# Catégorisation imprécise de postes de travail pour l'évaluation des risques sanitaires associés à des mélanges de polluants chimiques

# Imprecise categorization of workstations for the risk assessment related to chemical mixtures

Laura Calazans De Oliveira Costa[1] Violaine Antoine[1] Luiza De Oliveira Abrahão Reis[1] Pascal Petit[3] Dominique J. Bicout[2]
[1] Université Clermont Auvergne, Clermont Auvergne INP, UMR 6158 CNRS, LIMOS, Clermont-Ferrand, France
[2] Université Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France
[3] Université Grenoble Alpes, AGEIS, 38000 Grenoble, France

violaine.antoine@uca.fr

**Résumé :**
L'évaluation des expositions professionnelles aux hydrocarbures aromatiques polycycliques (HAP) représente un enjeu majeur pour la prévention des risques sanitaires en milieu professionnel. Cette étude propose d'analyser la similarité de divers postes de travail à travers l'utilisation de méthodes de classification non supervisée. Dans une première partie du travail, plusieurs méthodes de prétraitement et de classification sont testées et évaluées à l'aide d'une mesure d'entropie proposée spécifiquement pour le problème. La seconde partie du travail présente une chaîne de traitement basée sur les fonctions de croyance afin de considérer la possibilité qu'un poste de travail puisse appartenir simultanément à des degrés différents à plusieurs profils d'exposition. Les résultats obtenus sont en cohérence avec les connaissances métiers.

**Mots-clés :**
hydrocarbure aromatique polycyclique, évaluation de risques, classification automatique, fonctions de croyance, fusion d'information.

**Abstract:**
The assessment of occupational exposure to polycyclic aromatic hydrocarbons (PAHs) is a major issue for preventing health risks in the workplace. This study aims to analyze the similarity between various job positions using clustering methods. In the first part of the work, several preprocessing and clustering techniques are tested and evaluated using an entropy measure specifically designed for this problem. The second part introduces a pipeline based on belief functions to account for the possibility that a single job position may belong simultaneously with different degrees to multiple exposure profiles. The results obtained are consistent with expert domain knowledge.

**Keywords:**
polycyclic aromatic hydrocarbon, risk assessment, clustering, belief function, information fusion.

## 1 Introduction

L'exposition professionnelle aux polluants chimiques présents dans l'air constitue une préoccupation majeure de santé publique, en particulier en milieu professionnel. C'est notamment le cas des hydrocarbures aromatiques polycycliques (HAP), une famille de polluants cancérigènes classés comme prioritaires en raison de leur présence dans de nombreuses sources, tant environnementales (ex. : fumée de cigarette, feux de bois, barbecue, gaz d'échappement) que professionnelles (ex. : métallurgie, travaux publics) [1]. En France, plusieurs millions de travailleurs sont potentiellement exposés à ces substances. L'évaluation des risques sanitaires (ERS) liés aux HAP est à la fois essentielle et réglementaire. Toutefois, sa mise en œuvre reste complexe. En effet, les HAP sont généralement émis sous forme de mélanges complexes de gaz et de particules, dont la composition varie selon les sources d'émission et les procédés industriels [1]. Les HAP sont émis lors des processus de combustion incomplète de la matière organique mais aussi au cours de la distillation du charbon et du pétrole, ou de l'utilisation de produits dérivés de ces processus. Avant de pouvoir estimer les risques, il est donc indispensable de caractériser précisément ces situations de multi-exposition.

Une étape clé dans ce processus consiste à

constituer des groupes homogènes d'exposition (GHE), c'est-à-dire des ensembles d'individus supposés exposés de manière similaire à un même polluant [1]. Les GHE sont souvent fondés sur des thésaurus standardisés, mais aucun d'entre eux ne prend en compte simultanément les informations liées à l'exposition et celles liées au poste de travail. Or, deux postes pouvant paraître de nature très différente peuvent présenter des similarités en termes d'exposition aux mélanges de HAP. À ce jour, la constitution des GHE repose généralement sur un seul congénère du mélange de HAP, ce qui limite la représentativité de l'exposition réelle [1]. Une étude préalable a tenté de surmonter cette limitation en constituant des GHE prenant en compte plusieurs HAP à la fois, en s'appuyant sur une classification hiérarchique de Ward (CAH Ward), une méthode d'analyse traditionnelle [2]. Il semble alors intéressant d'étendre ce travail afin de 1) permettre une séparation non linéaire des groupes, et 2) considérer que certains postes peuvent appartenir à plusieurs groupes.

