Revue Internationale du Chercheur
ISSN: 2726-5889
Volume 7 : Numéro 1

Exploration des données douanières en RDC : Segmentation et détection d'anomalies dans les flux d'importation

Exploration of Customs Data in the DRC: Segmentation and Anomaly Detection in Import Flows

NYAMI NYATE Ruphin
Doctorant
Université de Lubumbashi / Faculté de Sciences et Technologie / Mathématiques-Informatiques
RDC

BULA KATENDI Axel
Professeur
Université de Lubumbashi / Faculté de Sciences et Technologie / Mathématiques-Informatiques
RDC

BALAULA MPATA Larry
Doctorant
Université Protestante de Lubumbashi / Faculté des Sciences Informatiques
RDC

KAMWANG-A-MUSAS Rose
Msc
Université Protestante de Lubumbashi / Faculté des Sciences Informatiques
RDC

TSHIBUABUA Franck
Msc
Université Protestante de Lubumbashi / Faculté des Sciences Informatiques
RDC

KUMAMBANGE KOMBE Francis
Assistant
Institut Supérieur de Commerce d'Ilebo
RDC

Date de soumission : 07/01/2026

Date d'acceptation : 16/02/2026

Pour citer cet article :

NYAMI NYATE. R & Al (2026) «Exploration des données douanières en RDC : Segmentation et détection d'anomalies dans les flux d'importation», Revue Internationale du chercheur «Volume 7 : Numéro 1» pp : 880-905.

Résumé

Pour passer d'un contrôle systématique à un ciblage intelligent des flux d'importation, cet article explore les données douanières afin de débusquer des profils cachés et des signaux de risque. Nous combinons l'analyse en composantes principales (ACP), la classification hiérarchique ascendante, le K-means et DBSCAN. Les flux douaniers sont complexes : très dimensionnels, déséquilibrés et sensibles aux fraudes variées, comme la sous-évaluation, la fragmentation des colis, des codes SH anormaux, des valeurs unitaires instables, des profils temporels suspects ou des comportements isolés. Les règles strictes actuelles butent sur la « malédiction de la dimensionnalité », entraînant des libérations erronées de déclarations frauduleuses, des saisies injustes et un gaspillage de temps. Nous proposons une approche innovante via une ingénierie de données ciblée. Des indicateurs multidimensionnels d'anomalies sont créés, normalisés, puis simplifiés par l'ACP pour révéler les dimensions clés. Le K-means et la classification hiérarchique segmentent les profils d'importateurs, validés par des tests de qualité et de stabilité. DBSCAN isole les cas atypiques, confirmés par analyse de sensibilité. Les résultats mettent en lumière des profils distincts et un petit nombre d'anomalies inhabituelles, idéales pour un contrôle ciblé. Cette chaîne ACP-clustering-détection d'anomalies fournit un outil clair et opérationnel pour l'analyse de risques douaniers.

Mots-clés : données douanières ; importations ; analyse en composantes principales ; clustering non supervisé ; DBSCAN ; segmentation ; détection d'anomalies ; analyse de risque.

Abstract

To transition from systematic customs control to intelligent targeting of import flows, this article explores customs data to uncover hidden profiles and risk signals. We combine Principal Component Analysis (PCA), hierarchical ascendant classification, K-means, and DBSCAN. Customs flows are complex: highly dimensional, imbalanced, and sensitive to diverse frauds, such as under-valuation, shipment fragmentation, anomalous HS codes, unstable unit values, suspicious temporal profiles, or isolated behaviors. Current strict rules falter against the "curse of dimensionality," leading to erroneous releases of fraudulent declarations, unjust seizures, and time wastage. We propose an innovative approach through targeted data engineering. Multidimensional anomaly indicators are created, normalized, and simplified via PCA to reveal key dimensions. K-means and hierarchical classification segment importer profiles, validated by quality and stability tests. DBSCAN isolates atypical cases, confirmed through sensitivity analysis. The results highlight distinct profiles and a small number of unusual anomalies, ideal for targeted control. This PCA-clustering-anomaly detection chain provides a clear and operational tool for customs risk analysis.

Keywords: customs data; imports; data mining; feature engineering; principal component analysis; unsupervised learning; DBSCAN; clustering; anomaly detection; risk analytics.

Introduction

Dans le sillage de la mondialisation, les frontières nationales ne sont plus de simples limites géographiques, mais des points névralgiques où se jouent la souveraineté économique et le développement social d'une nation. Pour la République Démocratique du Congo (RDC), ces frontières constituent les poumons commerciaux. La Direction Générale des Douanes et Accises (DGDA) déploie des bureaux comme le port sec de Kasumbalesa, le Port de Matadi, Mukambo, Lufu ou Zongo pour mobiliser les recettes de l'Etat. Des volumes de déclarations douanières sont en enregistrées chaque jour et chacune d'elles est le témoin d'une activité humaine et commerciale dont la juste perception des droits conditionne le financement des infrastructures et des services publics essentiels. Plus les années s'écoulent, ces masses de données sont considérées comme une simple archive administrative. En revanche, ces archives constituent une mine d'or douanière à explorer car, certaines importations peuvent se réaliser avec ou sans licence. Cela divise les regards : pour les juristes, c'est de la contrebande ; aux yeux des économistes cela est une hémorragie de recettes ; pour les sociologues, un cri de survie et un mécanisme vital face à la précarité des populations locales. Au regard du Code des Douanes congolais : toute importation sans licence modèle IB ayant une valeur au-delà de 2 500 USD, est considérée comme une importation irrégulière, saisissable ou passible d'amendes équivalant aux taxes éludées. Pourtant, une tolérance adoucit le sort des petits commerçants aux colis modestes, dédouanés via procédures simplifiées. De surcroît, le SEGUCE, guichet unique censé fluidifier les échanges, s'embourbe dans des inspections BIVAC complexes et une occasion pour des pratiques prohibées. D'où le recours aux pistes de brousse du « bilanga », qui consiste à fractionner une cargaison en lots infimes, portés par des individus, pour contourner licences onéreuses et bureaucratie asphyxiante. A la DGDA, deux types d'importateurs sont identifiés. D'un côté, les importateurs formels, ayant de Numéros d'Identification Fiscale (NIF) et inscrits au Registre du Commerce et du Crédit Mobilier : ils déploient des volumes réguliers, obtiennent leurs licences d'importation, et jouissent de régimes suspensifs d'entrepôts, admissions temporaires, exonérations. Ils sont considérés comme partenaires de l'État qui gonflent les recettes fiscales, néanmoins non sans risquer la sous-évaluation. De l'autre, l'informel : particuliers et micro-entrepreneurs acheminant de biens de première nécessité sous le seuil toléré, sans licence ou « prête-noms » et occasionnels, abusant ainsi de NIF génériques pour frauder via fractionnement.

Tout porte à croire que l'approche traditionnelle, souvent fondée sur des contrôles aléatoires, règles spécifiques ou une intuition empirique, est impuissante et limitée face à des schémas de fraude de plus en plus sophistiqués.

Par conséquent, la présente étude s'inscrit dans une volonté de moderniser ce contrôle par l'utilisation de l'intelligence artificielle et du Data Mining. Elle propose une démarche d'exploration des données douanières visant à caractériser les flux d'importation, à segmenter les opérateurs selon leurs comportements et à détecter des anomalies susceptibles de signaler des risques. Sous le prisme de l'apprentissage non supervisé, cette étude ne se contente pas de traiter des chiffres ; elle cherche à déceler les pépites et les comportements cachés et atypiques qui fusent la fraude. Il est avéré que l'usage l'intelligence artificielle (IA) dans les entreprises publique offre une perspective novatrice dans l'efficacité de contrôles (AMOUSSOU & BAMPOKY, 2025). Appliquée dans un secteur vital de l'économie nationale, l'IA apporte une vision claire sur la dynamique de ciblage et vigilance accrue. Dans le contexte de lutte contre les diverses formes de la fraude douanière, la segmentation apporte une réponse adéquate dans l'amélioration du ciblage de certains types de produits, la détection rapide des anomalies à savoir : le délai anormal, de valeurs unitaires inhabituelles, les fausses déclarations ou encore sur les profils extrêmes. Nonobstant la modernisation des infrastructures portée par le déploiement du système SYDONIAWorld et l'instauration du Guichet Unique Intégral du Commerce Extérieur (SEGUCE), le ciblage, le contrôle demeure statique, chronophage et dilatoire. Avouons-le : l'expérience des inspecteurs en douane reste irremplaçable. Néanmoins elle s'essouffle face à une masse de données qui s'accélère et à des stratégies d'importation de plus en plus sophistiquées. L'inspecteur en douane est appelé à combiner quatre dimensions complémentaires pour qualifier une déclaration de véridique ou fausse. Ces quatre dimensions de risque sont : (1) la dimension volumétrique et fiscale ; (2) la dimension logistique et diversité, où la complexité masque le camouflage ; (3) la dimension temporelle et dynamique ; et (4) les Signaux de Risque.

