Séminaires de modélisation statistique
Mort subite de l'adulte : Prévision de l'arrêt cardiaque

Juliet MEYNENT, Clémence BRACQ
Sous la surpevision de Younès YOUSSFI

April 2023

Contents

1 Introduction

La mort subite est un événement caractérisé par une mort inattendue et rapide. Malgré des efforts pour améliorer la prise en charge et la prévention, prédire la survenue de la mort subite reste un défi. Les facteurs de risque actuels et les méthodes de stratification du risque sont limités, et la plupart des patients ne peuvent pas bénéficier d'une évaluation personnalisée de leur risque.

Notre encadrant, Younès Youssfi, compare les données des cas de mort subite, provenant d'un centre d'expertise [1], à celles de trois groupes témoins : des patients ayant eu un syndrome coronarien aigu, des patients atteints d'insuffisance cardiaque chronique et des patients sans comorbidités.

L'objectif est d'identifier des facteurs prédictifs de survenue de la mort subite en analysant les différences entre les données des cas et celles des groupes témoins. L'objectif de la thèse de Younès Youssfi est plus précisément de construire un algorithme de prédiction de survenue de la mort subite. Cet algorithme prédictif doit permettre d'améliorer la capacité à identifier les patients à risque de mort subite et à mieux cibler les mesures préventives. Une meilleure compréhension des facteurs prédictifs de la mort subite pourrait permettre une stratification plus précise du risque et une prise en charge plus personnalisée des patients à haut risque.

La construction de cet algorithme est effectuée en deux phases. Dans la première phase, des algorithmes d'apprentissage non supervisés (clustering, modèles probabilistes, NLP) permettent d'identifier des sous-groupes de patients à risque. Dans la deuxième phase, Younès Youssfi souhaite utiliser des algorithmes de classification supervisée pour prédire la survenue de la mort subite par rapport aux groupes témoins.

Notre travail correspond à la première phase. Après une courte description de la base de données, nous effectuerons des statistiques descriptives sur la base afin de mettre en lumière les spécificités de la population victime de mort subite. Ensuite, nous mettrons en oeuvre une méthode de classification de séries temporelles. Nous utiliserons plus précisément l'algorithme des k-means avec la distance Dynamic Time Warping pour regrouper les trajectoires individuelles des patients en clusters représentant différents profils.

2 Les bases de données et leur mise en forme

Nous utiliserons deux bases de données différentes. La première concerne les patients ayant subi une mort subite ayant entre 18 et 55 ans au moment de

[1] Les cas de notre étude sont des patients inclus dans le registre du Centre d'Expertise Mort Subite, qui ont connu une mort subite entre mai 2011 et mai 2016. Le registre collecte des informations prospectives sur les événements, la prise en charge et le devenir des patients.

l'arrêt cardiaque en Ile-de-France. Elle contient diverses informations médicales sur les circonstances de l'évènement, leur sexe et leur âge.

La deuxième table correspond aux diagnostics et médicaments délivrés dans les cinq années précédent la mort subite [2]. Pour chaque soin, nous disposons du numéro d'enquêté à qui il a été administré, le nombre de jour avant la mort subite de celui-ci et la nature du médicament ou du diagnostic. La table dispose de cinq niveaux de précision pour caractériser ce soin. Les actes médicaux, en cas de délivrance médicamenteuse, suit la classfication ATC tandis que pour les diagnostics, les données reposent sur la classfication ICD-10.

Nous avons joint ces deux tables de données. Face à la difficulté que constitue une base de données aussi riche et importante, nous avons choisi de compter le nombre de soins (diagnostics et délivrance médicamenteuse). Cette approche est exploratoire et serait à affiner puisque nous n'avons pas pu différencier ces types de soins (à part dans la partie descriptive concernant justement le type de soins) et les maladies correspondantes. Il s'agit simplement de voir si oui ou non les personnes ont eu des problèmes de santé et de le quantifier grossièrement.

3 Caractéristiques des individus victimes de mort subite

3.1 Taux de survie

En moyenne sur la base de données, plus d'un tiers ont été admis vivant à l'hôpital mais seuls 11% ont suvécu à la mort subite.

3.2 Facteurs de risques sociodémographiques : le sexe, l'âge, et la précarité

3.2.1 L'âge, un facteur de risque, notamment après 40 ans

Le champ de l'étude concerne les personnes entre 18 et 55 ans. L'âge médian sur la base de données est de 46 ans, et on observe plus précisément (Figure 1) que les effectifs augmentent drastiquement avec l'âge, notamment après 40 ans.

3.2.2 Les hommes, 2.5 fois plus touchés par la mort subite que les femmes

Il y a moins de 30% de femmes dans la base de données. Les hommes sont donc 2.5 fois plus sujets à la mort subite (Figure 2).

