International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

Analyse Spatiale Et Classification Des Concentrations De Monoxyde De Carbone A Madagascar Par Techniques D'intelligence Artificielle

[Spatial Analysis And Classification Of Carbon Monoxide Concentrations Over Madagascar Using Artificial Intelligence Techniques]

Solofo RAFANOMEZANTSOA [a, 1], Jacques Chrysologue RATSIMAVO [b, 2], Adolphe Andriamanga RATIARISON [c, 3]

[a, b, c] Laboratoire de Dynamique de l'Atmosphère, du Climat et de l'Océan, Département de Physique et Applications, Faculté des Sciences et Technologies, Université d'Antananarivo et de Toliara, Madagascar
[1] fanantenanasolofo@gmail.com, [2] rat.chryso@gmail.com, [3] adolphe.ratiarison@univ-antananarivo.mg
Auteur Correspondant: Solofo RAFANOMEZANTSOA, fanantenanasolofo@gmail.com

Résumé – Le présent travail a pour objectif de réaliser une analyse spatiale et une classification des concentrations de monoxyde de carbone (CO) au-dessus de Madagascar à l'aide de techniques d'intelligence artificielle. La zone d'étude s'étend entre les longitudes 42° Est et 52° Est, et les latitudes 11° Sud et 27° Sud. Les données utilisées proviennent de la plateforme NASA GIOVANNI, au format NetCDF, et couvrent la période du 01 Janvier 1980 au 31 Décembre 2023, pour un total de 528 points de grille. L'approche méthodologique comporte trois étapes principales. La première consiste en une réduction de dimensionnalité à deux composantes pour permettre la visualisation spatiale des points de grille. Trois méthodes sont comparées : l'Analyse en Composantes Principales (ACP), t-distributed Stochastic Neighbor Embedding (t-SNE) et Linear Discriminant Analysis (LDA). La deuxième étape applique des algorithmes de clustering (k-Means, Fuzzy C-Means et DBSCAN) pour regrouper les points de grille aux concentrations similaires. La méthode du Coude permet de déterminer le nombre optimal de clusters. Enfin, la qualité des classifications est évaluée par les indices de Dunn (DI), Davies-Bouldin (DBI) et le coefficient de silhouette (SI). L'analyse révèle une classification optimale en trois zones distinctes de concentration de CO sur Madagascar.

Mots-clés : ACP, CO, DBSCAN, Méthode du coude, FCM, k-Means, LDA et t-SNE.

Abstract – The present work aims to perform a spatial analysis and classification of Carbon Monoxide (CO) concentrations over Madagascar using various artificial intelligence techniques. The study area is located between longitudes 42° East and 52° East and latitudes 11° South and 27° South. The data comes from the NASA GIOVANNI platform in NetCDF format. These data represent CO concentrations at 528 grid points from January 1, 1980, to December 31, 2023. The methodological approach consists of three main steps. The first involves dimensionality reduction to two components to enable spatial visualization of the grid points. Three methods are compared: Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Linear Discriminant Analysis (LDA). The second step applies clustering algorithms (k-Means, Fuzzy C-Means, and DBSCAN) to group grid points with similar concentrations. The Elbow method is used to determine the optimal number of clusters. Finally, the quality of the classifications is

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

evaluated using the Dunn Index (DI), Davies-Bouldin Index (DBI), and Silhouette Coefficient (SI). The analysis reveals an optimal classification into three distinct zones of CO concentration over Madagascar.

Keywords: CO, DBSCAN, Elbow method, FCM, k-Means, LDA, PCA and t-SNE

I. INTRODUCTION

L'analyse de la variabilité spatiale de concentration de monoxyde de carbone (CO) fait l'objet de nombreuses recherches scientifiques [1] et [2]. Le monoxyde de carbone (CO) n'est pas un gaz à effet de serre majeur, mais il joue un rôle indirect dans le réchauffement climatique en influençant la concentration du méthane et de l'ozone troposphérique [3] et [4]. Il est produit par la combustion incomplète du carbone en présence de dioxygène. Ce gaz est incolore, inodore et toxique à des concentrations élevées. La concentration de CO varie entre des niveaux faibles à élevés, provoquant divers effets et impacts sur la santé humaine, l'environnement et le climat (https://environnement.public.lu/fr/loft/air/Polluants_atmospheriques/CO/effets-CO.html). Selon l'Organisation Mondiale de la Santé (OMS), cette concentration peut atteindre entre 200 ppm et 12800 ppm, entraînant des symptômes tels que maux de tête, vertiges, nausées, fatigue, perte de conscience immédiate, et dans les cas les plus graves, le décès. (https://www.centreantipoisons.be/monoxyde-de-carbone/le-monoxyde-de-carbone-co-en-d-tail/quelles-sont-les-concentrations-toxiques-de CO). Il est donc important de connaître la variabilité spatiale de concentration de monoxyde de carbone.

L'objectif de cette étude est d'identifier les zones optimales et d'analyser la répartition spatiale des concentrations de CO à Madagascar en utilisant diverses techniques d'intelligence artificielle.

II. MATÉRIELS ET MÉTHODES

II.1. Données d'expérimentation

Les données d'expérimentation sont de données en points de grille sous format NetCDF (.nc) contenant les données de concentration de CO issus de plateforme NASA GIOVANNI (https://giovanni.gsfc.nasa.gov/giovanni/). Les latitudes et les longitudes sont bornés respectivement par 42° Est, 52° Est et 11° Sud, 27° Sud. La couverture temporelle s'étend du 01 Janvier 1980 et 31 Décembre 2023. La résolution des points de grille est de 0.5° x 0.625°.

II.2. Pré-traitements des données

La matrice initiale **M** des jeux de données est obtenue après les quatre étapes suivantes :

Étape 1 : Conversion des données du format NetCDF (.nc) au format .npy pour obtenir une matrice contenant les concentrations de CO en points de grille.

Étape 2 : Inversion des lignes de chaque matrice avec la commande Python flipud pour aligner correctement les données selon les latitudes et longitudes.

Étape 3 : Calcul, pour chaque point de grille, de la moyenne climatologique mensuelle.

Étape 4 : Construction de la matrice **M** considérée comme matrice initiale pour l'analyse spatiale et la classification. Les individus sont les points de grille et les variables sont les mois de l'année. La matrice **M** est donc de dimension 528 x 2.

