



UNIVERSITÉ DE RENNES 1 - ENSAI

RAPPORT DE STAGE

Classification de trajectoires de traitement du diabète

Nom de l'étudiante :

Romane LE GOFF

Tuteurs de stage :

Geoffray BIZOUARD

Oriane BRETIN

Responsable pédagogique :

Catherine BENJAMIN

Du 28 février 2022 au 31 août 2022



Table des matières

Remerciements	1
Introduction	2
1 Contexte du stage	3
1.1 IQVIA	3
1.2 Problématique	3
1.2.1 Format des données	3
1.2.2 Calcul de distances	4
1.3 Le diabète	5
1.4 Cadrage de l'étude	6
1.4.1 Sélection des patients	7
1.5 Construction des séquences de traitement	7
1.5.1 Catégorisation des traitements	7
1.5.2 Définition de la durée d'une délivrance	8
1.5.3 Définition d'un traitement continu	8
1.5.4 Discontinuation	8
1.5.5 Add-on	9
1.5.6 Switch	9
1.5.7 Lissage des séquences	10
2 Recherche bibliographique	11
2.1 Mesures de dissimilarités	11
2.2 Optimal Matching	12
2.2.1 Concept	12
2.2.2 Détails de la méthode	12
2.2.3 Variantes	13
2.3 Distances basées sur le compte des attributs communs entre les séquences	14
2.3.1 Distance de Hamming simple	14
2.3.2 Distance de Hamming Dynamique (DHD)	14
2.3.3 Distance basée sur la longueur de la plus longue sous-séquences commune (LCS)	15
2.3.4 Nombre de sous-séquences correspondantes (NMS)	15
2.4 Distances entre les distributions de probabilité	15
2.5 Analyse multi-séquences	16
2.6 Modèles mixtes à classes latentes	16
2.7 Bag-of-Words	17
3 Résultats	18
3.1 Optimal Matching	18
3.1.1 Choix des coûts	18
3.1.2 Visualisation	19
3.2 Variantes de l'OM	22
3.3 Distances basées sur le compte des attributs communs entre les séquences	23
3.3.1 Séquences de même longueur - Distances de Hamming	23
3.3.2 NMS	23
3.4 Distances entre les distributions d'états	24
3.5 Analyse multi-séquences	24
3.6 Modèles de mélange	25
3.6.1 Application de la méthode et limites	25
3.6.2 Résultats d'un clustering à 4 groupes	25

3.7	Bag-of-Words	26
3.8	Recommandations dans le choix des méthodes	26
3.9	Interprétation économique des résultats	27
3.10	Conclusion partielle	28
Conclusion		29
A Annexe		31
A.1	Flowchart	31
A.2	Traitements du diabète	31
A.2.1	Les familles de médicaments	31
A.2.2	Schémas thérapeutiques considérés	33
A.3	Algorithme de la Haute Autorité de Santé	33
A.4	Classification non-supervisée	34
A.4.1	Classification hiérarchique	34
A.4.2	Classification non-hiéarchique	34
A.4.3	Classification par densité	35
A.5	Vue d'ensemble des mesures de qualité d'un clustering	35
A.6	Indicateurs de qualité du clustering dans R	36
A.7	Variantes de l'OM	37
A.7.1	OM localisé (OMloc)	37
A.7.2	OM sensible à la durée des épisodes (OMslen)	37
A.7.3	OM entre des séquences d'épisodes (OMspell)	38
A.7.4	OM sur les transitions (OMstran)	39
A.7.5	Nombre de sous-séquences correspondantes (NMS)	39
A.8	Bag-of-Words	40
A.9	Implémentation dans R	41
A.9.1	Algorithme derrière l'Optimal Matching	41
A.9.2	La librairie TraMineR	44
A.10	Coûts utilisés pour l'application des méthodes dans R	47
A.11	Qualité du clustering et temps de calcul en fonction des méthodes et du nombre de classes	48
A.11.1	Valeur de l'ASW en fonction du nombre de groupes	48
A.11.2	Temps de calcul	49
A.12	Visualisation des résultats	50
A.12.1	Optimal Matching	50
A.12.2	OM sur des séquences de 41 états	53
A.12.3	Présentation de l'algorithme LRx de typage du diabète	56
A.12.4	Variantes de l'OM	57
A.12.5	Distances basées sur le nombre de sous-séquences communes	60
A.12.6	Distances entre les distributions d'états	65
A.12.7	Bag-of-Words	67
A.12.8	Analyse multi-séquences	68
A.12.9	Description des groupes	69
A.12.10	Modèles de mélange	71

Table des figures

1.1	Exemple d'une séquence de traitement	4
1.2	Calcul de la distance entre deux patients	4
1.3	Dendrogramme issu d'une Classification Ascendante Hiérarchique et exemple de regroupement de séquences de traitement	5
1.4	Définition d'un traitement continu	8
1.5	Illustration d'une discontinuation	8
1.6	Illustration d'un add-on d'un traitement B à un traitement A	9
1.7	Illustration des deux cas de figure considérés pour un switch	9
1.8	Aperçu des séquences de traitement du diabète	10
2.1	Opérations élémentaires utilisées en OM	12
3.1	Dendrogrammes issus de la CAH	19
3.2	Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM TRATE)	20
3.3	OM (CAH) - 5 groupes	20
3.4	OM PAM - 5 groupes	21
3.5	MCSA index plot empilé (CAH) - 5 groupes	24
3.6	Arbre de décision dans le choix des méthodes	27
A.1	Flowchart de la construction de la population	31
A.2	Algorithme retraçant les schémas thérapeutiques possibles du patient diabétique, adaptés en fonction de la cible HbA1c* (source : HAS)	33
A.3	Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM INDELSLOG)	50
A.4	Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM INDELS)	50
A.5	OM (CAH) - 7 groupes	51
A.6	OM (CAH) - 10 groupes	51
A.7	OM (PAM) - 7 groupes	52
A.8	OM (PAM) - 10 groupes	52
A.9	OM sur des séquences de 41 états (CAH) - 10 groupes	53
A.10	OM sur des séquences de 41 états (CAH) - 10 groupes (nombre d'états réduit à 6)	54
A.11	OM sur 6 périodes (CAH) - 7 groupes	55
A.12	OM sur 6 périodes (CAH) - 10 groupes	55
A.13	Algorithme LRx du typage du diabète	56
A.14	Variantes de l'OM (CAH) - 5 groupes	57
A.15	Variantes de l'OM (CAH) - 7 groupes	57
A.16	Variantes de l'OM (CAH) - 10 groupes	58
A.17	Variantes de l'OM (PAM) - 5 groupes	58
A.18	Variantes de l'OM (PAM) - 7 groupes	59
A.19	Variantes de l'OM (PAM) - 10 groupes	59
A.20	Distances de Hamming (CAH) - 5 groupes	60
A.21	Distances de Hamming (CAH) - 7 groupes	60
A.22	Distances de Hamming (CAH) - 10 groupes	61
A.23	Distances de Hamming (PAM) - 5 groupes	61
A.24	Distances de Hamming (PAM) - 7 groupes	62
A.25	Distances de Hamming (PAM) - 10 groupes	62
A.26	NMS (CAH) - 5 groupes	63
A.27	NMS (CAH) - 7 groupes	63
A.28	NMS (CAH) - 10 groupes	63
A.29	NMS (PAM) - 5 groupes	64
A.30	NMS (PAM) - 7 groupes	64

A.31 NMS (PAM) - 10 groupes	64
A.32 Distributions d'états (CAH) - 5 groupes	65
A.33 Distributions d'états (CAH) - 7 groupes	65
A.34 Distributions d'états (CAH) - 10 groupes	66
A.35 Bag-of-Words (CAH) - 5 groupes	67
A.36 Bag-of-Words (CAH) - 7 groupes	67
A.37 Bag-of-Words (CAH) - 10 groupes	68
A.38 MCSA tapis (CAH) - 5 groupes	68
A.39 Représentation des variables statiques dans chaque groupe (clustering avec OM TRATE)	69
A.40 Représentation des variables statiques dans chaque groupe dans un graphique à barres cumulées (clustering avec OM TRATE)	69
A.41 Proportion de chaque traitement et combinaison de traitement par cluster	70
A.42 Tapis des 4 groupes obtenu avec le modèle mixte à classes latentes	71
A.43 Chronogramme des 4 groupes obtenu avec le modèle mixte à classes latentes	71

Liste des tableaux

2.1	Distance de Hamming : exemple	14
2.2	Distance de Levenshtein II : exemple	15
3.1	Matrice de coûts de substitution obtenue avec OM TRATE	18
3.2	Description simple des clusters	22
3.3	Coûts moyen d'une monothérapie (2015)	28
A.1	Combinaisons possibles en trithérapie [1]	33
A.2	Exemple de distance OMloc entre trois séquences en utilisant un coût fixe de substitution de 1, $\alpha = 0.1$ et $\beta = 0.8$ [2]	37
A.3	Tableau récapitulatif (TF-IDF)	40
A.4	Matrice de scores $D(i, j)$	42
A.5	Initialisation de la matrice de scores $D(i, j)$	42
A.6	Remplissage de la matrice	42
A.7	Détermination du meilleur alignement (chemin parcouru visible en gras)	43
A.8	Détails des méthodes de calcul de coûts sur R	44
A.9	Coûts d'indel (OM INDELSLOG)	47
A.10	Coûts de substitution (OM INDELSLOG)	47
A.11	Coûts d'indel (OM INDELS)	47
A.12	Coûts de substitution (OM INDELS)	47
A.13	Comparaison de la qualité du clustering par variante d'OM	48
A.14	Comparaison de la qualité du clustering pour d'autres types de mesure de dissimilarité	48
A.15	Qualité du clustering (ASW) des autres méthodes explorées	49
A.16	Temps de calcul des variantes d'OM pour déterminer les distances dans R (sur 10 itérations)	49
A.17	Temps moyen de calcul des autres mesures de dissimilarité dans R (sur 10 itérations)	49
A.18	Temps moyen de calcul des autres méthodes explorées (sur 10 itérations)	49
A.19	Modèle obtenu : $\Lambda(t) = X(t)\beta + Z(t)u_i + w_i(t)$ (classe latente de référence : 1)	71
A.20	Effectifs par classe	72
A.21	Probabilité d'appartenir à une classe sachant que le patient a été classé dans une classe	72
A.22	Proportions de patients en fonction de leur probabilité (> 0.7 , > 0.8 ou > 0.9) d'appartenir à une classe	72

Liste des abréviations

Abréviation	Définition
AD	Anti-diabétique
AGLP1	Analogues de la Glucagon-Like Peptide
BoW	Bag-of-Words
CAH	Classification Ascendante Hiérarchique
CHI2	Distance du Khi2
CNAM	Caisse Nationale d'Assurance Maladie
DDD	Defined Daily Dose
DHD	Distance de Hamming Dynamique
DT1	Diabète de Type 1
DT2	Diabète de Type 2
EUCLID	Distance euclidienne
HAM	Distance de Hamming
HAS	Haute Autorité de Santé
iDPP4	inhibiteurs de la Dipeptidylpeptidase-4
IMC	Indice de Masse Corporelle
LCS	Plus longue sous-séquence commune
LRx/LTD	Longitudinal Rx Data
NMS	Nombre de sous-séquences correspondantes
OAD	Autre anti-diabétique
OM	Optimal Matching
OMloc	Optimal Matching localisé
OMslen	Optimal Matching sensible à la durée des épisodes
OMspell	Optimal Matching entre des séquences d'épisodes
OMstran	Optimal Matching sur les transitions
MET	Metformine
PAM	Partitioning Around Medoids
SGLT2	Sodium Glucose de Type 2
SNDS	Système National des Données de Santé
SVRspell	Représentation vectorielle des sous-séquences

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à ce que ce mémoire se passe dans les meilleures conditions possibles, et qui m'ont aidée lors de la rédaction du rapport.

Tout d'abord, je remercie mes tuteurs, Geoffray Bizouard et Oriane Bretin, d'avoir pris le temps de m'accompagner au cours de ce projet, et d'avoir veillé à la bonne réalisation des travaux de recherche. Je remercie également Pascale Rondeau, responsable de l'équipe du pôle Biométrie, pour son soutien tout le long de mon stage. Je remercie aussi Clément Chastagnol, Soline Leblanc et Raphaël Sigogne pour m'avoir partagé leur expertise en Data Science et sans qui le déroulement du stage n'aurait pas été aussi facile grâce à nos réunions mensuelles.

Je remercie aussi toute l'équipe du pôle Biométrie et DBSA, et plus particulièrement Oumon Diaby et Antoine Bessou pour le transfert de leurs connaissances en épidémiologie qui m'ont aidée à la construction de la base de données et à l'interprétation des résultats.

Enfin, merci à Catherine Benjamin, mon référent académique, qui a veillé au bon déroulement du projet.

Grâce à toutes ces personnes, ce sujet de mémoire aura été une très bonne expérience et une réussite, tant sur le plan professionnel que personnel.

Introduction

L'ensemble des actes tels que les visites et prescriptions médicales, ou la délivrance de médicaments tout au long du suivi du patient constituent son parcours de soins. Le parcours de soins est personnalisé et a pour objectif de faire bénéficier chaque patient d'un suivi médical adéquat, d'une gestion minutieuse du dossier médical et d'une prévention adaptée. La dimension longitudinale apportée par le suivi du patient permet de cartographier l'évolution des traitements et soins médicaux utilisés sur une période. Dans les études en vie réelle où le nombre de parcours distinct peut être très important, la recherche de similitudes ou de regroupement de trajectoires patient est primordiale pour mieux comprendre la population étudiée. Les études de vie réelle sont des études observationnelles sur base de données améliorant les connaissances médicales, les expériences des patients et des médecins, grâce à leur contribution à la conception de meilleurs traitements. Le recours à l'analyse des parcours de soins permet l'étude de l'impact, économique et social, d'une pathologie, sur la société (e.g. sur une population spécifique, sur les remboursements de l'assurance maladie...).

Le stage a pour objectif de faire l'état de l'art des méthodes de regroupement de trajectoires d'une pathologie donnée en utilisant une base de données française IQVIA de délivrances de médicaments en pharmacie, et de fournir une comparaison rigoureuse et méthodologique des méthodes sélectionnées. D'après Santé Publique France, la prévalence du diabète traité pharmacologique était de 5,3 % en 2020 en France, un chiffre en constante augmentation ces dernières décennies. La bonne connaissance de la maladie et la volumétrie disponible des données font du diabète la pathologie idéale pour l'application des méthodes étudiées dans le cadre du stage. Le sujet a ainsi un objectif de recherche et développement autour de cette problématique de classification. Les résultats de la recherche sont à destination de l'équipe encadrante, afin de les intégrer dans des projets clients.

La première partie du stage consiste en une récolte d'informations sur la maladie du diabète et sur la construction de la base de données. La seconde partie du stage s'étend à une recherche bibliographique des méthodes de classification de séquences pouvant être appliquées. Enfin, le stage se termine par la programmation des méthodes sélectionnées et la rédaction d'un guide méthodologique pour l'utilisation de ces dernières à destination de l'équipe. Tout le long du stage, un comité méthodologie, regroupant data scientists et statisticiens de différentes branches d'IQVIA, a été organisé mensuellement pour faire le point sur les avancées et problématiques du stage.

Contexte du stage

Dans cette partie, IQVIA et la branche Real World Solutions sera présentée dans un premier temps. La problématique de stage sera ensuite introduite avant de présenter la maladie du diabète. Enfin, les méthodes de sélection de la population pour la construction de la base de données seront expliquées.

1.1 IQVIA

IQVIA est un leader mondial dans l'utilisation de la science des données humaines qui intègre l'étude des sciences humaines par des innovations en Data Science et technologie pour faire avancer la santé. L'entreprise est divisée en plusieurs unités régionales, appelées RBU (pour Regional Business Units), dans le monde : EMA (Europe Middle-East Africa), Etats-Unis & Canada, Amérique Latine, Asie Pacifique et Japon. IQVIA France fait partie de l'EMA. De plus, trois unités opérationnelles composent IQVIA : Research, Development & Solutions (RDS) pour les essais cliniques, Technology & Commercial Solutions (TCS) pour la conception d'applications IT, et Real World Solutions (RWS) pour tout ce qui concerne les données en vie réelle. Les clients d'IQVIA sont majoritairement des sociétés pharmaceutiques, des sociétés de biotechnologie, des entreprises de santé grand public, des payeurs et fournisseurs ainsi que des agences gouvernementales.

Le stage se déroule au sein de la branche RWS, dans laquelle quatre différents pôles travaillent en collaboration : les pôles Health Technology Assessment and Advisory (HTAA), Primary Data Collection Studies (OPS), Database Studies and Analytics (DBSA) et Biométrie. L'expertise au sein de RWS s'occupe de la mise en place d'une stratégie d'accès au marché, d'une modélisation médico-économique, d'études sur bases de données secondaires, et d'études prospectives avec collecte de données primaires. Le stage est encadré par l'équipe Biométrie, qui est composée de statisticiens. Elle réalise majoritairement des études statistiques sur les bases de données secondaires d'IQVIA ainsi que sur les bases de données de patients publics et partenaires, en particulier sur la base SNDS (le Système National des Données de Santé). Biométrie travaille étroitement avec les épidémiologistes du pôle DBSA.

1.2 Problématique

Tout le long de son suivi, un patient pourra recevoir des prescriptions de traitements et/ou de soins, qui lui seront adaptés : le parcours de soins d'un individu est unique. Pouvoir identifier les trajectoires possibles d'une pathologie est essentiel pour caractériser l'évolution des thérapies dans le temps. Une trajectoire pourra se définir par une séquence de traitements, autrement dit par une suite d'événements (les traitements) ayant lieu dans un intervalle de temps fixé (la durée de suivi du patient). L'objectif du stage est de fournir et de comparer des méthodes permettant le regroupement de trajectoires de traitement similaires, afin d'identifier des "parcours-types" de patients.

1.2.1 Format des données

Le parcours d'un patient peut être représenté sous la forme d'une "séquence", c'est-à-dire d'une suite d'événements appelés "états". L'ensemble des états distincts retrouvés dans les séquences est appelé l'alphabet des séquences. Dans notre cadre, le terme "séquence" fait référence à un enchaînement de traitements délivrés pour un individu. Pour créer une séquence, des informations sur les dates et la durée couverte des délivrances des traitements du patient sont requises (cf. Figure 1.1). La durée d'une délivrance est définie par le conditionnement de la boîte de médicament délivrée et la posologie du médicament.

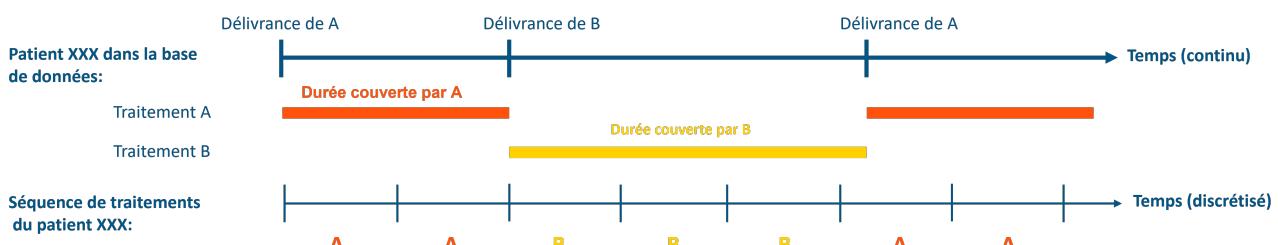


FIGURE 1.1 – Exemple d'une séquence de traitement

Le temps est ensuite discrétisé (cf. Figure 1.1) : une unité de temps, qui indique à quel moment les états (traitements) ont lieu, doit être choisie pour toutes les séquences (e.g. journalier, hebdomadaire, mensuel, annuel). Le choix de l'unité dépend de la durée du suivi des patients, de la fréquence à laquelle les patients changent de traitement ou passent de nouveaux examens, et de l'objectif de l'analyse.

1.2.2 Calcul de distances

Le regroupement de séquences par des méthodologies de clustering implique de tout d'abord calculer des "distances" entre patients.

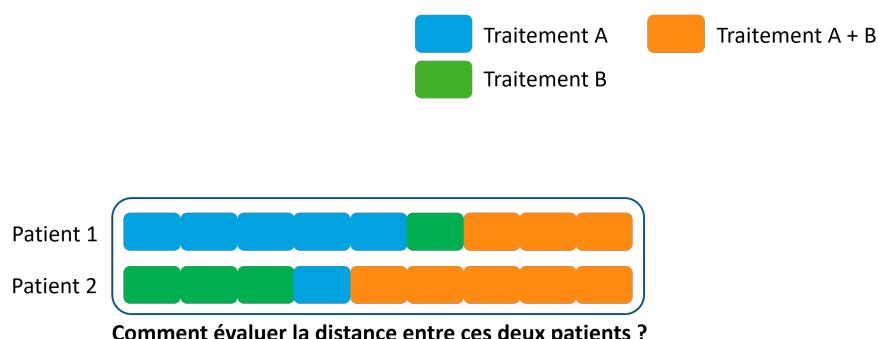


FIGURE 1.2 – Calcul de la distance entre deux patients

Pour ce faire, différentes méthodes de calcul de distance entre séquences (en réalité, de dissimilarité) existent et seront présentées dans ce document :

- Les distances éditées : l'Optimal Matching (OM) et ses variantes
 - Les distances basées sur le compte d'attributs communs
 - Les distances basées sur les distributions d'états

Ces distances vont être plus ou moins sensibles à certains aspects de la dimension temporelle des trajectoires, qui peuvent être :

- Le timing (calendrier) : le moment où l'état a lieu ; e.g. dans la Figure 1.2, faut-il accorder une certaine importance au fait que l'état "A+B" arrive chez le patient 1 plus tard que chez le patient 2 ?
 - La durée : le temps consécutif passé dans le même état ; e.g. dans la Figure 1.2, faut-il accorder une certaine importance au fait que l'état "A+B" dure plus longtemps chez le patient 2 ?
 - Le sequencing (séquençage) : l'ordre dans lequel les états sont expérimentés ; faut-il accorder une certaine importance au fait que l'état "A" ait lieu avant l'état "B" pour le patient 1, contrairement au patient 2 ?

Le choix de la méthode de calcul de distance appliquée se fera en fonction de la (les) dimension(s) d'intérêt. Une fois les distances entre les individus calculées, un algorithme de classification non-supervisée peut être appliqué pour regrouper les individus ayant des parcours similaires (illustration en Figure 1.3), autrement dit, les individus proches en termes de distances. Des détails sur la classification non-supervisée [3] ainsi que sur les critères de validation du clustering sont disponibles en annexe A.4, A.5 et A.6.

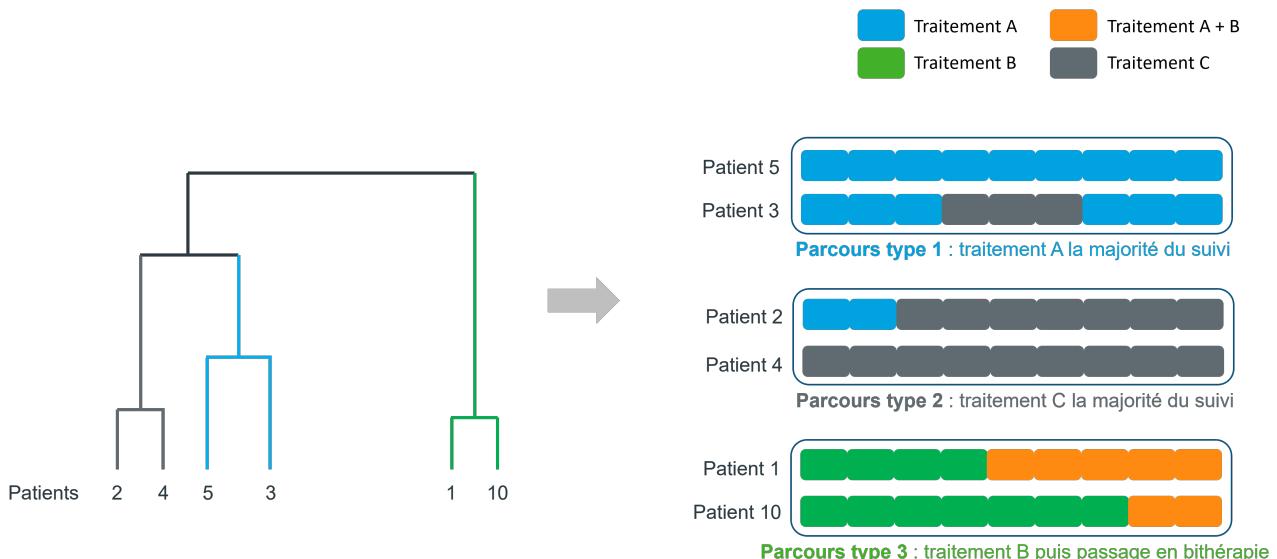


FIGURE 1.3 – Dendrogramme issu d'une Classification Ascendante Hiérarchique et exemple de regroupement de séquences de traitement

1.3 Le diabète

Le diabète est une maladie chronique causée par un dysfonctionnement du métabolisme de l'insuline, l'hormone responsable de la régulation du glucose dans le sang. Les deux principaux types de diabète en termes de fréquence sont le diabète du type 1 (DT1) et le diabète du type 2 (DT2).

Le DT1 (5-10% des cas) est causé par la destruction des cellules β du pancréas due à une combinaison de facteurs génétiques et environnementaux. Une incidence plus importante de ce type de diabète est observée chez les enfants et adolescents. Les patients sont traités par insuline - il en existe trois types : les analogues d'action rapide, les analogues d'action intermédiaire et les analogues d'action prolongée.

Le DT2 (90-95% des cas) est dû à une résistance des cellules à l'action de l'insuline. Les principaux facteurs de risque du diabète de type 2 sont l'âge, les antécédents familiaux et les facteurs liés au mode de vie. Il est précédé d'une période de prédiabète caractérisée par une glycémie élevée. Le pic de prévalence de ce type de diabète se situe entre 65 et 85 ans mais l'incidence chez les enfants commence à augmenter. Les patients peuvent être traités de différentes manières : en monothérapie sous forme d'anti-diabétique oral (la molécule de référence étant la metformine), en bi ou tri-thérapie associant différentes familles de molécules, et également par la combinaison d'insulines d'action intermédiaire et de longue durée.

En 2017, 451 millions de personnes étaient diagnostiquées diabétiques dans le monde [4], un nombre en forte augmentation ces dernières années. En France, la prévalence du diabète traité pharmacologiquement était de 5,3 % en 2020 (contre 4,6 % en 2012), d'après Santé Publique France. Elle est plus élevée chez les hommes et augmente avec l'âge.

Différents schémas thérapeutiques sont considérés de façon à caractériser le type de traitement pris par le patient diabétique (détails en Annexe A.2). Ce dernier peut être traité sous monothérapie (hors insuline), i.e. sous un seul traitement, qui est en général la metformine si le patient n'a pas de contre-indication d'utilisation du médicament. Une bithérapie voire une trithérapie ou plus peut également être considérée pour le patient. Dans ce cas, la metformine est combinée avec un ou plusieurs médicaments. En outre, le patient peut être mis sous insuline (DT1) ou sous combinaison d'insuline et d'autres anti-diabétiques oraux (DT2). Plus de détails concernant les différents schémas thérapeutiques considérés par la Haute Autorité de la Santé en annexe (cf. Figure A.2 en Annexe).

Le choix du diabète dans le cadre du stage se justifie par une bonne connaissance de la pathologie ainsi que par le nombre grandissant d'individus atteints de la maladie [5], que cela soit au niveau des DT1 ou des DT2, donnant une bonne volumétrie de séquences de traitement pour l'application des méthodes et facilitant l'interprétation et la validation des résultats.

1.4 Cadrage de l'étude

Cette étude est une analyse longitudinale rétrospective sur base de données de vie réelle. Les données proviennent d'une base de données IQVIA appelée LTD (Longitudinal Treatment Dynamics). Un panel de 9 600 pharmacies représentatives de la France métropolitaine constitue la base, regroupant ainsi 40 millions de patients. Ces données longitudinales, disponibles depuis 2012, correspondent à des délivrances de médicaments, sans diagnostic du patient. Il est possible de suivre le patient dans différentes pharmacies, à condition qu'elles fassent partie du panel.

L'étude consiste à suivre des patients diabétiques naïfs de traitement. Le point de départ du suivi correspond à la première date de délivrance d'un traitement anti-diabétique observée sur une période d'inclusion définie au préalable. Cette première date constitue notre "date index". A partir de celle-ci, le suivi continue jusqu'à la dernière date de délivrance observée, augmentée de la durée du traitement estimé, sur notre période d'observation.

La première ligne de traitement est définie par l'absence de délivrance de tous médicaments anti-diabétiques dans les 12 mois précédent cette première délivrance. On considère que le patient démarre son premier traitement à cette date ; le patient sera donc défini comme naïf de traitement. Ce design permet d'étudier une cohorte de patients reflétant la distribution actuelle des traitements anti-diabétiques en France afin de donner la photo la plus récente de la prise en charge médicamenteuse du diabète.

Les patients inclus dans l'étude sont préalablement identifiés dans la base de données à partir de critères d'inclusion. La phase d'inclusion commence au 1er janvier 2015 et s'achève au 31 décembre 2015, et la période de suivi s'étale jusqu'au 31 décembre 2021.

Périodes d'étude et définition de la date index : Les données sont extraites du 1er janvier 2014 au 31 décembre 2021. La période d'inclusion s'étend du 1er janvier 2015 au 31 décembre 2015.

Date index : La date index est définie comme la première date de délivrance d'un traitement anti-diabétique indiquée sur la période d'inclusion du 1er janvier 2015 au 31 décembre 2015. Cette date index marque le début du suivi.

Période de suivi : La période de ce suivi débute à la date index et s'achève :

- A la date d'identification de la dernière ligne de traitement anti-diabétique.
La dernière ligne de traitement anti-diabétique est définie par l'absence de délivrance de tout traitement anti-diabétique dans les 90 jours suivant la fin de la durée de la dernière délivrance.
- Ou à la date de fin d'étude, i.e. le 31 décembre 2021.

Périodes historiques : Une période pré-index d'un an, commençant 1 an avant la date index et se terminant à la date index, a été définie pour appliquer le critère de régularité du patient dans la pharmacie. Les données de tous les médicaments sur la période historique sont extraites afin d'identifier les patients du panel traités avec d'autres traitements que les médicaments anti-diabétiques. Cette période historique permet de vérifier la présence du patient dans le panel avant la première ligne de traitement (via des délivrances de médicaments quels qu'ils soient) et l'absence de délivrances de traitements anti-diabétiques sur cet historique.

Définition des cas : Les cas sont définis par les patients identifiés par les critères d'inclusion comme étant initiateurs de traitement anti-diabétique à l'inclusion (située entre le 1er janvier 2015 et le 31 décembre 2015).

1.4.1 Sélection des patients

Critères d'inclusion

La population source de l'étude est constituée de l'ensemble des patients répondant aux critères d'inclusion suivants :

- Présence d'au moins une délivrance d'un traitement anti-diabétique sur la période d'inclusion (du 1er janvier 2015 au 31 décembre 2015)
- Présence d'au moins trois délivrances de médicaments anti-diabétiques indiquées sur la période du 1er janvier 2015 au 31 décembre 2016
- Critère de régularité pré-index : au moins trois mois de présence enregistrés dans le panel, tout traitement confondu, avant la date index (période d'un an)
- Critère de régularité post-index : au moins trois mois de présence dans le panel, séparés au maximum de quatre mois d'intervalle, tout traitement confondu, 1 an après la date index
- Critère d'éligibilité pour les pharmacies : seuls les patients visitant des pharmacies ayant fourni des données tous les mois de l'étude sont conservés

Une régularité de trois mois de présence annuelle, séparés au maximum de quatre mois d'intervalle, est assurée tout le long du suivi. Si le patient ne remplit pas ce critère durant une année du suivi, la fin de son suivi est avancée à la dernière ligne de traitement enregistrée avant le dernier mois de présence.

Critère d'exclusion

La population source de l'étude est constituée de l'ensemble des patients répondant aux critères d'exclusion suivants :

- Tout patient ayant une délivrance de traitement anti-diabétique un an avant la date index est exclu. Ce critère permet de s'assurer que le patient est bien naïf de traitement.
- Tout patient ayant un suivi d'une durée de moins de 16 mois est exclu.

Le flowchart de la construction de la population est visible en Annexe A.1.

1.5 Construction des séquences de traitement

La base de données LTD contient les dates de délivrance, le nombre de boîtes et les ATC5 des médicaments délivrés. Afin de déterminer les séquences de traitement, il est nécessaire de savoir quel(s) médicament(s) un patient prend chaque jour. Cela implique de faire des hypothèses sur le temps couvert par chaque délivrance, à partir de différents indicateurs tels que la dose quotidienne définie (DDD ou Defined Daily Dose), le nombre de boîtes et le conditionnement du traitement délivré. La DDD correspond à la dose moyenne de traitement supposée par jour pour un médicament utilisé pour son indication principale chez l'adulte.

La construction des séquences de traitement, réalisée sur le logiciel statistique SAS, comprend plusieurs étapes.

1.5.1 Catégorisation des traitements

Les traitements seront catégorisés comme ci-dessous :

- Metformine
- Sulfamides + glinides
- iDDP4 : inhibiteurs de la dipeptidylpeptidase-4
- SLGT2 : sodium glucose de type 2
- AGLP1 : analogues de la glucagon-like peptide
- Autre type de monothérapie (hors insuline) : inhibiteurs des alpha-glucosidases
- Insuline à action rapide
- Autre type d'insuline

La définition d'une mono, bi, tri ou insulinothérapie se fait par la combinaison de ces traitements.

1.5.2 Définition de la durée d'une délivrance

L'hypothèse de durée d'une délivrance repose sur la DDD, le nombre de boîtes délivré par traitement ainsi que le conditionnement de la boîte (nombre de comprimés ou de stylos). Les DDD sont récupérées sur le site de WHO Drug et utilisées pour définir la posologie de chaque médicament. Le conditionnement de la boîte sert à déterminer la durée du traitement, en fonction de la posologie définie au préalable. Deux cas particuliers sont considérés :

- Dans le cas d'une combinaison de traitements, les DDD attribuées se basent sur le principe de compter la DDD du principe actif majoritaire (qui est généralement la metformine), quel que soit le nombre d'ingrédients actifs inclus dans la combinaison.
- Dans le cas de l'insuline, un stylo est compté comme un mois de traitement. Le nombre de stylos par boîte est utilisé pour calculer la durée de la délivrance.

Ainsi, les durées d'exposition au traitement reposent en grande partie sur les DDD. Or, la DDD est une moyenne de la posologie d'un traitement, ce qui ne correspond pas toujours à la vraie posologie du patient. Les durées de délivrance deviennent des approximations. C'est pourquoi il est possible d'observer des "trous" de traitement, qui ne sont en réalité pas de vrais arrêts de traitement. Pour définir un arrêt, un intervalle de temps de 90 jours est fixé et va nous permettre de définir un arrêt ou une continuité de traitement.

1.5.3 Définition d'un traitement continu

Un traitement est considéré comme continu si un écart de moins de 90 jours est observé entre la fin d'une délivrance (date de délivrance + durée théorique de couverture du médicament délivré) et le début de la délivrance suivante, comme illustré dans la figure 1.4.

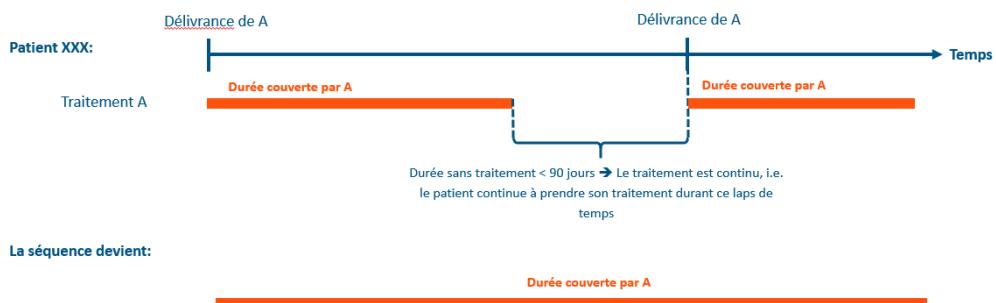


FIGURE 1.4 – Définition d'un traitement continu

1.5.4 Discontinuation

Une discontinuation du traitement est envisagée si un écart de plus de 90 jours est présent entre la fin d'une délivrance (date de délivrance + durée théorique de couverture du médicament délivré) et la délivrance suivante. La figure 1.5 illustre cette idée.

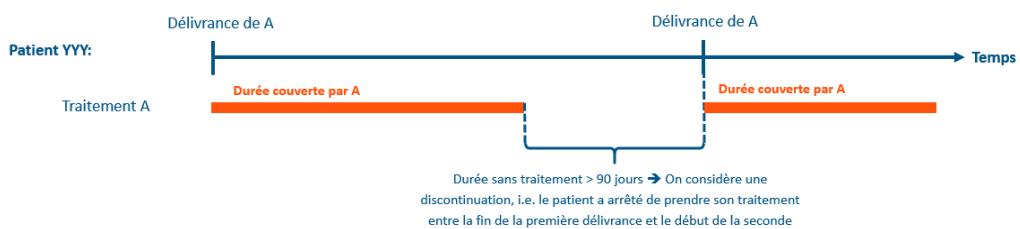


FIGURE 1.5 – Illustration d'une discontinuation

1.5.5 Add-on

Il est aussi possible d'observer deux traitements qui se chevauchent. Pour définir si le patient prend réellement ces deux traitements (add-on) en même temps ou s'ils sont pris successivement, l'intervalle de temps de 90 jours est une nouvelle fois utilisée.

