

Clustering des trajectoires thérapeutiques pour la bronchopneumopathie chronique obstructive (BPCO) : comparaison des méthodes d'encodage de séquences

Romane Péan*¹, Nina Temam¹, Marie Génin¹, Diane Vincent¹, Rachel Nadif², Sofiane Kab³, Nicolas Roche⁴, and Pauline Guilmin¹

¹Quinten Health – Quinten – 8 rue Vernier, 75017, Paris, France

²Centre de recherche en épidémiologie et santé des populations (CESP) – Assistance publique - Hôpitaux de Paris (AP-HP), Institut National de la Santé et de la Recherche Médicale – 16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France, France

³Unité Cohortes épidémiologiques en population (Constances) – Institut National de la Santé et de la Recherche Médicale - INSERM – 16 Av. Paul Vaillant Couturier, 94800 Villejuif, France

⁴Institut Cochin (IC UM3 (UMR 8104 / U1016)) – Institut National de la Santé et de la Recherche Médicale, Université Paris Cité – 22 rue Méchain, 75014 Paris, France

Cette étude a été réalisée dans le cadre d'un partenariat multiprojets avec Constances par le biais d'une convention Inserm « Inserm transfert ».

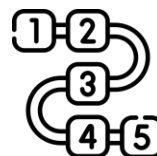


Analyse de séquences thérapeutiques : défis méthodologiques

Comprendre la prise en charge des patients



Utilisation des RWD
(séquences thérapeutiques /
événements médicaux)



Algorithmes de machine
learning classiques (e.g.
clustering)



Problématique

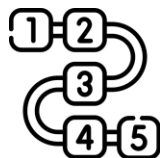
Les algorithmes de ML ne savent pas gérer la notion de
séquence ou ordre temporel entre les données

Analyse de séquences thérapeutiques : défis méthodologiques

Comprendre la prise en charge des patients



Utilisation des RWD
(séquences thérapeutiques /
événements médicaux)



Encodage des données
longitudinales en vecteurs
numériques



Algorithmes de machine
learning classiques (e.g.
clustering)



OBJECTIF : Comparer trois méthodes d'encodage de séquences combinées au clustering & évaluer leur capacité à regrouper les trajectoires thérapeutiques des patients



SeqMining



SeqToChar



Autoencoder

Overview de la méthode

1. Création de cohorte



- Cohorte Constances¹ + SNDS
- BPCO
- France
- 2008 – 2019
- 5 ans post 1er traitement BPCO²
- 1,926 patients
- Traitements de fond inhalés (mono, bi ou trithérapies)²

2. Création de séquences



Transformer les données brutes des prescriptions de traitements...



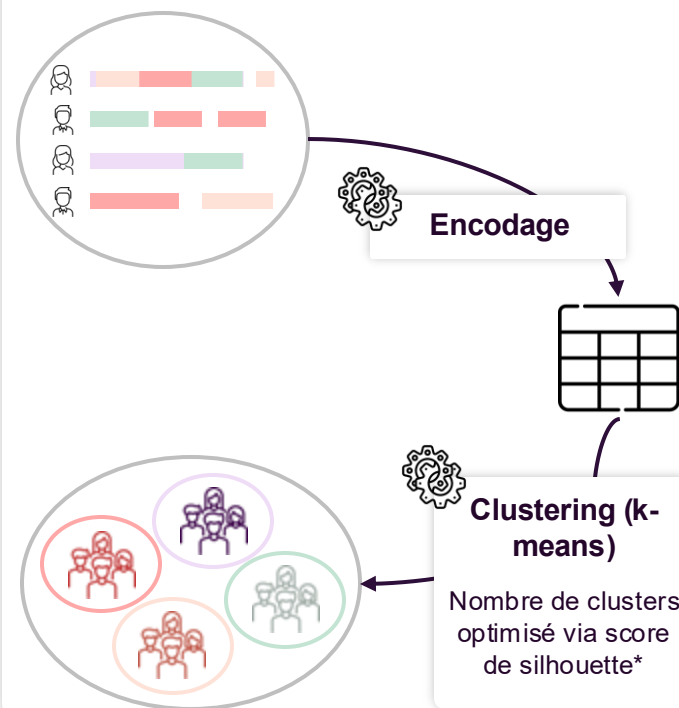
... en séquences de traitements interprétables et lisibles



