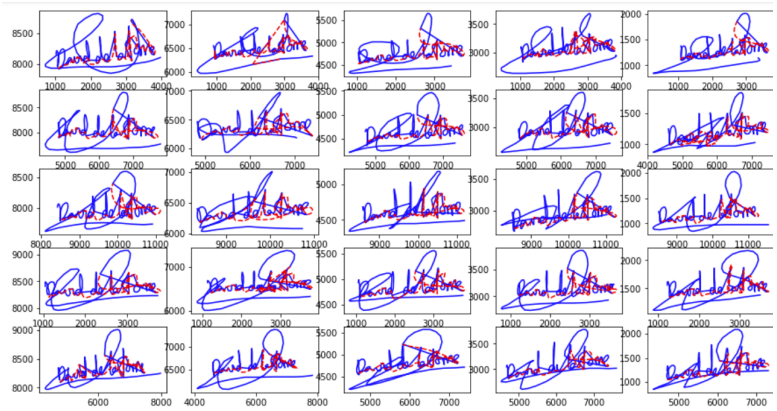
Juliette OULD-AKLOUCHE - Armand ROBERGE - Emilien VIMONT

Analyse de signatures manuscrites pour la vérification d'identité

Table des matières :

1. Introduction

Ce rapport porte sur l'étude de signatures manuscrites acquises sur des outils numériques. Pour cela, la base données MCYT-100 est utilisée, celle-ci réunit un ensemble de 2500 signatures pour 100 individus (chaque individu a rentré 25 fois sa signature), afin de permettre de catégoriser les individus ainsi que les signatures dans différentes classes. La complexité des 2500 signatures est également intégrée à cette base de données et a été calculée avec le modèle GMM (Gaussian Mixture Model) couplé à la mesure d'entropie différentielle. Le calcul de la complexité pour chaque signature est disponible pour différents nombres de gaussiennes NG : 4, 8, 24.
A l'aide de Pandas, on visualise les signatures de chaque individu :



Affichage des 25 signatures de l'individu 2, les points où la pression du stylo vaut 0 sont représentés en rouge et pointillés

A cette étape de visualisation, on dénote une variance pouvant être importante dans l'aspect des 25 signatures pour certains individus. On estime cependant pour la suite que la complexité du modèle GMM demeure l'unique critère objectif pour démontrer la complexité d'une signature au-delà de son aspect.

2. Classification des 100 personnes selon leur complexité moyenne

2.1 Visualisation des complexités moyennes

Dans le but de catégoriser chaque individu selon la complexité moyenne de ses 25 signatures, on décide de représenter graphiquement la complexité de chacune d'elle (figure 1). On constate graphiquement que plus le nombre de Gaussiennes utilisé augmente, plus la complexité diminue, on constate également une plus grande dispersion de la complexité moyenne pour 24 gaussiennes, ce modèle pourrait d'ores et déjà sembler plus intéressant puisqu'il s'avérait plus discriminant : des catégories émergent plus facilement.

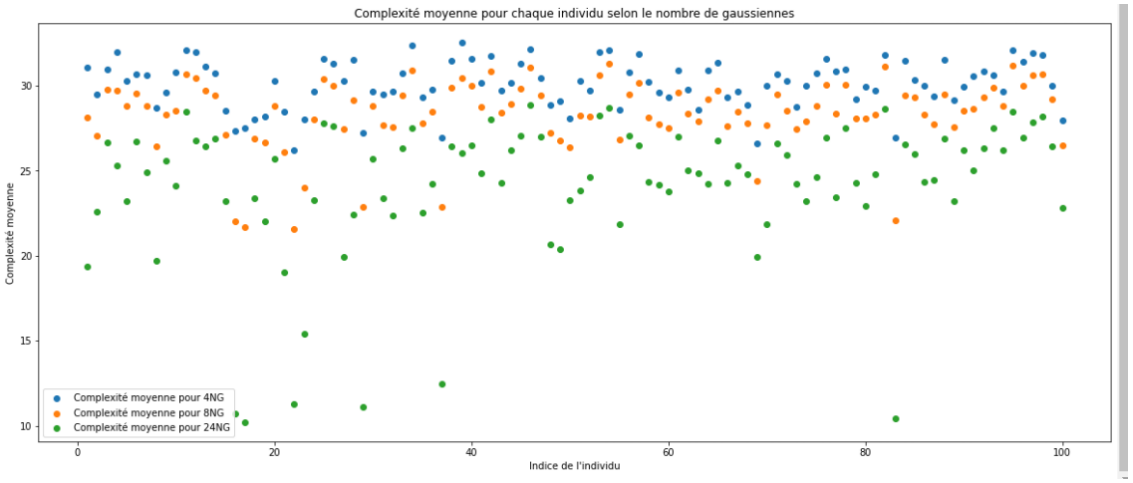


Figure 1 : Complexité moyenne pour chaque individu suivi son indice et le nombre de Gaussiennes utilisé pour le calcul de la complexité.

2.2 Clustering des individus par K-Means et K-Medoids

Étant donné le caractère abstrait de la complexité d'une signature, des méthodes de classification non supervisée semblent plus pertinentes dans l'optique de classifier les individus. Les algorithmes du K-Means et du K-Medoids sont donc particulièrement pertinents dans l'optique de déceler des similarités entre les individus.

L'utilisation des classes pré-définies par sci-kit learn a ici été privilégiée car ces dernières fournissent des résultats bien plus pertinents que les classes que nous avons nous-mêmes implémentées. Les deux algorithmes classifient alors nos données en 3 catégories (voir figure 2). On constate que la répartition pour le K-means tend à marginaliser le cluster 1 qui correspond en fait au cluster des signatures à complexité faible et ce quelque soit le nombre de gaussiennes.

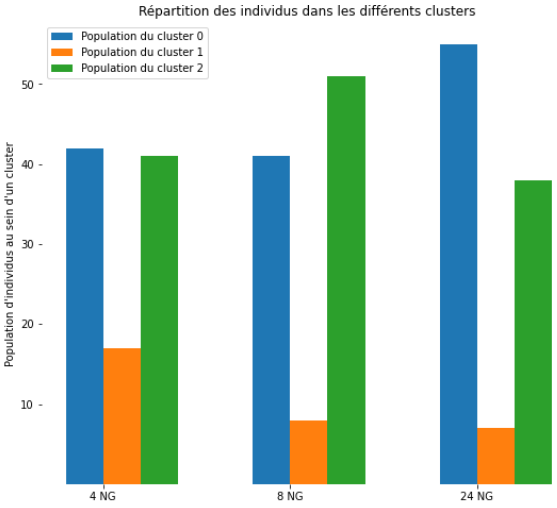


Figure 2 (a) K-Means

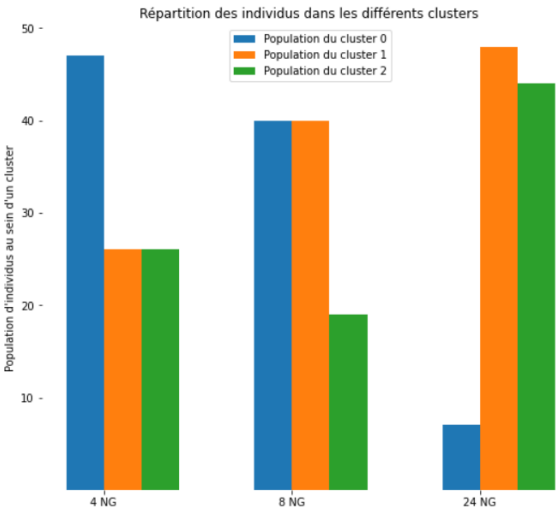


Figure 2 (b) K-Medoids

On constate également une diminution de l'effectif du cluster de complexité plus faible davantage progressive pour le K-medoid. Ce qui n'est pas visible graphiquement car les clusters sont positionnés aléatoirement.

