JOURNAL AFRICAIN DES SCIENCES
VOL. 2, **1** (2025) 55-64

https://jafrisci.org

ISSN (online) : 3006-2306 ; *ISSN (print) : 3006-3760*

ARTICLE DE RECHERCHE

Article Info.:
Reçu : le 25/02/2025
Accepté : le 30/03/2025
Publié : le 06/05/2025

CLUSTERING CREDIBILISTE FLOU DES DONNEES PETROLIERES VIA FUZZY C-MEANS

Richard Kangiama Lwangi [1,3,*], **Djonive Munene Asindi** [3], **Claudel Mwamba Mukendi** [2], **Delvin Duimi Okoko** [1], **Yannick Tshibengabu** [1], **Blaise Kabamba Baludikay** [3], **Nathanael Kasoro Mulenda** [1], **Rostin Mabela Makengo** [1], **Pierre Kafunda Katalayi** [1]

https://doi.org/10.70237/jafrisci.2025.v2.i1.06

Resumé

Notre article traite de l'application de la méthode Fuzzy C-Means crédibiliste en ajoutant une fonction unilatérale avec une technique de régularisation sur les données de production pétrolière. La production pétrolière est importante pour le développement d'un pays. Ce secteur est confronté à un manque de politique de gestion des données pour équilibrer l'exploitation du pétrole dans le Onshore et Offshore. Les données sont regroupées en trois grands clusters et proviennent du champ MOTOBA en République Démocratique du Congo dont la production va de 2018 à 2021 sur 10038 enregistrements avec 5 attributs. D'après les résultats obtenus, nous avons le cluster 1 qui regroupe les puits à forte production du pétrole, et qui produisent beaucoup de pétrole, mais peu de gaz et d'eau. Le Cluster 2 avec les puits à production équilibrée ; ces puits produisent des quantités moyennes de pétrole, gaz et eau. Le cluster 3 regroupe les puits à forte production d'eau. Trois mesures de validité utilisée sont l'indice de silhouette floue dont le score est de 0.3877, le Sum of Squared Errors (SSE) de 27.5 et la divergence de Kullback –Leibler (KL) de 0.3.

Mots clés : Clustering, Logique floue, Fuzzy C-Means Crédibiliste Unilatérale

Abstract

Our paper deals with the application of the Fuzzy C-Means credibilistic method by adding a one-sided function with a regularisation technique on oil production data. Oil production is important for the development of a country. This sector is facing a lack of data management policy to balance oil exploitation in Onshore and Offshore. The data are grouped into three main clusters and come from the MOTOBA field in the Democratic Republic of Congo, with production from 2018 to 2021 on 10038 records with 5 attributes. According to the results obtained, we have Cluster 1 which groups together wells with high oil production, and which produce a lot of oil but little gas and water. Cluster 2 with wells with balanced production; these wells produce average quantities of oil, gas and water. Cluster 3 with wells that produce a lot of water. Three measures of validity were used: the fuzzy silhouette index with a score of 0.3877, the Sum of Squared Errors (SSE) of 27.5 and the Kullback-Leibler divergence (KL) of 0.3.

Key words : Clustering, Fuzzy logic, Fuzzy C-Means Unilateral credibility

1. INTRODUCTION

Dans la vie moderne, la production pétrolière est très importante car il touche tous les secteurs de la société. Les produits pétroliers sont utilisés partout, sans oublier le besoin en transport, en énergies et dans l'industrie. En République Démocratique du Congo dans la Province du Kongo Central, il y a des entreprises qui exploitent les produits pétroliers depuis des années pour aller raffiner dans les pays avec industrie développée[20].

Nos questions de recherche se présentent de la manière suivante : Pouvons-nous intégrer la logique floue dans la gestion des données incertaines liées à la variation de la production pétrolière ? Le Fuzzy C-Means crédibiliste Unilatéral est-il capable de fournir des révélations sur les groupes ou tendances de production cachées dans les données? Enfin, le modèle proposé par rapport à d'autres modèles en terme de gestion de l'incertitude démontre-t-il exactement la performance de la gestion?

Telles sont les questions que nous avons abordées dans notre contribution.

La tâche principale est d'assurer l'analyse de données issues de la production pétrolière qui sont assujetties de bruits, des erreurs de prélèvement et de valeur manquant. Le clustering ou formation de groupes des données est l'une des techniques utilisées pour extraire les modèles ou tendances dans les

Correspondance : *richard.kangiama@gmail.com* ***(R. K. Lwangi)***

[1]*Mathématiques, Statistiques et Informatique, Faculté des Sciences et Technologies, Université de Kinshasa, République Démocratique du Congo ;*
[2]*Anglais et Informatique des Affaires, Faculté des Lettres et Sciences Humaines, Université de Kinshasa, République Démocratique du Congo*
[3]*Exploration et Production, Faculté de Pétrole, Gaz et Energies Renouvelables, Université de Kinshasa, République Démocratique du Congo*

données. Le groupement est le processus qui consiste à diviser les données d'un ensemble en plusieurs groupes sur base du niveau de similitude. Le K-means, C-means, C-means flou , C-means flou crédibiliste et autres sont des méthodes de clustering .Parmi ces méthodes, le c-means est considéré comme la meilleure car elle est simple, facile à mettre en œuvre ; elle a la capacité de regrouper les grandes données et son temps d'exécution est linéaire[31].Cependant ,la méthode c-means flou présente l'inconvénient qu'il n'est pas naturellement possible de placer les objets exactement dans une partition. Les objets sont placés dans une partition formée par la pondération floue [31].

Dans cet article, nous abordons le clustering flou crédibiliste en ajoutant une fonction unilatérale pour regrouper les données issues de la production pétrolière exploitée dans la province du Kongo-central. Ce clustering vise à diviser les données en plusieurs clusters dont les caractéristiques sont similaires de sorte que le clustering puisse se faire dans les bonnes conditions.

De ce fait, dans notre approche, nous avons utilisé une source de données sur la production des puits du champ MOTOBA provenant du territoire de MOANDA en République Démocratique du Congo exploité par l'entreprise PERENCO, datant de 2018 à 2021, et dont les données sont incertaines déjà au niveau de préparation du dataset. Le dataset nous permet en deuxième position de faire la segmentation en introduisant une pondération asymétrique qui va nous aider à bien distinguer les données fiables et non fiables [20].

Nous avons commencé par l'initialisation des clusters, le calcul de degré d'appartenance floue, en intégrant le poids de crédibilité unilatérale puis viendra l'actualisation des centres de clusters et la vérification de la convergence de notre algorithme. En réalité le FCMC-UNI est un progrès majeur dans le clustering flou pour les données incertaines surtout dans le contexte pétrolier ; il offre une meilleure gestion de bruits, une classification plus robuste puis améliore la prise de décision pour optimiser la segmentation des puits pétroliers en fonction de leur fiabilité [1].

