Représentation visuelle des clusters d'adhésion 1 et 5 avec des heatmaps de distributions quotidiennes selon les classes d'heures d'utilisation $[0; 0,01[$, $[0,01; 2[$, $[2; 4[$, $[4; 7[$ et $[7; 24]$. Figure exportée depuis l'application.

4.5 Représentation d'indicateurs statistiques individuels

Nous introduisons deux représentations graphiques dans cette section. Elles reposent sur le calcul d'indicateurs statistiques, qualifiés d'"individuels" car ils résument chacune des séries temporelles. La perspicacité d'un indicateur donné est dépendante du contexte d'étude des séries chronologiques. Dans le cas de l'adhésion à la PPC, nous jugeons pertinent le taux d'inutilisation du dispositif alors que d'autres pourraient préférer le taux de nuits avec utilisation en dessous de 30 minutes. Nous laissons la possibilité à l'utilisateur de représenter les indicateurs de son choix, sous condition d'importer un fichier ".Rdata" contenant le code des fonctions permettant le calcul de ses indicateurs (voir l'Annexe A pour plus de détails). Ce que ces graphiques peuvent révéler sur les clusters en termes de caractéristiques individuelles typiques et sur leurs temporalités dépend des indicateurs statistiques fournis. En conséquence, contrairement à la section précédente, nous ne saurions indiquer ce que chaque graphique illustre d'une manière générale.

4.5.1 Diagramme en radar

Ce type de représentation graphique admet de nombreuses appellations : diagramme de Kiviat, diagramme en toile d'araignée ou encore diagramme polaire. Un diagramme en radar représente des données quantitatives sur différents axes, admettant le point central du graphique pour origine. Chaque axe est en général associé à une variable. Au minimum $r = 3$ variables doivent être représentées, et l'angle formé par deux axes consécutifs est constant et égal à $\frac{360°}{r}$. À partir d'un résumé de l'ensemble des indicateurs calculés pour chaque cluster, ce graphique permet la comparaison de plusieurs indicateurs individuels, pour un ou plusieurs cluster(s). Les radars de plusieurs clusters peuvent être juxtaposés ou superposés. Les résumés proposés sont la moyenne des indicateurs individuels, leur écart-type, leur médiane, leur minimum ou leur maximum. Une mise à l'échelle est d'abord nécessaire pour ramener les valeurs des différents indicateurs individuels dans un même intervalle et faciliter la lisibilité sur un même graphique. Le procédé de standardisation min-max décrit ci-après permet de ramener les valeurs dans l'intervalle $[0, 1]$. On considère un indicateur $\{y_{1,j}, y_{2,j}, ..., y_{n,j}\}$ calculé sur l'ensemble des trajectoires classifiées. La valeur standardisée de $y_{i,j}$ est $y'_{i,j} = \frac{y_{i,j} - \min(y_{.j})}{\max(y_{.j}) - \min(y_{.j})}$.

La Figure 4.13 représente graphiquement les 6 clusters d'adhésion avec un diagramme en radar en disposition juxtaposée, selon les moyennes par cluster des trois indicateurs suivants : adhésion moyenne, écart-type et taux de zéro. Cela approuve des caractéristiques des trajectoires du cluster 2 (en bleu) : une haute moyenne du temps d'utilisation du dispositif, très peu de zéros, et une faible variabilité intra-trajectoire ; par opposition aux trajectoires du cluster 5 (en rouge) : basse moyenne, beaucoup de zéros, et une variabilité intra-trajectoire élevée par rapport à la moyenne.

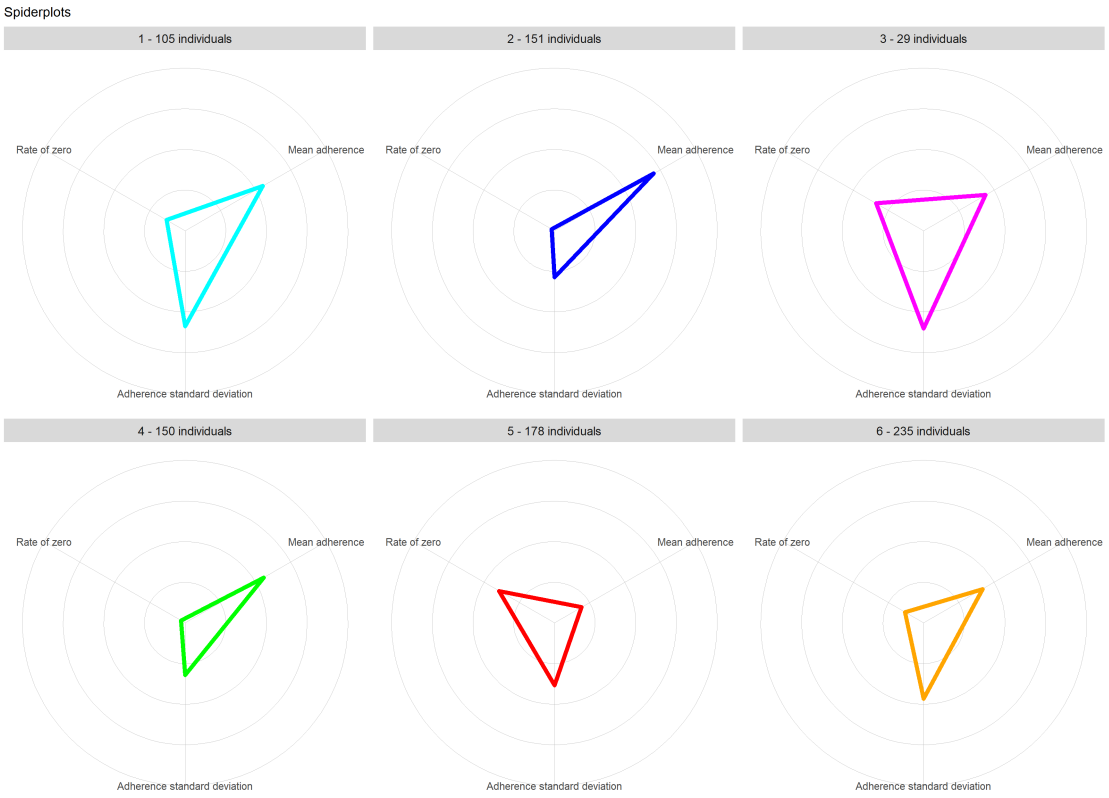


FIGURE 4.13 – Représentation visuelle des 6 clusters d'adhésion avec des diagrammes en radar juxtaposés, selon les moyennes par cluster de 3 indicateurs statistiques : la moyenne d'une trajectoire, son écart-type corrigé et le taux de nuits d'inutilisation de l'appareil. Figure exportée depuis l'application.