Un add-on est considéré si un recouvrement de plus de 90 jours est observé (cf. Figure 1.6) entre le début d'une nouvelle délivrance et la fin de la délivrance actuelle. Autrement dit, un nouveau médicament est ajouté au traitement actuel du patient à partir de la nouvelle date de délivrance enregistrée.

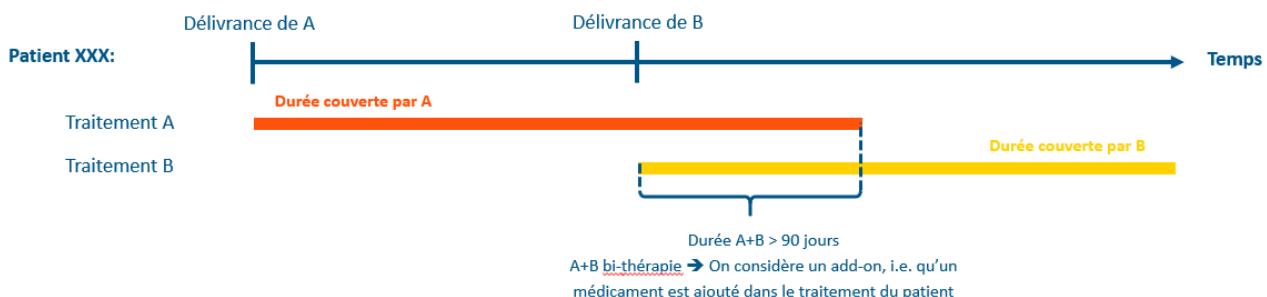


FIGURE 1.6 – Illustration d'un add-on d'un traitement B à un traitement A

1.5.6 Switch

Un switch est défini comme un changement d'une classe de traitement précédemment définie entre deux séquences de traitement consécutives. Deux cas de figure sont considérés (cf. Figure 1.7) :

1. Un switch est envisagé si un recouvrement de moins de 90 jours est observé entre le début d'une nouvelle délivrance et la fin de la délivrance actuelle.
 2. Un switch est envisagé si un écart de moins de 90 jours est observé entre le début d'une nouvelle délivrance et la fin de la délivrance actuelle.



La séquence devient:

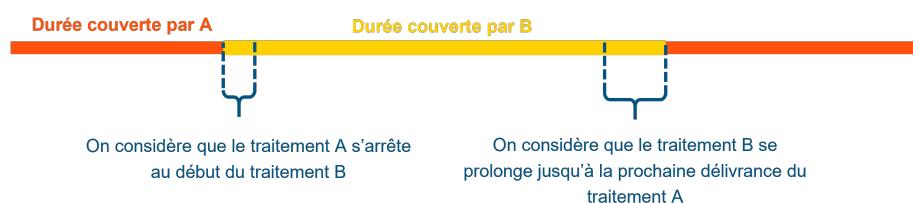


FIGURE 1.7 – Illustration des deux cas de figure considérés pour un switch

1.5.7 Lissage des séquences

Une dernière étape consiste en un lissage des séquences de traitement par période de 4 mois, pour un total de 18 périodes. Pour ce faire, la contribution en termes de durée couverte par chaque médicament, pour chaque période, est calculée. Le traitement affichant la contribution la plus élevée est sélectionné pour chaque période. Les périodes non traitées, codées par l'état "Sans Rien", c'est-à-dire "sans traitement", ne sont pas prises en compte dans le calcul des contributions.

Cela donne, dans les analyses de séquences présentées dans ce document, un suivi pouvant durer jusqu'à 6 ans par patient. Les séquences ne sont pas toutes de la même longueur, i.e. que tous les patients n'ont pas un suivi complet de 6 ans - mais ils ont un suivi minimal de 4 périodes. Pour simplifier la visualisation du clustering, les modalités des états ont été réduites aux grandes catégories de thérapies suivantes : monothérapie, bithérapie, trithérapie (ou plus), insuline, insuline combinée avec un anti-diabétique oral (OAD), sans rien (période sans traitement).

Le tapis de séquences ainsi que le chronogramme de ces données est visible dans la Figure 1.8 ci-dessous. Le tapis permet de visualiser la dimension individuelle des séquences : chaque ligne correspond à la séquence d'un patient. Le chronogramme présente une série de coupes transversales : pour chaque période les proportions d'individus dans les différentes situations sont visibles (monothérapie, bithérapie, etc.).

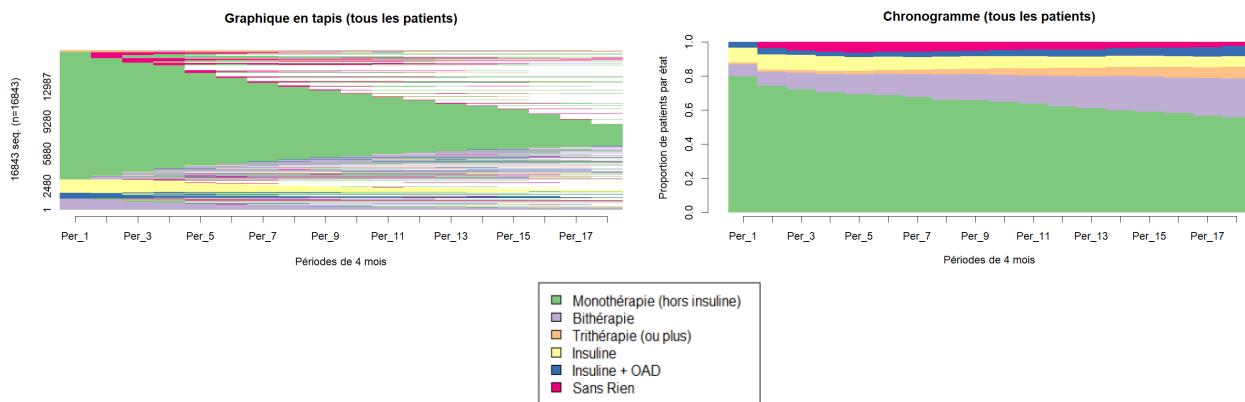


FIGURE 1.8 – Aperçu des séquences de traitement du diabète

En résumé, dans cette partie, l'environnement professionnel et la problématique du stage ont été introduits. Une collecte d'informations sur la maladie du diabète a été faite avant d'entamer la sélection de la population d'intérêt à partir d'une base de données IQVIA contenant les délivrances de traitements anti-diabétiques en pharmacie entre 2014 et 2021. Enfin, la base de données contenant les séquences de traitement a été construite de façon à ce qu'elle ressemble le plus possible aux réels suivis des patients.

Recherche bibliographique

Une recherche bibliographique a précédé la mise en oeuvre des méthodes sélectionnées. Dans cette partie, une revue de la littérature scientifique sur l'analyse de séquences sera présentée. Dans un premier temps, la définition d'une mesure de dissimilarité sera donnée avant de décrire les trois principales mesures présentes dans la littérature - l'Optimal Matching, le compte du nombre de sous-séquences communes, les distances basées sur les distributions des états dans les séquences. L'extension apportée par l'analyse multi-séquences sera ensuite introduite, avant de conclure par un tout autre type de méthode : les modèles mixtes à classes latentes.

2.1 Mesures de dissimilarités

En pratique, la comparaison de deux séquences se fait en utilisant la mesure de dissimilarité pour quantifier le niveau de ressemblance (ou non) entre les séquences. Les différentes mesures de dissimilarité ont chacune des propriétés spécifiques et une sensibilité plus ou moins forte à chacune des trois dimensions temporelles.

En classification, une dissimilarité est une fonction mesurant la ressemblance (ou dissemblance) entre les individus ou points d'un ensemble. Une fonction $\Delta(x, y)$ est une dissimilarité entre deux points x et y si elle remplit les trois conditions suivantes :

- $\Delta(x, y) = 0 \Leftrightarrow x$ et y sont les mêmes
- $\Delta(x, y) \geq 0$: les distances sont positives ou nulles
- $\Delta(x, y) = \Delta(y, x)$: la symétrie

De plus, une dissimilarité est une métrique (ou distance) si elle remplit la quatrième condition ci-dessous :

- $\Delta(x, y) \leq \Delta(x, z) + \Delta(z, y)$: l'inégalité triangulaire (la distance de x à y ne peut être plus grande que la distance de x à n'importe quel autre point *plus* la distance de ce point à y - i.e. qu'il n'y a pas raccourcis)

Les dissimilarités, qui ont déjà été présentées plus haut dans le rapport, peuvent se distinguer en trois classes :

1. Les distances "éditées", qui mesurent le coût des opérations qui sont nécessaires pour transformer une séquence en une autre
2. Les mesures basées sur le compte des attributs communs entre les séquences (i.e. sur le compte du nombre de points communs)
3. Les distances entre les distributions de probabilité

Il est important que les mesures de dissimilarité soient métriques, en particulier si l'objectif est de les utiliser dans le cadre d'outils d'analyse de clusters, qui reposent sur des dissimilarités nécessitant une relation cohérente globale, de façon à ce qu'elles soient utilisées pour construire un espace dans lequel les observations peuvent être mises en réseau.

2.2 Optimal Matching

2.2.1 Concept

L'Optimal Matching (OM) mesure la distance entre chaque paire de séquences en définissant le nombre d'opérations (ou éditions) nécessaires pour rendre deux séquences strictement identiques (insertions-délétions et substitutions). A chaque opération est associé un coût, qui peut varier en fonction des états. Le coût minimum de la transformation - i.e. le coût des opérations nécessaires pour rendre deux séquences identiques - correspondra à la distance entre les deux séquences.

2.2.2 Détails de la méthode

L'OM utilise trois opérations élémentaires pour calculer la distance entre deux séquences : l'insertion, la délétion et la substitution (cf. Figure 2.1). La délétion et l'insertion, aussi appelées opérations d'indel, reposent toutes les deux sur le concept de distorsion du temps (ou concept d'indel). Une délétion (respectivement une insertion) consiste en la suppression (respectivement l'ajout) d'un état dans la séquence. La substitution respecte la dimension temporelle de la séquence mais modifie la séquence des états, en remplaçant un état par un autre état différent au même instant t.

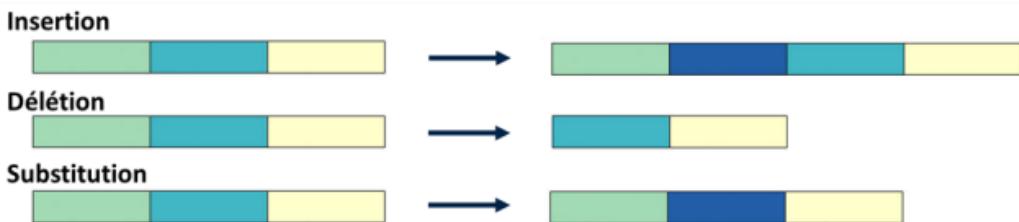


FIGURE 2.1 – Opérations élémentaires utilisées en OM

Les trois opérations décrites ci-dessus servent de base au calcul de la mesure de distance entre deux séquences. A chaque opération (indel ou substitution) est assigné un coût spécifique pouvant être constant ou pouvant varier en fonction des états subissant une transformation. La distance entre deux séquences correspond à la somme des coûts de chacune des opérations nécessaires à la transformation d'une séquence en une autre. L'algorithme cherche le chemin le plus court et le moins coûteux pour rendre deux séquences identiques. Dans l'exemple ci-dessous, il faudrait transformer les états apparaissant en gras.

Patient 1	A	A	B	B	C
Patient 2	A	B	B	C	C

Pour ce faire, une substitution de **A** vers **B** et de **B** vers **C** est possible dans la séquence du Patient 1 (ou inversement pour le Patient 2).

Patient 1	A	A	\rightarrow B	B	\rightarrow C	C
Patient 2	A	B		B	C	C

Patient 1	A	A	B	B	C	
Patient 2	A	B	\rightarrow A	B	\rightarrow B	C

Autre possibilité : une délétion de A dans la séquence du Patient 1 donnerait la séquence ABBC et une insertion de C à la fin de cette dernière rendrait la séquence identique à celle du Patient 2.

Patient 1		A	B	B	C
Patient 2	A	B	B	C	C

Patient 1	A	B	B	C	C
Patient 2	A	B	B	C	C

Les possibilités sont multiples. Par conséquent, le choix final des opérations effectuées dépendra du coût d'une insertion/délétion d'un état et de celui d'une substitution entre deux états.

La définition des coûts de substitution et d'indel est une étape majeure de l'OM.

La matrice de coûts de substitution obtenue est symétrique - substituer A par B a le même coût que de substituer B par A - et diagonale - substituer un état par lui-même a un coût nul. Il existe trois stratégies d'attribution des coûts de substitution :

- Coûts basés sur la théorie : utiliser des connaissances a priori pour déterminer un ordre de grandeur de similarité entre deux états, nous permettant de classer les substitutions possibles.
- Coûts basés sur les attributs des états : il s'agit de spécifier une liste d'attributs pour lesquels nous voulons évaluer la proximité entre les états - à condition que chaque état ait des particularités/attributs qui lui sont propres.
- Coûts déterminés par les données : utiliser directement la composition des séquences pour attribuer les coûts, e.g. en attribuant des coûts plus élevés lorsque la substitution entre états est rare et un coût faible lorsque des transitions fréquentes sont observées.

Contrairement aux coûts de substitution, le(s) coût(s) d'indel ne se présente(nt) pas sous la forme d'une matrice. Il existe deux stratégies d'attribution du coût d'indel :

- Coût d'indel unique : l'indel étant considéré comme un opérateur permettant de combler les écarts dans la longueur des séquences, la plupart des applications utilisent le même coût d'indel, quel que soit l'état d'insertion ou de déletion. De façon à respecter l'inégalité triangulaire [2], le coût attribué devra se situer entre la moitié du coût maximal de substitution et la moitié du coût maximal de substitution multiplié par la longueur maximale d'une séquence.
- Coûts d'indel dépendant de l'état : les états plus rares ont un coût plus élevé ; e.g. il est possible de définir un coût d'indel d'un état comme une fonction monotone - tel qu'un logarithme - de l'inverse de la fréquence totale observée de cet état.

Le coût d'indel s'il est unique est spécifié une fois les coûts de substitution définis. Dans le cas où plusieurs coûts d'indel sont définis, ils le sont avant les coûts de substitution. Des coûts d'indel élevés rendent la dissimilarité extrêmement sensible au temps, ce qui revient à privilégier l'utilisation de l'opération de substitution lorsque les séquences sont de même longueur. C'est pourquoi, lorsque l'objectif est d'identifier des portions de séquences identiques mais décalées dans le temps, il est préférable de définir des coûts d'indel faibles. Quand le coût d'indel est unique et est inférieur ou égal à la moitié du coût minimal de substitution, seules les opérations d'indel sont utilisées. Cela revient à utiliser la distance LCS (ou Levenshtein II), qui sera définie dans la section 2.3.3.

2.2.3 Variantes

Il est reproché à l'OM de ne pas suffisamment prendre en compte le contexte. Seules les valeurs d'une paire d'états, dans chaque séquence, sont prises en compte et leur environnement (états futurs, états précédents) est invisible. L'idée de ces méthodes est de modifier les coûts de façon à prendre en compte le contexte de l'édition. Des variantes de l'OM existent pour prendre en compte la position des états (i.e. le moment où ils ont lieu) lorsque des opérations de transformation sont appliquées. Ces variantes ont pour objectif de rendre les opérations plus sensibles au contexte en les faisant dépendre soit de la position dans la séquence quand l'opération s'applique, soit des patterns/états entourant cette position. Malheureusement, la plupart de ces approches ne respecte pas les propriétés d'une métrique (notamment l'inégalité triangulaire), limitant considérablement leur utilité. Le détail de ces méthodes est disponible en Annexe A.7.

OM localisé (OMloc)

L'OM localisé fait dépendre les coûts d'indel des états adjacents [2] [6]. La motivation d'une telle méthode vient du fait qu'insérer ou supprimer un état qui est similaire à ses voisins ne changerait que la longueur de l'épisode concerné sans affecter l'ordre d'apparition des états.

OM sensible à la durée des épisodes (OMslen)

Dans cette variante de l'OM [7], les coûts dépendent de la durée des épisodes dans laquelle a lieu l'opération. Autrement dit, lorsqu'il est nécessaire de supprimer, ajouter ou substituer un seul état présent dans un épisode, le coût de cette édition sera d'autant plus faible que l'épisode concerné est grand. Dans le cadre d'une séquence de traitement, cette méthode considère donc qu'il est moins coûteux de modifier un état provenant d'une longue période de traitement, que d'en modifier un provenant d'une très courte période de traitement.

OM sur les séquences d'épisodes

Ce ne sont plus des séquences d'états qui sont considérées mais des séquences d'épisodes. Un épisode correspond à une suite d'états identiques expérimentée par l'individu dans une séquence (e.g. AA, BB et C sont des épisodes dans la séquence AABBC). Cette variante [2] [6] [8] permet de substituer ou d'insérer/supprimer un grand nombre d'états identiques en une seule fois à faible coût.

OM sur les transitions d'états (OMstran)

Cette méthode recode les séquences sous forme de séquences de transitions avant d'utiliser l'OM sur ces nouvelles séquences [2] [6]. La séquence aabbb deviendrait par exemple aa-ab-bb-bb [8].

2.3 Distances basées sur le compte des attributs communs entre les séquences

Dans les années 1990, Dijkstra et Taris (1995) ont proposé une façon de regarder la similarité des données longitudinales qui se focalise sur l'ordre d'apparition des états. Dans ce cas, plus deux séquences expérimentent les mêmes états dans le même ordre, plus elles sont considérées comme similaires. Différentes approches sont présentées ci-dessous.

2.3.1 Distance de Hamming simple

Hamming (1950) a proposé de mesurer la dissimilarité entre deux séquences en utilisant le nombre d'états non-concordants, i.e. le nombre de périodes/positions où les états sont différents entre deux séquences. La distance de Hamming (HAM) repose sur la comparaison de la position des états dans une séquence et ne peut par définition être appliquée que sur des séquences de même longueur. Par exemple, la distance entre la séquence ABCD et la séquence CABC sera maximale, aucun état n'ayant lieu au même moment - alors que les patients ne sont peut-être pas si éloignés, ayant chacun l'enchaînement ABC.

La distance de Hamming simple peut s'apparenter à un OM classique n'utilisant que les opérations de substitution avec des coûts fixes égaux à 1. Il est tout de même possible de définir, comme en OM, des coûts de substitution dépendants des états, calculés automatiquement ou déterminés par l'utilisateur. Le tableau A.3 illustre deux exemples de calcul de distance entre deux séquences. Dans le premier exemple, les coûts sont fixes et égaux à 1, peu importe l'état substitué. Dans le second, les coûts sont dépendants des états, et une substitution de C vers B (ou inversement) est considérée comme plus coûteuse qu'une substitution de A vers B.

TABLE 2.1 – Distance de Hamming : exemple

Séquence 1	A	A	B	C	
Séquence 2	A	B	B	B	
Coûts fixes	0	1	0	1	Distance= 2
Coûts dépendants des états	0	0.5	0	1	Distance= 1.5

2.3.2 Distance de Hamming Dynamique (DHD)

Dans cette variante de la distance de Hamming, qui devient "dynamique", les coûts de substitution sont liés au temps, i.e. qu'ils dépendent de la période t où l'état a lieu dans la séquence.

Une matrice de coûts de substitution est calculée pour chaque période. Les coûts de substitution sont dérivés à la période t des taux de transition observés entre $t-1$ et t et entre t et $t+1$, nous donnant le coût de substitution de a vers b (ou inversement), dépendant de t , $\gamma_t(a,b)$ suivant [2] :

$$\gamma_t(a,b) = 4 - p_t(b|a) - p_t(a|b) - p_{t+1}(b|a) - p_{t+1}(a|b)$$

où $p_t(b|a)$ est la probabilité de passer de a à b entre $t-1$ et t et est estimée par $\frac{n_{t-1,t}(a,b)}{n_{t-1}(a)}$ où $n_{t-1,t}(a,b)$ est le nombre de transitions de a vers b de la période $t-1$ à t sur la population totale, et $n_{t-1}(a)$ est le nombre total de a à la période $t-1$.

2.3.3 Distance basée sur la longueur de la plus longue sous-séquences commune (LCS)

Une sous-séquence est obtenue en supprimant n'importe quel nombre d'états dans une séquence, en conservant l'ordre d'apparition de ces derniers, e.g. AAB et BD sont des sous-séquences de ABCD. La distance LCS (ou Levenshtein II) mesure la distance entre deux séquences en les comparant par rapport à la longueur de la plus longue sous-séquence commune.

Contrairement à la distance de Hamming, qui ne permet pas les indels, la distance LCS ignore les substitutions et mesure la distance avec le coût associé au nombre d'opérations d'indel nécessaires pour transformer une séquence en une autre. La distance de Levenshtein II s'apparente à la méthode d'OM classique avec une matrice de coûts de substitution où tous les coûts sont égaux à 2 et où le coût d'indel est unique et égal à 1, i.e. où les opérations de substitution ne sont jamais utilisées car cela revient toujours moins cher d'utiliser les indels que les substitutions.

Si $N(s_1, s_2)$ correspond à la longueur de la plus longue sous-séquence commune dans les séquences s_1 et s_2 , alors la distance LCS s'écrit : $d_{LCS}(s_1, s_2) = N(s_1, s_1) + N(s_2, s_2) - 2N(s_1, s_2)$, avec $N(s_1, s_1)$ le nombre d'états de s_1 et $N(s_2, s_2)$ le nombre d'états de s_2 . Lorsque les séquences s_1 et s_2 sont strictement identiques, $d_{LCS}(s_1, s_2) = 0$.

Un exemple de calcul de cette distance est visible dans le tableau 2.2, pour deux séquences de même longueur : AABC et AABB. Pour rendre les deux séquences identiques, il faudrait par exemple supprimer "C" dans la séquence 1 (opération "d" dans le tableau) et ajouter "B" dans cette même séquence (opération "i" dans le tableau). Par construction, cette méthode peut traiter des séquences de longueurs différentes.

TABLE 2.2 – Distance de Levenshtein II : exemple

Séquence 1	A	A	B	C	-	
Opérations				d	i	
Séquence 2	A	A	B	-	B	
Coûts fixes	0	0	0	1	1	Distance= 3

Puisqu'elle n'est pas basée sur des appariements relatifs à la position des états (comme pour les distances qui seront présentées en section 2.4), la distance LCS n'est pas sensible au timing mais aux différences dans la distribution des états et au séquençage. En particulier, la distance LCS est sensible à l'ordre des états les plus fréquents et, dans une moindre mesure, aux différences au niveau du temps passé dans les états distincts.

2.3.4 Nombre de sous-séquences correspondantes (NMS)

Dans cette méthode, comme pour LCS, une sous-séquence est obtenue en supprimant n'importe quel nombre d'états dans une séquence. L'idée de mesurer la distance par le nombre de sous-séquences correspondantes (NMS) est que, plus souvent un ordre donné d'états dans une séquence est observé dans une autre séquence, plus les deux séquences sont proches l'une de l'autre. La métrique de représentation vectorielle des sous-séquences (SVRspell) est une extension de cette méthode, qui permet de prendre en compte la longueur des sous-séquences dans le calcul des distances et peut donner plus d'importance à la durée des épisodes inclus dans les sous-séquences. Les détails de cette méthode sont disponibles en Annexe A.7.5.

2.4 Distances entre les distributions de probabilité

Distance entre les distributions d'états

L'approche met l'accent sur la distribution longitudinale des états dans chaque séquence, i.e. le temps passé dans chaque état au sein des séquences.

Par exemple, pour un alphabet {A,B,C,D}, le vecteur de distribution des états de la séquence ABACBB est (1/3,1/2,1/6,0). La dissimilarité entre séquences est mesurée par la distance entre les vecteurs de distribution en utilisant soit la distance euclidienne, soit la distance du khi-deux. La première distance tient compte des différences absolues dans la proportion de temps passé dans les états.

La seconde pondère les différences au carré pour chaque état par l'inverse de la proportion globale de temps passé dans l'état. Cette dernière donne plus d'importance à un état "rare" qu'à un état fréquent, tandis que la distance euclidienne est par définition plus sensible aux différences dans les épisodes de longue durée. Avec $p_{j|x}$ la proportion de temps passé dans l'état j dans la séquence x , $|\Sigma|$ le nombre total d'états présents dans les séquences, et p_j la proportion globale de temps passé dans l'état j , la distance du khi-deux des distributions d'états s'écrit :

$$d_{\text{chi}(x,y)}^2 = \sum_{j=1}^{|\Sigma|} \frac{(p_{j|x} - p_{j|y})^2}{p_j}$$

Cette méthode est sensible au temps passé dans les états et insensible à l'ordre et au timing des états. Il est possible d'y remédier en découplant la séquence en K périodes de temps et en calculant les distributions des états sur chacune des K périodes. La distance devient la somme des distances du khi-deux pour chaque période. Avec $p_{j|x_k}$ la proportion de temps passé dans l'état j dans la séquence x dans la période k et $p_{j|k}$ la proportion globale de temps passé dans l'état j dans la période k , la distance du khi-deux des distributions d'états sur K périodes successives s'écrit :

$$d_{\text{chi},K(x,y)}^2 = \sum_{k=1}^K \sum_{j=1}^{|\Sigma|} \frac{(p_{j|x_k} - p_{j|y_k})^2}{p_{j|k}}$$

Les périodes peuvent se chevaucher dans le calcul de la distance. Lorsque K est égal à la longueur des séquences, la distance correspond à un comptage pondéré des états non-concordants. La méthode devient très sensible aux timings non-concordants du fait de sa dépendance à la position des états et, par conséquent, gagne une certaine sensibilité au séquençage.

2.5 Analyse multi-séquences

Une autre approche utilisée dans la création d'une typologie de trajectoires est l'analyse multiséquences (MCSA) [9], qui est une extension de l'analyse de séquences unidimensionnelle présentée précédemment. Dans la MCSA, plusieurs séquences sont prises en compte par individu. Chaque séquence traduit une dimension de son parcours global. La MCSA suit les mêmes étapes que l'analyse de séquence unidimensionnelle, i.e. la construction des séquences, le calcul de la distance entre ces dernières grâce à l'OM et enfin leur classification.

Le nombre de séquences prises en compte par individu change : chaque séquence correspondra à une dimension du parcours du patient et possédera son propre nombre d'états. Dans le cadre de séquences de traitement, il pourrait être envisagé d'ajouter à chaque patient une séquence représentant une autre dimension de son parcours, comme ses rendez-vous médicaux et ses séjours hospitaliers.

Chaque dimension possédant ses propres états (i.e. qu'il y a un alphabet différent par dimension), autant de matrices de coûts de substitution et de valeurs d'indel seront dérivées qu'il y aura de dimensions dans le parcours global du patient. La distance entre deux parcours sera calculée en appliquant l'OM comme décrit précédemment, en utilisant les coûts issus de chaque dimension. Les distances obtenues vont pouvoir être utilisées pour identifier la typologie de parcours finale grâce à un algorithme de clustering comme la CAH.

2.6 Modèles mixtes à classes latentes

L'utilisation de modèles mixtes devient de plus en plus populaire pour l'analyse de données longitudinales. Alors que le modèle linéaire mixte assume que la population observée est homogène, et s'applique sur des marqueurs longitudinaux continus ayant des variabilités aléatoires gaussiennes, les modèles mixtes linéaires généralisés étendent cette théorie aux résultats longitudinaux binaires, ordinaux ou de Poisson. Les modèles mixtes à processus latent ont été conçus dans le but d'étudier des marqueurs longitudinaux non-gaussiens. Ce sont des modèles permettant l'identification de classes latentes suivant des parcours similaires dans leur développement temporel. Le package R *lcmm* [10] offre une série de fonctions pour adresser les différentes extensions du modèle mixte linéaire, en particulier lorsqu'une hétérogénéité non-observée existe dans la population observée. De plus, ces fonctions peuvent estimer des modèles statistiques basés sur des modèles mixtes adaptés pour des résultats longitudinaux (la variable d'intérêt) pouvant être gaussiens, binaires, ordinaux ou continus mais asymétriques (e.g. l'absence ou la présence de symptômes, une échelle psychologique).

L'approche consiste en la séparation du modèle structurel, qui décrit la quantité d'intérêt (le processus latent) en fonction du temps et des covariables, du modèle de mesure, qui lit la quantité d'intérêt aux observations [11]. Chaque phénomène dynamique peut ainsi être caractérisé par un processus latent, noté $\Lambda(t)$ qui évolue dans le temps t (continu ou non). La quantité d'intérêt $\Lambda(t)$ est définie en utilisant un modèle linéaire mixte standard fonction du temps :

$$\Lambda(t) = X(t)\beta + Z(t)u_i + w_i(t)$$

où :

- $X(t)$ et $Z(t)$ sont des vecteurs de covariables ($Z(t)$ est inclus dans $X(t)$) ;
- β sont les effets fixes, i.e. les effets moyens de la population observée ;
- u_i sont les effets aléatoires, i.e. les effets individuels normalement distribués de moyenne zéro et de matrice de covariance B ;
- $w_i(t)$ est un processus gaussien pouvant être ajouté au modèle pour assouplir la structure de corrélation intra-individu.

La relation entre le processus latent et les observations de la variable d'intérêt Y_{ij} pour l'individu i et l'événement j est définie, au temps de mesure t_{ij} , par un modèle de mesure flexible non-linéaire :

$$Y_{ij} = H(\Lambda(t_{ij}) + \epsilon_{ik}; \eta)$$

où :

- t_{ij} est le moment où l'événement j a lieu de pour l'individu i ;
- ϵ_{ij} est une erreur gaussienne indépendante de moyenne zéro ;
- H est la fonction de lien (paramétrée par η).

Différentes familles paramétriques sont utilisées. Quand la variable d'intérêt est continue, H^{-1} est une famille paramétrique de fonctions monotones croissantes. Quand la variable d'intérêt est discrète (binaire ou ordinaire) : H est une fonction "de seuil" (*thresholds function*), i.e. que chaque niveau de la variable d'intérêt Y correspond à un intervalle de $\Lambda(t_{ij}) + \epsilon_{ij}$ dont les bornes sont à estimer. Dans ce cas, un modèle probit (cumulatif) est appliqué. Par ailleurs, la complexité numérique du modèle avec cette fonction de lien est bien plus importante qu'avec les autres fonctions de lien, du fait de l'intégration numérique sur la distribution des effets aléatoires. Il faut faire attention au nombre d'effets aléatoires à inclure dans le modèle.

2.7 Bag-of-Words

Le Bag-of-Words (BoW ou "sac de mots") [12] est une tout autre méthode sortant du contexte de l'analyse de séquences. L'idée derrière cette méthode issue du traitement de langage est de voir l'ensemble des séquences comme des textes dans un corpus. Les états sont considérés comme des mots dans un document. L'occurrence des mots va être comptée dans chaque document puis réévaluer en fonction de leur fréquence d'apparition dans tout le corpus, de sorte que les scores des mots fréquents qui sont également fréquents dans tous les documents soient pénalisés. Un description plus détaillé de la méthode appliquée sur les séquences de traitement est visible en Annexe A.8.

En résumé, pour mieux comprendre les méthodes qui existent dans la littérature avant de les appliquer, une recherche bibliographique est une étape indispensable avant l'application des méthodes sélectionnées. Une première approche consiste à voir les parcours de soins comme des séquences et a effectué une analyse de séquences via des mesures de dissimilarité. Trois types de mesures de dissimilarité existent : les distances éditées (l'OM), les distances basées sur le nombre de sous-séquences communes et les distances entre distributions d'états. L'extension apportée par l'analyse multi-séquences a également été introduite, avant de conclure par deux autres approches sortant de l'analyse de séquences : les modèles mixtes à classes latentes, un type de modèle particulier s'appuyant sur les probabilités a posteriori d'appartenir à une classe, et les Bag-of-Words, un modèle considérant les séquences d'états comme des mots dans un texte.

Résultats

L'application des méthodes étudiées s'est faite avec le langage de programmation R. Cette partie présentera les principaux résultats obtenus suite à la revue de la littérature effectuée pour la sélection des méthodes, dans le même ordre que la partie précédente. Les résultats de l'OM seront présentés de façon plus détaillée que les autres méthodes. Enfin, le choix de la méthode la plus appropriée en fonction des objectifs de l'analyse de séquences sera traité, avant d'ouvrir sur l'interprétation économique qui peut être tirée des résultats.

3.1 Optimal Matching

L'algorithme implémenté dans R est celui de Needleman–Wunsch (1970). Le détail de celui-ci est ainsi que de la librairie TraMineR et des fonctions utilisées dans le cadre de l'analyse de séquences est disponible en Annexes A.9.1 et A.9.2.

3.1.1 Choix des coûts

Les méthodes de calcul de coûts en OM, "TRATE", "INDELSLOG" et "INDELS" proposées par TraMineR ont été utilisées à titre comparatif. La méthode "TRATE" est la plus classique de l'OM, où le coût d'indel est unique et les coûts de substitution sont calculés à partir des probabilités de transition entre états. Le coût de substitution $\gamma(a, b)$ de a vers b (ou de b vers a) est égal à $\gamma(a, b) = 2 - P(a|b) - P(b|a)$. La matrice des coûts de substitution calculés à partir des séquences de traitement du diabète est visible dans la table 3.1.

TABLE 3.1 – Matrice de coûts de substitution obtenue avec OM TRATE

	Monothérapie	Bithérapie	Insuline	Insuline + OAD	Trithérapie	Sans Rien
Monothérapie	0	1.909	1.992	1.950	1.970	1.712
Bithérapie	1.909	0	1.996	1.965	1.893	1.970
Insuline	1.992	1.996	0	1.903	1.998	1.951
Insuline + OAD	1.950	1.965	1.903	0	1.969	1.994
Trithérapie	1.970	1.893	1.998	1.969	0	1.992
Sans Rien	1.712	1.970	1.951	1.994	1.992	0

Le coût d'indel c_I unique est calculé à partir du coût maximal de substitution : $c_I = \max(\text{coûts de substitution})/2 = 0.999$. Les trois méthodes de calcul de coûts de l'OM seront par la suite notées OM TRATE, OM INDELS et OM INDESLOG.

Avec OM INDELSLOG, les coûts d'indel dépendent des états et sont calculés dans un premier temps à partir de leur fréquence relative. Une transformation logarithmique est ensuite appliquée sur les coûts. Les états plus rares ont un coût associé plus élevé que les autres. Les matrices de coûts d'indel et de substitution de cette méthode sont visibles dans les tableaux A.9 et A.10 de l'Annexe A.10.

Avec OM INDELS, les états plus rares sont plus coûteux que dans la méthode OM INDELSLOG, car les coûts ne subissent pas de transformation logarithmique. Par exemple, dans le tableau A.11 de l'Annexe A.10, le coût d'indel de trithérapie est plus de 20 fois plus coûteux que le coût d'indel de la monothérapie.

3.1.2 Visualisation

Dendrogramme

Une première visualisation du clustering par un dendrogramme permet d'avoir une vue d'ensemble sur les étapes de classification. Le dendrogramme Figure 3.1 provient d'une Classification Ascendante Hiérarchique (CAH) classique utilisant la méthode de Ward. Les parcours patients semblent au premier abord se distinguer en fonction de la typologie des thérapies suivies.