On visualise d'abord les trois clusters pour 4 gaussiennes. On remarque très peu de différences notables entre la classification du K-Medoids et du K-Mean si ce n'est que la classe des signatures de complexité faible du K-Medoid est plus importante. On constate cependant que la démarcation entre le cluster hautement complexe et le cluster moyennement complexe n'est pas nette pour le K-Mean. (figure 3)

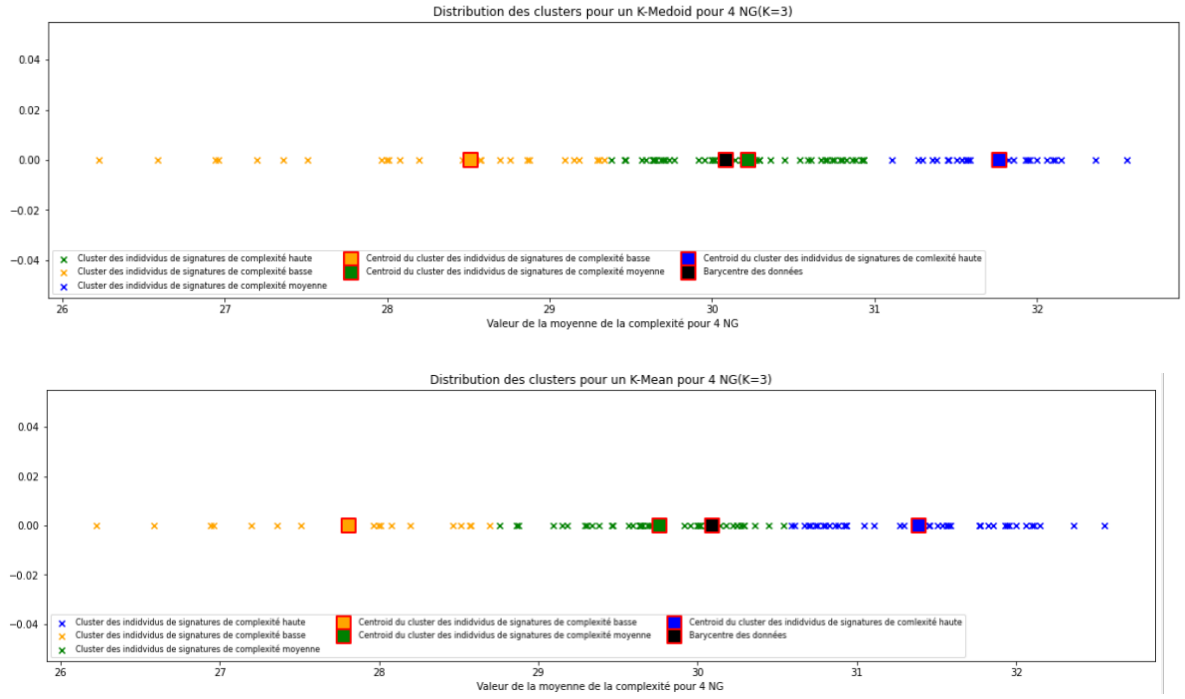


Figure 3 - Distribution des individus après clustering pour 4 Gaussiennes

Pour 8 gaussiennes, les résultats apparaissent nettement différents car le cluster des signatures peu complexes du K-Medoid est encore plus large, contrairement au K-Mean. (figure 4)

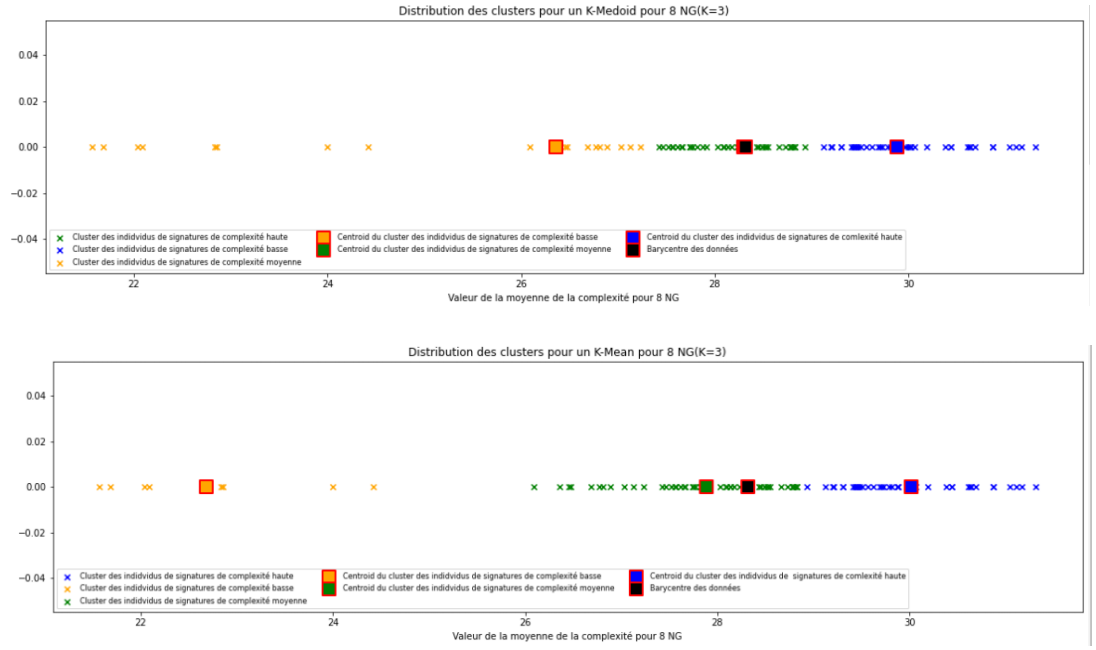


Figure 4 - Distribution des individus après clustering pour 8 Gaussiennes

Finalement pour 24 Gaussiennes, les deux algorithmes classifient les signatures les moins complexes de la même façon. Les différences apparaissent minimes si ce n'est que la classe la plus complexe du K-Mean a un effectif plus important. (figure 5)