— **La dimension volumétrique et fiscale**

La dimension volumétrique et fiscale constitue un indicateur central pour l'analyse du risque douanier. Elle révèle souvent la cohérence ou la rupture entre le poids réel des marchandises et la valeur déclarée au Trésor. Lorsqu'un flux important génère une contribution fiscale faible, le signal ne relève plus d'un simple écart statistique : il suggère une probable sous-évaluation.

Cette tension entre volume et fiscalité demeure au cœur du travail des contrôleurs, qui doivent percer l'opacité logistique afin de distinguer les opérateurs conformes de ceux qui cherchent à dissimuler leur véritable assiette taxable. En mettant en évidence ces déconnexions, l'analyse proposée ne se limite pas au calcul ; elle apporte un soutien décisionnel concret pour cibler les contrôles de manière plus juste et plus efficace.

— **Dimension logistique et diversité**

La dimension logistique et la diversité des produits offrent une lecture complémentaire de la conformité. Elles révèlent le rythme des transactions et la variété des codes SH mobilisés par un opérateur. Les contrôles traditionnels peinent face aux stratégies de dissimulation consistant à morceler la cargaison en des petits colis. De plus, les importateurs diluent des marchandises fortement taxées parmi des articles anodins pour esquiver le fisc congolais. Dans ce contexte, le système d'information douanier doit dépasser la simple fonction d'archivage pour aider à distinguer un opérateur régulier d'un acteur opportuniste dont la diversité devient suspecte. L'enjeu consiste à transformer ce foisonnement logistique en indicateurs fiables. En s'appuyant sur le clustering, l'analyse met en évidence des schémas atypiques et permet de cibler les contrôles avec précision, sans pénaliser les opérateurs conformes.

– **Dimension temporelle et dynamique**

Le véritable défi douanier se joue aussi dans le temps. La dimension temporelle révèle la cadence des déclarations c'est-à-dire leur rythme, leur rapidité, leurs ruptures. Quand l'urgence s'impose, la faille apparaît. Sous pression pour accélérer les opérations, certaines anomalies passent, entre autres : fréquences irrégulières, trajets anormalement courts, mouvements trop rapides pour être crédibles. Souvent, ces signaux traduisent un contournement discret des contrôles. Le clustering transforme alors ce désordre apparent en repères clairs. Il fait ressortir les profils dont la dynamique s'écarte des cycles habituels. La brigade ne subit plus le flux : elle anticipe, cible et intervient là où le tempo lui-même devient suspect.

— **La dimension des Signaux de Risque**

La limite majeure des systèmes classiques réside dans leur incapacité à capter ce qui n'est pas déclaré. Les signaux de risque montrent, pourtant, que les données « parlent ». La fraude actuelle n'est plus spectaculaire : elle se cache dans une normalité apparente. Un prix unitaire paraît banal, jusqu'à ce qu'on le compare à la médiane du secteur ou à l'historique de l'opérateur. C'est là que l'anomalie apparaît. L'enjeu de cette étude est précisément d'aller au-delà du contrôle transactionnel.

Il s'agit de passer à une analyse comportementale, capable de repérer des écarts subtils et de transformer ces signaux en appui concret pour la décision. Aujourd'hui, derrière chaque déclaration enregistrée dans la base de données se cache une réalité humaine et commerciale complexe qui ne demande qu'à être décryptée. Plutôt que de voir ces masses de données comme une simple archive administrative, nous proposons de les traiter comme un organisme vivant. Les méthodes de Data Mining agissent ici comme un nouveau regard : en analysant simultanément les volumes, les fréquences et la stabilité des valeurs, elles permettent de passer d'une surveillance réactive à une intelligence prédictive. L'enjeu n'est plus seulement de surveiller, mais de comprendre minutieusement les comportements pour instaurer une douane plus juste, capable de protéger les recettes de l'État sans entraver l'élan des opérateurs honnêtes. C'est dans cette perspective que s'inscrit cette étude, qui cherche à apprécier l'apport de cette analyse exploratoire à l'orientation des contrôles douaniers : ***Dans quelle mesure la caractérisation multidimensionnelle des comportements d'importation permet-elle d'identifier des profils atypiques et des schémas à risque que les dispositifs de ciblage conventionnels ne détectent pas ?***

L'objectif de cette étude est de montrer comment une analyse multidimensionnelle des données douanières peut contribuer à une meilleure compréhension des comportements d'importation et à l'identification précoce de signaux de risque. Pour ce faire, nous mettons en place une chaîne analytique combinant l'ingénierie de variables, l'analyse en composantes principales et des algorithmes d'apprentissage non supervisé, complétés par une détection d'anomalies basée sur la densité. Cette approche permet d'extraire des profils homogènes d'importateurs et de repérer des cas atypiques susceptibles d'orienter des vérifications ciblées.

Du point de vue méthodologique et au-delà de la démonstration empirique, cette étude s'inscrit dans une démarche quantitative fondée sur l'analyse de données d'importation douanières. La démarche adoptée vise à illustrer comment des outils de data science peuvent être intégrés, de manière pragmatique et interprétable, dans un dispositif opérationnel d'analyse de risque douanière. Le recours aux techniques de Machine Learning, principalement non supervisées, permet de segmenter et d'isoler les profils atypiques.

La suite l'étude est organisée en trois sections complémentaires. La première présente l'état de la question et situe cette recherche dans les travaux existants sur l'usage du data mining pour l'analyse des flux douaniers et la gestion du risque. La deuxième décrit la méthodologie retenue, en détaillant la construction des variables, l'application de l'analyse en composantes principales, puis l'utilisation des algorithmes de clustering et de détection d'anomalies.

La troisième section expose les principaux résultats empiriques, en mettant en évidence les profils d'importateurs identifiés et les cas atypiques qui émergent de l'analyse, ainsi que leurs implications potentielles pour le ciblage des contrôles.

1. Etat de la question

La littérature récente montre un constat clair : les méthodes traditionnelles de sélection et de contrôle ne suffisent plus. Les échanges se multiplient. Ils deviennent plus rapides, plus complexes. Jusqu'où ces systèmes peuvent-ils suivre sans se saturer ? C'est cette question qui pousse les administrations à se tourner vers des approches fondées sur les données et l'intelligence artificielle.

Plusieurs travaux soulignent cette transition (Hamid & Rahman, 2025; Vijayakumar, 2025). Les techniques de data analytics et de machine Learning renforcent la capacité à anticiper les risques, tout en maintenant la fluidité du commerce légitime. Là où l'on utilisait autrefois des règles fixes et des systèmes experts rigides, on voit désormais apparaître des modèles capables d'apprendre, de s'adapter et de détecter des comportements difficiles à formaliser.

Cependant, l'état de l'art rappelle aussi une limite : aucun outil n'est magique. L'efficacité de ces approches dépend de la qualité des données, de leur gouvernance et de la capacité des équipes à interpréter les résultats. En d'autres termes, l'IA n'efface pas la complexité douanière ; elle offre surtout de nouveaux moyens de la comprendre.