Légende

- Trt 1
- Trt 2
- Trt 3
- Trt 4

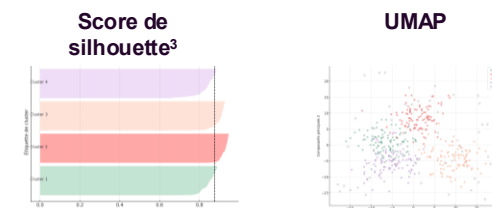
3. Modélisation



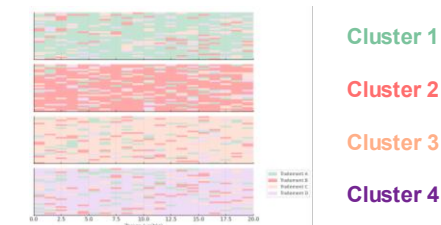
4. Evaluation



Evaluation de l'homogénéité des clusters pour chaque méthode d'embedding



Description clinique des clusters



¹ Zins, M.; Goldberg, M. The French CONSTANCES Population-Based Cohort: Design, Inclusion and Follow-Up. Eur J Epidemiol 2015, 30, 1317–1328. <https://doi.org/10.1007/s10654-015-0096-4> (Cette étude a été réalisée dans le cadre d'un partenariat multiprojets avec Constances par le biais d'une convention Inserm Inserm transfert.)

² Mirza, S.; Clay, R. D.; Koslow, M. A.; Scanlon, P. D. COPD Guidelines: A Review of the 2018 GOLD Report. Mayo Clin Proc 2018, 93 (10), 1488–1502. <https://doi.org/10.1016/j.mayocp.2018.05.026> (Corticostéroïdes inhalés (ICS), Bêta-agoniste à longue durée d'action (LABA), Anticholinergique à longue durée d'action (LAMA))

³ Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

* en utilisant la méthode du coude

Création de cohorte et séquences

Création de la cohorte

Patients BPCO inclus dans la cohort
Constances¹ via



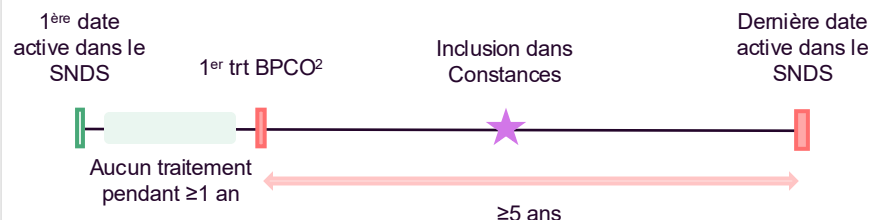
Questionnaire de
diagnostic auto
rempli

OU



Test de
Spirométrie

Patients suivis pendant au moins 5 ans après
leur premier traitement



Création des séquences de traitements

Exemple de séquence
de traitement brute



Combiner traitements
rapprochés



Fusionner les
combinaisons
similaires proches



Légende

✕ LAMA
✕ ICS

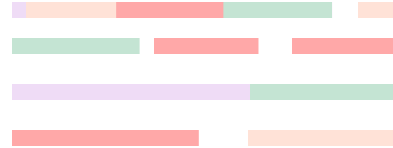
✕ LAMA/ICS
✕ LABA

¹ Zins, M.; Goldberg, M. The French CONSTANCES Population-Based Cohort: Design, Inclusion and Follow-Up. Eur J Epidemiol 2015, 30, 1317–1328. <https://doi.org/10.1007/s10654-015-0096-4> (Cette étude a été réalisée dans le cadre d'un partenariat multiprojets avec Constances par le biais d'une convention Inserm Inserm transfert.)

² Traitements de fond inhalés (mono, bi ou trithérapies) : Corticostéroïdes inhalés (ICS), Bêta-agoniste à longue durée d'action (LABA), Anticholinergique à longue durée d'action (LAMA)

3 méthodes d'encodage des séquences

Données
temporelles
brutes



Vecteurs numériques
utilisables par les algorithmes
de ML



(A) SeqMining

Extraction de sous-séquences récurrentes via l'algorithme SPADE¹.
Génération de vecteurs binaires indiquant la présence ou l'absence de motifs spécifiques.