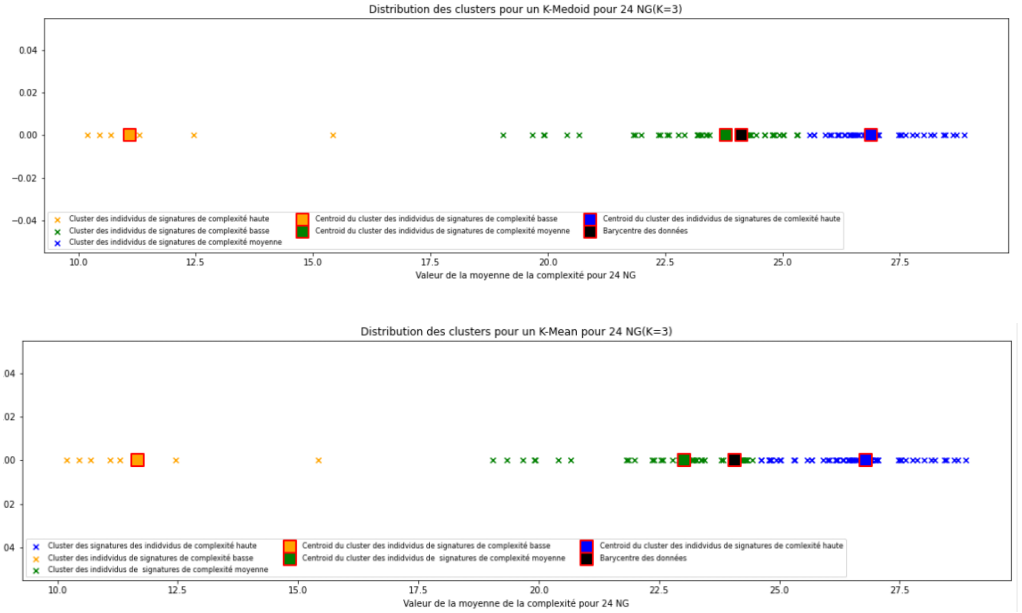
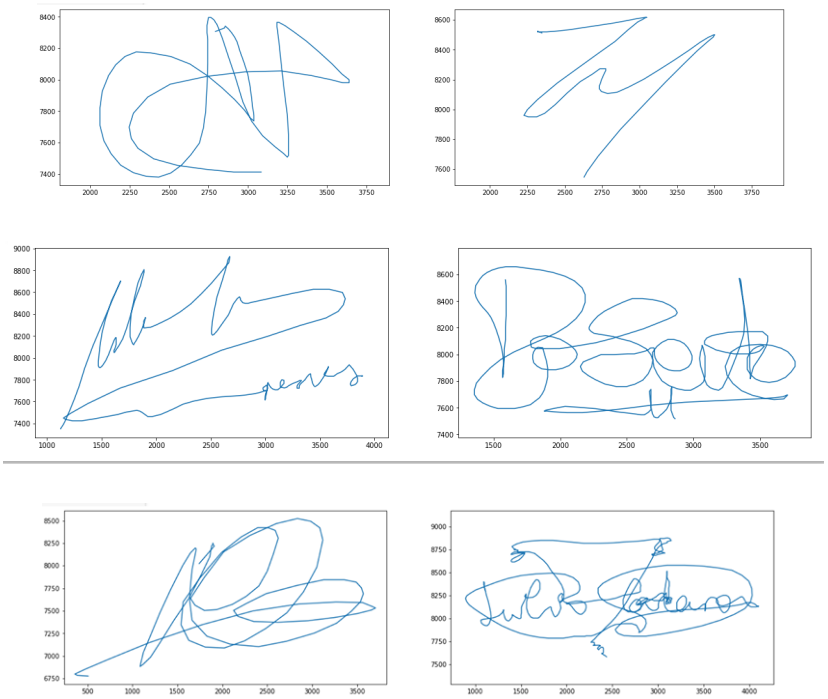


Figure 5 - Distribution des individus après clustering pour 24 Gaussiennes

On décrit ici les signatures des différents clusters, si il est clair qu'on observe une différence notables entre les signatures les moins complexes et les autres, il n'y a que peu de différences notables entre les signatures moyennement complexes et hautement complexe selon l'oeil humain, ce qui correspond à la représentation précédente puisqu'on avait du mal à définir une démarcation nette entre ces deux catégories. (figure 6)