L'ensemble des points de grille comprend 528 éléments. La nomenclature des points de grille est comme l'illustre à la figure 1.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

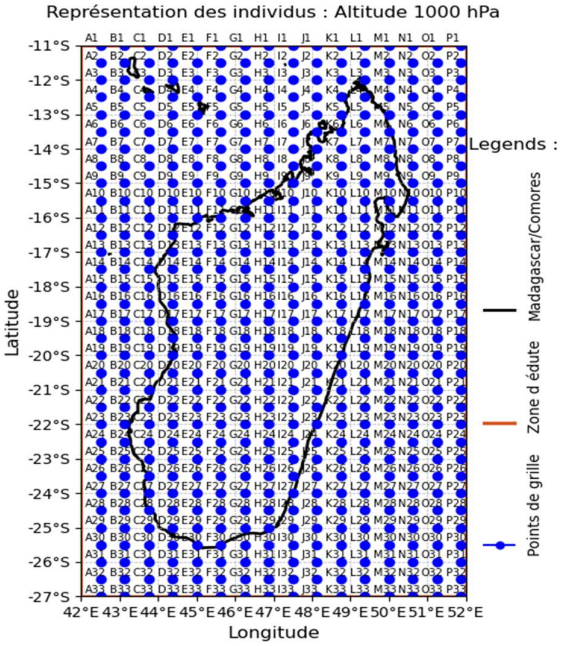

Figure 1. Représentation des points de grille sur la zone d'étude.

II.3. Méthodes

La démarche adoptée se subdivise en cinq (05) étapes :

— **Étape 1**: Réduction de la dimension des données initial à deux dimensions finales

— **Étape 2**: Visualisation des individus sur un plan

— **Étape 3**: Regroupement des individus par clustering

— **Étape 4**: Détermination le nombre optimal de clusters par la méthode de Coude (Elbow method)

— **Étape 5**: Evaluation des classifications obtenues par l'indice de Dunn, l'indice de Davies-Bouldin et l'indice de Silhouette.

a) Étape 1: Réduction de la dimension des données initial à deux dimensions finales

La matrice **M** est formée de 12 variables (les mois de l'année). Sa dimension initiale est donc 12. Les algorithmes ACP [5]; t-SNE [6] et LDA [7] permettent de réduire cette dimension à 2. L'intérêt de cette réduction est de permettre la visualisation des individus sur un plan, en préservant la proximité des points entre l'espace initial à haute dimension et l'espace réduit.

a.1) Analyse Composantes Principales (ACP)

L'ACP est une technique couramment utilisée pour réduire la dimensionnalité des hautes dimensions qui permettent de visualiser les individus sur un plan.

Les principes de l'ACP [5] sont de :

— Calculer la statistique descriptive des concentrations de CO : moyenne, l'écart type, valeurs minimale et maximale

— Centrer et réduire les données des concentrations de CO

— Déterminer de la matrice de corrélation puis en déduire le vecteur propre, la valeur propre

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

— Trouver les coordonnées des indices et variables et sélectionner les axes factoriels importantes

a.2) t-distributed Stochastic Neighbor Embedding (t-SNE)

Le t-SNE est une technique de machine learning de réduction de dimensionnalité particulièrement efficace pour visualiser des données de haute dimension [6]. L'algorithme t-SNE convertit les coordonnées en probabilités. Le principe est que les individus qui sont proche dans l'espace à haute dimension sont aussi proche dans l'espace à faible dimension finale.

Notons que $p_{j/i}$ est la probabilité conditionnelle définie par l'équation 1. La probabilité conditionnelle est élevée tandis que la séparation des points est plus large.

$$p_{j/i} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)} \quad (1)$$

Où σ_i est la variance de la gaussienne centrée sur le point de données x_i.

$\left\|x_i - x_j\right\|^2$: Distance entre les deux points, est grande alors $p_{j/i}$ est faible pour tous les j.

Et la probabilité conjointe symétrique p_{ij} est la somme $p_{j/i}$ et $p_{i/j}$ par deux fois n, définie par l'équation 2.

$$p_{ij} = \frac{p_{j/i} + p_{i/j}}{2n} \quad (2)$$

Pour notre cas, n = 12, appelée variables quantitatives représentant les 12 mois de l'année.

La probabilité en espace réduit que la représentation de x_i et x_j dans l'espace à deux dimensions soit proche est q_{ij} comme suit :

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq l} (1 + \left\|y_k - y_l\right\|^2)^{-1}} \quad (3)$$

Nous donnons l'algorithme t-SNE [6] comme suit :

Données:

- Charger les données (par exemple à 12 variables) $X = \{x_1, x_2, ..., x_{12}\}$
- Paramètre clé de la fonction coût : perplexité
- Paramètres d'optimisation : nombre d'itération, taux d'apprentissage, momentum $\alpha(t)$
- Résultat : représentation des données en basse dimension $\gamma^{(T)} = \{y_1, y_2\}$

Début:

1. Calculer les affinités par paires $p_{j/i}$ avec la perplexité par l'équation (1)

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

2. Définir l'équation (2)
3. Échantillonner la solution initiale $\gamma(0)=\{y_1,y_2\}$ à partir de $\mathrm{N}(0;10^{-4}I)$ où I désigne la matrice unitaire
4. Pour t = 1 à T faire :
   a. Calculer les affinités en basse dimension l'équation (3)
   b. Calculer le gradient $\frac{\delta C}{\delta y_i}=4\sum_j(p_{ij}-q_{ij})(y_i-y_j)(1+\|y_i-y_j\|^2)^{-1}$
   c. Mettre à jour $Y(t)=Y(t-1)+\eta\frac{\partial C}{\partial Y}+\alpha(t)\times(Y(t-1)-Y(t-2))$

   Où $Y(t)$ indique la solution à l'itération t, η indique le taux d'apprentissage ; $\alpha(t)$ représente l'écart à l'itération t
   d. Fin pour

Fin

a.3) Analyse Discriminante Linéaire (LDA)

LDA [7] est une technique de réduction de dimension **supervisée** qui cherche à maximiser la séparation entre les classes (S_C), et réduire les dimensions en projetant les individus qui sont proches sur un espace à haute dimension sont aussi proches sur un espace à faible dimension finale. En calculant la mesure la séparation entre les classes, nous utilisons la formule mathématique définie par l'équation (4) :

$$S_C=\sum_{j=1}^{C}N_j(\mu_j-\mu)(\mu_j-\mu)^T \quad (4)$$

Où

N_j : C'est le nombre d'échantillons dans la classe j.

