MAT 4102

Analyse de signatures pour l'identification

Léopold TERPEREAU, Charles FARHAT

Sommaire

I. Introduction et visualisation des données

Avec la digitalisation des signatures et leur reconnaissance comme méthode d'authentification, il devient important de détecter les possibles falsification de ces mêmes signatures. Pour cela, l'apprentissage machine peut être utilisé pour obtenir, transformer et classifier des données de signature. Le modèle proposé doit être robuste et capable de différencier une signature authentique de sa contrefaçon, même si elles sont très similaires. De plus, il doit être capable de reconnaître un auteur malgré de légères variations dans sa signature. Ce problème est actuellement complexe et certaines entreprises investissent actuellement dans la recherche sur ces problématiques.

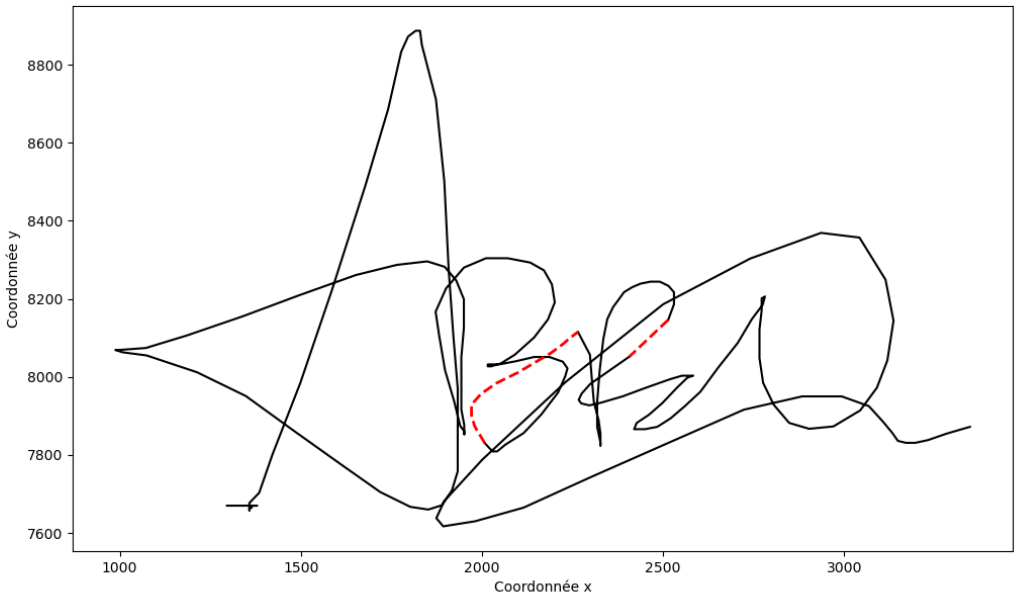
Dans le cadre de ce projet nous allons utiliser la base de données MCYT-100 qui comporte 25 signatures de 100 personnes.

Avant toute chose il vient de vérifier que les données de la base correspondent à ce que nous attendons (les images représentent bien les signatures, les grandeurs semblent cohérentes, les données sont dans le même format...)

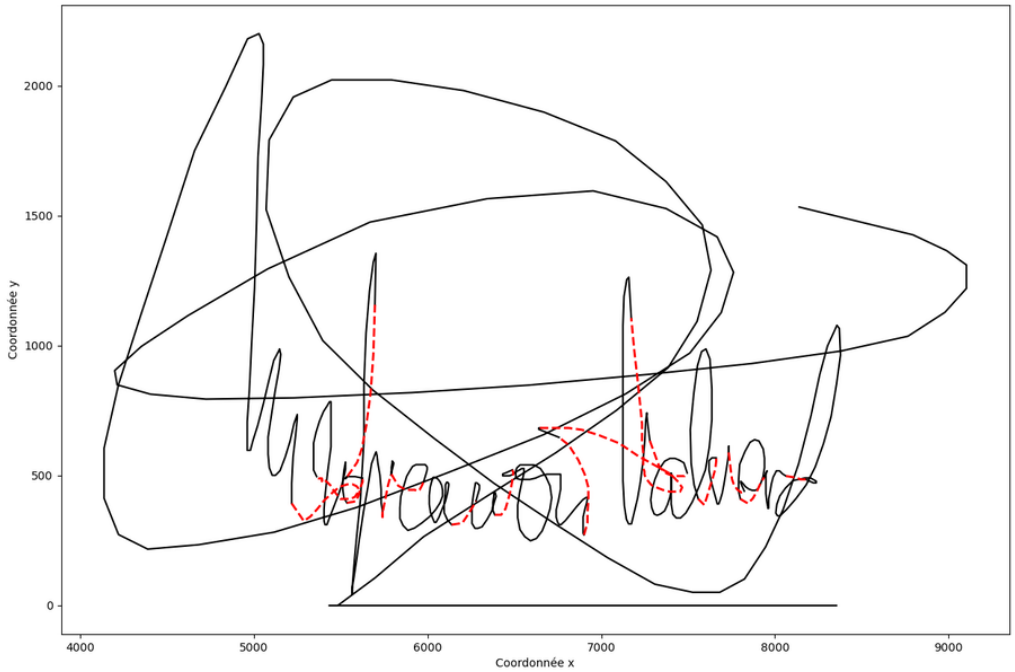
Quelques exemples de MCYT-100

Nous allons commencer par afficher quelques signatures de la base de données pour commencer à l'étudier. Voyons en trois.

Signature n°1 de l'individu n°1



Signature n°10 de l'individu n°10



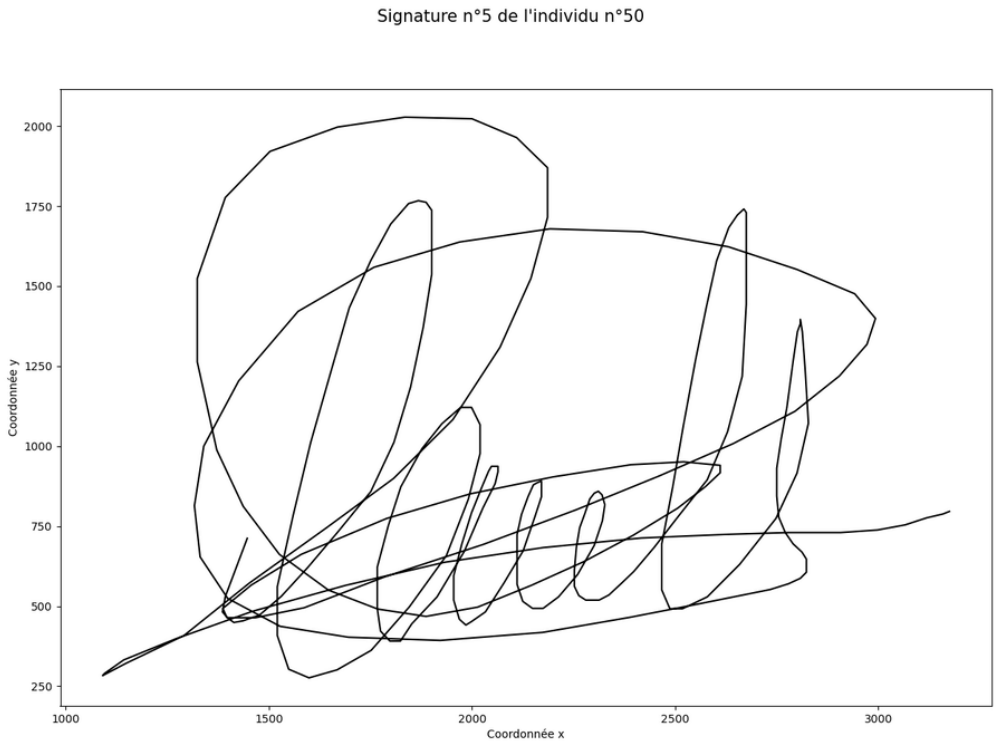


Figure – Quelques signatures pour se familiariser avec la base de données

Il vient plusieurs observations possibles à partir de ces données. Premièrement, elles semblent complètes et peu bruitées. En effet, les lignes sont directement issues de l'enregistrement de l'appareil, elles sont donc discrètes de longueur de la distance d'échantillonnage de l'appareil. Les données obtenues étant de bonnes qualités et d'une précision suffisante pour être par la suite traité, nous pouvons confirmer le bon choix de l'appareil de mesure.