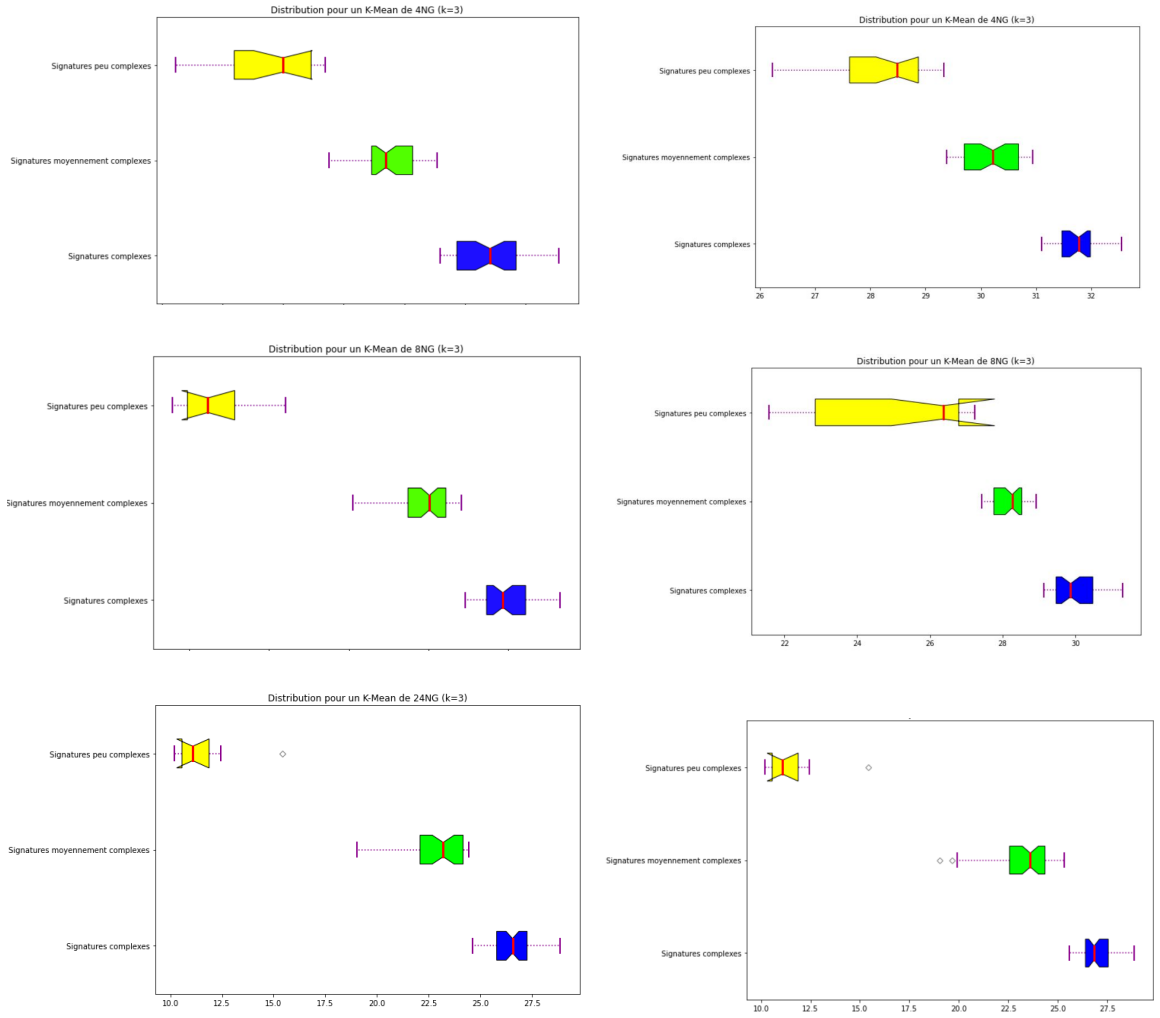
K-Mean *K-Medoids*

Figure 6 : De haut en bas : de la catégorie la moins complexe à la plus complexe

2.3 Comparaison des méthodes de clustering

On observe le clustering pour 4 gaussiennes avec des box plots pour expliciter le minimum, le maximum, les quartiles et la répartition de la population au sein des clusters.
Pour le K-Means à 4 Gaussiennes, on a des boxplots beaucoup plus resserrés pour les signatures peu complexes et très complexes. Les box plots permettent d'observer que les clusters tendent non seulement à s'éloigner mais également à se rétrécir lorsque le nombre de gaussiennes utilisées augmente. L'observation effectuée en introduction s'avère donc vérifiée, il semble que le GMM à 24 Gaussiennes est plus propice au clustering que les autres. (figure 7)

Pour le K-Medoid, on constate une amélioration du clustering pour 8 gaussiennes mais le meilleur clustering est également pour 24 gaussiennes. Le box plot nous permet d'observer aussi que l'algorithme de K-Medoid permet d'obtenir des clusters légèrement plus étroits pour le deuxième et troisième clusters.



K-Mean *K-Medoid*

Figure 7 - Box plots des deux méthodes de clustering

On s'intéresse enfin aux valeurs numériques pertinentes pouvant être tirées des clusters. On distingue premièrement les différents représentants (les centroids carrés entouré en rouge de la figure 5). Le barycentre des données est lui en noir. La distance utilisée est la distance euclidienne au carré.

On obtient donc l'inertie intra-classe et inter-classe, le meilleur modèle étant celui minimisant l'inertie intra-classe et maximisant l'inertie inter-classe, on constate d'une part que le K-Medoid est particulièrement inadapté pour un modèle à 8 gaussiennes. On note aussi que le modèle le plus adapté est le K-Medoid avec 24 Gaussiennes, on sélectionne donc ce modèle pour les autres classifications. (tableau 1 - tableau 2)

	K-Medoid	K-Mean
NG = 4 gaussiennes	36.72	30.19
NG = 8 gaussiennes	142.65	50.9
NG = 24 gaussiennes	185.34	190.53

Tableau 1 : Inertie Intra-classe

	K-Medoid	K-Mean
NG = 4 gaussiennes	5.34	6.99
NG = 8 gaussiennes	6.27	34.76
NG = 24 gaussiennes	177.66	162.73

Tableau 2 : Inertie Inter-classe

3. Classification non-supervisée des signatures sur le modèle à 24 NG

3.1. Etude du clustering des signatures

On décide maintenant de se focaliser sur le modèle à 24 Gaussiennes qui était le plus efficace pour les signatures moyennes par individu. Sur cette partie, l'étude se porte cependant sur l'ensemble des signatures c'est-à-dire 2 500 signatures. Etant donné le nombre beaucoup plus importants de données, ces dernières ont été normalisées (cf. figure 8), le K-Medoids étant le modèle s'étant avéré le plus performant lors de la partie précédente, celui-ci sera utilisé sur cette seconde partie.

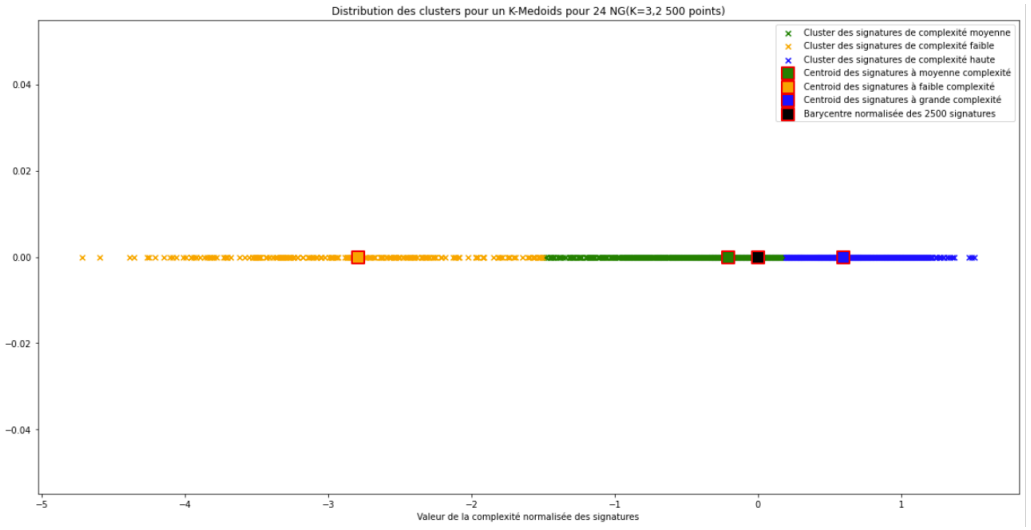


Figure 8 : Représentation des clusters et distribution des complexités des signatures normalisée avec $Xnormalisé = \frac{X-m}{\sigma}$

Indice de silhouette	0.58
Homogénéité	0.11
Inertie inter-classe	8.2
Inertie intra-classe pour le cluster des signatures les moins complexes	1413.65
Inertie intra-classe pour le cluster des signatures moyennement complexes	15771.73
Inertie intra-classe pour le cluster des signatures les plus complexes	957.96
Inertie intra-classe	18143.34

Tableau 3 : L'indice de silhouette étant assez proche de 1, le clustering est donc assez bon compte tenu de la quantité de signatures. L'inertie inter-classe est quant à elle très faible grâce à la normalisation des données et au volume de données significatif. L'importance de l'inertie intra-classe s'explique quant à elle par la haute population de chaque cluster et par l'utilisation de la norme au carré. On remarque de plus que l'inertie intra-classe est énorme pour les signatures moyennes ce qui s'explique par le nombre important d'individus au sein de ce cluster. (cf. figure 9)