Représentation numérique

PTID 1	1	1	1	1	1
PTID 2	0	1	0	0	0
PTID 3	1	0	0	0	0
PTID 4	0	1	1	0	0



(B) SeqToChar

Représentation des
séquences sous forme de
chaînes de caractères.

PTID 1	A - F - J - B - F
PTID 2	B - J - J
PTID 3	A - B
PTID 4	J - F

Calcul d'une matrice de similarité (Jaro²) par paires.

Représentation numérique

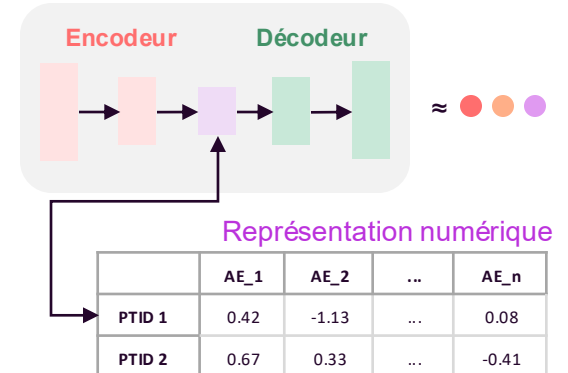
	PTID 1	PTID 2	PTID 3	PTID 4
PTID 1	1	0.63	0.43	0.97
PTID 2	0.63	1	0.24	0.33
PTID 3	0.43	0.24	1	0.12
PTID 4	0.97	0.33	0.12	1



(C) Autoencodeur

Modèle d'apprentissage profond apprenant des représentations continues abstraites (vecteurs) des séquences en prenant en compte leur durée, dans un espace latent de dimension réduite³.

PTID 1



	AE_1	AE_2	...	AE_n
PTID 1	0.42	-1.13	...	0.08
PTID 2	0.67	0.33	...	-0.41

¹ Zaki, M. J. SPADE: An Efficient Algorithm for Mining Frequent Sequences. 2021.

² Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association 1989, 84 (406), 414–420. <https://doi.org/10.1080/01621459.1989.10478785>.

³ Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; Zhou, J. Patient Subtyping via Time-Aware LSTM Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: Halifax NS Canada, 2017; pp 65–74. <https://doi.org/10.1145/3097983.3097997>.

Résultats – SeqToChar présente les clusters les plus homogènes

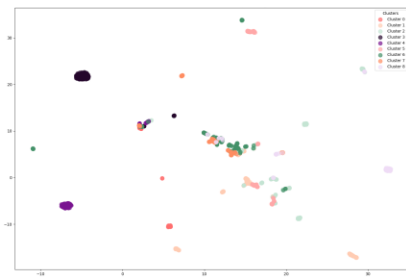
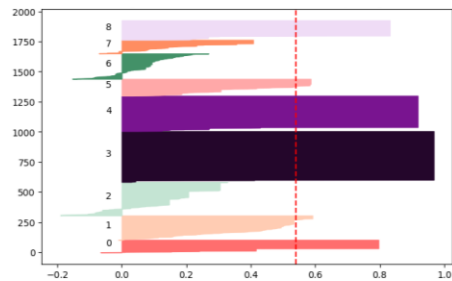


Interprétation du score de silhouette : (Excellent : 0,71-1 ; Bon : 0,51-0,7 ; Moyen : 0,26-0,5 ; Faible : 0-0,25 ; Mauvais : < 0)



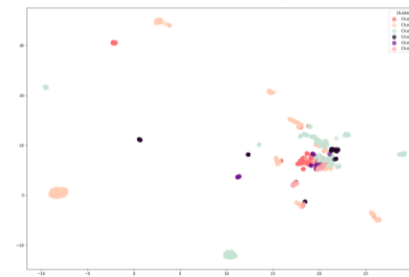
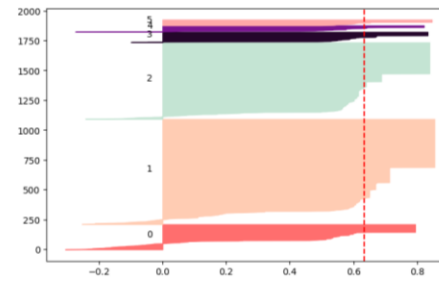
SeqMining