Par ailleurs, en moyenne, les hommes ont consommé 158 soins, contre 284 pour les femmes (respectivement en termes de médiane 64 contre 163). Plus précisément, le nombre médian de soins consommés est plus important pour les femmes à tout âge, en bleu, mais il est également plus variable, sauf pour les

[2] Par souci de concision, nous parlerons de soins pour parler des médicaments et diagnostics, que nous traiterons comme des proxys de l'état de santé.

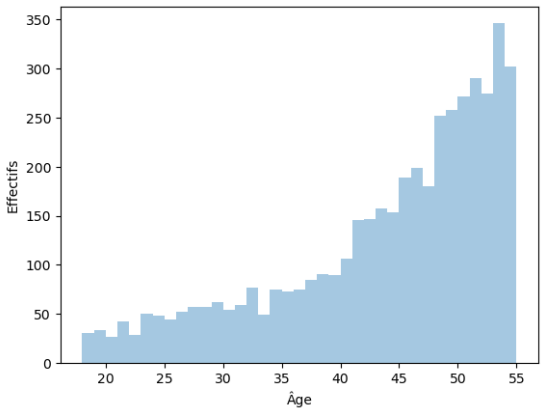

Figure 1: Effectifs par âge

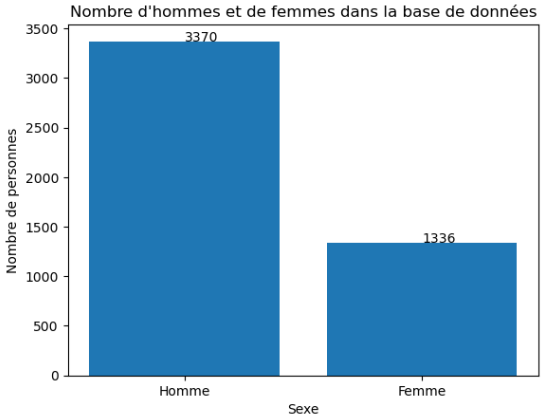

Figure 2: Effectifs par sexe

quarantenaires (Figure 3). La différence de variance est beaucoup plus importante pour les jeunes, entre 18 et 30 ans.

3.2.3 Les personnes en situation de précarité sont plus vulnérables aux morts subites

Nous utiliserons comme indicateur de précarité l'indicatrice qui signifie si un patient bénéficie ou non de la CMU. L'AME aurait également pu être un indicateur intéressant, mais très peu de personnes de la base en bénéficient (moins de 2%).

Dans la base de données, 15% des personnes sont bénéficiaires de la CMU contre 6,8% en population générale (Figure 4).

Ces inégalités de santé augmentent avec l'âge (Figure 5). Chez les plus je-

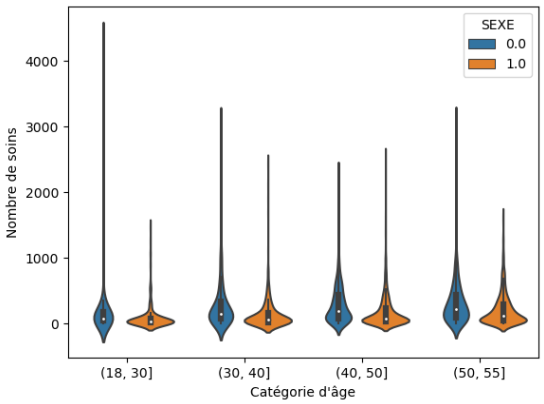

Figure 3: Comparaison entre les hommes et les femmes touchés par la mort subite au regard de leur consommation de soins et de leur catégorie d'âge

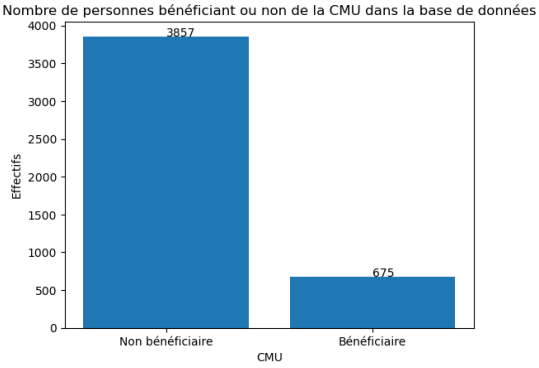

Figure 4: Effectifs des personnes bénéficiant ou non de la CMU

unes, les différences entre bénéficiaires de la CMU ou non sont assez faibles. Avec l'âge, les médianes s'écartent, avec une consommation de soins plus importante chez les personnes bénéficiaires de la CMU.