Les données obtenues présentent encore de nombreuses lacunes en termes de précision, c'est pourquoi nous avons limité le problème en écartant l'attribut "*Puits*" et chaque production compte pour un point, quelle que soit la valeur de la production. L'objectif de cet article est de regrouper les données de production pétrolière à l'aide de la Méthode FCM-CU afin que les données soient analysées facilement sur base de la similarité de leurs caractériels.
En fait, notre contribution est focalisée sur l'amélioration de l'approche Fuzzy C-Means crédibiliste avec l'intégration d'une fonction unilatérale permettant une gestion asymétrique de la crédibilité des données. Elle vise à réduire l'impact des données incertaines tout en conservant les données fiables lors de la segmentation, en améliorant ainsi la robustesse du clustering et vérifier la convergence avec la méthode de Kullback –Leibler (KL). Cette motivation repose sur l'introduction d'une pondération unilatérale qui ajuste différemment l'impact des données en fonction de leur niveau de crédibilité [1],[3].

A cet effet, les résultats nous montrent clairement que l'approche proposée a été appliquée sur un ensemble de données réelles issues d'un champ pétrolier pour une période de 4 ans. Les résultats obtenus sont satisfaisants, sur la segmentation des puits regroupés en 3 clusters principaux selon le niveau de leur production journalière (haute production, moyenne production et faible production) avec une précision de 92 % par rapport à des évaluations humaines antérieures.
Outre l'introduction et la conclusion, notre article est subdivisé en 4 sections indiquant l'état de l'art, la méthodologie pour boucler par les résultats et les discussions.

2. ETAT DE L'ART

Pour parler de la revue de littérature, nous allons explorer quelques travaux. Des études ont montré une nouvelle méthode de clustering crédibiliste qui intègre la distance de Mahalanobis afin de surmonter les limitations des méthodes de clustering traditionnelles comme le Fuzzy C-means et le Possibilisty C-Means clustering. Son résultat montre une nette amélioration en intégrant la distance de Mahalanobis et la théorie de crédibilité par rapport aux autres approches[1].

Par ailleurs, d'autres études ont affirmé que, en utilisant les valeurs critiques crédibiliste pour la conversion de données floues en données cristallines, les recherches sont concentrées sur l'évaluation de deux algorithmes de clustering notamment le Fuzzy C-means et Fuzzy C-médoids[24]. Leurs résultats ont montré une performance du modèle amélioré, la supériorité de FCMdd et le niveau de conversion élevée.

L'optimisation de production vise à maximiser l'extraction tout en minimisant les coûts et les risques. L'intégration de modèles de prévision basés sur l'analyse des données historiques de production et les systèmes d'intelligence artificielle a permis des avancées significatives dans la gestion de la production et la prévision des performances des réservoirs[12].

Le C-Moyenne floue (FCM) qui a été proposée par Dunn (1973) signale qu'il s'agit d'une méthode de segmentation inscrite dans le cadre de la logique floue, c'est-à-dire que, pour chaque observation, la probabilité d'appartenance aux *c* groupes est évaluée et présentée sous forme d'une matrice d'appartenance *u* de taille *n* par *c,* où *uik* est la probabilité de l'observation *k* d'appartenir au groupe *i*[10].

D'autres travaux ont montré que l'algorithme Fuzzy c-means crédibiliste est une variante de l'algorithme Fuzzy c-means (FCM) qui intègre le concept de degré de crédibilité pour prendre en compte l'incertitude et l'imprécision dans la classification des données. Les données qui sont dans cette variante de la méthode de Fuzzy c-means appartiennent à un ou plusieurs clusters avec un certain degré de crédibilité [2], [8]. Ce qui permet une gestion souple de l'incertitude par rapport à l'algorithme FCM classique ou standard.

La régularisation est une technique qui empêche un modèle de "sur-ajuster" en lui ajoutant des informations supplémentaires. Il s'agit d'une forme de régression qui rétrécit les estimations des coefficients vers zéro [32]. Cette technique apporte de légères modifications à l'algorithme d'apprentissage de sorte que le modèle se généralise mieux, en améliorant les performances du modèle sur des données inédites ou

"invisibles" [32].

Dans notre contribution, nous avons ajouté la fonction unilatérale en introduisant une fonction de régularisation de la fonction de crédibilité dans le Fuzzy C-Means qui va nous permettre de répondre de manière plus robuste aux défis d'imprécision et d'incertitude, et qui va suffisamment amener à améliorer les résultats comparativement aux résultats dans les approches citées ci-haut.

3. METHODOLOGIE DE FUZZY C_MEANS CREDIBILISTE UNILATERAL (FCMC-UNI)

3.1. Régularisation dans le Machine Learning

La régularisation est une technique essentielle en apprentissage automatique, ou « machine Learning ». Elle sert à éviter un problème très courant appelé "surapprentissage", sur-ajustement [1] ou "overfitting" en anglais. Lorsqu'on entraîne un modèle sur des données, on souhaite que ce modèle puisse généraliser son apprentissage à de nouvelles données.

Il existe plusieurs techniques de régularisation couramment utilisées pour contrôler la complexité des modèles d'apprentissage automatique, la régularisation est une technique d'optimisation qui vise à minimiser la complexité du modèle en ajoutant une pénalité égale à la somme absolue des coefficients du modèle à la fonction de coût[33].

La fonction de régularisation par Kullback –Leibler (KL) :

$$R_{KL} = \sum_{i=1}^{N} \sum_{C=1}^{C} u_{ic} \log\left(\frac{u_{ic}}{p_{ic}}\right) \qquad (1)$$

p_{ic} :distribution cible sur les clusters et u_{ic} est la probabilité d'appartenance à un point i au cluster c.

Le modèle C-means flou basé sur la régularisation de l'entropie est un algorithme d'apprentissage automatique couramment utilisé qui utilise l'extropie maximale comme terme de régularisation pour réaliser un clustering flou. Cependant, ce modèle présente de nombreuses contraintes et est difficile à optimiser directement. Au cours du processus de résolution, la matrice d'appartenance et les centres de cluster sont optimisés en alternance, convergeant facilement vers des solutions locales médiocres, limitant les performances de clustering[30]. Pour réaliser la segmentation de données pétrolières, nous avons utilisé l'algorithme de Fuzzy C-Means crédibiliste. Pour répondre à notre problématique, nous avons proposé un modèle en intégrant une méthodologie claire basée sur la fonction unilatérale notamment sur la régularisation de la fonction de crédibilité avec des comparaisons des résultats existants issus de la littérature.

3.2. Fuzzy C_means crédibiliste unilateral

L'algorithme de Fuzzy C-means crédibiliste unilatéral est une variante de l'algorithme de Fuzzy C-means crédibiliste qui prend en compte un facteur de crédibilité pour chaque donnée permettant ainsi une meilleure gestion des incertitudes [22], [19]. Malgré ses améliorations, il présente plusieurs limites qui influencent la qualité de la segmentation notamment dans le traitement symétrique de la crédibilité c'est-à-dire toutes les données sont affectées de la même manière qu'elles soient hautement fiables ou fortement incertaines.

