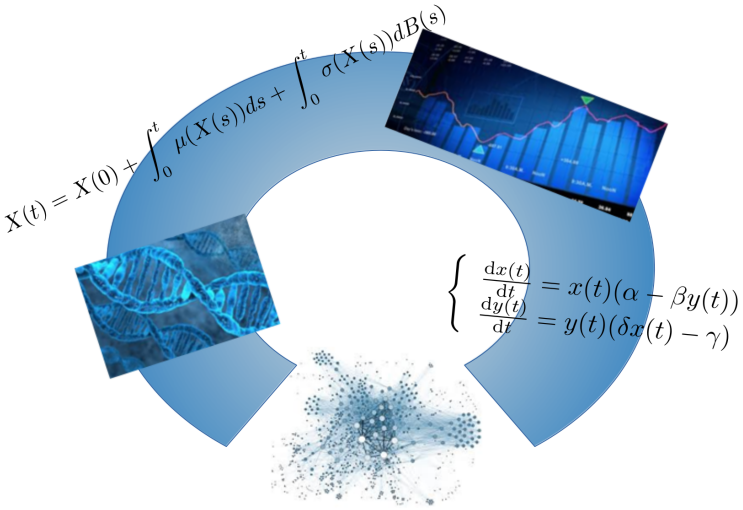
Classification de motifs d'intérêt en surveillance de l'environnement

24/05/2024

Rapport final



Equipe :	Client :	Référent :
Gabriel ABENHAIM	Elisabeth LAHALLE	Elisabeth LAHALLE
Nathan ALIMI		
Ghiles KEMICHE		
Seif ZAAFOURI		

Pôle n° 15
Modélisation Mathématique des Systèmes Complexes

Abstract

Ce rapport présente une étude sur la classification de motifs d'intérêt dans le cadre de la surveillance de l'environnement en utilisant des techniques de classification, principalement K-means et sa variante UK-means. Face aux vastes quantités de données générées par les avancées récentes en surveillance environnementale, il devient essentiel d'adopter des méthodes d'analyse sophistiquées. Notre approche combine l'analyse des données et la construction de modèles mathématiques pour générer et identifier des formes d'intérêt. L'analyse inclut également l'étude détaillée du bruit présent dans les données et son impact sur les performances des algorithmes de classification. Les résultats montrent une efficacité notable de ces algorithmes pour classifier les motifs environnementaux, même en présence de bruit, avec des mesures de performance évaluées par des métriques internes et externes. Cette méthode permet d'automatiser et de renforcer le processus de surveillance, améliorant ainsi la précision et l'efficacité des systèmes de surveillance environnementale et permettant d'aider les experts scientifiques dans la prise de décision face à ces problématiques.

Table des matières

1 Introduction

Ce rapport aborde l'étude de la classification de motifs d'intérêts dans la surveillance environnementale, une problématique centrale pour la compréhension et la prédiction des phénomènes naturels. Avec l'augmentation de la disponibilité des données environnementales, la nécessité de développer des méthodes efficaces pour leur analyse est devenue cruciale.

Jusqu'à présent, diverses techniques de traitement et de classification des données ont été explorées, allant des méthodes traditionnelles de statistique aux approches plus récentes basées sur l'apprentissage automatique. Les travaux antérieurs, tels que ceux de Smith et al. [**smith2020environmental**] et Doe [**doe2019data**], ont posé les fondations en utilisant des techniques de clustering pour catégoriser les types de données environnementales, soulignant l'importance de distinguer les formes spécifiques des séries temporelles pour une analyse précise.

L'objectif de cette étude est double. Premièrement, nous cherchons à améliorer la compréhension des différentes formes de données environnementales par une classification précise utilisant des méthodes de clustering avancées, telles que K-means et UK-means. Deuxièmement, nous visons à évaluer l'impact du bruit sur la fiabilité des classifications produites, une préoccupation majeure dans les études environnementales où les données sont souvent corrompues par des interférences externes.

Le corps du document est structuré comme suit :

— Le Chapitre 2 présente un état de l'art des techniques utilisées pour l'analyse des données environnementales, mettant en lumière les avancées récentes et les lacunes existantes.

— Le Chapitre 3 décrit en détail la méthodologie adoptée, incluant la préparation des données, la description des algorithmes de clustering utilisés et les techniques pour mesurer le bruit.

— Le Chapitre 4 détaille les résultats obtenus à travers les différentes configurations expérimentales, illustrant comment les variations de bruit affectent les performances des méthodes de clustering.

— Le Chapitre 5 discute de la pertinence des résultats et de leur implication pour des applications pratiques.

— La conclusion, au Chapitre 6, résume les contributions principales de l'étude et propose des directions pour les recherches futures.

Ces sections sont conçues pour guider le lecteur à travers les processus d'analyse, offrant à la fois une vue technique et pratique des défis et des solutions proposées dans la classification des données environnementales bruitées.

2 État de l'art

L'analyse des motifs environnementaux par classification automatique a connu des progrès significatifs avec l'adoption de techniques de clustering avancées. Le clustering K-means, bien que largement utilisé pour sa simplicité et son efficacité, présente des limites, notamment la nécessité de définir à l'avance le nombre de clusters et sa sensibilité aux valeurs initiales des centres de clusters. Ces défis sont partiellement surmontés par des extensions telles que K-means++ et des adaptations comme UK-means, qui propose une approche non supervisée capable de déterminer automatiquement le nombre optimal de clusters, comme discuté par Sinaga et Yang [1]. Cette méthode, qui incorpore un terme de pénalité d'entropie pour ajuster les biais, montre une avancée notable dans le domaine du clustering non supervisé, facilitant l'analyse de grandes bases de données sans intervention manuelle préalable.

Dans le cadre de cette revue sur le clustering de séries temporelles de la dernière décennie, Aghabozorgi et ses collaborateurs [2] ont mis en lumière des avancées importantes en termes d'efficacité, de qualité et de complexité des méthodes de clustering appliquées à des données chronologiques. Leurs travaux montrent une augmentation significative de la recherche sur des solutions non supervisées telles que les algorithmes de clustering, pour extraire des connaissances de vastes ensembles de données. Cette revue, en se concentrant spécifiquement sur les données de séries temporelles, expose quatre composantes principales du clustering de séries temporelles et vise à présenter une investigation actualisée sur les tendances d'améliorations des méthodes au cours de la dernière décennie, ouvrant ainsi de nouvelles voies pour des travaux futurs.

3 Méthodologie

Dans cette étude, l'application d'un algorithme de classification sur les données de séries temporelles ne permet pas de conclure sur la justesse des résultats en l'absence d'une référence claire. Par conséquent, une analyse visuelle et qualitative des séries temporelles est nécessaire. Cette analyse consiste à zoomer sur différentes plages temporelles pour identifier des familles de formes d'intérêt. Ensuite, nous générons différentes variantes de ces formes en ajustant des paramètres spécifiques tels que la largeur, la hauteur, et la fréquence des motifs. Nous appliquons donc les méthodes de classification, avant de les étendre aux données réelles, sur les données que nous avons générées. Cette approche est particulièrement pertinente car nous avons fixé la longueur des formes dans nos données, ce qui facilite l'évaluation initiale des résultats.

Une fois ces familles de formes identifiées, nous générons diverses formes pour chaque famille à l'aide de modèles mathématiques simples, permettant de moduler différents paramètres tels que la largeur et la hauteur des pics. En parallèle, différents niveaux de bruit sont ajoutés aux formes générées afin d'étudier leur effet sur la classification.

Les données ainsi générées sont ensuite testées avec l'algorithme de classification K-means, en variant le paramètre k. Cette étape est cruciale pour valider la méthode de classification. Cependant, même si les résultats sont satisfaisants, l'algorithme K-means a ses limites, notamment la nécessité de connaître à l'avance le nombre de classes. C'est pourquoi nous utilisons également l'algorithme UK-means, qui se passe de cette connaissance préalable tout en offrant une performance théorique supérieure et une complexité similaire.

La performance de ces algorithmes est évaluée à l'aide de métriques de précision internes et externes. Une fois validés, ces algorithmes sont appliqués aux données réelles pour obtenir les résultats finaux. Cette approche permet de s'assurer de la robustesse et de la fiabilité des méthodes de classification utilisées dans cette étude.

3.1 Analyse des données

3.1.1 Identification des familles des formes d'intérêt

Les motifs d'intérêts sont des phénomènes observés à l'oeil dans les données fournies sur plusieurs mois. Les indices mentionnés se réfèrent à des intervalles dans les données, et différentes formes de motifs sont identifiées et visualisées à l'aide de sous-graphes dans des figures.

Mois 1 - Février	
Indices	**Description**
1575 à 1581	Très grande émission dûe à une erreur de mesure.
3000 à 4500	Croissance lente suivie d'une décroissance rapide (CL-DR).
8000 à 8800	Forme de "cloche".
10300 à 11000	Forme de "cloche".
18300 à 19200	Motif inconnu.
39500	Motif incomplet.

TABLE 1 – Tableau récapitulatif des phénomènes d'intérêt pour le mois de Février

Mois 2 - Avril	
Indices	**Description**
200 à 1600	Croissance lente, pic rapide, et décroissance lente.
1700 à 3550	Trois petites cloches.
3800 à 4600	Forme de "M".
20200 à 20700	Croissance rapide puis décroissance lente (CR-DL).

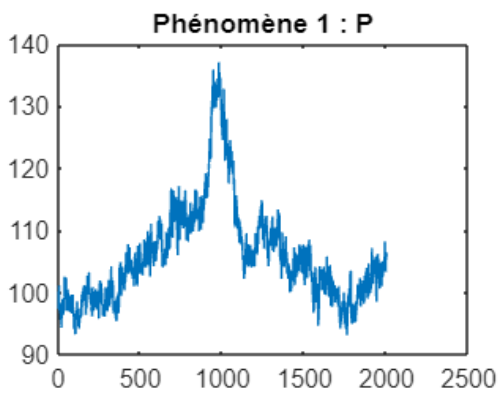
TABLE 2 – Tableau récapitulatif des phénomènes d'intérêt pour le mois d'Avril

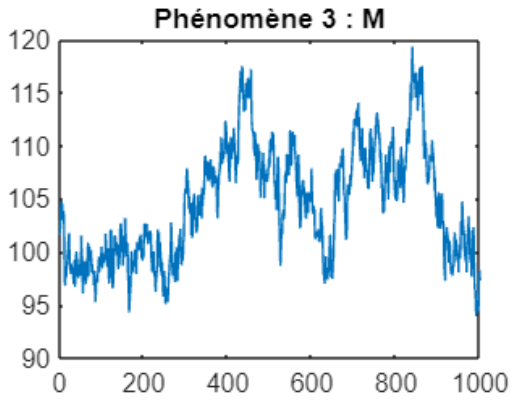
Mois 3 - Juin	
Indices	**Description**
4400 à 4500	Pic.
35400 à 36050	Deux CR-DL successifs.
38500 à 38900	CR-DL.

TABLE 3 – Tableau récapitulatif des phénomènes d'intérêt pour le mois de Juin

Mois 4 - Octobre	
Indices	**Description**
2600 à 3500	CL-DR.
10700 à 11400	Forme de "M".