Concernant les signatures en elles-mêmes, le nombre de points étant variant, il nous semble reconnaître trois classes de signatures : celles simples avec peu de détails (première figure), celles de complexité moyenne aux motifs courbes et celles de grande complexité à la précision aiguë. Au-delà des variations de signatures entre les individus, nous remarquons aussi une variabilité des signatures pour un même individu. Ceci était prévisible : la signature étant manuscrite, elle n'est jamais identique.

Pour appuyer notre propos, voici ci-dessous les signatures de l'individu 15 :

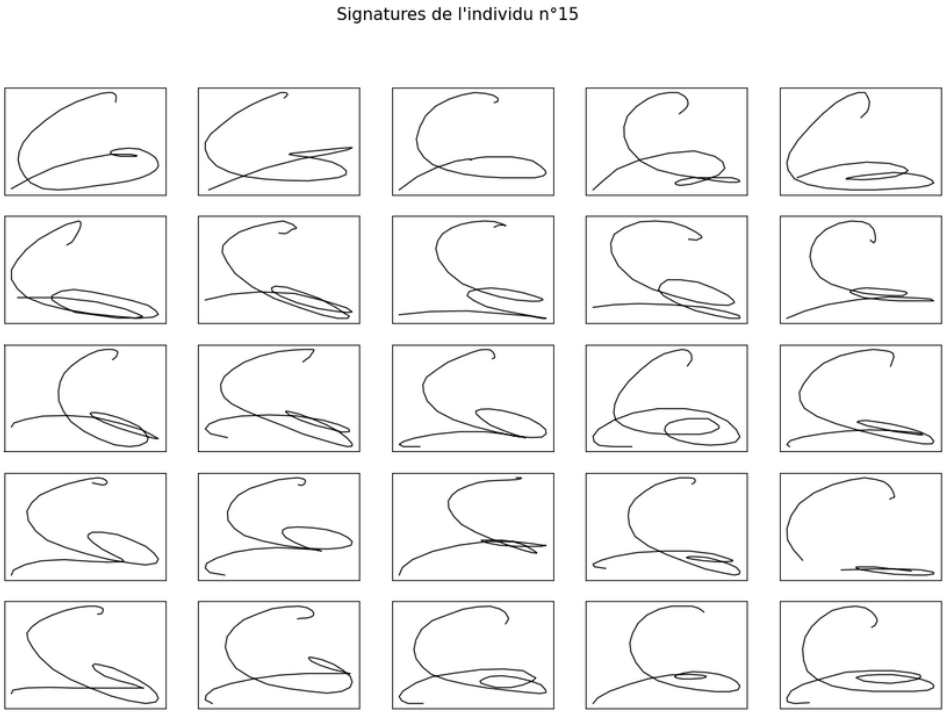


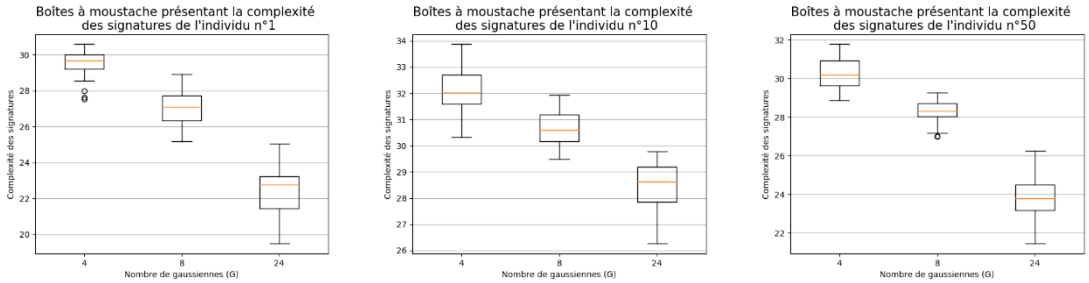
Figure – Les 25 signatures de l'individu n°25

Nous pouvons observer qu'ici la position relative du début de la signature varie fortement entre les différentes captations. En effet, cette caractéristique des données est cruciale puisque les signatures simples ne présentent pas beaucoup de données, ce qui rend ce caractère dispersé et non homogène pour un même individu difficile à traiter. Cette variabilité intrinsèque en est d'autant plus gênante, car un fort écart sur quelques données suffira à mal classifier une signature de faible complexité.

II. Classification non-supervisée des personnes

1. Le nombre de gaussiennes abaisse la complexité des signatures

Comme nous disposons de 3 bases de données issues de trois mélanges de gaussiennes. L'une en comporte 4, une autre 8 et la dernière 24. L'enjeu de cette partie est de déterminer si oui ou non l'augmentation du nombre de gaussiennes pour calculer l'entropie d'une signature a un impact sur la qualité du regroupement donné par des algorithmes de classification non-supervisé.



Figures – Boîtes à moustaches de la complexité des signatures de plusieurs individus, pour les 3 nombres de gaussiennes étudiés

Une tendance ressort directement de ces figures : la moyenne des entropies des signatures d'une personne semble baisser avec l'augmentation du nombre de gaussiennes. Il pourrait être intéressant de chercher à comprendre d'où provient cette tendance (ce qui revient à étudier l'utilisation du mélange de gaussiennes pour caractériser la complexité des signatures).

On va donc essayer d'afficher la complexité moyenne des signatures en fonction du nombre de gaussiennes, on a alors :

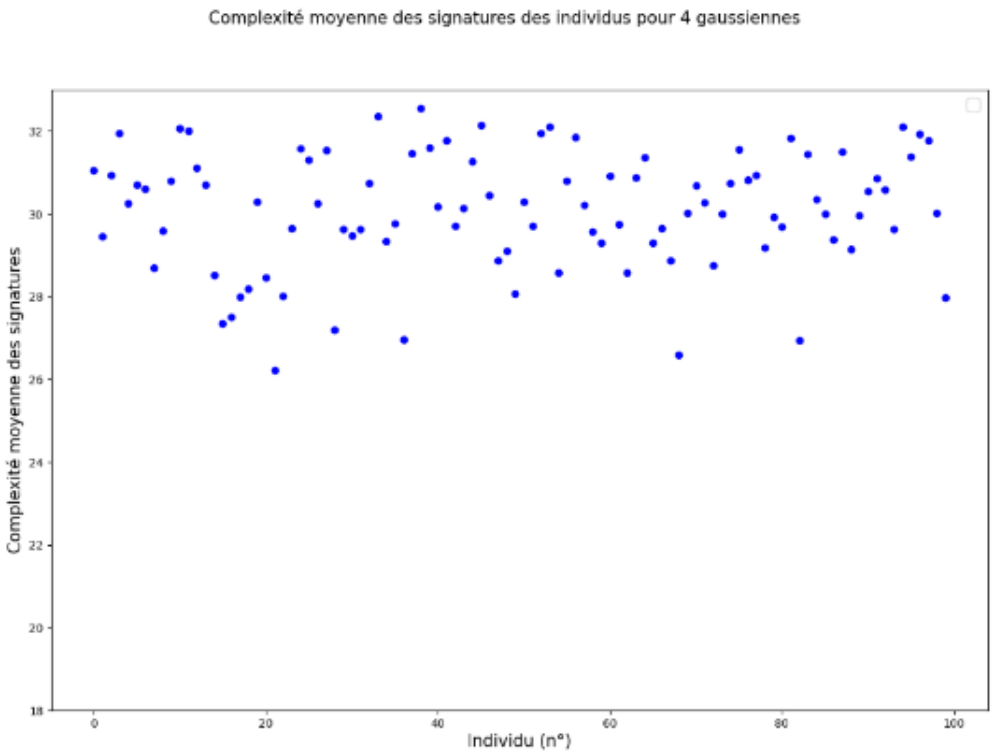


Figure – Nuage de points de la complexité moyenne des signatures des individus, NG=4