Tout comme dans l'étude de la complexité moyenne par individu, on constate de nombreuses signatures au niveau de la frontière des signatures à moyenne complexité et à haute complexité. On émet alors l'hypothèse que les signatures d'une même personne peuvent se retrouver dans un cluster différent dans le cas d'individus possédant des signatures d'un certain degré de complexité. Pour chaque individu, on compte alors le nombre de signatures qu'il possède dans chaque cluster et attribuons à cet individu le cluster dans lequel il a le plus de signatures. (cf. figure 9)

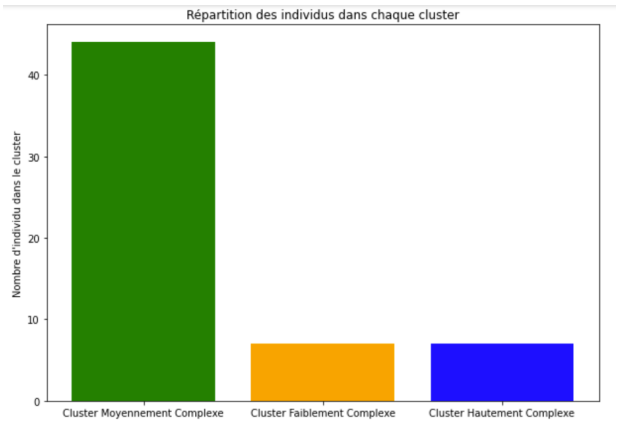


Figure 9 : On constate que le K-Medoids sur l'ensemble des signatures indique que la majorité des individus possèdent des signatures moyennement complexes.

3.2. Etude du clustering des individus

Cette représentation ne prend cependant pas en compte les individus mal classés par le K-Medoids. On décide alors de représenter la répartition des signatures des individus dits “mal classés” c’est-à-dire les individus dont l’ensemble des signatures ne correspond pas à un unique cluster. (cf. figure 10)

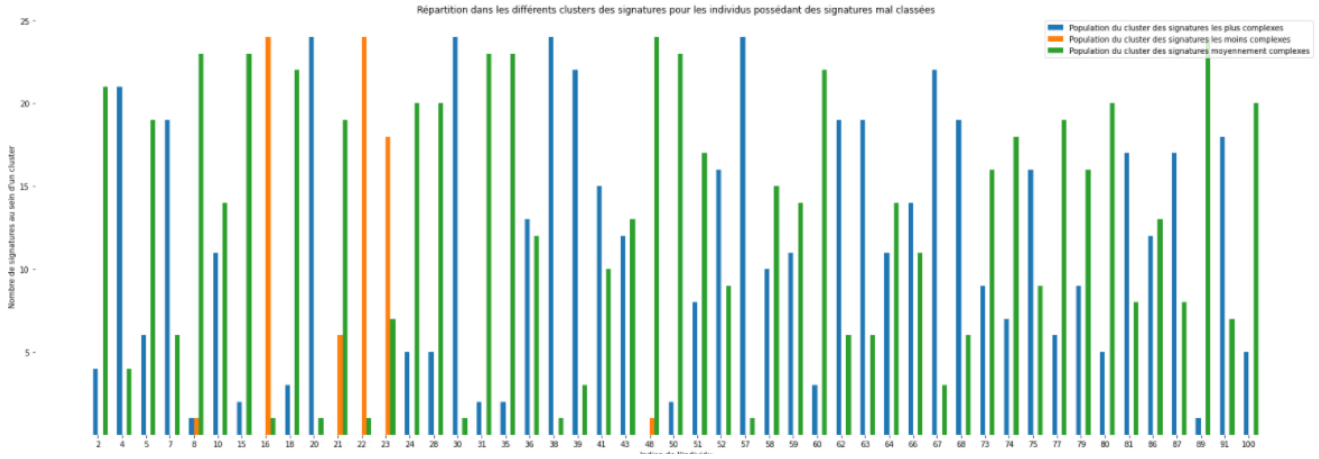


Figure 10 : En abscisse l’indice des individus dont les signatures n’appartiennent pas au même cluster confirme l’hypothèse que la plupart des individus mal classés possèdent une signature excédant une certaine complexité.

La Figure 10 montre ainsi la présence de certaines données aberrantes, on peut penser par exemple à l’individu 8 qui possède au moins une signature dans chaque cluster ou aux signatures d’individus dont plus de 20 signatures appartiennent à un cluster (l’individu 16 par exemple qui possède plus de 20 signatures dans le cluster faiblement complexe mais qui possède une signature dans le cluster moyennement complexe)

Une idée pour définir la pertinence de la complexité moyenne pour définir les signatures d’un individu serait d’examiner la variabilité des complexités de ses signatures autour de sa moyenne. On décide de mettre cela en œuvre en représentant la distribution de l’écart à la moyenne pour chaque signature d’un individu. (cf. figure 11)