TABLE 4 – Tableau récapitulatif des phénomènes d'intérêt pour le mois d'Octobre





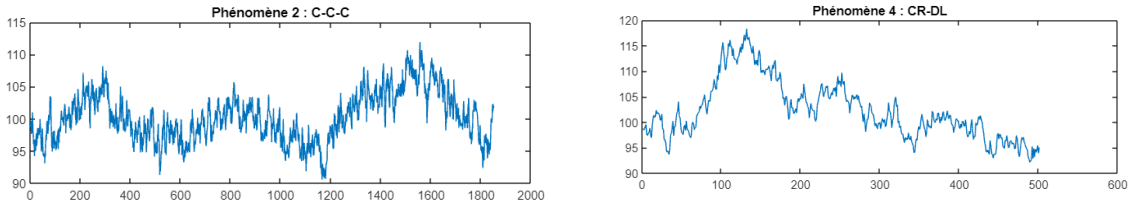


FIGURE 1 – Allure des différentes familles de formes identifiées : Pics, Cloches, Formes en M et Formes en Croissance Rapide Décroissance Lente

3.1.2 Génération Paramétrée des Formes Basée sur des Modèles Mathématiques

Dans le cadre de notre projet, nous avons développé un ensemble de méthodes pour générer des formes paramétrées basées sur des modèles mathématiques. Ces formes sont utilisées pour la classification et la détection de motifs d'intérêt dans les données de surveillance de l'environnement. L'une des formes les plus représentatives générées est celle de la "Croissance Rapide et Décroissance Lente" (CR-DL). Ce modèle est composé de deux segments : une croissance exponentielle et une décroissance linéaire, permettant de modéliser des phénomènes avec une montée rapide suivie d'une descente progressive.

3.1.3 Croissance lente Décroissance rapide

La croissance exponentielle est modélisée pour x appartenant à l'intervalle $[a, c]$ par l'équation :

$$y = A\left(e^{\alpha(x-a)} - 1\right)$$

où :

— A est l'amplitude de la croissance.
— α est un paramètre de l'exponentielle.
— a est le début de l'intervalle de croissance.
— c est la fin de l'intervalle de croissance.

La décroissance linéaire est modélisée pour x appartenant à l'intervalle $[c, b]$ par l'équation :

$$y = mx + k$$

où :

— m et k sont des paramètres déterminés pour assurer la continuité en $x = c$.
— b est la fin de l'intervalle de décroissance.

Les valeurs de m et k sont calculées en assurant la continuité des segments :

$$m = \frac{D - A(e^{\alpha(c-a)} - 1)}{b - c}$$

$$k = A(e^{\alpha(c-a)} - 1) - mc$$

où D est la valeur finale atteinte en $x = b$.

Cette modélisation permet de créer une courbe qui représente fidèlement les phénomènes de croissance rapide suivie de décroissance lente observés dans les données environnementales. La figure ci-dessous montre l'exemple de cette forme générée.

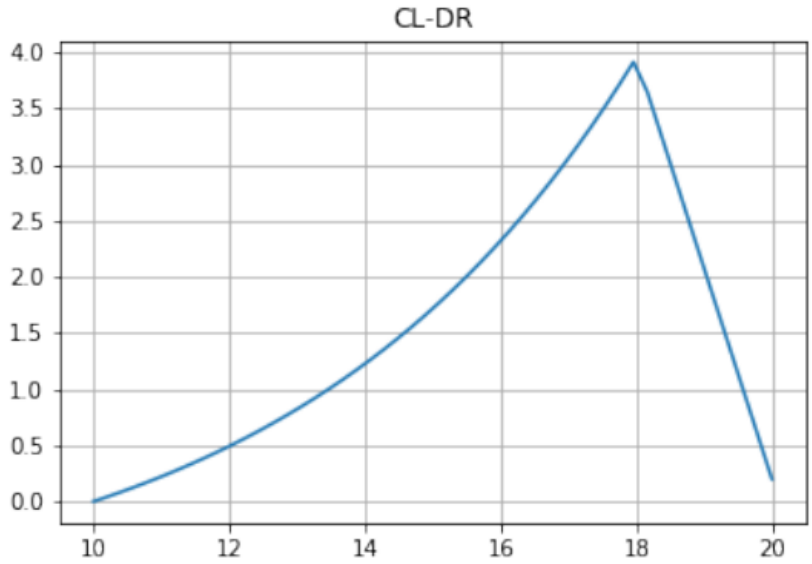


FIGURE 2 – Exemple de forme générée : Croissance Lente et Décroissance Rapide (CR-DL)

3.1.4 Forme en Cloche (Parabole)

La forme en cloche est modélisée par une parabole définie par l'équation :

$$y = a(x - h)^2 + k$$

où :

— a contrôle la largeur, la convexité/concavité (le signe) de la parabole.
— h est l'abscisse du sommet de la parabole.
— k est l'ordonnée du sommet de la parabole.

Cette modélisation permet de créer différentes variantes de cloches qui représentent en fonctions des paramètres les phénomènes symétriques de montée et descente observés dans les données environnementales.

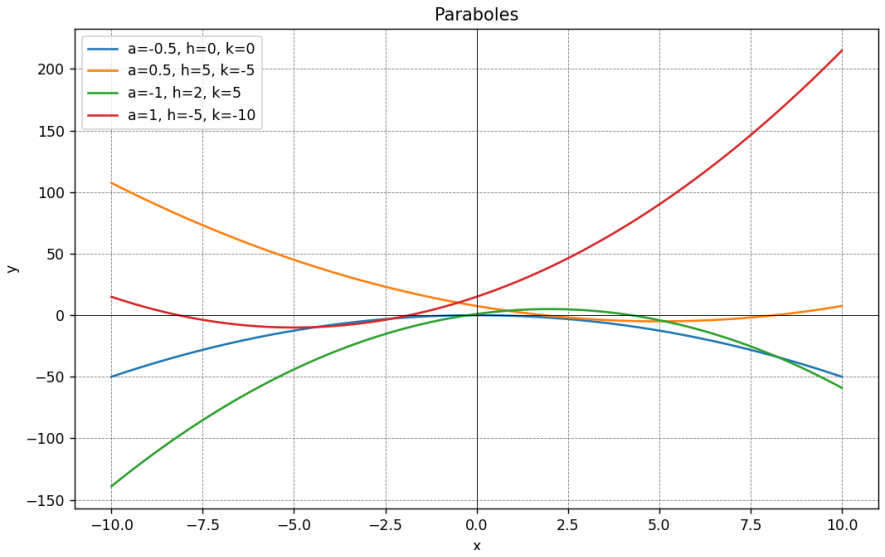


FIGURE 3 – Exemples de formes générées : Forme en Cloche (Parabole)

3.1.5 Forme de Pic

La forme de pointe est modélisée par une fonction triangulaire symétrique. Soit t le temps et h la hauteur maximale de la pointe. La fonction peut être définie par :

$$y(t) = \begin{cases} \frac{2h}{T}t & \text{pour } 0 \leq t < \frac{T}{2} \\ h - \frac{2h}{T}(t - \frac{T}{2}) & \text{pour } \frac{T}{2} \leq t \leq T \end{cases}$$

où T est la durée totale de la pointe. Cette équation assure une montée linéaire jusqu'à $t = \frac{T}{2}$, suivie d'une descente linéaire jusqu'à $t = T$.

La forme de pointe peut également être générée avec une durée spécifique. Soit T_d la durée de la pointe. La fonction est alors définie par :

$$y(t) = \begin{cases} \frac{2h}{T_d}t & \text{pour } 0 \leq t < \frac{T_d}{2} \\ h - \frac{2h}{T_d}(t - \frac{T_d}{2}) & \text{pour } \frac{T_d}{2} \leq t \leq T_d \end{cases}$$

où h est la hauteur maximale de la pointe, et T_d est la durée de la pointe.

Cette modélisation permet de créer une courbe qui représente fidèlement les phénomènes de montée et descente rapides observés dans les données environnementales. La figure ci-dessous montre l'exemple de cette forme générée.

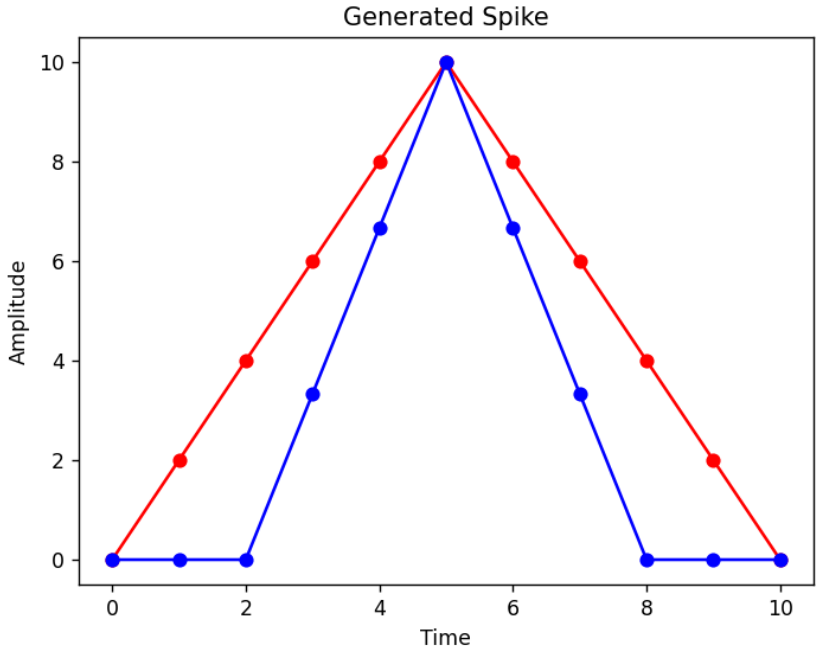


FIGURE 4 – Forme de Pic

3.1.6 Forme en "M"

La forme en "M" est modélisée en divisant la largeur totale width en quatre segments égaux et en définissant des pentes pour chaque segment. Soit x l'axe des abscisses et y l'axe des ordonnées, la fonction peut être définie par les segments suivants :

1. **Premier segment** (montée) :

$$y = \frac{4h}{\text{width}}x \quad \text{pour } 0 \leq x < \frac{\text{width}}{4}$$

2. **Deuxième segment** (descente) :

$$y = 2h - \frac{4h}{\text{width}}(x - \frac{\text{width}}{4}) \quad \text{pour } \frac{\text{width}}{4} \leq x < \frac{\text{width}}{2}$$

3. **Troisième segment** (montée) :

$$y = -2h + \frac{4h}{\text{width}}(x - \frac{\text{width}}{2}) \quad \text{pour } \frac{\text{width}}{2} \leq x < \frac{3\text{width}}{4}$$

4. **Quatrième segment** (descente) :

$$y = 4h - \frac{4h}{\text{width}}(x - \frac{3\text{width}}{4}) \quad \text{pour } \frac{3\text{width}}{4} \leq x < \text{width}$$

où h est la hauteur maximale de chaque segment.

Cette modélisation permet de créer une courbe qui représente les phénomènes de montée et descente rapides observés dans les données environnementales, formant un motif en "M". La figure ci-dessous montre l'exemple de cette forme générée.