$\mu_j=\frac{1}{N_j}\sum_{x_i\in C_j}x_i$: C'est la moyenne de la classe j

$\mu=\frac{1}{N}\sum_{i=1}^{N}x_i$: C'est la moyenne de toutes les données (la moyenne globale)

N : C'est le nombre total de données

x_i : C'est l'ensemble des données d'entrée de i (matrice M)

C_i : L'ensemble des échantillons appartenant à la classe j

C : Le nombre total de classes dans le jeu de données

$(\mu_j-\mu)$: Différence entre la moyenne de sa classe j et la moyenne globale

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

$(\mu_j - \mu)^T$: Transposée du vecteur précédent (sert pour former un produit matriciel)

$(\mu_j - \mu)(\mu_j - \mu)^T$: Matrice de covariance pour un point donné

Si la matrice de dispersion **intra-classe** S_I est **faible,** plus les classes sont **bien regroupées, donnée par la formule comme suit :**

$$S_I = \sum_{j=1}^{C} \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j)^T \quad \textbf{(5)}$$

Nous donnons l'algorithme LDA comme suit :

Données :

- Charger les données (par exemple à 12 variables) $x_i = \{x_1, x_2, ..., x_{12}\}$
- Résultat attendu : représentation des données en faible dimension

Début:

1. Calculer les moyennes des classes et globale
2. Déterminer la matrice de dispersion intra-classe en utilisant l'équation (5)
3. Trouver la mesure la séparation entre classe en utilisant l'équation (4)
4. Transformer la matrice initiale en matrice de projection avec maximiser la séparation entre les classes et la matrice de dispersion intra-classe
5. Projeter les données obtenues à faible dimension sur un plan

Fin

a) Étape 2 : Visualisation des individus sur un plan

Avec l'opération de réduction de dimension des données, la matrice de jeux de données initiale **M** de dimension 528 x 12 devient une matrice de dimension 528 x 2. Ainsi, pour chaque individu, c'est à dire pour chaque point de grille, nous avons deux composantes. Nous pouvons donc visualiser chaque individu sur un plan avec l'ACP (respectivement avec t-SNE et LDA). La visualisation se fait sur le plan ACP (respectivement sur le plan t-SNE et LDA).

b) Étape 3 : Regroupement des individus par clustering

Le clustering consiste à regrouper et classer les individus représentés sur un plan suivi leur centroïdes. Les algorithmes utilisées sont k-Means [8], Fuzzy C-Means [9] et DBSCAN [10].

c.1) L'algorithme de k-Means

L'objectif de l'algorithme de k-Means consiste à minimiser la somme des carrés intra-cluster. Cette somme est représentée par une fonction de coût appelée fonction objective à minimiser $J(V, X)$. Elle mesure la compacité des clusters.

$$J(X,V) = \sum_{j=1}^{k} J_i(x_i, v_j) = \sum_{j=1}^{k} \left(\sum_{i=1}^{m} u_{ij} . d^2(x_i, v_j) \right) \quad (6)$$

Où $X = \{x_1, x_2, x_3 ... x_n\}$ est l'ensemble des données

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

V est l'ensemble des centroïdes des données

k est le nombre de clusters

m est le nombre total de points dans l'ensemble de données

x_i est un point de données

v_j est le centroïde du cluster c_j

$d^2(x_i, v_j)$ est la distance euclidienne au carré entre x_i et v_j définie par:

$$d^2(x_i, v_j) = \left\| \sum_{k=1}^{n} x_k{}^i - v_k{}^j \right\|^2 \quad (7)$$

u_{ij} est une matrice d'appartenance indiquant si un point x_i appartient au c_j :

$$u_{ij} = \begin{cases} 1; si\ d^2(x_i, v_j) \leq d^2(x_i, v_{j^*}) j \neq j^*, \forall j^* = 1, ..., k \\ 0; sinon \end{cases} \quad (8)$$

L'équation de centroïde est donnée par la formule suivant :

$$v_j = \frac{1}{|c_j|} \sum_{i, x_i \in c_j}^{m} x_i \quad (9)$$

Avec $|c_j|$ est le nombre de points dans le cluster c_j

L'algorithme de k-Means [8] est comme suit :

Début:

1- Ensemble des données (dans notre recherche, les résultats obtenus par l'algorithme de l'ACP, t-SNE et LDA)

2- Déterminer le nombre de clusters (dans notre travail, clusters = 3)

3- Initialiser aléatoirement les centroïdes

4- Répéter :

- Déterminer la matrice d'appartenance u_{ij} en utilisant l'équation (8)
- Calculer la fonction objective J (X, V) en utilisant l'équation (6)
- Recalculer les centroïdes v_j de chaque cluster en utilisant l'équation (9)

5- Répéter jusqu'à ce que les centroïdes ne changent pas

6- Tracer les résultats du regroupement

Fin

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

c.2) L'algorithme de Fuzzy C-Means

L'algorithme Fuzzy C-Means ou FCM est un algorithme d'apprentissage **non supervisé** et consiste à partitionner les K observations en N classes de manière à minimiser la similarité des observations à l'intérieur de chaque classe [9]. Mathématiquement traduit comme suit :

$$J_{FCM} = \sum_{k=1}^{k}\sum_{i=1}^{N}(U_{ki})^m \left\|x_k - c_i\right\|^2 \quad (10)$$

Où

k : Nombre de point de données (pour notre cas k=528)

N : Nombre de cluster (pour notre cas N=3)

m : facteur de fuzzification ou facteur flou tel que $m > 1$

x_k : k^{ème} de données

c_i : i^{ème} de centre de cluster

U_{ki} : degré d'appartenance de x_k dans le i^{ème} groupe tel que $0 \le U_{ki} \le 1$.

Pour un point x_k, la somme des valeurs d'appartenance de tous les clusters est égale à 1. Si la valeur de m est proche de 1, les limites entre les groupes deviennent plus nettes.

Enfin, cet algorithme utilise de mettre à jour la matrice d'appartenance U_{ki} et les centres des clusters v_k définie par l'équation 9 et 10 :

$$U_{ki} = \left(\sum_{l=1}^{k} \left(\frac{\left\|x_i - v_k\right\|}{x_i - v_l} \right)^{-\frac{2}{m-1}} \right)^{-1} \quad (11)$$

U_{ki} : Degré d'appartenance du point x_i au cluster k

v_k : Centroïde du cluster k

N : Nombre total de points de données

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m} \quad (12)$$

$\left\|x_i - v_k\right\|$: Distance entre x_i et le centroïde v_k

m : Paramètre de flou ($m > 1$), qui contrôle le niveau de flou dans la classification (m est souvent pris entre 1.1 et 2.7).

L'algorithme FCM peut être résumé comme suit lors de la mise en cluster :

Début :

1. Ensemble des données (dans notre recherche, les résultats obtenus par l'ACP et t-SNE et LDA)
2. Fixer les paramètres : le nombre de classe N=3 ; le degré flou m (par défaut m=2)
3. Mettre à jour la matrice U_{ki} des degrés d'appartenance par la relation de l'équation (11)
4. Mettre à jour le vecteur des centres des classes par l'équation (12)
5. Calculer la fonction objective l'équation (10)
6. Répéter les étapes plus 2 jusqu'à ce que J_{FCM} s'améliore de moins d'un seuil minimal spécifié (par défaut, 0.01 entre deux itérations successives) ou après un nombre maximal d'itération spécifiée (par défaut, 25 itérations)
7. Tracer les résultats du regroupement

Fin

c.3) L'algorithme de DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering de données basé sur la densité [10]. Il trouve le nombre de clusters à partir de la distribution de densité estimée des nœuds correspondants. Cet algorithme calcule la densité entre deux points de données. La formule pour calculer la distance euclidienne entre les deux points sur un plan est définie par :

$$d^2(x_p, x_q) = \left\| \sum_{k=1}^{n} (x_k^p - x_k^q) \right\|^2 \quad (13)$$

Où n est le nombre de dimensions de chaque point de données

$x_p = (x_1^p, x_2^p, x_3^p, ..., x_n^p)$: Coordonnées du point p dans un espace à n dimensions

$x_q = (x_1^q, x_2^q, x_3^q, ..., x_n^q)$: Coordonnées du point q dans un espace à n dimensions.