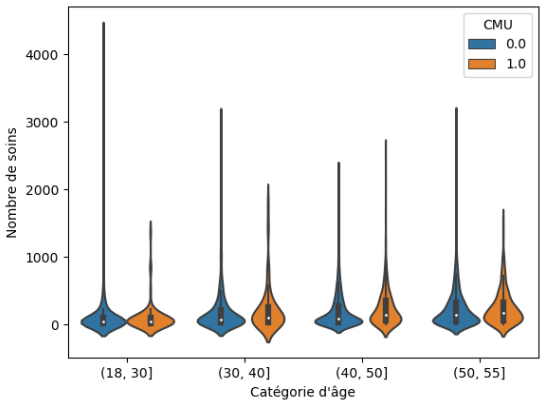

Figure 5: Relation entre bénéficiaire de la CMU, consommation de soins et catégorie d'âge chez les personnes touchées par la mort subite

3.3 Quantités et types de soins consommés

La quantité de soins consommés varie de 1 à 4302, avec une moyenne de 200 et une médiane de 93 soins.

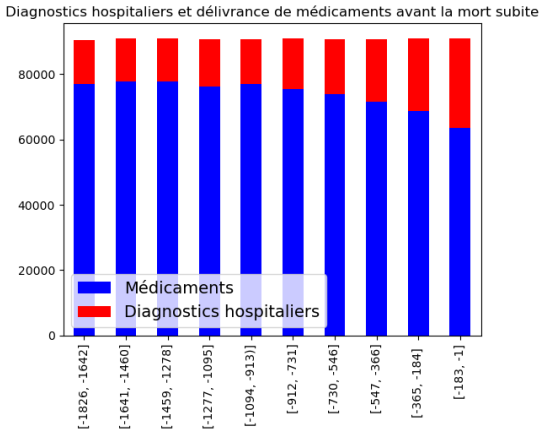

Figure 6: Volume et structure des consommations de soins par période avant la mort subite

Le volume global de consommations de soins ne semble pas varier fortement sur la période (Figure 6). A l'inverse, sa structure se modifie de manière très importante : les diagnostics représentent 15% des soins au début de la période (5 ans avant la mort subite), contre 30% à la fin de la période. Cette augmentation des diagnostics suggère que certaines pathologies se déclarent avant la mort subite et constituent ainsi des signes avant-coureurs.

4 Modélisation : Clustering de séries temporelles

4.1 Méthode de clustering et résultats

Nous allons appliquer l'algorithme des k-means pour établir des clusters (et donc des profils de patient). Cependant, ayant entre les mains une série temporelle, nous n'allons pas utiliser une distance euclidienne classique mais la "Dynamic Time Warping Distance" qui est plus adaptée au type de données qu'on traite. En effet, la distance euclidienne ne permet pas d'absorber l'aspect temporel des données. La DTW s'écrit de la manière suivante:

Soient les séries $X = (x_0, \ldots, x_n)$ et $Y = (y_0, \ldots, y_m)$,

$$DTW(X,Y) = min_\pi \sqrt{\sum_{(i,j)n\pi} d(x_i, y_j)^2} \tag{1}$$

telles que:

- d est la distance euclidienne
- où $\pi = [\pi_0, ..., \pi_K]$ est un chemin qui satisfait les propriétés suivantes.
  - π_k est un couple d'index (i_k, j_k) tels que $1 \leq i < n$ et $1 \leq j < m$
  - $\pi_0 = (0,0)$ et $\pi_K = (n-1, m-1)$
  - pour tout $k > 0$, π_k et π_{k-1} vérifient les relations suivantes:
    - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
    - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

On va utiliser une base de données qui contient le nombre de soins consommés par mois par enquêtés pour avoir des valeurs quantitatives afin de rendre possible l'implémentation du modèle.

Nous allons donc chercher à résoudre le problème suivant :

$$argmin_S \sum_{i=1}^{k} \sum_{x_j in S_i} DTW(x, \mu_i) \tag{2}$$

où

- $x = (x_1, ...x_{4706})$ correspond aux trajectoires individuelles de chaque patient
- $S = (S_1, ..., S_k)$ est une partition des trajectoires de patients
- μ_i est le barycentre des points dans S_i

Cependant, il est difficile de trouver cette partition. Ainsi, nous allons approcher la solution avec un algorithme greedy des k-means ++ . L'algorithme fonctionne de la manière suivante :

- On choisit k points qui représentent la position moyenne des partitions. Comme nous sommes dans le cas d'un k-means ++, l'initialisation se déroule de la manière suivante :
  - On choisit un premier point au hasard parmi les données.
  - Pour tous les points x que l'on a pas encore choisi, on calcule sa distance $D(x)$ avec le point initial.
  - On choisit alors un nouveau centre en utilisant une distribution de probabilité pondérée où un point x est choisi avec une probabilité proportionnelle à $D(x)^2$.
  - On recommence cette étape jusqu'à obtenir k centres.