4.5.2 Histogramme

Il s'agit de réaliser un histogramme de la distribution d'un indicateur individuel pour chaque cluster. Les histogrammes des différents clusters peuvent être superposés ou juxtaposés. La Figure 4.14 représente les histogrammes juxtaposés des distributions des moyennes des trajectoires pour les 6 clusters d'adhésion. On constate par exemple la grande étendue de la distribution des moyennes du cluster 3 des trajectoires d'arrêt de traitement (en magenta). Ceci peut s'expliquer par le fait que la moyenne d'une trajectoire est d'autant plus élevée que la séquence de zéros débute tard dans les trois premiers mois.

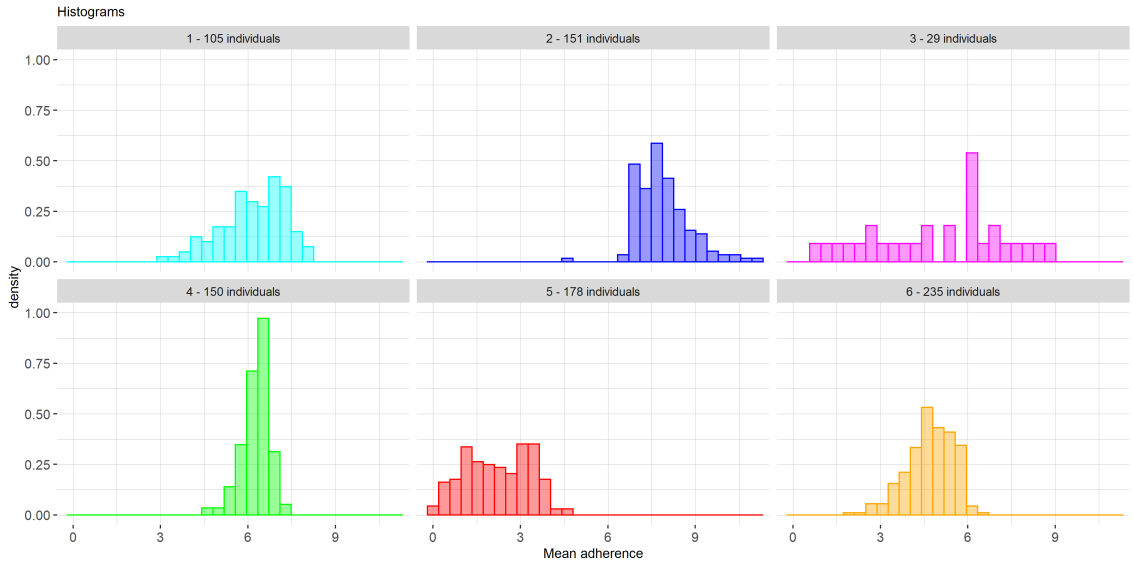


FIGURE 4.14 – Représentation graphique des histogrammes juxtaposés des distributions des moyennes des trajectoires pour les 6 clusters d'adhésion. Figure exportée depuis l'application.

4.6 Synthèse des différentes représentations

Les représentations graphiques juxtaposées se font systématiquement avec une échelle commune pour faciliter la comparaison des figures. Ce n'est pas toujours réalisé dans les communications scientifiques, ni prévu dans les représentations graphiques proposées dans certains packages. Aussi, la plupart des visualisations implémentées dans l'application utilisent un code couleur personnalisable pour l'identification des clusters. De même les types de lignes propres aux différents clusters peuvent être modifiés. Les graphiques étant exportables, l'aspect esthétique des figures produites se veut soigné et de nombreuses personnalisations sont prévues. Elles portent notamment sur les clusters à représenter (d'un à tous), sur le choix de leur affichage superposé (en un seul graphique) ou juxtaposé (un graphique par cluster) le cas échéant, ainsi que l'épaisseur des lignes à représenter. Les axes des graphiques peuvent aussi être personnalisés : titres, bornes et graduations. Les titres des graphiques peuvent être changés ainsi que les tailles des polices. La Table 4.1 propose un récapitulatif des représentations graphiques implémentées avec un comparatif sur

les possibilités de représentations visuelle en termes de juxtaposition/superposition des clusters, la nécessite pour l'utilisateur de l'application d'avoir des compétences en programmation R, ainsi que la préservation ou non des couleurs associés aux clusters.

Nom du graphique	Préservation des couleurs	Affichage		R requis	Figure(s) / Sous-section
		sup.	jux.		
Trajectoire individuelle		/	/		— / 4.4.1
Trajectoires aléatoires			✓		4.2, 4.3 / 4.4.1
Médoïde	✓	✓	✓		4.4 / 4.4.1
Diagramme spaghetti			✓		4.5 / 4.4.1
Trajectoire moyenne	✓	✓	✓		4.6, 4.7, 4.8.a / 4.4.2
Trajectoire médiane	✓	✓	✓		4.8.b / 4.4.2
Diagramme en boîtes par journée	✓		✓		— / 4.4.3
Diagramme en boîtes par intervalle de journées	✓		✓		— / 4.4.3
Diagramme en boîtes selon les jours de la semaine	✓		✓		4.9 / 4.4.3
Diagramme en boîtes selon les mois	✓		✓		— / 4.4.3
Heatmap des trajectoires			✓		4.10, 4.11 / 4.4.4
Heatmap des classes de valeurs			✓		4.12 / 4.4.4
Diagramme en radar	✓	✓	✓	✓	4.13 / 4.5
Histogramme	✓	✓	✓	✓	4.14 / 4.5

TABLE 4.1 – Tableau de synthèse des représentations graphiques proposées. Abréviations utilisées : *sup.* : superposé ; *jux.* : juxtaposé.