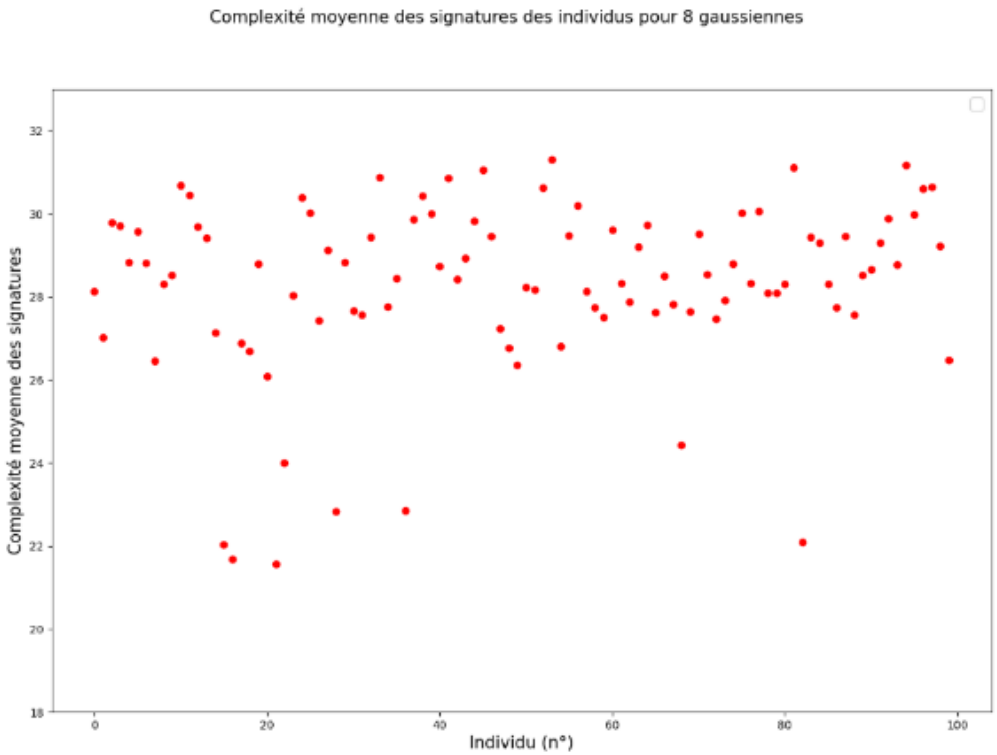


Figure – Nuage de points de la complexité moyenne des signatures des individus, NG=8

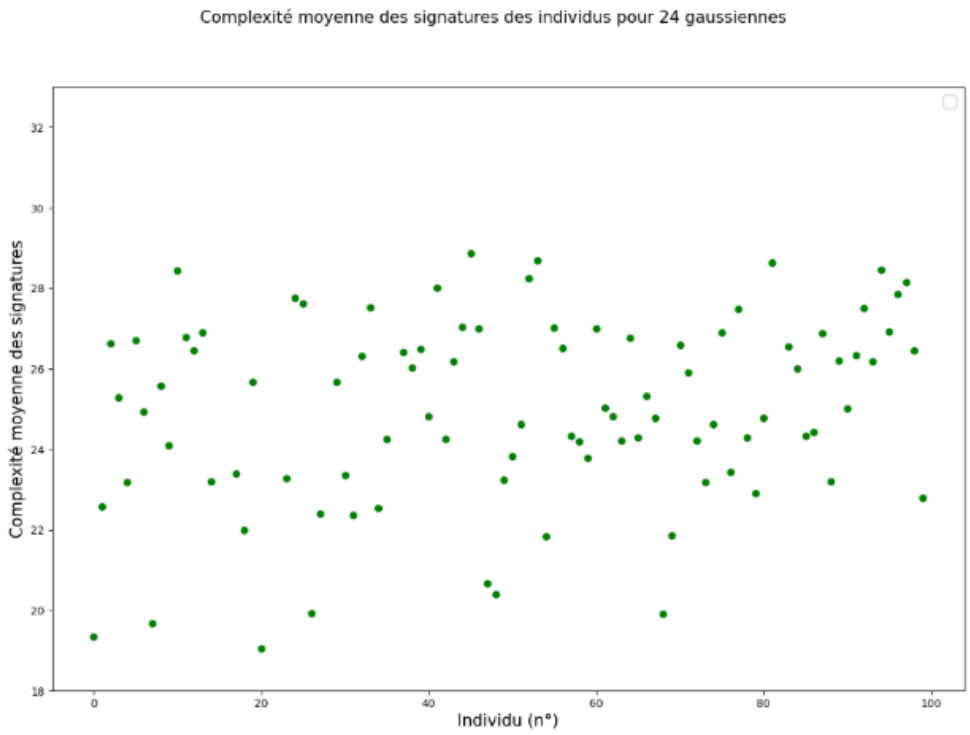


Figure – Nuage de points de la complexité moyenne des signatures des individus, NG=24

Ces figures nous permettent bien de confirmer notre hypothèse : sur cette base de données toutes les moyennes de complexité de signatures admettent diminuent avec le nombre de gaussiennes.

2. Classification classique : *K-means*

L'algorithme K-means

Kmeans est reconnu comme l'un des principaux algorithmes de classification non-supervisée. À partir des hyperparamètres wil détermine une répartition des classes qui minimise une grandeur appelée *Inertie*.

Pour $C \in \mathbb{N}^*$ le nombre de classes $n \in \mathbb{N}$ le nombre de données, $(X_i)_{1\leq i\leq n}$ le nuage de points et $(G_i)_{1\leq i\leq n}$ les barycentres des classes, l'inertie peut s'exprimer sous la forme :

$$I = \sum_i^C \sum_{j\in C_i} d(X_j, G_i)^2$$

Les classes sont initialisés aléatoirement puis recalculés à chaque itération afin de minimiser l'inertie. Il est alors, sous certaines conditions, démontrable que l'algorithme Kmeans converge toujours.

Les principaux problèmes de l'algorithme du Kmeans sont :

- L'inertie de Kmeans en tant que fonction des barycentres est non convexe, sa minimisation dépendra donc de son état initial. Pour être certain d'obtenir un minimum local appréciable l'algorithme devra être initialisé plusieurs fois.
- La vitesse de convergence quant à elle n'est pas assurée d'être rapide. Il est parfois nécessaire d'imposer un nombre d'itérations maximales à partir duquel le calcul des classes sera arrêté.

Certaines méthodes, comme kmeans++, propose des variantes à l'initialisation aléatoire - qui a été prouvée comme peu efficace et sensible aux données aberrantes. Cette approche consiste à éloigner le plus possible les nouveaux centres de ceux déjà choisi lors de l'initialisation...

Nos résultats sur la base de données des signatures

Pour notre expérimentation nous allons donc initialiser une classification non supervisée à l'aide des *Kmeans* de la façon suivante :

- Nombre d'initialisations : *n_init* = 10
- Nombre max d'itérations : *max_iter* = 300
- Algo d'initialisation : *Kmeans++*

Enfin le nombre de classes à utiliser est fixé par l'énoncé : il y aura 3 classes à attribuer (faible, moyenne et forte complexité). Ce point simplifie déjà fortement notre étude : chercher le nombre de groupes optimal pour un nuage de points donné nécessite une analyse à part entière (par méthode du coude (*Elbow method*) par exemple...).