9 clusters
Score de silhouette³ = 0.54



SeqToChar ★

6 clusters
Score de silhouette³ = 0.63



Autoencodeur

9 clusters
Score de silhouette³ = 0.58

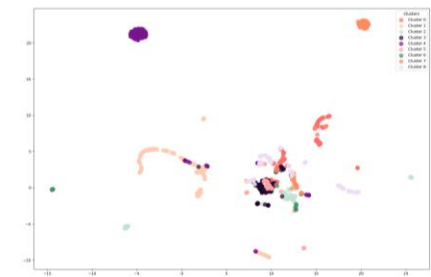
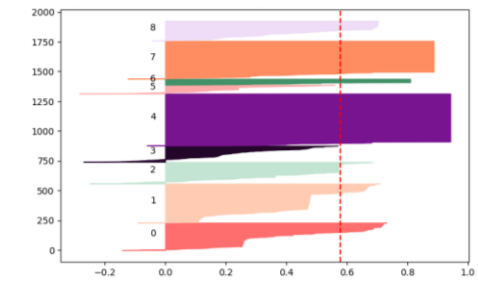


Diagramme de
silhouette¹

UMAP²

1 Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

2 McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 18, 2020. <https://doi.org/10.48550/arXiv.1802.03426>

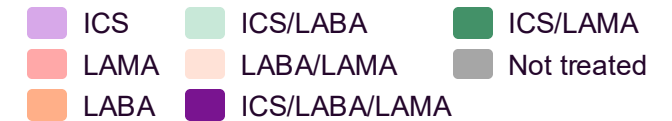
3 Kaufman & Rousseeuw (1990), Finding Groups in Data, pp. 79–81 – (Excellent : 0,71-1 ; Bon : 0,51-0,7 ; Moyen : 0,26-0,5 ; Faible : 0-0,25 ; Mauvais : < 0)

Hyperparamètres : SeqMining (min_support=0.3) ; SeqToChar (event_same_visit_strategy="alphabetical_order" ; edit_distance="jaro") ; Autoencoder object (hidden_dim=16; hidden_dim2=32; hidden_dim3=30) ; Autoencoder embeddings (epochs=10; val_split=0.2; fix_learning_rate=0.01; starter_learning_rate=None; decay_rate=0.1; decay_steps=5)

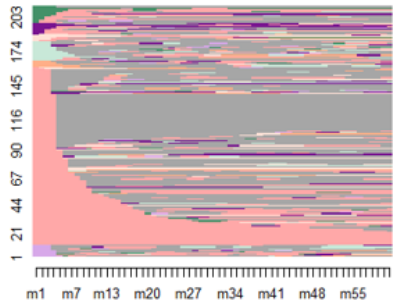
Résultats - description clinique des clusters



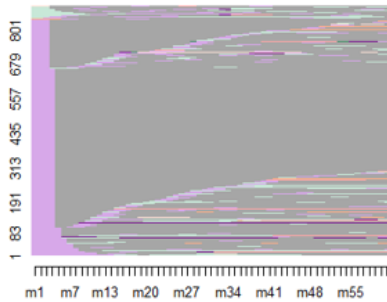
Légende



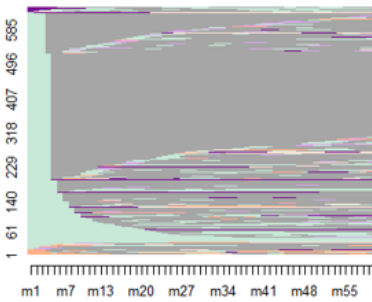
Cluster ① (N=211)



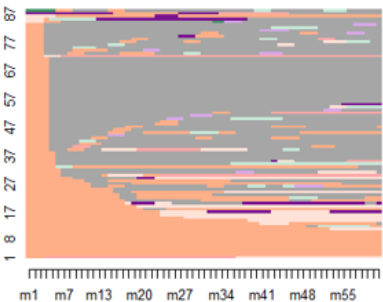
Cluster ② (N=882)