Chapitre 5

Conclusion et perspectives

Les travaux présentés dans cette thèse sont précurseurs pour le laboratoire HP2 dans la description des comportements d'observance à la PPC à partir des données du télésuivi. Les contributions apportées concernent des thématiques plus larges que celles exposées dans la littérature. Ces dernières sont focalisées sur l'analyse statistique des séquences d'observance alors que nos développements couvrent le traitement des données transmises par un PSAD et la visualisation des clusters de trajectoires. Nous ne proposons pas d'avancée de la connaissance dans le domaine de la classification non supervisée de séries chronologiques, mais avons fourni un effort pour identifier une méthodologie statistique adaptée à la question clinique, et permettre son utilisation par les cliniciens.

Dans le Chapitre 2 nous nous sommes intéressés à la qualité des données et avons proposé des recommandations pour favoriser la production de séquences d'observance fidèles aux comportements des sujets télésuivis. Une discussion à propos de la pertinence et de la représentativité des données permet de mettre en garde contre de mauvaises utilisations de ces données au regard de questions cliniques motivant l'identification des comportements typiques d'observance.
Dans le cas où une question clinique pourrait, et souhaiterait considérer conjointement les trajectoires d'IAH résiduels associées aux trajectoires d'observance, des méthodes existent pour étendre l'utilisation de la dissimilarité DTW à des séries chronologiques multivariées. Shokoohi-Yekta et al. [74] comparent deux approches dans le cadre de la classification supervisée.
Aussi il est possible, selon les fabricants d'appareils de PPC, d'obtenir les données d'observance avec une granularité plus fine. Un même nombre d'heures d'utilisation quotidienne de la PPC peut recouvrir plusieurs comportements d'observance quotidiens distincts, liés au fractionnement du sommeil ou à l'utilisation de la machine lors d'éventuelles siestes. Grewe et al. [75] ont montré que les sujets qui ont une utilisation quotidienne supérieure à 4 heures plus de 70% des jours fractionnement moins leurs utilisations quotidiennes de la PPC. Soose et al. [76] ont étudié les comportements d'observance à la stimulation des voies respiratoires supérieures qui est une thérapie du SAOS alternative à la PPC, également soumise à des problèmes d'observance. Ils ont réalisé une classification à l'aide d'un MMG ajusté sur un ensemble d'indicateurs statistiques incluant le nombre d'interruptions ou les heures de début et fin de traitement. Il y a un intérêt récent à prendre en compte les motifs quotidiens d'utilisation dans le traitement du SAOS, ce qui requiert de mobiliser une méthodologie de clustering différente de celle présentée dans le Chapitre 3.

Si l'on se satisfait de l'étude des données d'observance quotidiennement agrégées, la publication incluse dans le Chapitre 3 montre que la dissimilarité DTW semble adaptée pour comparer les comportements d'observance lorsque les trajectoires sont regroupées par la CAH de Ward et que le nombre de clusters est choisi à l'aide de l'indice de Dunn. Cependant en comparant des partitions ayant entre 2 et 15 clusters, la méthode de clustering a produit respectivement 2 et 7 clusters sur deux échantillons de données supplémentaires supposés indépendants de respectivement 222 et 193 sujets. Imposer l'égalité des nombres de clusters entre les échantillons ne permet pas non plus de construire des clusters similaires. Les deux jeux de données proviennent d'un même PSAD, ont subi les mêmes pré-traitements et les échantillons ne diffèrent que par le fait que les patients sont suivis ou non par les médecins du laboratoire HP2. Ce manque de reproductibilité des résultats encourage à proposer des améliorations ou changements de la méthode de clustering.
Un premier prolongement est d'investiguer la validation de la classification. En effet, les scores ARI moyens associés aux partitions sélectionnées par l'indice de Dunn sur simulations (Table 2 de l'article inséré en Section 3.2) ne sont pas optimaux relativement aux scores associés aux meilleures partitions accessibles par CAH avec le lien de Ward et la dissimilarité DTW (Table 1 de l'article). La capacité de l'indice de Dunn à choisir les partitions pourrait être impactée par la présence de trajectoires atypiques, et cet indice est celui dont les scores sont les plus dispersés.
Un deuxième prolongement concerne la méthode de classification. Dans le cas euclidien, la méthode de Ward tend à identifier des clusters sphériques et de mêmes volumes par minimisation de la WSS. Cette contrainte peut être relâchée via ajustement de MMG en cherchant à construire des clusters toujours organisés autour de centres mais ayant des formes ellipsoïdales quelconques. Cependant avec une dissimilarité telle que DTW, la méthode de Ward ne minimise que la somme des inerties intra-clusters. On ne sait pas quelle est la "forme" des clusters lorsque l'on perd les propriétés de séparation et d'inégalité triangulaire. Il serait intéressant d'utiliser des méthodes de positionnement multidimensionnel [77] pour projeter les trajectoires dans un espace euclidien conformément à la géométrie induite par la dissimilarité DTW, et réaliser la classification des trajectoires dans un tel espace via par exemple des MMG.
Un troisième prolongement est de travailler directement avec la décomposition de l'inertie pour éviter les calculs des centroïdes et les approximations correspondantes. Cette quantité pourrait être utilisée pour calculer des indices de validation de classification internes tel que l'indice de Calinski-Harabasz [58], ou pour développer un algorithme alternatif aux k-moyennes.
Les prolongements listés précédemment reposent sur l'hypothèse qu'il convient de minimiser l'inertie pour la dissimilarité DTW. Cette dernière possède de nombreuses variantes et extensions qu'il peut être intéressant d'explorer. Aussi, il semble que l'apparition des zéros d'observance soit informative sur les comportements d'observance. Une première manière de mieux prendre en compte les valeurs nulles dans la constitution des clusters est de les remplacer par des valeurs négatives. En créant une discontinuité entre les petites durées d'utilisation les inutilisations, cela inciterai DTW à aligner les séries en fonction des jours de non utilisations de la PPC. Une seconde manière est de s'orienter vers des approches de classification modélisant l'apparition des zéros au sein des séries chronologiques. Dans tous les cas cela doit encourager à être rigoureux lors de l'imputation des valeurs manquantes par des