1.1. Le paradigme de la détection d'anomalies de valeur

La détection d'anomalies occupe une place centrale en science des données. Elle vise à repérer des observations dont le comportement s'écarte significativement de la structure générale des données. Dans des environnements à forte volumétrie comme les bases d'importations douanières, elle permet d'identifier des cas rares, inattendus ou potentiellement à risque. La littérature rappelle notamment que la manipulation de la valeur déclarée demeure un vecteur majeur de fraude, l'écart par rapport aux prix médians constituant un indicateur robuste (Chalendard et al., 2023). Toutefois, ces outils n'ont pas vocation à remplacer l'expertise humaine : ils agissent comme un révélateur d'incohérences devenues invisibles face à la masse d'informations. Dans des postes frontaliers très fréquentés, la détection d'anomalies représente ainsi une réponse pragmatique aux limites des systèmes fondés sur des règles statiques, souvent contournées. Elle permet de signaler des incohérences telles que des valeurs sous-estimées récurrentes, des volumes atypiques ou des envois artificiellement fragmentés.

Sur le plan méthodologique, la littérature distingue plusieurs paradigmes de détection d'outliers notamment : statistiques, distance, densité et modèles. L'efficacité de chacune dépend étroitement de la nature des données (dimensionalité, bruit, déséquilibre) (Domingues et al., 2018; Thomas & E. Judith, 2019). Les synthèses récentes insistent sur la nécessité d'articuler ces approches avec des techniques de réduction de dimension pour traiter des jeux complexes et hétérogènes (Hu et al., 2023) (Wang et al., 2019). Dans une forêt dense de données bruitées, les méthodes basées sur la densité, telles que DBSCAN ou LOF, deviennent utiles. Elles identifient des points isolés dans des régions de faible densité, priorisant ainsi les observations selon leur degré d'isolement (Breunig et al., 2000). Des comparaisons empiriques récentes montrent toutefois qu'aucun algorithme ne domine systématiquement : les performances varient selon la structure des données et le compromis recherché entre rappel, faux positifs et robustesse (Fuhnwi et al., 2023). Dans ce contexte, la démarche adoptée vise à passer d'un contrôle généralisé à une priorisation basée sur des signaux d'anomalie, afin de cibler les profils les plus atypiques tout en préservant la fluidité des échanges.

1.2. La puissance du clustering non supervisé

Le clustering non supervisé constitue un pilier du data mining lorsqu'il s'agit d'explorer des données non étiquetées et de faire émerger des structures latentes. En l'absence de classes prédéfinies, ces techniques regroupent les observations selon leurs similarités et offrent ainsi un outil pertinent pour analyser les flux douaniers dépourvus d'indicateurs de risque explicites. Des travaux tels que ceux de (Wei et al., 2024)Jallepalli et al. (2020) montrent que des algorithmes de clustering permettent de dégager des signatures comportementales. Parmi ces techniques, on recense plusieurs familles de méthodes à savoir : centroïdes (K-means), hiérarchiques, basées sur la densité (DBSCAN) ou appuyées sur l'apprentissage profond. Leur pertinence varie selon la structure des données, le bruit et la dimensionnalité.

Plus récemment, les avancées en deep clustering ont élargi le champ des applications. En combinant représentations apprises par réseaux neuronaux et mécanismes de regroupement, ces approches capturent des structures non linéaires et améliorent la séparation entre groupes dans des données complexes ou hétérogènes (Wei et al., 2024). Des cadres hybrides optimisent simultanément la transformation des données et la formation des clusters, rendant les partitions plus robustes au bruit ou aux formes non sphériques (Yin et al., 2023), tandis que de nouveaux travaux intègrent auto-encodeurs, contraintes d'interprétabilité et apprentissage profond (Dewoprabowo et al., 2025). Dans l'ensemble, la littérature converge vers l'idée que la valeur

du clustering réside autant dans la qualité de la séparation que dans la capacité à produire des partitions explicables et directement mobilisables pour l'aide à la décision.

1.2.1. Analyse multidimensionnelle et clustering non supervisés

L'essor des bases de données douanières massives a conduit à recourir à des techniques de réduction de dimension pour synthétiser une information abondante et corrélée. L'analyse en composantes principales (ACP) occupe une place prédominante. Elle condense plusieurs variables en quelques axes interprétables tout en limitant la perte d'information. Les travaux fondateurs (Jolliffe & Cadima, 2016) montrent qu'elle facilite la mise au jour de structures latentes et soutient l'analyse exploratoire comme prédictive. Plusieurs applications récentes confirment ce rôle consistant à la fois de réduire la redondance, de préparer des pipelines d'apprentissage (Khojasteh Aliabadi et al., 2022) et de construire des indicateurs synthétiques pour comparer les performances institutionnelles (Fuhnwi et al., 2023; Zamora Torres & Navarro Chávez, 2015, 2015). Dans ce cadre, l'ACP n'est plus seulement un outil de synthèse, mais une étape structurante pour des dispositifs d'analyse plus complexes. Associée à des techniques d'apprentissage, elle aide à repérer des comportements atypiques lorsque les écarts aux composantes dominantes signalent des anomalies potentielles (Crépey et al., 2022).

Notre étude s'inscrit dans cette logique : l'ACP est utilisée pour réduire la redondance des variables d'importation, puis ses composantes alimentent une démarche non supervisée visant à segmenter les comportements et à détecter des profils atypiques exploitables en amont des dispositifs de gestion du risque. Enfin, les travaux récents consacrés au clustering soulignent l'intérêt de combiner plusieurs familles d'algorithmes. La CAH (Ward) éclaire la structure emboîtée des populations ; K-means offre des partitions rapides et interprétables ; DBSCAN identifie des densités irrégulières et isole le bruit ; des variantes comme HDBSCAN renforcent la détection d'anomalies. Les comparaisons montrent qu'aucune méthode n'est optimale en toutes circonstances et que la stabilité dépend du choix des métriques et paramètres. Dans cette perspective, notre étude adopte une architecture hybride : ACP, CAH, K-means et DBSCAN afin de capturer des structures complémentaires et de produire des signatures de risque robustes dans un contexte douanier complexe.

1.3. La complexité et la diversité comme vecteurs de risque

Les travaux récents confirment que la structure du panier d'importation constitue un signal central dans l'analyse du risque. Vanhoeyveld et al. (2020) montrent que la diversité des codes SH peut accompagner des tentatives de dissimulation de produits fortement taxés. Dans le

même esprit, plusieurs études soulignent l'intérêt d'approches non supervisées : le clustering permet de faire émerger des régularités, d'identifier des regroupements atypiques et de détecter des comportements suspects dans des bases massives (Choi, 2019; Molina Narváez et al., 2024a; Vanhoeyveld et al., 2020, 2020). Par ailleurs, la littérature converge sur l'importance de passer d'un contrôle systématique à une sélectivité fondée sur l'analyse de risque. Geourjon et Laporte (2004) montrent qu'une stratégie de ciblage bien conçue peut améliorer l'efficacité des contrôles sans compromettre les recettes. Lorsque des labels fiables sont disponibles, plusieurs auteurs recourent à des modèles supervisés afin d'optimiser la détection d'événements rares ou coûteux (Nelson, 2020 ; Zhou, 2019), en intégrant parfois le coût différencié des erreurs dans le processus décisionnel (Geourjon & Laporte, 2004; Nelson, 2020).

Notre contribution se distingue en proposant un cadre non supervisé intégré combinant ACP, CAH, K-Means et DBSCAN. Cette architecture permet d'extraire des « signatures de risque » à partir des données d'importation, sans hypothèse préalable sur la nature des anomalies. Elle constitue un outil d'exploration et de structuration des flux susceptible d'alimenter, en amont, des dispositifs de ciblage et des modèles prédictifs ultérieurs, notamment dans des environnements opérationnels complexes.

2. Méthodologie

Notre approche repose sur une chaîne analytique intégrée. D'abord, les indicateurs métier sont sélectionnés et les données sont normalisées. Ensuite, l'ACP réduit la dimension et prépare la segmentation. K-means identifie des groupes homogènes, tandis que la CAH vérifie la cohérence des partitions. Enfin, DBSCAN repère les anomalies à partir des densités locales.

Cette démarche permet de comprendre la structure des flux d'importation. Elle fait émerger des profils récurrents et met en évidence des cas atypiques susceptibles de signaler un risque.

Les résultats attendus sont doubles : mieux segmenter les comportements et détecter plus tôt les anomalies. L'approche fournit ainsi des outils opérationnels pour prioriser les contrôles et soutenir une gestion du risque plus ciblée et plus équitable.