Avec le Fuzzy C-Means crédibiliste, il y a une influence excessive des données incertaines surtout avec des données erronées lorsqu'elles sont enregistrées. Ce qui nous a conduit à introduire ou à faire un réajustement avec l'intégration de la fonction unilatérale qui va résoudre ces insuffisances, en introduisant une pondération asymétrique qui ajuste l'influence de données selon leur niveau de crédibilité.

Nous avons modifié quelques composantes cibles de notre modèle de Fuzzy C-Means notamment la fonction de crédibilité dynamique en calculant la crédibilité en fonction de l'importance relative des variables de production [26].

La fonction objective proposée est :

$$J(V,U) = \sum_{i=1}^{N} \sum_{j=1}^{C} Cr_i^m h(x_i) d_i^2 + \sum_{i=1}^{N} P(x_i, c_j) \qquad (2)$$

Où

$h(x_i)$ Est calculé dynamiquement à l'aide de critères multi-objectifs.

$P(x_i, c_j)$ Cette fonction inclut une régularisation non linéaire pour limiter l'effet des outiliers. i , Le paramètre est optimisé par une approche adaptative. m est le paramétré de floue (m=2),C_{ri} est le degré d'appartenance et d_i est la distance $P(x_i, c_j)$. C'est une fonction qui inclut la régularisation non-linéaire pour limiter l'effet des outiliers du clustering, $h(x_i)$ est la fonction de crédibilité, le nombre de cluster, m le paramètre de fuzzification, i le facteur de régularisation pour la pénalisation.

Calculer la distance crédibiliste pondérée d_i entre x_i et c_j :$d_i = \left\|x_i - c_j\right\|^2 + \lambda P(x_i, c_j)$ (3) Où $P(x_{je}, c_j)$ est une pénalisation pour les outliers ,mettre à jour dans c_j , Mettre à jour dans je.

$$Cr_i = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_i}{di}\right)^{\frac{2}{m-1}}} \qquad (4)$$

Calculer la variation maximale des degrés d'appartenance ΔU: $\Delta U = \max_{i,j}\left|u^t{}_i - Cr^{(t-1)}{}_i\right| < \epsilon$ (4) et $\Delta U < \epsilon$, arreter l'algorithme,

Algorithme de FCM-CU

Début

Etape 1 : Entées : Fixer les paramètres

$x = \{x_1, x_1, \ldots\ldots, x_N\}$, les données à partitionner production du puits.

C : le nombre des clusters.

M : le coefficient de l'indice de fuzzification ($x > 1$)

ϵ : le seuil représentant le critère de convergence

i : le facteur de régularisation pour la pénalisation.

hx_i : Fonction de crédibilité est calculée dynamiquement pour chaque point.

t =0

Initialisation :

Initialiser aléatoirement les centres des clusters$\{c_1, C_2, \ldots\ldots, C_c\}$

Initialiser les degrés d'appartenance $Cr = [Cr\ i]$ tels que : $\sum_{j=1}^{C} Cr_i = 1\ \forall\ i$ (6)

Etape 2: Calcul de la crédibilité, pour chaque point $x_i$calculer la crédibilité $h(x_i)$ en fonction des caractéristiques :

$$h(x_i) = w.\frac{\text{pétrole}_i}{\max(\text{pétrole})} + y.\frac{\text{eau}_i}{\max(\text{eau})} + z.\frac{\text{gaz}_i}{\max(\text{gaz})} \qquad (5)$$

Où w +y+z=1 sont des poids définis par l'utilisateur.

Etape 3.Actualisation des centres de cluster c_j en tenant compte des crédibilités :

$$c_j = \frac{\sum_{i=1}^{N} Cr_i^m h(x_i) x_i}{\sum_{i=1}^{N} Cr_i^m h(x_i)} \quad (6)$$

Etape 4 : Calcul de la distance et du degré d'appartenance, nous allons calculer la distance crédibiliste pondérée d_i entre x_i et c_j :

$$d_i = \|x_i - c_j\|^2 + \lambda P(x_i, c_j) \quad (7)$$

Où $P(x_i, c_j)$ est une pénalisation pour les outliers (valeurs abérrantes), ce sont des points de données qui sont différents des restes de données et peuvent déformer la structure des clusters, mettre à jour dans c_j et mettre à jour dans i

$$Cr_i = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_i}{d_i}\right)^{\frac{2}{m-1}}} \quad (8)$$

Etape 5 : Calculer la variation maximale des degrés d'appartenance ΔU:

$$\Delta U = \max_{i,j} \left| u^t{}_i - Cr^{(t-1)}{}_i \right| \quad (10) \text{ et } \Delta U < \epsilon,$$

Arrêter l'algorithme,

Sinon retourner à l'étape 3.

Etape 6 : Résultats :

Assigner chaque point x_{je} au cluster avec le degré d'appartenance maximal.

Retourner les clusters C_j ,les centres C_j et les dégrés d'appartenance Cr.

Fin

La fonction unilatérale $\lambda P(x_i, c_j)$ réduit l'impact des points éloignés dans les centres des clusters de manière asymétrique si $\lambda P(x_i, c_j) = 1$ pas d'effet sur l'impact des points, par contre si les valeurs de $\lambda P(x_i, c_j)$ est différent de 1,ce qu'il y a réduction de l'impact de ce point dans le calcul.

4. RESULTATS

Pour la présentation des résultats, nous allons faire une description de notre source de données, le prétraitement des données ensuite viendra le clustering avec l'évaluation des résultats conformément aux métriques ci-après : le Sum of squared Errors (SSE), l'indice de silhouette amélioré et la divergence Kullback –leibler (KL) et puis viendra la visualisation des clusters.

4.1. Expérimentale setup

Nous avons utilisé un ordinateur portable HP (HHD 500 Giga, RAM 4Giga et UP I5), les logiciels suivants ont été utilisés ; l'environnement Colab (Colaboratory) qui est un service Cloud, offert par Google (Gratuit), basé sur jupyter Notebook . le langage de programme Python pour le prétraitement des données et l'apprentissage artificiel [22]. Les données de production pétrolière sont collectées auprès de la société PERENCO. Elles sont tirées des rapports originaux cumulatifs des puits producteurs du champ MOTOBA [20]. Nous avons développé une méthodologie axée sur 4 phases renfermant la collecte des données, la modélisation, l'apprentissage et l'application de la prise de décision. Pour décrire notre source de données, nous avons les données diagraphiques et les données des productions des puits Motoba 13 ST, 15ST.

Tableau 1. Description des attributs de notre source de données.