zéros.

La restitution des clusters obtenus sur les données réelles à l'aide d'indicateurs statistiques individuels ou à l'aide des représentations graphiques des médoïdes ou des trajectoires médianes telle qu'opérée dans l'article du Chapitre 3 ne permet pas une interprétabilité suffisante des résultats de la classification. Dans le Chapitre 4 nous diversifions les visualisations des clusters. Nous présentons différents graphiques que nous avons implémenté dans une application web ne nécessitant pas de connaissance en programmation statistique pour ses utilisateurs. L'outil est développé pour être utilisé dans le contexte clinique de la thèse. Néanmoins son champ d'utilisation se veut général. L'application permet d'explorer visuellement des groupements quelconques de séries chronologiques de valeurs quotidiennes de mêmes longueurs. La finalité de l'application est de permettre l'interprétation des clusters par les cliniciens. Cet objectif est finalement difficile à atteindre car certains graphiques peuvent mener à des déductions erronées et requièrent des compétences en analyse des données. Les médoïdes s'interprètent en fonction de la mesure de dissimilarité utilisée. Les trajectoires moyenne, médiane et de quartiles d'un cluster peuvent conduire à des conclusions hâtives. La comparaison de l'évolution de ces trajectoires semble informative notamment pour montrer les tendances stables caractéristiques des clusters et mériterait d'être investiguée.
Des modifications de l'application informatique sont nécessaires avant son déploiement en ligne. L'accessibilité doit être améliorée pour l'utilisateur, avec une meilleure ergonomie, davantage de retours d'informations, et l'amélioration des rubriques d'aide. La structure du code doit également être revue pour faciliter les mises à jour ultérieures de l'application comme l'amélioration et l'enrichissement des représentations graphiques. Pour faciliter les publications scientifiques, l'export des figures selon d'autres formats est à prévoir et la création d'un système de sauvegarde de session utilisateur permettrai de reprendre les figures à l'identique pour toute modification souhaitée. Enfin une extension des fonctionnalités liées aux indicateurs statistiques individuels serait d'intégrer la réalisation et l'export des tableaux descriptifs et comparatifs des groupes de trajectoires, selon des covariables et les indicateurs statistiques calculés par l'application.

Ces travaux permettront aux cliniciens de mieux appréhender les trajectoires d'observance des patients télésuivis à partir des données collectées.

Bibliographie

[1] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J.-L. Pépin, "Obstructive sleep apnoea syndrome," *Nature Reviews Disease Primers*, p. 15015, jun 2015.

[2] A. J. M. Hirsch Allen, N. Bansback, and N. T. Ayas, "The effect of osa on work disability and work-related injuries." *Chest*, vol. 147, pp. 1422–1428, May 2015.

[3] S. Ryan, C. Arnaud, S. F. Fitzpatrick, J. Gaucher, R. Tamisier, and J.-L. Pépin, "Adipose tissue as a key player in obstructive sleep apnoea." *European respiratory review : an official journal of the European Respiratory Society*, vol. 28, Jun. 2019.

[4] S. Javaheri, F. Barbe, F. Campos-Rodriguez, J. A. Dempsey, R. Khayat, S. Javaheri, A. Malhotra, M. A. Martinez-Garcia, R. Mehra, A. I. Pack, V. Y. Polotsky, S. Redline, and V. K. Somers, "Sleep apnea," *Journal of the American College of Cardiology*, vol. 69, no. 7, pp. 841–858, feb 2017.

[5] A. Zinchuk and H. K. Yaggi, "Phenotypic subtypes of OSA," *Chest*, vol. 157, no. 2, pp. 403–420, feb 2020.

[6] J. Fleetham, N. Ayas, D. Bradley, K. Ferguson, M. Fitzpatrick, C. George, P. Hanly, F. Hill, J. Kimoff, M. Kryger *et al.*, "Directives de la société canadienne de thoracologie : Diagnostic et traitement des troubles respiratoires du sommeil de l'adulte," *Canadian Respiratory Journal*, vol. 14, no. 1, pp. 31–36, 2007.

[7] W. T. McNicholas, "Diagnosis of obstructive sleep apnea in adults," *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 154–160, feb 2008.

[8] P. Escourrou, N. Meslier, B. Raffestin, R. Clavel, J. Gomes, E. Hazouard, J. Paquereau, I. Simon, and E. O. Frija, "Quelle approche clinique et quelle procédure diagnostique pour le sahos ?" *Revue des maladies respiratoires*, vol. 27, pp. S115–S123, 2010.

[9] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi, "Rules for scoring respiratory events in sleep : Update of the 2007 AASM manual for the scoring of sleep and associated events," *Journal of Clinical Sleep Medicine*, vol. 08, no. 05, pp. 597–619, oct 2012.

[10] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea : a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, aug 2019.

[11] E. M. Wickwire, S. E. Tom, A. Vadlamani, M. Diaz-Abad, L. M. Cooper, A. M. Johnson, S. M. Scharf, and J. S. Albrecht, "Older adult US medicare beneficiaries with untreated obstructive sleep apnea are heavier users of health care than matched control patients," *Journal of Clinical Sleep Medicine*, vol. 16, no. 1, pp. 81–89, jan 2020.