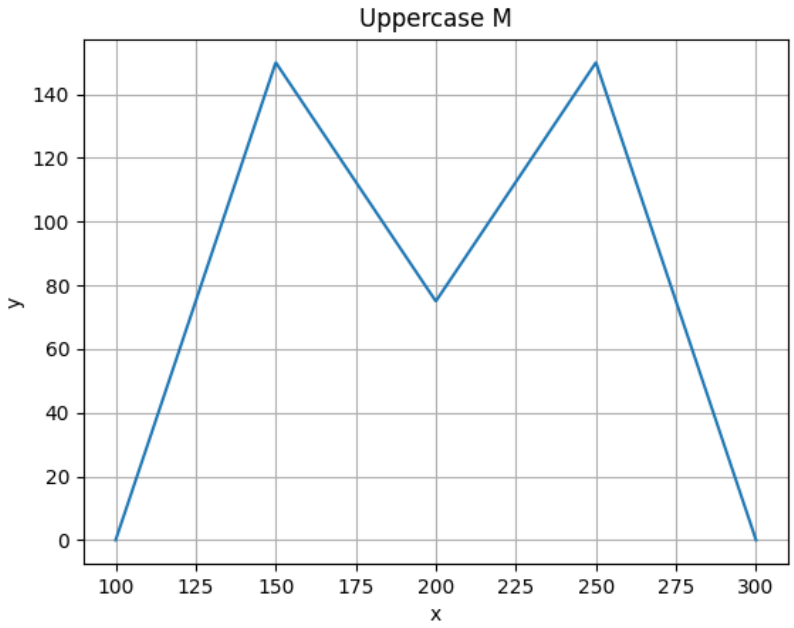
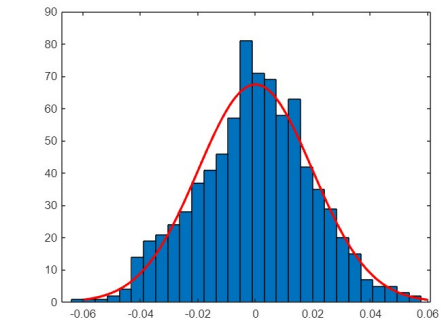


FIGURE 5 – Exemple de forme générée : Forme en "M"

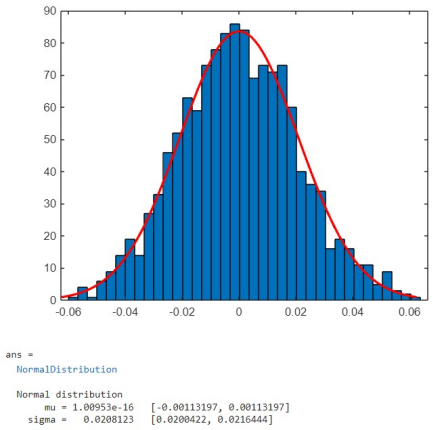
3.1.7 Analyse du bruit des données mesurées

L'étude du bruit dans les données temporelles est cruciale pour l'optimisation des processus de clustering. Afin de cerner avec précision les caractéristiques du bruit inhérent à nos données, nous avons employé une méthode spécifique qui consiste à analyser des segments où les séries temporelles sont statistiquement constantes. Cela nous permet d'isoler le bruit en soustrayant la valeur moyenne du signal sur chaque intervalle sélectionné.
Les segments analysés ont été soigneusement choisis pour représenter des périodes où l'activité mesurée reste stable, ce qui rend l'évaluation du bruit plus fiable.

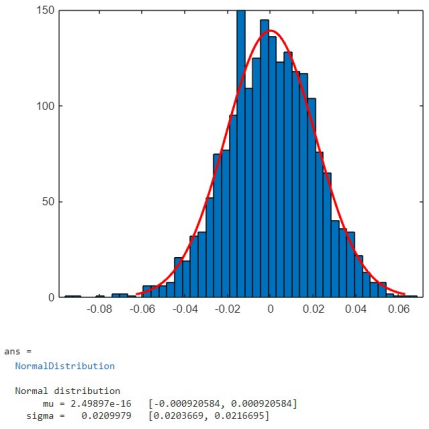
Cette approche méthodique assure que le bruit analysé est représentatif des fluctuations aléatoires du signal et non des variations dues à des changements dans les conditions mesurées. Les résultats obtenus sont visualisés dans les histogrammes suivants (Figure 6), avec les ajustements de distributions normales pour chaque période analysée.



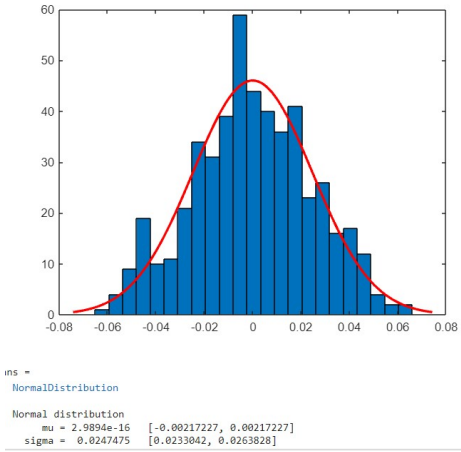
(a) Premier intervalle



(b) Deuxième intervalle



(c) Troisième intervalle



(d) Quatrième intervalle

FIGURE 6 – Analyse de la distribution du bruit pour différents intervalles de temps

Les résultats montrent que la variance du bruit reste cohérente à travers les différents intervalles, validant ainsi notre hypothèse de la stabilité du bruit au sein des séries temporelles analysées. Cette constance est essentielle pour l'application de techniques de clustering robustes et efficaces. On note une valeur de 0.02 pour la variance du bruit, valeur qui sera utilisée plus tard lors des tests des algorithmes de clustering.

3.2 K-means

La méthode K-means est une technique de classification non supervisée qui partitionne les observations en k clusters, dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche. Cette méthode est largement utilisée pour son efficacité et sa simplicité. Elle s'appuie sur la minimisation de la fonction de coût :

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

où :
— k est le nombre de clusters,
— C_i est le i-ème cluster,
— μ_i est le centroïde du i-ème cluster.
On utilisera ces mêmes notations dans la suite du rapport.

Cependant K-means est sensible aux initialisations et nécessite qu'on lui fournisse le nombre de classe. En ce sens, il n'est pas réellement non supervisé. C'est pourquoi nous allons développer l'algorithme UK-means qui ne requiert pas le nombre de classe.

3.3 U-Kmeans

La méthode U-Kmeans est une variante de l'algorithme K-means qui introduit des contraintes supplémentaires pour améliorer la qualité des clusters, notamment en tenant compte des incertitudes dans les données. Cet algorithme utilise un terme de pénalité d'entropie pour ajuster le biais et un schéma d'apprentissage pour trouver le nombre de clusters.

On explique ici le principe de l'algorithme développé en [2].

Soit $X = \{x_1, x_2, \ldots, x_n\}$ un ensemble de données dans un espace euclidien d-dimensionnel $\mathbb{R}^d$. Soit $A = \{a_1, a_2, \ldots, a_c\}$ les centres des clusters et $z = [z_{ik}]_{n \times c}$, où z_{ik} est une variable binaire indiquant si le point de données x_i appartient au k-ième cluster.

L'objectif de l'algorithme k-means est de minimiser la fonction objective $J(z, A)$:

$$J(z, A) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2$$

On utilise donc $-\ln \alpha_k$ comme l'information sur l'appartenance d'un point de données à la k-ième classe, et donc $-\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ devient la moyenne de l'information. En fait, le terme $-\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ représente l'entropie des proportions α_k. Lorsque $\alpha_k = 1/c, \forall k = 1, 2, \ldots, c$, nous disons qu'il n'y a pas d'information sur α_k. Nous ajoutons un ce terme à la fonction objective du k-means $J(z, A)$ en tant que pénalité. Nous construisons ensuite un schéma pour estimer α_k en minimisant l'entropie pour obtenir le plus d'informations pour α_k. Minimiser $-\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ équivaut à maximiser $\sum_{k=1}^{c} \alpha_k \ln \alpha_k$. Pour cette raison, nous utilisons $\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ comme terme de pénalité pour la fonction objective du k-means $J(z, A)$. Ainsi, nous proposons une nouvelle fonction objective comme suit : $\beta \geq 0$

$$J_{UKM1}(z, A, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k$$

Pour déterminer le nombre de clusters, nous considérons ensuite un autre terme d'entropie. Nous combinons les variables d'appartenance z_{ik} et la proportion α_k. En utilisant les bases de la théorie de l'entropie, nous proposons un nouveau terme sous la forme de $z_{ik} \ln \alpha_k$. Ainsi, nous proposons la fonction objective du k-means non supervisé (U-k-means) comme suit :

$$J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k - \gamma \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \ln \alpha_k$$

Nous combinons les variables d'appartenance z_{ik} et la proportion α_k. En utilisant la théorie de l'entropie, nous suggérons un nouveau terme sous la forme $z_{ik} \ln \alpha_k$. Ainsi, nous proposons la fonction objective des k-means non supervisés (U-k-means) comme suit :

$$J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k - \gamma \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \ln \alpha_k$$

La minimisation de cette fonction conduit aux équations :

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 - \gamma \ln \alpha_k = \min\limits_{1 \le k \le c} \|x_i - a_k\|^2 - \gamma \ln \alpha_k \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$\alpha_k^{(t+1)} = \sum_{i=1}^{n} \frac{z_{ik}}{n} + \frac{\beta}{\gamma} \alpha_k^{(t)} \left(\ln \alpha_k^{(t)} - \sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \right) \tag{2}$$

$$\beta^{(t+1)} = \min \left(\frac{\sum_{k=1}^{c} \exp(-\eta n \left| \alpha_k^{(t+1)} - \alpha_k^{(t)} \right|)}{c}, \frac{1 - \max\limits_{1 \le k \le c} \left(\frac{1}{n} \sum_{i=1}^{n} z_{ik} \right)}{- \max\limits_{1 \le k \le c} \alpha_k^{(t)} \sum_{k'=1}^{c} \ln \alpha_{k'}^{(t)}} \right) \tag{3}$$

$$\alpha_k^* = \frac{\alpha_k^*}{\sum_{s=1}^{c^{(t+1)}} \alpha_s^*} \tag{4}$$

$$z_{ik}^* = \frac{z_{ik}^*}{\sum_{s=1}^{c^{(t+1)}} z_{is}^*} \tag{5}$$

On en déduit donc le pseudo-code de l'algorithme ukmeans :

Algorithm 1 Algorithme de clustering U-k-means

Require: $\epsilon > 0$, donné initialement $c^{(0)} = n$, $\alpha_k^{(0)} = \frac{1}{n}$, $a_k^{(0)} = x_i$, et les taux d'apprentissage initiaux $\gamma^{(0)} = \beta^{(0)} = 1$. Fixer $t = 0$.