On obtient alors, après entrainement sur notre base de données, un pourcentage d'individu dans chaque cluster en fonction du nombre de gaussienne dans la mixture :

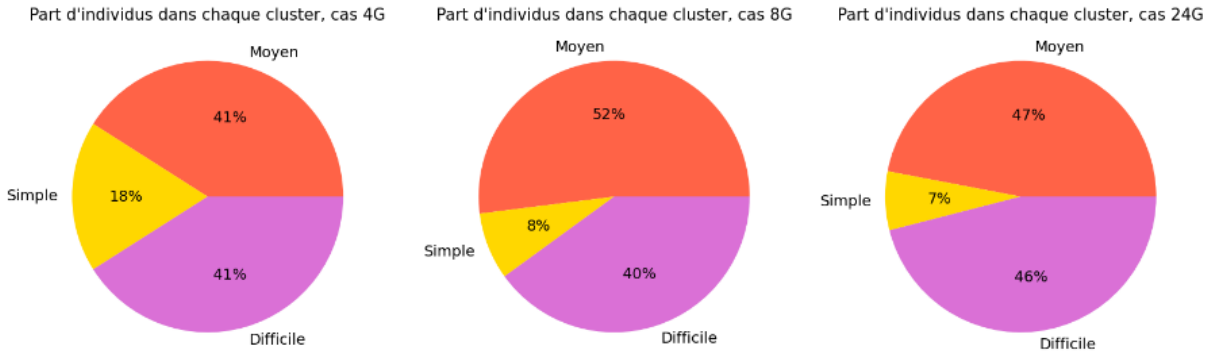


Figure – Distribution des individus dans chaque cluster, pour chaque nombre de gaussiennes étudié

Nous remarquons tout d'abord que la classification donne des groupes de taille inégale. Ce trait semble également s'accentuer avec le nombre de gaussiennes utilisées dans le mélange. Il est pour l'instant impossible de conclure sur une quelconque interprétation, mais le fait qu'un groupe (celui des signatures de complexité moyennes) soit si peu représenté doit nous amener à penser que le nombre de classes n'était pas forcément adapté. Il nous vient alors de nous intéresser à la représentation de chaque cluster.

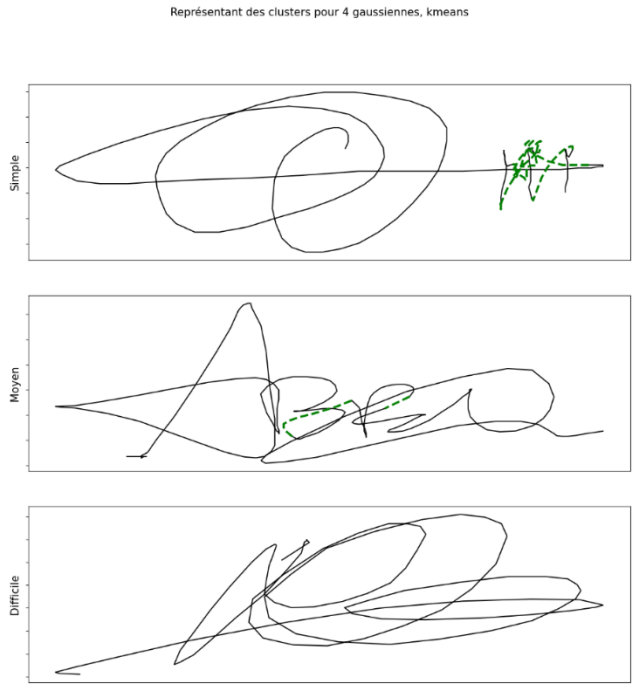


Figure – Représentants de chaque cluster, cas 4 gaussiennes

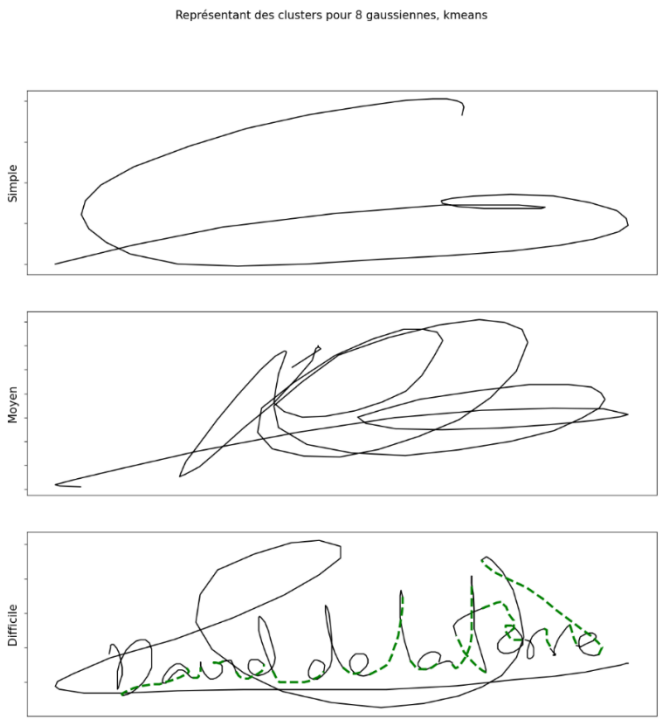


Figure – Représentants de chaque cluster, cas 8 gaussiennes

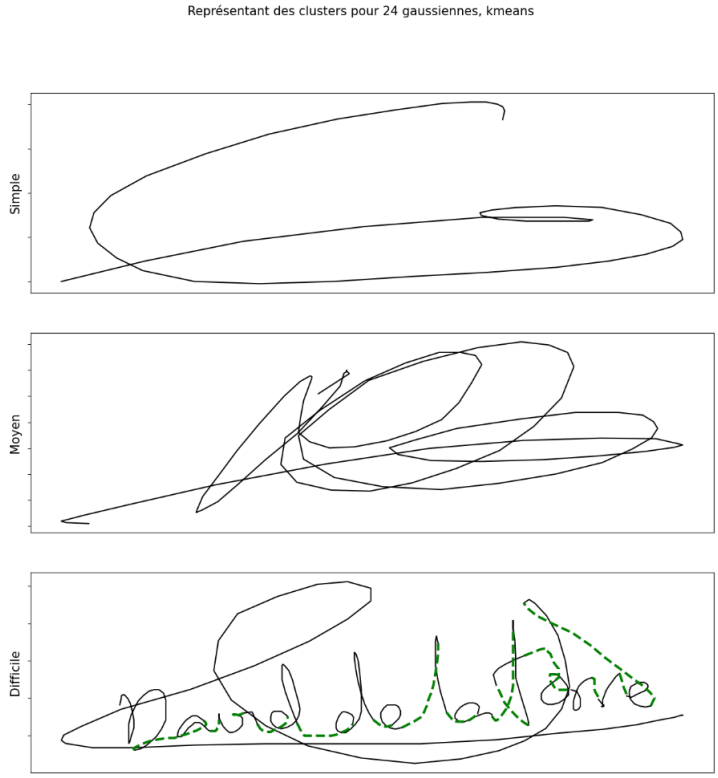


Figure – Représentants de chaque cluster, cas 24 gaussiennes

Enfin, nous calculons les variances inter et intra clusters avec Kmeans (dont l'objectif est de maximiser la variance inter et minimiser la variance intra) :

	4 gaussiennes	8 gaussiennes	24 gaussiennes
Intra	30.11	50.85	183.04
Inter	493.11	2596.95	12764.75

Tableau – Variances inter et intra clusters avec *Kmeans*

Plusieurs observations peuvent être faites à partir de ces résultats :

- Le mélange à 4 gaussiennes retourne des résultats parfois surprenants. À l'œil nu, il semblerait même que la majorité des signatures n'aient pas été correctement classées. Nous pouvons déjà avancer que le modèle à 4 signatures est inadapté.
- La classe de certaine signature semble converger avec le nombre de gaussiennes.
- En revanche la classes d'autres signatures semble assez incertaine. Pour statuer il serait nécessaire d'entrainer plusieurs fois l'algorithme pour confirmer une appartenance.
- Une bonne nouvelle cependant est que nous observons déjà des accords entre les 3 mélanges sur certaines signatures. *Kmeans* semble donc reconnaitre un schéma constant sur une partie des données.