$x_k^p - x_k^q$: Différence entre les coordonnées du point p et du point q selon la dimension k.

L'ensemble des points situés dans le voisinage de x_p est donnée par l'équation comme suit:

$$N_\varepsilon(x_p) = \left\{ x_q \in X \middle| d^2(x_p, x_q) \leq \varepsilon \right\} \quad (14)$$

X est l'ensemble des points de données.

ε est le rayon du voisinage autour d'un point donné x_p

η est le nombre minimum de points requis pour former un cluster, l'équation donnée par :

$$\eta \leq \left| N_\varepsilon(x_p) \right| \quad (15)$$

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

L'algorithme DBSCAN réalise un clustering automatique à partir des paramètres ε (eps) ≥ 0 et η (MinPts) ≥ 5. Pour chaque combinaison des valeurs de ε (eps) et η (MinPts), l'algorithme génère une classification des données. Lorsqu'un point de données est évalué, si son voisinage ε contient suffisamment de points, un cluster est initié. Sinon, ce point est étiqueté comme bruit.

L'algorithme DBSCAN [8] nous donne à résumer comme suit lors de la mise en cluster :

Début :

1 Charger les données (dans notre recherche, les résultats obtenus par la réduction de dimensions de l'ACP, t-SNE et LDA)
2 Fixer le nombre de clusters
3 Entrer les paramètres de DBSCAN : ε (eps) et η (MinPts)
4 Initialiser les centroïdes
5 Calculer la valeur de la fonction de DBSCAN avec les données d'expérimentation
6 Recalculer le centroïde de chaque cluster en utilisant l'équation (9)
7 Visualiser les individus sur un plan et le centroïdes

Fin

c) Étape 4: Détermination du nombre optimal de clusters par la méthode de Coude

À la fin des étapes précédentes, il existe 3 méthodes pour déterminer le nombre optimal de clusters :

— Par la visualisation des individus sur un plan t-SNE avec de paramètre perplexité
— L'algorithme de DBSCAN avec les paramètres clés eps et MinPts
— L'algorithme d'Elbow point (méthode du Coude)

La méthode du coude (Elbow method) [11] qui consiste à déterminer le nombre optimal de clusters dans un algorithme de clustering. Cette l'algorithme repose sur l'analyse de l'inertie intra-cluster ou fonction de distortion (somme des carrés des distances des points au centroïde de leur cluster). L'inertie totale pour un nombre de clusters k est donnée par l'équation (16) comme suit :

$$J(k) = \sum_{i=1}^{k} \sum_{x \in c_i} \left\| x - u_i \right\|^2 \quad (16)$$

Où k est le nombre de clusters

c_i est le cluster i

u_i est le centroide du cluster i

x est un point de données appartenant au clusters c_i

$\left\| x - u_i \right\|^2$ est la distance euclidienne au carré entre un point et le centroïde de son cluster

Le « **coude** » de la courbe correspond au point à partir duquel la réduction de la fonction de coût J devient peu significative lorsque le nombre de clusters augmente. Ce point est considéré comme le nombre optimal de clusters. Autrement dit, il s'agit du premier point où la variation de l'inertie commence à décroître lentement. La diminution de la variation de l'inertie $\Delta(k)$ est exprimée par l'équation suivante :

$$\Delta(k) = J(k) - J(k+1) \quad (17)$$

Nous donnons l'algorithme du Coude comme suit :

Début:

1. Charger les données à partir de résultat des techniques de réduction dimension (matrice M)
2. Initialiser inertie J
3. Plage de k à tester : k_{min} à k_{max} (par défaut, $k_{min} = 1$, $k_{max} = 11$)
4. Calculer l'inertie pour chaque k :

   Pour chaque k dans (k_{min} , k_{max}) :

   - Exécuter avec l'un de clustering avec k clusters
   - Stoker l'inertie J

   Fin pour
5. Trouver k en utilisant l'équation (16) et (17)
6. Retourner k optimal et afficher la courbe représentant le résultat k optimal

Fin

d) Étape 5: Evaluation des classifications

À l'issue des étapes 3 et 4, nous avons généré neuf (09) modèles de classification distincts, résultant des combinaisons suivantes :

- **Modèle 1** : ACP (Analyse en Composantes Principales) + k-Means
- **Modèle 2** : ACP + FCM (Fuzzy C-Means)
- **Modèle 3** : ACP + DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- **Modèle 4** : t-SNE (t-Distributed Stochastic Neighbor Embedding) + k-Means
- **Modèle 5** : t-SNE + FCM
- **Modèle 6** : t-SNE + DBSCAN
- **Modèle 7** : LDA (Linear Discriminant Analysis) + k-Means
- **Modèle 8** : LDA + FCM
- **Modèle 9** : LDA + DBSCAN

Pour sélectionner le meilleur modèle de classification, nous évaluons chaque modèle candidat à l'aide de trois indicateurs à savoir l'Indice de Dunn (DI), l'Indice de Davies-Bouldin (DBI) et le coefficient de Silhouette (SI) [12] et [13]. Dans notre étude, nous accordons la priorité au SI, puis au DBI et enfin au DI.

e.1) Indice de Dunn

L'indice de Dunn [12] est une mesure de la qualité d'un partitionnement des données. Il cherche la distance minimale qui sépare deux classes dans la partition tout en tenant compte de la distribution des éléments à l'intérieur des classes. La valeur de l'indice de

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

Dunn est un nombre réel positif. La valeur de l'indice de Dunn élevée indique une meilleure classification. Nous donnons la formule générale de l'indice de Dunn par l'équation 15.

$$DI = \frac{\min_i \left\{ \min_{i \neq j} d_\alpha(c_i, c_j) \right\}}{\max_h s_\alpha(c_h)} \tag{18}$$

Où c_i ou c_j : Individu ou point ;

$d_\alpha(c_i, c_j)$: Distance euclidienne entre les deux individus c_i et c_j ;

$s_\alpha(c_h)$: Diamètre du groupe ;

$h \geq 2$, le nombre de groupe (Clusters) que l'on veut former, dans notre recherche h = 3.