Étape 1 : Fixer $\epsilon > 0$. Donner initialement $c^{(0)} = n$, $\alpha_k^{(0)} = \frac{1}{n}$, $a_k^{(0)} = x_i$, et les taux d'apprentissage initiaux $\gamma^{(0)} = \beta^{(0)} = 1$. Fixer $t = 0$.
Étape 2 : Calculer $z_{ik}^{(t+1)}$ en utilisant $a_k^{(t)}$, $\alpha_k^{(t)}$, $c^{(t)}$, $\gamma^{(t)}$, $\beta^{(t)}$ par (1).
Étape 3 : Calculer $\gamma^{(t+1)}$ par $\gamma^{(t)} = e^{-c^{(t)}/250}$.
Étape 4 : Mettre à jour $\alpha_k^{(t+1)}$ avec $z_{ik}^{(t+1)}$ et $\alpha_k^{(t)}$ par (2).
Étape 5 : Calculer $\beta^{(t+1)}$ avec $\alpha^{(t+1)}$ et $\alpha^{(t)}$ par (3).
Étape 6 : Mettre à jour $c^{(t)}$ à $c^{(t+1)}$ en supprimant les clusters avec $\alpha_k^{(t+1)} \leq \frac{1}{n}$ et ajuster $\alpha_k^{(t+1)}$ et $z_{ik}^{(t+1)}$ par (4) et (5).
if $t \geq 60$ et $c^{(t-60)} - c^{(t)} = 0$ **then**
    Poser $\beta^{(t+1)} = 0$.
end if
Étape 7 : Mettre à jour $a_k^{(t)}$ avec $c^{(t+1)}$ et $z_{ik}^{(t+1)}$ par $a_k = \frac{\sum_{i=1}^{n} z_{ik} x_{ij}}{\sum_{i=1}^{n} z_{ik}}$.
Étape 8 : Comparer $a_k^{(t+1)}$ et $a_k^{(t)}$.
if $\max_{1 \leq k \leq c^{(t)}} \left\| a_k^{(t+1)} - a_k^{(t)} \right\| < \epsilon$ **then**
    Arrêter.
else
    $t = t + 1$ et revenir à l'Étape 2.
end if

3.4 Métriques d'évaluation

Dans le cadre de l'évaluation de la performance de nos algorithmes pour la classification des motifs en surveillance de l'environnement, il est essentiel d'utiliser des métriques de performance qui permettent de quantifier l'efficacité et la précision de la méthode de clustering choisie. Ces métriques se divisent en deux catégories principales : les métriques d'évaluation internes et les métriques d'évaluation externes, chacune ayant ses spécificités et son domaine d'application.

Les **métriques d'évaluation externes** requièrent une vérité terrain, c'est-à-dire des étiquettes pré-définies, pour comparer les clusters obtenus. Ces métriques permettent d'évaluer la concordance entre les clusters formés par l'algorithme de clustering et les classifications pré-existantes, offrant ainsi une mesure de la précision de l'assignation des points aux clusters. L'accuracy, l'Adjusted Rand Index (ARI), et la Normalized Mutual Information (NMI) sont parmi les métriques externes couramment adoptées pour évaluer la correspondance entre le résultat du clustering et les étiquettes réelles.

En contraste, les **métriques d'évaluation internes** sont utilisées pour mesurer la qualité du clustering sans nécessiter d'informations externes. Elles se basent uniquement sur les données elles-mêmes et l'output du clustering. Ces métriques évaluent principalement à quel point les clusters sont cohérents et bien séparés. Parmi les plus utilisées, on trouve le Score de Silhouette, l'Index de Calinski-Harabasz, et

l'Index de Davies-Bouldin, qui fournissent des indices sur la compacité et la séparation des clusters formés.

L'objectif de ces métriques est de fournir une évaluation globale de la performance de l'algorithme de clustering. Les métriques internes nous permettent de comprendre l'intégrité structurelle des clusters sans considération pour les étiquettes, tandis que les métriques externes jugent l'efficacité de la classification en relation avec un standard ou une attente prédéfinie. Ensemble, elles offrent une vue complète de la performance du clustering, relevant à la fois les forces et les faiblesses de l'approche adoptée.

Les sections suivantes détaillent chacune des métriques qui seront utilisées pour évaluer nos modèles.

3.4.1 Métriques d'évaluation externes

Les métriques d'évaluation externes sont utilisées pour mesurer à quel point la classification correspond aux étiquettes préétablies. Les métriques suivantes ont été prises en compte pour évaluer la performance de K-Means dans nos tests :

— **Accuracy :** Indique la proportion de prédictions correctes (c'est-à-dire le nombre de formes correctement classifiées par rapport au total des formes). C'est la mesure la plus intuitive de la performance.

$$\text{Accuracy} = \frac{\text{nombre de prédictions correctes}}{\text{nombre total de prédictions}}$$

— **Précision (Precision) :** La précision est le ratio du nombre de vrais positifs sur le nombre total de cas classés positifs par le modèle. Elle est définie comme suit :

$$\text{Précision} = \frac{TP}{TP + FP}$$

où :

— TP (True Positives) est le nombre de paires correctement assignées ensemble.

— FP (False Positives) est le nombre de paires incorrectement assignées ensemble.

— **Rappel (Recall) ou Sensibilité :** Le rappel est le ratio du nombre de vrais positifs sur le nombre total de cas qui sont réellement positifs. Il est défini comme suit :

$$\text{Rappel} = \frac{TP}{TP + FN}$$

où FN (False Negatives) est le nombre de paires qui appartiennent à la même classe mais ne sont pas assignées ensemble.

— **Score F1 (F1 Score) :** Le score F1 est la moyenne harmonique de la précision et du rappel. Il prend en compte à la fois la précision et le rappel pour calculer le score. Le score F1 est particulièrement utile lorsque les distributions des classes sont déséquilibrées. Il est défini comme suit :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

— **Rand Index (RI) :** Le Rand Index est une mesure de la similarité entre deux affectations de clustering, calculée en considérant toutes les paires de points et en évaluant si les décisions de regroupement sont identiques ou différentes entre les deux clustering. Notons C le *vrai* clustering, et K celui résultant d'un algorithme de classification. Alors le RI est défini comme suit :

$$RI = \frac{a+b}{\binom{n}{2}}$$

où :
— a est le nombre de paires d'éléments qui sont dans le même ensemble dans C et dans le même ensemble dans K.
— b est le nombre de paires d'éléments qui sont dans des ensembles différents dans C et dans des ensembles différents dans K.
— $\binom{n}{2}$ est le nombre total de paires possibles dans le dataset
Bien que le RI soit une mesure intuitive et facile à comprendre pour évaluer la qualité du clustering, il a une limitation majeure : il ne prend pas en compte la possibilité que des correspondances correctes puissent survenir par hasard. Nous préférerons donc utiliser l'ARI.

— **Adjusted Rand Index (ARI) :** Mesure la similarité entre deux affectations en tenant compte de la chance. Il est défini par la formule suivante :

$$ARI = \frac{\text{RI} - \text{Expected RI}}{\max(\text{RI}) - \text{Expected RI}}$$

— **Normalized Mutual Information (NMI) :** Évalue l'information partagée entre les étiquettes prédites et les vraies étiquettes ; des valeurs plus élevées indiquent un plus grand degré de correspondance.

$$NMI(X,Y) = \frac{2 \times I(X;Y)}{H(X) + H(Y)}$$

où $I(X,Y)$ est l'information mutuelle, $H(X)$ et $H(Y)$ sont les entropies. Nous ne rentrerons pas dans les définitions de ces termes, issues de la théorie de l'information. Nous calcurons simplement cette métrique à titre indicatif, à l'aide de `scikit-learn`.

— **Fowlkes-Mallows Index (FMI) :** Calculé comme la moyenne géométrique de la précision et du rappel des clusters ; il mesure la tendance des paires d'éléments à être assignées ensemble de manière correcte.

$$FMI = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}$$

— **Jaccard Score :** Évalue la similarité entre les membres du même cluster par rapport aux véritables étiquettes, fournissant une mesure de l'intersection sur l'union des ensembles de cluster et d'étiquettes.

$$Jaccard = \frac{TP}{TP+FP+FN}$$

L'usage combiné de ces métriques externes offre une perspective complète sur la performance de l'algorithme de clustering, reflétant non seulement la précision de l'assignation des classes mais aussi la pertinence des groupements effectués par rapport aux classes réelles. Cependant, nous focaliserons notre étude sur la précision, le rappel et surtout l'accuracy car c'est la métrique la plus intuitive.

3.4.2 Métriques d'évaluation internes

Les métriques d'évaluation internes sont cruciales pour évaluer la cohérence interne et la qualité des clusters formés par un algorithme de clustering. Elles ne nécessitent pas d'informations externes telles que des étiquettes de classe et sont basées uniquement sur les données de clustering elles-mêmes.

— **SSE (Somme des Carrés des Erreurs) :** Mesure l'erreur de clustering en calculant la somme des carrés des distances entre chaque point et le centre de son cluster. Un SSE faible indique que les clusters sont compacts et bien définis.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

où C_i est l'ensemble des points dans le cluster i et c_i est le centre du cluster i.

— **Score de Silhouette :** Cette métrique mesure à quel point chaque point dans un cluster est proche des points dans son cluster par rapport aux points dans le cluster le plus proche. Un score élevé indique que les clusters sont bien séparés et densément regroupés. Il est défini pour chaque point i comme suit :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

où $a(i)$ est la distance moyenne entre le point i et tous les autres points dans le même cluster, et $b(i)$ est la distance moyenne du point à son cluster le plus proche.

— **Indice de Calinski-Harabasz :** Il est défini comme le rapport entre la dispersion inter-cluster (BCSS) et la dispersion intra-cluster (WCSS), normalisé par leur nombre de degrés de liberté, pour un ensemble de données de n points répartis en k clusters :

$$CH(k) = \frac{\text{BCSS}}{\text{WCSS}} \times \frac{n-k}{k-1}$$

où :

— $\text{BCSS} = \sum_{i=1}^{k} n_i \|c_i - c\|^2$ est la Somme des Carrés Inter-Clusters (avec n_i le nombre de points dans le cluster C_i, c_i le centroïde du cluster C_i, et c le centroïde global des données.)

— $\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$ est la Somme des Carrés Intra-Clusters.

— **Index de Davies-Bouldin :** Cette métrique évalue la moyenne du ratio de la distance intra-cluster à la distance inter-cluster. Des valeurs plus faibles de l'index indiquent une meilleure séparation des clusters.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

où σ_i est la dispersion moyenne des points dans le cluster i, c_i est le centroïde du cluster i, et $d(c_i, c_j)$ est la distance entre les centroïdes des clusters i et j.

Ces métriques permettent d'évaluer non seulement la compacité des clusters mais aussi leur isolation les uns par rapport aux autres, fournissant ainsi une vue générale de la performance de clustering.

4 Résultats

4.1 K-means

L'analyse des performances de l'algorithme K-Means a été effectuée en considérant trois scénarios distincts concernant le niveau de bruit : sans bruit, bruit similaire aux données réelles (variance = 0.02), et un niveau de bruit élevé (variance = 0.05). Nous avons utilisé un nombre fixe de clusters ($k = 5$) pour toutes les expérimentations lorsque ce n'est pas précisé.

Pour chaque scénario de bruit, nous avons calculé des métriques d'évaluation internes et externes afin de quantifier l'efficacité de l'algorithme K-Means dans la classification des formes environnementales variées.