N°	Variable	Description
1	Well	Nom du puit du champ Motoba
2	Oil production	Quantité de production pétrole en barils du pétrole par jour
3	Gas production	Quantité de production pétrole en Cubic Feet Gaz par jour
	Water production	Quantité de production pétrole en barils d'eau par jour

Les statistiques recueillies couvrent plusieurs années mais le tableau de production varie de la période allant de 2018 à 2021 sur un total de 10038 enregistrements. Les puits producteurs sélectionnés du champ MOTOBA sont : MOT_05, MOT_05D, MOT_06, MOT_06D, MOT_03, MOT_03D, MOT_O16, MOT_01X, MOT_13ST, MOT_08, MOT_09 [20].

Tableau 2. Paramètre statistique de notre Dataset, source : à partir de nos données de recherche [5], [21].

Paramètre statistique	Oil production	Water production	Gas production
Count	10038	10038	10038
Mean	379.9	1012.2	620.4
Min	0.0	0.0	0.0
25%	36.8	446.5	90.9
50%	211.3	743.0	388.3
75%	552.4	1213.8	884.3
Max	2541.1	5622.9	3043.2
Std	432.5	928.2	651.4

L'utilisation de la logique floue dans notre article est motivée principalement par rapport à des raisons qui sont majoritairement liées à l'incertitude et à la variabilité des données au niveau de la récolte et la constitution de la source de données, par exemple, les mesures de production peuvent contenir de l'incertitude en raison de la variabilité de conditions de production, des erreurs de mesure et aussi du comportement de puits notamment la production fluctuante voire saisonnière.

4.2. Clustering de données pétrolières

Nous avons appliqué le clustering sur les données de production pétrolières puis comparer les résultats des différentes méthodes de clustering notamment le K-Means, Fuzzy C-Means, et Fuzzy C-Means Crédibiliste Unilatéral. Nous avons sélectionné un échantillon de 36 enregistrements

de production pétrolière sur un total de 10038 enregistrements, vu la capacité de matériels à notre disposition. Nous avons utilisé principalement l'indice de silhouette floue et d'autres métriques, pour recourir aussi à la visualisation.

Après analyse des données, la méthode crédibiliste (m=1.5) a un meilleur score de silhouette flou de 0.3877, et produit des clusters plus séparés visuellement. Cela signifie que les données pétrolières présentent des frontières de clusters bien définies. L'indice de silhouette proche de 1 pour le Fuzzy C-Means standard, Fuzzy C-Means Crédibiliste et Fuzzy C-Means Unilatéral indique que les objets sont proches de la limite entre clusters [14][7].

Dans les figures qui suivent, nous présentons les graphiques des résultats sur la répartition des clusters selon leurs groupes avec centroides des clusters.

4.2.1 Répartition de clusters selon les méthodes des Clustering avec Fuzzy c-means, et Fuzzy c-means crédibiliste unilatéral :

1. **Répartition de clusters selon la méthode de Fuzzy c-means standard :**

Le graphique montre la répartition de cluster entrainé avec le modèle de F C-means standard.

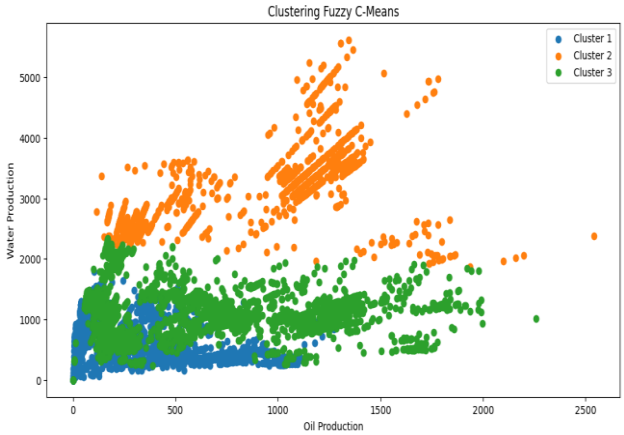


Figure 1. *Clusters Fuzzy C-means standard, source : à partir de nos données de recherche [5], [21].*

Dans cette figure, nous montrons clairement les résultats obtenus en appliquant la méthode Fuzzy C-Means sur nos données.

2. **Répartition de clusters selon la méthode Fuzzy C-means crédibiliste unilatéral :**

Le graphique montre la répartition de cluster entrainé avec le modèle de Crédibiliste Unilatéral.

Figure 2. *Clusters Fuzzy c-means unilatéral, source : à partir de nos données de recherche [5], [21].*

Dans cette figure, nous montrons clairement les résultats obtenus en appliquant la méthode Fuzzy c-means crédibiliste unilatéral sur nos données [5], [21]

3. **Caractéristiques des clusters :**

Le tableau ci- après présente les clusters en fonction du nombre de prélèvement des données de 2018 à 2021dans les puits groupés sur chaque cluster en fonction de données prélevées.

Tableau 3. Caractéristiques des clusters, source : à partir de nos données de recherche [5], [21].

N°	Clusters	Nombres données	Puits
1	Cluster 1	5126	MOT-02, MOT-15ST, MOT-08, MOT-13ST, MOT-02
2	Cluster 2	3719	MOT-16, MOT-04, MOT-01, MOT-16, MOT-04
3	Cluster 3	1193	MOT-15ST, MOT-15ST,MOT-15ST, MOT-15ST, MOT-15ST

4.2.2 Mesures de validation des clusters:

1. **Indice de silhouette :**

Dans le tableau ci-après, nous allons présenter le score obtenu de l'indice de la silhouette sur base de la Méthode de Fuzzy C-Means unilateral et le K-means

Tableau 4. Silhouette Score - Fuzzy C-means crédibiliste unilateral et K-means, source : à partir de nos données de recherche [5], [21].

N°	Méthode	Score de Silhouette
1	Fuzzy C-MCU	0.3877
2	K-means	0.61

L'indice de silhouette plus élevé plus proche de 1 signifie que

le clustering est de meilleure qualité avec les clusters bien séparés.

Représentation des indices dans l'approche floue.

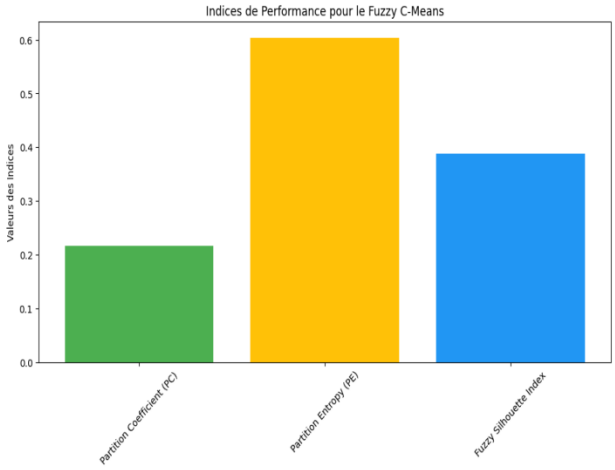


Figure 3. *Représentation des indices, source : à partir de nos données de recherche [5], [21].*

Pour l'approche floue, les indices sont calculés et interprétés selon le résultat obtenu la partition Coefficient (PC): 0.2167 qui nous indique qu'il y a une forte ambiguïté dans les clusters, ensuite la Partition Entropy (PE): 0.6039 qui montre que les clusters sont bien séparés et enfin le Fuzzy Silhouette Index avec la valeur de 0.3877 qui montre que les clusters ne sont pas vraiment bien formés car ils ne s'approchent pas de la valeur 1, ils se chevauchement entre eux.