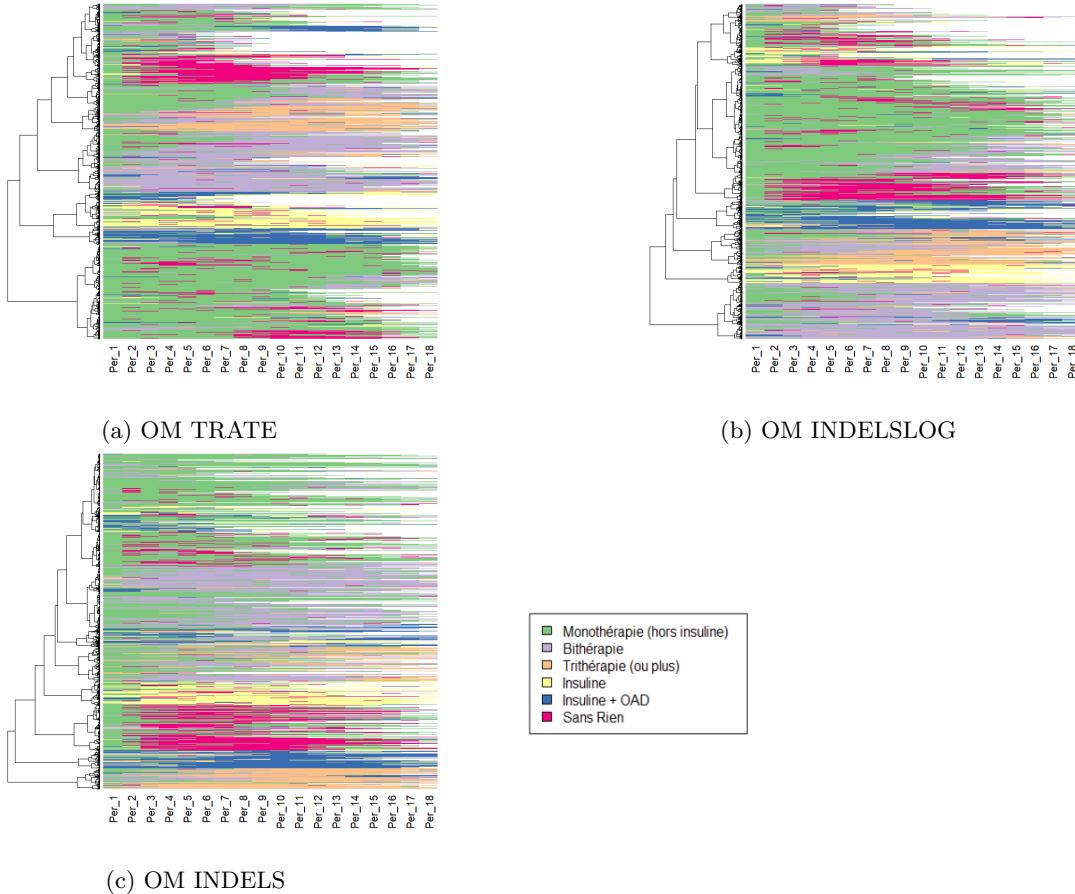


FIGURE 3.1 – Dendrogrammes issus de la CAH

Evolution des indicateurs de qualité

Les indicateurs de qualité du clustering listés en Annexe A.6 sont représentés ensemble pour la méthode OM TRATE dans la Figure A.4 sous forme de graphique, en fonction du nombre de clusters choisi. Ils sont normalisés de façon à être comparables. Les mêmes graphiques pour les méthodes OM INDELSLOG et OM INDELS sont visibles en Annexe A.12.1. Pour OM TRATE, les indicateurs montent significativement (ou diminuent pour le C de Hubert) jusqu'à 5 groupes puis continuent d'augmenter doucement, suggérant un clustering à 5-6 groupes. OM INDELS suit la même tendance tandis que pour OM INDELSLOG les indicateurs commencent à diminuer (à augmenter pour le C de Hubert) à partir de 5-6 groupes.

Tapis de séquences

La visualisation des clusters par tapis de séquences est indispensable pour conclure de la qualité des classes créées et comparer les méthodes testées.

L'épaisseur de chaque ligne des tapis de séquences est proportionnelle au poids de la séquence dans la stratification, i.e. à sa fréquence d'apparition dans le cluster. Parmi les 16843 séquences obtenues après la sélection de la population, 4933 séquences distinctes sont considérées dans la classification. Les lignes sont ordonnées en fonction du coefficient de silhouette. Cela signifie que, plus une séquence est située en haut du tapis, plus elle est proche du centre de la classe et plus elle est "représentative" du cluster (cf. Annexe A.5 pour plus d'informations sur le coefficient de silhouette). L'axe des ordonnées indique le nombre de séquences

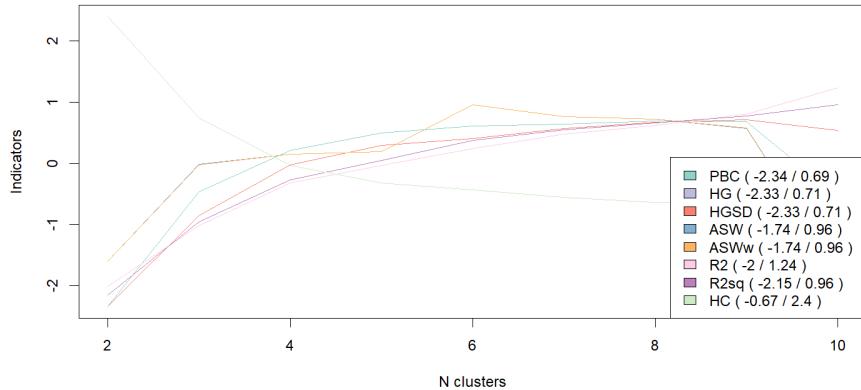


FIGURE 3.2 – Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM TRATE)

distinctes par groupe, tandis que le nombre inscrit entre parenthèse au dessus de chaque tapis correspond au nombre total de patients classés dans le groupe.

Deux algorithmes de classification non-supervisées ont été testées : la CAH (méthode de Ward) et l'algorithme Partitioning Around Medoids (PAM) qui est initialisé à partir des résultats de la CAH (cf. Annexe A.4). Les classes obtenues pour 5 groupes sous forme de tapis de séquences sont visibles dans les Figures 3.3 et 3.4.

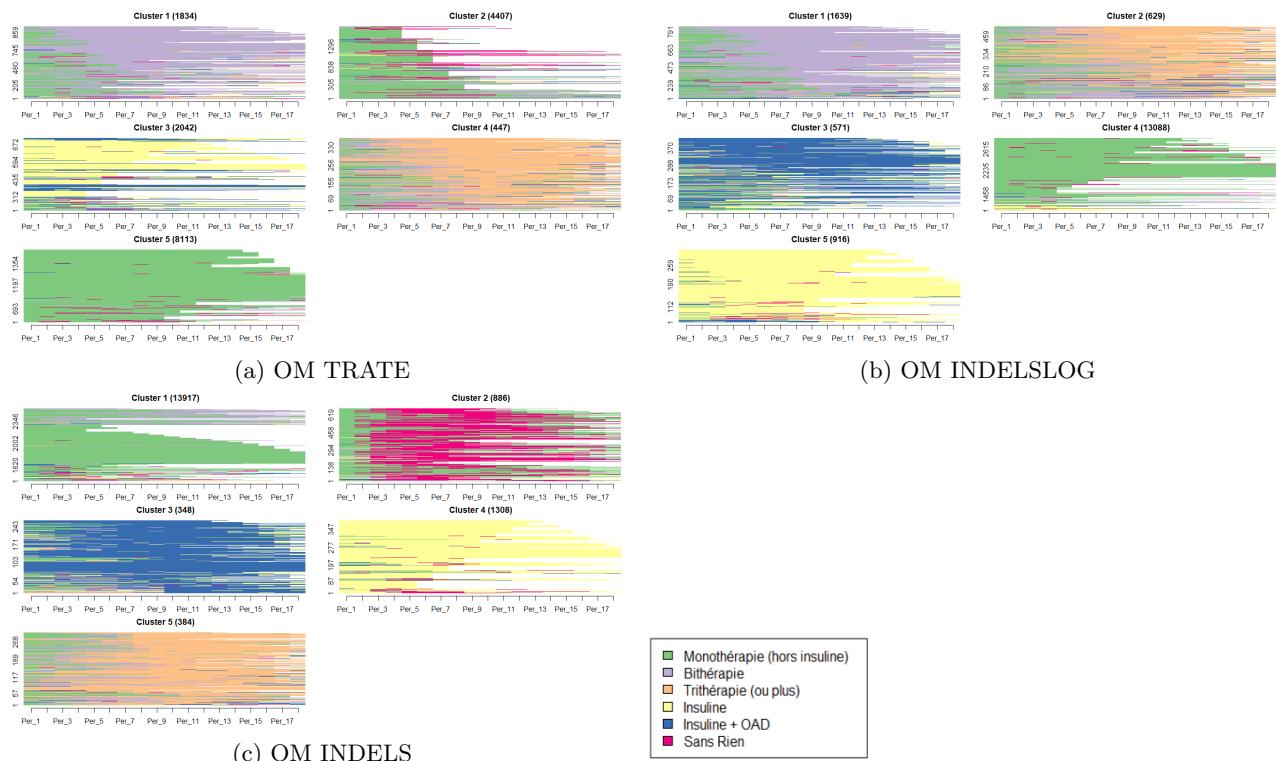


FIGURE 3.3 – OM (CAH) - 5 groupes

Pour 5 groupes avec la CAH, OM TRATE semble distinguer un groupe où les patients restent en monothérapie tout le long de leur suivi, un contenant les patients sous bithérapie, un sous insuline, un où les suivis sont très courts, et enfin un où la trithérapie domine le suivi. Quant à OM INDELS et OM INDELSLOG, ces méthodes semblent avoir mis l'accent sur les suivis plus rares : une distinction des patients sous insuline de ceux sous insuline + OAD est visible, ainsi qu'une distinction des patients dont le suivi est discontinu, i.e. contenant beaucoup de périodes de "Sans Rien". Avec la PAM, les classes obtenues sont moins nettes, notamment pour OM TRATE et OM INDELSLOG. Deux groupes contenant un mélange de suivis courts et de patients sous insuline + OAD sont distingués. Les classes issues d'OM INDELS sont plus satisfaisante :

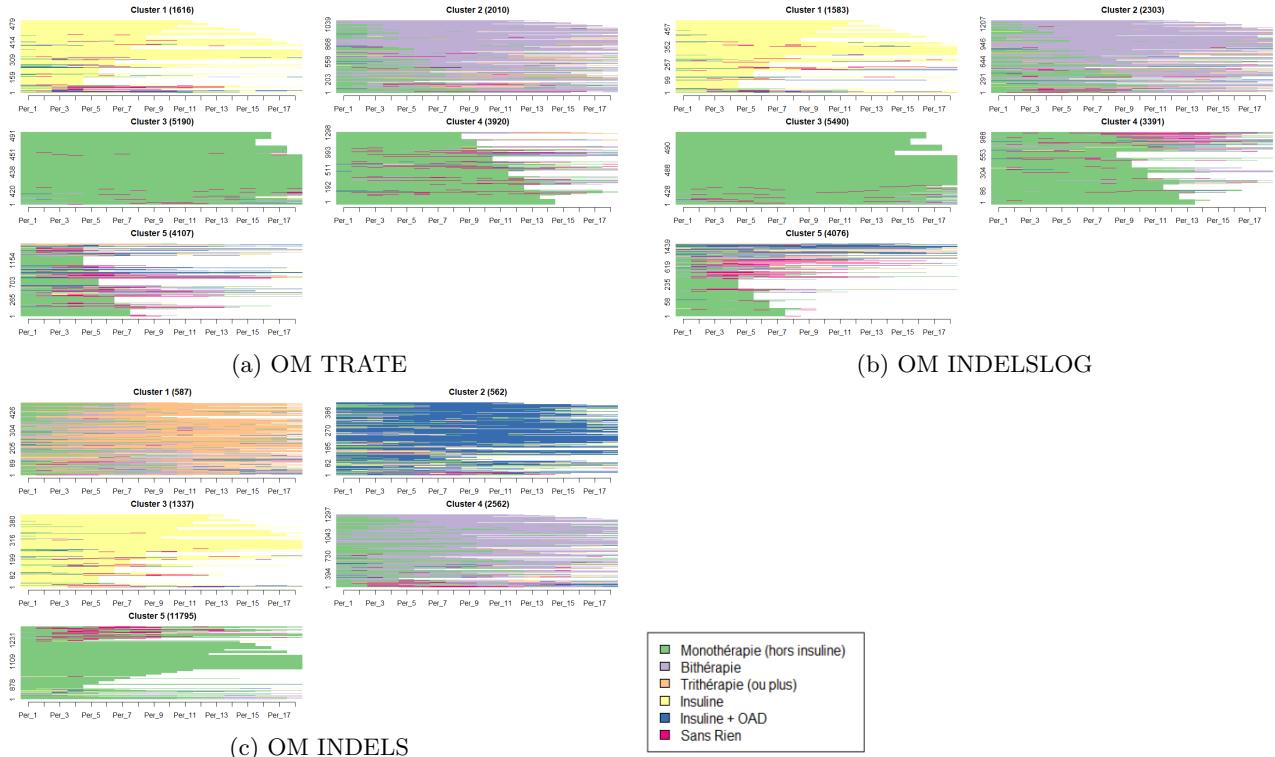


FIGURE 3.4 – OM PAM - 5 groupes

En ayant sous les yeux les classifications de ces trois méthodes de coût pour 7 et 10 groupes (cf. Figures A.5, A.6, A.7 et A.8 en Annexe), une interprétation des méthodes peut être déduite :

- OM TRATE semble séparer les suivis courts parce que le coût d'indel est unique. Autrement dit, combler les séquences courtes est une opération coûteuse.
- OM INDELSLOG permet de donner de l'importance aux états rares, ce qui explique pourquoi les trithérapies sont distinguées ainsi que les patients sous combinaison d'insuline et d'anti-diabétique oral.
- OM INDELS donne encore plus d'importance aux états rares (i.e. *Trithérapie, Insuline + OAD, Sans Rien*) et sépare aussi les "Sans Rien".

D'un point de vue épidémiologique, le clustering à 7 et 10 classes semble plus intéressant car il distingue des sous-groupes pertinents : une classe contenant les suivis discontinus, des classes distinguant les bithérapies commençant au début du suivi ou en fin de suivi, une distinction des patients uniquement sous insuline de ceux qui sont à la fois sous insuline et sous un traitement anti-diabétique oral.

Description des groupes

Afin de mieux comprendre la classification, il est d'intérêt de représenter visuellement les facteurs caractérisant la composition des groupes. Dans le cadre de l'exemple sur le diabète et de l'utilisation des données de la base LRx pour obtenir les délivrances de traitements, il est possible de connaître l'âge et le sexe du patient. Un algorithme de typage du diabète a été appliqué sur la base de données de façon à déterminer, à partir des délivrances, si le patient est DT1, DT2 ou s'il s'agit d'un diabète gestationnel. Le fonctionnement de cet algorithme est détaillé dans la figure A.13 en Annexe.

Pour le clustering à 5 groupes obtenu avec OM TRATE, une simple table descriptive peut être dérivée (cf. Table 3.2). L'âge médian des patients varie de 63 à 70 ans dans les groupes, et les hommes sont plus représentés que les femmes, comme observé dans la population diabétique en général.

Une représentation graphique plus complète des variables statiques par groupe est visible dans les Figures A.39 et A.40 en Annexe. Globalement, les proportions affichées dans ces graphiques semblent cohérentes avec la réalité. Par exemple, la grande majorité (82.9 %) des patients de moins de 18 ans présents dans les séquences font partie du "Cluster 3" qui correspond à la classe des patients sous insuline (les DT1). Les DT2 sont en effet plus âgés en moyenne (plus de 60 ans). De plus, les diabètes gestationnels sont tous classés dans le "Cluster 3", ce qui est également cohérent compte tenu de la prohibition de la metformine pour une patiente enceinte.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Nombre de patients	1834	4407	2042	447	8113
Sexe (% de femmes)	36.4 %	43.3 %	51.8 %	34.9 %	42.6 %
Âge médian à la fin du suivi	66	70	70	63	71

TABLE 3.2 – Description simple des clusters

Il est également possible de regarder la proportion des classes de médicament pris par les patients lors de leur suivi en fonction du groupe auquel ils appartiennent, en ramenant les états des séquences aux molécules. En Figure A.41, deux illustrations sont visibles où le détail est donné à l'année et où seules les classes de médicament représentant plus de 1 % des patients du groupe apparaissent sur le graphique. Les molécules affichées sont les suivantes : MET (metformine), SUL (sulfamides ou glinides), GLP (analogue GLP1), DPP (inhibiteur DPP4), SGL (SGLT2), RIN (insuline à action rapide), AIN (autre type d'insuline), AUT (autre type de molécule), Sans Rien. Lorsque deux molécules ou plus sont séparées d'une virgule apparaît dans la légende, cela signifie que les molécules sont combinées pour faire un médicament (bi ou trithérapie). Le groupe "Cluster 1" correspond aux patients passant une grande partie de leur suivi en bithérapie, le groupe "Cluster 3" correspond aux patients sous insuline ou combinaison d'insuline et OAD et la classe "Cluster 5" rassemble les patients sous monothérapie la majorité de leur suivi.

Indicateurs de qualité

Quant à l'interprétation statistique du clustering, elle peut être donnée par les indicateurs de qualité introduits précédemment (Annexe A.6). L'ASW pour OM TRATE à 5 groupes avec la CAH s'élève à 0.233 (0.311 pour OM INDELSLOG et 0.341 pour OM INDELS), ce qui est considéré comme statistiquement médiocre (< 0.5), quelque soit la méthode. Les indicateurs pour la PAM ne changent que peu de ceux de la CAH (cf. Table A.13c). Les résultats visuels fournis par la représentation des tapis de séquences sont plus déterminants dans l'évaluation de la qualité de la classification.

Variation de certains paramètres

Le résultat de l'application de l'OM (coûts TRATE) sur des séquences contenant les combinaisons de molécules utilisées dans les traitements anti-diabétiques (41 états considérés au lieu de 6) sont visibles en Annexe A.12.2. Augmenter le nombre d'états conduit à une classification plus précise en termes de parcours mais rend la tâche de l'algorithme plus difficile pour différencier certaines séquences atypiques. En effet, cela résulte en une classe hétérogène contenant les suivis éloignés des autres.

L'OM sur des séquences réduites à 6 périodes d'1 an (au lieu de 18 de 4 mois), est visible en Annexe A.12.2. Lorsque le nombre de périodes est réduit à 6, le nombre de séquences distinctes diminue considérablement. Ce dernier passe de 4933 à 664. Les groupes obtenus sont moins nets. De l'information est en effet perdue avec un lissage sur 6 périodes, ne facilitant pas le calcul des distances. Néanmoins, les groupes obtenus ont une allure qui reste similaire aux groupes obtenus avec des séquences dont les suivis sont détaillés sur plus de périodes.

L'analyse des groupes donnée par la suite sera concentrée sur les résultats de la CAH, les résultats de la PAM différents peu ou n'étant pas aussi intéressants visuellement - ils restent disponibles dans les annexes associées à la visualisation du clustering des méthodes.

3.2 Variantes de l'OM

Les tapis de séquences des différents clustering appliqués sur les variantes de l'OM sont visibles en Annexe A.12.4.

Les figures A.15 et A.16 (a) présentent le résultat de l'OM sensible à la longueur des épisodes(OMslen). Dans cette méthode, les coûts sont d'autant plus faibles que les épisodes impactés sont longs (paramètre $h = 1/2$).

Les figures A.15 et A.16 (b) présentent le résultat de l'OM localisé (OMloc), où les coûts dépendent des états adjacents (paramètres $\alpha = 0.8$ et $\beta = 0.1$).

Les figures A.15 et A.16 (c) présentent le résultat de l'OM entre séquences d'épisodes (OMspell), où ce ne sont plus les séquences d'états qui sont considérés, mais les séquences d'épisodes - des épisodes de longueur 1 à 11

sont considérés. Le paramètre δ est fixé 0.5 et correspond au coût d'expansion ou de compression d'une séquence par une unité de temps.

Enfin, les figures A.15 et A.16 (d) présentent le résultat de l'OM sur les transitions d'états (OMstran), où les séquences sont recodées sous forme de transitions. En appliquant cette méthode, R détermine 36 états de transition distincts. Le paramètre $w = 0.5$ est fixé et contrôle le compromis entre le coût lié à l'état d'origine et la transition.

Ces méthodes ne semblent pas améliorer les résultats du clustering. La complexité de ces variantes de l'OM fait dépendre les coûts du contexte, de façon à mettre davantage l'accent sur le séquençage (OMloc, OMstran) ou la durée des états (OMslen, OMspell). Cette complexité n'impacte que peu la composition des clusters et ne semble donc pas nécessaire en pratique pour ce type de séquence. Les clustering obtenus avec OMstran et OMslen sont similaires à ceux obtenus avec l'OM classique. Quant à OMspell, elle semble davantage différencier les suivis courts de monothérapie et les suivis contenant des "Sans Rien" des autres séquences. Enfin, OMloc distingue différents ordres d'apparition des états dans ses groupes (patients sous insuline + OAD en fin de suivi ou en début de suivi, trithérapie précédée d'une monothérapie, trithérapie précédée d'une bithérapie...). A noter que la sensibilité des méthodes aux paramètres n'a pas été testée. Le choix de ces derniers s'est fait à partir de ce qui a été observé dans la littérature [2] [6], et des valeurs par défaut dans R.

3.3 Distances basées sur le compte des attributs communs entre les séquences

3.3.1 Séquences de même longueur - Distances de Hamming

Lorsque les séquences sont toutes de même longueur, les distances de Hamming et Hamming Dynamique peuvent être utilisées (cf. Figures A.20, A.21 et A.22 en Annexe). Il est intéressant de noter que pour 10 groupes quelque soit la distance choisie, les patients ayant un suivi comprenant un schéma thérapeutique "Insuline + OAD" sont distingués en deux sous-groupes. Le premier contient les patients commençant presque directement en thérapie "Insuline + OAD" tandis que dans le second les patients expérimentent d'autres thérapies (mono, bi et trithérapie et parfois insuline seule) avant d'intégrer l'insuline dans leur traitement. Ce sont deux groupes toutefois "rares" contenant peu de patients. L'OM donne des résultats visuellement similaires à Hamming et DHD.

Globalement, en comparaison avec les résultats de l'OM sur des séquences dont la longueur des suivis varie, la classification est robuste à l'intégration de séquences de longueurs différentes. La seule différence notable est que, pour un clustering à 10 groupes, un troisième sous-groupe de bithérapie remplace le sous-groupe de patients sous insuline avec un suivi court.

3.3.2 NMS

En calculant les distances à partir du nombre de sous-séquences correspondantes (NMS), ou à partir de sa généralisation vectorielle (SVRspell), et en appliquant une CAH avec la distance du saut moyen - plutôt que la distance de Ward qui est plus adaptée aux distances de type euclidienne - on s'éloigne des résultats attendus de clustering. Les résultats de ces méthodes sont disponibles en Annexe A.12.5. Ces méthodes semblent très sensibles au séquençage. NMS semble repérer les séquences similaires dans l'ordre d'apparition des états mais dès lors que les séquences diffèrent beaucoup des autres, elle va leur attribuer une distance plus élevée vis-à-vis des autres, résultant d'un cluster regroupant toutes ces séquences "atypiques". Ce cluster représente plus de 10000 séquences (cf. Figures A.26, A.27 et A.28 (a) des résultats de la CAH en Annexe). Dans les figures A.26, A.27 et A.28, SVRspell est paramétré de façon à accorder plus d'importance aux sous-séquences partagées contenant des épisodes de longue durée et à mettre l'accent sur les sous-séquences partagées de grande longueur : plus la séquence est longue, plus le poids (\sqrt{l} où l est la longueur de la séquence) qui lui est attribué est important. Dans ce cas, les résultats ne semblent pas non plus concluants : presque toutes les classes ne sont composées que d'une seule séquence distincte.

3.4 Distances entre les distributions d'états

Pour les distances basées sur les distributions d'états, le choix d'application des distances a été fait sur 4 types de distances : les distances du CHI2 et euclidienne sur des séquences découpées en 18 périodes (tout le suivi) et les distances du CHI2 et euclidienne sur des séquences découpées en 4 périodes successives pouvant se chauffer. Les tapis de clusters sont présentés en Annexe A.12.6. L'application sur les distances sans utiliser le découpage en K périodes n'a pas donné de résultats satisfaisants visuellement et ne sera pas présentée.

Pour les tapis dans les figures A.32 (a), (b), A.33 (a), (b) et A.34, (a), (b) présentent les distances du CHI2 et euclidienne appliquées sur la distribution des états sur les 18 périodes des séquences (équivaut à $K = 18$ dans la section 2.4). Dans ce cas de figure, les suivis courts ne sont pas distingués. La proportion de temps passée dans les états étant utilisée pour mesurer la distance entre les séquences, l'accent est mis sur la durée passée dans les états, mais aussi sur le timing des états au fur et à mesure que le nombre de groupes augmente. Contrairement à la distance du khi-deux, la distance euclidienne distingue les séquences contenant de longues périodes de "Sans Rien".

Pour les distances calculées sur des intervalles de 4 périodes successives (EUCLID overlap et CHI2 overlap), pouvant se chevaucher (Figures A.32 (c), (d), A.33 (c), (d) et A.33, (c), (c)), la version euclidienne semble davantage distinguer les suivis courts tandis que la distance du khi-deux semble plutôt mettre l'accent sur le séquençage. En effet, la version du khi-deux fait apparaître des classes avec différents ordres d'apparition des états : e.g. pour 10 groupes, trois groupes de trithérapies se distinguent : un où les patients semblent commencer directement en trithérapie, un autre où la transition est plus lente avec un début de suivi en monothérapie puis bithérapie, et un dernier groupe où les patients passent en trithérapie assez rapidement après avoir tester une mono et/ou bithérapie sur une courte période de temps.

3.5 Analyse multi-séquences

Les trajectoires des patients sous traitement anti-diabétique peuvent être séparées en huit séquences différentes, i.e. une séquence ou dimension par traitement : metformine, sulfamide (ou glinide), AGLP1, iDPP4, SGLT2, insuline à action rapide, autre type d'insuline et autre molécule (inhibiteur alpha-glucosidase). Par exemple, si un individu est traité sous metformine tout le long de son suivi, il aura une dimension contenant une trajectoire de 18 périodes de metformine, et ses 7 autres dimensions seront "vides" (mais prises en compte par l'algorithme). Les séquences considérées ont été extraites dans une étape de la construction des séquences précédant la gestion des cas particuliers (switch, "Sans Rien" sur de courtes périodes...), afin de réaliser un calcul de distances et un clustering sur des séquences peu transformées.

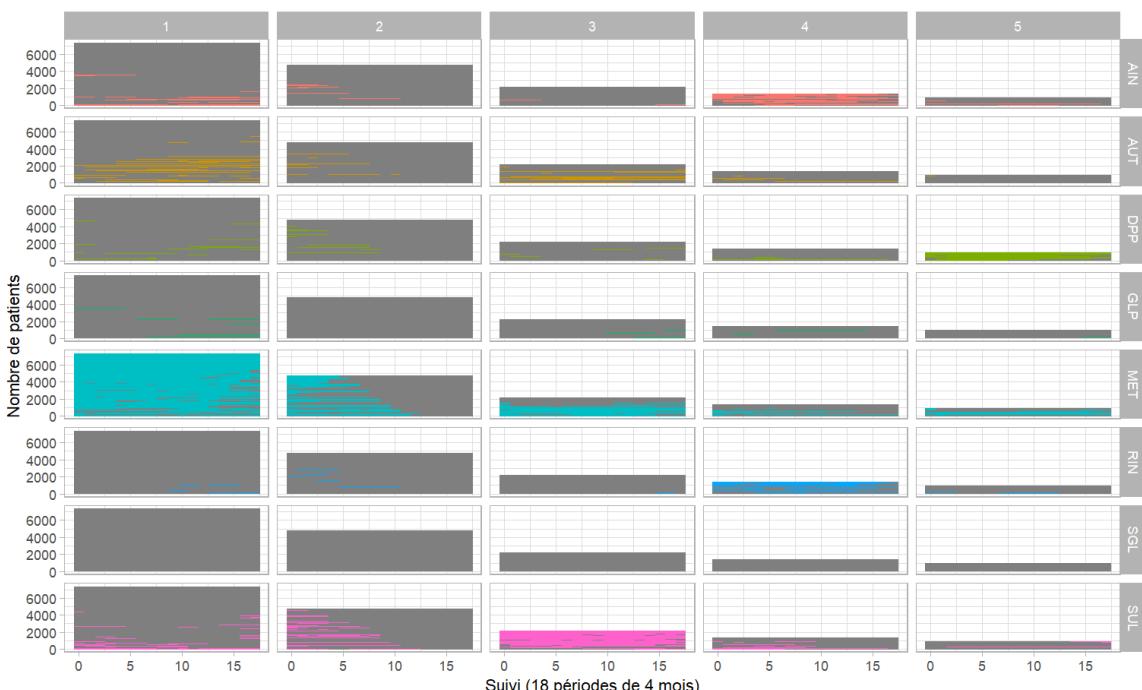


FIGURE 3.5 – MCSA index plot empilé (CAH) - 5 groupes

L'OM classique avec la méthode TRATE réalisée sur une MCSA donne les résultats visibles en figure 3.5 pour 5 groupes. La visualisation est obtenue en empilant les tapis des 8 dimensions sur un graphique. Chaque colonne du graphe représente un groupe, numéroté de 1 à 6, et chaque ligne une dimension. La taille des tapis est proportionnelle au nombre de séquences présentes dans la classe. Les valeurs manquantes sont représentées par la couleur grise : si une séquence est colorée cela signifie que la molécule est (plus ou moins) présente dans la classe où la couleur apparaît. La plus grande classe (groupe 1), comprend près de 8000 patients, et correspond à un cluster contenant principalement des monothérapies sous metformine. La seconde classe contient les suivis courts. La troisième semble contenir les bithérapies de sulfamide et metformine. Le groupe 5 contient la majorité des patients sous IDPP4 dans leur suivi tandis que le groupe 4 contient les patients sous insuline.

Les tapis de séquences des patients, sans distinction des dimensions sont également visibles dans la Figure A.38 en Annexe. Les observations faites avec l'index plot empilé ne sont pas retrouvées avec cette visualisation, hormis pour le groupe 4 contenant les patients sous insuline.

A noter que, les temps de calcul étaient légèrement plus longs avec la MCSA par rapport à l'OM. La visualisation et l'interprétation des résultats est plus compliquée à mettre en place au vue du cadre multidimensionnelle de l'analyse. Au final, cette méthode ne semble pas convenir dans le cadre de séquences de traitement. Cette approche pourrait être pertinente lorsque d'autres types de séquences sont à disposition, comme la chronologie des rendez-vous médicaux du patient, ou son parcours dans un hôpital et ses examens reçus.

3.6 Modèles de mélange

3.6.1 Application de la méthode et limites

La variable d'intérêt de l'étude est ordinaire et contient 6 niveaux (monothérapie, bithérapie, trithérapie, insuline, insuline + OAD, sans rien). Ce type de variable nécessite l'utilisation de la fonction de lien *thresholds* qui témoigne d'une plus grande complexité numérique que les autres fonctions de lien. Afin de limiter le temps de calcul, la classification des patients a été faite en réduisant le nombre de périodes à 6 périodes d'une durée d'un an, et en ne sélectionnant qu'un échantillon de 1000 patients. La méthode grid search [13] a également été testée pour permettre l'estimation d'un modèle à partir d'une grille de valeurs initiales aléatoires afin de réduire les chances d'une convergence vers un maximum local. Néanmoins, malgré ces précautions, les calculs effectués avec différentes définitions des paramètres et du nombre de groupes ont pris énormément de temps (jusqu'à 15 heures) sans pour autant converger à chaque tentative. Au total, 5 tentatives de clustering ont été effectuées :

1. 4 groupes (a convergé)
2. 2 groupes avec une méthode grid search (a convergé)
3. 4 groupes avec une méthode grid search (n'a pas convergé)
4. 7 groupes (n'a pas convergé)
5. 7 groupes avec une méthode grid search (n'a pas convergé)

Pour chacune de ces méthodes, une limite de 100 itérations maximum a été définie. Dans le cadre du grid search, le départ est défini pour 50 valeurs initiales aléatoires différentes et son initialisation s'est faite à partir d'un modèle comprenant un seul groupe.

3.6.2 Résultats d'un clustering à 4 groupes

Pour classer les patients, le modèle effectue la classification à partir des probabilités a posteriori qu'il aura calculé pour chaque patient. Il effectue la classification du patient en sélectionnant le groupe pour lequel la probabilité d'appartenance est la plus élevée. La première tentative de clustering à 4 groupes a produit des résultats intéressants. Le modèle linéaire mixte obtenu est visible en annexe A.12.10.

La figure A.43 en Annexe présente les 4 clusters sous forme de chronogramme. Pour chacun des groupes obtenus les périodes du suivi (de 0 à 5) sont visibles en abscisse et en ordonnée les proportions de patients en fonction du type de thérapie suivi dans chacune des périodes. La figure A.42 en Annexe présente les clusters sous forme de tapis de séquences par groupe. Une ligne du tapis représente la trajectoire d'un seul patient.

Les effectifs et proportions de patients par cluster sont visibles dans la table A.20 de l'Annexe A.12.10, e.g. la 2ème classe comprend 79 % des patients a posteriori. La table A.21 indique que les patients ont été classés dans la classe 1 avec une probabilité a posteriori moyenne de 94 %. Autrement dit, la probabilité a posteriori

d'un patient, qui a été classé dans la classe 1, d'appartenir à cette même classe est de 0.941. Ces probabilités sont toutes supérieures à 0.85 pour chacun des groupes. Enfin, la table A.22 de l'Annexe A.12.10 donne des informations sur les proportions de patients par classe en fonction de leur probabilité a posteriori d'appartenir à cette même classe. Ainsi, 100 % des patients ont été classés dans la classe 1 avec une probabilité a posteriori supérieure à 0.7, tandis que "seulement" 74 % des patients ont été classés dans la classe 3 avec une probabilité a posteriori supérieure à 0.9.

Au final, l'utilisation de cette méthode n'est pas adaptable dans le cadre d'une variable d'intérêt ordinaire et ne peut donc être appliquée à l'analyse de séquences, du fait du temps de calcul et de la difficulté d'atteindre la convergence malgré des données réduites en volume.

3.7 Bag-of-Words

Pour cette dernière méthode, les groupes obtenus semblent mettre l'accent sur la durée passée dans les états, et distinguent les suivis courts. Au fur et à mesure que le nombre de groupes augmente, les parcours expérimentant en grande majorité un seul état dans leur suivi se démarquent, mais les suivis courts, peu importe leur contenu, ou les parcours légèrement différents des autres finissent par être classés dans un groupe de séquences "atypiques" (cf. Annexe A.12.7). Contrairement aux distances entre distributions d'états (Figures A.34 (a) et (b)), les suivis courts de monothérapie par exemple ne sont pas classés avec les suivis longs de monothérapie, et sont placés dans un cluster à part.

3.8 Recommandations dans le choix des méthodes

Il est intéressant de noter que le clustering obtenu avec la distance NMS est celui dont l'indicateur de qualité ASW (> 0.8) est le plus élevé (cf. Tables A.13 et A.14), alors que cette méthode s'est révélée être la moins satisfaisante de toutes. Cela s'explique par la composition des clusters : ils sont tous très homogènes et contiennent peu de séquences, sauf un. Par conséquent, il est important de souligner que la valeur des indicateurs de qualité du clustering est subjective. La qualité de la classification devrait avant tout être évaluée par une représentation graphique des groupes en tapis de séquences ainsi que par la vérification de leur composition exacte : si l'ensemble des séquences paraît homogène au sein de leur classe, qu'elles contiennent un nombre suffisant et cohérent d'individus, et que les classes semblent différentes les unes des autres, alors le clustering pourra être défini comme "bon".

Les indicateurs de qualité du clustering de la PAM sont (presque) tous inférieurs à ceux de la CAH pour chacun des clustering (cf. Tables A.13 et A.14), reflétant la préférence visuelle accordée à la CAH dans la présentation des résultats. Les temps de calcul des distances des différentes méthodes (pour 4333 séquences distinctes) varient d'une méthode à l'autre, comme illustrés dans les tables A.16 et A.17, mais restent assez faibles (moins de 30 secondes en général), le temps le plus long étant d'environ 90 secondes pour NMS. Les distances de Hamming, euclidienne et du khi-2 sont les clustering nécessitant le moins de temps de calcul.

Le choix de la méthode et, pour l'OM, des coûts utilisés dépend du type de séquence analysé, de la dimension temporelle d'intérêt (timing, séquençage, durée passée dans un épisode) et des attentes dans la composition des classes (accent sur les états rares, la longueur des suivis...).

L'avantage de l'OM réside dans sa simplicité d'application et sa flexibilité dans le choix des coûts. En choisissant des coûts d'indel faibles, la sensibilité à la durée passée dans les états sera d'autant plus forte. Dans le cadre de l'analyse de séquences de traitement anti-diabétique présenté dans ce document, les variantes de l'OM prenant en compte le contexte n'ont pas semblé fournir d'amélioration majeure au niveau du clustering. Ces méthodes pourraient être pertinentes sur d'autres types de parcours, ou avec davantage d'états considérés dans l'analyse.

Des recommandations d'application des méthodes décrites dans ce document, en fonction des objectifs de l'utilisateur et des observations tirées des résultats dans le cadre du diabète, sont présentées sous forme d'arbre de décision dans la Figure 3.6. La sensibilité des méthodes aux trois dimensions temporelles est spécifiée via les lettres T (Timing), D (Durée) et S (Séquençage) à côté du nom des distances.