4.1.1 Résultats Sans Bruit

Pour la *vraie* valeur $k = 5$:

Dans le scénario sans bruit, l'algorithme K-Means a été appliqué avec $k = 5$ clusters. Les résultats montrent une bonne séparation des différents motifs, comme illustré par les métriques de performance et confirmé visuellement par la Figure 7, où quelques courbes sont représentées, avec une coloration correspondant au cluster identifié.

Les métriques internes et externes obtenues sont les suivantes :

— Accuracy : 0.86
— Adjusted Rand Index (ARI) : 0.738
— Normalized Mutual Information (NMI) : 0.804
— Fowlkes-Mallows Index (FMI) : 0.789
— Jaccard Score : 0.776
— Score de Silhouette : 0.496
— Index de Calinski-Harabasz : 64.335
— Index de Davies-Bouldin : 0.892
— SSE : 21275

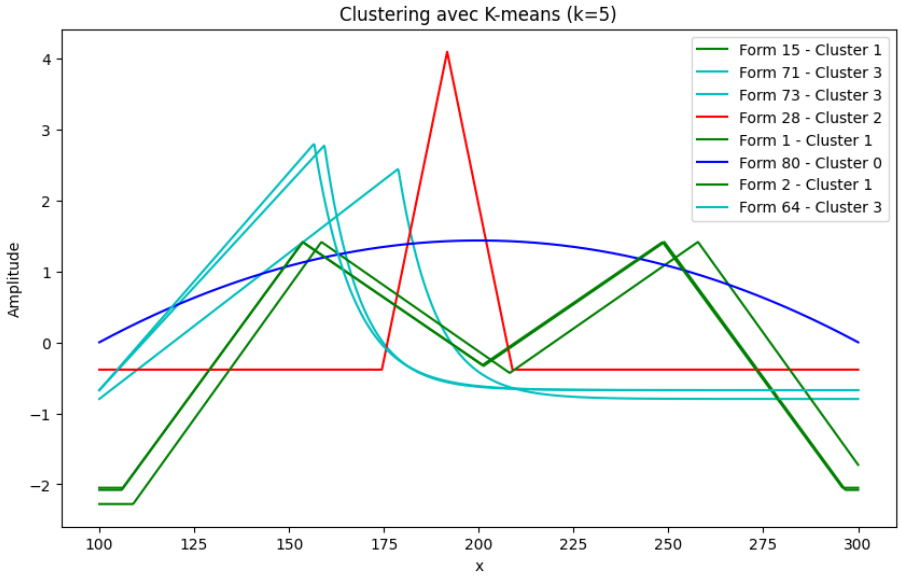


FIGURE 7 – Visualisation du clustering sans bruit

Nous avons aussi calculés la précision, le rappel, et le score F1 pour chaque classe. Ces valeurs sont regroupées dans la Table 5.

Classe	Précision	Rappel	Score F1
M	1.000	1.000	1.000
Spike	0.833	1.000	0.909
CLDR	0.667	0.600	0.632
CRDL	0.778	0.700	0.737
Parabola	1.000	1.000	1.000

TABLE 5 – Précision, rappel et score F1 pour chaque classe (sans bruit)

Un bon moyen d'évaluer la capacité du modèle à différencier les formes est la matrice de confusion. Elle fournit une représentation visuelle de la précision avec laquelle le modèle a classé les instances dans les clusters correspondants aux catégories réelles. Chaque ligne de la matrice représente les instances dans une classe réelle, tandis que chaque colonne représente les instances dans une classe prédite. Cette disposition nous permet de voir facilement le nombre de classifications correctes et incorrectes, offrant ainsi un aperçu direct de la performance du modèle pour chaque classe. La matrice de confusion pour notre scénario sans bruit est présentée ci-dessous, dans la Figure 8.

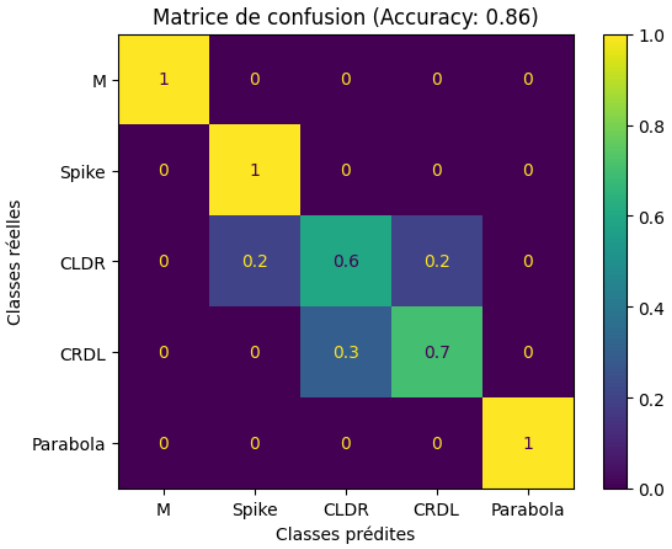
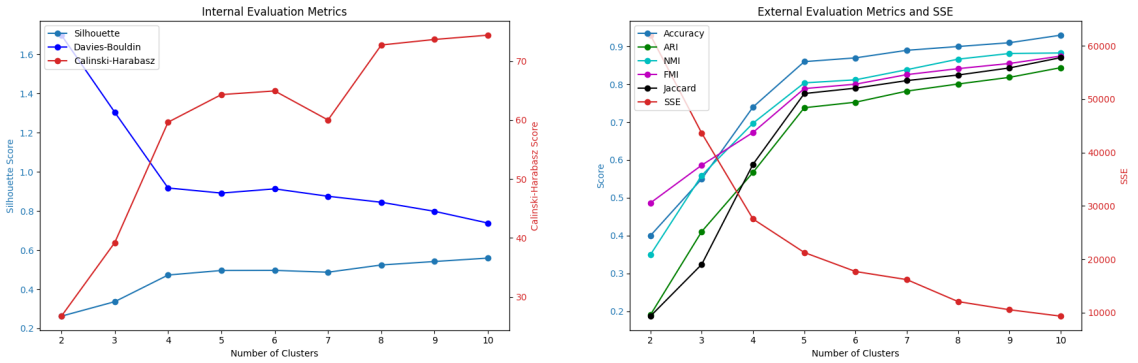


FIGURE 8 – Matrice de confusion pour des données non bruitées

Les résultats sans bruit indiquent une performance excellente pour les classes 'M' et 'Parabola', avec une précision et un rappel parfaits. Les défis demeurent pour les classes 'Spike', 'CRDL' et 'CLDR', où le rappel est haut mais la précision pourrait être améliorée. Ces résultats soulignent l'efficacité de K-Means dans des conditions idéales sans bruit.

Influence du nombre de classes k

Nous avons réalisé ce même test pour différentes valeurs du nombre de clusters k, afin d'en étudier son influence. Les métriques (externes et internes) ont été évaluées pour déterminer la qualité des clusters obtenus. Les résultats sont illustrés dans les Figures 9a et 9b.



(a) Métriques d'évaluation internes (b) Métriques d'évaluation externes et SSE

FIGURE 9 – Métriques en fonction de k (sans bruit)

Les métriques de performance, tant externes qu'internes, montrent des tendances distinctes à mesure que le nombre de clusters k varie. Pour les métriques externes,

notamment l'accuracy, l'ARI, la NMI, la FMI, et le score Jaccard, nous observons une amélioration générale des scores avec l'augmentation de k, atteignant un maximum pour $k = 10$. Il y a une amélioration significative jusqu'à $k = 6$, après quoi les scores tendent à se stabiliser, indiquant une saturation en termes de bénéfice apporté par l'ajout de clusters supplémentaires.

Pour les métriques internes, le Score de Silhouette augmente légèrement jusqu'à $k = 5$, puis se stabilise, ce qui suggère que la cohésion interne maximale est atteinte à ce point. L'Index de Davies-Bouldin, qui diminue avec l'augmentation de k, indique une meilleure séparation des clusters. En parallèle, l'Index de Calinski-Harabasz montre une augmentation avec k, ce qui confirme que les clusters deviennent plus distincts et mieux dispersés.

Point de transition à $k = 5$: Les résultats suggèrent un point de transition à $k = 5$, où ajouter des clusters améliore significativement les métriques. Au-delà de ce point, les gains deviennent marginaux.

Conclusion : Un nombre de clusters entre $k = 5$ et $k = 6$ semble optimal, capturant efficacement les structures sous-jacentes tout en conservant une bonne cohésion et séparation des clusters. Choisir k dans cette gamme offre un bon compromis entre la précision et la complexité du modèle de clustering. Cette observation est cohérente avec le fait que nous avons généré cinq familles distinctes de formes dans les données.

4.1.2 Avec un Bruit Similaire aux Données Réelles (variance = 0.02)

Pour ce scénario, l'Accuracy obtenue a été de 0.83, avec un ARI de 0.69 et un NMI de 0.77. Ces résultats montrent une performance respectable de K-Means malgré la présence de bruit. Les clusters obtenus pour quelques formes sont présentés dans la Figure 10.

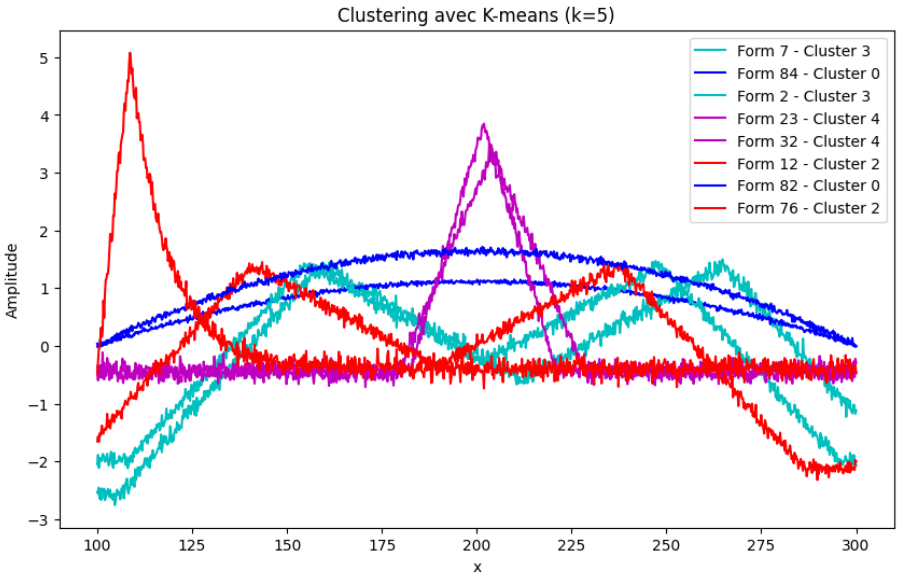


FIGURE 10 – Visualisation des clusters avec un bruit similaire aux données réelles

La matrice de confusion, Figure 11, montre une bonne précision de classification pour les formes M et Parabola, mais une confusion notable entre les formes Spike et CLDR/CRDL.