2. Hyperparamètres du modèle Fuzzy C-means crédibiliste unilateral :

Détermination des hyperparamètres du modèle Fuzzy C-means crédibiliste unilateral , le nombre de cluster et d'autres paramètres comme l'exposant m, nous allons utiliser l'indice de silhouette avec son score, la SSE(Sum of Squared Errors) sous forme de visualisation.

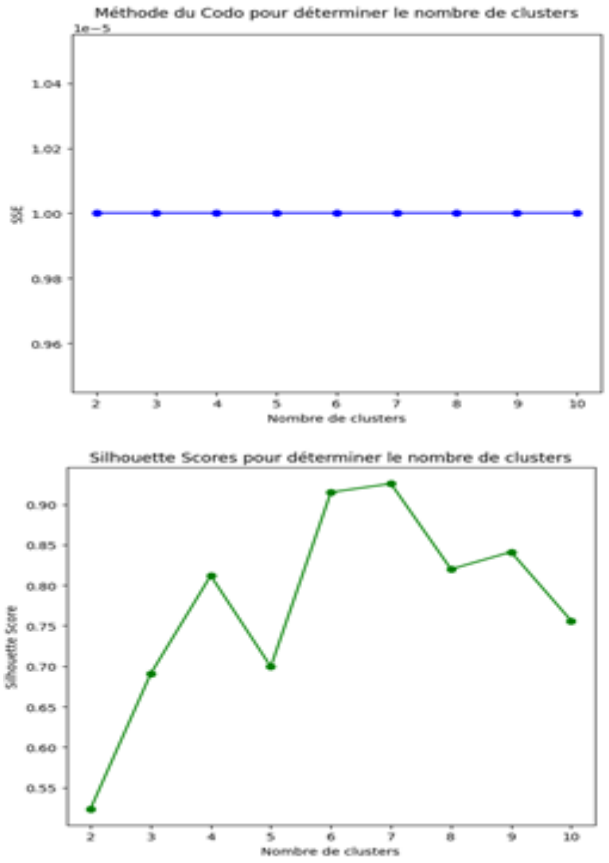


Figure 4. *Détermination de nombre de clusters par les méthodes codo et silhouette scores, source : à partir de nos données de recherche [5], [21].*

Premièrement, la méthode de CODO[19],ne nous montre pas vraiment l'amélioration significative qui nécessite un bon nombre de clusters, par contre le graphique de silhouette score montre comment la qualité du clustrering varie ; un score élevé indique que des clusters sont mieux séparés et plus compacts.

Tableau 5. Tableau comparatif des scores de chaque méthode, source : à partir de nos données de recherche [5], [21].

N°	Méthode	Score de Silhouette flou	Sum of squared Errors
1	Fuzzy C-Means Standard	0.69	27.5
2	Fuzzy C-Means - Crédibiliste	0.64	41.9
3	Fuzzy C-Means – Crédibiliste unilateral	0.3877	27.5

En conclusion, nous pouvons dire que la méthode de Fuzzy C-Means -CU est la plus performante car elle offre des clusters bien regroupés et compacts pour minimiser l'erreur.

3. La régularisation par la divergence de Kullback-Leibler (KL):

Régularisation vise à limiter le surapprentissage (overfitting) et contrôler l'erreur de type variance pour aboutir à de meilleures performances.

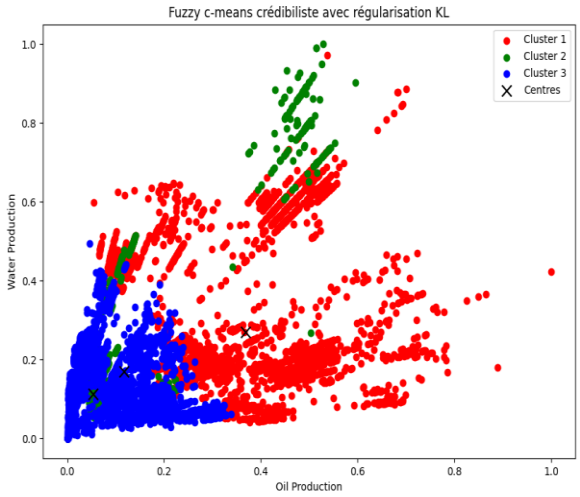


Figure 5. *Visualisation divergence de Kullback-Leibler, source : à partir de nos données de recherche [5], [21].*

Le graphique nous montre le classement des clusters en fonction de la régularisation KL de variable production en eau et production en oïl. Calcul de la divergence de Kullback-Leibler (KL): les valeurs sont 0.3, qui signifient que notre modèle est converge.

Par exemple, le cluster 2 était sous représenté et a été mieux ajusté après régularisation, la figure suivante présente l'évolution de la divergence de KL.

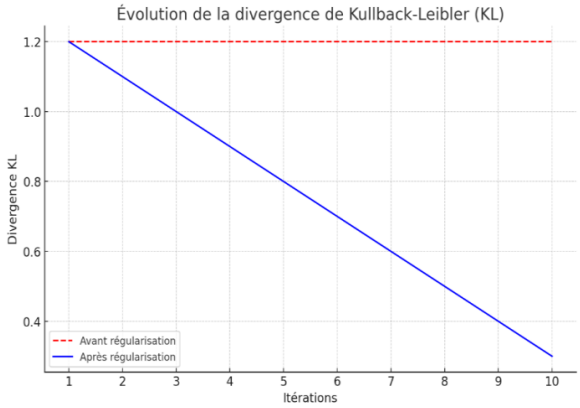
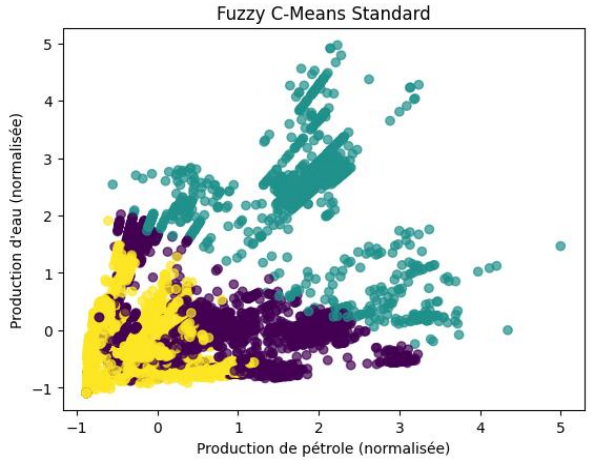
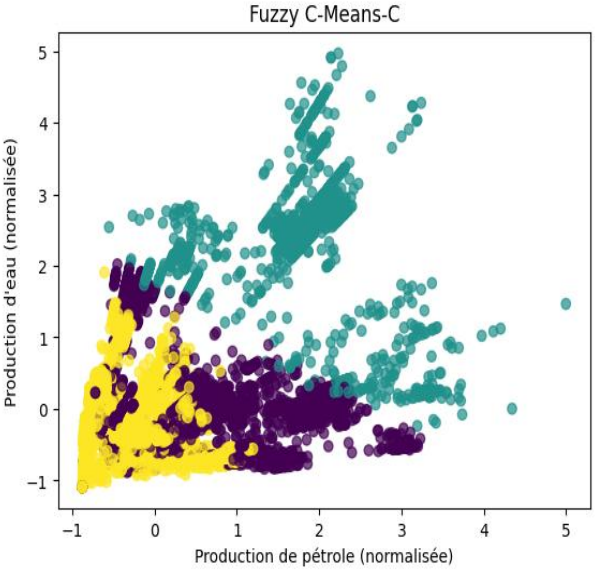