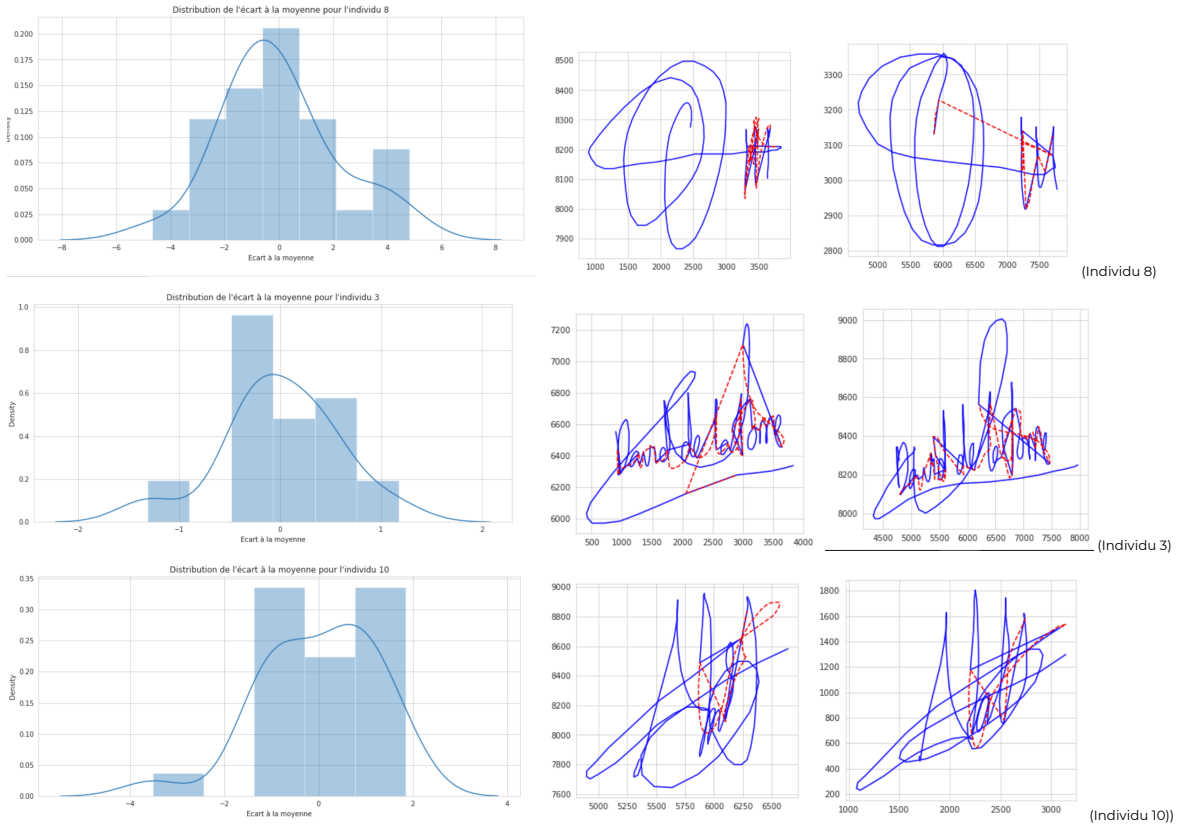


Figure 11 : A gauche : La distribution de l'écart à la moyenne pour les 25 signatures d'un individu. Au centre : Un exemple d'écart à la moyenne important pour l'individu. A droite : Un exemple d'écart à la moyenne fort opposé.

On étudie le facteur de variabilité au sein d'un individu à l'aide de trois exemples : Premièrement l'individu 8 qui possède effectivement une signature dont la complexité varie fortement (cela est visible surtout sur la partie droite de sa signature qui peut être parfois simple ou parfois complexe) mais qui demeure cependant très majoritairement classé comme possesseur d'une signature moyenne du fait de sa complexité moyenne se situant au proche du centre de la classe des signatures moyennes. La très haute variabilité de ce signataire (gaussienne d'écart-type 4 contre environ 1 ou 2 pour un individu moyen), montre que classer un individu sur la base de sa signature moyenne peut s'avérer dangereux.

L'individu 3 représente un second type d'individu rencontré, l'ensemble de ses signatures sont homogènes (gaussienne d'écart-type 1) et classifiés dans le cluster des signatures les plus complexes, pour ce genre d'individu, la représentation de la complexité de l'individu par la moyenne de ses complexités est très pertinente, ces individus sont les individus dits "bien classés" représentant la moitié des individus de la base de données.

Enfin l'individu 10 est quant à lui un individu dit "mal classé", 14 de ses signatures sont moyennement complexes contre 12 classées très complexes (cf. figure 10). On constate sur la figure 11 que quelques signatures sont moins complexes que la majorité de ses signatures. Ce sont ces quelques signatures qui font basculer l'individu 10 dans la catégorie des individus à signatures moyennement complexes.

3.3 Pertinence de la classification par complexité moyenne

En conclusion, il apparaît que la classification des individus selon leur complexité moyenne s'avère plus simple mais moins efficace que la classification se basant sur l'ensemble des signatures de l'individu. 50% des individus sont en effet indéniablement bien classés avec cette méthode. Le reste des individus dits "mal classés" sont quant à eux victimes de certains biais comme cela a pu être observé à travers l'individu 10, cependant ces biais n'affectent pour la plupart pas la bonne classification d'un individu comme l'exemple de l'individu 8 le montre.

4. Classification par apprentissage des signatures sur le modèle à 24 NG

La classification non supervisée des signatures s'avère donc assez intéressante pour déterminer la complexité de la signature d'un individu. Il demeure cependant qu'environ 50% des individus possèdent des signatures qui n'appartiennent pas au cluster auquel l'individu est destiné. La classification de l'ensemble des signatures induit également une importante inertie intra-classe (tableau 3) découlant du nombre important d'individus possédant des signatures de complexité moyenne. La base de données demeurant assez conséquente, on se propose d'améliorer la classification des individus et des signatures en utilisant une classification par apprentissage : Un K-Medoid sera appliqué à 50% de la base de données afin de récupérer les représentants de chaque clusters et le reste de la base de données sera catégorisé dans différents clusters par l'utilisation de l'algorithme du K-Nearest-Neighbour (KNN) avec un k fixé à 1 afin que chaque signature ne soit associée qu'à un unique cluster.

La pertinence de la classification par apprentissage étant dû à la manière dont les données sont séparées, nous proposons trois techniques de séparation des données.

4.1. Séparation aléatoire des signatures

Afin de représenter au mieux la base de données, une première approche consiste à séparer le jeu de données en deux parties aléatoires. (figure 12)