  Puisqu'il s'agit d'un algorithme greedy, on va crée plusieurs échantillons pour choisir le meilleur groupe de k barycentres.

- On affecte à chaque observation le groupe dont la distance DTW avec sa moyenne est la plus faible. C'est à-dire qu'on crée les partitions suivantes lors de la t-ième itération de l'algorithme:

$$S_i^{(t)} = \{x_p : DTW(x_p, m_i) \leq DTW(x_p, m_j) \ \forall j, 1 \leq j \leq k\} \tag{3}$$

  Si plusieurs groupes sont possibles, l'observation est affectée à un unique groupe.

- On recalcule alors les barycentres:

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{4}$$

- On réitère l'algorithme le nombre de fois souhaitées.

Nous allons utiliser 10 itérations car cela est souvent suffisant pour assurer la convergence de l'algorithme.

Nous allons déterminer a priori le nombre de clusters le plus adéquat pour représenter les données. Pour ce faire, nous allons appliquer le modèle pour tout $K \in N \cap [2, 11]$ avec 10 itérations et calculer le coefficient de Silhouette de chacun des modèles. Ce coefficient vise à traduire l'homogénéité moyenne des clusters et est défini de la manière suivante :

$$coeff_{sil} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} s_{sil}(i) \tag{5}$$

où

- K est le nombre de clusters
- I_k l'ensemble des points appartenant au cluster k

- $s_{sil}(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$ pour chaque point i dans le cluster k tels que :
  - $a(i) = \frac{1}{|I_k|-1} \sum_{j \in I_k, i \neq j} d(x^i, x^j)$ (La distance moyenne du point à son groupe)
  - $b(i) = \min_{k' \neq k} \frac{1}{|I_{k'}|} \sum_{i' \in I_{k'}} d(x^i, x^{i'})$ (La distance moyenne du point à son groupe voisin)

Ainsi, plus les clusters sont homogènes entre eux, plus le coefficient est élevé. Ainsi, nous choisissons de créer le nombre de clusters associé au coeffecient le plus élevé. Générer 2 clusters est la solution la plus optimale.

Graphiquement, nous obtenons alors les clusters suivants représentés dans la figure 7.

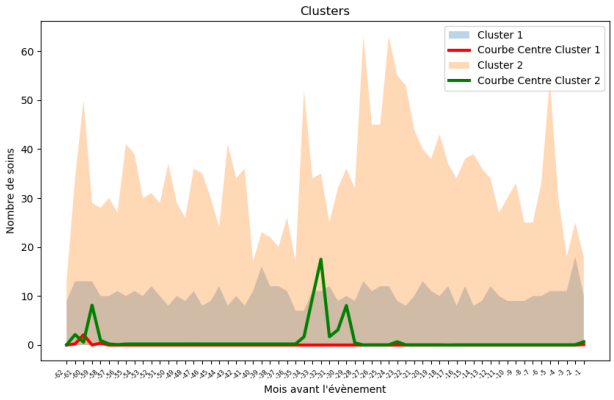

Figure 7: Clustering

On voit que nos clusters se distinguent par un groupe de patients qui n'ont presque jamais recours à des soins ou à des diagnostics. L'autre groupe se caractérise par une consommation de soins faibles en moyenne, même si le groupe contient dans ses valeurs extrêmes des patients pouvant avoir une consommation élevée. On remarque d'ailleurs dans ce groupe un pic de la moyenne du recours à des soins ou à des diagnostics dans les deux années et demi précédent l'arrêt cardiaque avec un centre des données non pas à 0 soin mais à une quinzaine de soins. Cependant, cette dernière interprétation est discutable car l'aspect aléatoire de l'algorithme amène à un aléa sur la valeur des barycentres.

Afin de mieux comprendre ce qui peut différencier ces deux clusters, nous allons effectuer une analyse a posteriori des variables socio-démographiques au sein des clusters.