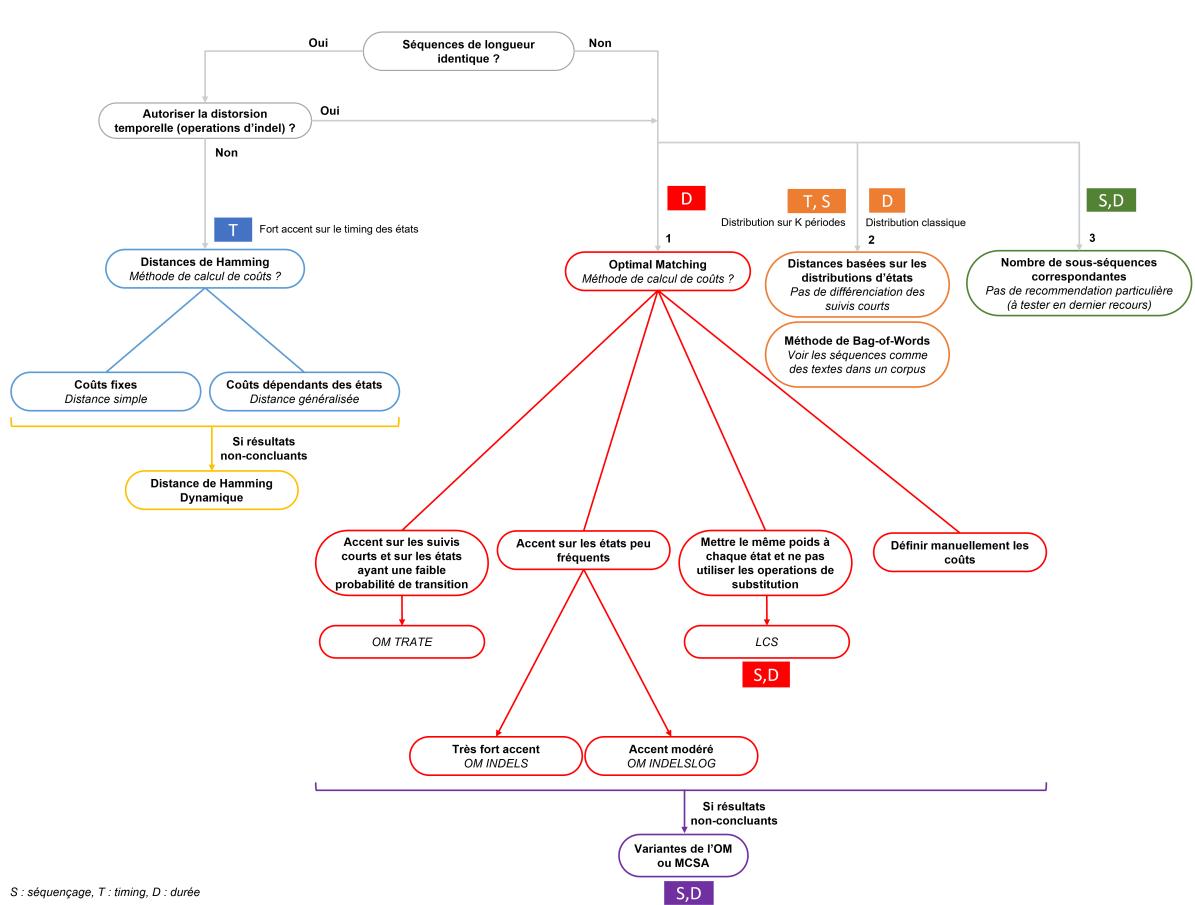


FIGURE 3.6 – Arbre de décision dans le choix des méthodes

Il est indispensable de tester plusieurs nombres de groupes pour comparer les différentes classifications possibles et choisir la plus pertinente. Plus le nombre d'états considérés est grand, plus il y aura de parcours patients atypiques, et plus il est possible d'obtenir une classe contenant des parcours hétérogènes (un groupe "fourre-tout"). Dans ce cas, il ne faudra pas hésiter à augmenter le nombre de groupes de façon à minimiser la taille de ce groupe de séquences atypiques, et éventuellement le supprimer.

3.9 Interprétation économique des résultats

A l'aide des résultats d'une classification et de la description détaillée des classes obtenues, il est possible de se servir de ces données pour évaluer l'impact économique du diabète par type de parcours, en fonction des différents facteurs associés, sur la société. Dans le cadre de l'exemple du diabète, le coût total du DT2 représentait 8,5 milliards d'euros en 2013, soit environ 5 % des dépenses de santé pour l'année. [14] Avec les données de l'Assurance Maladie et les montants des remboursements des traitements et soins à disposition, il serait envisageable d'évaluer le coût moyen d'un patient en fonction de la classe à laquelle il appartient, mais aussi de mesurer l'évolution des coûts de la maladie, afin de pouvoir contribuer par exemple au ciblage des politiques de prévention qui pourraient être mises en place. En effet, le diabète étant aujourd'hui à la fois un enjeu de santé publique et un enjeu économique, il est, selon Santé Publique France [15], primordial face à cette situation d'améliorer la prévention de la pathologie, par exemple par la promotion d'une alimentation adaptée et d'une activité physique régulière ou par l'accompagnement et l'éducation des patients.

Une illustration de l'étude de l'impact économique du diabète peut être donnée par l'utilisation de la metformine dans le cadre d'une monothérapie. La metformine est le médicament de référence pour le traitement du DT2 et son utilisation est moindre par rapport à ce qui serait attendu au niveau des recommandations données par la HAS. La CNAM souligne en effet que la sous-utilisation de 2008 à 2017 de la metformine a un impact économique conséquent [16]. Elle a ainsi quantifier l'impact économique qu'aurait un meilleur suivi des recommandations concernant la metformine par une simulation de la hausse de son utilisation sur les 2,4 millions de personnes diabétiques affiliées au régime général de l'Assurance Maladie, et ayant eu au moins une

délivrance de traitement AD fin 2015. En 2015, parmi l'ensemble de ces patients, 71 % des monothérapies [16] étaient à base de metformine - un chiffre proche de ce qui a été obtenu dans à la description des clusters visible en Annexe, Figure A.41 (c), dont la différence réside dans le fait que la classe représentée décrit la proportion de metformine du cluster comprenant le plus de monothérapie et non la proportion de metformine parmi toutes les monothérapies de l'ensemble des séquences (la classe ne concerne par ailleurs que les patients naïfs de traitement).

Dans l'étude de la CNAM, le taux de recours à la metformine en monothérapie a été simulé à hauteur de 80 %, en conservant 20 % des monothérapies hors metformine pour prendre en compte des contre-indications éventuelles. Ci-dessous, la Table 3.3 résume les coûts observés d'une monothérapie en fonction de l'utilisation ou non de la metformine en 2015.

Monothérapie	Coût moyen sur 4 mois
Metformine	24 euros
Toutes molécules confondues	31 euros

TABLE 3.3 – Coûts moyen d'une monothérapie (2015)

En simulant un taux de recours à la metformine de 80 %, le coût moyen d'une monothérapie sur quatre mois en 2015, toutes molécules confondues, passeraient à 29 euros. Par conséquent, un meilleur respect des recommandations de la HAS conduirait à une diminution des dépenses remboursées à hauteur de 6,6 millions d'euros [16]. Une telle baisse permettrait de pouvoir attribuer cette somme au financement d'autres conditions de prise en charge. A noter que cette simulation ne tient pas compte du bénéfice attendu de l'utilisation d'autres molécules, notamment en termes de l'équilibre glycémique qui pourrait être amélioré, de la moindre fréquence d'effets secondaires survenus, ou éventuellement de complications.

3.10 Conclusion partielle

Cette partie englobe les principaux résultats obtenus à partir des méthodes présentées dans la revue de la littérature. Toutes les méthodes testées ont conduit à des résultats plus ou moins satisfaisants, à l'exception de la distance basée sur le nombre de sous-séquences correspondantes et des modèles mixtes à classes latentes. L'OM classique a bien fonctionné pour mesurer une dissimilarité entre les séquences permettant un regroupement cohérent avec la CAH. La méthode semble être sensible à la durée des états dans les séquences, en particulier lorsque les coûts d'indel sont faibles. La complexité de ses variantes ne semble pas nécessaire pour obtenir une classification plus pertinente. Les distances de Hamming, pour les séquences de même longueur, et les distances basées sur les distributions d'états ont également donné des résultats satisfaisants et semblent sensibles à d'autres aspects temporels des séquences. Les distances de Hamming semblent sensibles au timing, tandis que les distances euclidiennes et du Chi-2 basées sur la distribution des états, mesurée sur l'ensemble des 18 périodes des séquences ; semblent mettre un accent encore plus fort sur la durée passée dans les états, et sur le timing. Par ailleurs, la vision textuelle apportée par le Bag-of-Words s'est révélée être une bonne alternative à l'analyse de séquences. Le choix du nombre de groupes et de la méthode dépend principalement des classes attendues et de la sensibilité souhaitée aux différentes dimensions temporelles des séquences. C'est pourquoi il est recommandé de tester plusieurs nombres de groupes avant de sélectionner le clustering final. Ces méthodes sont avant tout exploratoires et nécessitent une connaissance a priori de la pathologie étudiée, ou une interaction avec des experts en épidémiologie, pour pouvoir interpréter les résultats.

Conclusion

La problématique du stage visait à la recherche et à la comparaison de méthodes permettant le regroupement de séquences de traitement. Ces méthodes ont ensuite été décrites et comparées pour l'équipe Biométrie à l'aide de la rédaction d'un guide pratique d'application des méthodes, sur le langage de programmation R.

Plusieurs étapes ont été conduites pour répondre à la problématique de stage, dont la construction d'une base de données sur SAS, à partir de délivrances de traitements en pharmacie. Une revue de la littérature scientifique a ensuite été effectuée avant l'application des méthodes sélectionnées sur le logiciel statistique RStudio.

L'étude de l'hétérogénéité des parcours de soins peut être réalisée grâce à l'analyse de séquences. Différentes méthodes, dont l'Optimal Matching, peuvent être utilisées pour calculer les distances entre des individus suivant des trajectoires plus ou moins similaires. Ces distances vont permettre de regrouper les patients dont les parcours se ressemblent, via des algorithmes de clustering (CAH, PAM). Le choix de la méthode et des coûts utilisés dépend du type de séquences analysé, de la dimension temporelle d'intérêt (timing, séquençage, durée passée dans un épisode) et des attentes dans la composition des classes (accent sur les états rares, la longueur des suivis...). Une fois le clustering réalisé, il est possible de compléter la description des classes obtenues en s'intéressant à d'autres variables caractérisant les patients et n'ayant pas été utilisées dans le calcul des distances (e.g. l'âge, le sexe, l'IMC...). D'autres méthodes, comme les modèles mixtes à classes latentes ou le concept de Bag-of-Words peuvent être utilisées dans le cadre de l'analyse de données longitudinales. Les modèles mixtes sont peu adaptées à ce type de données et demandent beaucoup de temps de calcul, tandis que la vision textuelle apportée par le Bag-of-Words donne une approche alternative concluante de l'analyse de séquences.

Par la suite, des pistes d'amélioration des méthodes pourraient être explorées :

- En testant d'autres méthodes de coûts dans l'Optimal Matching qui ne sont pas calculées automatiquement par R mais définies par l'utilisateur
- En testant la sensibilité des paramètres de certaines méthodes (pour les variantes de l'OM notamment).
- En faisant une analyse de sensibilité du pas de temps choisi pour la conception des séquences.
- En définissant une métrique ou une méthodologie permettant de faire le compromis dans le choix du nombre d'états pris en compte dans l'analyse de séquences.
- En gérant les groupes "fourre-tout" contenant les séquences atypiques issus de certaines méthodes, avec une augmentation du nombre de classes ou avec une différente méthode de classification.
- En automatisant la description des groupes obtenus sur R pour faciliter l'interprétation.
- Si suffisamment de variables sont disponibles pour les patients, un algorithme de machine learning supervisé pourrait utilisé pour déterminer quelles variables sont les plus discriminantes dans la classification.
- En étendant l'analyse à une analyse multi-séquences (MCSA) avec d'autres types de séquences que les traitements.

Au final, ce stage de recherche a abouti à la rédaction d'un guide technique à destination de l'équipe encadrante. Il sera utilisé pour aider dans la prise en main de ces méthodes dans des projets client, en particulier pour des études en épidémiologie analysant des trajectoires de soins, et pour des études médico-économiques. L'identification des parcours-types dans une étude permet une analyse plus ciblée des patients. Elle améliore la compréhension des différents profils afin d'adapter au mieux les recommandations sanitaires, ou, dans un cadre économique, pour la mise en place d'un plan d'action adapté. Le clustering réalisé dans le cadre du stage s'est appuyé sur une base de données regroupant des traitements délivrés en pharmacie incluant peu d'informations sur le patient, compliquant la description des groupes identifiés. L'application de ces méthodes sur la base de données SNDS contenant tous les remboursements de dépenses en santé, et, dans certains cas, l'appariement avec la base de données Electronic Medical Records (EMR) d'IQVIA comprenant les diagnostics de médecins, pourra être très utile pour une analyse plus poussée des classes identifiées dans le futur.

Bibliographie

- [1] HUG Dr O. Braillard, Service de médecine de premier recours. Prise en charge thérapeutique du diabète de type 2. 2017.
- [2] M. Studer and G. Ritschard. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 179(2) :481–511, 2016.
- [3] D. Delaunay. Apprentissage non-supervisé. In *Introduction au Machine Learning*. Master 2 MAS Université de Rennes 1, 2021.
- [4] Cosson E. Mandereau-Bruno L. et al. Fuentes, S. Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. 2019.
- [5] Fédération Française des Diabétiques. Les chiffres du diabète en France, 2021.
- [6] B. Halpin. Three Narratives of Sequence Analysis. *Advances in Sequence Analysis : Theory, Method, Applications*, 2013.
- [7] B. Halpin. Optimal matching analysis and life-course data : The importance of duration. *Sociological Methods Research*, 38 :365–388, 04 2010.
- [8] N. Robette. Mesurer la dissemblance entre trajectoires. In *L'analyse statistique des trajectoires : Typologies de séquences et autres approches*. INED, 2021.
- [9] J. Roux. *Parcours de soins des patients atteints de sclérose en plaques à partir des données médico-administratives en France*. Theses, Université Rennes 1, November 2018.
- [10] CRAN. How to estimate a latent process mixed model using lcmm function.
- [11] C. Proust-Lima, V. Philipps, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *arXiv : Computation*, 2015.
- [12] J. Brownlee. A Gentle Introduction to the Bag-of-Words Model, 2017.
- [13] RDocumentation. lcmm package - gridsearch: Automatic grid search.
- [14] B. Charbonnel, D. Simon, J. Dallongeville, I. Bureau, J. Gourmelen, and B. Detournay. Coût du diabète de type 2 en france : une analyse des données de l'egb. *Médecine des Maladies Métaboliques*, 11 :IIS24–IIS27, 2017. Approches médico-économiques de la prise en charge du diabète de type 2 : quelles alternatives ?
- [15] Ricci P., Chantry M., Detournay B., Poutignat N., Kusnik Joinville O., Raimond V., Thammavong N., and Weill A. Coûts des soins remboursés par l'Assurance maladie aux personnes traitées pour diabète : Études Entred 2001 et 2007, 2009.
- [16] CNAM Et Pr Pierre F. Vice-Président SFD. Diabète de type 2 : Comment évolue le recours aux anti-diabétiques les huit premières années qui suivent l'instauration d'un traitement anti-diabétique ? 2018.
- [17] DSREB. Metformine : Effets secondaires. 2022.
- [18] Bibliographie du projet ROCHE DIABETE CARE. La place des différents traitements médicamenteux dans les stratégies thérapeutiques du diabète de type 2. 2019.
- [19] N. Singh Chauhan. DBSCAN Clustering Algorithm in Machine Learning, 2022.
- [20] M. Studer. Weightedcluster library manual: a practical guide to creating typologies of trajectories in the social sciences with R. *Sociological Methods Research*, pages 1–32, 2013.
- [21] Wikipedia. Needleman–Wunsch algorithm, 2013.
- [22] V. Likic. The Needleman-Wunsch algorithm for sequence alignment, 2020.
- [23] A. Gabadinho, G. Ritschard, N. Müller, R. Burgin, P. A. Fonta, and G. Ritschard. Package TraMineR, 2022.

Annexe

A.1 Flowchart

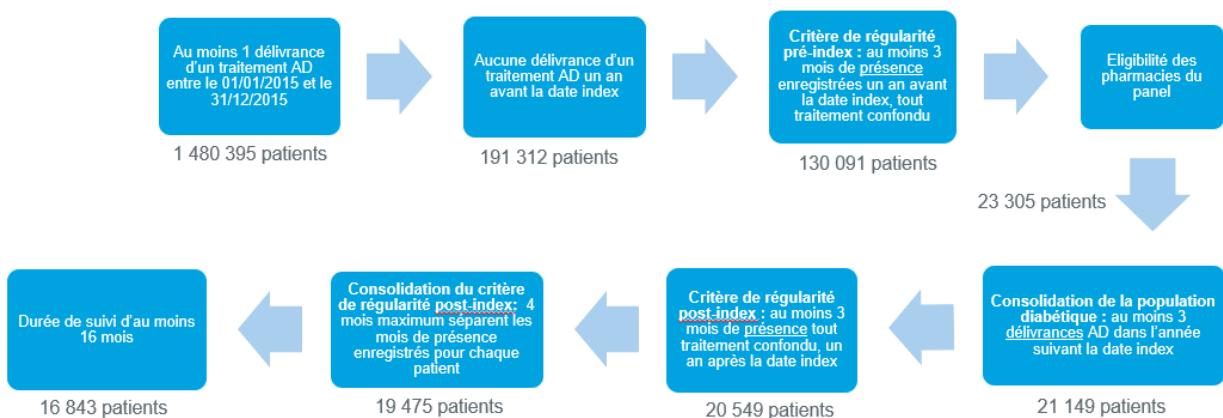


FIGURE A.1 – Flowchart de la construction de la population

A.2 Traitements du diabète

A.2.1 Les familles de médicaments

Les médicaments utilisés pour le traitement du diabète de type 2 n'ont pas tous les mêmes effets sur le diabétique. Nous retrouvons des médicaments qui (1) améliorent la sensibilité à l'insuline, (2) stimulent la production d'insuline, (3) réduisent l'absorption des sucres, (4) agissent par le biais des incrétines et enfin (5) des médicaments qui favorisent l'élimination des sucres. De plus, des injections d'insuline (6) peuvent être prescrites pour contrôler la glycémie. Ci-dessous, les traitements utilisés selon leurs actions sur le corps :

La metformine (1) :

- Fait partie de la famille des **biguanides**, qui sont des anti-diabétiques oraux **améliorant l'efficacité de l'insuline**, en particulier au niveau des muscles et du foie (qui constituent les réserves de sucre).
- Effets indésirables courants : nausée, douleur à l'estomac, vomissement, diarrhée, appétit réduit [17]
- Médicaments : GLUCOPHAGE® (médicament référent) et génériques, STAGID®.
- Utilisation ancienne, efficacité et sécurité bien connues.
- L'insuffisance rénale doit être un facteur de précaution lors du recours à la metformine, mais son indication est aujourd'hui bien définie. [18]
- 71 % des monothérapies étaient à base de metformine en 2015 [18]

Les sulfamides hypoglycémiants et les glinides (2) :

- Deux familles d'anti-diabétiques agissant en **stimulant la libération d'une plus grande quantité de l'insuline** par le pancréas.
- Effet indésirable courants : hypoglycémie, en particulier chez les patients âgés et/ou souffrant d'insuffisance rénale.

- Les **sulfamides hypoglycémiants** sont proposés le plus souvent en association avec la metformine, lorsqu'un seul médicament ne suffit pas pour équilibrer le diabète ; médicaments : AMAREL®, DAONIL®, DIAMICRON® et génériques, OZIDIA®.
- Les **glinides** ont une durée d'action plus courte que celles des sulfamides hypoglycémiants (ils doivent être pris immédiatement avant le repas) ; médicaments : NOVONORM® et génériques.

L'acarbose (3) :

- Appartient à la famille des **inhibiteurs des alpha-glucosidases**, médicaments qui réduisent l'absorption des sucres en retardant la digestion ainsi qu'en ralentissant le passage des sucres dans le sang après les repas.
- Effets indésirables courants : flatulence, douleur abdominale, diarrées (digestifs).
- Médicaments : GLUCOR® et génériques.

Les gliptines et les analogues de la glucagon-like peptide (GLP-1) (4) :

- Deux familles d'antidiabétique agissant par le biais des **incrétines**, qui sont deux hormones (GLP-1 et GIP) intestinales contrôlant la sécrétion d'insuline par le pancréas après les repas.
- Médicaments de la famille des gliptines (ou inhibiteurs de la dipeptidylpeptidase-4 (iDDP4), anti-diabétiques oraux) : GALVUS®, JANUVIA®, ONGLYZA®, XELEVIA®.
- Médicaments de la famille des AGLP-1 (voie injectable sous-cutanée) : BYDUREON®, BYETTA®, OZEMPIC®, TRULICITY®, VICTOZA®.
- Les gliptines : Elles sont utilisées par voie orale pour contrôler le diabète en cas d'échec des mesures hygiéno-dietétiques, en association avec d'autres anti-diabétiques oraux (de préférence avec la metformine, parfois avec les sulfamides hypoglycémiants) ou avec de l'insuline. Effets secondaires : des **angioédèmes** ont été rapportés chez des patients traités par gliptines. Ils peuvent avoir des conséquences graves.
- AGLP-1 : Ils sont utilisés en association avec un anti-diabétique oral (metformine ou sulfamide hypoglycémiant, par exemple) ou avec l'insuline lorsque ces traitements n'ont pas été suffisamment efficaces pour contrôler la glycémie.

La dapagliflozine et l'empagliflozine (5) :

- Font partie de la famille la plus récente d'anti-diabétiques oraux, qui est celle des **gliflozines** ou inhibiteurs du co-transporteur sodium glucose de type 2 (SGLT2), ils favorisent l'élimination du glucose dans l'urine, ce qui réduit le taux de sucre dans le sang.
- Effets indésirables fréquents : augmentation des infections des voies urinaires, hypoglycémie (liés au mécanisme d'action du médicament).
- Médicament avec pour p dapagliflozine : FORXIGA®.
- Médicament de type empagliflozine : JARDIANCE®.

Les injections d'insuline (6) :

- Utilisé lorsque les traitements oraux ne sont pas suffisamment efficaces pour contrôler le taux de sucre dans le sang, le médecin peut prescrire des injections d'insuline (voir liste des insulines).
- En général, le passage à l'insuline commence en associant des anti-diabétiques oraux à une injection quotidienne d'insuline retard (à action lente) avant le coucher. La dose d'insuline est progressivement augmentée jusqu'à ce que la glycémie à jeun (au lever) soit inférieure à 1,10 g/l.
- Si l'association d'anti-diabétiques oraux et d'insuline ne suffit pas à atteindre l'objectif glycémique, alors le patient peut passer à une insulinothérapie seule. Les injections peuvent avoir lieu tout d'abord au moment du coucher, puis avant ou après un ou plusieurs repas. Plusieurs combinaisons d'insuline(s) peuvent être utilisées (lente, intermédiaire, rapide, ultra-rapide).
- Le passage à l'insuline est également nécessaire lorsqu'une femme souffrant de diabète de type 2 est enceinte. Il y a en effet un risque malformatif de la metformine, qui est contre indiquée pendant la grossesse.

A.2.2 Schémas thérapeutiques considérés

Les schémas thérapeutiques généralement considérés pour le traitement du diabète sont listés ci-dessous.

- Monothérapie (hors insuline) : metformine (sauf si insuffisance rénale)
- Bithérapie : metformine combinée avec un autre médicament (sulfamides hypoglycémiants, SGLT2, glipfine, AGLP1 ou insuline basale)
- Trithérapie ou plus : metformine combinée avec deux (ou plus) autres médicaments (cf. Table A.1)
- Insuline seule
- Pas de schéma thérapeutique

Metformine +				
Sulfamides hypo-glycémiants +	SGLT2 +	Gliptine +	AGLP1 +	Insuline basale
Gliptine SGLT2 AGLP1 Insuline basale	Sulfamides Gliptine Insuline basale	Sulfamides SGLT2 Insuline basale	Sulfamides Insuline basale	Sulfamides Gliptine SGLT2 AGLP1

TABLE A.1 – Combinaisons possibles en trithérapie [1]

A.3 Algorithme de la Haute Autorité de Santé

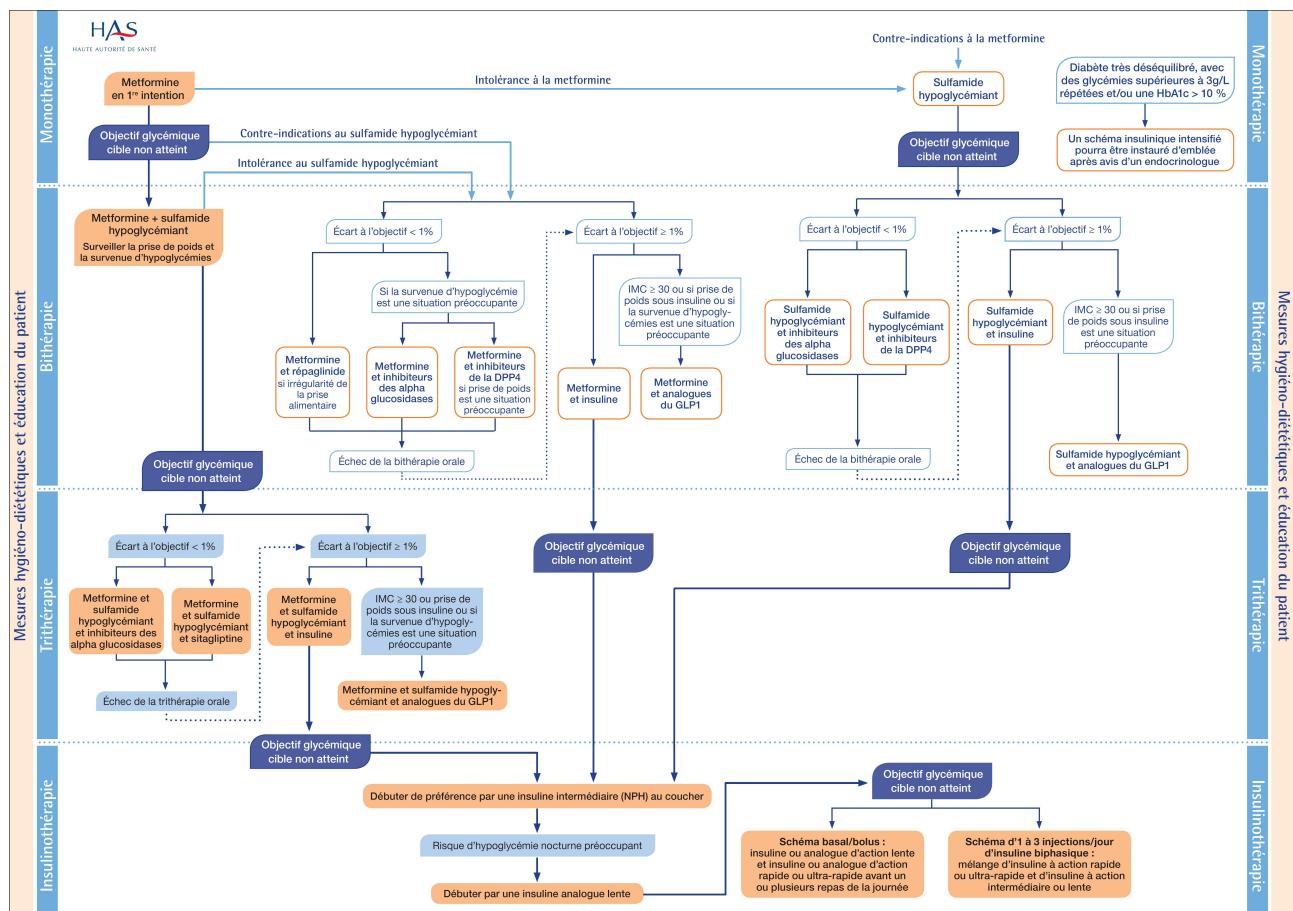


FIGURE A.2 – Algorithme retracant les schémas thérapeutiques possibles du patient diabétique, adaptés en fonction de la cible HbA1c* (source : HAS)

*HbA1c : test d'hémoglobine glyquée, mesure de la forme glyquée de l'hémoglobine pour obtenir la moyenne de la glycémie sur trois mois

A.4 Classification non-supervisée

A.4.1 Classification hiérarchique

Deux méthodes possibles : l'algorithme ascendante (agglomératif) et descendant (divisif). Le premier construit les classes par aggrégation successives des observations 2 à 2 tandis que le deuxième les construit par dichotomies progressives de l'ensemble des observations. Ces deux algorithmes aboutissent à la construction d'un dendrogramme, un arbre rassemblant des observations de plus en plus dissemblables au fur et à mesure que l'on s'approche de sa racine. La méthode agglomérative reste la plus utilisée, en particulier la CAH, dont les avantages et inconvénients sont cités ci-dessous.

Avantages :

- Permet de choisir le nombre de classes de façon optimale grâce aux indicateurs de qualité de la classification en fonction du nombre de classes
- S'adapte aux différentes formes de classes par le choix de la distance

Inconvénients :

- Complexité algorithmique non-linéaire
- Non adapté aux grosses volumétries (conséquence du 1er point) : quelques milliers de points, pas plus
- 2 observations classées dans des classes différentes ne sont plus jamais comparées

A.4.2 Classification non-hiéarchical

Le principe de la classification non-hiéarchical est de réaliser une partition de l'ensemble des individus décrits dans R^p en K classes, en maximisant la variance inter-classes. Parmi les méthodes de clustering non-hiéarchical, nous retrouvons :

- L'algorithme des centres mobiles (Forgy) : on tire au hasard K centres de classes parmi les points et on affecte chaque point à une classe selon la distance euclidienne, ensuite on actualise le nouveau centre classe et on réaffecte des points, le processus s'arrête lorsque les barycentres ne changent plus ;
- Les K-Means (Mc Queen) : les barycentres ne sont pas recalculés à la fin des affectations mais à la fin de chaque allocation d'un individu à une classe ; l'algorithme est ainsi plus rapide ;
- Les K-Medoids : similaire aux K-Means, sauf qu'une classe ne va plus être définie par une valeur moyenne (le centroïd) mais par un représentant le plus central, le médoïd ; cette méthode cherche à minimiser l'erreur quadratique moyenne (i.e. la distance entre les points de la classe et le point central), lui donnant une plus grande robustesse vis-à-vis des données extrêmes par rapport aux K-Means.

Avantages

- Très rapide
- Simple à programmer
- Convient aux gros jeux de données

Inconvénients

- Algorithme stochastique
- Nombre de classes fixé a priori
- L'utilisation de la distance euclidienne conduit l'algorithme à détecter des formes sphériques dans les données
- Sensible aux valeurs extrêmes : un point éloigné peut former une classe d'un individu

A.4.3 Classification par densité

DBSCAN (density-based spatial clustering of applications with noise) est un algorithme de partitionnement de données fondé sur la densité [19]. Il s'appuie sur la densité estimée des clusters pour effectuer le partitionnement. Dans les implémentations naïves, la classification est réalisée par itération sur **chaque point** du jeu de données en calculant sa distance par rapport à tous les autres points puis en associant chaque point à ses voisins.

Avantages :

- Adapté aux formes non convexes de clusters
- Pas besoin de fixer le nombre de clusters a priori
- Capable de gérer les données aberrantes en les éliminant du processus de partitionnement
- Peut trouver des clusters de taille et de forme arbitraire

Inconvénients :

- Ne fonctionne pas très bien quand les clusters ont des densités trop différentes (paramétrage de ϵ et $minPoints$ difficile)
- Algorithme non déterministe : il peut produire des résultats différents selon les points qui sont visités en premier
- Peu adapté aux données avec des dimensions importantes car le seuil de distance devient difficile à estimer
- Ne peut pas être utilisé sur un échantillon car l'échantillonnage peut modifier les caractéristiques de densité des données

Dans le cadre de l'exemple présenté dans ce document, cette méthode de clustering ne s'est pas révélée concluante. Aucune classe n'est identifiée.

A.5 Vue d'ensemble des mesures de qualité d'un clustering

Sans supervision, il est délicat d'évaluer la qualité de l'algorithme. Il existe plusieurs mesures possibles :

1. La forme des clusters :

- L'homogénéité intra-classe : homogénéité globale $T = \frac{1}{K} \sum_{k=1}^K T_k$ où l'homogénéité T_k d'un cluster k , noté C_k est défini comme la moyenne des distances de chacun des points contenus dans ce cluster à son centroïde u_k : $T_k = \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \mu_k)$
 - L'hétérogénéité inter-classes : critère de séparabilité globale $S = \frac{2}{K(K-1)} \sum_{k=1}^K \sum_{l=k+1}^K S_{k,l}$ se calculant comme la moyenne des séparabilités $S_{k,l}$ des clusters deux à deux (où $S_{k,l}$ est le critère de séparabilité entre deux clusters k et l quantifiant à quel point les 2 clusters sont distants les uns des autres)
 - Le coefficient de silhouette d'un point x : compris entre -1 et 1, plus il est proche de 1, plus l'assignation de x à son cluster est satisfaisante
2. La stabilité des clusters : est-ce que les clusters restent les mêmes lorsque l'on fait tourner plusieurs fois le code ?
3. La connaissance des experts : l'algorithme est souvent évalué "à l'oeil", en regardant si les clusters proposés ont du sens

Le troisième point est le principal point de comparaison des méthodes car il permet d'évaluer la cohérence avec les connaissances en épidémiologie sur la maladie du diabète.

A.6 Indicateurs de qualité du clustering dans R

La librairie *WeightedCluster* offre des procédures pour faciliter le choix d'une solution de clustering particulière et définir le nombre de groupes. En particulier, *wcClusterQuality* est une fonction de ce package qui donne un certain nombre de statistiques sur la qualité d'un clustering, résumant le 1er point de l'annexe A.5 :

- Point Biserial Correlation (**PBC**) : mesure de la capacité du clustering à reproduire les distances, valeur oscillant dans l'intervalle $[-1, 1]$ (**une bonne partition maximise cette mesure**)
- Hubert's Gamma ((**HG**) : Mesure de la capacité du clustering à reproduire les distances (intervalle $[-1, 1]$, **maximisation**)
- Hubert's Somers D (**HGSD**) : Mesure de la capacité du clustering à reproduire les distances avec prise en compte des égalités sur les distances (intervalle $[-1, 1]$, **maximisation**)
- Hubert's C (**HC**) : Ecart entre la partition obtenue et la meilleure partition qu'il serait théoriquement possible d'obtenir avec ce nombre de groupes et ces distances
- Average Silhouette Width (**ASW**) : Cohérence des assignations (lorsqu'elle est élevée cela indique des distances inter-groupes élevées et une forme homogénéité intragroupe)
- Average Silhouette Width (weighted) (**ADWw**) : ASW pondéré
- Calinski-Harabasz index (**CH**) : Pseudo F calculé à partir des distances
- Calinski-Harabasz index (distances au carré) (**CHsq**) : Pseudo F calculé à partir des distances au carré
- Pseudo R^2 (**R2**) : Part de la dispersion expliquée par la solution de clustering (comparaison avec des partitions avec un nombre de groupes identiques)
- Pseudo R^2 (distances au carré) ((**R2sq**) : R2 mais avec des distances au carré

Parmi ces indicateurs, ASW semble être l'indicateur le plus intéressant pour déterminer la qualité d'un clustering [20]. L'interprétation des valeurs pour une partition obtenue peut être donnée de la façon suivante :

- < 0.25 : Aucune structure identifiée dans le clustering.
- $0.26 - 0.50$: La structure est faible et pourrait être artificielle.
- $0.51 - 0.70$: Structure raisonnable identifiée.
- > 0.71 : Structure forte identifiée.

A.7 Variantes de l'OM

A.7.1 OM localisé (OMloc)

L'OM localisé fait dépendre les coûts d'indel des états adjacents. La motivation d'une telle méthode vient du fait qu'insérer ou supprimer un état qui est similaire à ses voisins ne changerait que la longueur de l'épisode concerné sans affecter l'ordre d'apparition des états.

Dans la séquence aab, insérer b tel que la séquence devient aabb ne modifierait que la durée de l'épisode b et non l'ordre des états. Cela implique que les insertions rallongeant un épisode coûteront moins chères que des insertions modifiant l'ordre d'apparition des états. Dans sa définition des coûts, OMloc considère les opérations en termes d'insertion [6], la définition étant facilement transposable à la délétion. Le coût d'indel d'insérer l'élément c entre les éléments a et b d'une séquence est calculé comme suit [2] :

$$c_I(c|a, b) = \alpha \frac{\gamma_{a,c} + \gamma_{b,c}}{2} + \beta \gamma_{max}$$

Dans le cas où l'insertion de l'élément se fait à la fin d'une séquence :

$$c_I(c|a) = \alpha \gamma_{a,c} + \beta \gamma_{max}$$

$\gamma_{i,j}$ indique le coût de substitution (autrement dit la distance entre les deux états) et γ_{max} est le coût de substitution maximal. Les paramètres α et β sont fixés par l'utilisateur. Le premier paramètre est un coût fixe pour l'insertion tandis que le second pondère la distance entre l'élément inséré et ses voisins. α contrôle la pénalisation des différences avec les états environnants mesurées par les coûts de substitution, tandis que β pénalise la distorsion temporelle. Tant que $1 - 2\beta \leq \alpha$, la méthode empêche également l'OM d'utiliser une paire d'indels au lieu d'une substitution, et fournit ainsi un moyen de permettre d'importantes distorsions du temps tout en préservant l'efficacité des coûts de substitution.