Figure 7. *Evolution de la divergence de Kulback-Leiber, source : à partir de nos données de recherche [5], [21].*

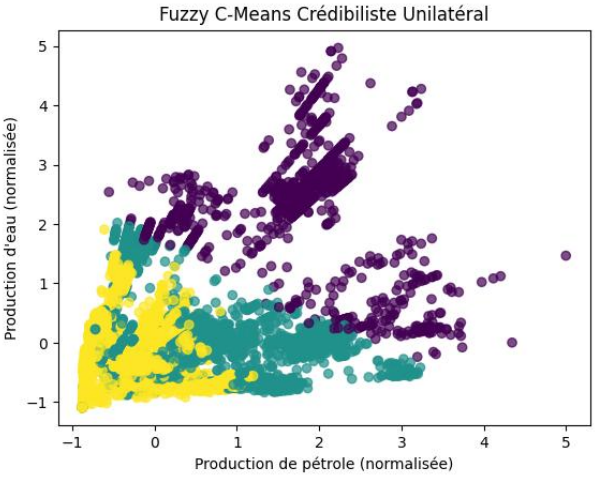
4.2.3. Visualisation de production d'eau en fonction du pétrole avec les trois méthodes.



(a)



(b)



(c)

Figure 6. Visualisation de production d'eau en fonction du pétrole avec les trois méthodes normalisées avec le score de silhouettes floue de Fuzzy C-Means standar, Fuzzy C-Means - Crédibiliste et Fuzzy C-Means - crédibiliste unilatéral, source : à partir de nos données de recherche [5], [21].

Les graphiques montrent les clusters formés avec des points

colorés qui représentent les puits pétroliers assignés à chaque cluster. Les clusters sont plus flous avec des recouvrements entre groupes et certains puits pourraient appartenir à plusieurs groupes avec des degrés d'appartenance similaires.

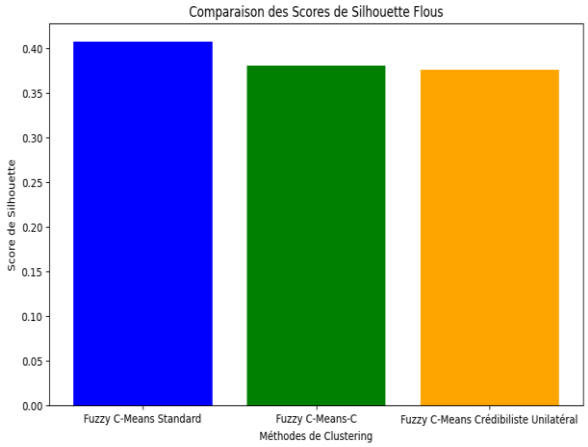


Figure 7. *Comparaison des scores de silhouette de Fuzzy C-Means standar, Fuzzy C-Means - Crédibiliste et Fuzzy C-Means - Crédibiliste Unilatéral, source : à partir de nos données de recherche [5], [21].*

Par contre, avec le Fuzzy C-Means - Crédibiliste en introduisant la crédibilité, la séparation des clusters est meilleure. Les puits avec des valeurs extrêmes sont mieux pris en compte tout en évitant les erreurs d'affectation. Le grand problème se pose aux puits proches des frontières des clusters, raison pour laquelle le Fuzzy C-Means – Crédibiliste Unilatéral fait la transformation accentuée des hauteurs d'appartenances et diminue les faibles appartenances.

Donc, il y a une meilleure séparation des clusters, les centres des clusters sont éloignés les uns des autres montrant une meilleure stabilité du modèle. En résumé, nous pouvons dire que le Fuzzy C-Means - Crédibiliste Unilatéral produit des clusters nets bien séparés et mieux définis. Cette méthode est plus robuste pour l'analyse des profils de production des puits pétroliers.

Tableau 6. Tableau comparatif des scores de chaque méthode, source : à partir de nos données de recherche [5], [21].

Méthode	Score de silhouette flou
Fuzzy C-Means Standard	0.4078
Fuzzy C-Means-C	0.3807
Fuzzy C-Means Crédibiliste unilatéral	0.3763

En conclusion, nous avons remarqué que le score de silhouette flou est moins adapté par rapport à ce type de données ; c'est pourquoi nous allons utiliser d'autres

4.2.4. Comparaison des résultats par rapport à d'autres métriques de Fuzzy C-Means standar, Fuzzy C-Means -C et Fuzzy C-Means - Crédibiliste unilatéral.

a. Analyse temporelle

Métriques adaptées au clustering flou le coefficient de partition, nous allons aussi intègrer la dimension temps et voir comment analyser la relation entre les trois types de production.