4.2 Analyse a posteriori des caractéristiques socio-démographiques des clusters

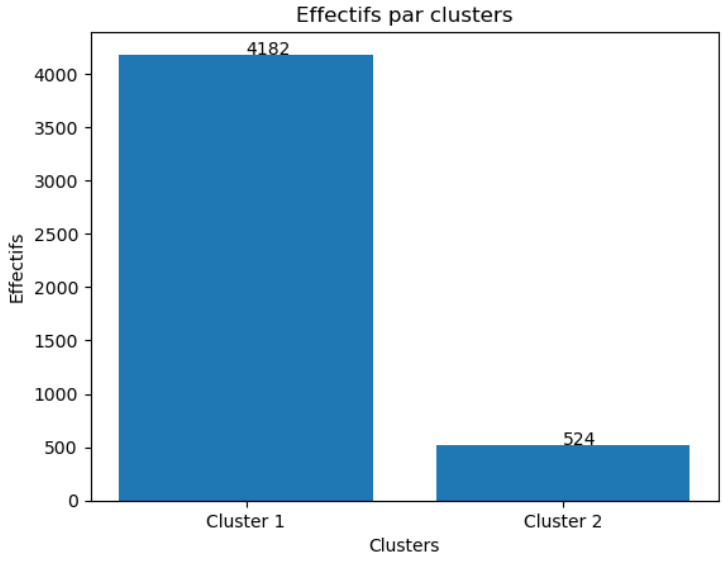

Figure 8: Effectifs par cluster

Le cluster 2, qui regroupe les personnes ayant un nombre important de consommation de soins, est largement minoritaire et ne représente qu'un peu plus de 10% de la population étudiée.

Dans le cluster 1, le groupe dans lequel les individus ont peu recours aux soins et/ou peu de problèmes de santé, les femmes sont moins représentées. Elles représentent 26% contre 45% dans le cluster 2 (graphique 9).

Les personnes les plus âgées de l'échantillon sont également moins nombreuses que dans l'échantillon initial, alors que les plus jeunes sont surreprésentées. Les 18-30 ans représentent ainsi 13% du cluster 1 contre 4% du cluster 2, alors que les 50-55 ans ne représentent que 31% du cluster 1 contre 41% dans le cluster 2 (graphique 10).

Ces deux résultats paraissent cohérents avec la modélisation : les femmes ont plus souvent recours à des soins que les hommes, notamment pour la prévention. Les personnes jeunes ont moins souvent de médicaments ou de diagnostics de pathologies que les personnes plus âgées.

Les personnes du cluster 1 arrivent plus souvent vivants à l'hôpital et survivent plus souvent que ceux du cluster 2. 39% d'entre eux arrivent vivant à l'hôpital, contre 26% dans le cluster 2 (graphique 10), et 12% survivent dans le cluster 1, contre 5 dans le cluster 2 (graphique 11).

On peut supposer que le taux de survie du groupe 1 est lié à l'âge moyen des personnes qui la compose. On peut le vérifier par une régression linéaire.

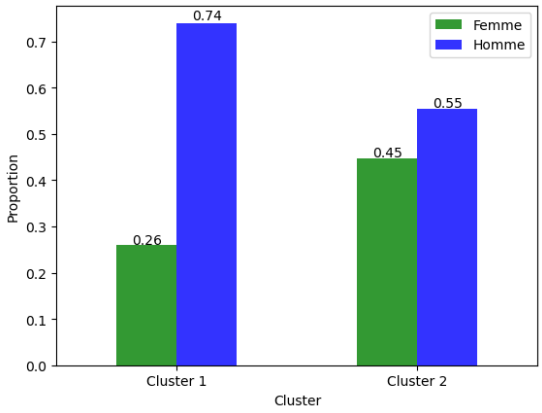

Figure 9: Proportions d'hommes et de femmes dans les clusters

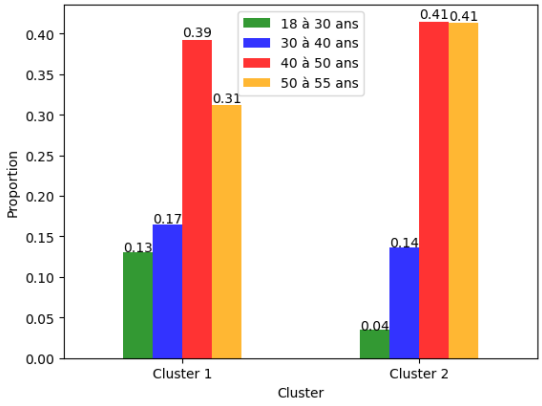

Figure 10: Proportions par catégorie d'âge dans les clusters

Quand on régresse la survie sur les clusters et l'âge (variable continue ici), on obtient un coefficient non-significatif pour l'âge et un coefficient de 0.07 pour le cluster. Appartenir au premier cluster augmente donc la probabilité de survie de 7% en contrôlant l'effet de l'âge.

Enfin, on remarque que les bénéficiaires de la CMU sont relativement plus nombreux dans le cluster 2 (16% contre 14% dans le cluster 1, graphique 13). On peut supposer que des personnes malades ont plus de chances de demander la CMU pour justement consommer des soins. Du fait d'une autre pathologie, leurs chances de survie sont probablement moins importantes.

Les résultats obtenus ainsi avec deux clusters correspondent aux caractéristiques observées en population générale : par exemple, les femmes et les personnes âgées tendent à consommer plus de soins que les hommes et les jeunes.