Par conséquent, les contributions proposées dans cette nouvelle étude sont les suivantes :

— la création d'une mesure d'évaluation semi-supervisée en considérant que les observations d'un même poste doivent être regroupées dans une même classe, et l'utilisation de cette mesure pour étudier les meilleurs prétraitements et méthodes de classification non supervisée à appliquer sur le jeu de données,
— l'utilisation d'une classification évidentielle non supervisée [3] sur les observations, suivie d'une méthode de fusion d'information [4] sur les masses des observations pour un même poste, afin de détecter des groupes de postes avec la possibilité d'avoir des postes imprécisément assignés entre plusieurs groupes.

La suite de l'article est organisée comme suit : la section 2 présente le contexte d'étude, c'est-à-dire l'existant en termes d'études des HAPs et les données utilisées dans le cadre de ce travail. La section 3 décrit une première analyse des données avec l'utilisation d'algorithme de clustering standard générant une partition dure. La section 4 présente une chaîne de traitement basée sur l'utilisation des fonctions de croyance pour la modélisation d'imprécision et de doute quant à l'affectation d'un poste à une classe. Finalement, la section 5 présente une conclusion avec une proposition de perspectives à ce travail.

# 2 Contexte d'étude

## 2.1 Analyses existantes des HAPs

Les HAP constituent un sujet d'étude populaire et majeur avec une tendance à la hausse ces deux dernières décennies [5, 6]. En effet, les HAP représentent l'une des familles de composés organiques les plus toxiques connus à ce jour. Malgré des similarités structurales, tous les HAP ne possèdent pas la même toxicité et leur potentiel cancérigène varie grandement. Du fait de leurs propriétés mutagène et cancérigène, les HAP sont classés comme polluants prioritaires par l'Union Européenne. La plupart des recherches actuelles se focalisent sur la caractérisation des expositions aux HAP et leurs effets associés sur la santé humaine [7, 8, 9]. Parmi ces études, certaines groupent les expositions des individus (ex. : villes, industries, etc.) pour conduire l'ERS [10, 11, 2]. La plupart du temps, les études utilisent la classification hiérarchique avec une distance de Ward [10, 11, 2] permettant une visualisation en dendrogramme des résultats.

## 2.2 Données E-HAP

Les données de l'étude préalable [2] et issues d'Exporisq-HAP (E-HAP) [1] ont été utilisées, incluant un total de 3600 observations, 16 variables quantitatives continues (i.e., 16 concentrations de HAP) et une variable catégorique qui désigne le poste de travail. La distribution des postes est équilibrée, avec 100 observations pour chacun des 36 postes distincts.

Les variables présentant des distributions log-normales, il a été décidé de réaliser une transformation logarithmique afin de réduire l'influence des valeurs extrêmes. La figure 1 présente les histogrammes correspondant aux 16 variables transformées. Les données sont normalisées pour être centrées-réduites.

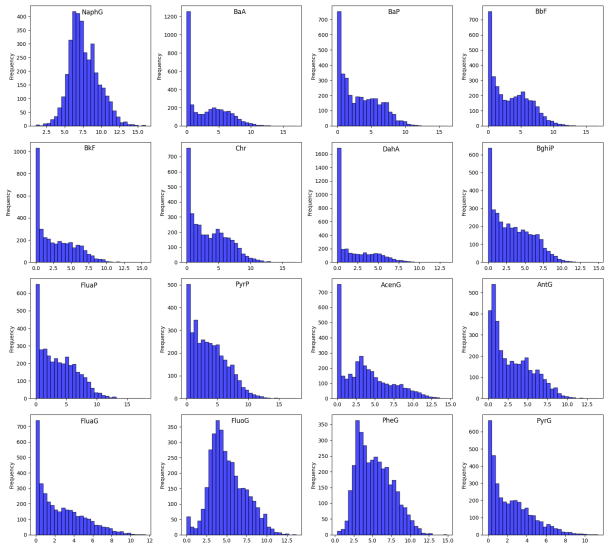


FIGURE 1 – Histogramme des variables après transformation logarithmique.