Une valeur de DI est plus élevée, cela signifie que les clusters sont bien séparés et faiblement dispersés, ce qui indique une classification plus distincte. A l' inverse, faible separation.

e.2) Indice de Davies-Bouldin

L'indice de Davies-Bouldin [14] est une métrique couramment utilisée en apprentissage automatique pour évaluer la qualité d'une partition d'un ensemble de données. Cet indice tente de minimiser la distance moyenne entre les points d'un cluster tout en maximisant la séparation entre les clusters. La formule mathématique est définie comme suit :

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \leq j \leq k, j \neq i} \left(\frac{diam(c_i) + daim(c_j)}{dis(c_i, c_j)} \right) \tag{19}$$

Où k : C'est le nombre total de clusters que nous avons dans notre partition de données, dans notre travail k = 3 ;

$diam(c_i)$: C'est le diamètre du cluster i, qui mesure la compacité du cluster. Une faible valeur de $diam(c_i)$ indique que les points du cluster sont proches les uns des autres, ce qui est un signe de bonne compacité ;

$diam(c_j)$: C'est le diamètre du cluster j ;

$dis(c_i, c_j)$: C'est la distance entre les centroïdes des clusters i et j. Cette distance mesure à quel point les clusters sont séparés les uns des autres. Une grande valeur de $dis(c_i, c_j)$ indique une bonne séparation entre les clusters.

Un bon clustering tend à minimiser l'indice de Davies-Bouldin. Plus cet indice est faible, meilleure est la qualité de la partition (clusters).

e.3) Coefficient de Silhouette (SI)

Le coefficient de Silhouette (ou indice de Silhouette) [15] est une métrique permettant d'évaluer la qualité d'un clustering (regroupement). Pour chaque point de données, il mesure la cohésion et la séparation. Le coefficient de Silhouette $S(i)$ est défini comme :

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{20}$$

Où, i est l'ensemble des tous les points

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

$a(i)$ est la distance moyenne entre le point i et les autres points du même cluster (cohésion intra-cluster).

$b(i)$ est la distance moyenne entre le point i et les points du cluster le plus proche (séparation inter-cluster).

Nos recherches, nous utilisons l'indice global de Silhouette est la moyenne des coefficients de Silhouette de tous les points :

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S(i) \tag{21}$$

Où n est le nombre des données.

Si la moyenne est proche de **1**, le point est bien classé dans son cluster, bonne séparation. En revanche, la moyenne est proche de -**1**, le point est probablement mal classé.

Plus cette moyenne est élevée, meilleur est la partition des clusters.

III. RÉSULTATS DES DONNÉES ET INTERPRÉTATIONS

III.1. Résultats de la visualisation des données

La figure 2 nous compare les résultats avec les trois méthodes de la visualisation des individus ACP, t-SNE et LDA. La visualisation sur le plan de l'ACP, t-SNE et LDA est différente forme. À gauche et à droite, le plan ACP et LDA montre les individus répartis dans les quatre quadrants, avec cette représentation, le nombre de cluster n'est pas évident. Par contre, les résultats avec l'algorithme t-SNE nous donnent un nombre de cluster évident. En effet, à chaque valeur du paramètre perplexité correspond une visualisation sur le plan t-SNE. La simulation a été faite avec perplexité allant de 5 à 50 par pas de 1. Avec la valeur de 13, la meilleure visualisation est obtenue avec nombre de cluster 3.

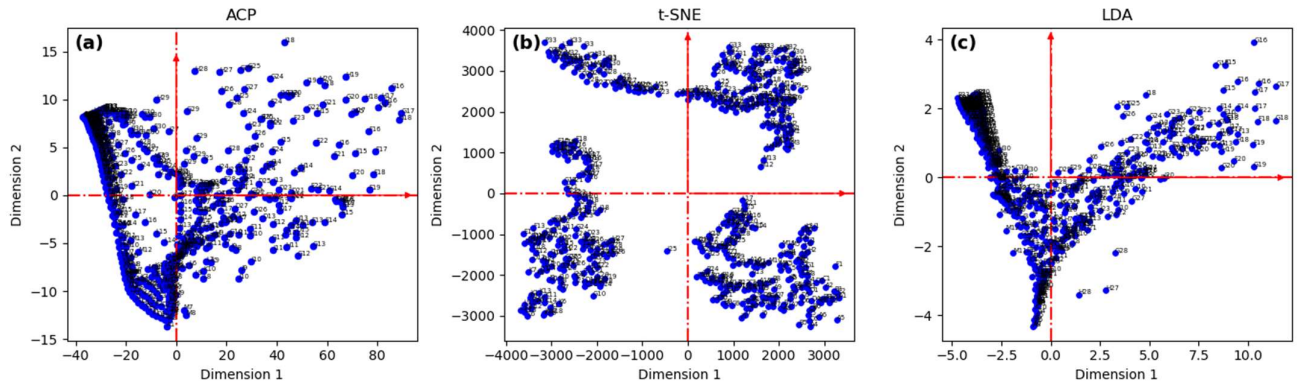

Figure 2 : Visualisation des individus (points de grille) sur le plan ACP, t-SNE et LDA.

III.2. Résultats de détermination le nombre optimal de clusters (Elbow method)

a. Résultats du modèle ACP, t-SNE et LDA + k-Means

L'algorithme t-SNE produit les nombres de clusters évident par la visualisation sur un plan de t-SNE avec le paramètre importante perplexité allant de 5 à 50 par pas de 1. Avec la **valeur de 13,** on obtient le nombre de cluster K = 3.

La figure 3 illustre la courbe du coude obtenue à partir des combinaisons ACP + k-Means, t-SNE + k-Means et LDA + k-Means. Cette courbe montre que l'inertie décroît rapidement lorsque le nombre de clusters augmente, avant de se stabiliser progressivement. La méthode du coude identifie le point où la **pente** commence à s'aplatir, indiquant ainsi le nombre optimal de clusters. Dans chaque

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

cas, les résultats suggèrent la formation de **3 clusters** (K = 3). De plus, la méthode de visualisation t-SNE projetée sur un plan et la méthode du coude convergent vers le même résultat, confirmant la cohérence de cette estimation.

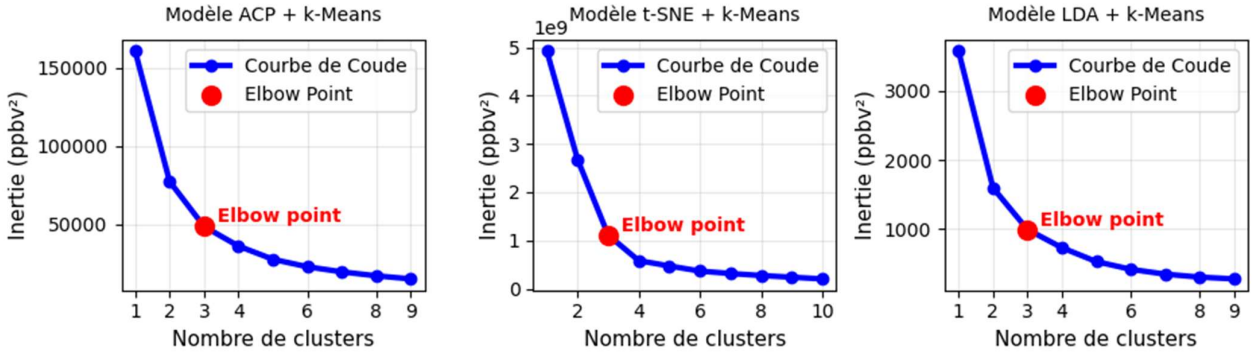

Figure 3 : Résultats de la méthode du Coude pour ACP, t-SNE et LDA avec k-Means.

b. Résultats du modèle ACP, t-SNE et LDA + Fuzzy C-Means

Comme dans l'interprétation précédente, la combinaison de l'ACP, du t-SNE et de la LDA avec le FCM a permis d'obtenir un nombre de clusters égal à 3 (C = 3), comme le montre la figure 4.