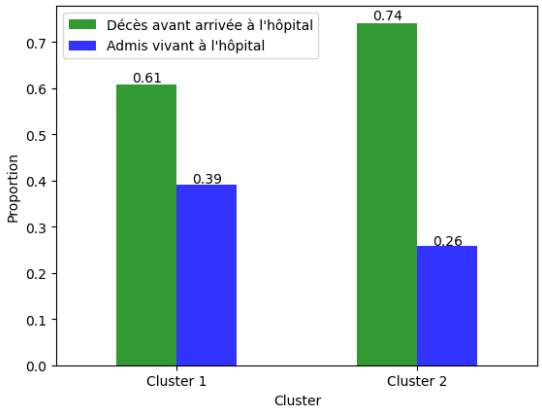

Figure 11: Proportions de personnes décédées avant l'arrivée à l'hôpital par cluster

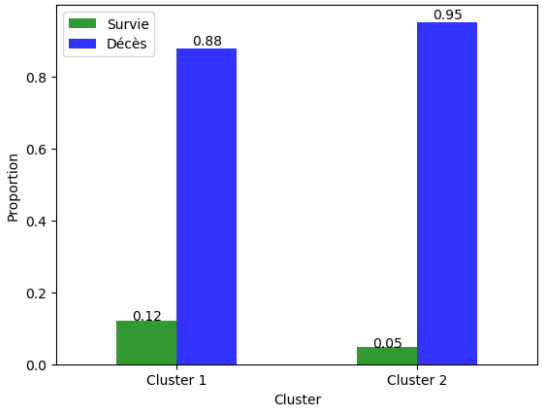

Figure 12: Proportions de personnes décédées par cluster

4.3 Résultats avec une deuxième modélisation et trois clusters

Pour aller plus loin et essayer de comprendre les caractéristiques spécifiques des personnes qui sont victimes de mort subite, nous avons voulu implémenter un modèle avec plus de deux clusters. D'après le critère de maximisation du coefficient de Silhouette, le nombre de trois clusters était optimal.

Nous obtenons ainsi les clusters représentés Figure 14.

On constate que le cluster 1 correspond aux personnes avec une forte consommation de soins, notamment trois ans et demi avant la mort subite. Le cluster 2 est caractérisé par une consommation moindre, avec un pic 2 ans avant

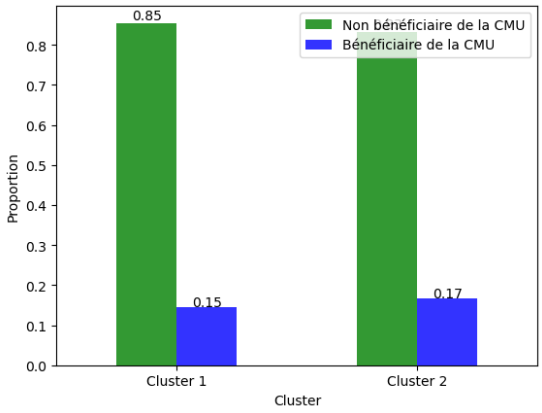

Figure 13: Proportions de personnes bénéficiant de la CMU par cluster

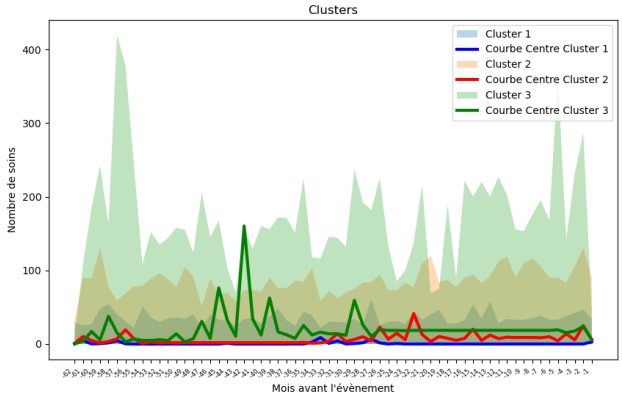

Figure 14: Clustering à 3 clusters

l'événement 3 fois moins élevé que le pic du premier groupe trois ans et demi avant la mort subite. Enfin, le dernier cluster est constitué d'une population qui ne recourt pas ou peu aux soins.

On remarque ici que le cluster 3 ne représente que 15 individus, outliers qui ont consommé un très grand nombre de soins. La modélisation ne permet pas de modifier radicalement l'analyse précédente sur les deux premiers clusters mais seulement d'analyser les caractéristiques spécifiques de ce dernier cluster [3].

Le cluster 3 regroupe beaucoup de personnes jeunes : un tiers ont moins de 30 ans et 60% moins de 40 ans. Ce n'est pas étonnant, on retrouve ici les tendances données par les deux autres clusters, également vraies en population

[3]Les graphiques correspondant à ceux réalisés pour la première modélisation sont présentés en annexe 6.3.