## 3 Clustering dur des données

L'objectif de l'étude est de regrouper les postes de travail les plus similaires. Pour réaliser cet objectif, une possibilité est d'utiliser un algorithme de clustering à partition dure qui regroupe les individus et qui analyse a postériori, à l'aide d'une mesure d'évaluation, l'intégrité d'un poste à une classe.

### 3.1 Mesure d'évaluation semi-supervisée

Soient $\Omega = \{\omega_1, \ldots, \omega_c\}$ l'ensemble de classes obtenue par clustering, $X = \{x_i\}$ l'ensemble des $n$ individus du jeu de données et $P = \{P_\ell\}$ l'ensemble des $l$ postes. La probabilité conditionnelle qu'un poste $P_\ell$ appartienne à une classe $\omega_k$ est définie par :

$$Pr(\omega_k|P_\ell) = \frac{|\{x_i \in P_\ell \cap \omega_k\}|}{|\{x_i \in P_\ell\}|}, \qquad (1)$$

avec $|.|$ la cardinalité de l'ensemble. L'entropie de la distribution d'un poste dans les différentes classes est donc :

$$H(P_\ell) = -\sum_{\omega_k \in \Omega} Pr(\omega_k|P_\ell) \log Pr(\omega_k|P_\ell). \qquad (2)$$

La métrique finale d'évaluation du clustering est définie comme l'entropie moyenne sur l'ensemble des postes :

$$H = \frac{1}{l}\sum_{\ell=1}^{l} H(P_\ell). \qquad (3)$$

L'entropie $H \geq 0$ est optimale quand elle prend la valeur 0.

### 3.2 Protocole expérimental et résultats

Deux méthodes de clustering générant des partitions dures ont été testées : k-means et la classification hiérarchique en utilisant la distance de Ward. De plus, six stratégies de prétraitement sont comparées : la première se contente de la transformation logarithmique et de la normalisation (Raw), alors que les suivantes appliquent, en plus de cette transformation : une ACP (PCA) afin de compresser l'information, une transformation non linéaire UMAP [12], une transformation non linéaire t-SNE [13].

Les combinaisons ACP et UMAP, ainsi que ACP et t-SNE sont également testées. Le diagramme des différences critiques [14] est alors employé (cf. Figure 2) pour comparer les résultats par prétraitement. Chaque prétraitement utilise ensuite k-means ou la classification hiérarchique de Ward avec différentes valeurs d'hyperparamétrisation (nombre de classes variant de 2 à 6, nombre de voisins égal à 15, 30 ou 50 pour UMAP, et perplexité de 30, 50 ou 100 pour t-SNE).

Le diagramme montre qu'une transformation non linéaire en prétraitement (UMAP et t-SNE) présente des résultats significativement meilleurs en termes de distinction des postes dans les classes. Le tableau 1 présente un

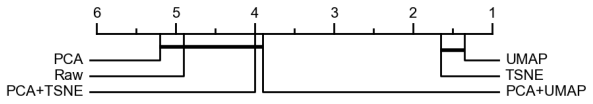


FIGURE 2 – Diagramme de différences critiques pour six stratégies de prétraitement, avec $\alpha = 0.05$ pour les différents tests statistiques.

détail des résultats obtenus pour UMAP et t-SNE. Il permet également de vérifier la stabilité des résultats, notamment pour les transformations non linéaires qui sont des algorithmes non déterministes. Le tableau 1 permet également d'observer que la mesure d'entropie est sensible au nombre de classes. Ainsi, moins il y a de classes et meilleure sera l'entropie. Le choix du nombre de classes a donc été réalisé par l'utilisation de l'indice de silhouette. Au final pour la suite des expériences, nous avons retenu le prétraitement présenté à la figure 3 et k-means avec un nombre de classes fixé à 3.