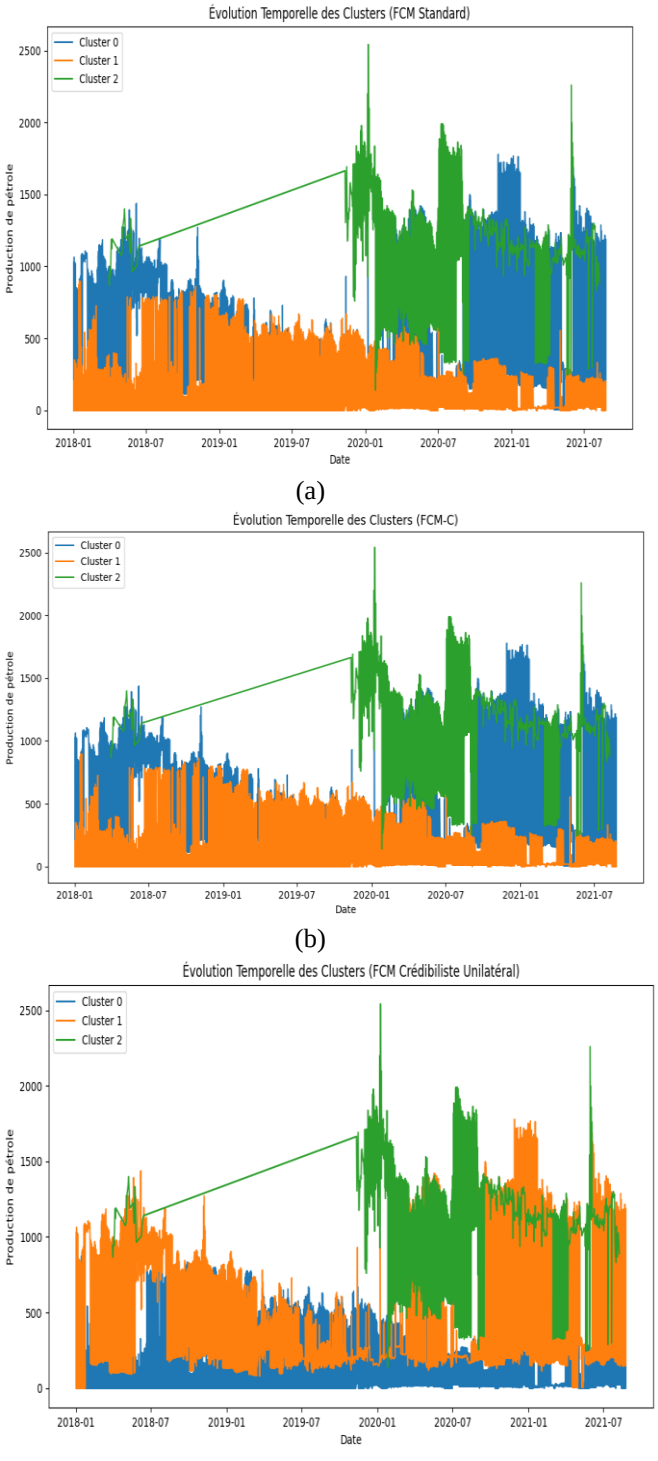


Figure 8. *Analyse temporelle des clusters, source : à partir de nos données de recherche [5], [21].*

b. Matrice de corrélation :

Tableau 7. Matrice de corrélation, source : à partir de nos données de recherche [5], [21].

	Oil production	Water production	Gas production

Oil production	1	0.49	0.36
Water production	0.49	1	0.47
Gas production	0.36	0.47	1

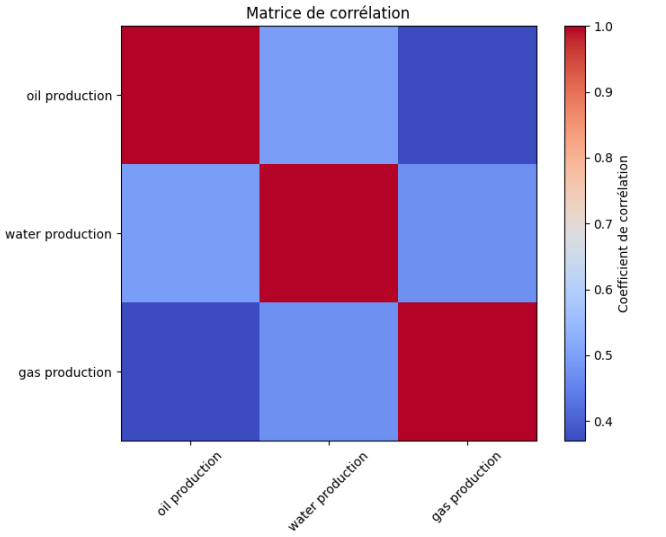


Figure 9. *Matrice de corrélation , source : à partir de nos données de recherche [5], [21].*

5. DISCUSSIONS

Dans cet article, nous avons comparé les résultats produits par quatre approches ; k-Means, FCM, FCM-C et FCM-CU sur un jeu des données réelles de la production du champ pétrolier MOTOBA. Nous avons notamment constaté que les résultats corroborent avec les progrès détectés dans les recherches précédentes tout en offrant une contribution distincte et innovante ; le score de la silhouette obtenue avec la méthode de FCM-CU de 0,69 est supérieur au score de la méthode FCM-C 0 ,64 ; cela démontre clairement que le FCM-CU produit des clusters mieux séparés et aussi la valeur de SSE(Sum of squared Errors) de 27,5 est plus faible par rapport au FCM-C de 41,9 et par rapport aux autres méthodes. En plus, les points sont proches de leur centre de cluster donc la méthode de FCM –CU réduit mieux l'erreur quadratique totale et regroupe les données de manière précise. En conclusion, nous pouvons dire que la méthode de FCM-CU est la plus performante, car elle offre des clusters bien séparés et compacts pour minimiser l'erreur.

De ce fait, notre modèle par surcroit introduit une méthode unilatérale qui renferme la capacité de parallélisme entre clusters. Cela est confirmé par l'indice de silhouette qui atteste une meilleure séparation des clusters par rapport aux méthodes classiques et antérieures. Par contre, dans le volet gouvernance des données sur base des observations d'Ismail et al. (2006), la gouvernance efficace des données pétrolières dépend de la qualité des informations et de la capacité à les exploiter pour la prise de décision.
Les résultats observés notamment le score de silhouette de 0,69 et le SSE de 27.5, dans ce cadre, indiquent que l'utilisation du FCM crédibiliste unilatéral favorise cette gouvernance en améliorant l'organisation des données pétrolières. Grâce au traitement de 1038 observations, nous avons pu classer les puits. Nous avons identifié les clusters qui représentent les différents types de puits, par exemple le cluster 1 (puits à forte production du pétrole) regroupe les puits qui produisent beaucoup de pétrole, mais peu de gaz et d'eau. Cluster 2 (puits à production équilibrée); ces puits produisent des quantités moyennes et équilibrées de pétrole, gaz et eau.

Les cluster 3 (puits à forte production d'eau) produisent beaucoup d'eau. Ce qui peut signaler une structuration du réservoir ou un problème d'extraction. Pour optimiser la production, il faut identifier les puits qui sont les plus rentables.

6. CONCLUSION

En conclusion, sur base des résultats et de la discussion, les résultats finaux regroupent des données avec la méthode de fuzzy c-means crédibiliste, en intégrant la régularisation unilatérale et la divergence de Kullback-Leibler pour son évaluation [26] qui montre que la fonction unilatérale ont convergée en moyenne.

Le cluster 1 qui regroupe les puits à forte production du pétrole, mais peu de gaz et d'eau. Le Cluster 2 les puits à production équilibrée, ces puits produisent des quantités moyennes et équilibrées de pétrole, gaz et eau. Le cluster 3 puits à forte production d'eau, ce qui peut signaler une structuration du réservoir ou un problème d'extraction.