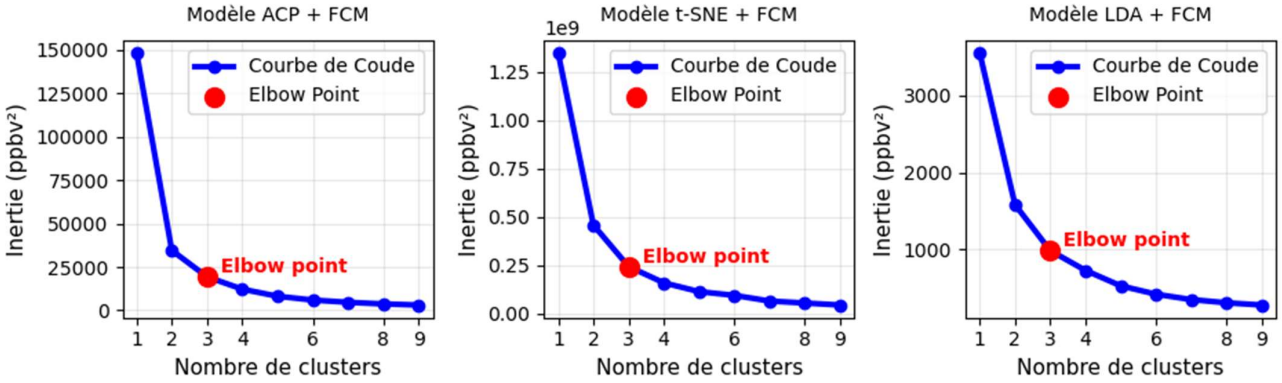

Figure 4 : Résultats de la méthode du Coude pour ACP, t-SNE et LDA avec Fuzzy C-Means.

c. Résultats du modèle ACP, t-SNE et LDA + DBSCAN

L'algorithme DBSCAN génère automatiquement le nombre de clusters. Chaque combinaison de deux paramètres clés (eps et MinPts) permet d'obtenir un nombre spécifique de clusters. Le premier résultat, obtenu avec l'ACP et DBSCAN, correspond à (eps = 2.7, MinPts = 10), le deuxième, avec t-SNE et DBSCAN, à (eps = 0.4, MinPts = 14), et le dernier, également avec DBSCAN, à (eps = 0.6, MinPts = 18). À chaque classification, trois nombres de clusters différents sont produits.

En résumé, l'algorithme t-SNE, la méthode du coude et la combinaison de l'ACP, du t-SNE et de la LDA avec DBSCAN produisent **trois clusters** (03) à retenir pour la classification finale de la concentration de CO.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

III.3. Résultats du regroupement des individus par clustering

a. Résultats du regroupement des individus par k-Means

La figure 5 nous compare les résultats du clustering k-Means sur trois types de projections des individus (données de CO). Elle représente trois clusters respectivement plus proche leurs centroïdes par le critère de la méthode du coude et l'algorithme t-SNE.

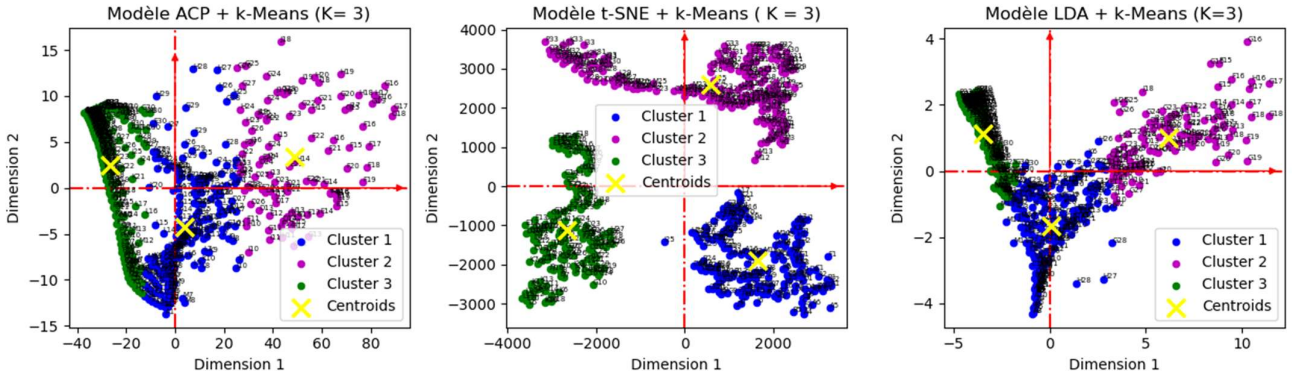

Figure 5 : Regroupement des individus sur le plan ACP, t-SNE et LDA par k-Means.

b. Résultats du regroupement des individus par Fuzzy C-Means

La figure 6 nous illustre les regroupements des individus par combinaison de ACP, t-SNE et LDA avec Fuzzy C-Means. Chaque représentation a un paramètre flou m de FCM, première résultat m est 2.7, le deuxième est 1.3 et le dernier est respectivement 1.2. Comme le critère de la méthode de Coude et l'algorithme t-SNE, les nombres de clusters obtiennent est 3 plus proche de leur centroïdes respectif montrant une plus grande homogénéité.

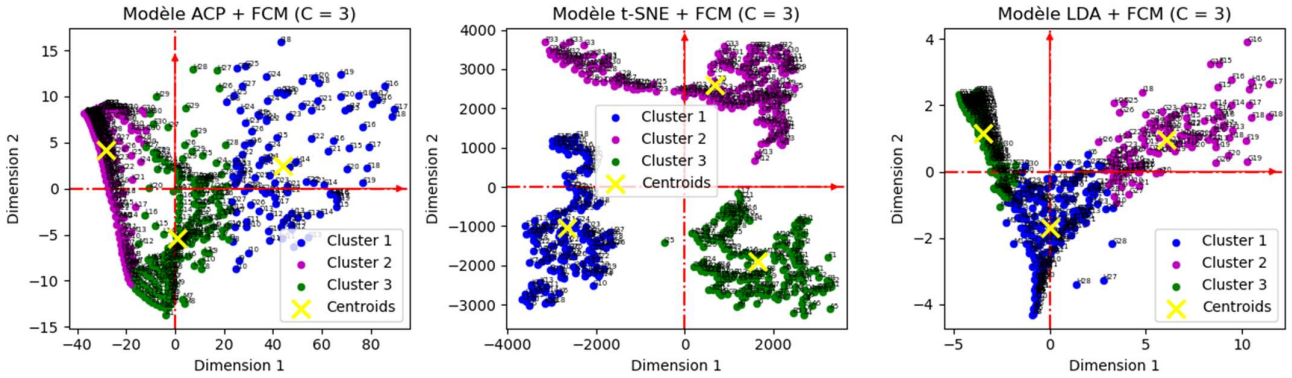

Figure 6 : Regroupement des individus sur le plan ACP, t-SNE et LDA par Fuzzy C-Means.