[12] J.-L. Pépin, S. Bailly, P. Rinder, D. Adler, D. Szeftel, A. Malhotra, P. Cistulli, A. Benjafield, F. Lavergne, A. Josseran, R. Tamisier, and P. H. and, "CPAP therapy termination rates by OSA phenotype : A french nationwide database analysis," *Journal of Clinical Medicine*, vol. 10, no. 5, p. 936, mar 2021.

[13] C. R. Davies and J. J. Harrington, "Impact of obstructive sleep apnea on neurocognitive function and impact of continuous positive air pressure," *Sleep Medicine Clinics*, vol. 11, no. 3, pp. 287–298, sep 2016.

[14] S. B. Montesi, B. A. Edwards, A. Malhotra, and J. P. Bakker, "The effect of continuous positive airway pressure treatment on blood pressure : A systematic review and meta-analysis of randomized controlled trials," *Journal of Clinical Sleep Medicine*, vol. 08, no. 05, pp. 587–596, oct 2012.

[15] I. H. Iftikhar, C. W. Valentine, L. R. Bittencourt, D. L. Cohen, A. C. Fedson, T. Gíslason, T. Penzel, C. L. Phillips, L. Yu-sheng, A. I. Pack, and U. J. Magalang, "Effects of continuous positive airway pressure on blood pressure in patients with resistant hypertension and obstructive sleep apnea," *Journal of Hypertension*, vol. 32, no. 12, pp. 2341–2350, dec 2014.

[16] I. H. Iftikhar, C. M. Hoyos, C. L. Phillips, and U. J. Magalang, "Meta-analyses of the association of sleep apnea with insulin resistance, and the effects of CPAP on HOMA-IR, adiponectin, and visceral adipose fat," *Journal of Clinical Sleep Medicine*, vol. 11, no. 04, pp. 475–485, apr 2015.

[17] J. M. Marin, S. J. Carrizo, E. Vicente, and A. G. Agusti, "Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure : an observational study," *The Lancet*, vol. 365, no. 9464, pp. 1046–1053, mar 2005.

[18] R. D. McEvoy, N. A. Antic, E. Heeley, Y. Luo, Q. Ou, X. Zhang, O. Mediano, R. Chen, L. F. Drager, Z. Liu, G. Chen, B. Du, N. McArdle, S. Mukherjee, M. Tripathi, L. Billot, Q. Li, G. Lorenzi-Filho, F. Barbe, S. Redline, J. Wang, H. Arima, B. Neal, D. P. White, R. R. Grunstein, N. Zhong, and C. S. Anderson, "CPAP for prevention of cardiovascular events in obstructive sleep apnea," *New England Journal of Medicine*, vol. 375, no. 10, pp. 919–931, sep 2016.

[19] J.-L. Pépin, S. Bailly, P. Rinder, D. Adler, A. V. Benjafield, F. Lavergne, A. Josseran, P. Sinel-Boucher, R. Tamisier, P. A. Cistulli, A. Malhotra, and P. Hornus, "Relationship between CPAP termination and all-cause mortality," *Chest*, feb 2022.

[20] W. T. McNicholas, C. L. Bassetti, L. Ferini-Strambi, J. L. Pépin, D. Pevernagie, J. Verbraecken, W. Randerath, W. T. McNicholas, C. L. Bassetti, L. Ferini-Strambi, J. L. Pépin, D. Pevernagie, J. Verbraecken, M. R. Bonsignore, R. Farre, L. Grote, J. Hedner, M. Kohler, M. A. Martinez-Garcia, S. Mihaicuta, J. Montserrat, F. Pizza, O. Polo, R. L. Riha, S. Ryan, and W. Randerath, "Challenges in obstructive sleep apnoea," *The Lancet Respiratory Medicine*, vol. 6, no. 3, pp. 170–172, mar 2018.

[21] G. Labarca, J. Dreyse, L. Drake, J. Jorquera, and F. Barbe, "Efficacy of continuous positive airway pressure (CPAP) in the prevention of cardiovascular events in patients with obstructive sleep apnea : Systematic review and meta-analysis," *Sleep Medicine Reviews*, vol. 52, p. 101312, aug 2020.

[22] S. Schiza, P. Lévy, M. A. Martinez-Garcia, J.-L. Pepin, A. Simonds, and W. Randerath, "The search for realistic evidence on the outcomes of obstructive sleep apnoea," *European Respiratory Journal*, vol. 58, no. 4, p. 2101963, oct 2021.

[23] B. W. Rotenberg, D. Murariu, and K. P. Pang, "Trends in CPAP adherence over twenty years of data collection : a flattened curve," *Journal of Otolaryngology - Head & Neck Surgery*, vol. 45, no. 1, aug 2016.

[24] C. H. K. Lee, L. C. Leow, P. R. Song, H. Li, and T. H. Ong, "Acceptance and adherence to continuous positive airway pressure therapy in patients with obstructive sleep apnea (OSA) in a southeast asian privately funded healthcare system," *Sleep Science*, vol. 10, no. 2, pp. 57–63, 2017.

[25] M. Tsuyumu, T. Tsurumoto, J. Iimura, T. Nakajima, and H. Kojima, "Ten-year adherence to continuous positive airway pressure treatment in patients with moderate-to-severe obstructive sleep apnea," *Sleep and Breathing*, vol. 24, no. 4, pp. 1565–1571, feb 2020.

[26] M. Kohler, A.-C. Stoewhas, L. Ayers, O. Senn, K. E. Bloch, E. W. Russi, and J. R. Stradling, "Effects of continuous positive airway pressure therapy withdrawal in patients with obstructive sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 184, no. 10, pp. 1192–1199, nov 2011.