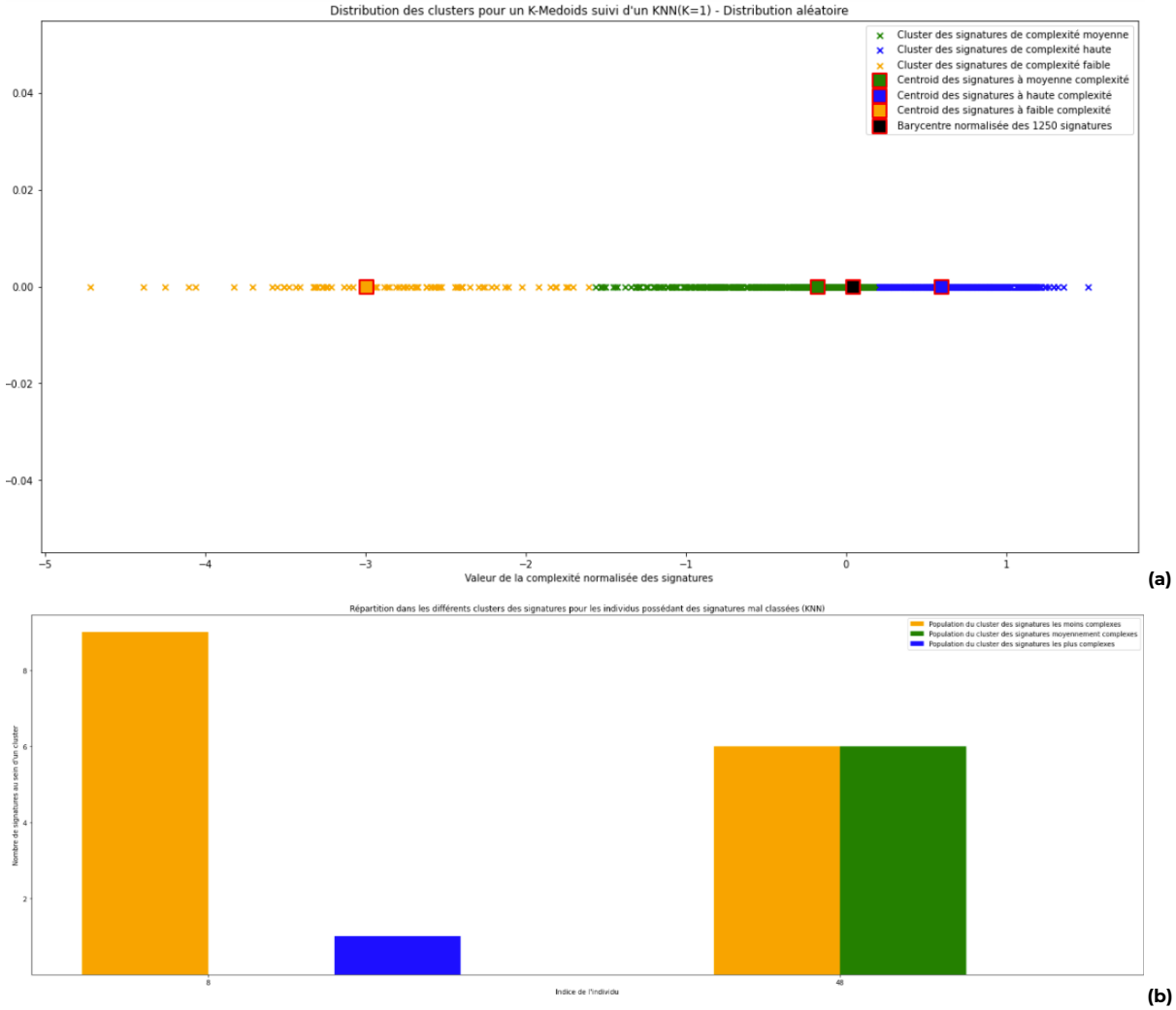


Figure 12 - a) Distribution des signatures dans chaque clusters prédéfinis par le K-Medoid.
Figure 12 - b) Répartition des signatures des individus dont les signatures n'appartiennent pas à une unique catégorie dans les différents clusters.

Une séparation aléatoire des clusters permet visiblement de bien mieux classer les individus dans les différents clusters puisque seuls 2 individus sur les 100 possèdent des signatures pouvant être dans des clusters différents. On notera que l'individu 8 est toujours présent ce qui confirme la grande variabilité de sa signature. La classification par apprentissage semble cependant classer différemment l'individu 48 qui ne possédait qu'une signature mal classée avec un K-Medoid unique (voir figure 10).

Les valeurs numériques fournies plus bas par le tableau 4 indiquent que le clustering est aussi efficace que pour 2500 signatures d'après l'indice de silhouette qui n'évolue pas entre le K-Medoid pour 2500 signatures et pour 1250. L'inertie inter-classe augmente naturellement étant donné que le nombre de données est divisé par 2. On constate enfin une inertie intra-classe beaucoup plus faible, en effet, seul l'inertie intra-classe du cluster des signatures les moins complexes augmente par rapport à l'autre modèle, cela s'explique par une population plus importante de signatures peu complexes

dans le modèle de classification par apprentissage. Globalement la méthode de classification par apprentissage semble meilleure pour classifier les individus avec cette technique de séparation aléatoire.

4.2. Séparation représentative des signatures

Une seconde méthode envisageable pour mieux classer les signatures serait de séparer le jeu de données de manière à ce que la base d'apprentissage soit la plus représentative possible de la base de données. Pour effectuer cela, on visualise premièrement la complexité de chaque signature. (figure 13)

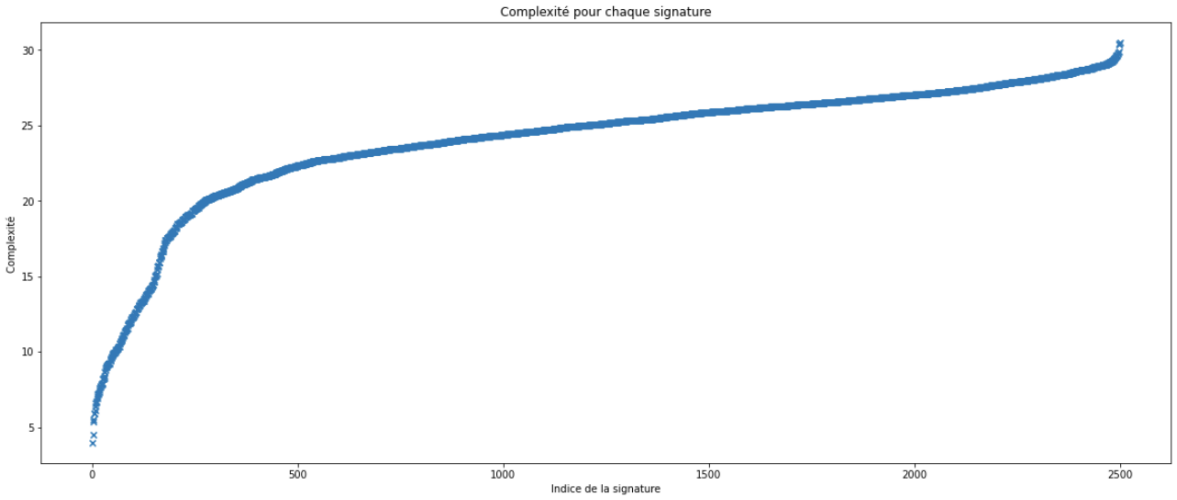
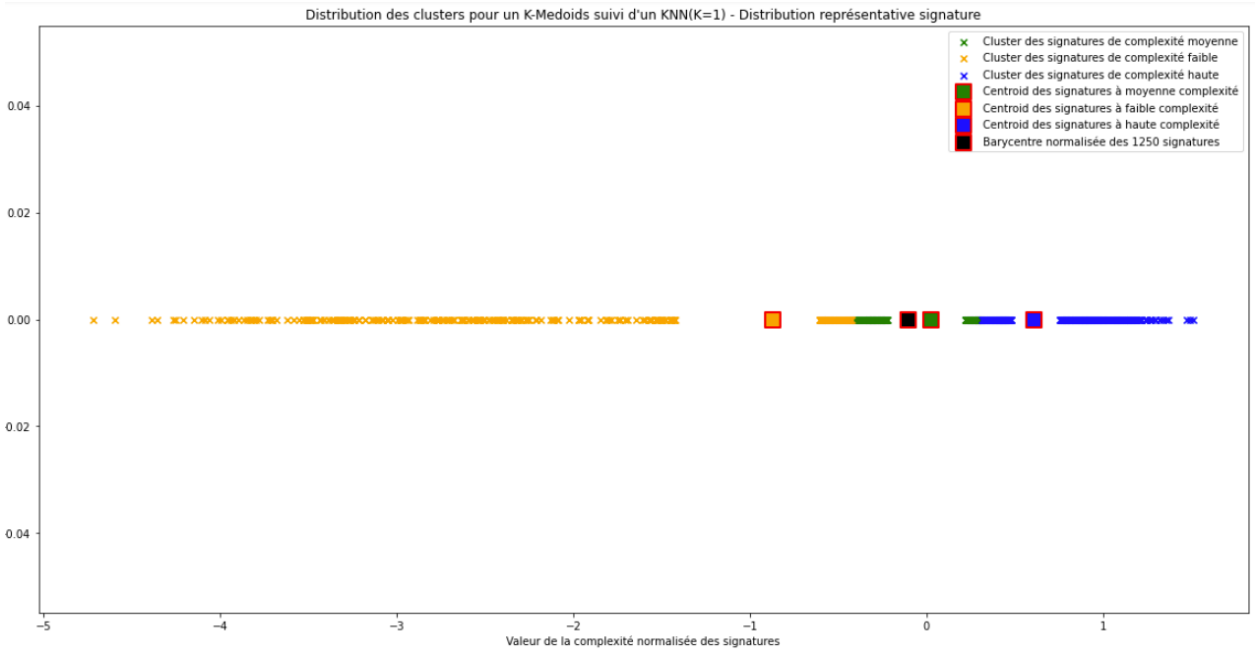