Cette variante de l'optimal matching est une dissimilarité ne garantissant pas l'inégalité triangulaire. En effet, les coûts d'indel pour chaque état à chaque position sont calculés au début et ne changent pas pendant le processus d'alignement, i.e. même quand les états environnants sont changés, alors que cela aurait plus de sens d'adapter le coût d'indel après chaque opération. En considérant les séquences S_1 et S_3 dans la table A.2, une fois que a est substitué pour b à la première position dans S_3 (S_3 devient abbb) pour un coût égal à 1, insérer un second a en face du a substitué (pour donner la séquence aabbb) devrait coûter 0.1 ($\gamma(a, a)$ étant égal à 0) et supprimer un des derniers b devrait également coûter 0.1. Opérer de cette façon donne un coût total égal à $1 + 0.1 + 0.1 = 1.2$ (au lieu de 2 avec deux substitutions) et respecterait ainsi l'inégalité triangulaire.

	aabb	abbb	bbbb
S_1	aabb	0	
S_2	abbb	0.2	0
S_3	bbbb	2.0	1

TABLE A.2 – Exemple de distance OMloc entre trois séquences en utilisant un coût fixe de substitution de 1, $\alpha = 0.1$ et $\beta = 0.8$ [2]

A.7.2 OM sensible à la durée des épisodes (OMslen)

Dans cette variante de l'OM, les coûts dépendent de la durée des épisodes dans laquelle a lieu l'opération. Autrement dit, lorsqu'il est nécessaire de supprimer, ajouter ou substituer un seul état présent dans un épisode, le coût de cette édition sera d'autant plus faible que l'épisode concerné est grand. Dans le cadre d'une séquence de traitement, cette méthode considère donc qu'il est moins coûteux de modifier un état provenant d'une longue période de traitement, que d'en modifier un provenant d'une très courte période de traitement.

La stratégie la plus simple introduite de modification des coûts est de diviser les coûts par la racine carrée de la longueur de l'épisode. Cette modification des coûts peut se généraliser par l'introduction d'un facteur au coût égal à $1/t^h$ où t est la durée de l'épisode et h (valeur entre 0 et 1) est un exposant appliquant un poids au temps ($h = 1/2$ pour la racine carrée). Soit a_t , un état a dans un épisode de longueur t et le coût d'indel basique c_I , le nouveau coût d'indel $c_I^H(a_t)$ de a_t est calculé comme suit [2] :

$$c_I^H(a_t) = \frac{c_I(a)}{t^h}$$

Le coût de déletion d'un état dans un épisode de longueur t sera d'autant plus faible que t est grand. Par ailleurs, supprimer un état dans un épisode de longueur $t = 1$ n'a aucun impact sur le coût. Ainsi, le coût de supprimer un état b dans la séquence aabbba sera bien plus faible avec OMslen ($\frac{c_I(b)}{4^h}$) qu'avec l'OM classique ($c_I(b)$).

Pour le coût de substitution, il peut être réduit en fonction du plus long des deux épisodes qu'il affecte. Soit $\gamma(a, b)$ le coût de substitution basique entre a et b , le nouveau coût de substitution $\gamma^H(a_{t_1}, b_{t_2})$ de a_{t_1} et b_{t_2} est calculé comme suit [2] :

$$\gamma^H(a_{t_1}, b_{t_2}) = \gamma(a, b) \frac{1}{\max\{t_1^h, t_2^h\}}$$

La moyenne arithmétique ou géométrique de t_1^h et t_2^h peut également être utilisée à la place de la fonction \max .

Comme l'OM localisé, cette variante de l'optimal matching est une dissimilarité ne garantissant pas l'inégalité triangulaire (plus de détails dans l'article [6]).

A.7.3 OM entre des séquences d'épisodes (OMspell)

Ce ne sont plus des séquences d'états qui sont considérées mais des séquences d'épisodes (cf. tableau ?? où la ligne SPELL correspond à ce format). Cette variante permet de substituer ou d'insérer/supprimer un grand nombre d'états identiques en une seule fois à faible coût.

L'idée générale de cette méthode est de considérer pour chaque durée possible d'un épisode, un épisode dans l'état a d'une durée de t unités de temps comme un élément distinct, noté a_t , de l'alphabet. Par exemple, la séquence aabbc deviendrait $a_2b_2c_1b_1$. Cela a pour conséquence d'augmenter considérablement la taille de l'alphabet et le nombre de coûts d'indel et de substitution. De plus, un facteur $\delta \geq 0$ pondérant la durée de l'épisode est inclus dans le coût et permet d'avoir un certain contrôle sur la distorsion du temps.

Soit $c_I^S(a_t)$ le coût d'indel de a_t et $c_I(a)$ le coût d'indel de a :

$$c_I^S(a_t) = c_I(a) + \delta(t - 1)$$

Soit $\gamma^S(a_{t_1}, b_{t_2})$ le coût de substitution des états a_{t_1} et b_{t_2} et $\gamma(a, b)$ le coût de substitution de l'état a vers l'état b (ou inversement), les coûts de substitution sont définis comme suit :

$$\begin{aligned} \gamma^S(a_{t_1}, b_{t_2}) &= \gamma(a, b) + \delta(t_1 + t_2 - 2) \text{ si } a \neq b \\ \gamma^S(a_{t_1}, b_{t_2}) &= \delta|t_1 - t_2| \text{ si } a = b \end{aligned}$$

Le paramètre δ correspond au coût d'expansion ou de compression d'une séquence par une unité de temps, tandis que la substitution entre deux épisodes de durée t correspond alors au coût de compresser chaque épisode en un seul état d'une durée $t = 1$, plus la substitution entre les deux états concernés.

Le coût de supprimer l'épisode bbbb de aabbba (recodé a_2b_4) sera plus faible avec OMspell ($c_I(a) + \delta \times 3$) qu'avec l'OM classique ($4 \times c_I(a)$). L'idée générale de cette méthode est donc de permettre de substituer ou d'insérer/supprimer un grand nombre d'états identiques en une seule fois à faible coût.

Lorsque $t = 1$, les coûts appliqués reviennent à ceux de l'OM classique. Dans le cas où $\delta = 0$, l'OM des épisodes devient la distance d'OM des séquences d'états distinctes, i.e. que les indices t de l'alphabet ne sont plus considérés.

Cette variante d'OM est par construction sensible aux différences dans les durées des épisodes, mais aussi à l'ordre des états, et elle permet un certain contrôle sur la distorsion du temps grâce à la pénalité attribuée par le facteur δ .

A.7.4 OM sur les transitions (OMstran)

Cette méthode recode les séquences sous forme de séquences de transitions avant d'utiliser l'OM sur ces nouvelles séquences. La séquence aabbb deviendrait par exemple aa-ab-bb-bb.

Recoder les séquences sous forme de séquences de transitions augmentant considérablement la taille de l'alphabet (i.e. du nombre d'états considéré). Une matrice des coûts de transition est générée à partir de ces données "augmentées". En considérant la transition $a \rightarrow b$ qui est faite à partir de l'état d'origine a et d'un type de transition (vers un nouvel état ou le même état), les coûts d'indel et de substitution d'une transition sont exprimés comme combinaison linéaire du coût d'indel de l'état d'origine et du coût $c_T(a \rightarrow b)$ associé au type de transition. Avec $c_I(a)$ le coût d'indel de l'état d'origine a et $\gamma(a, c)$ le coût de substitution entre les états d'origine a et c , les coûts d'indel $c_I(a \rightarrow b)$ et de substitution $\gamma(a \rightarrow b, c \rightarrow d)$ sont définis comme suit [2] :

$$c_I(a \rightarrow b) = w c_I(a) + (1 - w) c_T(a \rightarrow b)$$

$$\gamma(a \rightarrow b, c \rightarrow d) = w \gamma(a, c) + (1 - w)(c_T(a \rightarrow b) + c_T(c \rightarrow d))$$

Les coûts dépendent d'un coefficient, $w \in [0, 1]$, contrôlant le compromis entre le coût lié à l'état d'origine et le coût lié au type de transition (changement d'état ou non). L'OM des séquences de transitions est par construction sensible au séquençage.

A.7.5 Nombre de sous-séquences correspondantes (NMS)

La première version de cette méthode consiste en la suppression de tous les états non-communs et répétés avant d'effectuer le compte des sous-séquences partagées dans les deux séquences comparées. Soit $\text{emb}_s(u)$ le nombre de fois qu'une sous-séquence u apparaît dans la séquence s , le nombre de sous-séquences correspondantes, $A_{NMS}(s_1, s_2)$, entre s_1 et s_2 est [2] :

$$A_{NMS}(s_1, s_2) = \sum_{u \in S(s_1, s_2)} \text{emb}_{s_1}(u) \text{emb}_{s_2}(u)$$

où $S(s_1, s_2)$ correspond à l'ensemble des sous-séquences communes distinctes. Par exemple, les séquences $s_1 = A A B B$ et $s_2 = A B B C A A C C B A B A$ seraient tout d'abord ramenées à $s_1 = AB$ et $s_2 = ABABABA$ - les états non-communs et répétés sont supprimés. Les sous-séquences partagées seraient alors A, B et AB. s_1 et s_2 partagent $S(s_1, s_2) = 3$ sous-séquences distinctes et $A_{NMS} = 1 \times 4 + 1 \times 3 + 1 \times 3 = 10$ (A, B et AB apparaissent une fois dans s_1 tandis que A apparaît 4 fois dans s_2 , B apparaît 3 fois dans s_2 et AB apparaît 3 fois dans s_2). La distance entre les deux séquences est ensuite calculée de la façon suivante :

$$d_{NMS}(s_1, s_2) = A_{NMS}(s_1, s_1) + A_{NMS}(s_2, s_2) - 2 \cdot A_{NMS}(s_1, s_2)$$

Une telle méthode ne considérera que l'ordre d'apparition des états et ne prend pas en compte la durée totale des épisodes.

Une généralisation de la méthode permet de mesurer chaque sous-séquence correspondante u par sa longueur l_u ou par une transformation l_u^a de celle-ci, où a est un poids attribué à la longueur de la sous-séquence u . Cette généralisation tient compte de la durée de chaque état présent dans la sous-séquence. Par exemple, dans l'exemple précédent, $s_1 = A A B B$ et $s_2 = A B B C A A C C B A B A$, qui peut être réécrit dans les séquences A^2-B^2 et $A^1-B^2-C^1-A^2-C^2-B^1-A^1-B^1-A^1$, la somme de la durée de la sous-séquence AB est égale à $2 + 2 = 4$ dans la première séquence et à $1 + 2 + 2 + 1 + 1 + 1 = 8$ dans la seconde séquence. Pour $a = 1$, cela donne un poids de $4^{11} = 32$ à la sous-séquence commune AB lors de la comparaison des deux séquences.

La métrique de représentation vectorielle des sous-séquences (SVRspell) étend cette méthode. Elle prend aussi en compte la longueur l des sous-séquences correspondantes via un paramètre appliqué sur la longueur (a). Un autre paramètre, b , est introduit et donne plus d'importance à la durée t des épisodes (t^b) inclus dans les sous-séquences. Lorsque $b = 0$ et $a = 0$, nous retrouvons la distance NMS. SVRspell offre une grande versatilité dans la pondération des features (a), la distorsion temporelle (b), et pour faire face aux séquences de longueurs différentes.

A.8 Bag-of-Words

Le modèle du "sac de mots" est une représentation de simplification utilisée dans le traitement du langage naturel et la recherche d'information. Ce type de modèle est connu pour être très efficace dans la classification de documents. Dans cette approche, le compte des mots dans chaque texte est utilisé comme une variable.

Pour tenir compte des états rares, la statistique TF-IDF (Term Frequency - Inverse Document Frequency) peut être utilisée. Cette mesure permet de réévaluer la fréquence des mots en fonction de leur fréquence d'apparition dans tous les documents, de sorte que les scores des mots fréquents qui sont également fréquents dans tous les documents soient pénalisés. Soit $f_{t,d}$ la fréquence du terme t dans un document d , et soit D un corpus de textes :

$$TF(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$IDF(t, D) = \log \frac{N}{|d \in D : t|}$$

$$TFIDF(t, d, D) = TF(t, d).IDF(t, D)$$

Le TFIDF de chaque terme sera calculé dans tous les textes. Ensuite, la distance pourra être appliquée sur la matrice TFIDF obtenue.

Ainsi, pour l'adapter à l'analyse de séquences, l'occurrence de l'alphabet (des états) doit être déterminée dans chaque séquence. Ci-dessous, un exemple d'application de cette méthode en utilisant la métrique TF-IDF sur deux séquences $S1 = ABBBC$ et $S2 = AABBB$, $D = \{S1, S2\}$.

$$TF(A, S1) = 1/5$$

$$IDF(A, D) = \log(2/2) = 0$$

$$TFIDF(A, S1, D) = 0$$

Les TF-IDF de A et B pour S2 seront aussi égaux à 0.

$$TF(C, S1) = 1/5$$

$$IDF(A, D) = \log(2/1) = 0.3$$

$$TFIDF(A, S1, D) = 0.06$$

Et ainsi de suite.

TABLE A.3 – Tableau récapitulatif (TF-IDF)

Etats	A	B	C
TF-IDF (S1)	0	0	0.06
TF-IDF (S2)	0	0	0

La distance euclidienne entre S1 et S2 est alors égale à $\sqrt{((0 + 0 + 0.06)^2)} = 0.06$. Cette méthode, comme pour les distributions d'états, ne prend pas en compte l'ordre dans laquelle apparaissent les états, et met l'accent sur la durée passée dans chaque état.

A.9 Implémentation dans R

A.9.1 Algorithme derrière l'Optimal Matching

L'algorithme implémenté dans R est celui de Needleman-Wunsch (1970) [21] [22]. L'idée de l'algorithme est de construire le **meilleur alignement** de deux séquences en utilisant les alignements optimaux de plus petites sous-séquences. L'algorithme de Needleman-Wunsch divise essentiellement un grand problème (e.g. la séquence complète) en une série de petits problèmes et utilise les solutions des petits problèmes pour trouver une solution optimale au grand problème.

Prenons l'exemple des séquences ATGGCGT et ATGAGT. Les séquences n'étant pas de même longueur, nous pourrions les aligner de cette façon : ATGGCGT et ATG-AGT, ou de cette façon : ATGGCGT et A-TGAGT. Un **score d'alignement** est calculé pour chacun des deux alignements possibles et permet de déterminer lequel est le "meilleur". Le but de l'algorithme est de trouver tous les alignements possibles ayant le score le plus élevé, ou le plus faible, suivant le **schéma de score** choisi. Le schéma de score est un ensemble de lois qui assigne un score d'alignement à un alignement donné de deux séquences. Un score est attribué à chaque opération possible entre deux états d'une séquence : les **scores de substitutions** (match/mismatch), plus les **pénalités pour les écarts** (indel). Le score d'alignement est la somme des scores de substitutions et des pénalités d'indel.

Un exemple de schéma de score peut être simplement :

- Match (les deux états sont identiques) : 1
- Mismatch (les deux états sont différents, revient à la substitution) : -1
- Indel (opération d'insertion ou de délétion) : -1

Dans ce cas, plus le score d'alignement est bas, plus la distance éditée est grande. Pour un tel système de scoring, un score élevé est souhaité. Dans l'exemple de séquences donné ci-dessus, un score de $+1+1+1+0-1+1+1 = 4$ serait donné pour l'alignement ATGGCGT et ATG-AGT, et un score de $+1+0-1+1-1+1+1 = 2$ pour l'alignement ATGGCGT et A-TGAGT. Le premier alignement est donc le meilleur.

A contrario, un score faible est désiré pour le système de scoring suivant :

- Match (les deux états sont identiques) : 0
- Mismatch (les deux états sont différents, revient à la substitution) : 1
- Indel : 1

Pour réduire le nombre de possibilités à considérer et tout de même garantir de trouver la meilleure solution, l'algorithme de Needleman-Wunsch propose une stratégie de programmation dynamique.

Une matrice de distances $D(i, j)$ indexée par les "états" (ou lettres) de chaque séquence est construite de façon récursive, de sorte que :

$$D(i, j) = \max (\text{ou min}) \begin{cases} D(i - 1, j - 1) + s(i, j) \\ D(i - 1, j) + g \\ D(i, j - 1) + g \end{cases}$$

$s(i, j)$ est le score de substitution (match/mismatch) pour les états i et j , et g est la pénalité attribuée aux écarts (indels). La matrice $D(i, j)$ retranscrit ainsi dans chaque cellule le score maximum pour chaque alignement possible. Toutes les paires possibles sont considérées au sein des deux séquences.

Exemple d'une telle matrice, avec les mots SEND et AND, dans la table A.4. Trois possibilités sont à considérer pour calculer les scores candidats de $D(i, j)$:

- Le chemin partant de la cellule supérieure ou de gauche représente un appariement indel, il faut donc prendre les scores de la cellule de gauche et de la cellule supérieure, et ajouter le score d'indel à chacun d'eux.
- Le chemin en diagonal représente un match/mismatch, il faut prendre donc le score de la cellule en diagonale supérieure gauche et ajoutez le score de match si les bases (lettres) correspondantes de la ligne et de la colonne sont concordantes ou le score de mismatch si elles ne le sont pas.

		S	E	N	D
	D(1,1)	D(1,2)	D(1,3)	D(1,4)	D(1,5)
A	D(2,1)	D(2,2)	D(2,3)	D(2,4)	D(2,5)
N	D(3,1)	D(3,2)	D(3,3)	D(3,4)	D(3,5)
D	D(4,1)	D(4,2)	D(4,3)	D(4,4)	D(4,5)

TABLE A.4 – Matrice de scores $D(i, j)$

Le score résultant pour la cellule est le score le plus élevé (ou le plus faible) des trois scores candidats. Etant donné qu'il n'y a pas de cellule supérieure ou supérieure gauche sur la première ligne, la cellule existante à gauche peut être utilisée pour calculer le score de chaque cellule. La même idée s'applique pour la première colonne, comme seul le score existant au dessus de chaque cellule peut être utilisé. Cette première ligne et première colonne correspondent à l'initialisation de la matrice.

Le schéma de score se rapprochant davantage de l'OM, où match = 0 et mismatch/indel = 1, nous donne l'initialisation de la matrice en table A.5, e.g. $D(2, 1) = D(1, 1) + \text{coût d'indel} = 1$.

	S	E	N	D	
	0	1	2	3	4
A	1				
N	2				
D	3				

TABLE A.5 – Initialisation de la matrice de scores $D(i, j)$

Ensuite, $D(2, 2)$ peut être calculé, suivi de $D(2, 3)$ (ou $D(3, 2)$), $D(2, 4)$ (ou $D(4, 2)$), etc. La matrice comprenant tous les $D(i, j)$ calculés est visible dans la table A.6. Exemple du calcul de $D(2, 2)$ et de $D(4, 5)$:

$$D(2, 2) = \min \begin{cases} D(1, 1) + s(2, 2) = 0 + 1 = 1 \\ D(1, 2) + 1 = 2 \\ D(2, 1) + 1 = 2 \end{cases}$$

$$D(4, 5) = \min \begin{cases} D(3, 4) + s(4, 5) = 2 + 0 = 2 \\ D(3, 5) + 1 = 4 \\ D(4, 4) + 1 = 4 \end{cases}$$

	S	E	N	D	
	0	1	2	3	4
A	1	1	2	3	4
N	2	2	2	2	3
D	3	3	3	3	2

TABLE A.6 – Remplissage de la matrice

Ensuite, pour retrouver l'alignement optimal, le chemin le plus court doit être tracé en partant de la cellule en bas à droite de la matrice ($D(4, 5)$) jusqu'à arriver en haut à gauche ($D(1, 1)$). On parle de "retour à l'origine", nous donnant le meilleur alignement pour chacune des séquences. Pour ce faire, on se déplace de cellules en cellules (soit en diagonal, soit à gauche), en choisissant la cellule avec le score le plus faible. Une fois le chemin trouvé, on déduit l'alignement comme suit :

- Si déplacement en diagonal : cela correspond à un match/mismatch, alors la lettre de la colonne (ou la lettre de la ligne) de la cellule d'origine va s'aligner.
- Si déplacement vertical ou horizontal : cela correspond à un indel, alors le déplacement vertical va aligner un gap (le "-") à la lettre correspondant à la ligne de la cellule (pas de "-" pour la lettre de la colonne), tandis que le déplacement horizontal va aligner un gap à la lettre correspondant à la colonne de la cellule (pas de "-" pour la lettre de la ligne).

		S	E	N	D
	0	1	2	3	4
A	1	1	2	3	4
N	2	2	2	2	3
D	3	3	3	3	2

TABLE A.7 – Détermination du meilleur alignement (chemin parcouru visible en **gras**)

Dans notre cas (cf. table A.7) cela nous donne :

- D → ND → END → END → SEND
- D → ND → AND → -AND

SEND et -AND est donc le meilleur alignement, avec un score d'alignement de $2 + 2 + 2 + 1 = 7$. En OM, ce score correspondrait à la distance (dissimilité) entre les deux séquences.

A.9.2 La librairie TraMineR

La librairie TraMineR [23] offre un panel de possibilités pour appliquer les méthodes de calcul de distances entre séquences - ou mesures de dissimilarité - introduites dans la section 2.1.

Calcul des coûts

Pour calculer les coûts de substitution et d'indel, TraMineR dispose de la fonction `seqcost()`. L'argument "method =" de cette fonction donne le choix entre différentes méthodes de calcul de coût, listées ci-dessous.

CONSTANT : Coûts constants. Autres arguments associés :

- *cval* : définition du coût unique de substitution (défaut = 2)

TRATE : Coûts dérivés à partir des taux de transition observés. Autres arguments associés :

- *time.varying*
- *cval* : pour cette méthode, *cval* doit être une valeur de base de laquelle sont soustraite les probabilités de transition ; si l'utilisateur ne met rien (*NULL*), alors la valeur mise par défaut sera *cval*=2
- *state.features* : dataframe contenant des attributs/features associés à chaque état (*peu d'informations disponibles concernant l'utilisation de cet argument*)

Avec cette méthode, le coût de substitution $\gamma(a, b)$ entre les états a et b est calculé comme suit :

$$\gamma(a, b) = cval - P(a|b) - P(b|a)$$

où $P(a|b)$ est la probabilité de transition de l'état a sachant b et *cval* est une constante fixée par l'utilisateur. Le coût d'indel c_I unique est ensuite déterminé à partir du coût maximal de substitution :

$$c_I = \max(\text{coûts de substitution})/2$$

INDELS et **INDELSLOG** : Coûts basés sur les coûts d'indel (un coût est calculé par état). Autre argument associé :

- *state.features*

Avec ces méthodes, les coûts des indels sont d'abord dérivés des fréquences relatives des états f_a . Un récapitulatif du calcul des coûts avec les différentes méthodes est visible dans le tableau A.8.

L'argument *time.varying=TRUE* peut être utilisé dans le cadre de la méthode DHD (section 2.3.2) pour le calcul d'une matrice de substitution dépendante de la période t .

Un récapitulatif du calcul des coûts avec les différentes méthodes est visible dans le tableau A.8.

Méthode	Coût d'indel de l'état a	Coût de substitution des états a et b
CONSTANT	$cval/2$	$cval$
TRATE	$\max(\text{coûts de substitution})$	$cval - P(a b) - P(b a)$
INDELS	$1/f_a$	$c_I(a) + c_I(b)$
INDELSLOG	$\log[2/(1 + f_a)]$	$c_I(a) + c_I(b)$

TABLE A.8 – Détails des méthodes de calcul de coûts sur R

Optimal Matching

Une fois les coûts calculés, la fonction `seqdist()` du package **TraMineR** permet d'appliquer une méthode de calcul de dissimilarité entre les séquences. L'argument "`method =`" dispose d'un large choix de méthodes, notamment l'optimal matching et ses variantes, dont les paramètres associés sont listés ci-dessous.

OM : Optimal Matching classique. Autres arguments associés :

- `indel` : la matrice des coûts d'indel (donnée par `seqcost()`)
- `sm` : la matrice des coûts de substitution (donnée par `seqcost()`)
- `norm` : la normalisation à utiliser - soit d la distance et m la distance maximale possible étant donné les longueurs p et d de deux séquences, et k la longueur de la séquence la plus longue :
 1. $maxlength : \frac{d}{k}$ (normalisation d'Abbott)
 2. $gmean : 1 - \frac{m-d}{p \times q}$ (normalisation d'Elzinga)
 3. $maxdist : \frac{d}{m}$
 4. $YujianBo : 2 \times \frac{d}{m+d}$ (normalisation de Yujian et Bo)
 5. `auto` : un défaut est associé à chaque méthode

OMspell : Optimal Matching des séquences d'épisodes. Autres arguments associés :

- `indel` : la matrice des coûts d'indel (donnée par `seqcost()`)
- `sm` : la matrice des coûts de substitution (donnée par `seqcost()`)
- `tpow` : coefficient exponentiel appliqué pour transformer les durées des épisodes (défaut = 1, i.e. que les longueurs sont prises en compte comme elles sont)
 - si $tpow = 0$: la durée des épisodes est ignorée (OMspell devient l'OM des séquences d'états distincts)
 - si $tpow = 0.5$: la racine carrée de la durée des épisodes est considérée (e.g. augmenter la longueur d'un épisode de 1 à 2 coûterait davantage que de l'augmenter de 3 à 4)
- `expcost` : le coût de la transformation de la durée d'un épisode ($\delta \geq 0$ dans la section A.7.3), i.e. le coût d'étendre ou de compresser la durée d'un épisode d'une unité de temps (défaut = 0.5)
 - la durée de l'épisode prise en compte par `expcost` est celle qui a été transformée au préalable avec l'argument `tpow`
 - si `expcost = 0` : la durée des épisodes est ignorée (OMspell devient l'OM des séquences d'états distincts)
- `norm`

OMloc : Optimal Matching localisé. Autres arguments associés :

- `sm` (pas de matrice de coûts d'indel à entrer, cette dernière étant calculée à partir de `sm` et des arguments ci-dessous)
- `expcost` : β dans la section A.7.1, le coût fixe pour l'insertion (dépendant du coût de substitution maximal)
- `context` : α dans la section A.7.1, la pondération de la distance entre l'élément inséré et ses voisins, i.e. le coût d'une insertion locale (défaut = $1 - 2 * expcost$)
- `norm`

OMslen : Optimal Matching sensible à la longueur des épisodes. Autres arguments associés :

- `indel` : la matrice des coûts d'indel (donnée par `seqcost()`)
- `sm` : la matrice des coûts de substitution (donnée par `seqcost()`)
- `link` : la fonction utilisée pour (re)calculer les coûts de substitution pour cette méthode, qui peut être soit `mean` (moyenne arithmétique) ou `gmean` (moyenne géométrique)
- `h` : le poids attribué à la longueur d'un épisode (défaut = 0.5) (h dans la section A.7.2)

OMstran : Optimal Matching des séquences de transition

- `indel` : la matrice des coûts d'indel (donnée par `seqcost()`)
- `sm` : la matrice des coûts de substitution (donnée par `seqcost()`)
- `transindel` : méthode pour calculer les nouveaux coûts d'indel de transition, 3 possibilités : "`constant`" (coût fixé à 1.0), "`subcost`" (basé sur les coûts de substitution), ou "`prob`" (basé sur les probabilités de transition)
- `otto` : le poids (w dans la section A.7.4) associé au compromis origine-transition (entre 0 et 1)
- `previous` : indique si l'on veut prendre compte la transition de l'état précédent (`TRUE` ou `FALSE`)
- `add.column` : est-ce que la dernière colonne (et la première colonne quand `previous = TRUE`) devrait être dupliquée quand les séquences ont des longueurs différentes
- `norm`

Autres méthodes

L'argument "*method* = " dispose d'autres choix de méthodes que l'OM et ses variantes.

HAM, DHD : les distances de Hamming et de Hamming Dynamique

- *sm* : facultatif car la matrice de coûts peut être dérivée automatiquement (R fixera dans ce cas des coûts de substitution égaux à 1)
- *norm*

LCS : la distance basée sur la plus longue sous-séquence commune

- *norm*

NMS : le nombre de sous-séquences correspondantes. Autres arguments associés :

- *kweights* : (facultatif) les poids attribués aux longueurs des sous-séquences (défaut = vecteur de 1), sa longueur doit être égale au nombre de colonnes du jeu de données : *kweights* contient à la position *k* le poids attribué aux sous-séquences de longueur *k* (*a* dans la section A.7.5)

SVRspell : la distance de représentation vectorielle des sous-séquences (une paramétrisation de l'argument *kweights* identique dans le cadre de SVRspell et NMS donne les mêmes résultats). Autres arguments associés :

- *kweights* : les poids attribués aux longueurs des sous-séquences (défaut = vecteur de 1), sa longueur doit être égale au nombre de colonnes du jeu de données : *kweights* contient à la position *k* le poids attribué aux sous-séquences de longueur *k* (*a* dans la section A.7.5)
- *tpow* : *b* dans la section A.7.5

CHI2 : Distance du khi-2 entre les distributions d'états. Autres arguments associés :

- *breaks* : la liste des périodes pouvant se chevaucher (défaut = *NULL*)
- *step* : la longueur des intervalles de temps utilisée pour le calcul de la distribution entre états (défaut = 1, i.e. que les distributions vont être comparées période par période); pour considérer tout le suivi il faudra spécifier *K* = *max(longueurs des séquences)*
- *overlap* : les périodes doivent-elles se chevaucher ? (défaut = *FALSE*)
- *weighted* : les distributions des états doivent-elles tenir compte des poids des séquences ? (défaut = *TRUE*)
- *global.pdotj* : le vecteur de proportions des états à utiliser comme distribution marginale (défaut = *NULL*)

EUCLID : Distance euclidienne entre les distributions d'états. Autres arguments associés :

- *breaks* : la liste des périodes pouvant se chevaucher (défaut = *NULL*)
- *step* : la longueur des intervalles de temps utilisée pour le calcul de la distribution entre états (défaut = 1, i.e. que les distributions vont être comparées période par période); pour considérer tout le suivi il faudra spécifier *K* = *max(longueurs des séquences)*
- *overlap* : les périodes doivent-elles se chevaucher ? (défaut : = *FALSE*)

A.10 Coûts utilisés pour l'application des méthodes dans R

TABLE A.9 – Coûts d'indel (OM INDELSLOG)

Monothérapie	Bithérapie	Insuline	Insuline + OAD	Trithérapie	Sans Rien
0.286	0.592	0.638	0.666	0.669	0.666

TABLE A.10 – Coûts de substitution (OM INDELSLOG)

	Monothérapie	Bithérapie	Insuline	Insuline + OAD	Trithérapie	Sans Rien
Monothérapie	0	0.878	0.924	0.952	0.955	0.946
Bithérapie	0.878	0	1.230	1.258	1.261	1.252
Insuline	0.924	1.230	0	1.304	1.307	1.298
Insuline + OAD	0.952	1.258	1.304	0	1.335	1.326
Trithérapie	0.955	1.261	1.307	1.335	0	1.329
Sans Rien	0.946	1.252	1.298	1.326	1.329	0

TABLE A.11 – Coûts d'indel (OM INDELS)

Monothérapie	Bithérapie	Insuline	Insuline + OAD	Trithérapie	Sans Rien
1.990	9.375	17.604	36.731	40.887	29.543

TABLE A.12 – Coûts de substitution (OM INDELS)

	Monothérapie	Bithérapie	Insuline	Insuline + OAD	Trithérapie	Sans Rien
Monothérapie	0	11.366	19.594	38.721	42.877	31.534
Bithérapie	11.366	0	26.979	46.106	50.262	38.919
Insuline	19.594	26.979	0	54.334	58.490	47.147
Insuline + OAD	38.721	46.106	54.334	0	77.617	66.274
Trithérapie	42.877	50.262	58.490	77.617	0	70.430
Sans Rien	31.534	38.919	47.147	66.274	70.430	0

A.11 Qualité du clustering et temps de calcul en fonction des méthodes et du nombre de classes

A.11.1 Valeur de l'ASW en fonction du nombre de groupes

Statistique	OM I ^a	OM IL ^b	OMloc	LCS	OM LS ^c	OM T ^d	OMslen	OMstran	OM 6P ^e	OMspell
ASW (CAH)	0.341	0.311	0.300	0.297	0.284	0.233	0.226	0.210	0.122	0.103
ASW (PAM)	0.301	0.111	0.254	0.148	0.378	0.160	0.109	0.130	0.133	0.102

(a) Qualité du clustering avec 5 groupes

- a. INDELS
- b. INDELSLOG
- c. Long suivi
- d. TRATE
- e. 6 périodes

Statistique	OM I	OM IL	OMloc	LCS	OM LS	OM T	OMslen	OMstran	OM 6P	OMspell
ASW (CAH)	0.375	0.252	0.233	0.214	0.261	0.228	0.165	0.180	0.136	0.129
ASW (PAM)	0.313	0.232	0.215	0.165	0.330	0.162	0.127	0.115	0.133	0.118

(b) Qualité du clustering avec 7 groupes

Statistique	OM I	OM IL	OMloc	LCS	OM LS	OM T	OMslen	OMstran	OM 6P	OMspell
ASW (CAH)	0.292	0.215	0.249	0.183	0.220	0.168	0.139	0.129	0.082	0.129
ASW (PAM)	0.289	0.188	0.172	0.167	0.253	0.165	0.104	0.139	0.103	0.138

(c) Qualité du clustering avec 10 groupes

TABLE A.13 – Comparaison de la qualité du clustering par variante d'OM

Statistique	SVRspell	NMS	EUCLID	CHI2	HAM	DHD	CHI2 overlap	EUCLID overlap
ASW (CAH)	0.838	0.589	0.378	0.354	0.275	0.260	0.221	0.0951
ASW (PAM)	0.145	0.469	0.231	0.304	0.217	0.232	-0.0400	0.0826

(a) Qualité du clustering avec 5 groupes

Statistique	SVRspell	NMS	EUCLID	CHI2	HAM	DHD	CHI2 overlap	EUCLID overlap
ASW (CAH)	0.834	0.592	0.355	0.365	0.205	0.224	0.0884	0.0851
ASW (PAM)	0.103	0.465	0.377	0.370	0.266	0.274	-0.0272	0.0707

(b) Qualité du clustering avec 7 groupes

Statistique	SVRspell	NMS	EUCLID	CHI2	HAM	DHD	CHI2 overlap	EUCLID overlap
ASW (CAH)	0.812	0.591	0.311	0.349	0.180	0.225	0.103	0.0868
ASW (PAM)	0.081	0.345	0.286	0.338	0.217	0.215	0.0409	0.0741

(c) Qualité du clustering avec 10 groupes

TABLE A.14 – Comparaison de la qualité du clustering pour d'autres types de mesure de dissimilarité

Nombre de groupes	BoW	MCSA
5	0.401	0.400
7	0.469	0.338
10	0.377	0.317

TABLE A.15 – Qualité du clustering (ASW) des autres méthodes explorées

A.11.2 Temps de calcul

	OMloc	OMstran	OM I	OM IL	OMslen	LCS	OM T	OM LS	OMspell	OM 6P
Temps (sec)	41	35	24	23	21	14	10	7	5	1

TABLE A.16 – Temps de calcul des variantes d'OM pour déterminer les distances dans R (sur 10 itérations)

	NMS	SVRspell	CHI2 overlap	EUCLID overlap	CHI2	EUCLID	DHD	HAM
Temps (sec)	91	11	4	4	1	1	1	1

TABLE A.17 – Temps moyen de calcul des autres mesures de dissimilarité dans R (sur 10 itérations)

	BoW	MCSA
Temps (sec)	7	54

TABLE A.18 – Temps moyen de calcul des autres méthodes explorées (sur 10 itérations)

A.12 Visualisation des résultats

A.12.1 Optimal Matching

Indicateurs de qualité du clustering

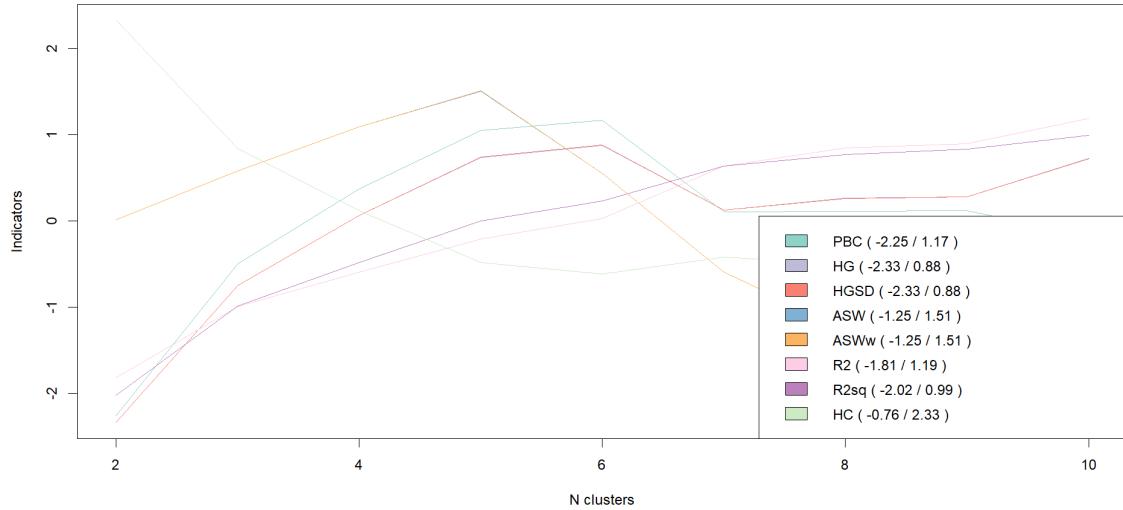


FIGURE A.3 – Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM INDELSLOG)