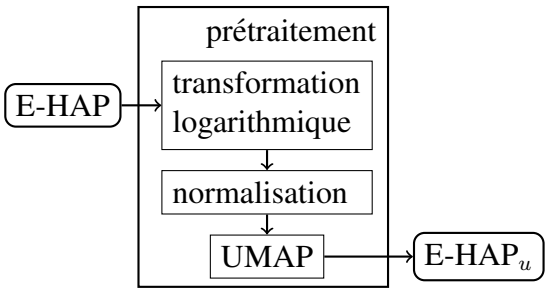


FIGURE 3 – Prétraitement sélectionné.

L'utilisation de partitions dures en clustering a permis d'exploiter la mesure d'entropie eq. (3) pour orienter le choix du prétraitement et de la méthode de clustering. Toutefois, cette approche ne permet pas de détecter si certains postes sont localisés entre plusieurs classes. Nous proposons donc d'utiliser une méthode de clustering évidentiel afin de pallier cette limite.

# 4 Clustering évidentiel

## 4.1 Théorie des fonctions de croyance

La théorie de l'évidence de Dempster-Shafer, également appelée théorie des fonctions de croyance [15], permet de représenter le doute dans un raisonnement. Elle s'applique sur un ensemble d'états fini $\Omega = \{\omega_1, \ldots, \omega_c\}$. La connaissance partielle concernant la valeur d'une variable $\omega$ est représentée par une fonction de masses $m$, qui est une appplication de l'ensemble des parties de $\Omega$ dans l'intervalle $[0, 1]$ telle que

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{4}$$

Les sous-ensembles $A \subseteq \Omega$ tels que $m(A) > 0$ sont appelés les éléments focaux de $m$. La quantité $m(A)$ mesure la croyance allouée exactement à l'hypothèse $A$. Quand $A = \Omega$, $m(\Omega)$ mesure le degré d'ignorance, et une valeur de 1 pour cette quantité représente l'ignorance totale sur la valeur de $\omega$. Enfin, $m(\emptyset) = 1$ peut être interprétée comme la croyance que la valeur réelle de $\omega$ n'appartient pas à $\Omega$.

Lorsqu'une décision doit être prise pour la valeur de $\omega$, il est possible de transformer la fonction de masse $m$ en probabilité pignistique [16].

Lorsqu'il existe plusieurs fonctions de croyance pour la valeur réelle de $\omega$ en provenance de diverses sources d'informations, il est possible de fusionner les informations de ces sources afin de prendre une décision. De nombreuses combinaisons d'informations ont été proposées dans le cadre évidentiel [17]. Les plus connues sont la règle conjonctive, qui est appliquée lorsque toutes les sources d'informations sont fiables, et la règle disjonctive, qui s'applique quand il existe au moins une source non fiable. Ces deux règles sont peu efficaces dans le cas de la fusion de nombreuses sources. Ainsi, dans la référence [4], les auteurs ont proposé une règle de combinaison conjonctive, nommé LNS-CR (Large Number of Sources), qui renforce la croyance sur les éléments focaux dont la majorité des sources sont en accord.

## 4.2 Partition crédale

Le clustering évidentiel consiste en la génération d'une partition crédale $\mathbf{M} = (\mathbf{m}_{x_i})$