2.1. Collecte de données

La collecte systématique des données douanières est essentielle à l'examen de profils et de risques en douane. Les données utilisées dans cette étude proviennent du système d'information douanier chargé de l'enregistrement des déclarations d'importation. Elles couvrent une période allant de Novembre 2014 au Décembre 20200 et correspondent aux opérations réalisées dans 17 bureaux. Le jeu de données final comprend :

— 38764 observations (déclarations ou opérateurs selon l'agrégation),
— 100 variables décrivant les caractéristiques administratives, logistiques, fiscales et économiques des importations.

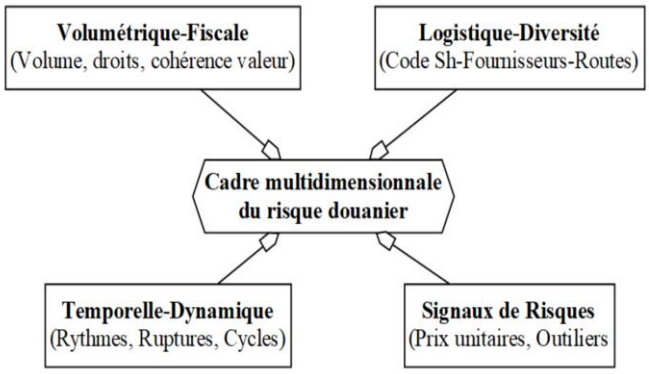
2.2.Prétraitement des données

Après contrôles de cohérence et filtrage des doublons, seules les observations complètes et exploitables ont été conservées pour l'analyse. Les variables ont été sélectionnées en fonction de leur pertinence pour l'analyse de risque douanière. Elles ont ensuite été transformées sous forme d'indicateurs comportementaux. Le pipeline proposé suit une logique suivante :

— Chargement de données : harmoniser les noms de colonnes, Conversion des dates en format datetime, conversion des montants et poids en numérique, suppression des champs purement administratifs, gestion des valeurs manquantes, agrégations logiques par déclaration et par importateur

Notre approche consiste à agréger les données selon quatre dimensions critiques :

Figure 1—Schéma de risques douaniers



Source : Auteurs

a) Dimension Volumétrique et Fiscale : les variables retenues dans cette catégorie sont :

— CIF_TOTAL & CIF_MOYEN : Représentent l'assiette fiscale globale et la taille moyenne des transactions.
— POIDS_TOTAL : Indicateur de volume physique, essentiel pour vérifier la cohérence avec la valeur déclarée.
— DROITS_PAYES_TOTAL : Contribution réelle aux recettes de l'État.

— DROITS_PAR_CIF_MOY : Ratio de pression fiscale. Une valeur anormalement basse par rapport à la moyenne du secteur indique une suspicion de sous-évaluation ou de détournement d'exonération.

b) Dimension Logistique et Diversité : cette dimension apporte un bon nombre de connaissances et d'indices notamment la logistique (Fréquence des déclarations + Mode de transport) et la diversité (Nombre de chapitres SH par déclaration).

Dans cette dimension, nous retenons donc les caractéristiques métiers suivants :

— NB_DECLARATIONS : Mesure de l'intensité de l'activité de l'importateur.
— DELAI_MOYEN : Temps moyen de traitement/dédouanement, pouvant révéler des facilitations ou des goulots d'étranglement.
— DIVERSITE_HS & DIVERSITE_PAYS **:** Nombre de types de marchandises (codes SH) et de pays d'origine. Une trop grande diversité peut indiquer un profil de commissionnaire en douane, tandis qu'une spécialisation outrancière facilite la fraude sur l'espèce.
— PART_TOP_HS : Degré de concentration sur un produit phare.

c) **Dimension Temporelle et Dynamique :** les caractéristiques métiers retenus sont :

— FREQ_MENSUELLE : Régularité de l'opérateur sur l'année.
— SAISONNALITE : Coefficient de variation (std/mean). Une forte saisonnalité peut indiquer des importations d'opportunité, souvent moins bien suivies par l'administration.
— TENDANCE_CIF : Direction de l'évolution des valeurs déclarées (croissance ou décroissance suspecte du volume d'affaires).

Afin de corriger l'asymétrie prononcée des données douanières et de stabiliser la variance, des transformations logarithmiques ont été appliquées à certaines variables. De plus, des variables de structure (ratios fiscaux et pondéraux) ont été créées pour isoler les indicateurs de performance douanière de l'effet de taille des opérations, permettant ainsi une détection plus fine des anomalies de comportement.

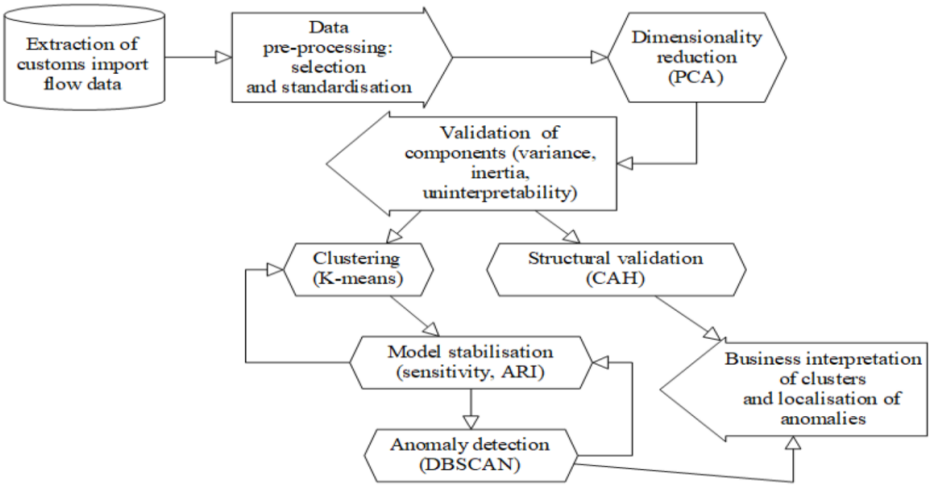
2.3. Logique générale de l'analyse

L'étude adopte une approche quantitative et exploratoire. Elle exploite des données réelles d'importation. Faute d'étiquettes fiables de fraude, nous utilisons des méthodes non supervisées. Elles permettent d'identifier des structures latentes et des comportements

atypiques dans un volume important de transactions. La démarche est progressive : simplifier, segmenter, puis détecter. L'objectif est d'appuyer le ciblage et l'aide à la décision.

Notre démarche s'appuie sur quatre outils qui travaillent ensemble plutôt qu'isolément. L'ACP nous permet d'abord de résumer les 16 variables en quelques composantes clés. *Les composantes issues de l'ACP* alimentent ensuite le K-means pour regrouper les opérateurs. La classification hiérarchique vient confirmer cette architecture en révélant la façon dont les profils s'organisent entre eux. Enfin, DBSCAN joue le rôle de détecteur d'alertes : en ajustant ses paramètres de densité, il met en évidence les opérateurs qui ne ressemblent vraiment à aucun autre et qui méritent une analyse plus attentive. Nous avons combiné la méthode du coude (inertie intra-classe) et le score de silhouette pour k compris entre 1 et 10. La solution retenue correspond au meilleur compromis entre cohérence interne et lisibilité économique des profils obtenus. Pour DBSCAN, les paramètres ont été calibrés à partir de la courbe k-distance afin d'identifier le seuil critique de densité. Les heuristiques usuelles (min_samples ≈ ln(n) et ≈ 2d) ont guidé une plage de tests, complétée par des analyses de sensibilité pour vérifier la stabilité des anomalies. Le Local Outlier Factor confirme que les profils atypiques se situaient bien dans des zones faiblement denses.