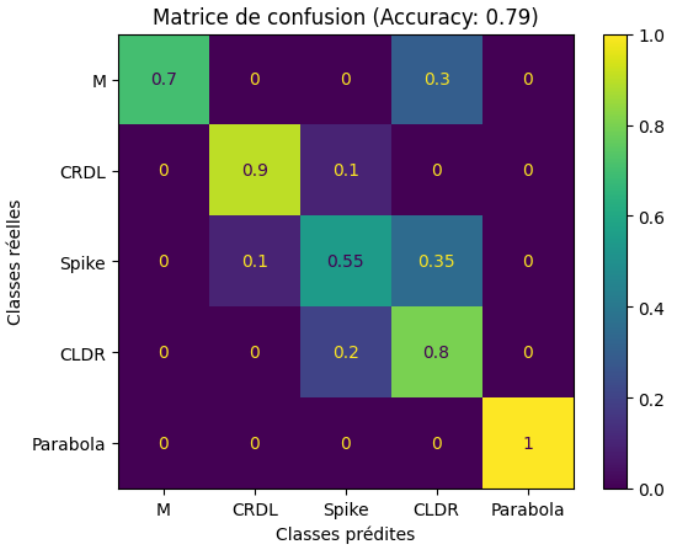


FIGURE 11 – Matrice de confusion avec un bruit similaire aux données réelles

Le tableau suivant résume les performances de précision, de rappel, et de score F1 pour chaque classe lors du clustering avec un bruit similaire aux données réelles. Ces métriques fournissent une mesure détaillée de l'efficacité de l'algorithme K-Means à bien classifier les formes en présence de bruit.

Classe	Précision	Rappel	Score F1
M	1.00	0.70	0.82
Spike	0.90	0.90	0.90
CLDR	0.65	0.55	0.59
CRDL	0.55	0.80	0.65
Parabola	1.00	1.00	1.00

TABLE 6 – Précisions, rappels, et scores F1 obtenus pour chaque classe avec un bruit similaire aux données réelles.

Les résultats montrent que bien que les classes M et Parabola affichent une excellente précision et un parfait rappel, les classes Spike, CLDR, et CRDL présentent des défis, en particulier pour le rappel des formes CLDR et la précision des formes CRDL. Cela souligne les difficultés du modèle à séparer efficacement certaines classes en présence de bruit, même si la performance globale reste bonne.

4.1.3 Avec Plus de Bruit (variance = 0.05)

Ce scénario, testé avec un niveau de bruit plus élevé (variance = 0.05), a étonnamment montré une performance de classification supérieure par rapport à un niveau de bruit légèrement inférieur (variance = 0.02). L'accuracy obtenue est de 0.83, une amélioration par rapport à l'accuracy de 0.79 observée avec moins de bruit (mais toujours plus faible que l'accuracy sans bruit qui était de 0.86).

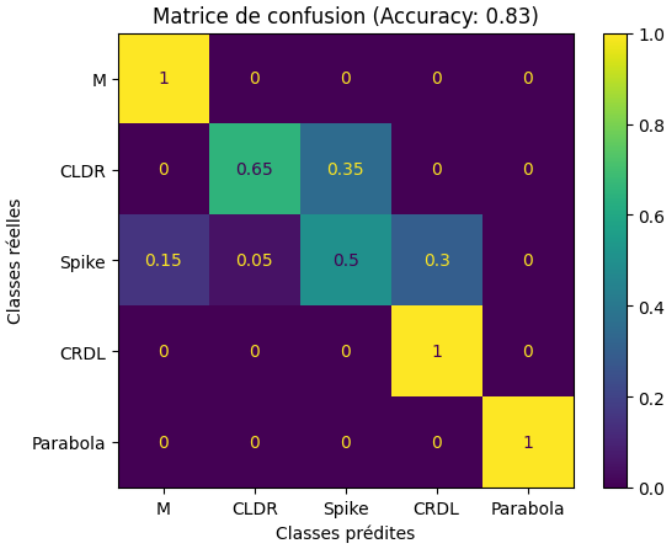


FIGURE 12 – Matrice de confusion avec un niveau élevé de bruit (*variance* = 0.05)

Comme le montre la Figure 12, bien que le bruit ait été augmenté, la précision de classification pour les formes M et Parabola reste parfaitement stable avec une précision de 1.0 pour chaque. Cependant, il existe une confusion notable entre les formes Spike et CLDR/CRDL, semblable à ce qui a été observé dans les scénarios précédents. Il est intéressant de noter que malgré l'augmentation du bruit, les performances pour les classes CRDL et Spike ont légèrement augmenté par rapport au scénario précédent, ce qui peut indiquer une certaine robustesse de l'algorithme K-Means. Cette observation suggère que dans certains cas, une augmentation modérée du bruit pourrait ne pas détériorer la performance de classification autant que prévu.

Après avoir identifié le nombre optimal de clusters k, nous approfondissons notre analyse en examinant l'impact direct du bruit sur la performance du clustering. En fixant k à une valeur optimale basée sur les analyses précédentes, nous pouvons isoler l'effet du bruit et observer ses répercussions spécifiques sur la précision, le rappel, et d'autres métriques importantes. La section suivante détaille donc les résultats obtenus avec différents degrés de bruit, offrant une perspective approfondie sur la robustesse de l'algorithme face à des défis environnementaux variables.

4.1.4 Influence du niveau de bruit

Cette section examine comment différents niveaux de bruit influencent l'accuracy de l'algorithme K-Means lors du clustering des données. L'objectif est de comprendre la robustesse de l'algorithme face à des perturbations variées et d'identifier le nombre optimal de clusters k qui maximise la précision dans des conditions bruitées.

Analyse de l'Accuracy en Fonction du Niveau de Bruit

Les résultats montrent une variation intéressante de l'accuracy en fonction du niveau de bruit pour différentes valeurs de k, comme le montre la Figure 13. Chaque ligne représente un nombre différent de clusters, permettant d'évaluer l'impact de k sur la résilience du modèle face au bruit.

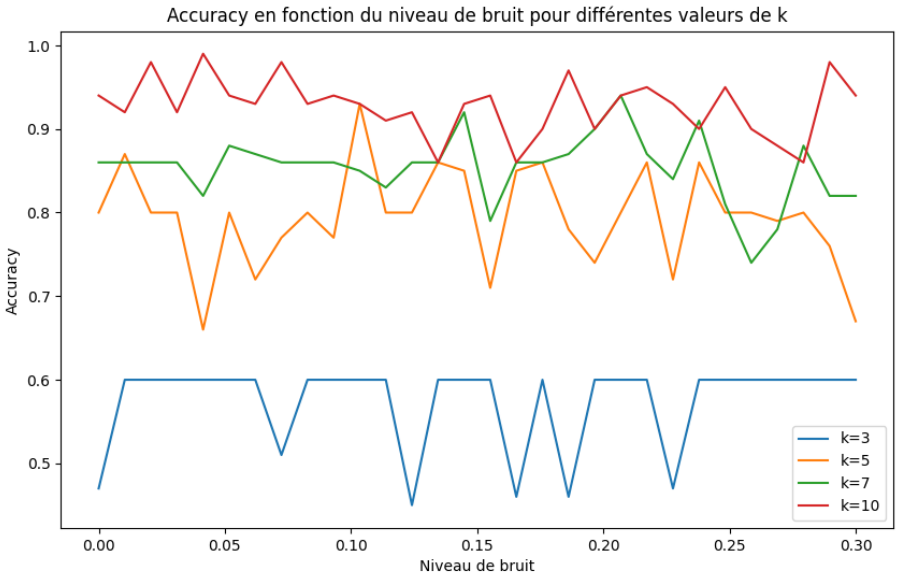


FIGURE 13 – Accuracy en fonction du niveau de bruit pour différentes valeurs de k

Il est évident que l'accuracy fluctue avec l'augmentation du bruit pour tous les k. Notamment, pour $k = 5$ et $k = 7$, l'accuracy reste relativement stable jusqu'à un certain seuil de bruit avant de diminuer. Cela suggère une tolérance jusqu'à un point critique au-delà duquel le bruit commence à impacter négativement la classification.

Visualisation 3D de l'Accuracy

Pour une compréhension plus profonde, une visualisation 3D de l'accuracy en fonction de k et du niveau de bruit est présentée dans la Figure 14. Cette représentation aide à visualiser simultanément les effets combinés de k et du bruit sur l'accuracy.

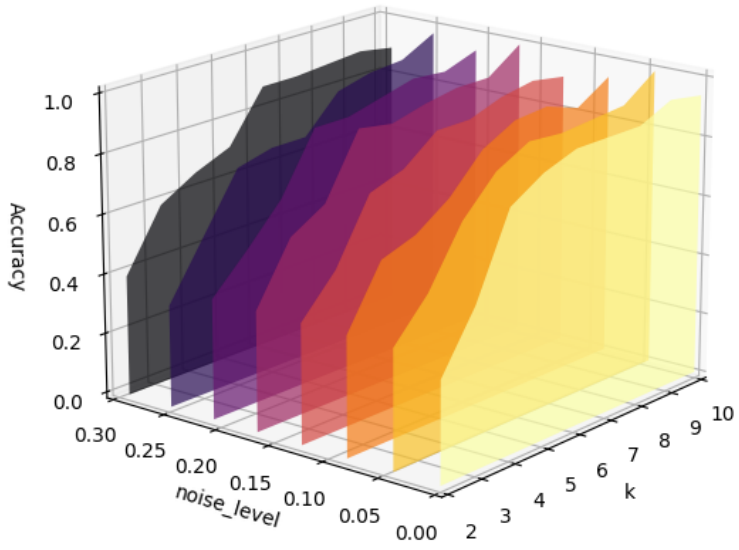


FIGURE 14 – Graphique 3D montrant l'accuracy en fonction de k et des différents niveaux de bruit

Les courbes montrent que l'augmentation du nombre de clusters jusqu'à $k = 6$ améliore généralement l'accuracy, indiquant que des clusters plus nombreux capturent mieux les nuances des données même en présence de bruit. Cependant, au-delà de $k = 6$, les bénéfices diminuent, ce qui peut indiquer une suradaptation ou une division excessive de données véritablement similaires.

Conclusion :

L'analyse suggère que choisir un k adapté peut considérablement améliorer la résistance du clustering face au bruit. La sélection d'un k entre 5 et 7 semble être optimale pour équilibrer précision et robustesse dans des conditions bruitées. Ces observations ouvrent la voie à des recherches futures sur l'ajustement des paramètres de clustering en présence de bruit et la conception de techniques plus résilientes.

Après avoir étudié en détail l'impact du niveau de bruit sur l'accuracy de l'algorithme K-Means et identifié les valeurs optimales de k pour la meilleure performance dans des conditions bruitées, il devient pertinent d'explorer des alternatives qui pourraient offrir une robustesse accrue. L'algorithme UKmeans, une variante de K-Means, promet d'aborder certains des défis identifiés avec K-Means, notamment la détermination du paramètre k optimal.

Dans la section suivante, nous allons explorer les caractéristiques de l'algorithme UKmeans, en examinant comment il modifie l'approche de clustering standard pour mieux capturer les structures sous-jacentes dans des ensembles de données complexes et bruitées. Cette analyse vise à comprendre si UKmeans pourrait servir de meilleure solution pour des applications spécifiques où la précision et la résilience au bruit sont primordiales.