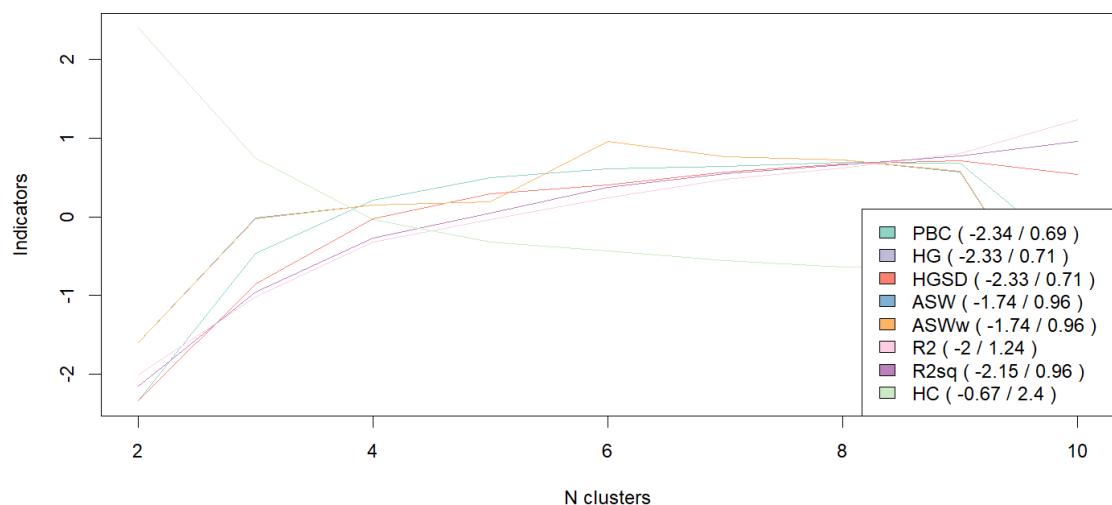


FIGURE A.4 – Evolution des indicateurs de qualité du clustering jusqu'à 10 groupes (OM INDELS)

Tapis de séquences (CAH)

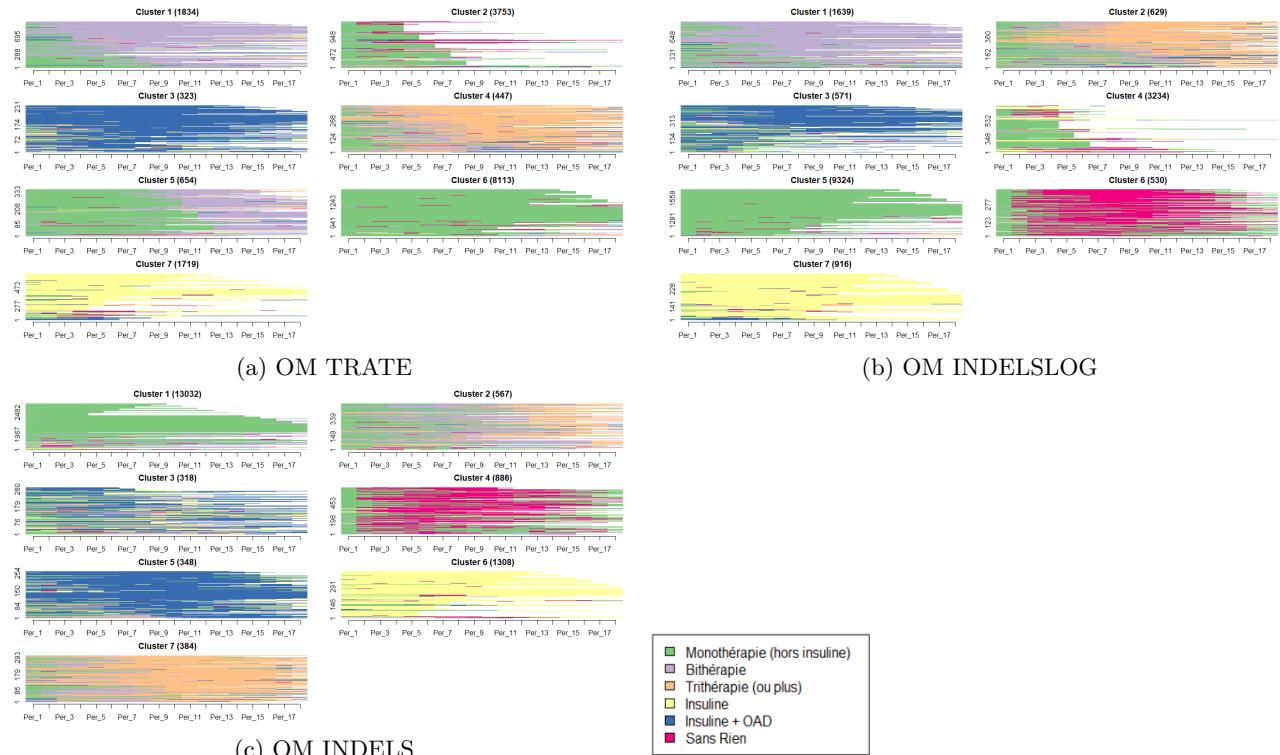


FIGURE A.5 – OM (CAH) - 7 groupes

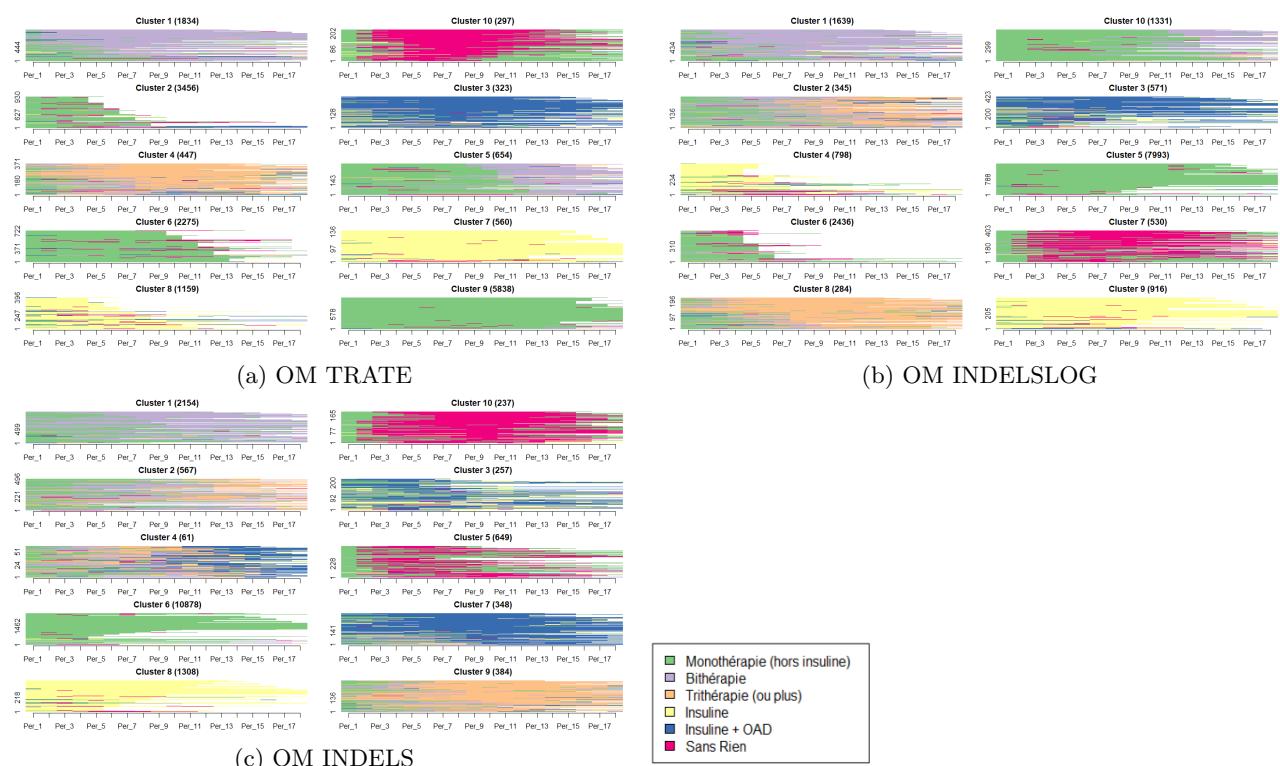


FIGURE A.6 – OM (CAH) - 10 groupes

Tapis de séquences (PAM)

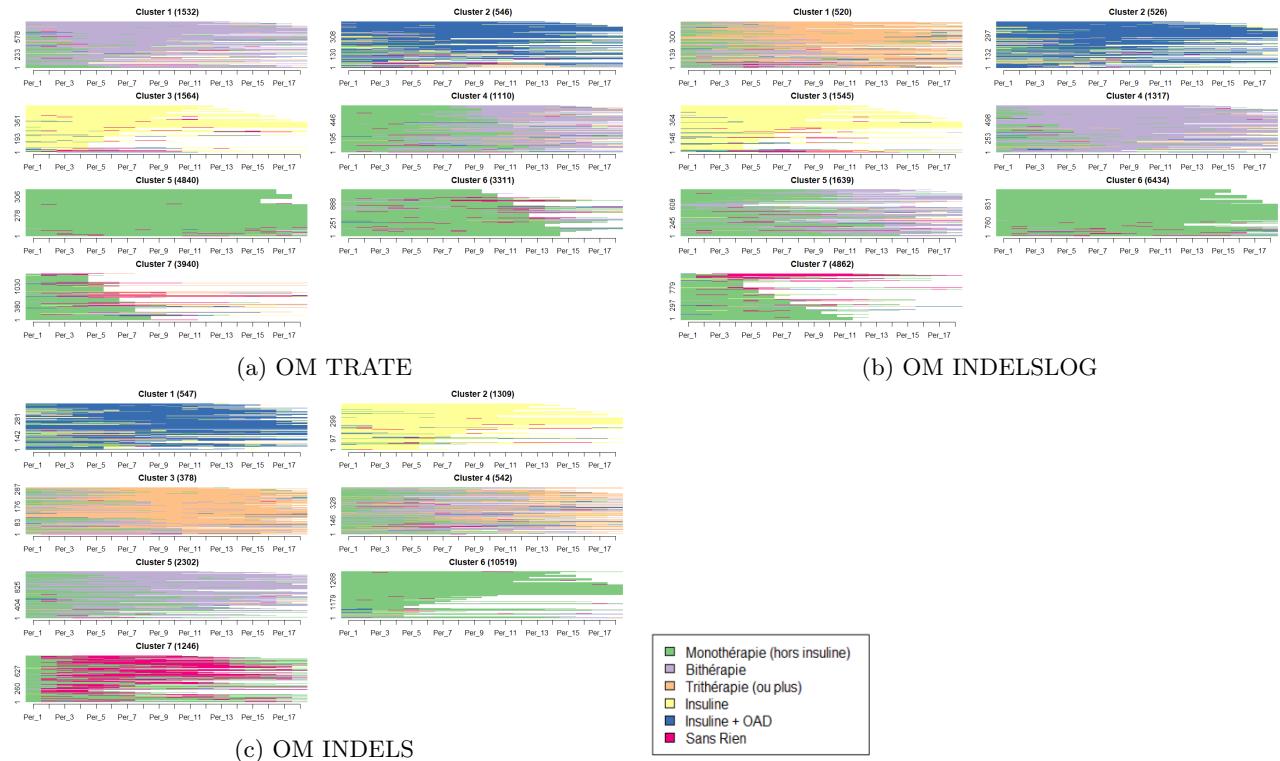


FIGURE A.7 – OM (PAM) - 7 groupes

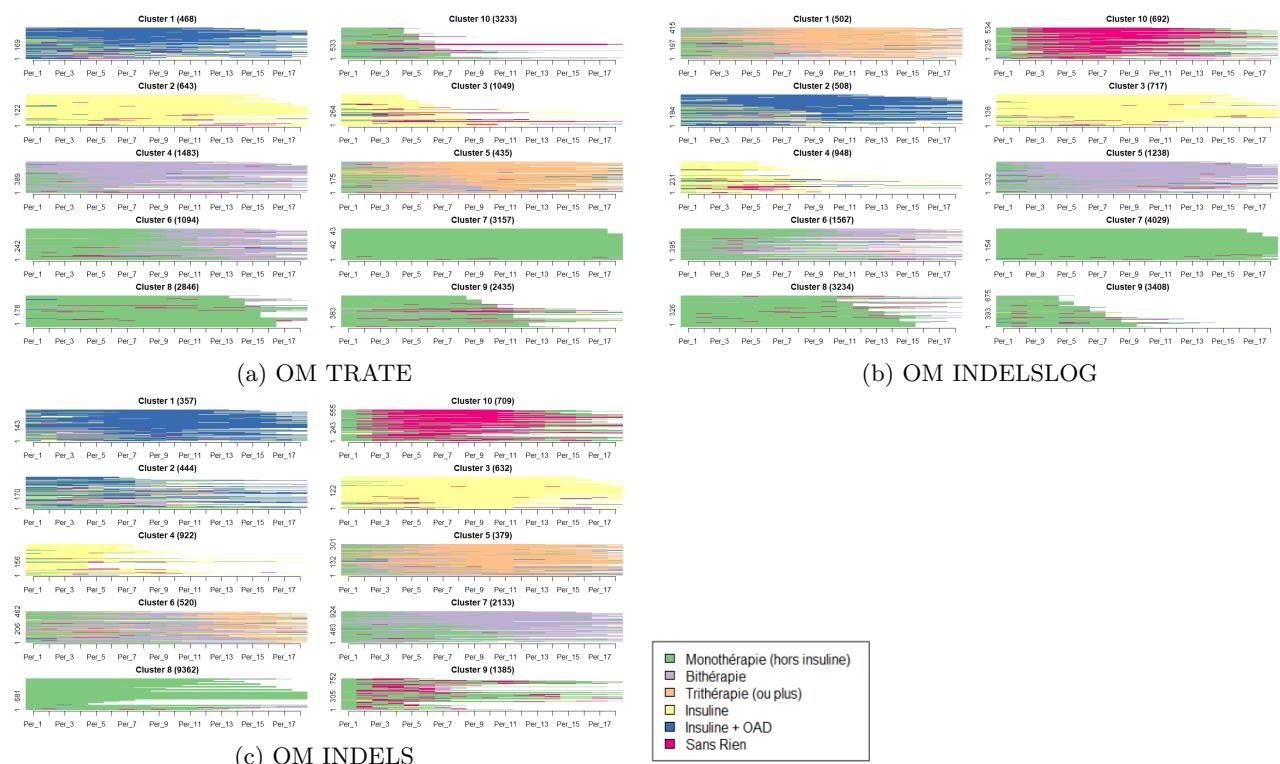


FIGURE A.8 – OM (PAM) - 10 groupes

A.12.2 OM sur des séquences de 41 états

Les méthodes appliquées jusque là ont été réalisées sur des états étiquetés par type de thérapie de façon à simplifier la visualisation et surtout l'interprétation des clusters. Il est donc intéressant de voir ce qu'une mesure de dissimilarité comme l'OM est capable de faire sur des séquences peu transformées, dont les états correspondent aux noms de molécules et aux combinaisons de celles-ci. De cette façon, 41 états sont considérés (metformine, sulfamide, SGLT2, iDPP4, AGLP1, insuline à action rapide, autre type d'insuline, autre molécule ou combinaisons de plusieurs d'entre elles). Au maximum 3 molécules sont combinées. Une fois l'OM effectué avec la méthode de calcul de coûts "TRATE" et le clustering appliqué, la visualisation en tapis de séquences reste possible mais devient compliquée à interpréter (cf. Figure A.9 pour un exemple de visualisation à 10 clusters).

Il est par ailleurs possible de reconstruire les clusters avec les états simplifiés - les types de thérapies - pour les visualiser et ensuite de regarder le détail de la composition des groupes par molécule (combinaison de molécules) et par année, comme présenté dans l'Annexe A.12.9. La représentation graphique simplifiée du clustering avec 10 classes est visible dans la Figure A.10. Les clusters suivants sont obtenus :

- Groupe 1 : Insuline de type autre insuline (733 patients)
- Groupe 2 : Monothérapie sous metformine (7937 patients)
- Groupe 3 : Insuline avec combinaison rapide et autre insuline (855 patients)
- Groupe 4 : Monothérapie (mais pas que...) suivi court (1772 patients), correspond au groupe de séquences atypiques
- Groupe 5 : Monothérapie sous IDPP4 (718 patients)
- Groupe 6 : Insuline+OAD (374 patients)
- Groupe 7 : Monothérapie sous sulfamide (2058 patients)
- Groupe 8 : Bithérapie sous metformine + autre (inhibiteur alpha-glucosidase) (546 patients)
- Groupe 9 : Suivi discontinu (614 patients)
- Groupe 10 : Bithérapie sous metformine + sulfamide (516 patients)

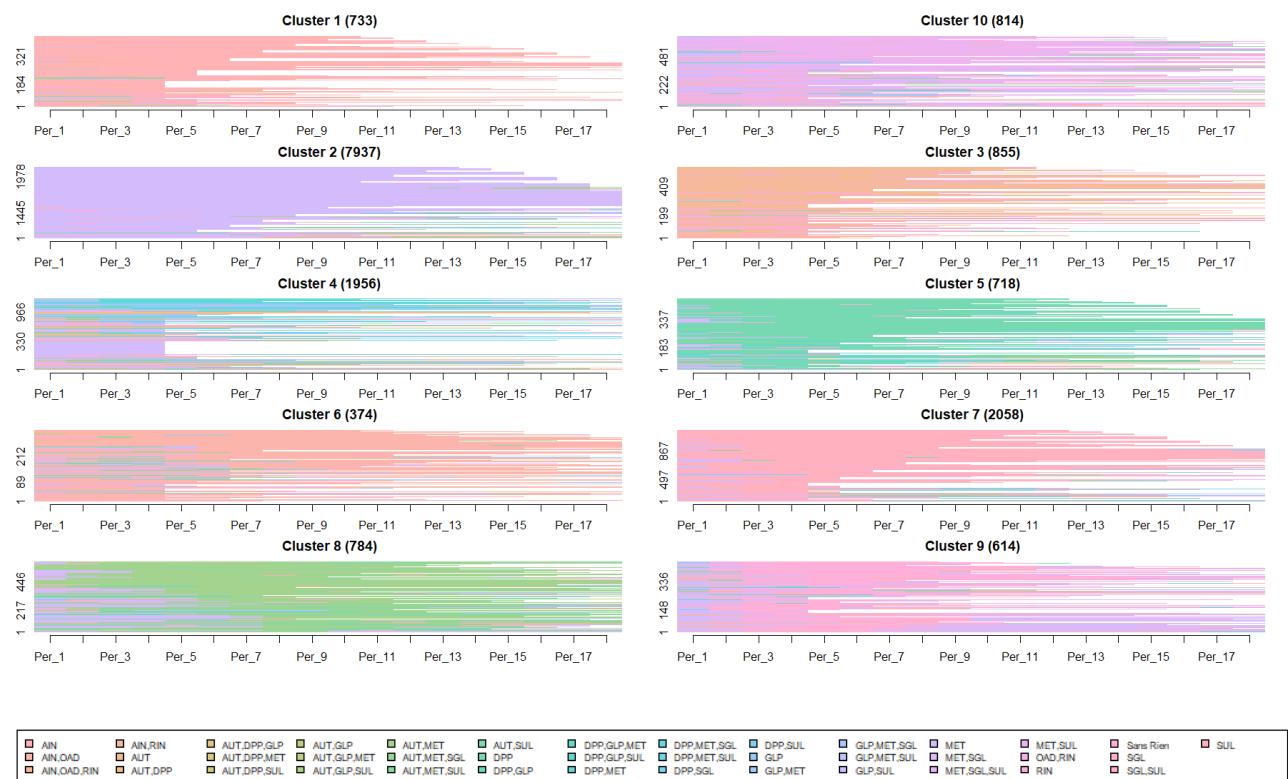


FIGURE A.9 – OM sur des séquences de 41 états (CAH) - 10 groupes

La principale différence avec un clustering à 6 états réside dans la distinction des deux principales molécules utilisées en mono et bithérapie : metformine et sulfamide, donnant une classification mettant l'accent sur le type de molécule et non plus le type de thérapie, même si cette dernière reste visible. Augmenter le nombre d'états conduit par ailleurs à l'apparition d'un groupe contenant des séquences atypiques (le cluster 4 dans la Figure A.9), que la CAH n'a pas réussi à classer dans d'autres classes. Pour contrer ce problème, il est envisageable d'augmenter le nombre de groupes, jusqu'à que ce cluster disparaisse (ou non). Il faut garder en tête que la classification est une cartographie de la base de données, et que tous les patients ne peuvent malheureusement pas toujours faire partie d'un "parcours-type", tous les parcours étant, en théorie, uniques.

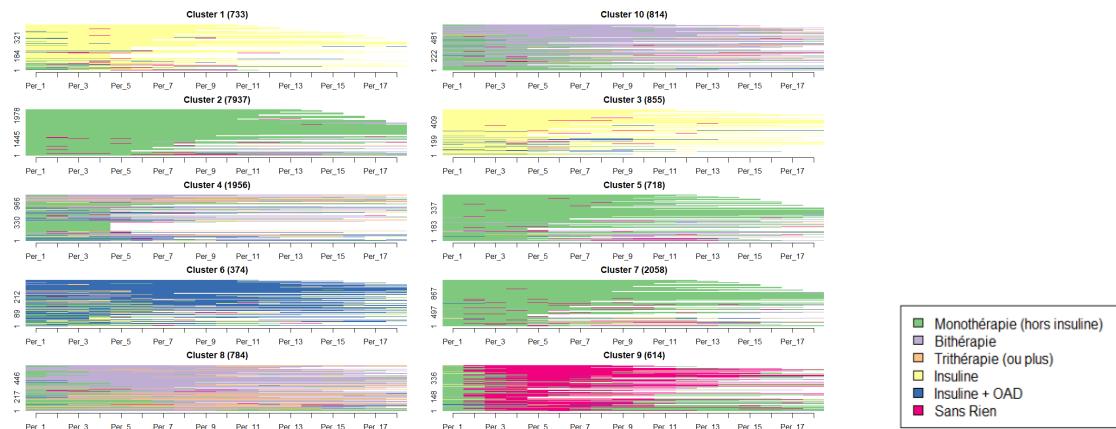


FIGURE A.10 – OM sur des séquences de 41 états (CAH) - 10 groupes (nombre d'états réduit à 6)

Tapis de séquences (sur 6 périodes d'1 an)