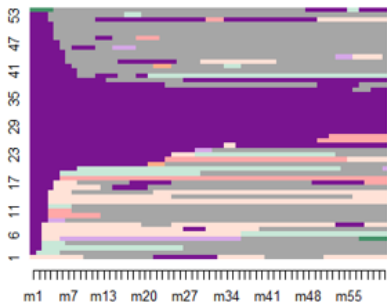
Cluster ③ (N=642)



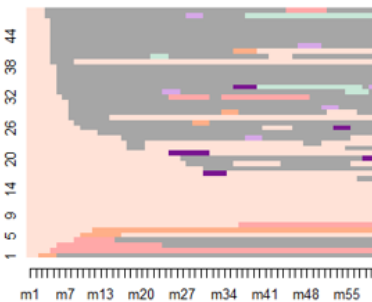
Cluster ④ (N=88)



Cluster ⑤ (N=54)



Cluster ⑥ (N=49)



• **Trajectoires de traitement hétérogènes**, nombreux arrêts précoces sans relais thérapeutique → problème d'**adhérence** connu

• **Clusters principalement regroupés selon le traitement initial**

• **Cluster ②** : majoritairement initié par une monothérapie ICS (non recommandée en usage isolé), semblant correspondre à un usage ponctuel

• **Cluster ⑤ et Cluster ⑥** : débutant par une trithérapie ICS/LABA/LAMA ou une bithérapie LABA/LAMA, présentant les trajectoires les plus durables dans le temps

• **Cluster ③** : initié par ICS/LABA, moindre persistance au traitement (mais avec relais partiel vers d'autres catégories de traitement)

• **Cluster ① et Cluster ④** : débutant par des monothérapies LABA ou LAMA présentant une faible continuité dans le temps, mais transitions fréquentes vers d'autres catégories de traitement

Corticostéroïdes inhalés (ICS), Bêta-agoniste à longue durée d'action (LABA), Anticholinergique à longue durée d'action (LAMA)

Discussion

Des résultats prometteurs

- ✓ Groupes **homogènes** sur le traitement initial
- ✓ Pas de différence majeure de performance observée entre les 3 méthodes d'encodage
- ✓ **Cohérence** (score de silhouette de 0.63) et **séparation visible** des clusters
- ✓ Reflètent **problèmes d'adhérence** connus dans la BPCO

Limites

- ~ Durée de traitement **imputée** → biais
- ~ Diagnostic de BPCO posé avant ou après le début du traitement (n.b. exclusion des patients asthmatiques)
- ~ Clustering sur le **premier événement**

Pour aller plus loin

- > Explorer d'autres méthodes de **clustering** et **d'encodage**
- > Enrichir le clustering avec des **variables additionnelles**
- > Trouver un **équilibre** entre performance statistique et pertinence médicale dans le choix de la méthode



Des questions ?

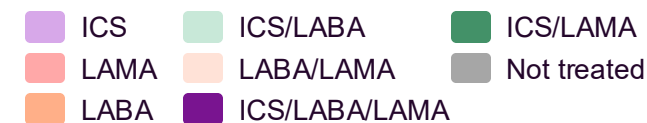
Bibliographie

1. Zins, M.; Goldberg, M. The French CONSTANCES Population-Based Cohort: Design, Inclusion and Follow-Up. *Eur J Epidemiol* 2015, 30, 1317–1328. <https://doi.org/10.1007/s10654-015-0096-4>
2. Mirza, S.; Clay, R. D.; Koslow, M. A.; Scanlon, P. D. COPD Guidelines: A Review of the 2018 GOLD Report. *Mayo Clin Proc* 2018, 93 (10), 1488–1502. <https://doi.org/10.1016/j.mayocp.2018.05.026>
3. Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Comput. Appl. Math.* 20, 53-65. *Journal of Computational and Applied Mathematics.* 20. 53-65. 10.1016/0377-0427(87)90125-7
4. Zaki, M. J. SPADE: An Efficient Algorithm for Mining Frequent Sequences. 2021.
5. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 1989, 84 (406), 414–420. <https://doi.org/10.1080/01621459.1989.10478785>
6. Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; Zhou, J. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: Halifax NS Canada, 2017; pp 65–74. <https://doi.org/10.1145/3097983.3097997>.
7. Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Comput. Appl. Math.* 20, 53-65. *Journal of Computational and Applied Mathematics.* 20. 53-65. 10.1016/0377-0427(87)90125-7.
8. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* September 18, 2020. <https://doi.org/10.48550/arXiv.1802.03426>
9. Kaufman & Rousseeuw (1990), *Finding Groups in Data*, pp. 79–81
10. Guo, Y.; Guo, S.; Jin, Z.; Kaul, S.; Gotz, D.; Cao, N. Survey on Visual Analysis of Event Sequence Data. *arXiv* June 25, 2020. <http://arxiv.org/abs/2006.14291> (accessed 2023-11- 23)