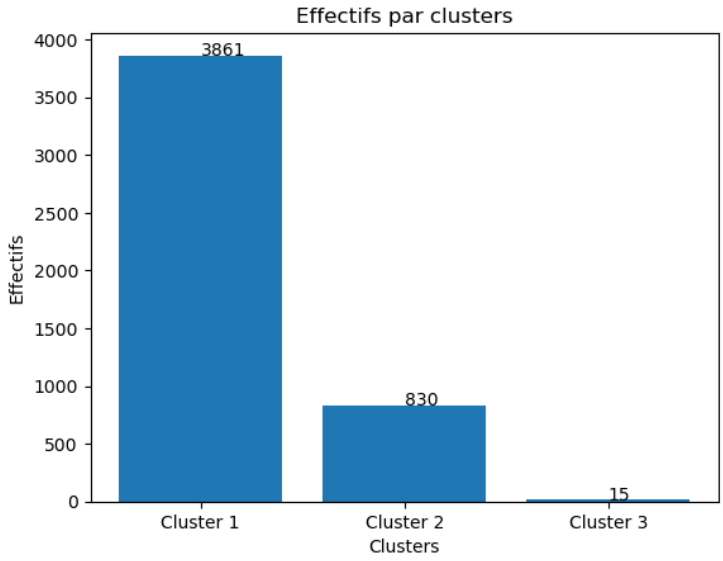

Figure 15: Effectifs par cluster

générale, selon lesquelles les jeunes consomment moins de soins.

Il regroupe également 2 fois plus de femmes que dans la base de données générale, et moitié moins de bénéficiaires de la CMU. Ces caractéristiques sont plus étonnantes et font la spécificité de ces outliers.

Enfin, ce cluster a un taux de survie équivalent au cluster 1, c'est-à-dire le groupe consommant le plus de soins et ayant le taux de survie le plus important, alors même que les individus qui la composent sont moins souvent admis vivant qu'en moyenne (de manière comparable au cluster 2).

5 Conclusion

À travers l'analyse de bases de données médicales, nous avons pu déterminer plusieurs tendances et observations importantes pour prédire l'arrêt cardiaque chez les adultes.

Tout d'abord, il est apparu que l'âge est un facteur de risque majeur, avec une augmentation significative des cas d'arrêt cardiaque après 40 ans. De plus, les hommes étaient plus touchés par la mort subite que les femmes, avec une proportion de femmes beaucoup plus faible dans notre échantillon. Les différences de consommation de soins entre les sexes étaient également notables, avec une consommation médiane plus élevée chez les femmes, en particulier dans la tranche d'âge des quarantenaires.

En ce qui concerne les facteurs socio-économiques, nous avons constaté que les personnes en situation de précarité, bénéficiaires de la Couverture Maladie Universelle (CMU), étaient plus vulnérables aux morts subites. Ces inégalités de santé se sont accentuées avec l'âge, les personnes bénéficiaires de la CMU ayant une consommation de soins plus élevée que les autres.

En examinant la quantité et les types de soins consommés, nous avons observé une variabilité importante, avec un nombre moyen de soins consommés de 200 et une médiane de 93. De plus, la structure des consommations de soins a évolué au fil du temps, avec une augmentation significative des diagnostics dans les cinq années précédant l'arrêt cardiaque, suggérant l'existence de signes avant-coureurs de la mort subite.

Enfin, nous avons abordé la modélisation des données en utilisant une méthode de clustering de séries temporelles. En appliquant l'algorithme des k-means avec la distance Dynamic Time Warping, nous avons pu regrouper les trajectoires individuelles des patients en clusters représentant différents profils. Nous avons également utilisé le coefficient de Silhouette pour évaluer la qualité des clusters, ce qui nous a permis de déterminer le nombre optimal de clusters. L'implémentation et la compréhension de ce type de modélisation complexe constitue une réussite en soi. Cependant, les résultats de cette clusterisation sont décevants, puisqu'ils ne nous apprennent que peu sur les signes avant-coureurs de la mort subite. La description d'un cluster d'outliers en dernière partie permet de mettre en lumière des individus atypiques : malgré leur faible taux d'admission vivant à l'hôpital, ce groupe a un taux important de survie par rapport aux autres.

Cette étude serait à approfondir par des travaux différenciant les différents types de soins et diagnostics. Les clusters obtenus seraient ainsi plus complexes et permettraient de mieux comprendre le comportement du groupe d'outliers.