| | Méthode | 2 classes | 3 classes | 4 classes | 5 classes | 6 classes |
|---|---|---|---|---|---|---|
| kmeans | raw | 0.198 ± 0.00 | 0.382 ± 0.00 | 0.550 ± 0.00 | 0.561 ± 0.00 | 0.757 ± 0.00 |
| | UMAP (n=15) | 0.113 ± 0.00 | **0.128 ± 0.00** | 0.243 ± 0.03 | 0.400 ± 0.01 | 0.496 ± 0.01 |
| | UMAP (n=30) | 0.118 ± 0.00 | 0.140 ± 0.00 | 0.274 ± 0.04 | 0.420 ± 0.01 | 0.516 ± 0.01 |
| | UMAP (n=50) | 0.118 ± 0.00 | 0.182 ± 0.05 | 0.280 ± 0.04 | 0.442 ± 0.00 | 0.548 ± 0.01 |
| | t-SNE (p=30) | 0.111 ± 0.00 | 0.161 ± 0.00 | 0.326 ± 0.00 | 0.431 ± 0.01 | 0.514 ± 0.00 |
| | t-SNE (p=50) | **0.109 ± 0.00** | 0.148 ± 0.00 | 0.334 ± 0.00 | 0.399 ± 0.00 | 0.507 ± 0.00 |
| | t-SNE (p=100) | 0.117 ± 0.00 | 0.153 ± 0.00 | 0.351 ± 0.00 | 0.436 ± 0.00 | 0.540 ± 0.00 |
| CAH Ward | raw | 0.153 ± 0.00 | 0.314 ± 0.00 | 0.378 ± 0.00 | 0.482 ± 0.00 | 0.680 ± 0.00 |
| | UMAP (n=15) | 0.121 ± 0.01 | 0.145 ± 0.02 | **0.222 ± 0.04** | 0.384 ± 0.03 | **0.478 ± 0.02** |
| | UMAP (n=30) | 0.123 ± 0.01 | 0.146 ± 0.01 | 0.271 ± 0.03 | 0.427 ± 0.02 | 0.511 ± 0.02 |
| | UMAP (n=50) | 0.123 ± 0.01 | 0.149 ± 0.01 | 0.284 ± 0.03 | 0.431 ± 0.02 | 0.531 ± 0.02 |
| | t-SNE (p=30) | 0.113 ± 0.02 | 0.143 ± 0.02 | 0.296 ± 0.02 | **0.381 ± 0.02** | 0.484 ± 0.03 |
| | t-SNE (p=50) | 0.122 ± 0.02 | 0.148 ± 0.03 | 0.297 ± 0.02 | 0.389 ± 0.03 | 0.493 ± 0.04 |
| | t-SNE (p=100) | 0.132 ± 0.01 | 0.205 ± 0.07 | 0.354 ± 0.03 | 0.447 ± 0.02 | 0.542 ± 0.02 |

TABLEAU 1 – Entropie moyenne ± écart type obtenue pour E-HAP avec k-means et la CAH de Ward pour différents paramétrages d'UMAP et t-SNE. En gras sont représentés les meilleurs résultats et en souligné les seconds meilleurs résultats.

de taille $n$ par $2^c$. Chaque masse $\mathbf{m}_{x_i}$ exprime le doute concernant l'affectation de l'individu $x_i$ à une classe. La partition crédale peut être ensuite transformée en partition crédale dure en assignant chaque objet au sous-ensemble dont la croyance est la plus forte.

Il existe de nombreux algorithmes de classification non supervisée qui génère une partition crédale [18]. Le plus populaire est la version évidentielle de k-means, nommé ECM [3], qui minimise l'inertie intra-classe suivante :

$$J = \sum_{i=1}^{n} \sum_{A \subseteq \Omega} |A|^{\alpha} m_{x_i}^{\beta}(A) d_{iA}^2 + \sum_{i=1}^{n} \delta^2 m_{x_i}(\emptyset),$$

avec $d_{iA}$ la distance euclidienne entre l'objet $x_i$ et le sous-ensemble $A$, $\alpha$ et $\beta$ des hyperparamètres qui permettent d'ajuster l'incertitude et l'imprécision obtenue par la partition crédale, $\delta$ une distance fixée permettant de gérer les individus atypiques. La distance entre un objet et une classe est calculée de la même manière que k-means, avec la représentation de la classe par un centre correspondant à son inertie. Dans le cas d'un sous-ensemble de $A$ de cardinalité supérieure à 1, le centre est calculé comme l'isobarycentre des centres associés aux classes composant $A$.

La métrique de non-spécificité quantifie le degré d'imprécision de la partition crédale générée [19] :

$$N(m) = \sum_{A \subseteq \Omega,\, A \neq \emptyset} m(A) \log |A| + m(\emptyset) \log |\Omega|$$

La moyenne normalisée de la non-spécificité sur l'ensemble des fonctions de masses de la partition crédale, nommée $N^{\star}$, peut alors être utilisée comme un indice de validité d'une partition crédale. Cet indice, qui a ses plages de valeurs entre 0 et 1, doit être minimisé.