Figure 2—Démarche méthodologique



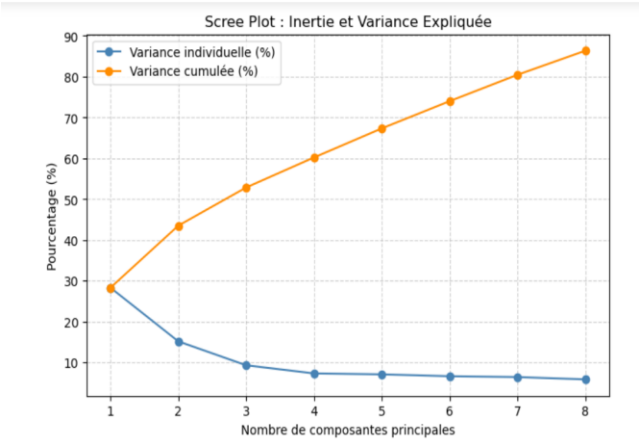
Source : Auteurs

3. Résultats

Quand on analyse les résultats de cette comparaison entre l'ACP, la Classification Ascendante Hiérarchique (CAH), K-means et DBSCAN, des motifs vraiment intéressants ressortent des

données douanières. L'ACP a montré que 74,02 % de variance totale est expliquée par les six premières composantes principales. Ceci confirmant la présence d'une structure multidimensionnelle comme illustré dans le Figure 3 suivante.

Figure 3—Variance expliquée par les composantes

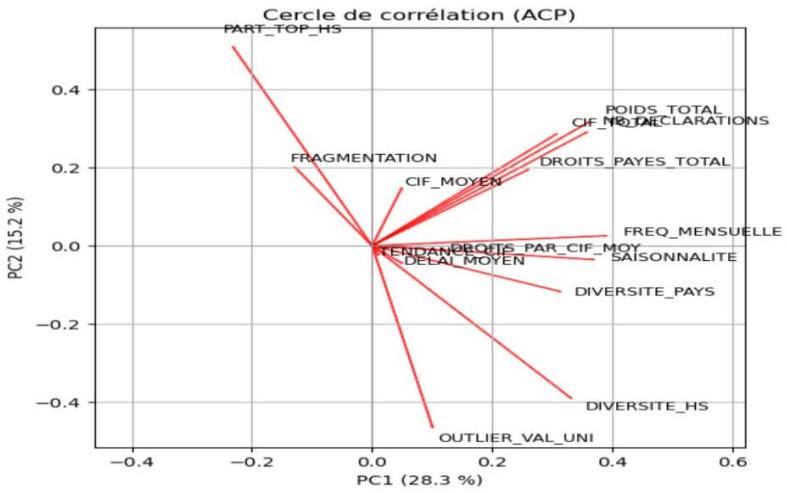


Source : Auteurs

Les deux premières composantes expliquent respectivement 28.34 % et 15.19 % de la variance expliquée. Cela indique une faible dominance des directions principales et nécessite l'utilisation d'un nombre plus important d'axes pour une représentation fiable.

Dans cette continuité, la Figure 4 suivante illustre le cercle de corrélation obtenue de notre ACP.

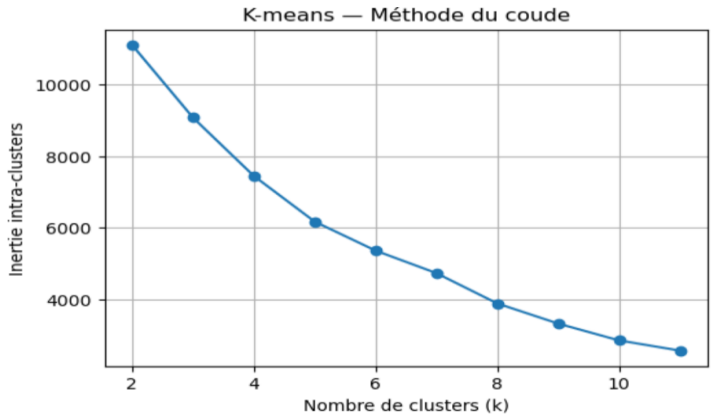
Figure 4—Cercle des corrélations



Source : Auteurs

Sur cette base, les composantes retenues (≈80–90 % de variance) alimentent un K-means plus stable. Le coude d'inertie et le score de silhouette convergent vers 4 à 6 clusters, cohérents avec les profils métier identifiés.

Figure 5——Méthode de Coude (Elblow Method)



Source : Auteurs

Sur le graphique (Figure 5), un coude net émerge vers k=4 k=4 ou 5. Justification : de k=2 à 4, l'inertie plonge de ≈11 000 à ≈7 500, capturant les grandes structures ; au-delà de k=5, la décroissance s'essouffle en pente linéaire, signalant un risque de sur-segmentation d'aires déjà homogènes. Bien que mathématiquement le coude soit visible à 4, le choix de k=5 est le plus pertinent stratégiquement pour la douane car il permet d'isoler des comportements métiers spécifiques comme illustré dans le Tableau—1 suivant :

Tableau 1—Description de cluster K-means (identité métier – unité de segementation)

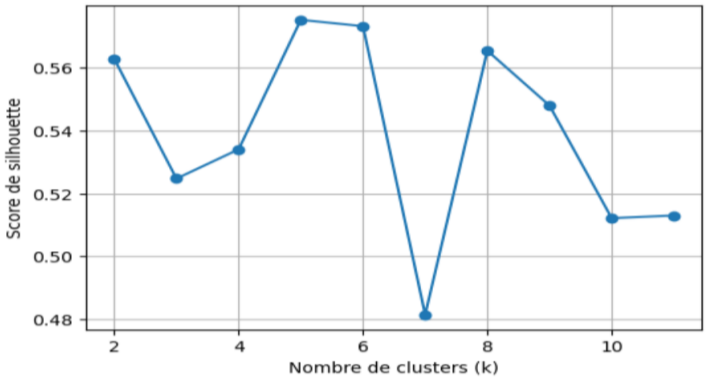
Cluster	Identité Métier Identifiée	Utilité de la Segmentation
Cluster 0	Commerce Transfrontalier	Flux de masse avec mixité formel/informel (sans NIF, sans licence).
Cluster 1	Moteur Fiscal	Importateurs formels standards (Flux licite).
Cluster 2	Risque Technique	Anomalies prioritaires (Suites suspensives, sans NIF).
Cluster 3	Cas Exceptionnels	Déclarations ultra-atypiques ou erreurs de saisie.
Cluster 4	Secteur Minier	Flux à haut volume et fortes exonérations.

Source : Auteurs

L'application de la méthode du coude révèle une inflexion majeure de l'inertie intra-classe pour k=5. Ce paramètre constitue l'équilibre optimal entre la robustesse statistique (minimisation de la variance interne) et la pertinence opérationnelle. Il permet de distinguer les cinq grandes catégories d'opérateurs économiques de la région (Miniers, Commerçants formels, Informels, Transitaires) tout en isolant un cinquième groupe d'individus atypiques pour le ciblage des contrôles.

Le score de silhouette mesure à quel point un objet est semblable à son propre cluster par rapport aux autres clusters. Il varie de -1 à 1, où un score élevé indique que les clusters sont bien isolés et cohérents. Ainsi, pour valider la méthode de Coude, nous montrons le score de silhouette.

Figure 6—Score de silhouette de cluster K-Means



Source : Auteurs

Les scores de silhouette valident notre découpage : si k=2 on a donc une séparation plus tranchée mathématiquement, k=5 l'emporte sur le terrain avec une cohésion stable (~0.57), assez fine pour pister les risques clés, fraude sur valeurs, détournement d'exonérations, fragmentation repérée par l'ACP.

Tableau 2—Niveau de Risque / cluster

Niveau de Risque	Cluster	Cible Prioritaire	Action Recommandée
Critique	Cluster 2	Valeur et NIF	Contrôle physique et audit de valeur
Élevé	Cluster 4	Exonérations	Audit a posteriori (conformité Code Minier)
Modéré	Cluster 0	Quantités	Scanner et contrôles aléatoires
Faible	Cluster 1	Conformité standard	Facilitation et Circuit Vert

Source : Auteurs

— **Résultat de clustering avec K-means**

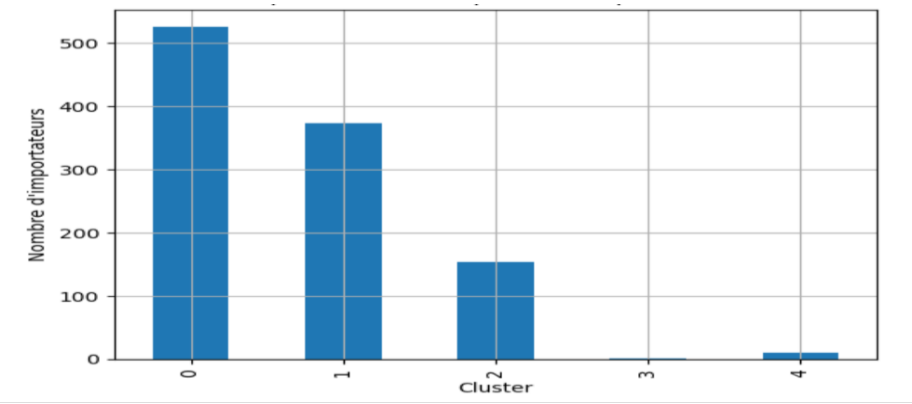
Pour évaluer la stabilité de la partition, nous appliquons K-means plusieurs fois, puis nous mesurons le taux d'accord moyen entre partitions.