4.2 U-Kmeans

Dans cette étude, nous appliquons l'algorithme UK-means à des données générées de manière synthétique, représentant cinq familles de formes différentes. Nous procédons par étapes :

1. Application de l'algorithme UK-means sur des données non bruitées.
2. Validation visuelle des clusters formés.
3. Utilisation de métriques internes et externes pour valider les résultats.
4. Étude de l'influence du bruit ajouté sur les résultats.
5. Analyse du nombre de clusters trouvés par UK-means par rapport au nombre réel de familles générées.

4.2.1 Résultats sur les Données Non Bruitées

Nous commençons par appliquer UK-means aux données non bruitées. Voici les résultats obtenus :

Validation Visuelle :

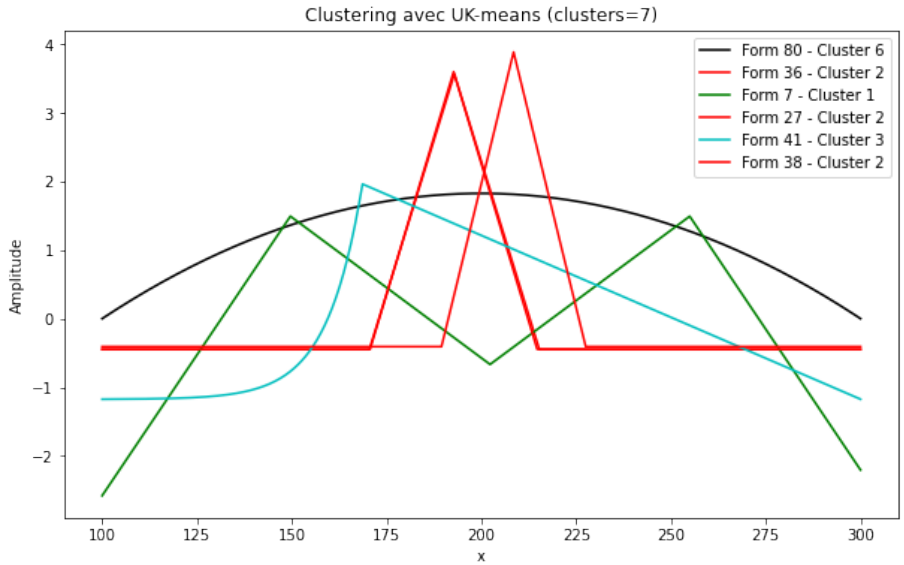


FIGURE 15 – Clusters formés par UK-means sur les données non bruitées.

La figure 15 montre que visuellement, les formes semblent bien séparées selon leurs familles. UK-means a identifié 7 clusters parmi les 5 familles de formes générées, indiquant une possible séparation fine des sous-groupes au sein des familles.

Métriques Internes :

— **Score de Silhouette** : 0.496
— **Index de Calinski-Harabasz** : 75.105

— **Index de Davies-Bouldin** : 0.749

Les métriques internes indiquent une bonne séparation et cohésion des clusters. Un score de silhouette de 0.496 montre que les clusters sont bien séparés et cohésifs. L'index de Calinski-Harabasz élevé (75.105) suggère une bonne densité de clusters. Un index de Davies-Bouldin bas (0.749) indique des clusters bien distincts les uns des autres.

Métriques Externes :

Les métriques externes, basées sur les classes réelles et prédites pour chaque classe, sont résumées dans le tableau ci-dessous :

Classe	**Précision**	**Rappel**	**F1 Score**
CLDR	0.905	0.950	0.927
CRDL	0.950	0.950	0.950
M	1.000	0.950	0.974
Parabola	1.000	1.000	1.000
Spike	1.000	1.000	1.000

TABLE 7 – Métriques externes pour les données non bruitées.

Voici les scores généraux :

— **Accuracy** : 0.970
— **Recall** : 0.970
— **Precision** : 0.971
— **F1 Score** : 0.970
— **Jaccard Score** : 0.944
— **Silhouette Score** : 0.524
— **Calinski-Harabasz Index** : 81.364
— **Davies-Bouldin Index** : 0.734

Les métriques externes indiquent une très bonne performance de UK-means, avec des scores de précision, rappel et F1 élevés pour la plupart des classes. Les scores généraux montrent une grande exactitude et une forte similarité entre les classes réelles et prédites.

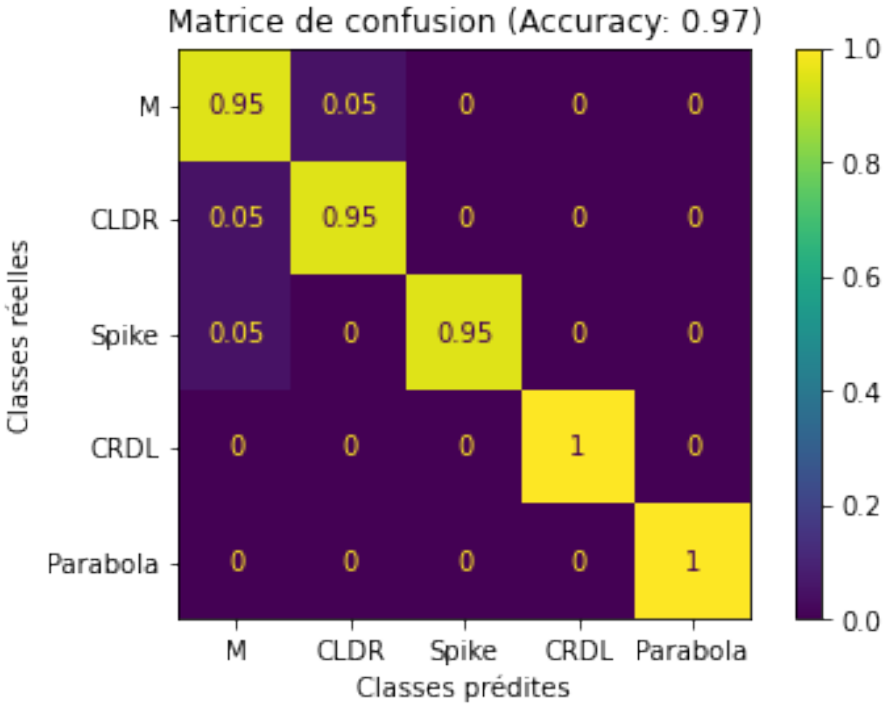


FIGURE 16 – Matrice de confusion pour les données non bruitées.

4.2.2 Influence du Bruit sur les Résultats

Ensuite, nous ajoutons progressivement du bruit aux données pour observer l'influence sur les résultats de UK-means.

On commence par une variance du bruit de 0.02 (variance réelle des données)

Validation Visuelle :

La figure 17 montre que malgré l'ajout de bruit (variance = 0.02), les formes restent visuellement relativement bien séparées par UK-means.

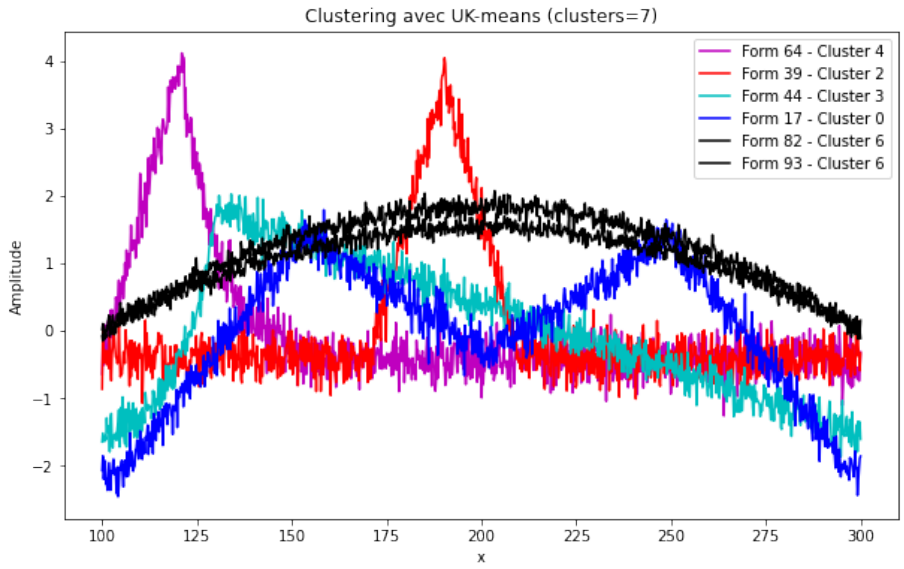


FIGURE 17 – Clusters formés par UK-means avec variance = 0.02.

Métriques Internes :

— **Score de Silhouette** : 0.432
— **Index de Calinski-Harabasz** : 54.648
— **Index de Davies-Bouldin** : 0.879

Les métriques internes indiquent une bonne séparation et cohésion des clusters, même avec du bruit ajouté.

Métriques Externes :

Les métriques externes, basées sur les classes réelles et prédites pour chaque classe, sont résumées dans le tableau ci-dessous :

Classe	Précision	Rappel	F1 Score
CLDR	0.613	0.950	0.745
CRDL	1.000	0.700	0.824
M	1.000	1.000	1.000
Parabola	1.000	1.000	1.000
Spike	0.933	0.700	0.800

TABLE 8 – Métriques externes pour les données avec variance = 0.02.

Voici les scores généraux :

— **Accuracy** : 0.870

— **Recall** : 0.870
— **Precision** : 0.909
— **F1 Score** : 0.874
— **Jaccard Score** : 0.792
— **Silhouette Score** : 0.432
— **Calinski-Harabasz Index** : 54.648
— **Davies-Bouldin Index** : 0.879

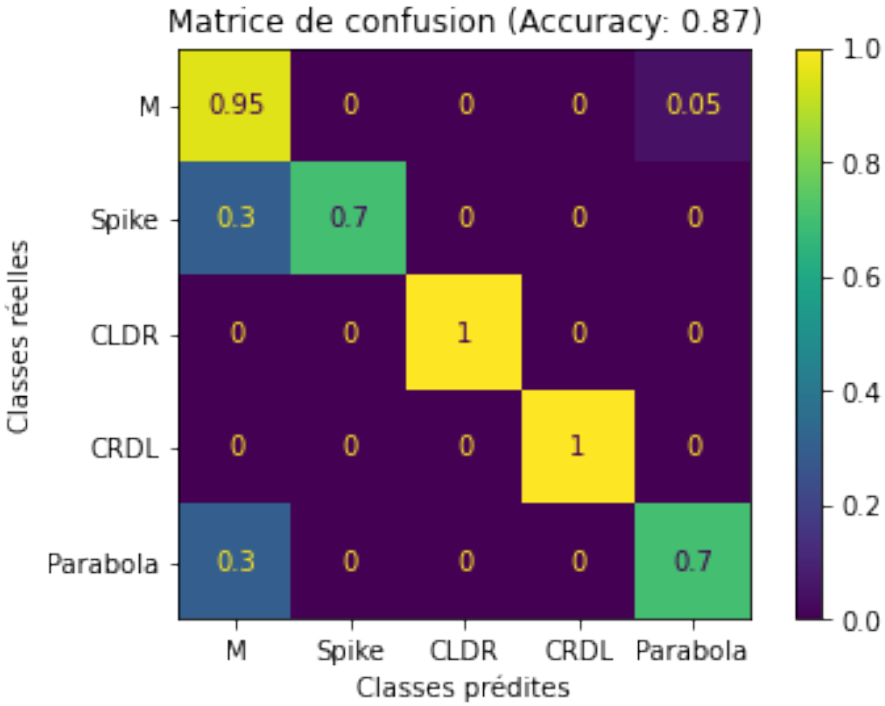


FIGURE 18 – Matrice de confusion pour les données avec variance = 0.02.