Pour déterminer quel mélange de gaussiennes semble le plus adapté nous choisissons de nous baser sur une classification humaine : la "vraie" classe de la signature sera celle attribuée après une observation à l'oeil nu. Même si cette approche risque d'admettre des biais, elle reste une méthode simple et efficace pour analyser nos résultats.

Pour ce type de métrique - qui peut donc être critiqué du fait de sa grande subjectivité - nous concluons donc qu'il est préférable de travailler avec un GMM à 24 gaussiennes. La raison est non seulement que cela permet plus de degré de liberté dans le calcul de la complexité - l'information portée est donc plus fine -, mais également qu'après plusieurs observations de représentants (que nous n'afficherons pas ici pour des raisons de clarté), les classifications de ce mélange s'approchaient le plus de nos classifications humaines.

3. Autre classification : *K medoids*

L'algorithme Kmedoids

Pour obtenir une comparaison nous proposons maintenant d'utiliser un deuxième algorithme de classification non supervisée : *Kmedoids*. Cet algorithme nécessite des préparations identiques à celle de *Kmeans*, ce qui nous simplifie son utilisation. Cette approche se base également sur une minimisation de distance aux centres de classe, la différence tient au fait que les centres (de classe) en question seront toujours des points de la base de données (contrairement aux barycentres de *Kmeans*). La dépendance aux conditions initiales est là encore à prendre en considération, la fonction de coût n'étant toujours pas convexe, il nous faudrait relancer l'apprentissage plusieurs fois.

Nos résultats

Le premier résultat concerne comme précédemment le pourcentage d'individus dans chaque cluster:

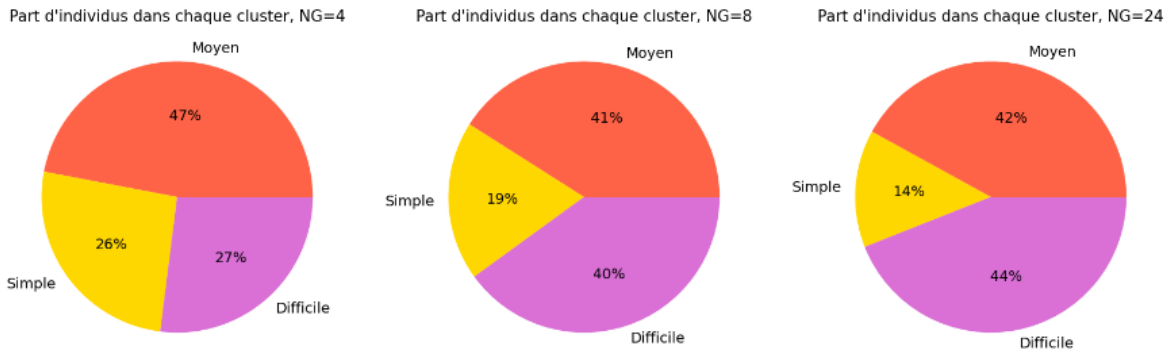


Figure – Distribution des individus dans chaque cluster, pour chaque nombre de gaussiennes étudié

Nous remarquons un meilleur équilibre entre les différents clusters, même si l'augmentation du nombre de gaussiennes tend à faire diminuer le groupe des individus "simples" (c'est à dire dont la moyenne des complexité des signatures est basse).

Nous pouvons de la même façon afficher les centres de chaque clusters pour les différents mélanges de gaussiennes :

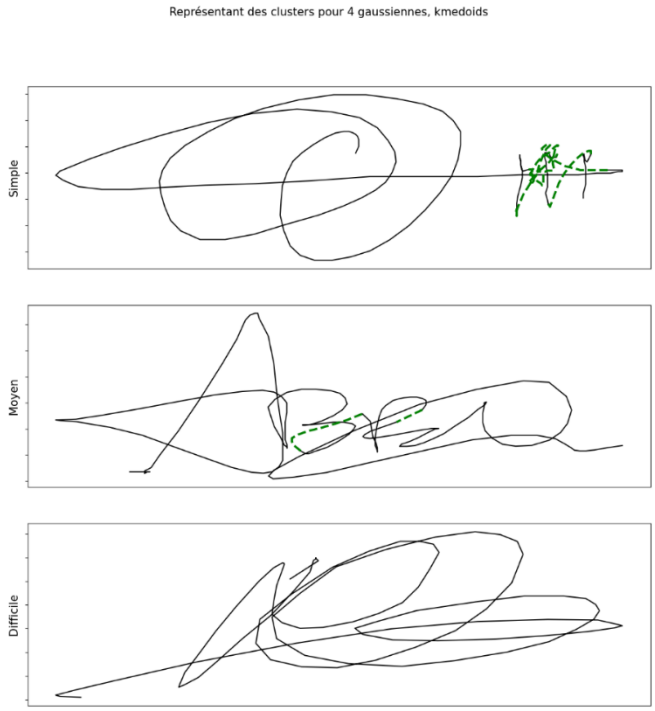


Figure – Représentants de chaque cluster, cas 4 gaussiennes

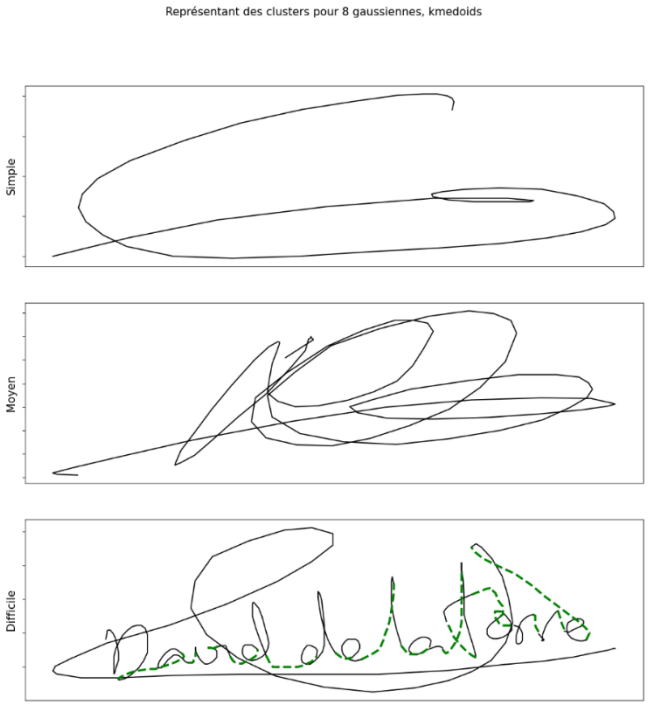


Figure – Représentants de chaque cluster, cas 8 gaussiennes

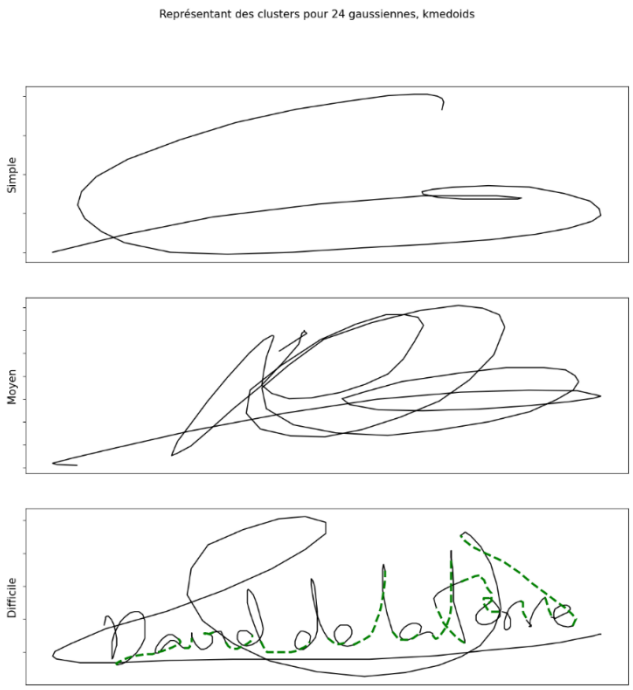


Figure – Représentants de chaque cluster, cas 24 gaussiennes

Finalement nous affichons aussi les différentes variances inter et intra clusters avec *kmedoids*. Ici aussi nous cherchons à maximiser la variance inter et minimiser la variance intra :