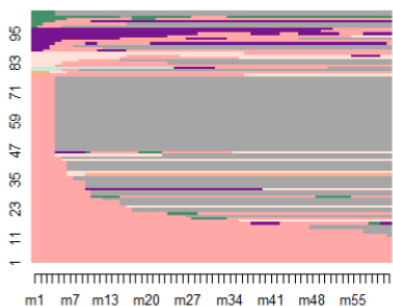
Résultats - description clinique des clusters



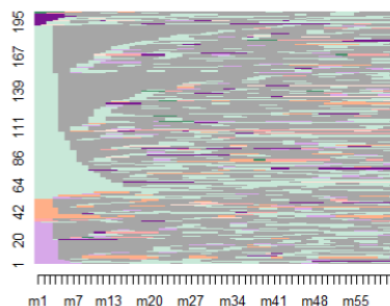
Légende



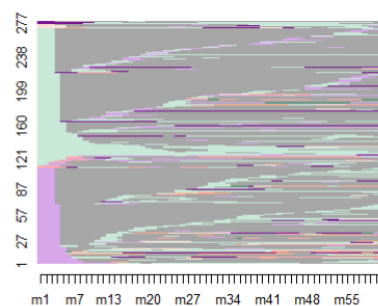
Cluster ① (N=104)



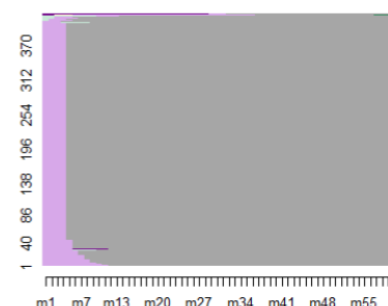
Cluster ② (N=202)



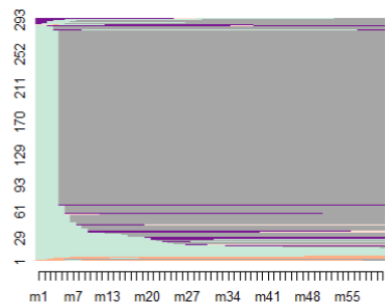
Cluster ③ (N=278)



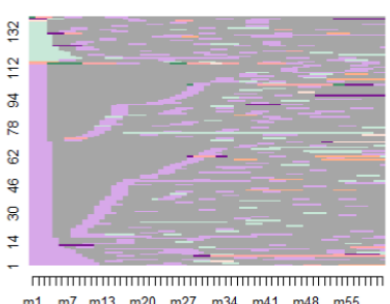
Cluster ④ (N=422)



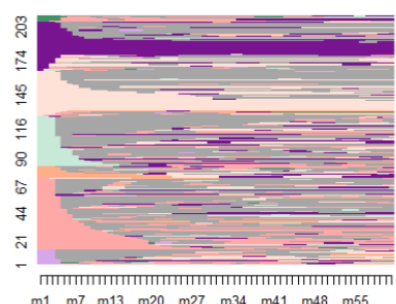
Cluster ⑤ (N=294)



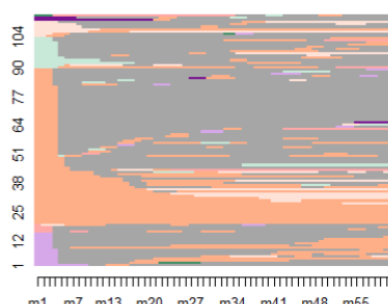
Cluster ⑥ (N=140)



Cluster ⑦ (N=49)



Cluster ⑧ (N=49)



Cluster ⑨ (N=49)