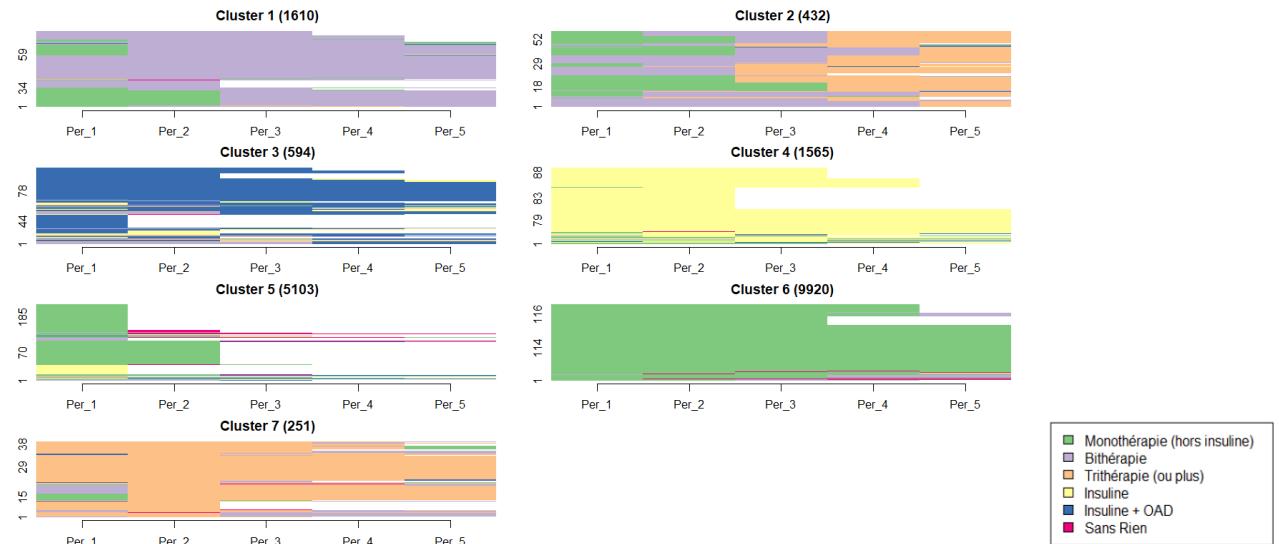


FIGURE A.11 – OM sur 6 périodes (CAH) - 7 groupes

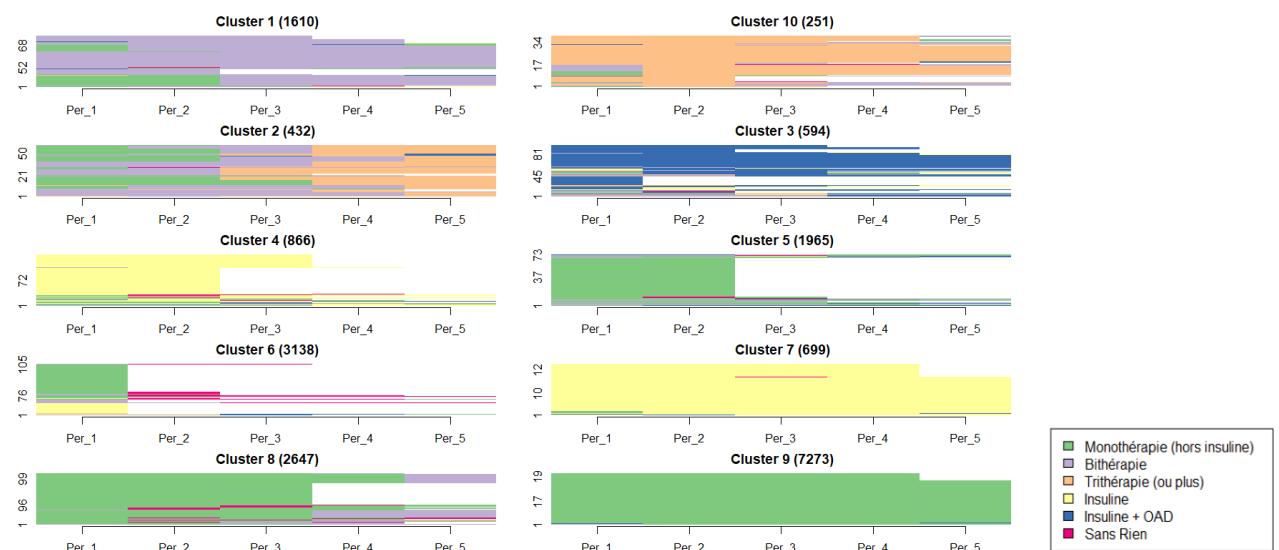


FIGURE A.12 – OM sur 6 périodes (CAH) - 10 groupes

A.12.3 Présentation de l'algorithme LRx de typage du diabète

Présentation de l'algorithme



FIGURE A.13 – Algorithme LRx du typage du diabète

A.12.4 Variantes de l'OM

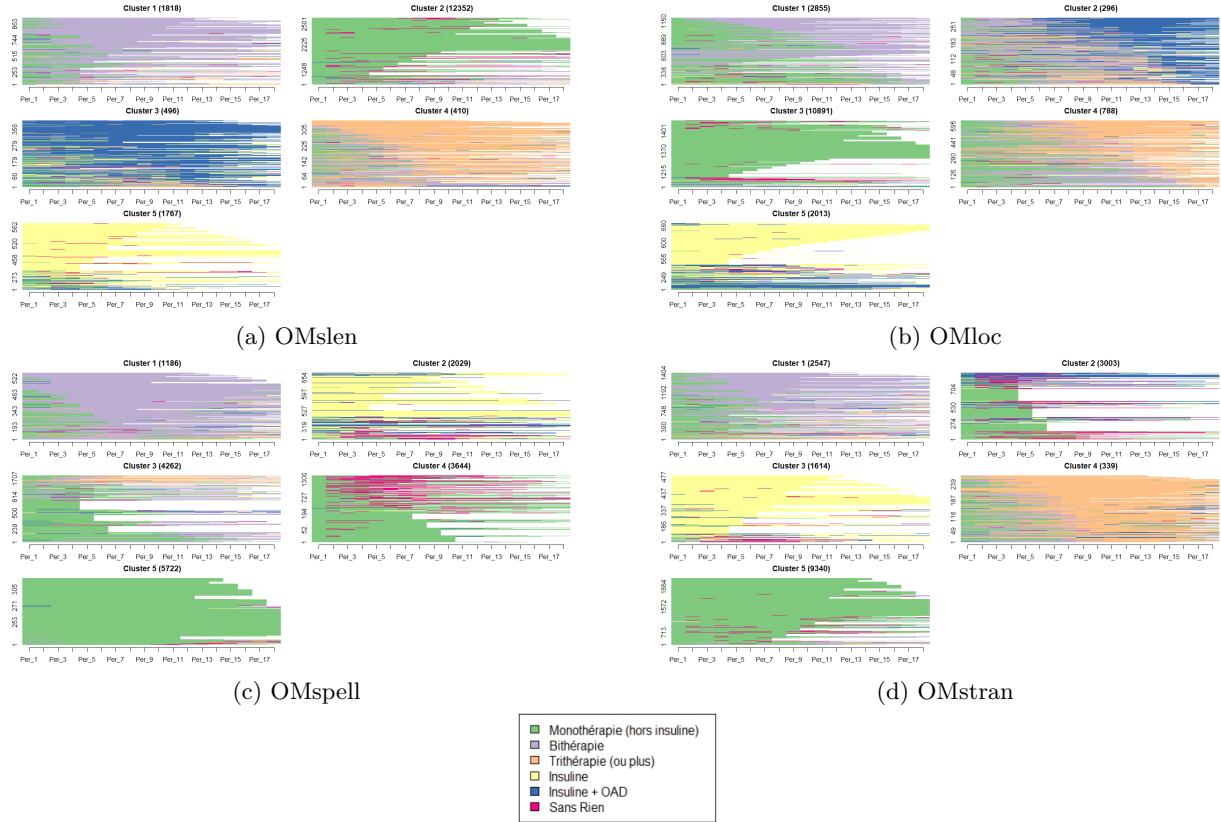


FIGURE A.14 – Variantes de l'OM (CAH) - 5 groupes

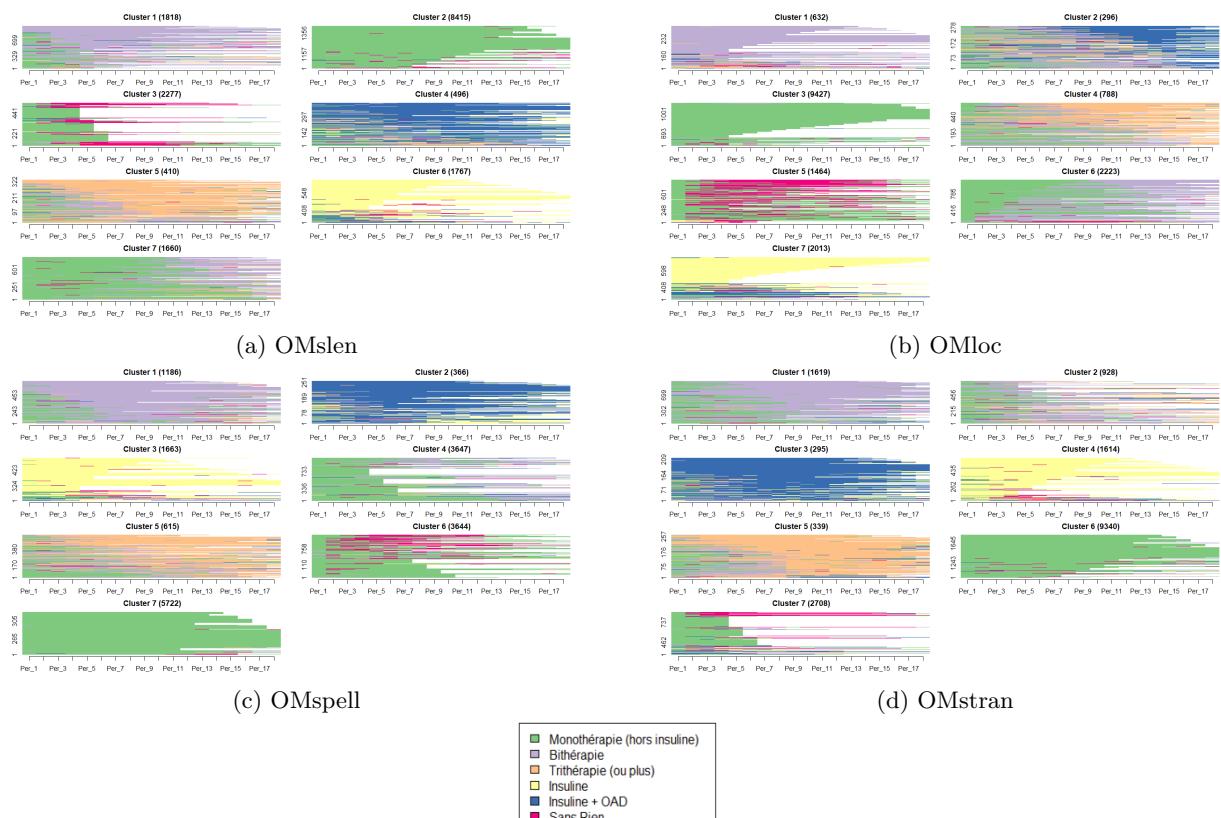


FIGURE A.15 – Variantes de l'OM (CAH) - 7 groupes

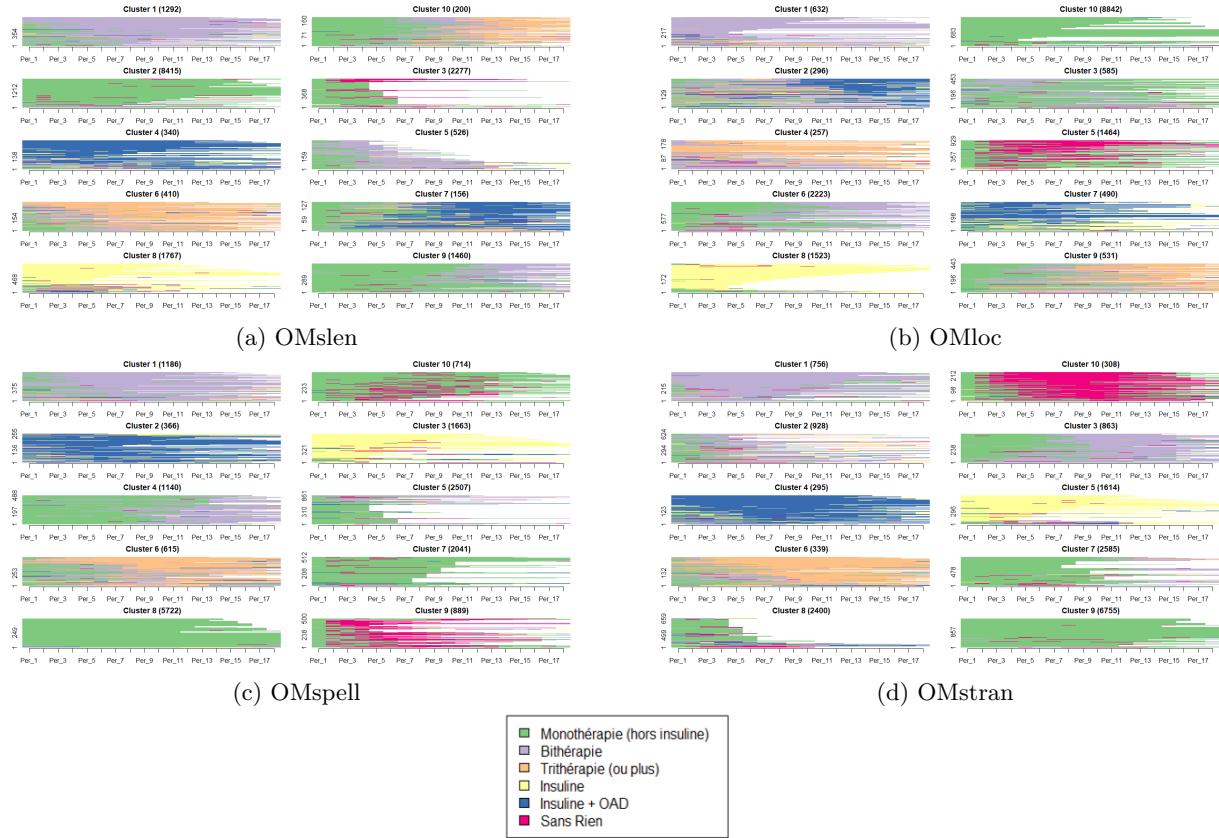


FIGURE A.16 – Variantes de l'OM (CAH) - 10 groupes

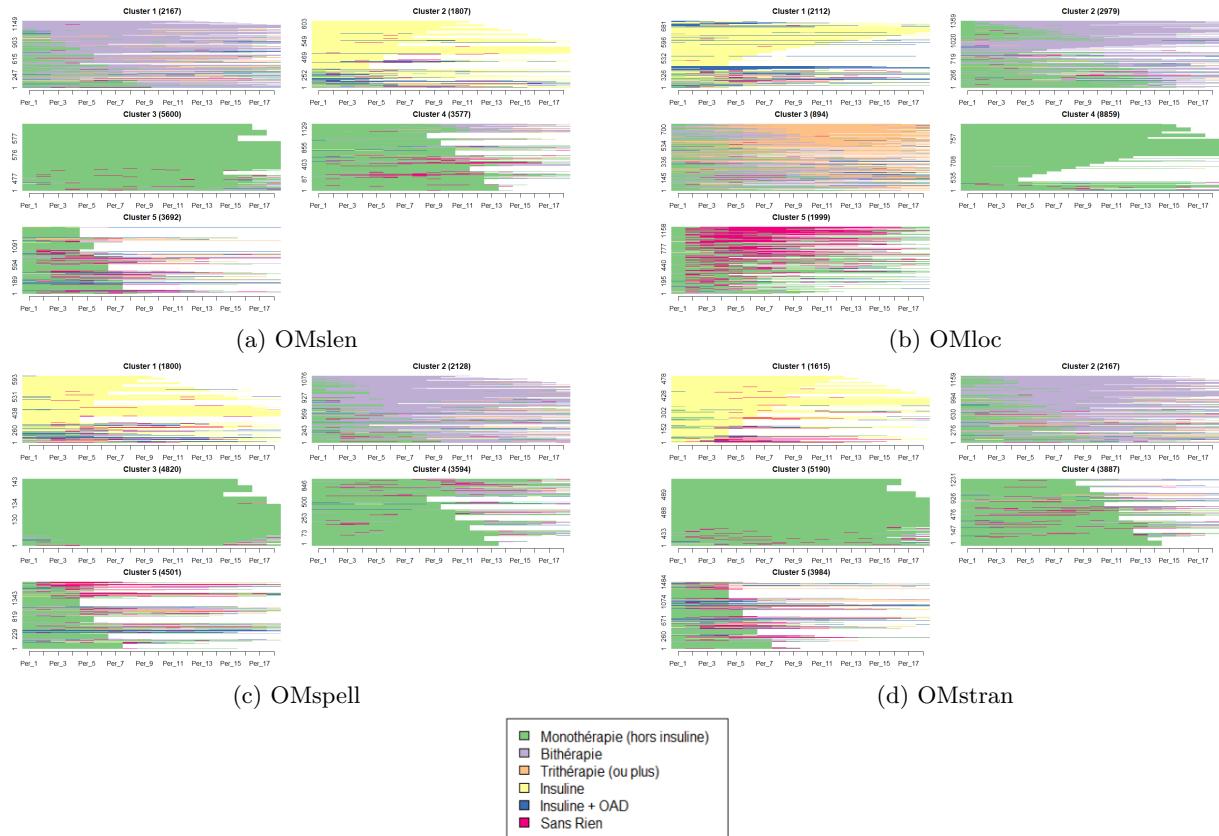


FIGURE A.17 – Variantes de l'OM (PAM) - 5 groupes

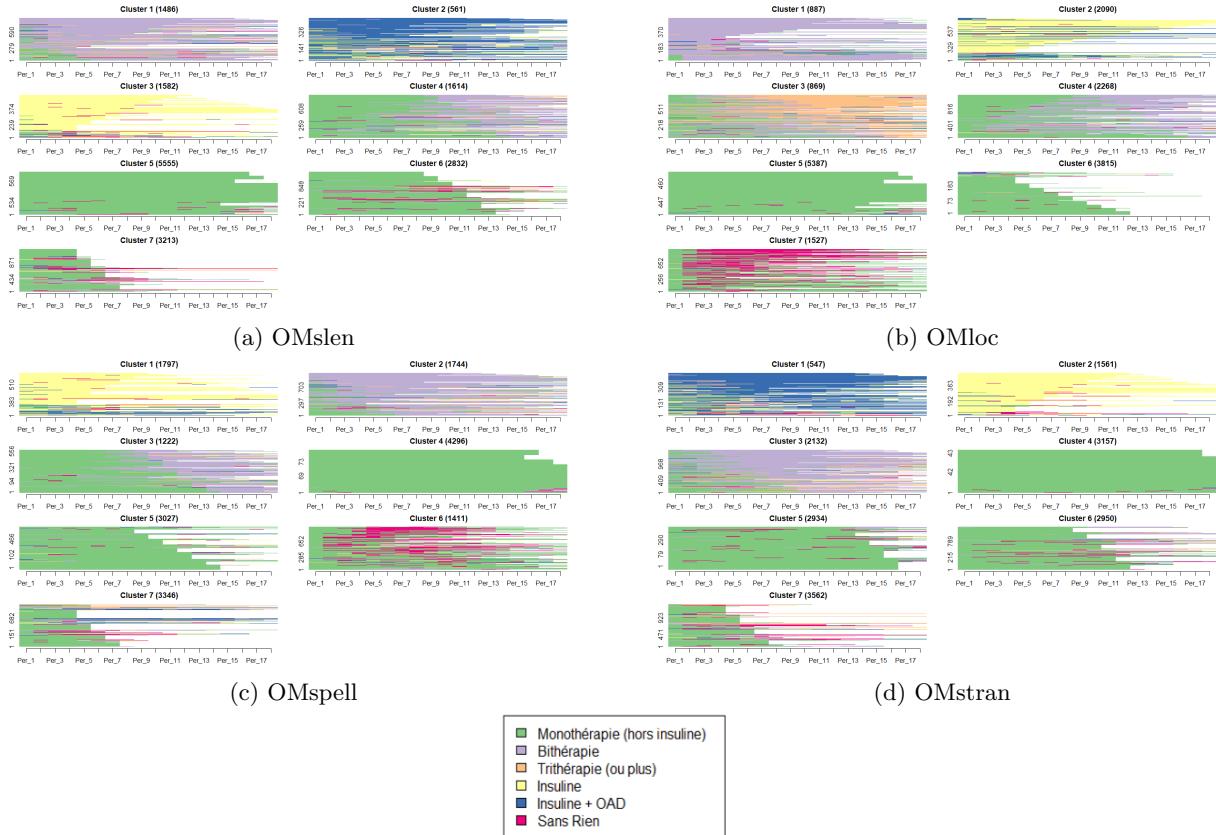


FIGURE A.18 – Variantes de l'OM (PAM) - 7 groupes

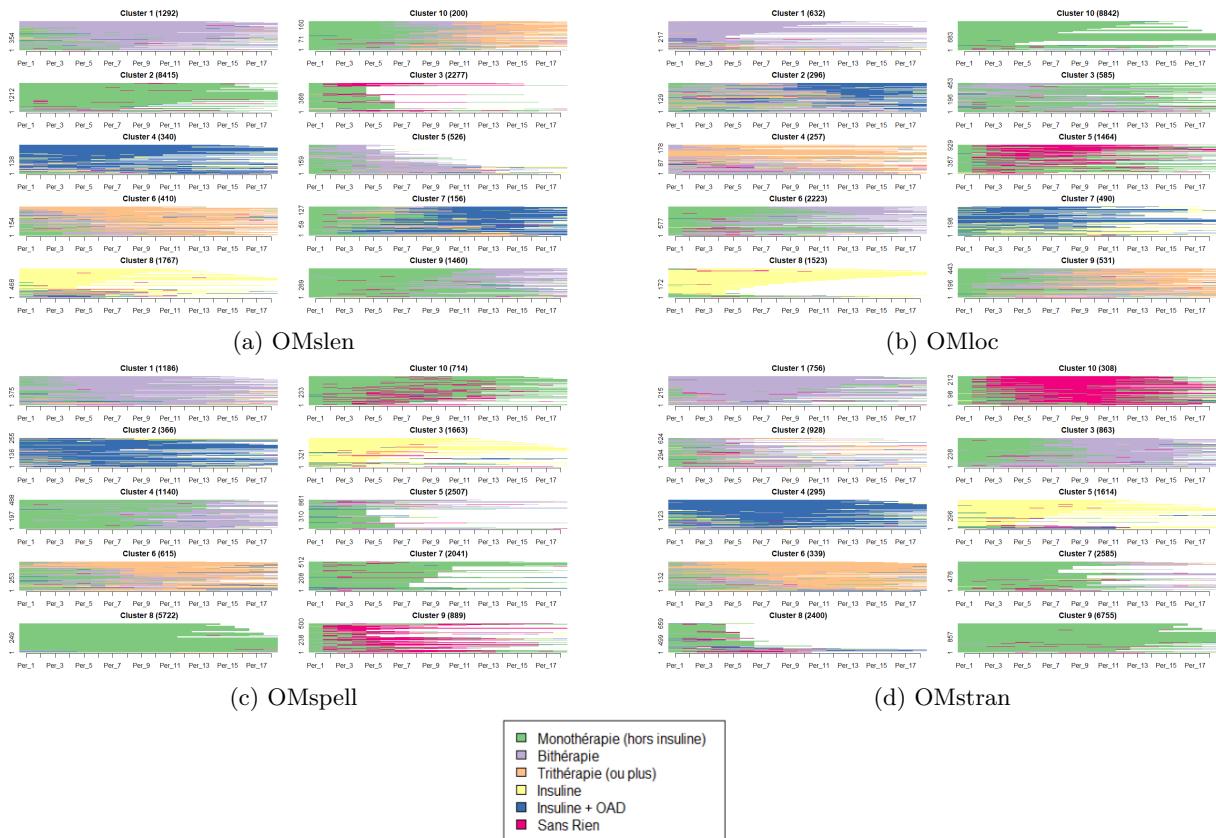


FIGURE A.19 – Variantes de l'OM (PAM) - 10 groupes

A.12.5 Distances basées sur le nombre de sous-séquences communes

Suivi de même longueur : Distances de Hamming vs OM

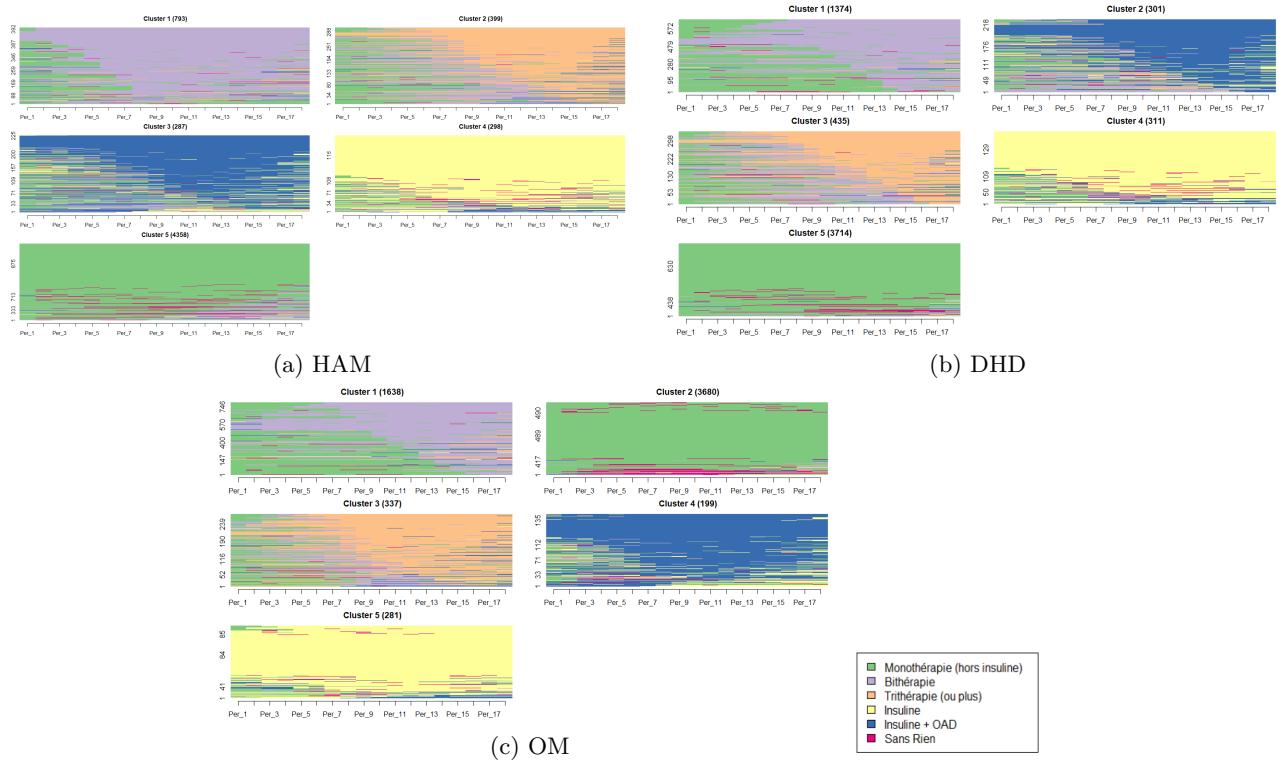


FIGURE A.20 – Distances de Hamming (CAH) - 5 groupes

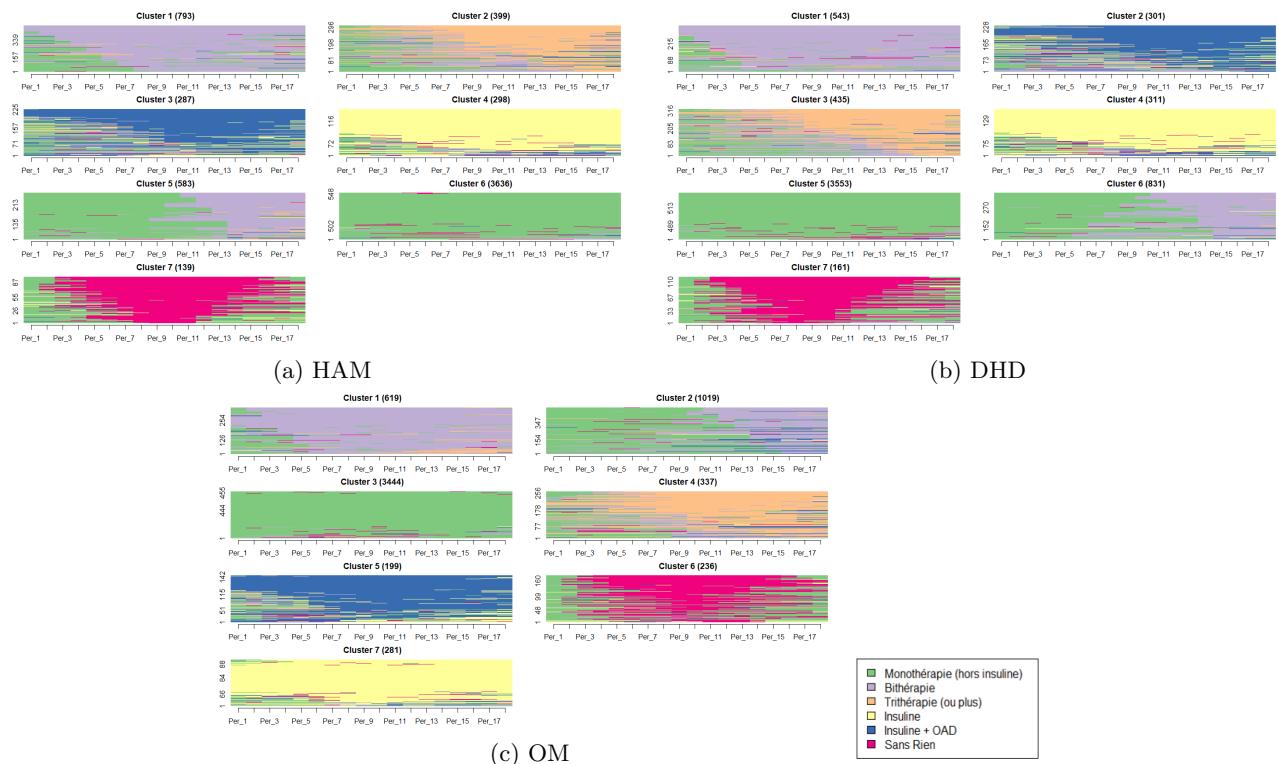


FIGURE A.21 – Distances de Hamming (CAH) - 7 groupes

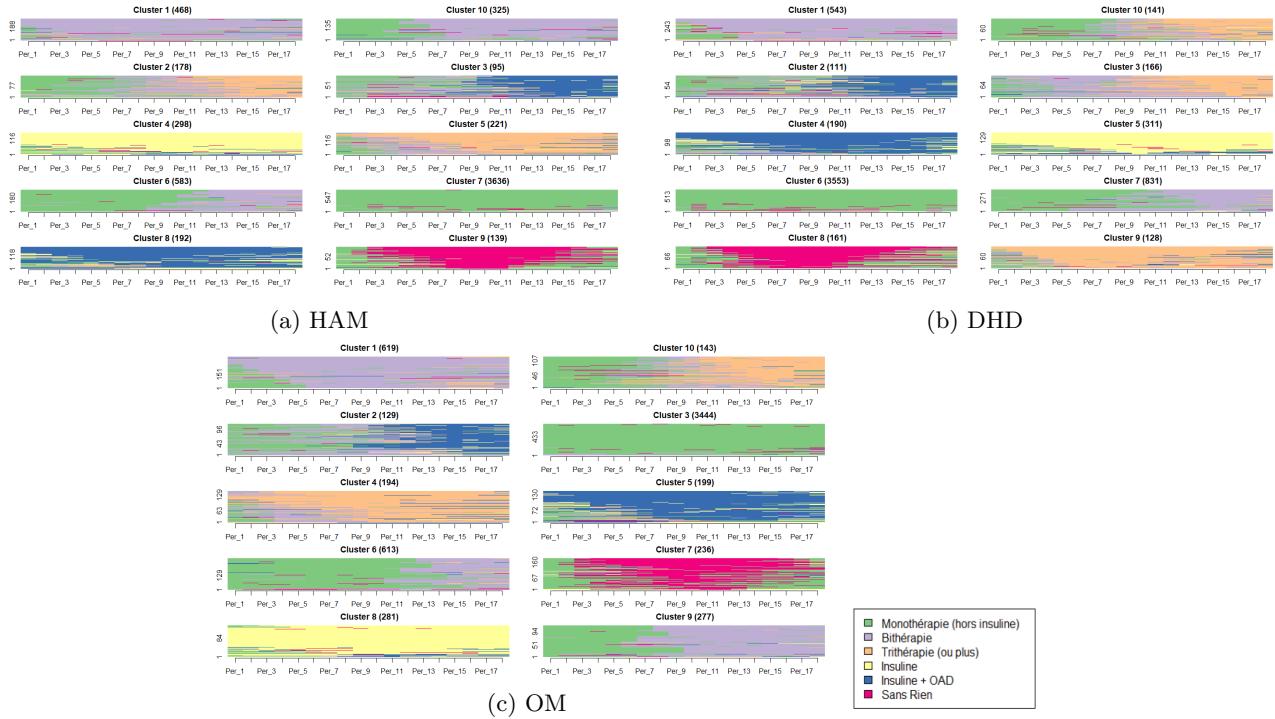


FIGURE A.22 – Distances de Hamming (CAH) - 10 groupes

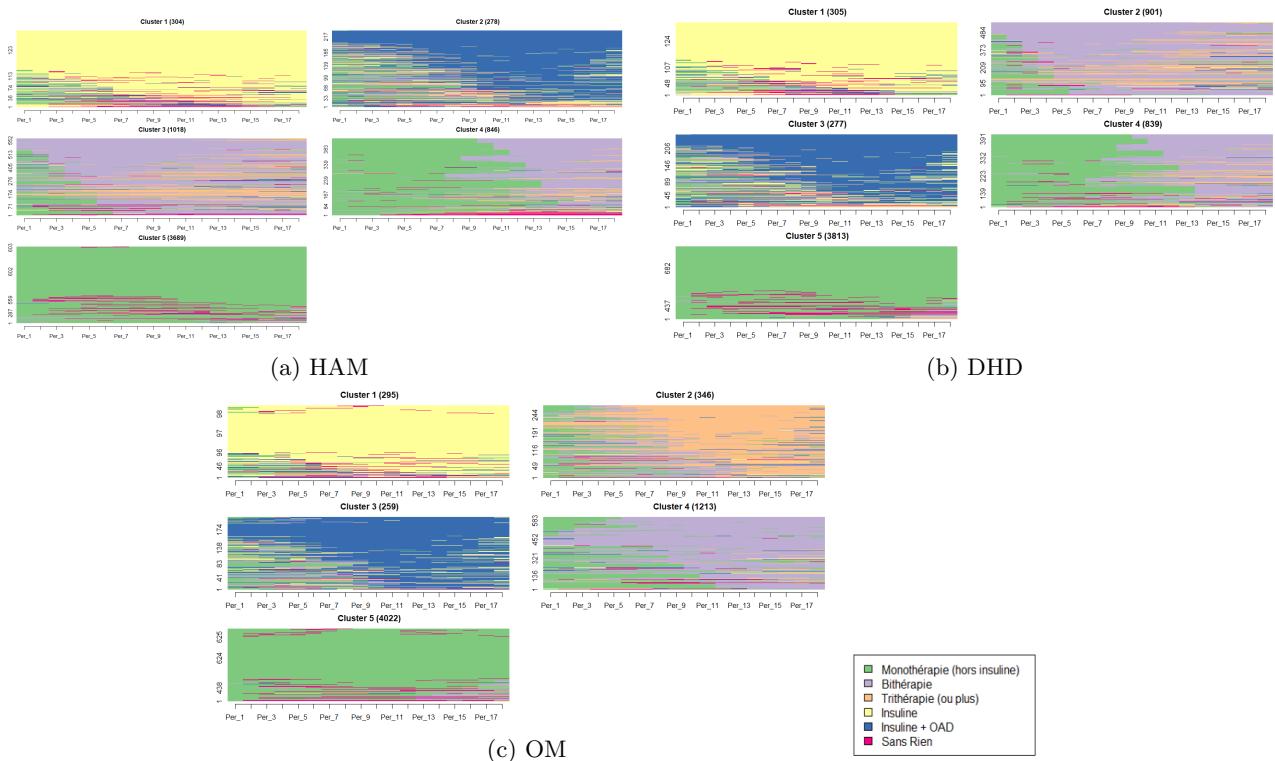


FIGURE A.23 – Distances de Hamming (PAM) - 5 groupes

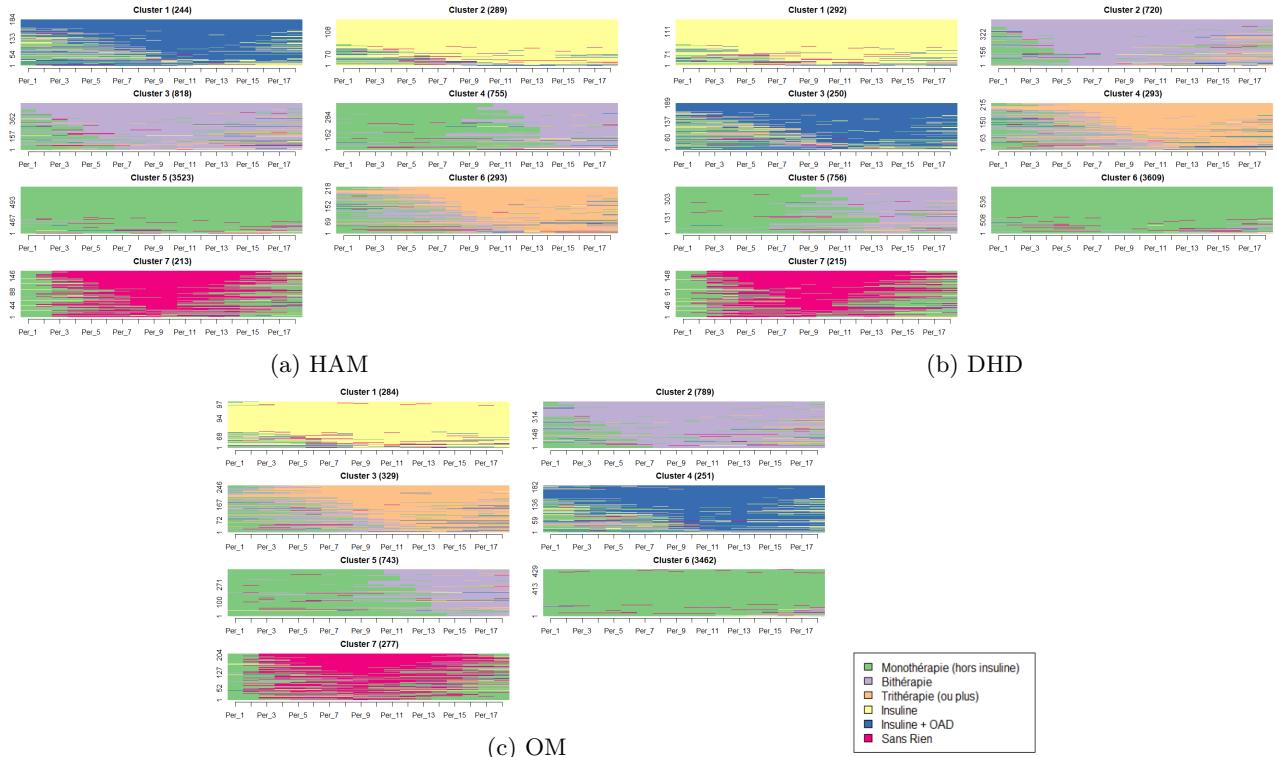


FIGURE A.24 – Distances de Hamming (PAM) - 7 groupes

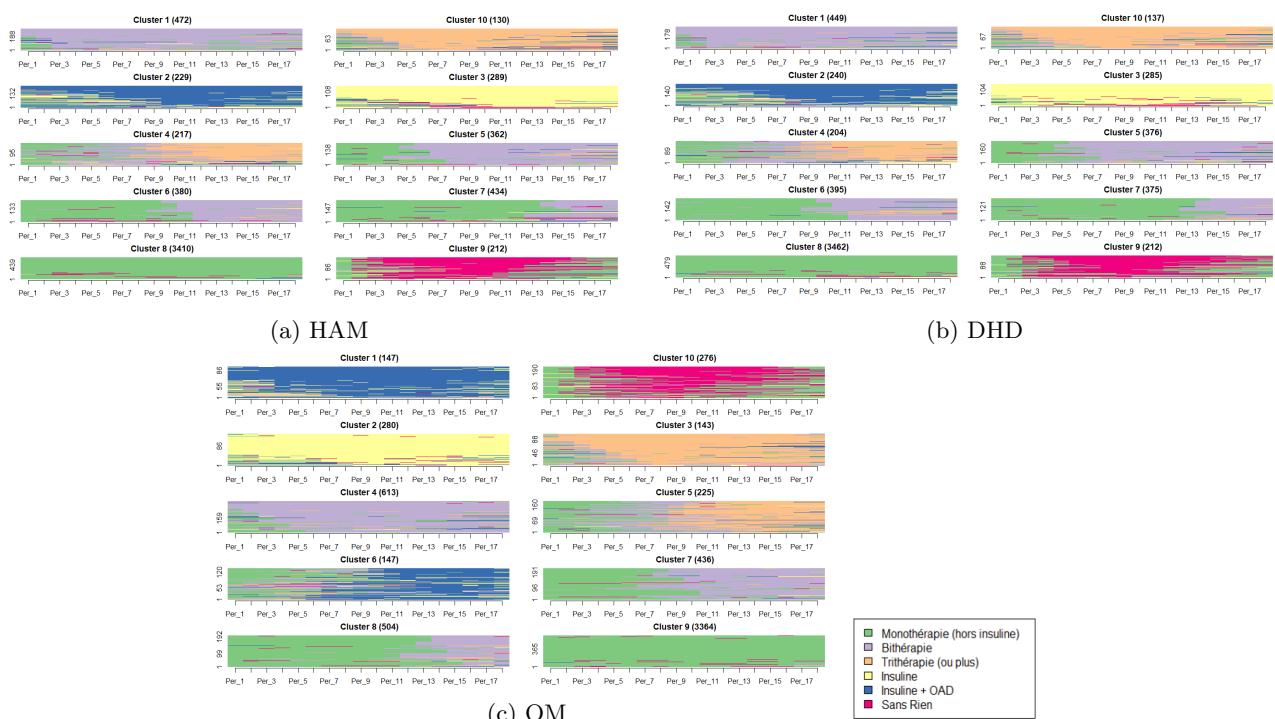


FIGURE A.25 – Distances de Hamming (PAM) - 10 groupes

NMS

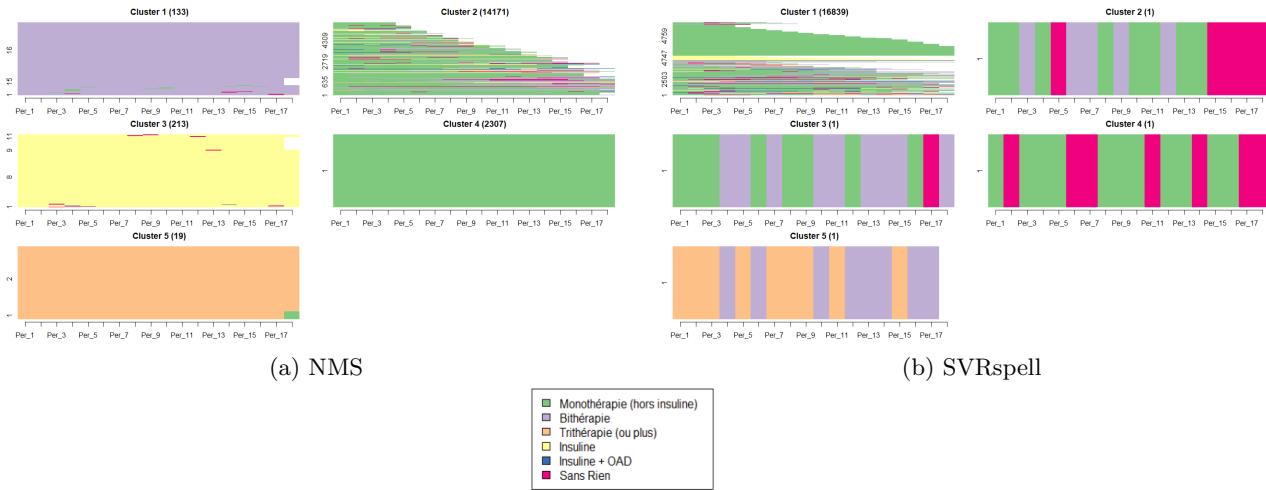


FIGURE A.26 – NMS (CAH) - 5 groupes

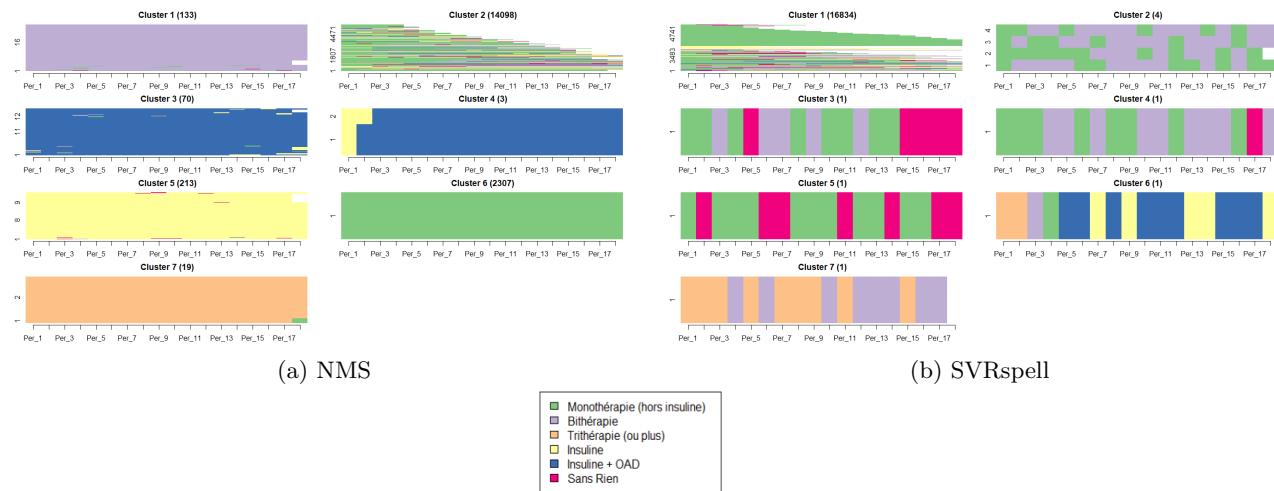


FIGURE A.27 – NMS (CAH) - 7 groupes

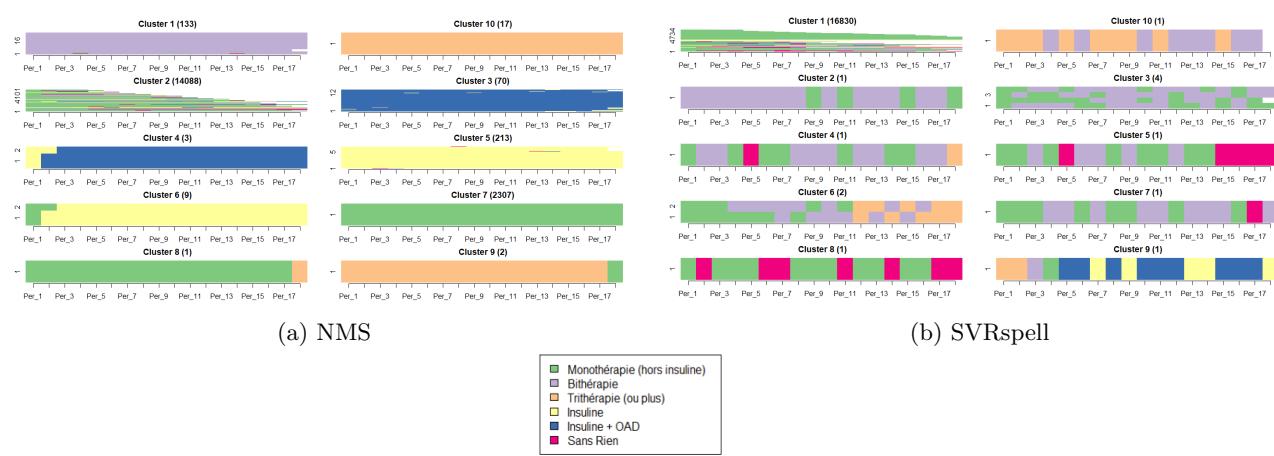


FIGURE A.28 – NMS (CAH) - 10 groupes

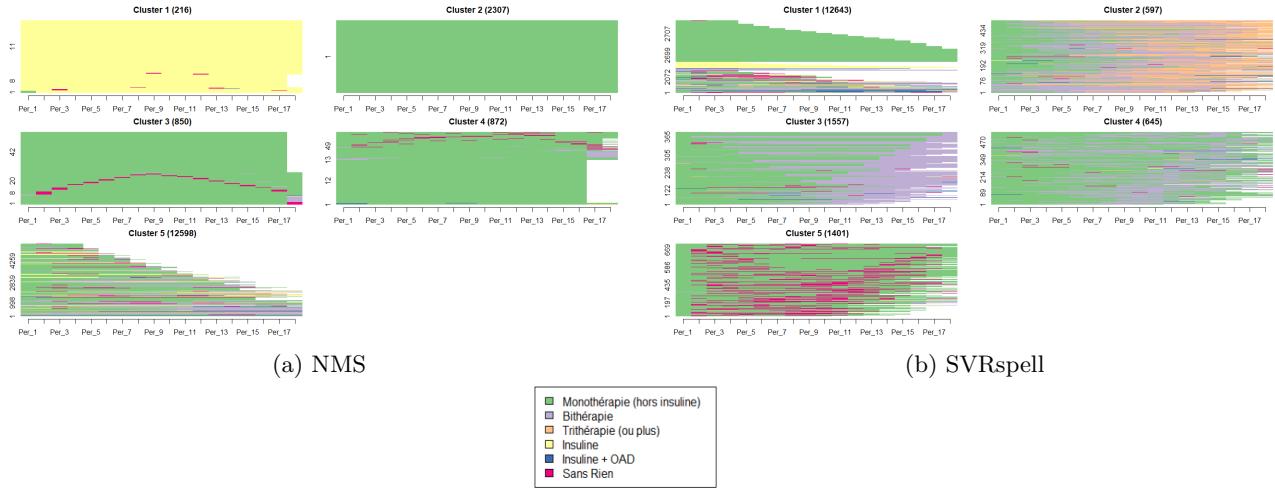


FIGURE A.29 – NMS (PAM) - 5 groupes

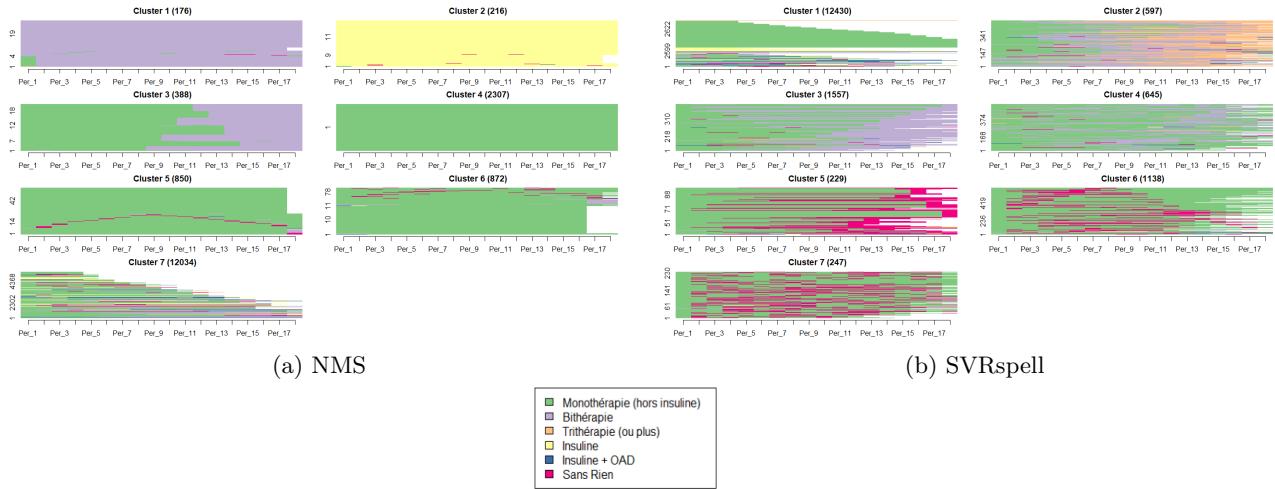


FIGURE A.30 – NMS (PAM) - 7 groupes

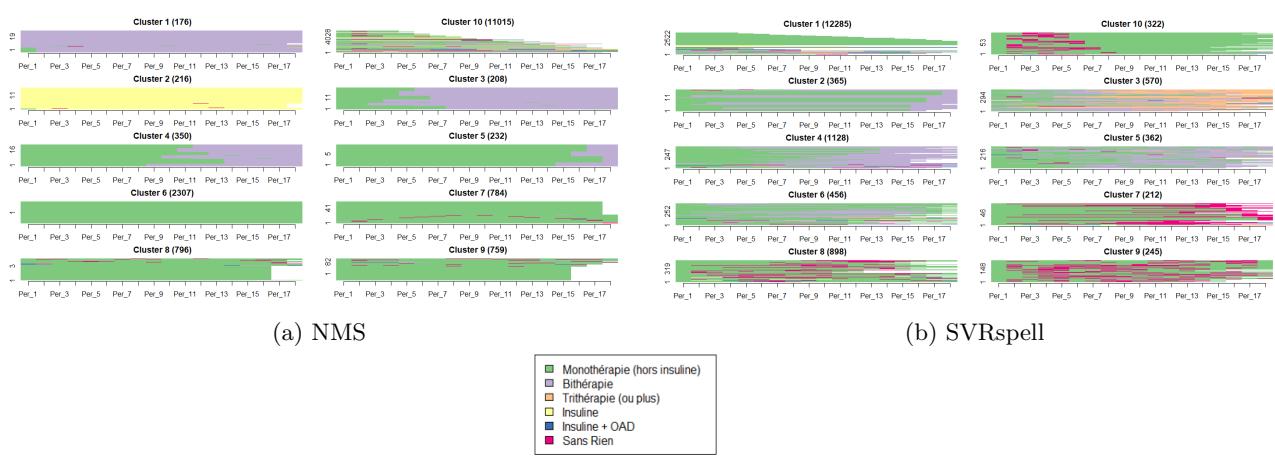


FIGURE A.31 – NMS (PAM) - 10 groupes

A.12.6 Distances entre les distributions d'états

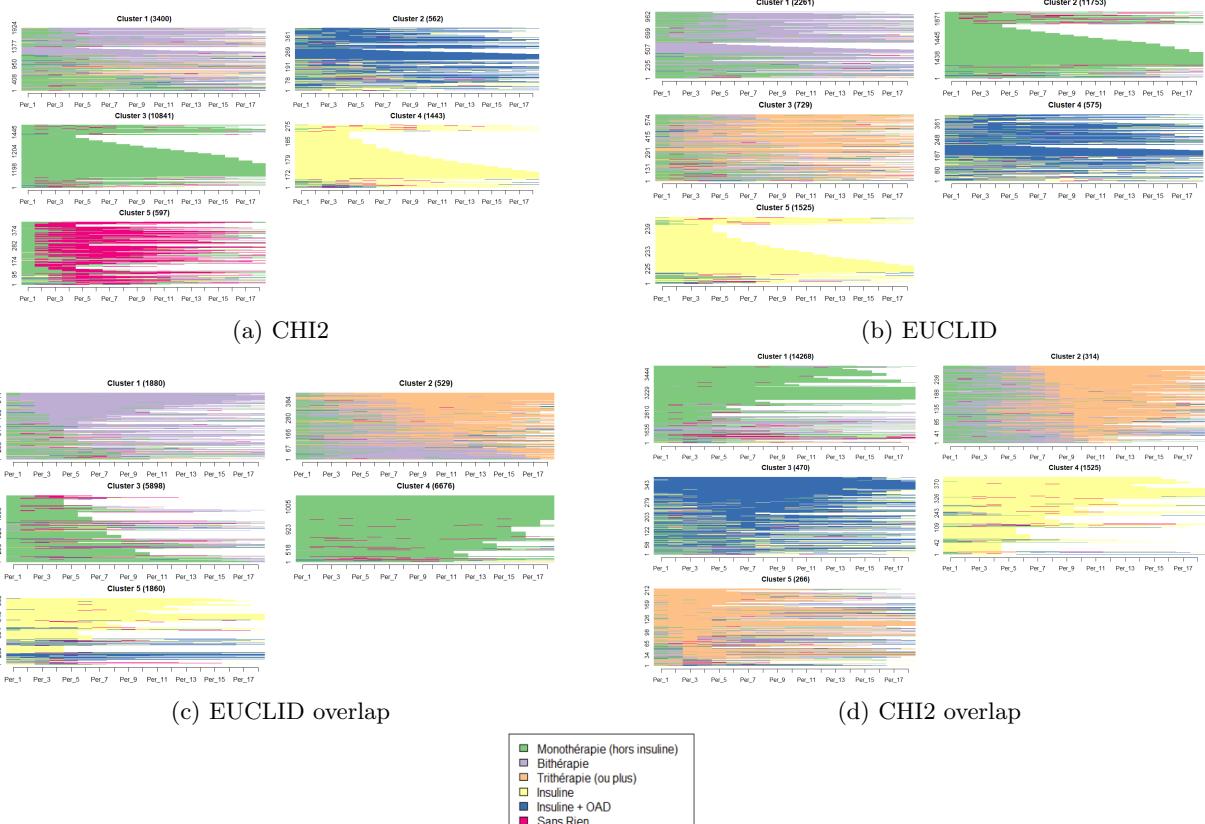


FIGURE A.32 – Distributions d'états (CAH) - 5 groupes

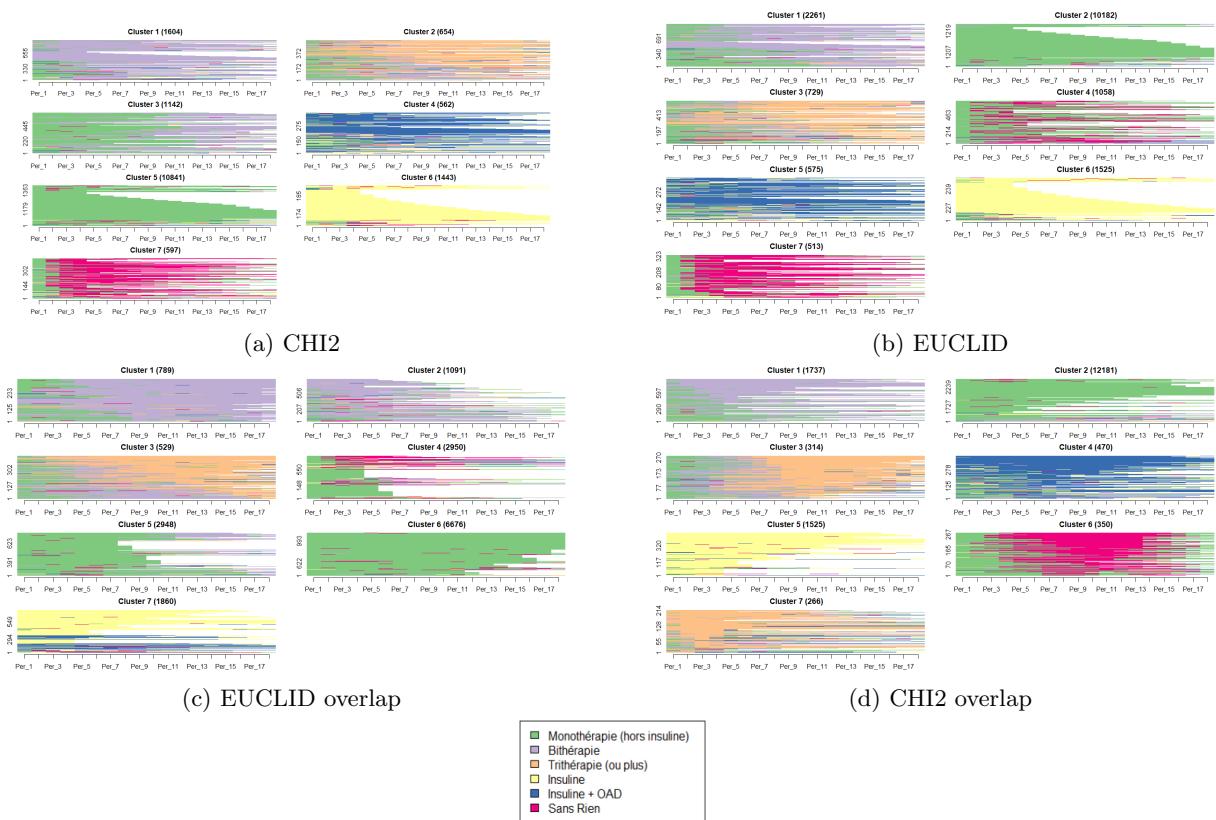


FIGURE A.33 – Distributions d'états (CAH) - 7 groupes

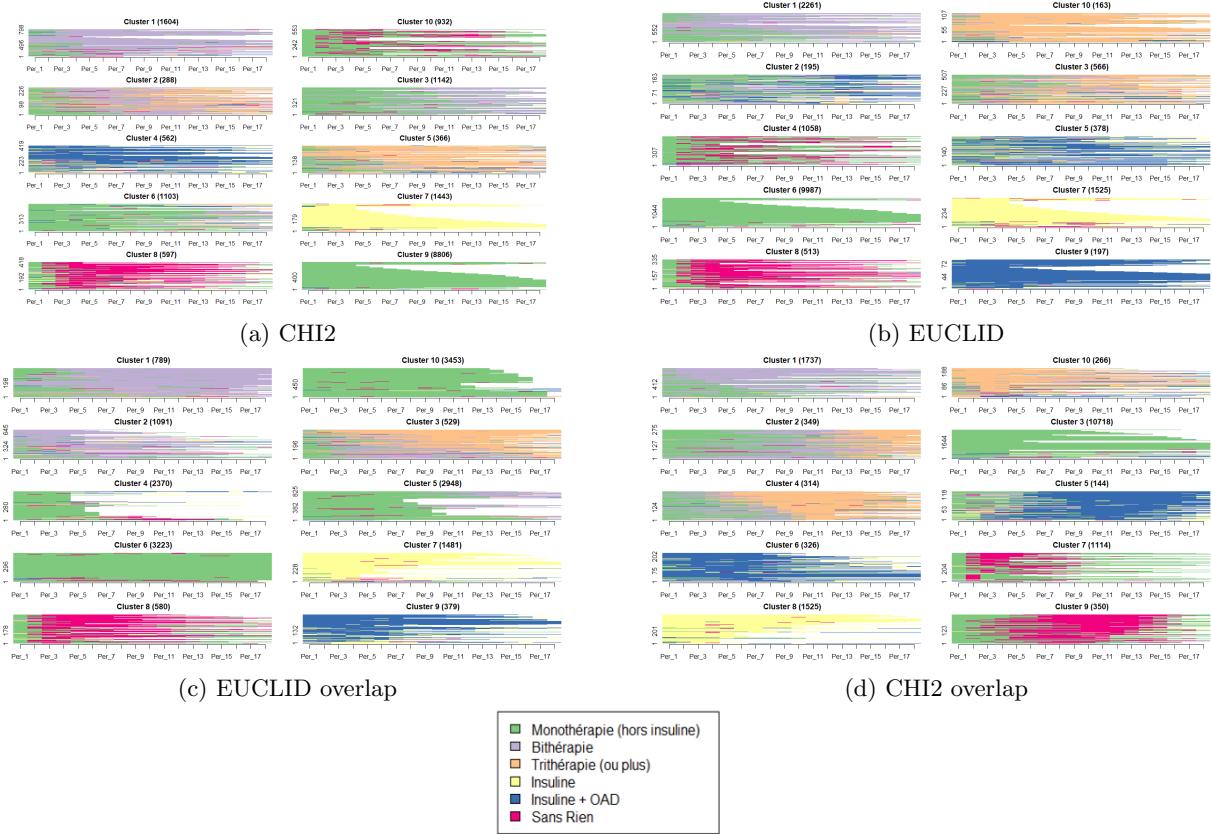


FIGURE A.34 – Distributions d'états (CAH) - 10 groupes

A.12.7 Bag-of-Words

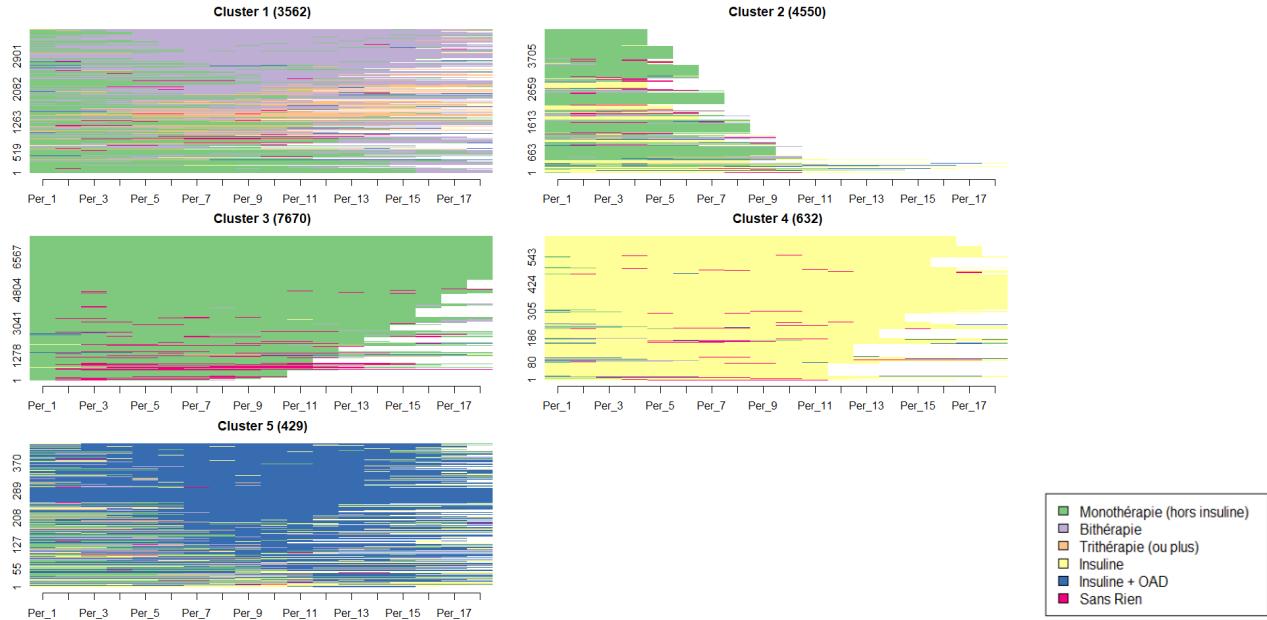


FIGURE A.35 – Bag-of-Words (CAH) - 5 groupes

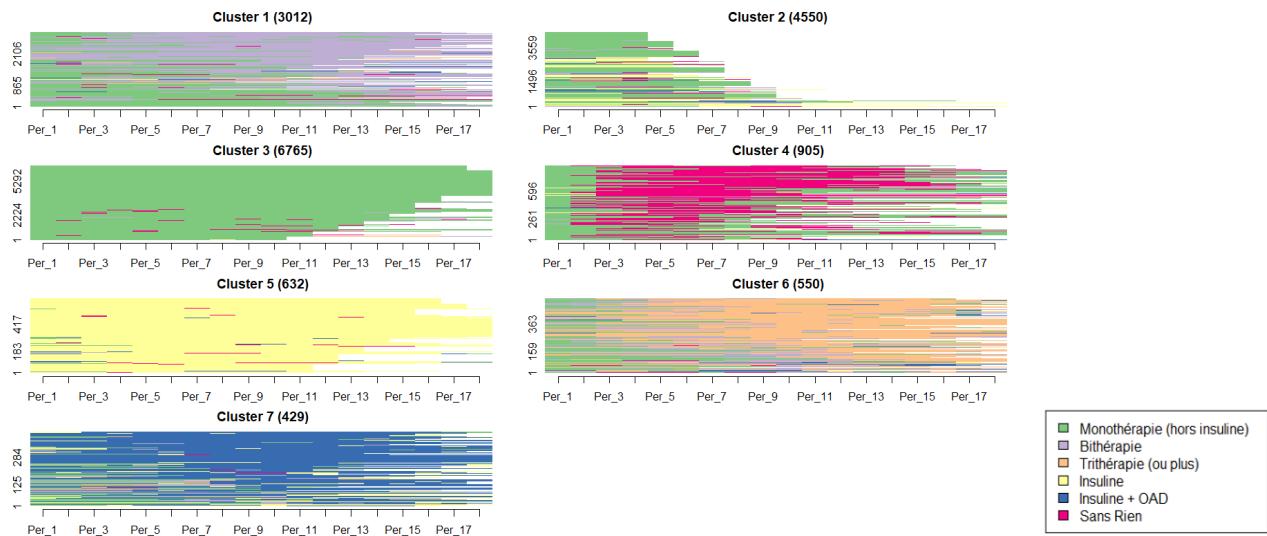


FIGURE A.36 – Bag-of-Words (CAH) - 7 groupes

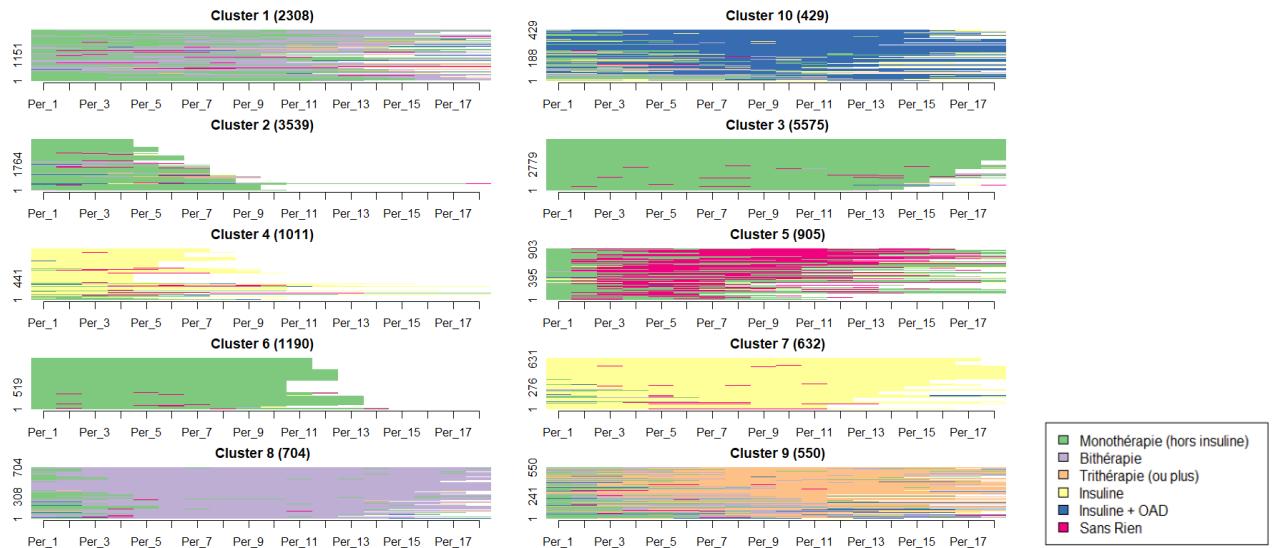


FIGURE A.37 – Bag-of-Words (CAH) - 10 groupes

A.12.8 Analyse multi-séquences

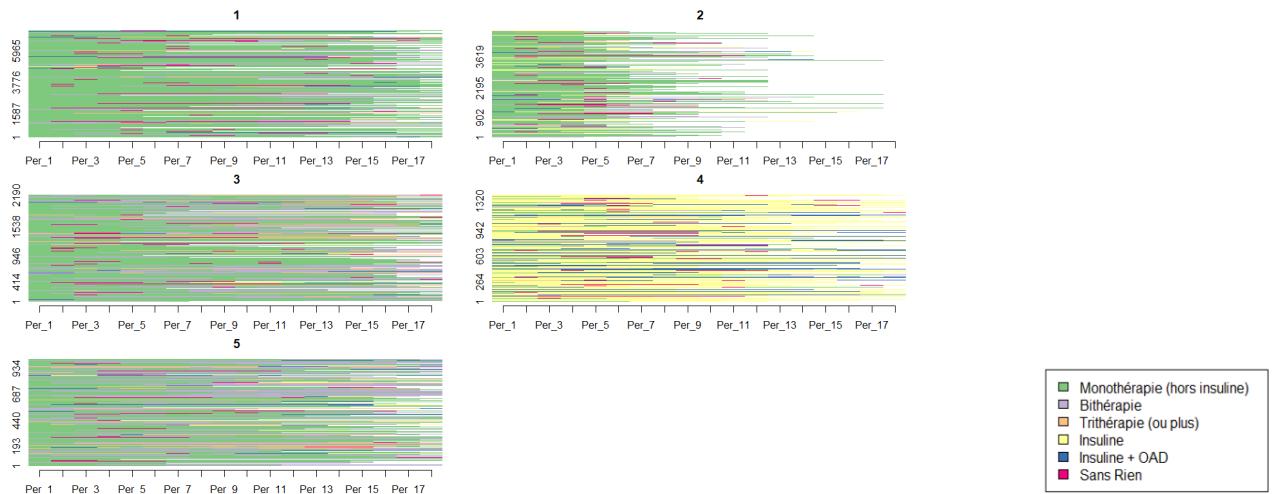


FIGURE A.38 – MCSA tapis (CAH) - 5 groupes

A.12.9 Description des groupes

Characteristic	Cluster 1 (1834), N = 1,834 ⁷	Cluster 2 (4407), N = 4,407 ⁷	Cluster 3 (2042), N = 2,042 ⁷	Cluster 4 (447), N = 447 ⁷	Cluster 5 (8113), N = 8,113 ⁷	Total, N = 16843 ⁷