	4 gaussiennes	8 gaussiennes	24 gaussiennes
Intra	47.13	69.83	122.26
Inter	488.49	487.03	3391.19

Tableau – Variances inter et intra clusters avec *Kmedoids*

Ces résultats nous indique les aspects suivants :

- Pour 4G nous remarquons que les classifications réalisées par Kmeans et par *Kmedoids* sont extrêmement similaires.
- Nous retrouvons cependant toujours quelques résultats aberrants à l'œil nu pour ce mélange. Encore une fois ces erreurs avaient été constatées pour *kmeans*.
- On observe également des points d'accords entre les 3 mélanges pour *Kmedoids*, preuve que certaines signatures sont particulièrement faciles à classifier (c'est à dire éloignées des autres).

- Nous remarquons de plus des signatures dont la catégorie semblait incertaine pour *Keans* qui apparaissent stables pour *Kmedoids*.

4. Comparaison des deux méthodes de clustering

Pour conclure sur cette première comparaison nous devrions trancher sur le choix du meilleur classificateur. Ce choix n'est cependant pas aisé pour plusieurs raisons. D'un point de vue mathématique, l'inertie intra cluster de *Kmeans* est plus faible et l'inertie extra de *Kmedoids* est la plus grande. Comme nous recherchons davantage une forte séparabilité entre les classes, *Kmedoids* pourrait sembler le plus adapté.

	4 gaussiennes	8 gaussiennes	24 gaussiennes
Intra	30.11	50.85	183.04
Inter	493.11	2596.95	12764.75

Tableau – Variances inter et intra clusters avec Kmeans

	4 gaussiennes	8 gaussiennes	24 gaussiennes
Intra	47.13	69.83	122.26
Inter	488.49	487.03	3391.19

Tableau – Variances inter et intra clusters avec Kmedoids

Par ailleurs d'un point de vue "réalité physique" il peut être plus utile d'avoir un centre de classe correspondant à une signature, notamment pour des raisons de représentation.

Enfin notre base de données étant particulièrement petite (pour des algorithmes tels que *Kmeans* et *Kmedoids*) nous ne nous sommes pas intéressés aux temps de calcul. Or cette base reste relativement incomplète vis à vis de la variété des signatures existantes. Également, si le procédé doit à terme être implémenté dans une société. La base de données sera régulièrement mise à jour ce qui impliquera des entrainements et vérifications réguliers, donc potentiellement beaucoup de temps de calcul.

Enfin, quel que soit l'algorithme les résultats sont sans appel concernant la meilleure base de données : celle à 24 gaussiennes représente bien mieux les signatures que les deux autres. La comparaison avec une classification à l'œil nu soutient particulièrement cette hypothèse.

III. Classification non-supervisée des signatures

Dans la partie précédente nous nous sommes concentrés sur la complexité moyenne des 25 signatures des individus. Bien que cette approche simplifie l'étude, elle supprime la variabilité intrinsèque d'un individu lors de sa signature, en passant de dimension 25 à 1. Nous allons à présent étudier la base de données complète des signatures des individus et non plus seulement la moyenne de complexité des signatures d'un individu, en nous limitant aux données du mélange de 24 gaussiennes dans tout ce qui suit.

1. Classification par *Kmeans*

Commençons par visualiser la variabilité de complexité des signatures des individus. Dans le nuage de points qui suit, chaque couleur représente un individu, et la croix représente la moyenne de complexité de ses signatures.

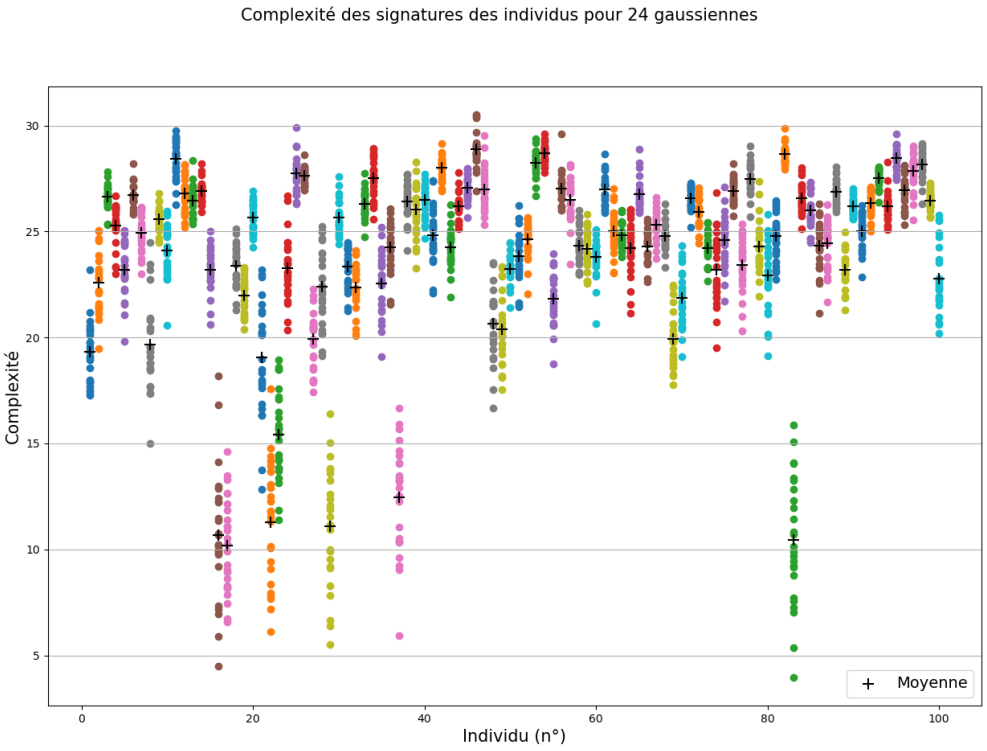


Figure - Complexité des signatures des individus, 24G

Tout d'abord, on peut remarquer que les signatures de faible complexité ont une variance assez importante, ce qui est lié au plus faible nombre de points/positions qui les composent. Cependant, leur éloignement des signatures de plus haute complexité rend leur classification relativement simple.

A présent, appliquons l'algorithme de *Kmeans* sur les 2500 (25x100) signatures, en demandant de récupérer 3 clusters. La figure qui suit présente la classification des signatures de chaque individu. A chaque unité d'abscisse (x entre 1 et 100), on observe les 25 signatures de l'individu x.