Tableau 3—Validation globale du clustering

Indicateur	Ce qu'il mesure	Résultat obtenu	Interprétation
WCSS (méthode du coude)	Compacité intra-cluster	Coude observé autour de k = 5	Gain marginal au-delà de 5 → évite la sur-segmentation
Silhouette moyenne	Séparation entre clusters	Maximum local autour de k = 5	Bon compromis cohésion / séparation
ARI (stabilité)	Reproductibilité entre exécutions	0,93 (σ = 0,9)	Partition très stable, peu sensible aux initialisations

Source : Auteurs

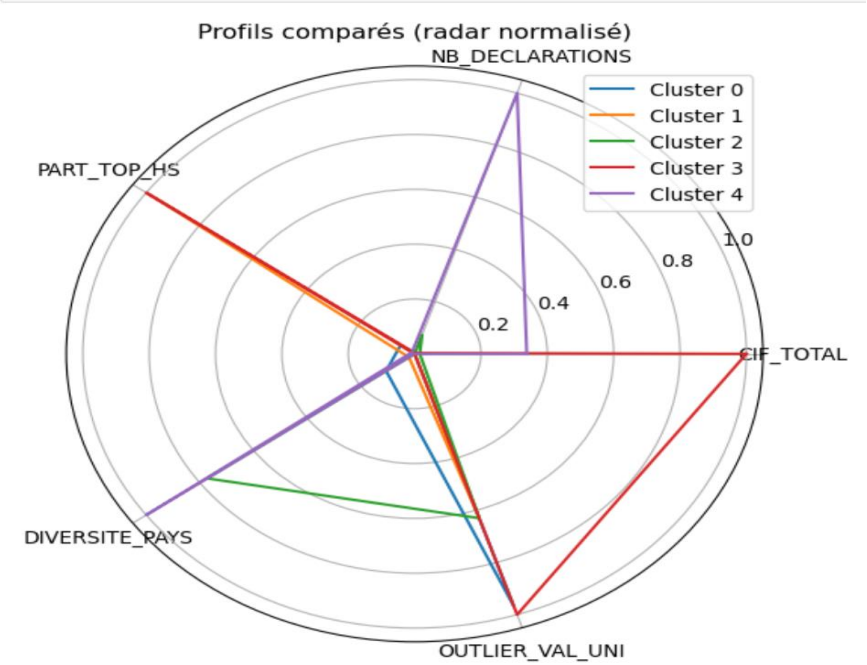
Figure 7—Répartition des importateurs par cluster



Source : Auteurs

La distribution des importateurs par cluster montre que deux profils dominent largement. Les clusters 0 et 1 regroupent à eux seuls l'essentiel des opérateurs, correspondant aux petits importateurs spécialisés et aux généralistes multiproduits. À l'opposé, les clusters 3 et 4 représentent des profils rares mais potentiellement stratégiques : leur faible effectif suggère soit des activités de niche, soit des opérateurs à forte intensité. Cette asymétrie justifie une approche différenciée des contrôles. Pour comparer la signature comportementale des clusters, nous illustrons dans la Figure 7 le radar normalisé.

Figure 8—Radar normalisé

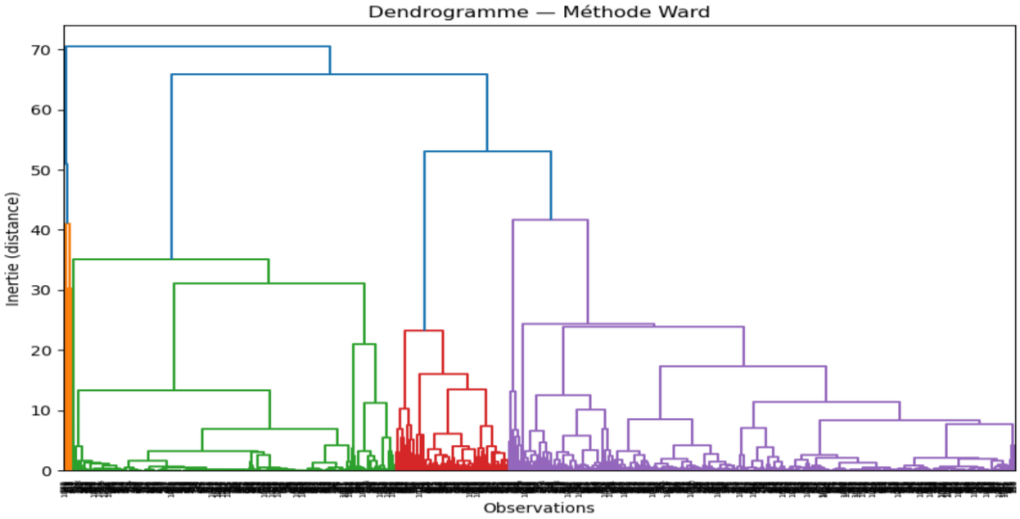


Le radar normalisé (0 à 1) permet de lire, d'un coup d'œil, la signature de chaque cluster.

— **Résultat de clustering avec CAH**

Sur cette base, nous appliquons ensuite une CAH afin de dégager les clusters. Le dendrogramme issu de la CAH met en évidence une structure fortement contrastée des distances inter-individuelles. Le résultat de cluster est illustré dans le dendrogramme de la Figure 8 suivante.

Figure 9—Dendrogramme de classification des données douanières



Source : Auteurs

Pour valider les clusters CAH face à ceux de K-Means, une sainte alliance en validation croisée est nécessaire. Si les deux méthodes disent la même chose, leurs clusters se chevauchent avec harmonie.

Tableau 4–Les matrices de correspondance entre la CAH et K-means

CLUSTER_PROFIL	0	1	2	3	4
CLUSTER_CAH					
1	0.000	0.000	0.000	0.0	1.0
2	0.000	0.000	0.000	1.0	0.0
3	0.006	0.000	0.994	0.0	0.0
4	0.079	0.921	0.000	0.0	0.0
5	0.920	0.051	0.028	0.0	0.0

Source : Auteurs

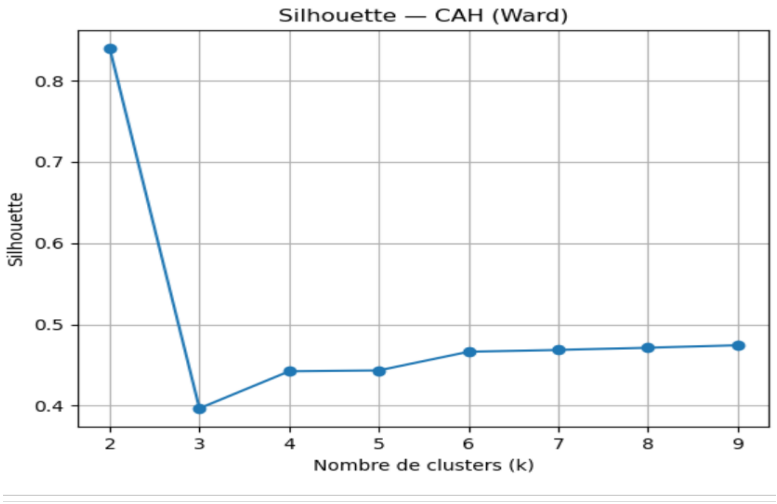
Le tableau croisé confirme une belle harmonie CAH-K-Means :

— CAH 1 : 100% en K-Means 4 – duo parfait pour ces géants stables.
— CAH 2 : 100% en K-Means 3 – alignement total.
— CAH 3 : 99,4% en K-Means 2 – quasi fusion.
— CAH 4 : 92% en K-Means 1 (quelques échappées vers 0) – cohérent, mais aux contours un peu flous.