### 4.3 Chaîne de traitement pour E-HAP

Dans un premier temps, les données E-HAP sont transformées en suivant les étapes présentés à la figure 3. Ensuite, l'algorithme ECM est exécuté afin d'obtenir une fonction de masses pour chaque individu. Ces individus peuvent être considérés comme des sources d'informations pour les postes. Nous appliquons alors une méthode de fusion de ces sources afin d'obtenir une fonction de masse par poste.

Il faut noter que les concentrations sont collectées sur plusieurs années, sur différentes saisonnalités et sur plusieurs employés. Il existe donc une variabilité intrinsèque dans les données pour chaque poste. Dans ce sens, et parce qu'il existe une multitude de sources par poste, nous choisissons d'employer la règle de fusion LNS-CR. La figure 4 illustre la chaîne de traitement réalisée après prétraitement des données. On note $m_{p_\ell}$ la masse attribuée pour le poste $\ell$.

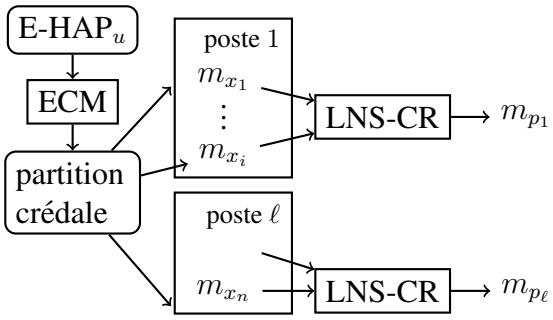


FIGURE 4 – Mise en place de la combinaison d'information en sortie d'ECM.

### 4.4 Résultats

Pour fixer les hyperparamètres de l'algorithme ECM, nous nous sommes appuyés sur les résultats obtenus avec k-means. Ainsi, le nombre de classes est fixé à 3 et l'hyperparamètre $\delta$ est fixé à 10 afin d'obtenir une partition crédale ayant peu de croyance sur l'ensemble vide. Les hyperparamètres $\alpha$ et $\beta$ sont ensuite testés pour diverses valeurs. La non-spécificité $N^\star$ ainsi que l'indice de Rand ajusté (ARI) [20] obtenus sont reportés au tableau 2.

| $\alpha \vert \beta$ | 2 | 2.5 | 3 |
|---|---|---|---|
| 1.1 | 0.24 (0.73) | 0.33 (0.73) | 0.38 (0.73) |
| 1.2 | 0.23 (0.73) | 0.32 (0.73) | 0.37 (0.73) |
| 1.3 | 0.23 (0.73) | 0.31 (0.73) | 0.37 (0.73) |
| 1.4 | **0.22 (0.73)** | 0.31 (0.76) | 0.36 (0.72) |
| 1.5 | 0.21 (0.68) | 0.30 (0.76) | 0.36 (0.70) |

TABLEAU 2 – Non-spécificité et ARI en fonction des hyperparamètres $\alpha$ et $\beta$. L'ARI est notée entre parenthèse.

L'ARI est une métrique permettant de mesurer le degré de concordance entre deux partitions. Si ARI=1, alors les deux partitions sont identiques. Dans notre cas, nous avons comparé la partition produite par k-means avec la partition dure de ECM obtenu en conservant le maximum de probabilité pignistique de la partition crédale. Les valeurs d'ARI élevées traduisent une plus grande similarité entre les deux solutions. Le principe est de sélectionner une solution qui correspond à un compromis entre les deux mesures, $N^\star$ devant être minimisé pour choisir une partition plus certaine et l'ARI devant être maximisé pour que la solution soit en cohérence avec les résultats obtenus par k-means (dans le but de conserver la relative homogénéité des classes par rapport aux postes). Dans cette optique, nous avons sélectionné $\beta = 2$ et $\alpha = 1.4$.

La règle LNS-CR est ensuite appliquée aux fonctions de masses de chaque poste avec $\eta = 0.5$. Une partition crédale dure est ensuite calculée afin de présenter les résultats sous forme d'un diagramme de Venn (cf. figure 5).