On augmente la variance du bruit à 0.05.

Métriques Internes :

— **Score de Silhouette** : 0.142
— **Index de Calinski-Harabasz** : 18.511
— **Index de Davies-Bouldin** : 1.733

Métriques Externes :

Les métriques externes, basées sur les classes réelles et prédites pour chaque classe, sont résumées dans le tableau ci-dessous :

Classe	Précision	Rappel	F1 Score
CLDR	0.650	0.650	0.650
CRDL	0.929	0.650	0.765
M	1.000	1.000	1.000
Parabola	1.000	1.000	1.000
Spike	0.769	1.000	0.870

TABLE 9 – Métriques externes pour les données avec variance = 0.05.

Voici les scores généraux :

— **Accuracy** : 0.860
— **Recall** : 0.860
— **Precision** : 0.870
— **F1 Score** : 0.857
— **Jaccard Score** : 0.774
— **Silhouette Score** : 0.142
— **Calinski-Harabasz Index** : 18.511
— **Davies-Bouldin Index** : 1.733

Avec une variance élevée de 0.3

Métriques Internes :

— **Score de Silhouette** : 0.174
— **Index de Calinski-Harabasz** : 14.043
— **Index de Davies-Bouldin** : 2.152

Métriques Externes :

Les métriques externes, basées sur les classes réelles et prédites pour chaque classe, sont résumées dans le tableau ci-dessous :

Classe	Précision	Rappel	F1 Score
CLDR	0.500	0.450	0.474
CRDL	0.583	0.700	0.636
M	1.000	0.950	0.974
Parabola	1.000	1.000	1.000
Spike	0.895	0.850	0.872

TABLE 10 – Métriques externes pour les données avec variance = 0.3.

Voici les scores généraux :

— **Accuracy** : 0.790
— **Recall** : 0.790
— **Precision** : 0.796
— **F1 Score** : 0.791
— **Jaccard Score** : 0.700
— **Silhouette Score** : 0.174
— **Calinski-Harabasz Index** : 14.043
— **Davies-Bouldin Index** : 2.152

L'ajout de bruit aux données affecte les performances de l'algorithme UK-means, mais il gère globalement bien différentes variances :

— **Variance nulle (0.0)** : Excellente séparation et cohésion des clusters avec des métriques très élevées.
— **Variance faible (0.02)** : Les scores restent élevés malgré une légère baisse, et les clusters restent bien définis.
— **Variance modérée (0.05)** : Les métriques montrent une certaine dégradation, mais les performances restent acceptables avec des clusters encore relativement bien définis.
— **Variance élevée (0.3)** : Malgré une dégradation notable des métriques, UK-means parvient à maintenir des scores raisonnables et une structure de cluster exploitable. Cependant, une dégradation de la séparation et de la cohésion des clusters est claire.

En résumé, UK-means démontre une robustesse notable face au bruit, maintenant des performances raisonnables même avec des niveaux de bruit élevés. Cependant, il perd notablement sa capacité à bien séparer les clusters dans l'espace, comme le montrent les métriques internes.

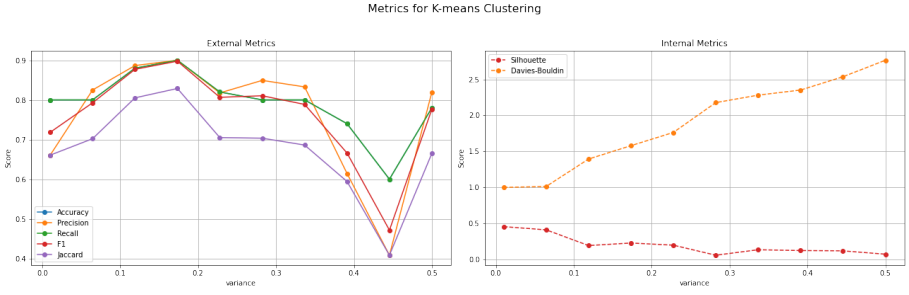


FIGURE 19 – Influence de la variance du bruit sur les métriques de UK-means.

La figure 19 montre clairement l'influence de la variance du bruit sur les performances de l'algorithme UK-means. On observe une tendance à la diminution des scores de précision, rappel et F1 avec l'augmentation de la variance, bien que UK-means maintienne des performances raisonnables. Les métriques internes comme le

score de silhouette et l'index de Davies-Bouldin illustrent la dégradation progressive de la séparation et de la cohésion des clusters. Cette analyse souligne la robustesse de UK-means face au bruit, tout en montrant ses limites dans des environnements fortement bruités.

4.2.3 Étude du Nombre de Clusters Trouvés par UK-means

L'objectif est d'analyser comment l'algorithme UK-means réagit lorsque certaines classes de formes sont enlevées du dataset, en comparant le nombre de clusters trouvés par l'algorithme avec le nombre réel de classes générées. Nous allons également étudier l'impact du bruit sur ces résultats.

Démarche :

- **Données sans bruit** :
  - Nous commençons par générer les données sans bruit, puis enlevons successivement une, deux, puis trois classes de formes.
  - Nous appliquons UK-means pour chaque scénario et observons le nombre de clusters trouvés par rapport au nombre réel de classes restantes.
- **Données avec bruit (variance = 0.02)** :
  - Nous répétons la même procédure avec des données contenant un bruit de variance 0.02, représentant un scénario réaliste.
- **Données avec bruit (variance = 0.2)** :
  - Enfin, nous appliquons la même procédure avec un bruit plus élevé (variance = 0.2) pour évaluer l'impact d'un bruit significatif sur le nombre de clusters trouvés.

Résultats :

Scénario	Nb Clusters Trouvés	Nb de Clusters
Sans bruit, sans 1 classe	7	4
Sans bruit, sans 2 classes	3	3
Sans bruit, sans 3 classes	4	2
Bruit 0.02, sans 1 classe	7	4
Bruit 0.02, sans 2 classes	3	3
Bruit 0.02, sans 3 classes	3	2
Bruit 0.2, sans 1 classe	4	4
Bruit 0.2, sans 2 classes	3	3
Bruit 0.2, sans 3 classes	3	2

TABLE 11 – Comparaison du nombre de clusters trouvés par UK-means et le nombre réel de classes.

Conclusion :

L'algorithme UK-means montre une certaine robustesse dans l'identification du nombre de clusters, même en présence de bruit. Quelques observations importantes peuvent être faites :

- **Sans bruit** :
  - Lorsque des classes sont enlevées, UK-means tend à sur-segmenter (ex. trouver 7 clusters au lieu de 4) lorsque la réduction du nombre de classes n'est pas très significative.
  - En enlevant des classes supplémentaires, l'algorithme s'ajuste mieux au nombre réel de clusters.
- **Avec bruit (variance 0.02)** :
  - Les résultats montrent une tendance similaire à celle sans bruit, avec une légère augmentation du nombre de clusters trouvés.
  - Malgré le bruit, UK-means parvient à identifier un nombre raisonnable de clusters.
- **Avec bruit (variance 0.2)** :
  - Le nombre de clusters tend à diminuer quand la variance augmente, ce qui fait que l'algorithme a plus de difficulté à détecter des sous-structures fines dans les données.
  - Cependant, le nombre de clusters trouvés reste raisonnablement proche du nombre réel.

5 Discussions

Cependant, ces méthodes sophistiquées nécessitent une gestion rigoureuse de la qualité des données et une interprétation prudente des résultats pour éviter les erreurs dans la classification des données environnementales. L'intégration future de l'apprentissage profond avec les techniques de clustering traditionnelles pourrait offrir une solution plus robuste pour traiter les défis de la surveillance environnementale, en améliorant la précision des prédictions et en réduisant le besoin de supervision humaine dans le processus de classification des données.

6 Conclusion

Dans cette étude, nous avons exploré l'application de techniques de classification pour l'analyse de motifs d'intérêt en surveillance environnementale, en particulier à travers les algorithmes K-means et UK-means. Nous avons commencé par une analyse des données, identifiant différentes familles de formes d'intérêt et générant des variantes paramétrées de ces formes. Cette approche a permis de tester et de valider les algorithmes de classification sur des données synthétiques avant de les appliquer aux données réelles.

Nous avons démontré que l'algorithme K-means, bien qu'efficace dans des scénarios sans bruit, présente des limitations en raison de la nécessité de définir le nombre de clusters à l'avance et de sa sensibilité aux valeurs initiales des centres de clusters. Pour surmonter ces défis, nous avons utilisé l'algorithme UK-means, qui offre une meilleure performance théorique et une robustesse accrue face au bruit, tout en permettant de déterminer automatiquement le nombre optimal de clusters.

Les résultats obtenus montrent que UK-means est capable de maintenir une performance élevée même en présence de bruit, bien que des niveaux de bruit très élevés affectent inévitablement la qualité de la classification. Nos tests ont révélé que, pour des niveaux de bruit modérés, l'algorithme conserve une précision et une cohésion des clusters acceptables.

En conclusion, l'intégration de méthodes de classification avancées, telles que UK-means, dans les systèmes de surveillance environnementale peut significativement améliorer la précision et l'efficacité de la détection de motifs d'intérêt. Cependant, pour maximiser les avantages de ces techniques, il est essentiel de continuer à affiner les modèles mathématiques utilisés pour générer des données de test et de développer des algorithmes encore plus robustes face aux variabilités des données réelles. À l'avenir, l'incorporation de techniques d'apprentissage profond pourrait offrir des solutions encore plus puissantes pour surmonter les défis de la surveillance environnementale, en améliorant la précision des prédictions et en réduisant le besoin de supervision humaine.

L'étude d'Hundman et de ses collaborateurs ([3]) illustre l'utilisation de réseaux LSTM pour détecter des anomalies dans les systèmes complexes, une méthode qui peut être adaptée pour surveiller et analyser les anomalies environnementales. Cette convergence entre les méthodes traditionnelles de clustering et les techniques d'ap-

prentissage en profondeur ouvre de nouvelles avenues pour la surveillance environnementale, permettant une réaction rapide et précise aux événements écologiques critiques.

Bibliographie

[1] Kristina P Sinaga et Miin-Shen Yang. "Unsupervised K-means clustering algorithm". In : *IEEE access* 8 (2020), p. 80716-80727.

[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi et Teh Ying Wah. "Time-series clustering–a decade review". In : *Information systems* 53 (2015), p. 16-38.

[3] Kyle Hundman et al. "Detecting spacecraft anomalies using lstms and non-parametric dynamic thresholding". In : *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, p. 387-395.