— CAH 5 : 92% en K-Means 0, aspirant aussi 1 et 2 – un sac plus hétéro, frôlant les frontières.

Avec le score de silhouette illustré dans le Figure 10 suivante :

Figure 10—Score de silhouette CAH (Ward)

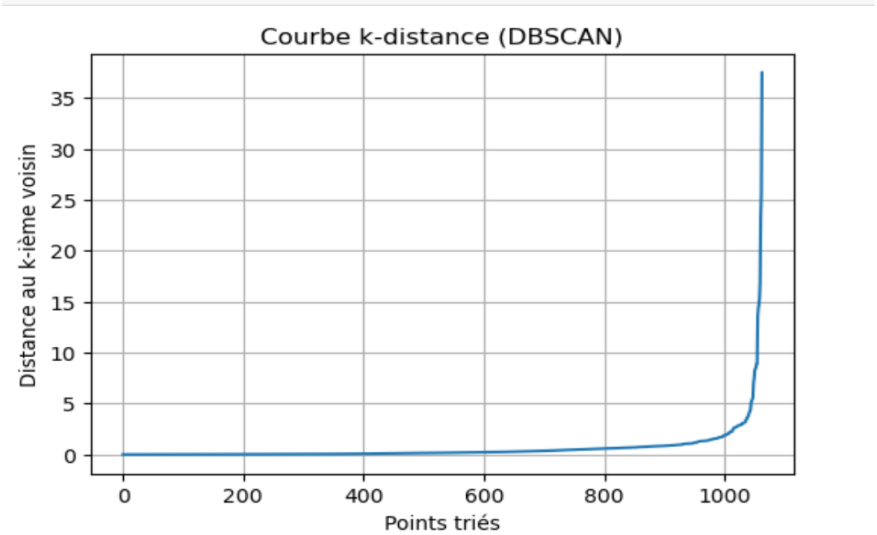


Source : Auteurs

— **Résultat avec le DBSCAN**

Comme pour K-Means et CAH, DBSCAN s'appuie sur les composantes ACP – ces perles non corrélées qui stabilisent le tout. On retient 8 d'entre elles, capturant 90% de la variance expliquée : un choix circonspect pour plonger en apnée dans les densités.

Figure 11—Score de Silhouette montrant les k-distances (DBSCAN)



Source : Auteurs

Résultat pragmatique : une plage [7-16] s'impose comme sweet spot théorique. Le résultat obtenu est illustré dans le Tableau 5 suivant :

Tableau 5—Résultat de clusters (DBSCAN)

Label DBSCAN	Interprétation	Effectif
0	Points normaux (dans une zone dense)	1045
−1	Anomalies / outliers détectés	18

Source : Auteurs

L'exécution de DBSCAN, aboutit à l'identification de 18 importateurs considérés comme anomalies structurelles, soit environ 2 % de la population. Ces opérateurs présentent des comportements isolés par rapport aux autres (montants extrêmement élevés, fréquence inhabituelle, ou valeurs unitaires atypiques). La majorité (98 %) des importateurs se situent dans des zones denses, confirmant la stabilité globale des profils identifiés. Un tableau croisé avec K-Means lève le voile : la moitié squatte le cluster 4, ces mastodontes économiques.

Tableau 6*—Tableau de localisation des anomalies*

Cluster	**% Anomalies**
0	0.4%
1	1.1%
2	2.0%
3	100%
4	100%

Les outliers DBSCAN affluent dans les clusters high-stakes (50% en 4 : volumes massifs, faible diversité). Ceci entraine des actions de ciblage douanier suivantes : **3/4** : Suivi intense, audits prix/qualité, **1** : Contrôles sporadiques (marginal). **0/2** : Fluidité, aléatoire sauf alertes.

— **Carte décisionnelle finale (profil → risque → action douanière)**

L'ACP, K-Means, CAH et DBSCAN fusionnent en une carte décisionnelle affûtée pour les contrôles douaniers. Le risque budgétaire se cristallise chez une élite – clusters 3 et 4, flaggés atypiques par DBSCAN. À l'inverse, 0 et 1 respirent la régularité, appelant facilitation. Cette priorisation optimise ressources et précision du ciblage.

Tableau 7—Carte décisionnelle finale

Cluster	Profil métier (dominant)	Caractéristiques clés	Risque principal	Action douanière recommandée
0	Petits importateurs opportunistes	Faible CIF, peu déclarations, produits variés	Faible impact budgétaire	*Facilitation + contrôles aléatoires*
1	Importateurs réguliers et diversifiés	Volume stable, plusieurs produits/pays, déclarations fréquentes	Risque modéré (erreurs, sous-déclaration ponctuelle)	*Contrôles ciblés selon indicateurs de risque*
2	Spécialisés par produit (quincaillerie / pièces / matériaux)	CIF moyen, forte dépendance à un HS unique	Risque de manipulation de valeur	*Contrôles thématiques (classement, valeur, origine)*
3	Opérateurs atypiques à faible volume	Très peu d'opérateurs mais structure « hors-normes »	Signal DBSCAN = profil anormal	*Audit documentaire systématique*
4	Grands importateurs dominants (carburants / gros équipements)	Très fort CIF, peu d'acteurs, influence budgétaire élevée	Risque élevé (fraude organisée / optimisation illégale)	*Contrôles renforcés + suivi permanent*

Source : Auteurs

Le tableau—8 synthétise la logique opérationnelle issue de l'ACP + K-means + CAH + DBSCAN :

Tableau 8—Tableau comparatif de résultats

Méthode	Paramètres	Variance Cumulée	Dim1 – Dim2 (%)	Autres	Clusters	Silhouette	Outliers
ACP	Dim = 6	90%	28.6% / 15.2%	Utilisée pour réduction dimensionnelle			
K-means	k = 5			Clusters sphériques	5	0.57	
CAH	Ward, 5 classes			Basée sur la distance	5	0.87	
DBSCAN	eps = 3.5, min_samples = 7			Méthode par densité	2	0.56	18

Source : Auteurs

4. Discussion

Cette étude exploite les données d'importation douanière afin d'analyser les comportements des opérateurs économiques et d'identifier des structures latentes dans les flux. L'approche combine réduction de dimension, segmentation et détection d'anomalies. Les résultats de l'analyse en composantes principales montrent que l'information n'est pas concentrée sur un facteur unique : la première composante n'explique que 28 % de la variance et six composantes sont nécessaires pour atteindre 90 %. Cette dispersion confirme que le risque et les comportements atypiques relèvent de dimensions multiples, où interagissent volumes financiers, fragmentation déclarative, diversité des fournisseurs et instabilité des valeurs unitaires. Elle justifie ainsi le recours à des méthodes analytiques véritablement multidimensionnelles. Le partitionnement par K-means révèle une structure stable et interprétable. Le critère du coude situe un point d'inflexion à k = 5, tandis que le score de silhouette (≈ 0,57) indique un bon équilibre entre cohésion interne et séparation entre groupes. Les segments identifiés correspondent à des profils métier distincts, des petits opérateurs irréguliers aux acteurs à forte intensité économique, en passant par des cas atypiques. La stabilité du partitionnement (ARI > 0,9) suggère que ces regroupements traduisent des régularités structurelles. La classification hiérarchique (méthode de Ward) apporte une vue complémentaire : cinq groupes principaux se déclinent en sous-profils homogènes, et certains opérateurs ne fusionnent qu'à des niveaux élevés d'inertie, signe de comportements singuliers. Cette perspective hiérarchique facilite la définition de seuils gradués d'intervention et l'articulation entre facilitation et contrôle. Parallèlement, DBSCAN isole un noyau dense d'activités régulières et détecte 18 anomalies robustes aux variations de paramètres. Plusieurs ne concernent pas des volumes élevés, mais des stratégies discrètes, comme la fragmentation des déclarations, la volatilité des prix ou une spécialisation excessive. Ce résultat illustre l'intérêt d'une détection fondée sur plusieurs indicateurs, plutôt que sur un signal unique. Dans l'ensemble, la convergence des analyses issues de l'ACP, du clustering et de DBSCAN met en évidence un cadre analytique cohérent pour la gestion du risque douanier : il permet de segmenter les opérateurs, de prioriser les investigations et de cibler les contrôles sur les profils à probabilité accrue d'irrégularité, tout en préservant la fluidité des flux conformes. Les travaux récents confirment cette orientation. Molina Narváez et al., (2024b) montrent, à partir de données réelles, que k-means permet d'identifier des anomalies temporelles et structurelles validées ensuite par expertise métier. Le projet européen PROFILE (European Commission, 2020) démontre qu'une intégration de données externes et de mécanismes de rétroaction

améliore la capacité prédictive des systèmes de ciblage. De même, Cao & Zheng, (2024) décrivent l'usage de modèles supervisés (CatBoost, XGBoost) en contexte douanier pour produire des scores de risque quasi temps réel et optimiser les inspections.