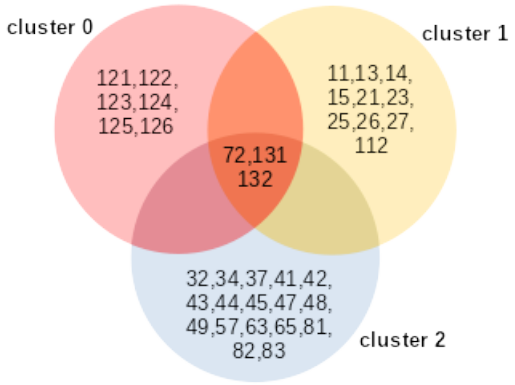


FIGURE 5 – Diagramme de Venn de la partition crédale dure des postes. Les numéros de poste sont des variables catégoriques, le premier chiffre correspondant au secteur industriel.

Les résultats sont cohérents avec les connaissances métier : le cluster 0 regroupe des usines de cokerie, avec des émissions de dérivés de houille et des niveaux élevés de HAP. Le cluster 1 présente également des émissions de dérivés de houille, mais avec des niveaux moyens

de HAP. Le cluster 2 se caractérise par des émissions issues du pétrole et un faible niveau d'exposition. Enfin, les postes 72, 131 et 132 correspondent à des tâches très variées, générant des niveaux de HAP variables mais comparables à ceux de certains groupes.

La figure 6 présente chaque individu dans UMAP et, en couleur, la classe (ou sous-ensemble de classe) dans lequel se trouve le poste d'un individu. Cela permet d'observer que bien que la plupart du temps les individus (et subséquemment les postes) sont regroupés dans une même région, certains individus se retrouvent isolés de leurs groupes. Ceci peut s'expliquer soit par la variabilité intrinsèque des mélanges d'HAP pour chaque poste, soit par l'utilisation de UMAP qui préserve uniquement les distances locales pour sa transformation en deux dimensions.

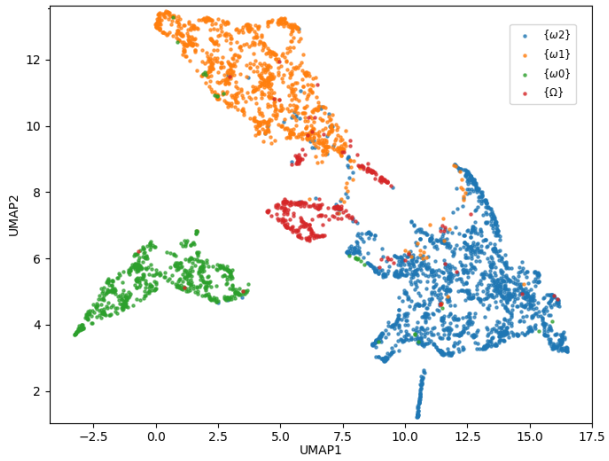


FIGURE 6 – UMAP des résultats de ECM suivi de l'agrégation des observations des postes, selon la règle LNS-CR ($\eta = 0.5$).

## 5 Conclusion

L'analyse des expositions professionnelles aux polluants organiques persistants, en particulier les HAP, est un enjeu important pour la santé des travailleurs. Cette étude identifie des similitudes entre certains postes exposés en utilisant une chaîne de traitement validée par la création d'une mesure d'évaluation semi-supervisée et permettant la génération d'incertitude et d'imprécision quant à l'affectation d'un poste à un groupe. Cette notion de doute, modélisée par la théorie des fonctions de croyances, permet d'intégrer la possibilité qu'un poste soit concerné par plusieurs profils d'exposition. Les résultats obtenus, notamment le diagramme de Venn, concordent avec les connaissances des experts.

Plusieurs perspectives sont envisagées : premièrement, des tests statistiques doivent être mis en place pour renforcer l'interprétation des résultats. Ensuite, il semble intéressant de réaliser une exploration des méthodes de classification non supervisée et non linéaire afin de supprimer du prétraitement l'utilisation de UMAP. Enfin, une dernière perspective consiste à modifier la mesure d'évaluation de l'entropie afin de prendre en entrée des fonctions de masse.

**Remerciements :**

Le travail de Pascal Petit a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-IACL-0006 et du programme Investissements d'avenir (ANR-10-AIRT-05 and ANR-15-IDEX-02).