[27] B. Salepci, B. Caglayan, N. Kiral, E. T. Parmaksiz, S. S. Comert, G. Sarac, A. Fidan, and G. A. Gungor, "CPAP adherence of patients with obstructive sleep apnea," *Respiratory Care*, vol. 58, no. 9, pp. 1467–1473, feb 2013.

[28] F. Barbé, J. Durán-Cantolla, M. S. de-la Torre, M. Martínez-Alonso, C. Carmona, A. Barceló, E. Chiner, J. F. Masa, M. Gonzalez, J. M. Marín, F. Garcia-Rio, J. D. de Atauri, J. Terán, M. Mayos, M. de la Peña, C. Monasterio, F. del Campo, J. M. Montserrat, for the Spanish Sleep, and B. Network, "Effect of continuous positive airway pressure on the incidence of hypertension and cardiovascular events in nonsleepy patients with obstructive sleep apnea," *JAMA*, vol. 307, no. 20, may 2012.

[29] C. Borriboon, J. Chaiard, C. Tachaudomdach, and S. Turale, "Continuous positive airway pressure adherence in people with obstructive sleep apnoea," *Journal of Clinical Nursing*, dec 2021.

[30] M. S. Aloia, J. T. Arnedt, M. Stanchina, and R. P. Millman, "How early in treatment is PAP adherence established ? revisiting night-to-night variability," *Behavioral Sleep Medicine*, vol. 5, no. 3, pp. 229–240, aug 2007.

[31] G. M. Nixon, R. Mihai, N. Verginis, and M. J. Davey, "Patterns of continuous positive airway pressure adherence during the first 3 months of treatment in children," *The Journal of Pediatrics*, vol. 159, no. 5, pp. 802–807, nov 2011.

[32] D. Ghosh, V. Allgar, and M. W. Elliott, "Identifying poor compliance with CPAP in obstructive sleep apnoea : A simple prediction equation using data after a two week trial," *Respiratory Medicine*, vol. 107, no. 6, pp. 936–942, jun 2013.

[33] T. Gentina, "Does CPAP use in the first 15 days predict its use after 4 months ? a prospective french cohort study," *Journal of Sleep Disorders & Therapy*, vol. 05, no. 01, 2015.

[34] M. S. Aloia, M. S. Goodwin, W. F. Velicer, J. T. Arnedt, M. Zimmerman, J. Skrekas, S. Harris, and R. P. Millman, "Time series analysis of treatment adherence patterns in individuals with obstructive sleep apnea," *Annals of Behavioral Medicine*, vol. 36, no. 1, pp. 44–53, aug 2008.

[35] T. E. Weaver, N. B. Kribbs, A. I. Pack, L. R. Kline, D. K. Chugh, G. Maislin, P. L. Smith, A. R. Schwartz, N. M. Schubert, K. A. Gillen, and D. F. Dinges, "Night-to-night variability in CPAP use over the first three months of treatment," *Sleep*, vol. 20, no. 4, pp. 278–283, apr 1997.

[36] R. Sampaio, M. G. Pereira, and J. C. Winck, "A new characterization of adherence patterns to auto-adjusting positive airway pressure in severe obstructive sleep apnea syndrome : clinical and psychological determinants," *Sleep and Breathing*, vol. 17, no. 4, pp. 1145–1158, feb 2013.

[37] W. K. Wohlgemuth, D. A. Chirinos, S. Domingo, and D. M. Wallace, "Attempters, adherers, and non-adherers : Latent profile analysis of CPAP use with correlates," *Sleep Medicine*, vol. 16, no. 3, pp. 336–342, mar 2015.

[38] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 355–378, mar 2019.

[39] M. R. Weiss, M. L. Allen, J. S. Landeo-Gutierrez, J. P. Lew, J. K. Aziz, S. S. Mintz, C. M. Lawlor, B. J. Becerra, D. A. Preciado, and G. Nino, "Defining the patterns of PAP adherence in pediatric obstructive sleep apnea : a clustering analysis using real-world data," *Journal of Clinical Sleep Medicine*, vol. 17, no. 5, pp. 1005–1013, may 2021.

[40] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, Mar. 1963.

[41] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method : Which algorithms implement ward's criterion ?" *Journal of Classification*, vol. 31, no. 3, pp. 274–295, oct 2014.

[42] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, vol. 49, no. 3, p. 803, sep 1993.

[43] V. Batagelj, "Generalized ward and related clustering problems," *Classification and Related Methods of Data Analysis*, 01 1988.

[44] M. K. R. Baddam, M. Araujo, and J. Srivastava, "Defining and monitoring patient clusters based on therapy adherence in sleep apnea management," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, jun 2021.

[45] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[46] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern Recognition*, vol. 28, no. 5, pp. 781–793, may 1995.

[47] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering – a decade review," *Information Systems*, vol. 53, pp. 16–38, oct 2015.

[48] Y. Kang and V. V. Prabhu, "Data mining for characterizing obstructive sleep apnea treatment adherence trends," in *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2013, p. 1600.

[49] N. G. P Den Teuling, E. R. van den Heuvel, M. S. Aloia, and S. C. Pauws, "A latent-class heteroskedastic hurdle trajectory model : patterns of adherence in obstructive sleep apnea patients on cpap therapy." *BMC medical research methodology*, vol. 21, p. 269, Dec. 2021.

[50] Y. Li, Y. Wang, G. Alan, Y. Chai, J. Luo, X. Niu, B. Hai, and J. Qin, "Pre- and in-therapy predictive score models of adult OSAS patients with poor adherence pattern on nCPAP therapy," *Patient Preference and Adherence*, p. 715, may 2015.

[51] J. Bros, C. Poulet, J. E. Methni, C. Deschaux, M. Gandit, P. J. Pauwels, and M. Charavel, "Determination of risks of lower adherence to CPAP treatment before their first use by patients," *Journal of Health Psychology*, vol. 27, no. 1, pp. 223–235, aug 2020.