Type de diabète						
Diabète Gestационnel	0 (0%)	0 (0%)	56 (2.7%)	0 (0%)	0 (0%)	56 (0.3%)
DT1	2 (0.1%)	18 (0.4%)	734 (36%)	0 (0%)	4 (<0.1%)	758 (4.5%)
DT2	1,831 (100%)	4,353 (99%)	1,118 (55%)	447 (100%)	8,038 (99%)	15,787 (94%)
Indeterminé	1 (<0.1%)	36 (0.8%)	134 (6.6%)	0 (0%)	71 (0.9%)	242 (1.4%)
Sexe						
Femme	668 (36%)	1,908 (43%)	1,058 (52%)	156 (35%)	3,454 (43%)	7,244 (43%)
Homme	1,139 (62%)	2,400 (54%)	939 (46%)	287 (64%)	4,554 (56%)	9,319 (55%)
Non connu	27 (1.5%)	99 (2.2%)	45 (2.2%)	4 (0.9%)	105 (1.3%)	280 (1.7%)
Âge						
>70	406 (22%)	1,533 (35%)	876 (43%)	52 (12%)	3,030 (37%)	5,897 (35%)
55-70	905 (49%)	1,820 (41%)	506 (25%)	236 (53%)	3,934 (48%)	7,401 (44%)
35-55	504 (27%)	946 (21%)	414 (20%)	155 (35%)	1,108 (14%)	3,127 (19%)
18-35	18 (1.0%)	99 (2.2%)	178 (8.7%)	4 (0.9%)	37 (0.5%)	336 (2.0%)
<18	1 (<0.1%)	9 (0.2%)	68 (3.3%)	0 (0%)	4 (<0.1%)	82 (0.5%)
⁷ n (%)						

FIGURE A.39 – Représentation des variables statiques dans chaque groupe (clustering avec OM TRATE)

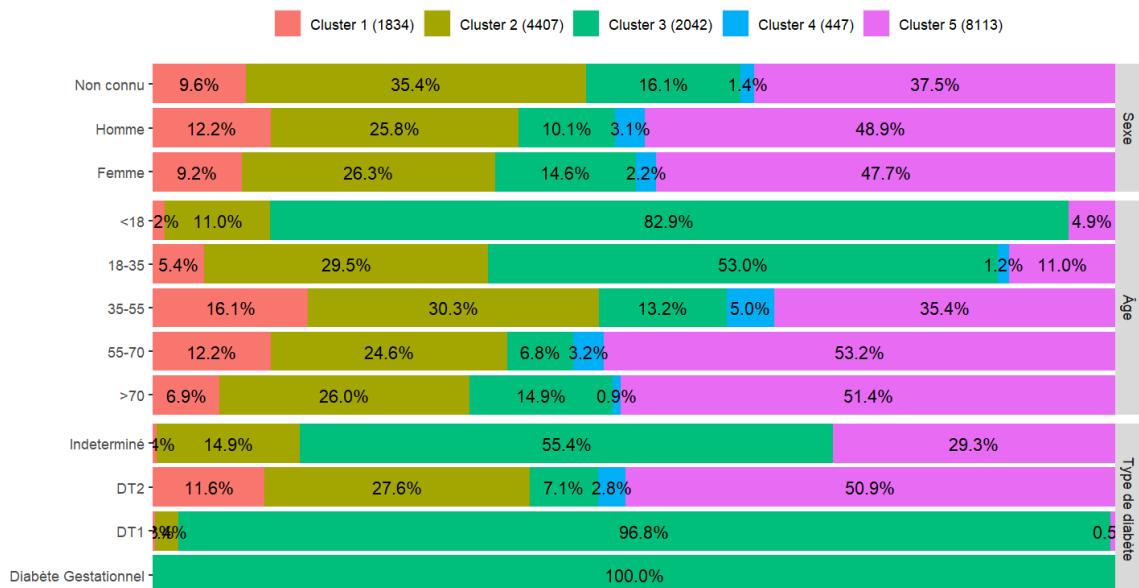


FIGURE A.40 – Représentation des variables statiques dans chaque groupe dans un graphique à barres cumulées (clustering avec OM TRATE)

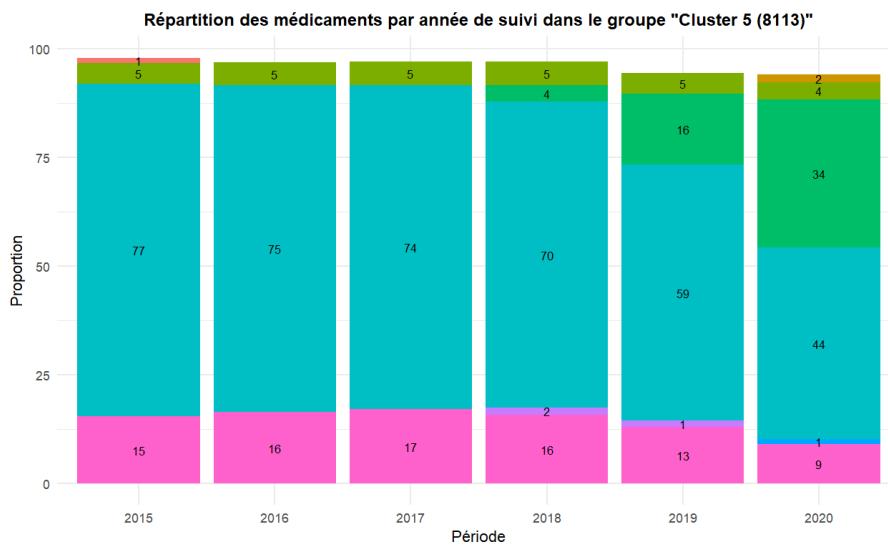
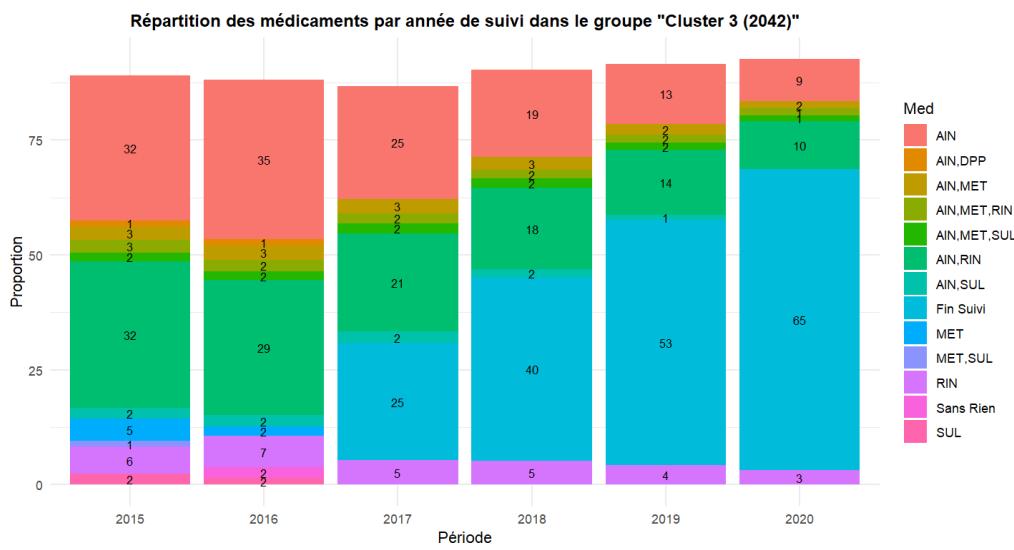
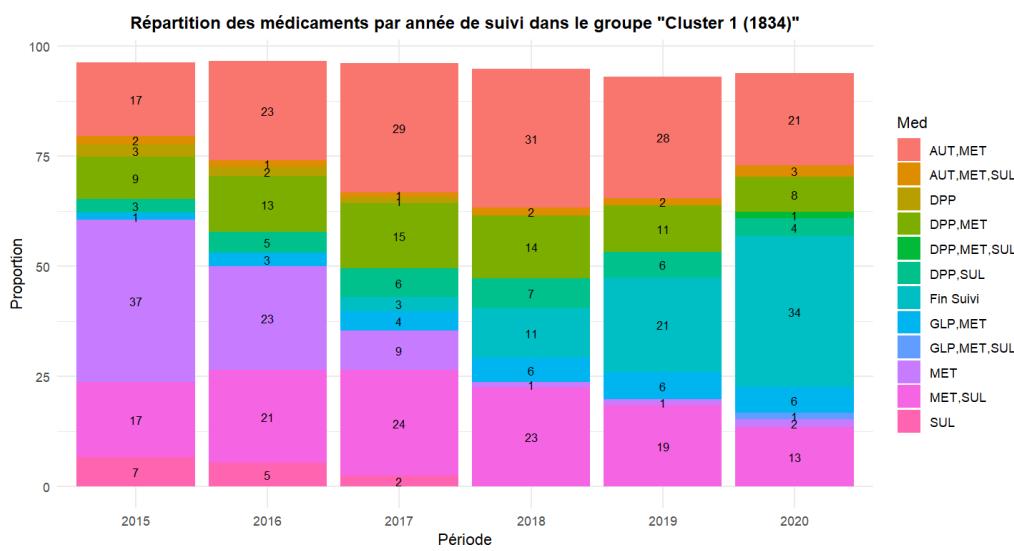


FIGURE A.41 – Proportion de chaque traitement et combinaison de traitement par cluster

A.12.10 Modèles de mélange

Modèle linéaire mixte (classification avec 4 groupes, 1000 patients)

TABLE A.19 – Modèle obtenu : $\Lambda(t) = X(t)\beta + Z(t)u_i + w_i(t)$ (classe latente de référence : 1)

	coef	Ecart-type	Wald	p-value
intercept class1 (not estimated)	0			
intercept class2	-3.983	0.501	-7.946	0
intercept class3	-6.234	0.515	-12.110	0
intercept class4	-12.389	0.884	-14.008	0
time class1	-1.035	0.150	-6.884	0
time class2	-0.287	0.023	-12.480	0
time class3	0.277	0.033	8.342	0
time class4	1.000	0.178	5.613	0

$\Lambda(t)$ est le type de thérapie suivi par le patient (monothérapie (hors insuline), bithérapie...), $Z(t)$ (effets aléatoires) et $X(t)$ (effets fixes) correspondent à la période (0 à 5) où le patient suit la thérapie (*time*).

Tapis de séquences et chronogramme

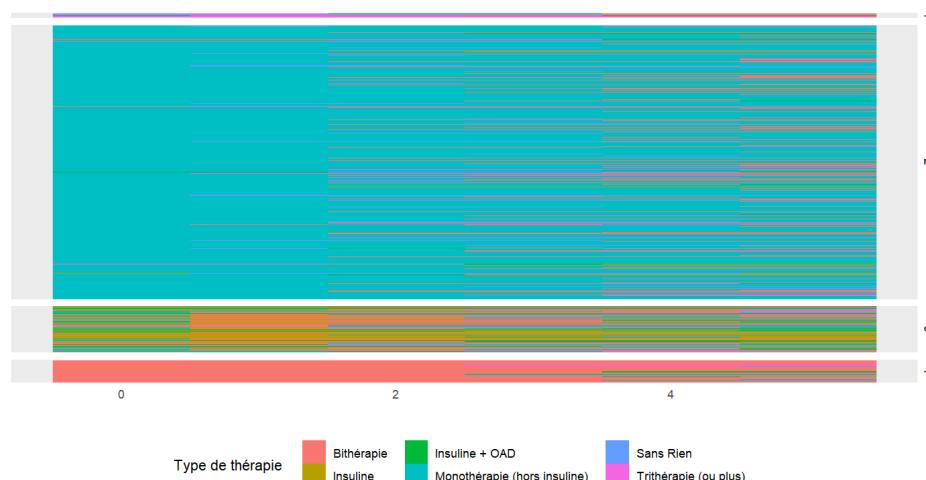


FIGURE A.42 – Tapis des 4 groupes obtenu avec le modèle mixte à classes latentes

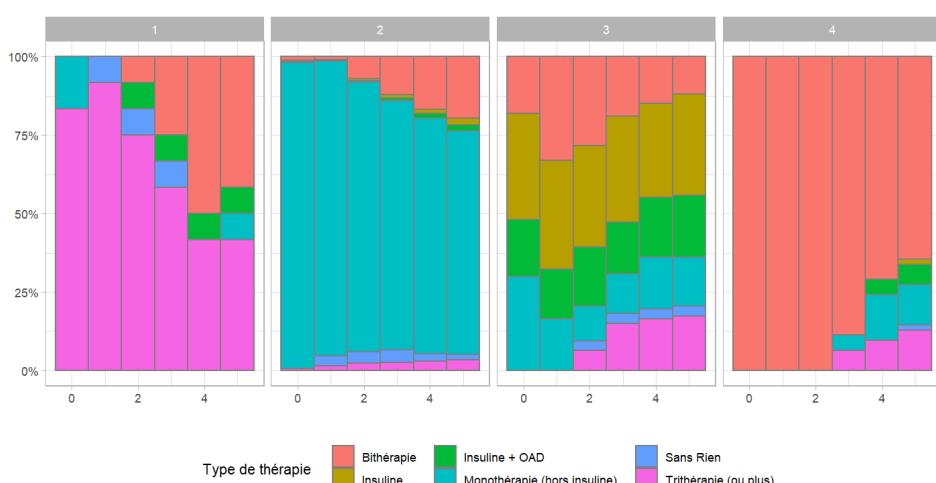


FIGURE A.43 – Chronogramme des 4 groupes obtenu avec le modèle mixte à classes latentes

Outputs du modèle mixte

TABLE A.20 – Effectifs par classe

	class1	class2	class3	class4
N	12	756	127	62
%	1.250	79	13.270	6.480

TABLE A.21 – Probabilité d'appartenir à une classe sachant que le patient a été classé dans une classe

	prob1	prob2	prob3	prob4
class1	0.941	0.059	0	0
class2	0.0005	0.962	0.038	0
class3	0	0.137	0.859	0.004
class4	0	0.0003	0.089	0.911

TABLE A.22 – Proportions de patients en fonction de leur probabilité (> 0.7 , > 0.8 ou > 0.9) d'appartenir à une classe

	class1	class2	class3	class4
prob>0.7	100	98.540	75.590	100
prob>0.8	91.670	97.880	74.020	93.550
prob>0.9	83.330	95.630	67.720	83.870