6 Annexe : graphiques pour la deuxième modélisation

6.1 Age

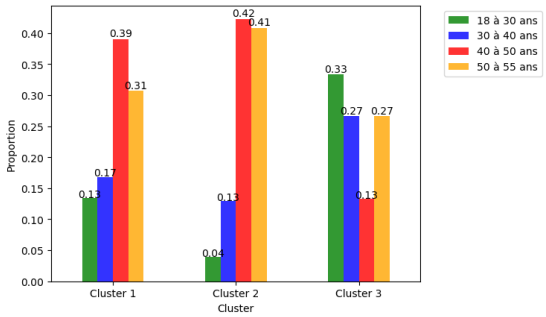

Figure 16: Proportions par catégorie d'âge dans les clusters

6.2 CMU

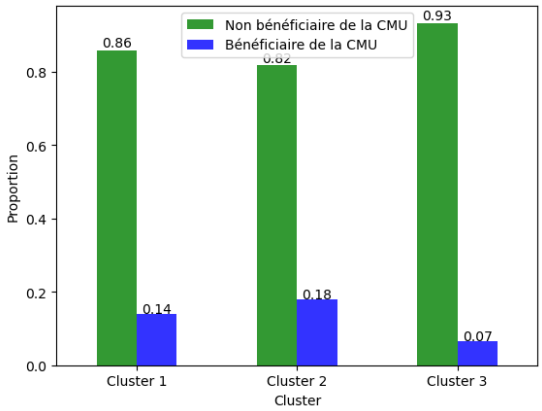

Figure 17: Proportions de personnes bénéficiant de la CMU par cluster

6.3 Sexe

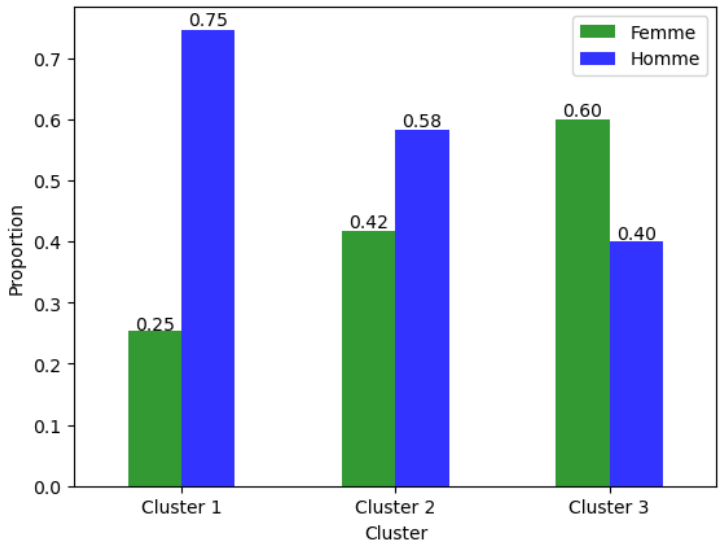

Figure 18: Proportions d'hommes et de femmes par cluster

6.4 Décès

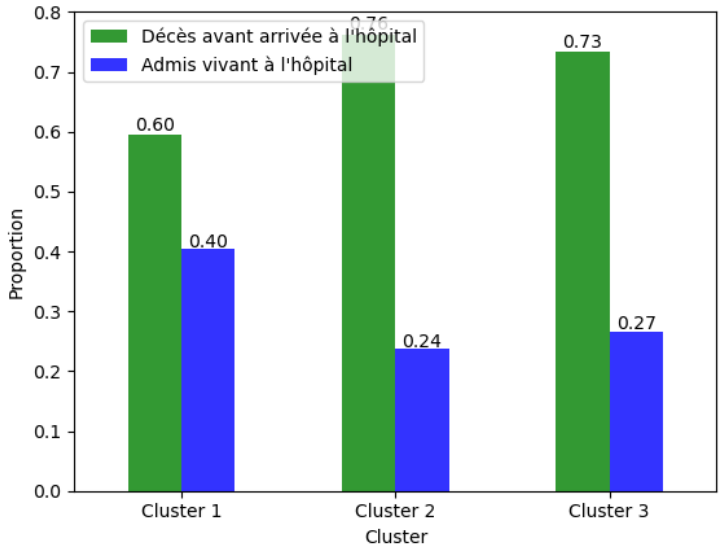

Figure 19: Proportions de personnes décédées avant l'arrivée à l'hôpital par cluster

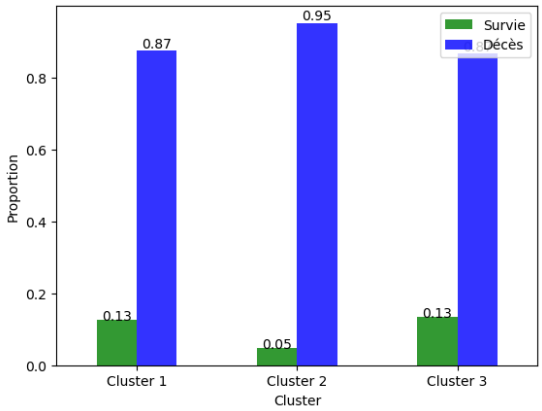

Figure 20: Proportions de personnes décédées par cluster