[52] S. F. Babbin, W. F. Velicer, M. S. Aloia, and C. A. Kushida, "Identifying longitudinal patterns for individuals and subgroups : An example with adherence to treatment for obstructive sleep apnea," *Multivariate Behavioral Research*, vol. 50, no. 1, pp. 91–108, jan 2015.

[53] F. Portier, E. O. Frija, J.-M. Chavaillon, L. Lerousseau, O. R. Degat, D. Léger, and J.-C. Meurice, "Traitement du SAHOS par ventilation en pression positive continue (PPC)," *Revue des Maladies Respiratoires*, vol. 27, pp. S137–S145, oct 2010.

[54] M. Müller, *Information Retrieval for Music and Motion*. Springer, 2007.

[55] C. Genolini, R. Ecochard, M. Benghezal, T. Driss, S. Andrieu, and F. Subtil, "kmlShape : An efficient method to cluster longitudinal data (time-series) according to their shapes," *PLOS ONE*, vol. 11, no. 6, p. e0150738, jun 2016.

[56] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, Jan. 1973.

[57] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J. M. Pérez, and J. I. Martín, "Towards a standard methodology to evaluate internal cluster validity indices," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 505–515, feb 2011.

[58] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[59] A. Sardá-Espinosa, "Time-series clustering in r using the dtwclust package," *The R Journal*, vol. 11, no. 1, p. 22, 2019.

[60] K. Pearson, "LIII. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, nov 1901.

[61] P. D'Urso, L. D. Giovanni, and V. Vitale, "Spatial robust fuzzy clustering of COVID 19 time series based on b-splines," *Spatial Statistics*, p. 100518, may 2021.

[62] S. Levantesi, A. Nigri, and G. Piscopo, "Clustering-based simultaneous forecasting of life expectancy time series through long-short term memory neural networks," *International Journal of Approximate Reasoning*, vol. 140, pp. 282–297, jan 2022.

[63] M. Maleki, H. Bidram, and D. Wraith, "Robust clustering of COVID-19 cases across u.s. counties using mixtures of asymmetric time series models with time varying and freely indexed covariates," *Journal of Applied Statistics*, pp. 1–15, jan 2022.

[64] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time series data," *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1–6, 2020. [Online]. Available : http://jmlr.org/papers/v21/20-091.html

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn : Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available : http://jmlr.org/papers/v12/pedregosa11a.html

[66] P. Montero and J. A. Vilar, "TSclust : AnRPackage for time series clustering," *Journal of Statistical Software*, vol. 62, no. 1, 2014.

[67] U. Mori, A. Mendiburu, and A. L. Jose, "Distance measures for time series in r : The TSdist package," *The R Journal*, vol. 8, no. 2, p. 451, 2016.

[68] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, *shiny : Web Application Framework for R*, 2021, r package version 1.6.0. [Online]. Available : https://CRAN.R-project.org/package=shiny

[69] P.-N. Tan, M. Steinbach, V. Kumar, and A. Karpatne, *Introduction to Data Mining, Global Edition.* Pearson Education Limited, May 2019. [Online]. Available : https://www.ebook.de/de/product/32492337/pang_ning_tan_michael_steinbach_vipin_kumar_anuj_karpatne_introduction_to_data_mining_global_edition.html

[70] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, apr 1979.

[71] M. Kim and R. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2353–2363, nov 2005.

[72] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J. I. Martín, J. Muguerza, J. M. Pérez, and I. Perona, "SEP/COP : An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index," *Pattern Recognition*, vol. 43, no. 10, pp. 3364–3373, oct 2010.

[73] L. Kaufmann and P. Rousseeuw, "Clustering by means of medoids," *Data Analysis based on the L1-Norm and Related Methods*, pp. 405–416, 01 1987.

[74] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing DTW to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 1–31, feb 2016.

[75] F. A. Grewe, M. Bradicich, T. Gaisl, M. Roeder, S. Thiel, N. A. Sievi, and M. Kohler, "Patterns of nightly CPAP usage in OSA patients with suboptimal treatment adherence," *Sleep Medicine*, vol. 74, pp. 109–115, oct 2020.

[76] R. J. Soose, M. Araujo, K. Faber, A. Roy, K. Lee, Q. Ni, J. Srivastava, and P. J. Strollo, "Cluster analysis of upper airway stimulation adherence patterns and implications on clinical care," *Sleep*, mar 2022.

[77] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*. Springer Berlin Heidelberg, 2008, pp. 315–347.

Annexes

A Présentation des fonctionnalités de l'application

L'application comporte 3 "rubriques" avec des fonctionnalités interconnectées, chaque rubrique étant composée de différents onglets. Cette annexe décrit chacune des rubriques dans des sous-sections spécifiques.

A.1 Fonctionnalités liées au chargement du jeu de données (rubrique "data")

Cette première partie de l'application permet d'une part de charger les données des clusters de trajectoires, de choisir les groupements de trajectoires à représenter et d'effectuer des opérations de datamanagement.

A.1.1 Onglet "file upload"

La première étape nécessaire à l'utilisation de l'application est le dépôt du fichier de données dans l'onglet "file upload" de la rubrique "data Main". L'extension ".csv" est attendue et le fichier doit contenir les trajectoires individuelles univariées réparties en clusters, leurs dates de début, et éventuellement des covariables individuelles. Nous précisons ici que seule la trajectoire est une donnée longitudinale, les autres métadonnées devant être constantes au cours du temps. Un affichage du tableau de données est proposé. L'utilisateur doit ensuite renseigner les correspondances entres les colonnes du tableau et les rôles prédéfinis dans l'application : identifiant des trajectoires, valeurs, temps de réalisation des valeurs (en journées relatives depuis le début de la trajectoire), dates de commencement des trajectoires, clusters et covariables continues et/ou discrètes.