c. Résultats du regroupement des individus par DBSCAN

La figure 7 présente le modèle de classification obtenu par la combinaison des trois techniques de réduction de dimensionnalité des données de CO avec l'algorithme de clustering DBSCAN. Chaque représentation repose sur deux paramètres clés de DBSCAN : eps et MinPts, utilisés pour déterminer le nombre de clusters. Les valeurs retenues sont respectivement : (2.7, 10) pour la première méthode, (0.4, 14) pour la deuxième, et (0.6, 18) pour la dernière.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

Chaque classification a permis d'identifier trois clusters, chacun associé à un centroïde. Les individus ne pouvant être affectés à aucun cluster sont représentés comme des points de bruit. Le résultat obtenu par la combinaison de t-SNE et DBSCAN est identique à celui obtenu avec l'algorithme k-Means (ou FCM).

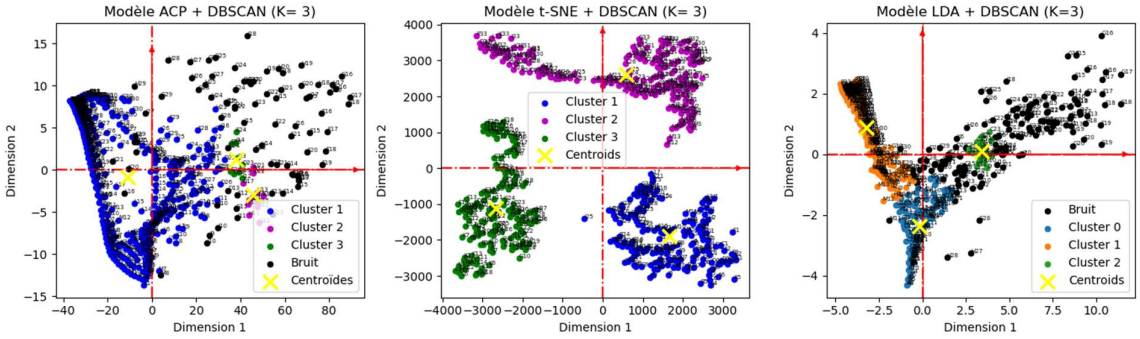

Figure 7 : Regroupement des individus sur le plan ACP, t-SNE et LDA par DBSCAN.

III.4. Résultats de l'évaluation des classifications par DI et DBI et Silhouette

Enfin, nous comparons la qualité de nos différentes méthodes de regroupement en calculant trois scores : l'indice de Dunn (DI), l'indice de Davies-Bouldin (DBI) et coefficient de Silhouette. Ces indices nous permettent d'évaluer si les groupes formés sont bien séparés et homogènes. Plus l'indice de Dunn est grand et plus l'indice de Davies-Bouldin est petit et coefficient de Silhouette est élevé, meilleure est la classification. Les résultats de ces évaluations sont résumés dans le Tableau 1.

Les classifications issues de la combinaison de t-SNE avec k-Means, FCM et DBSCAN sont plus proches de celles obtenues avec LDA combiné à k-Means et FCM, bien que les indices différentient.

Dans notre recherche, nous sélectionnons les meilleures classifications obtenues par la combinaison de l'algorithme t-SNE avec k-Means, FCM et DBSCAN. Cette combinaison représente le meilleur équilibre entre la séparation et la compacité des clusters, avec des clusters plus éloignés.

Tableau 1 : Résultats de l'analyse statistique de l'évaluation des classifications.

Réduction et visualisation des données	Clustering	Elbow point (K ou C)	Indice de Dunn	Indice de Davies Bouldin	Indice de Silhouette
Représentation sur le plan PCA	k-Means	3	0.014	0.578	0.552
	FCM	3	0.017	0.571	0.551
	DBSCAN	3	0.009	0.540	0.443
Représentation sur le plan t-SNE	**k-Means**	**3**	**0.153**	**0.561**	**0.598**
	FCM	**3**	**0.153**	**0.561**	**0.598**
	DBSCAN	**3**	**0.153**	**0.561**	**0.598**
Représentation sur le plan LDA	k-Means	3	0.017	0.570	0.570
	FCM	3	0.017	0.570	0.570
	DBSCAN	3	0.017	0.469	0.349

III.5. Résultat final de la classification de concentration de CO

Selon le critère de l'indice de Dunn, l'indice de Davies-Bouldin et l'indice de Silhouette, nous avons trois zones de monoxyde de carbone. La figure 8 nous présente le résultat final de la régionalisation en zone de concentration de CO selon notre propre démarche. La zone d'étude est distinguée clairement en trois zones :

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

La zone 1 (Z_1) : Elle est caractérisée par une concentration de CO très élevée allant de 58 à 80 ppbv (figure 9), est principalement localisée dans les régions centrales et occidentales de Madagascar (presque sur la terre). Les points de grille correspondants sont représentés en points rouges sur la figure 8.

La zone 2 (Z_2) : C'est l'ensemble des points colorées en vert sur la figure 8. Ces points ont des concentrations moyennes de CO allant de 50 à 65 ppbv (figure 9). Elle se situe généralement dans canal de Mozambique et la zone côtière ou périphérique de Madagascar.

La zone 3 (Z_3) : Colorée en bleu sur la figure 8, correspond aux régions présentant les plus faibles concentrations de CO allant de 45 à 58 ppbv (figure 9). Elle s'étend principalement dans l'Océan Indien de Madagascar.

La figure 9 illustre et confirme la qualité de la classification de la concentration de CO. À travers les courbes de chaque zone, on observe que la régionalisation de cette concentration est bien distinguée. De plus, ces courbes montrent une tendance générale à la hausse au fil du temps (1980 - 2023).