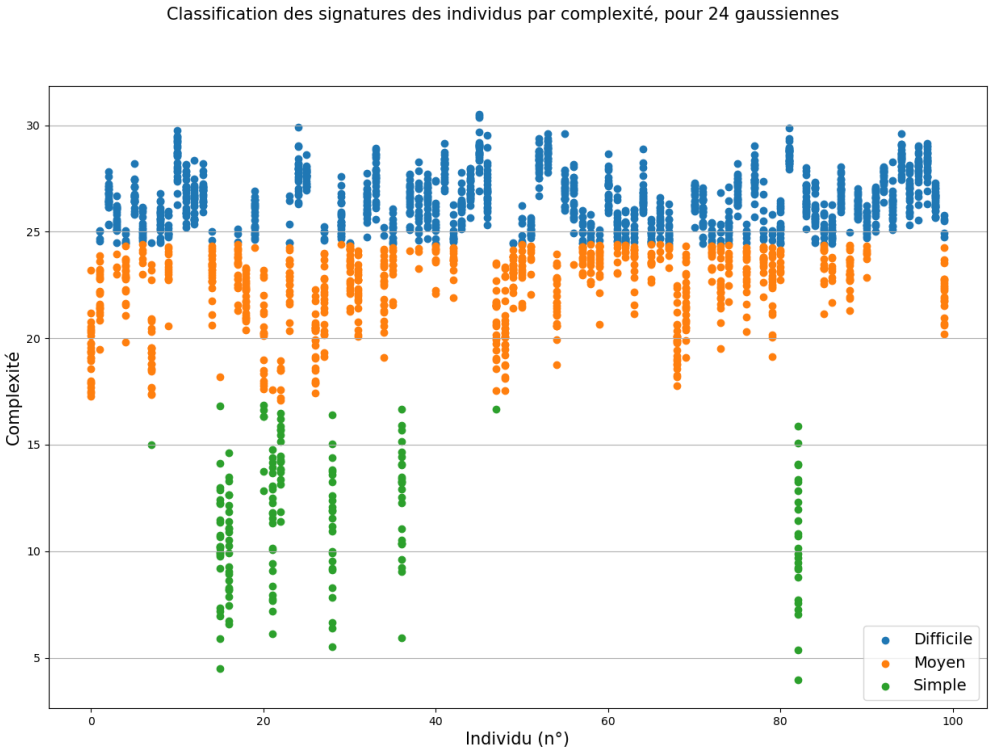


Figure – Classification des signatures des individus, selon leur complexité

Tout d'abord, on remarque que la majorité des individus voient leurs signatures appartenir à deux classes, mais jamais aux trois (ce qui est rassurant, si l'on souhaite classifier les individus et non plus leurs signatures).

La figure suivante décrit la répartition des signatures dans les clusters.

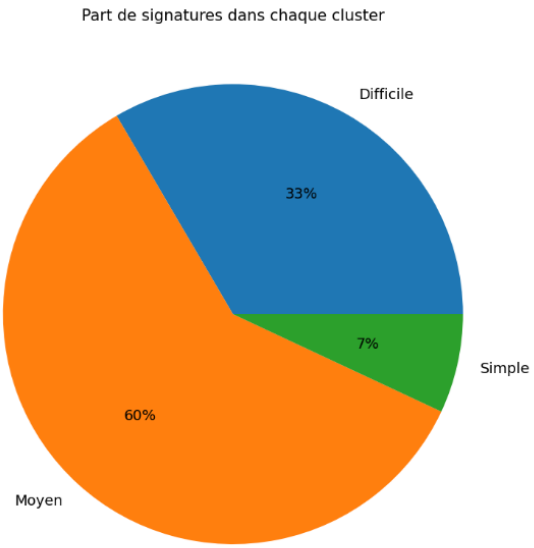


Figure – Proportion des 2500 signatures dans chaque cluster

Enfin, il peut être intéressant de voir la distribution des signatures dans les clusters, pour chaque individu. La figure suivante présente cela par ordre de complexité croissante (selon les clusters attribués), et non pas selon l'ordre "naturel" des individus dans la base de données. Il nous permet d'estimer rapidement la "complexité" d'un individu.

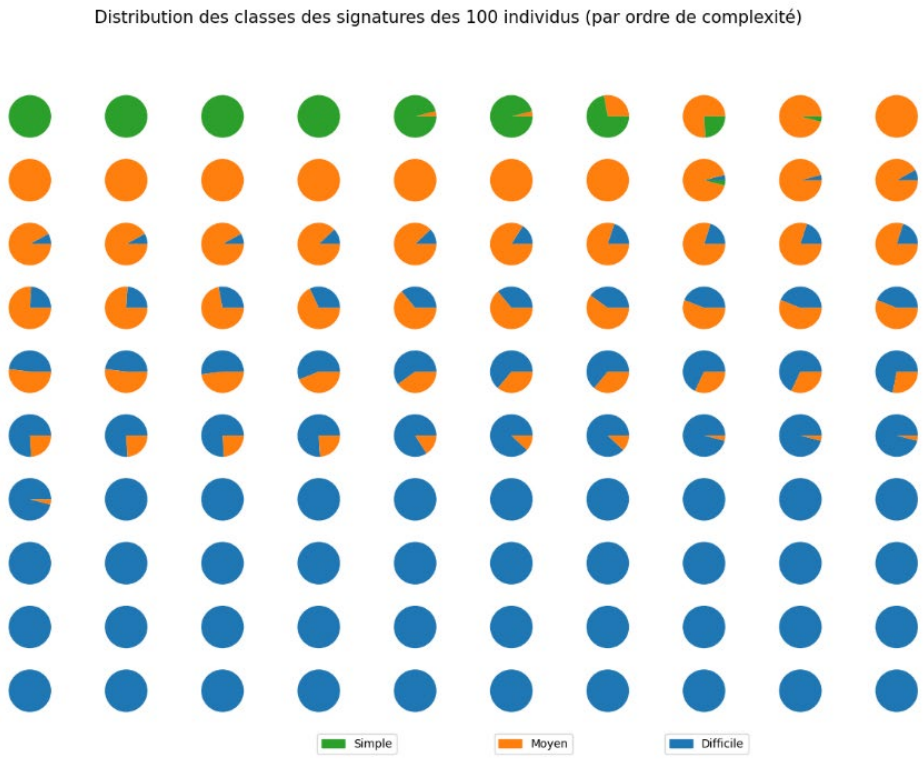


Figure – Proportion des signatures de chaque individu dans les clusters

A partir de la proportion de signatures d'un individu dans les clusters on pourrait proposer une classification des individus selon une règle très simple : chaque individu est classé dans le cluster qui détient le plus de ses signatures.

2. Interprétation des résultats

Le graphique précédent monte que les 25 signatures d'un individu ne sont pas toujours classées dans le même cluster : la "colonne" de points, qui représentent la complexité des 25 signatures d'un individu, on voit pour quasiment chaque individu des points de couleurs différentes, donc appartenant à des classes différentes

Ce sont généralement les individus qui ont des signatures un peu éloignées entre elles chez qui on observe une variabilité dans la catégorisation de leurs signatures. Par exemple pour l'individu n°35, la classification non-supervisée donne un ratio d'environ 50/50 pour la classe de ses signatures (entre la classe 'Moyen' et 'Difficile'). On voit que ses 25 signatures diffèrent de manière significative.

Figure – Signatures de l'individu n°35

Pour certains individus, on remarque que leurs signatures sont très proches, mais que leur complexité varie fortement, en particulier dans le cas de NG=4.

3. Apprentissage et généralisation

A présent, nous allons appliquer l'algorithme *Kmeans* sur la moitié de la base de données (1250 signatures) pour identifier les centres des classes, puis nous allons généraliser la classification au reste de la base (1250 autres signatures) en rattachant chaque signature au cluster ayant un représentant de complexité la plus proche. En l'occurrence, nous choisissons de diviser la base selon 50/50 individus, notamment pour simplifier la visualisation des résultats.