La matrice de dissimilarité préalablement calculée entre les trajectoires et stockée au format ".csv" peut également être chargée dans l'application dans un emplacement spécifique. À noter que les en-têtes de lignes et de colonnes de la matrice de dissimilarité doivent comporter les identifiants des trajectoires.

La deuxième étape concerne les réglages des trajectoires. Il est possible d'aligner les trajectoires selon un jour de la semaine, de sorte que toutes les trajectoires commencent à partir de la première valeur réalisée le jour demandé. La fenêtre temporelle peut être modifiée, et seules les trajectoires possédant toutes les valeurs entre les bornes de cette fenêtre seront incluses. Le nombre de trajectoires finalement retenues est indiqué dans un encadré.

A.1.2 Onglet "groups and parameters"

Cet onglet se décompose en trois parties. Dans la partie supérieure gauche, l'utilisateur devra spécifier le groupement des trajectoires à représenter. Ce groupement consiste en un croisement d'une classification (parmi celles sélectionnées dans l'onglet "file upload"), avec de zéro à plusieurs covariable(s) (parmi celle(s) présélectionnée(s)). La partie supérieure droite permet de changer l'ordre des différents groupes (les modalités de croisement) ainsi que le changement des paramètres d'affichage graphique associés (couleurs et types de lignes). Enfin, dans la partie inférieure de l'onglet, un diagramme en barres représente les effectifs des groupes qui résultent du croisement, avec les paramètres graphiques sélectionnés. Cette représentation permet de décider si des opérations de datamanagement sont nécessaires comme le regroupement de modalités ou la discrétisation d'une variable continue.

A.1.3 Onglet "data management"

Cet onglet propose trois panneaux pour la réalisation du datamanagement des variables de clusters, et des covariables discrètes ou continues. Les panneaux alloués aux variables de clusters et aux covariables discrètes proposent les mêmes fonctionnalités. Premièrement la sélection d'une variable sur laquelle effectuer un datamanagement. Ensuite il est possible de renommer cette variable, de grouper et/ou renommer certaines de ces modalités et enfin de restaurer les modalités de la variable à l'état initial. Un diagramme en barres affiche les effectifs de la variable en cours de modification. Concernant le panneau des covariables continues, un histogramme représente la distribution de la variable, et le regroupement de modalités laisse place à la spécification des valeurs selon lesquelles opérer la discrétisation. Les classes peuvent être renommées et un diagramme en barres affiche leurs effectifs.

A.2 Fonctionnalités de représentations graphiques des clusters (rubrique "trajectories")

Cette rubrique contient différents onglets, dont certains constitués de plusieurs panneaux. Diverses représentations graphiques des clusters de trajectoires sont proposées. Elles se répartissent en 4 catégories : des trajectoires d'évolution d'indicateurs statistiques, des diagrammes en boîte, des trajectoires individuelles (dont les médoïdes, nécessitant le chargement préalable de la matrice de dissimilarité entre les trajectoires du jeu principal) et des heatmaps. Les paramètres graphiques alloués aux différents clusters dans l'onglet "groups and parameters" sont préservés pour la plupart des représentations lorsque cela est possible. Les graphiques produits sont personnalisables et exportables.

A.2.1 Représentation de trajectoires d'évolution quotidienne d'un indicateur de tendance centrale

Il s'agit des trajectoires moyenne et médiane. Elles sont situées dans la rubrique "trajectories", sur des onglets propres respectivement intitulés "Mean trajectories" et "Median trajectories".

A.2.2 Représentation avec des diagrammes en boîte

Les diagrammes en boîte sont intégrés à l'onglet "Boxplots" de la rubrique "trajectories". Les 4 types de graphique sont présentés sur 4 panneaux distincts : la représentation par journée sur le panneau "By temporal units", la représentation par intervalles de journées sur le panneau "By periods", et les deux dernières représentations sur les onglets "By weekdays" et "By months".

A.2.3 Représentation de trajectoires individuelles

Ces représentations sont situées sur plusieurs panneaux de l'onglet "Individuals" dans la rubrique "trajectories". Le premier panneau "Individual Trajectories" permet la représentation d'une trajectoire individuelle choisie. Le second panneau est "Medoids". Les trajectoires aléatoires sont représentées sur le panneau "Random individual trajectories". Enfin les diagrammes spaghetti sont réalisables sur le panneau "Spaghetti plots".

A.2.4 Heatmaps

Les heatmaps sont intégrées à l'onglet "Heatmaps" de la rubrique "trajectories" dans deux panneaux distincts. Le panneau "Trajectories" représente les heatmaps avec les trajectoires en ligne et le panneau "Counting individuals" représente la distribution des valeurs selon des classes définies par l'utilisateur.

A.3 Fonctionnalités liées aux indicateurs statistiques (rubrique "statistical indicators")

Cette rubrique permet de représenter les trajectoires à travers des indicateurs statistiques individuels. Son utilisation nécessite quelques compétences en programmation R. Un premier onglet nommé "Load indicators" requiert le dépôt d'un fichier ".Rdata" contenant des fonctions codées en langage R pour le calcul d'un ensemble d'indicateurs statistique. Ces fonctions doivent prendre une trajectoire en entrée pour en sortie renvoyer un nombre. L'utilisateur sélectionne ensuite les indicateurs qu'il souhaite calculer. Les deux onglets suivants sont dédiés à la représentation de ces indicateurs via un diagramme en radar (onglet "Spiderplots") ou un histogramme (onglet "Histograms"). Ces graphiques sont également personnalisables, exportables, et préservent les paramètres graphiques propres à chaque cluster. Nous attirons ici l'attention sur le rendu visuel après export des diagrammes en radar. Chaque radar doit avoir une largeur égale à sa hauteur, ce qui empêche l'ajustement automatique des dimensions des graphiques. Pour éviter la présence de bandes blanches latérales (ou en-dessus/dessous de la figure), nous suggérons d'ajuster le ratio hauteur/largeur de l'export graphique.