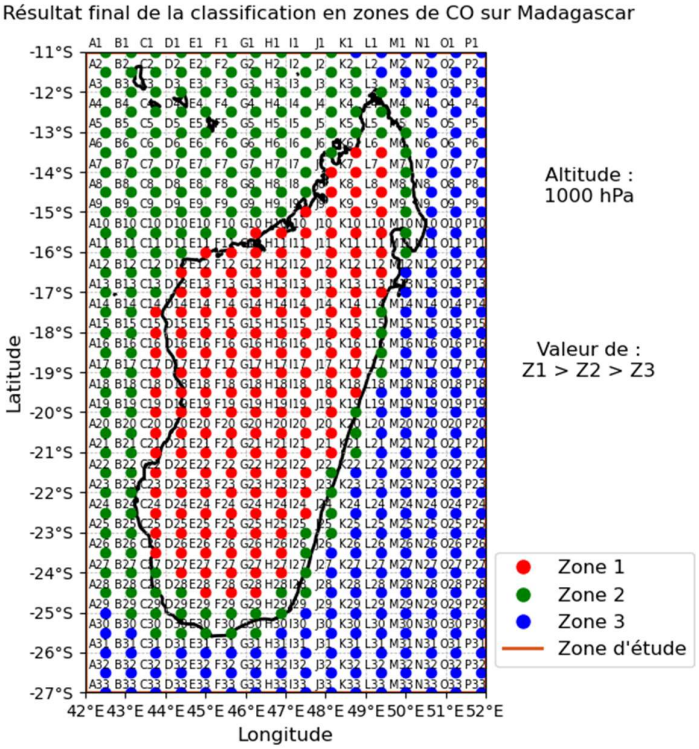

Figure 8 : Résultat final de la régionalisation en zones de monoxyde de carbone.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-019.

https://ijpsat.org/
Vol. 57 No. 2 June 2026, pp. 300-318

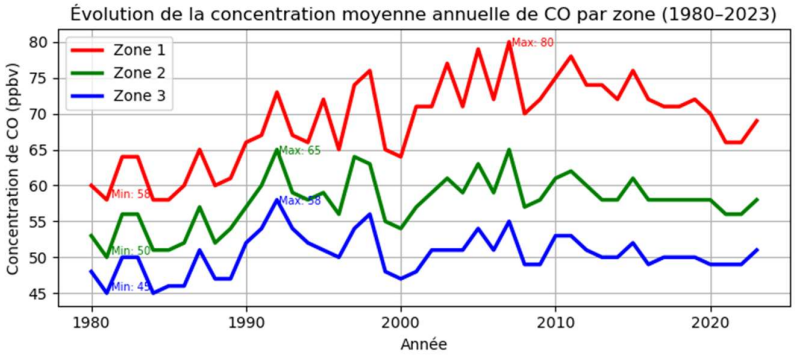

Figure 9 : Courbes d'évolution de la concentration moyenne annuelle de CO par zone.

IV. DISCUSSION

Plusieurs recherches scientifiques [16] et [17] procèdent à la régionalisation de points de grille en utilisant des paramètres climatiques tels que le vent, les précipitations, l'humidité et la température. Les méthodes employées reposent généralement sur une ou deux techniques d'intelligence artificielle.

Notre approche est similaire à celles de [16] et [17], mais se distingue par l'utilisation du monoxyde de carbone comme variable d'étude. Nous combinons également plusieurs techniques : algorithmes de réduction de dimension, algorithmes de clustering, et indices d'évaluation de classification. Enfin, la méthode du coude (Elbow method) a été utilisée pour déterminer le nombre optimal de clusters.

Notons que cette étude se limite à un nombre fixe de clusters (au final, trois clusters).

V. CONCLUSION

L'analyse de la concentration de CO à Madagascar met en évidence une forte hétérogénéité spatiale. Trois grandes zones distinctes ont été identifiées. En ce qui concerne les méthodes de visualisation, l'algorithme t-SNE s'est révélé plus performant que l'ACP et la LDA pour représenter les points de grille.
En perspective, nous proposons de mener une étude visant à prédire la concentration de CO par zone à l'aide des techniques d'intelligence artificielle avancées (réseaux de neurones profonds).

REFERENCES

[1] W. P. Anderson, P. S. Kanaroglou, et E. J. Miller, « Urban form, energy and the environment: a review of issues, evidence and policy », *Urban Stud.*, vol. 33, nº 1, p. 7-35, 1996.

[2] D. A. Aliyu *et al.*, « Optimization techniques for asthma exacerbation prediction models: A systematic literature review », *IEEE Access*, vol. 12, p. 110862-110890, 2024.

[3] Convention-Cadre des Nations Unies sur les Changements Climatiques, « Convention-Cadre des Nations Unies sur les Changements Climatiques », *Vingt Après Rio Avant-Goût Avenir*, p. 111-113, 2011.

[4] E. E. Ukpebor, E. O. Akpaja, J. E. Ukpebor, et J. I. Odiase, « Spatial and Diurnal Variations of Carbon Monoxide (CO) Pollution from Motor Vehicles in an Urban Centre », *J. Appl. Sci. Environ. Manag.*, vol. 13, nº 4, p. 15-20, 2009, doi: 10.4314/jasem.v13i4.55395.

[5] I. Jolliffe, « Principal components as a small number of interpretable variables: some examples », *Princ. Compon. Anal.*, 2002.

[6] L. van der Maaten et G. Hinton, « Visualizing Data using t-SNE », *J. Mach. Learn. Res.*, vol. 9, nº 86, p. 2579-2605, 2008.

[7] S. Zhao, B. Zhang, J. Yang, J. Zhou, et Y. Xu, « Linear discriminant analysis », *Nat. Rev. Methods Primer*, vol. 4, nº 1, p. 70, 2024.

[8] M. F. A. Z. Ansari, « Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions », *ISSN 2221-0741*, vol. 1, nº 5, p. 217-226, 2011.

[9] C. James et others, « FCM: The fuzzy c-means clustering algorithm », *Comput. Geosci.*, vol. 10, nº 2-3, p. 191-203, 1984.

[10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et others, « A density-based algorithm for discovering clusters in large spatial databases with noise », *kdd*, vol. 96, 1996.

[11] C. Shi, W. Wang, et J. Wang, « A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm », *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, nº 1, p. 31, 2021, doi: 10.1186/s13638-021-01912-w.

[12] U. Maulik et S. Bandyopadhyay, « Performance evaluation of some clustering algorithms and validity indices », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, nº 12, p. 1650-1654, 2003.

[13] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, et I. Perona, « An extensive comparative study of cluster validity indices », *Pattern Recognit.*, vol. 46, nº 1, p. 243-256, 2013.

[14] A. A. R. Fernandes *et al.*, « Comparison of cluster validity index using integrated cluster analysis with structural equation modeling the Warp-PLS approach », *J. Theor. Appl. Inf. Technol.*, vol. 99, nº 18, p. e1-e18, 2021.

[15] P. J. Rousseeuw, « Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis », *J. Comput. Appl. Math.*, vol. 20, p. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[16] J. Vertu, H. Jean, E. F. Randrianarivelo, T. O. Hanta, R. Randrianandrasanarivo, et D. Maxwell, « Regionalization of Precipitation Data in the Sofia Region », *Int. J. Adv. Res. Innov. Ideas Educ. IJARIIE*, vol. 10, nº 4, p. 3730-3739, 2024.

[17] I. BOUSRI, A. BOUCETTA, et S. SAHABI-ABED, « Régionalisation des normales annuelles des températures en Algérie par la méthode K-means », *JAMA*, vol. 6, p. 40-46, 2022.