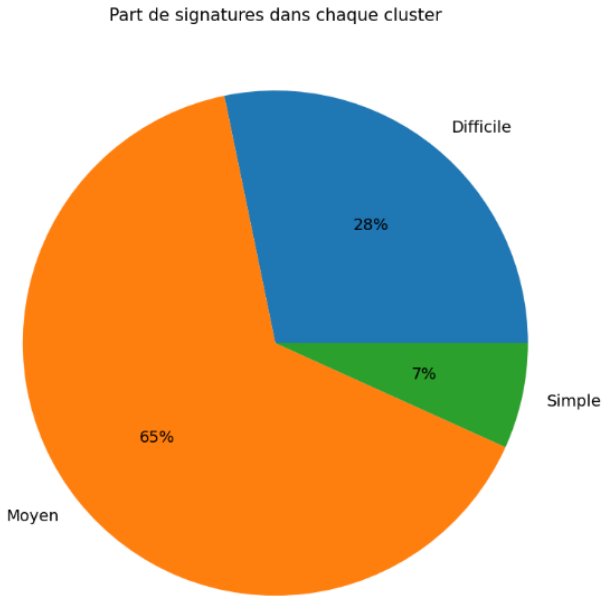


Figure – Proportion des 2500 signatures dans chaque cluster

En reprenant la méthode précédente mais en divisant la base d'individus en deux parties, on obtient une classification légèrement différente de la précédente :

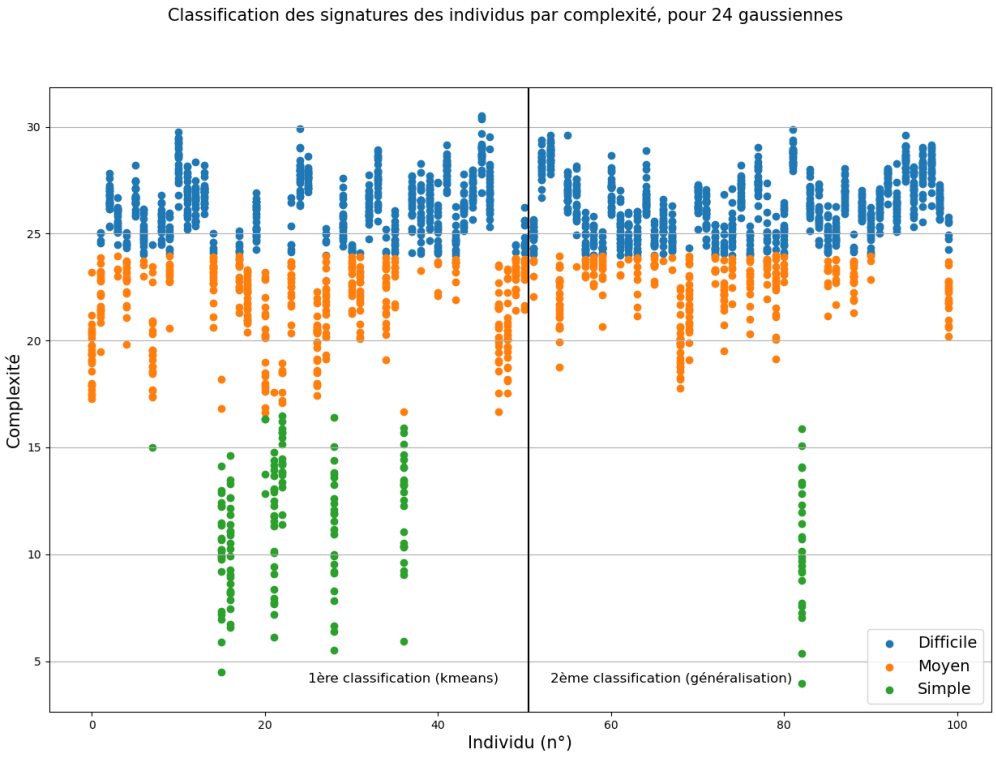


Figure – Classification des signatures des individus, selon leur complexité

Classifier la signature d'une personne selon sa complexité peut être très intéressant pour renforcer la sécurité de son identité numérique. En effet, on peut imaginer un cas d'application de la généralisation où une personne qui "invente" sa signature est capable de savoir si la complexité de sa signature est suffisante, où s'il doit la complexifier afin d'entrer dans la classe "Difficile" (ou "Moyen" au minimum).

En réalisant une double exposition avec les images des graphiques on peut facilement mettre en avant (en orange) les signatures qui ont été classées différemment avec la méthode de généralisation :

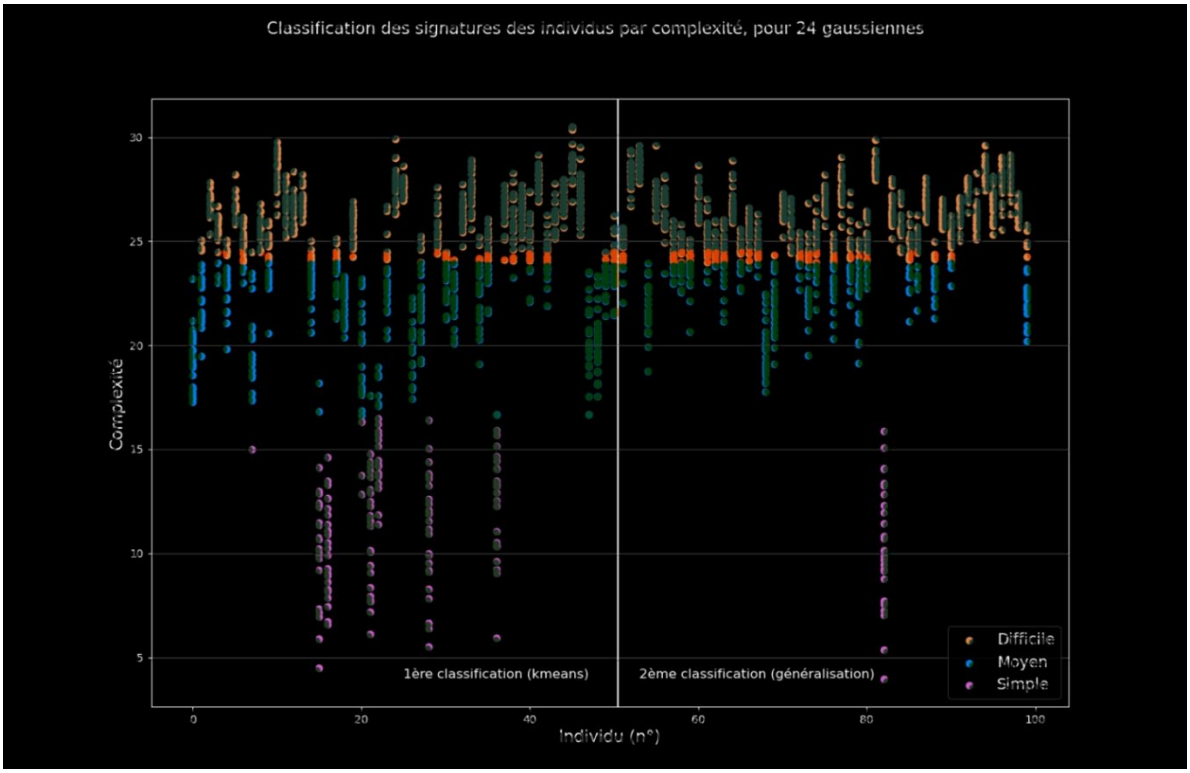


Figure – Signatures classifiées différemment que dans la partie précédente (en orange)

Lors de la première phase de classification avec *Kmeans* on voit que la frontière de décision Difficile/Moyen n'est pas fixée exactement au même endroit : certaines signatures classifiées "Difficile" dans la partie précédente sont ici classifiées "Moyen". Cela est simplement dû à la différence des exemples donnés lors du fit du modèle. Cette différence de frontière de décision se retrouve lors de la deuxième phase, la généralisation, qui classifie de la même façon un certain nombre d'exemples comme "Moyen" alors qu'ils étaient classifiés "Difficile" dans la partie précédente.

4. Conclusion

En s'appuyant sur la totalité des données disponible nous avons remarqué quelques données étonnantes plus "qu'aberrantes". En effet, les signatures de faible complexité admettent de grandes variations lorsque nous en venons au calcul d'entropie et même les signatures majoritairement catégorisées comme "Moyen" ou "Difficile" présentent des variations qui les amènent parfois à passer d'une classe à l'autre. Ainsi il n'est clairement pas possible d'assigner une classe unique (en passant par un calcul de moyenne ou les données brutes en tout cas) à toutes les signatures. Si telle est l'objectif d'une étude il faudrait chercher d'autres représentations de la base de données (x, y, pression, altitude, azimut). Au-delà de ça, on pourrait penser à des outils de réduction de dimension comme l'ACP plutôt qu'un mélange de gaussiennes.