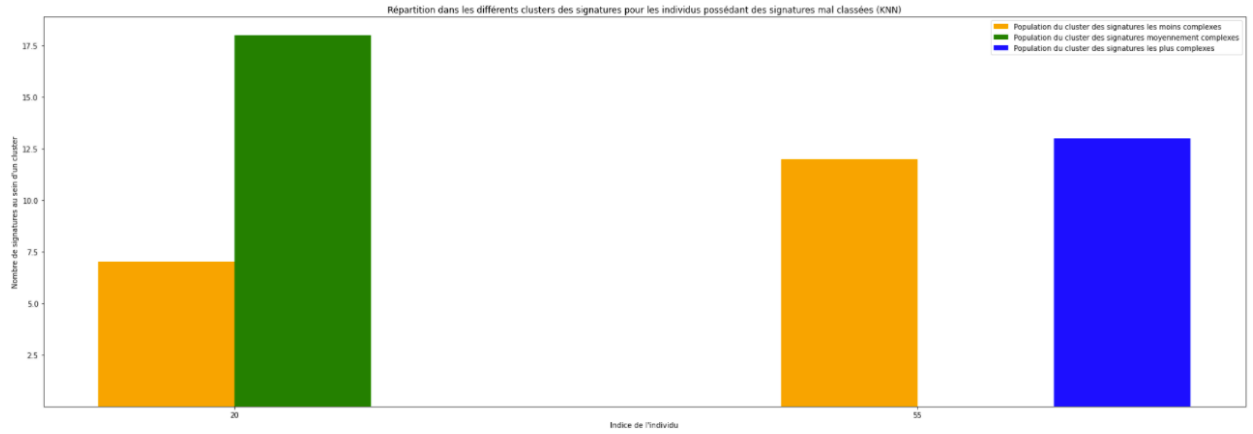
Figure 13 : Complexité de chaque signature de la base de données pour 24 NG

L'allure de la courbe confirme donc un des commentaires de la partie précédente : La majorité des signatures semble avoir une complexité moyenne.
Pour obtenir une base d'apprentissage correcte, on classe les signatures par complexité comme représentées sur le graphique et sélectionnons arbitrairement les 200 à 400 signatures les moins complexes pour représenter les signatures les moins complexes (on ne choisit pas les 200 premières pour éviter les outliers). Les signatures hautement plus complexes étant plus nombreuses que les signatures les moins complexes, on choisit les signatures indexées à 1600 jusqu'à 2100. Enfin, on complète la base de données d'apprentissage en ajoutant les 550 signatures dont l'écart à la moyenne sur l'ensemble des signatures est le plus faible. (norme euclidienne au carré utilisée)

La figure 14 montre les résultats obtenus pour ce type de séparation.



(a)



(b)

Figure 14 - a) Distribution des signatures dans chaque clusters prédéfinis par le K-Medoid.
Figure 14 - b) Répartition des signatures des individus dont les signatures n'appartiennent pas à une unique catégorie dans les différents clusters.

On constate premièrement le manque de représentation de certains individus par rapport à une séparation aléatoire en comparant les figures 14 - a) et 12 -a).

Le manque de représentation de certains individus possèdent cependant des avantages puisqu'on constate que, tout comme pour la séparation aléatoire, très peu d'individus sont mal classés (ratio de 2 individus mal classés pour 52 individus dont la signature est représentée).

Le désavantage majeur de cette représentation réside dans le fait que les informations obtenues par cette classification ne portent que sur les 52 individus.

Toutefois, un avantage certain est que les individus considérés comme bien classés sont bien classés avec plus de sûreté qu'avec une séparation aléatoire. Les individus bien classés avec la représentation de la figure 14

possèdent en effet, pour la grande majorité d'entre eux, leurs 25 signatures dans le cluster. La classification d'un individu dans un cluster est donc bien plus sûre avec cette méthode de séparation. La méthode de séparation aléatoire classe quant à elle l'ensemble des signatures étudiées d'un individu dans un même cluster, mais ces signatures sont cependant moins nombreuses (moyenne ≈ 10).

Paradoxalement, le tableau 4 montre finalement que si l'inertie intra-classe est beaucoup plus faible pour une séparation représentative des signatures, l'inertie inter-classe est aussi beaucoup plus faible que pour la séparation aléatoire. Ce problème n'est cependant pas très impactant étant donné que le choix arbitraire de privilégier certains représentants dans la base d'apprentissage a permis de classifier de manière efficace les individus. L'indice de silhouette montre cependant que cette séparation est moins efficace que la méthode du K-Medoid sur l'ensemble des signatures pour classifier les 1250 signatures.

	Séparation aléatoire	Séparation représentative des signatures
Indice de silhouette	0.58	0.41
Inertie inter-classe	9.57	1.11
Inertie intra-classe pour le cluster des signatures les moins complexes	3633.21	851.23
Inertie intra-classe pour le cluster des signatures moyennement complexes	100.29	28.41
Inertie intra-classe pour le cluster des signatures les plus complexes	49.78	157.07
Inertie intra-classe	3783.28	270.71

Tableau 4

4.3 Comparaison des méthodes de séparation

Une séparation représentative non plus des signatures mais des individus aurait également pu être intéressante pour la comparer aux deux méthodes mises en place. Finalement, il apparaît que la classification des signatures apparaît semblable par l'utilisation d'une méthode de classification par apprentissage. La classification des individus est cependant bien plus efficace par le biais de cette méthode quel que soit le type de séparation des bases utilisé. Les deux méthodes possèdent des avantages et des inconvénients comme il a été vu précédemment. Une séparation adéquate pourrait cependant être mise en place afin de privilégier une meilleure classification des signatures.

5. Conclusion générale :

L'approche consistant à classer les signatures sur la base de leurs complexités moyennes apparaît donc raisonnablement efficace pour le clustering de signatures. Cela permet aussi de conclure quant à une meilleure efficacité du K-medoid pour une classification non supervisée des signatures. Cette classification ne prend cependant pas en compte la variabilité de complexité des signatures des individus qui a été montrée dans la partie 2. Une classification sur l'ensemble des signatures se montre également efficace pour la classification des signatures mais peine cependant à classer correctement les individus du fait du volume important de données (entre autres). La classification par apprentissage fournit alors les meilleurs résultats de classification des individus et des signatures pour une séparation adéquate des données.