## Références

[1] P. Petit, D. J. Bicout, R. Persoons, V. Bonneterre, D. Barbeau, and A. Maître, "Constructing a database of similar exposure groups : the application of the exporisq-HAP database from 1995 to 2015," *Annals of Work Exposures and Health*, vol. 61, no. 4, pp. 440–456, 2017.

[2] P. Petit, A. Maître, R. Persoons, and D. J. Bicout, "Modeling the exposure functions of atmospheric polycyclic aromatic hydrocarbon mixtures in occupational environments," *Science of The Total Environment*, vol. 584, pp. 1185–1197, 2017.

[3] M.-H. Masson and T. Denœux, "ECM : An evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

[4] K. Zhou, A. Martin, and Q. Pan, "A belief combination rule for a large number

of sources," *Journal of Advances in Information Fusion*, vol. 14, no. 1, pp. 22–39, 2019.

[5] L. Lu and R. Ni, "Bibliometric analysis of global research on polycyclic aromatic hydrocarbons and health risk between 2002 and 2021," *Environmental Science and Pollution Research*, vol. 29, no. 56, pp. 84034–84048, 2022.

[6] X. Zhang, L. Yang, H. Zhang, W. Xing, Y. Wang, P. Bai, L. Zhang, K. Hayakawa, A. Toriba, Y. Wei, *et al.*, "Assessing approaches of human inhalation exposure to polycyclic aromatic hydrocarbons : A review," *International journal of environmental research and public health*, vol. 18, no. 6, p. 3124, 2021.

[7] K. Szramowiat-Sala, M. Marczak-Grzesik, M. Karczewski, M. Kistler, A. K. Giebl, and K. Styszko, "Chemical investigation of polycyclic aromatic hydrocarbon sources in an urban area with complex air quality challenges," *Scientific Reports*, vol. 15, no. 1, p. 6987, 2025.

[8] F. Barbosa Jr, B. A. Rocha, M. C. Souza, M. Z. Bocato, L. F. Azevedo, J. A. Adeyemi, A. Santana, and A. D. Campiglia, "Polycyclic aromatic hydrocarbons (PAHs) : updated aspects of their determination, kinetics in the human body, and toxicity," *Journal of Toxicology and Environmental Health, Part B*, vol. 26, no. 1, pp. 28–65, 2023.

[9] Y.-W. Chen, K.-T. Liu, H. T. P. Thao, M.-Y. Jian, and Y.-H. Cheng, "Insight into the diurnal variations and potential sources of ambient pm2. 5-bound polycyclic aromatic hydrocarbons during spring in Northern Taiwan," *Journal of Hazardous Materials*, vol. 476, p. 134977, 2024.

[10] M. Callén, J. López, A. Iturmendi, and A. Mastral, "Nature and sources of particle associated polycyclic aromatic hydrocarbons (PAH) in the atmospheric environment of an urban area," *Environmental Pollution*, vol. 183, pp. 166–174, 2013.

[11] G. A. Hasan, F. Rinky, A. K. Das, K. S. Ahmed, and K. Sikdar, "Assessment of polycyclic aromatic hydrocarbon (PAH) levels and health risks in kitchen dust from wood, kerosene, and gas cooking systems in cumilla, bangladesh," *Journal of Hazardous Materials Advances*, vol. 15, p. 100457, 2024.

[12] L. McInnes, J. Healy, and J. Melville, "Umap : Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv :1802.03426*, 2018.

[13] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[15] G. Shafer, *A mathematical theory of evidence*, vol. 42. Princeton university press, 1976.

[16] P. Smets and R. Kennes, "The transferable belief model," *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[17] Z. Liu and S. Letchmunan, "Representing uncertainty and imprecision in machine learning : A survey on belief functions," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 1, p. 101904, 2024.

[18] Z. Zhang, Y. Zhang, H. Tian, A. Martin, Z. Liu, and W. Ding, "A survey of evidential clustering : Definitions, methods, and applications," *Information Fusion*, vol. 115, p. 102736, 2025.

[19] G. Klir and M. Wierman, *Uncertainty-based information : elements of generalized information theory*, vol. 15. Springer Science & Business Media, 1999.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.