Ces références situent la portée de la présente étude. Malgré une segmentation robuste (ACP, K-means, CAH) et une détection d'anomalies stable, la principale limite tient à l'absence de liens directs avec des indicateurs de fraude validés.

Sans données issues de contrôles, de saisies ou d'audits, il n'est pas possible d'estimer précisément les taux de détection et les faux positifs/négatifs.

Par ailleurs, l'analyse repose uniquement sur des données déclaratives internes, sans enrichissement par des sources externes susceptibles d'améliorer la détection des signaux faibles. Enfin, les méthodes mobilisées demeurent en grande partie linéaires ou basées sur des distances euclidiennes, ce qui peut limiter la capture de schémas complexes ou d'interactions non linéaires.

Conclusion

L'objectif de cet article était de montrer qu'une imbrication d'ACP, de techniques de clustering et d'algorithmes de détection d'anomalies peut constituer un cadre cohérent et exploitable de ciblage du risque douanier, fondé sur des indicateurs véritablement multidimensionnels.

L'analyse conjointe de l'ACP, du clustering et de DBSCAN confirme que les comportements atypiques résultent d'interactions entre plusieurs dimensions : volumes financiers, fragmentation déclarative, diversité des partenaires et instabilité des valeurs unitaires. L'identification de structures stables et interprétables permet de segmenter les opérateurs, de hiérarchiser les priorités d'intervention et d'orienter les contrôles vers les profils présentant une probabilité accrue d'irrégularité, tout en préservant la fluidité des flux conformes. Ces résultats soulignent la pertinence d'outils de fouille de données multidimensionnels pour la gestion moderne du risque douanier. Ceci atteste que la valeur ajoutée tient moins à chaque algorithme pris isolément qu'à leur articulation au sein d'un processus transparent, reproductible et explicable.

L'approche proposée constitue un socle analytique crédible pour soutenir des politiques de contrôle plus ciblées, plus transparentes et plus efficientes. Elle ouvre la voie à une intégration progressive de solutions data-driven dans les environnements opérationnels, contribuant à mieux concilier facilitation des échanges et protection durable des recettes publiques.

Les travaux futurs se concentreront sur l'intégration d'étiquettes basées sur l'audit pour évaluer les performances prédictives, enrichir les modèles avec des sources de données externes et explorer des techniques avancées telles que l'analyse graphique et le regroupement profond. Un déploiement progressif dans des environnements opérationnels sera également évalué afin de mesurer l'impact, l'interprétabilité et l'acceptation par les utilisateurs.

Références

AMOUSSOU, C., & BAMPOKY, B. (2025). Intelligence artificielle et performance opérationnelle des entreprises publiques au Bénin. *Revue Belge*, *11*(130), 201-237.

Cao, Q., & Zheng, X. (2024). Application of Artificial Intelligence Technology in the Supervision of Customs Clearance Machine Inspection. *World Customs Journal*, *18*(2). https://doi.org/10.55596/001c.122754

Choi, Y. S. (2019). Identifying trade mis-invoicing through customs data analysis. *World Customs Journal*, *13*(2), 59-76.

Crépey, S., Lehdili, N., Madhar, N., & Thomas, M. (2022). Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks. *Algorithms*, *15*(10), 385. https://doi.org/10.3390/a15100385

Dewoprabowo, R., Stefanus, L. Y., & Saptawijaya, A. (2025). Explainable clustering : Methods, challenges, and future opportunities. *Journal of Intelligent Systems*, *34*(1), 20240477. https://doi.org/10.1515/jisys-2024-0477

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms : Experiments and analyses. *Pattern Recognition*, *74*, 406-421. https://doi.org/10.1016/j.patcog.2017.09.037

European Commission. (2020). PROFILE – Privacy-Preserving Risk Analysis for Custom Law Enforcement (Horizon 2020, Grant No 786748). Publications Office of the European Union.

Fuhnwi, G. S., Agbaje, J. O., Oshinubi, K., & Peter, O. J. (2023). An Empirical Study on Anomaly Detection Using Density-based and Representative-based Clustering Algorithms. *Journal of the Nigerian Society of Physical Sciences*, 1364. https://doi.org/10.46481/jnsps.2023.1364

Geourjon, A.-M., & Laporte, B. (2004). L'analyse de risque pour cibler les contrôles douaniers dans les pays en développement : Une aventure risquée pour les recettes? *Politiques et management public*, *22*(4), 95-109.

Hamid, I., & Rahman, M. M. H. (2025). AI, machine learning and deep learning in cyber risk management. *Discover Sustainability*, *6*(1), 389. https://doi.org/10.1007/s43621-025-01012-3

Hu, R., Lau, Y.-Y., & Wang, R. (2023). Evaluation of Customs Supervision Competitiveness Using Principal Component Analysis. *Sustainability*, *15*(3), 1833. https://doi.org/10.3390/su15031833

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis : A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Khojasteh Aliabadi, H. A., Daei-Karimzadeh, S., Iranpour Mobarakeh, M., & Zamani Boroujeni, F. (2022). Developing a model for managing the risk assessment of import declarations in customs based on data analysis techniques. *Advances in Mathematical Finance and Applications, 7*(4). https://doi.org/10.22034/amfa.2021.1942827.1644

Molina Narváez, P. X., Orellana, M., Lima, J.-F., & Zambrano-Martinez, J. L. (2024a). Vigilancia Inteligente del Comercio Exterior : Detección de Anomalías en las Importaciones del Ecuador con Minería de Datos. *Revista Tecnológica - ESPOL, 36*(E1), 12-24. https://doi.org/10.37815/rte.v36nE1.1208

Molina Narváez, P. X., Orellana, M., Lima, J.-F., & Zambrano-Martinez, J. L. (2024b). Vigilancia Inteligente del Comercio Exterior : Detección de Anomalías en las Importaciones del Ecuador con Minería de Datos. *Revista Tecnológica - ESPOL, 36*(E1), 12-24. https://doi.org/10.37815/rte.v36nE1.1208

Nelson, C. (2020). Machine Learning for Detection of Trade in Strategic Goods : An Approach to Support Future Customs Enforcement and Outreach. *World Customs Journal, 14*(2). https://doi.org/10.55596/001c.116422

Thomas, R., & E. Judith, J. (2019). A survey on outlier detection methods in data mining. *International Journal of Engineering & Technology, 7*(4), 6309-6312. https://doi.org/10.14419/ijet.v7i4.23153

Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Customs fraud detection : Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue. *Pattern Analysis and Applications, 23*(3), 1457-1477. https://doi.org/10.1007/s10044-019-00852-w

Vijayakumar, S. (2025). Technology-centric and Data-Driven Customs Risk Management for Supply Chain Security. *World Customs Journal, 19*(1). https://doi.org/10.55596/001c.131745

Wei, X., Zhang, Z., Huang, H., & Zhou, Y. (2024). An overview on deep clustering. *Neurocomputing, 590*, 127761. https://doi.org/10.1016/j.neucom.2024.127761

Yin, J., Wu, H., & Sun, S. (2023). Effective sample pairs based contrastive learning for clustering. *Information Fusion, 99*, 101899. https://doi.org/10.1016/j.inffus.2023.101899

Zamora Torres, A. I., & Navarro Chávez, J. C. L. (2015). Competitiveness of the customs administration in the international trade frame. *Contaduría y administración, 60*(1), 205-228.