Nous avons noté aussi que l'indice de silhouette flou dont le score est de 0.3877 et le SSE de 27.5 qui sont valables par rapport aux autres études, vu que nous avons utilisé les données réelles.

Dans cet article, les caractéristiques qui sont regroupées présentent un niveau élevé de similitude ou d'homogénéité. Ce qui permettra de les développer davantage en les appliquant à des domaines hétérogènes.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Ahad Rafati & Shahin Akarpour,2018.A new credibilistic clustering Method with Mahalanobis Distance,I .j.Mathematical Sciencccesss and Computing,2018,4,1-8(www.mecs-press.net/ijmsc)
[2] Baoding Liu, Yingjie Liu , 2002. "Expected Value of Fuzzy Set and Credibility Theory", Journal of Mathematical Analysis and Applications.
[3] C. Bezdek 1, Robert Ehrlich 2, William Full 3 , FCM: The fuzzy *c*-means clustering algorithm, Computers & Geosciences,Volume 10, Issues 2–3, 1984, Pages 191-203 ;
[4] Chakraborty, S. Nanda, D. Dutta Majumder, Narosa Publishing House, N.Delhi ,2007.Fuzzy Set and Writer Shibram Chakraborty, (with S. Ganguli), in, Fuzzy Logic and its Application in Technology and management, eds,68-70.
[5] Dataset et application,, 2025,https: //colab.research.google.com/drive/ 1w1GpdwKmobp69RilSx00Cpa-BskTKzIT#scrollTo=ZY9k3EPgZfT-

[6] Dominique Pastre, *Module Intelligence Artificielle*, Université Paris 5 - Maîtrise de mathématiques - Maîtrise MASS - MST ISASH 1999/2000(2000)
[7] Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3, 32-57. http://dx.doi.org/10.1080/01969727308546046(1973)
[8] Esmaeil Mehdizadeh ,Sadi-Nezhad Soheil,Reza Tavakkoli-Moghaddam, October 2008.Optimization of fuzzy clustering criteria by a hybrid PSO and fuzzy c-means clustering algorithm, Iranian Journal of Fuzzy Systems.
[9] Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, 2013 .M. Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems, 29, 1645-1660. Wireless Sensor Network, Vol.7 No.6, June 29, 2015.
[10] James C. Bezdek,1961.*Pattern Recognition with Fuzzy Objective Function Algorithms*, https://doi.org/10.1007/978-1-4757-0450-1, Media New York 1981, 272
[11] Jérémy Gelb et Philippe Apparicio , 2021.Apport de la classification floue c-means spatiale en géographie : essai de taxinomie socio-résidentielle et environnementale à Lyon ,Contribution of the spatial c-means fuzzy classification in geography: a socio-residential and environmental taxonomy in Lyon ,https://doi.org/10.4000/cybergeo.36414 .europ éen journal of geographie.
[12] Jian Su ,Shanglin Yao & He Liu , 2022 .Data Governance Facilitate Digital Transformation of Oil and Gas Industry, Frontiers in Earth Science
[13] Kizzy Nkem Elliot and Levi Damingo,application of artificial intelligence in the oil and gas industry,2024,International Research Journal of Modernization in Engineering Technology and Science 6(5):2582-5208,DOI: ,10.56726/IRJMETS57687
[14] Kizzy Nkem Elliot,Levi Damingo ,Application of artificial intelligence in the oil and gas industry,international research journal of modernization in engineering technology and science 6(5):2582-5208 doi: 10.56726/irjmets57687
[15] MathWorks,2023. *MATLAB Documentation*. https://www.mathworks.com/help/matlab/
[16] Mazyar Zahedi-Seresht,Bahram Sadeghi Bigham ,Shahrzad Khosravi &Hoda Nikpour ,2024. Oil Production Optimization Using Q-Learning Approach, Processes, *Processes* 2024, *12*(1), 10; https://doi.org/10.3390/pr12010110
[17] OECD,2021https://www.connaissancedesenergies.org/fiche-pedagogique/formation-du-petrole.
[18] Orhan Torkul ,Ismail hakki Cedimoglu,A. K. Geyik,2006.An application of fuzzy clustering to manufacturing cell design,journal of intelligent & Fuzzy Systems, 17(2):173-181.
[19] PENRECO,2021. Rapport sur les puits producteurs au niveau de Pinda Supérieur du champ MOTOBA. 40-45.
[20] Python Software Foundation,2023.*Python 3.11* Documentation. https://docs.python.org/3/.
[21] S.Samath et Senthil Kumar, 2013.Fuzzy Clustering using credibilistic Critical Values,International Journal of Computational Intelligence and Informatics,Vol.3 :N°3,October-December 2023.
[22] Salar Askari,2017.Oil reservoirs classification using fuzzy clustering, International Journal of Engineering, IJE TRANSACTIONS C: Aspects Vol. 30, No. 9, 1391-1400.
[23] Sampath et Al., 2018.An new credibilistic clustering Method With Mahalanobis Distance,I.J.Mathematical Sciences and Computing,4,1-18 .
[24] Trevor Hastie , Robert Tibshirani et Jerome Friedman, 2009. The *Elements of Statistical Learning*,Latest edition,XXII, 745 ,Springer New York, NY.
[25] W. Pedrycz, May 1996. "*Conditional Fuzzy C-Means*," Pattern Recognition Letters, Vol. 17, No. 6, pp. 625-632.
[26] Zadeh, L.A.,1965.Fuzzy Sets. Information Control, 8, 338-353. http://dx.doi.org/10.1016/S0019-9958(65)90241-X.
[27] Theodore Kotsilieris, ,(2022).Regularization Techniques for Machine Learning and Their Applications,mdpi.
[28] Feiping Nie, 2024. Moyennes floues C sans contrainte basées sur la régularisation de l'entropie : un modèle équivalent, Transactions IEEE sur l'ingénierie des connaissances et des donnéesPP(99):1-12, DOI:10.1109/TKDE.2024.3516085
[29] Abdulrahman Ibraheem, 2025.Regularizing cross entropy loss via minimum entropy and K-L divergence, DOI: 10.48550/arXiv.2501.13709
[30] EKA INDRA CAHYA, 2017. Application de la méthode fuzzy c-means sur le regroupement des agents d'impulsion au serveur cellulaire INDRA,Université Brawijaya MALANG,Indonesie.
[31] Chirag Goyal, 2021.Complete guide to regularization techniques in machine learning, https://www.analyticsvidhya.com/blog/2021/05/complete-guideto-regularization-techniques-in-machine-learning/.
[32] Hoerl, Arthur E and Kennard, Robert W,1970. *Ridge regression : Biased estimation for nonorthogonal problems, Technometrics 8* 27–51.