IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

Projet UV ODATA

Catégorisation de capteurs de pollution en fonction de leur environnement

Ribouchon Louison
Tuesta Léandre

FISE 2025

Sommaire

Introduction

Le présent projet a été initié en réponse aux besoins spécifiques d'une société spécialisée dans l'analyse des polluants chimiques de l'air. Face à la croissance continue de la demande en matière de surveillance de la qualité de l'air, la société a récemment déployé un réseau de capteurs de pollution dans la ville d'Aix-en-Provence, avec des projets d'expansion imminents vers d'autres grandes villes en France et à l'étranger.

L'objectif majeur de ce projet est de réaliser une classification non supervisée des capteurs de pollution en fonction de leur positionnement dans la ville, c'est-à-dire en fonction de leur environnement spécifique. Il est largement reconnu que l'environnement joue un rôle crucial dans les concentrations de polluants atmosphériques. Par conséquent, la catégorisation des capteurs en groupes présentant des mesures similaires devrait permettre une optimisation significative dans divers aspects opérationnels de la surveillance de la qualité de l'air.

Les principaux objectifs du projet sont les suivants :

1. **Optimisation du déploiement des capteurs** : Identifier les environnements-types pour un déploiement efficace des capteurs dans les villes.

2. **Optimisation de la Calibration** : Améliorer la calibration des capteurs en catégorisant par groupe plutôt qu'individuellement.

3. **Détection d'Anomalies** : Analyser les mesures pour détecter des anomalies dans les concentrations de polluants, pouvant indiquer des dysfonctionnements des capteurs ou des événements particuliers.

Au cours de ce projet, nous explorerons différentes méthodes de clustering, dont le K-means, la Classification Ascendante Hiérarchique (CAH), le Modèle de Mélange de Gaussiennes, DBSCAN, et le Partitionnement Spectral. Une attention particulière sera accordée à la méthode Spectral Clustering, qui n'a pas été étudiée en cours. Cette exploration sera complétée par une analyse approfondie des résultats obtenus et la proposition de catégories d'environnements-types avec leurs caractéristiques prédominantes.

1 - Etudes préalables

1.1 - Méthode de réduction de dimension t-SNE

La méthode t-SNE cherche à représenter des données multidimensionnelles de manière à préserver les similarités entre les points. Elle fonctionne en deux étapes :

Construction des probabilités de similarité : Elle commence par calculer les probabilités de similarité entre chaque paire de points dans l'espace d'origine. Ces probabilités sont basées sur la distribution des distances entre les points.

Construction des probabilités de similarité dans l'espace de projection : Ensuite, t-SNE crée un espace de projection (généralement en deux dimensions) et calcule à nouveau des probabilités de similarité pour les points dans cet espace. L'objectif est de minimiser la divergence Kullback-Leibler entre les probabilités de similarité de l'espace d'origine et celles de l'espace de projection. La divergence de Kullback-Leibler est une mesure de dissimilarité entre deux distributions de probabilités.

Les avantages de t-SNE :

- Capable de révéler des structures complexes et non linéaires dans les données.
- Généralement efficace pour la visualisation de données de haute dimension.
- Convient particulièrement bien pour l'exploration visuelle de clusters ou de groupes de données similaires.

Les limites de t-SNE :

- La projection obtenue n'est pas déterministe, ce qui signifie que des exécutions différentes peuvent donner des résultats légèrement différents.
- t-SNE peut être sensible aux hyperparamètres, ce qui nécessite des ajustements soignés.
- Il n'est pas adapté à la préservation des distances euclidiennes globales entre les points.
- Les calculs peuvent être intensifs en termes de temps et de mémoire, ce qui peut rendre son utilisation problématique pour de grandes quantités de données.

1.2 - Méthode de clustering Spectral Clustering

Le Spectral Clustering est une méthode de clustering qui ne se base pas directement sur la distance euclidienne entre les points, mais plutôt sur la structure spectrale des données. Voici un résumé du principe, des avantages et des limitations de cette méthode :

Création de la matrice d'affinité : Tout d'abord, une matrice d'affinité est construite à partir des données en mesurant la similarité entre les points. Cette matrice capture les relations de voisinage entre les points.

Décomposition spectrale de la matrice d'affinité : Ensuite, une décomposition spectrale est effectuée sur la matrice d'affinité. Cela consiste à extraire les vecteurs propres et les valeurs propres de cette matrice.

Réduction de dimension : Les vecteurs propres correspondant aux valeurs propres les plus importantes sont utilisés pour réduire la dimension des données, généralement en projetant les points dans un espace de dimension inférieure.

Application d'une méthode de clustering : Une méthode de clustering classique, comme K-means, est appliquée dans l'espace de dimension réduite pour regrouper les points en clusters.

Les avantages du Spectral Clustering :

- Capable de détecter des clusters de formes complexes et non linéaires.
- Peut gérer des données de haute dimension.
- Ne nécessite pas de présupposer la forme ou la densité des clusters, ce qui le rend plus flexible.
- Peut identifier des clusters de tailles inégales.

Les limites du Spectral Clustering** :

- Le choix du nombre de clusters (K) est souvent nécessaire, ce qui peut être délicat.
- La méthode est sensible aux hyperparamètres, tels que le type de matrice d'affinité et le nombre de vecteurs propres à conserver.
- Elle peut être intensive en termes de calcul et de mémoire, en particulier pour de grandes quantités de données.
- La compréhension de la décomposition spectrale peut être complexe pour les utilisateurs non initiés.

2 - Catégorisation des capteurs de pollution

2.1 - Examen des données

Fichier : Donnees_environnement_capteurs.xlsx

- 97 Lignes / 11 Colonnes
- Données majoritairement qualitatives
- Décrit le fonctionnement de capteurs en fonction de leur environnement

```
RangeIndex: 97 entries, 0 to 96
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   INDEX                                   97 non-null     object
 1   Typologie de la zone                    97 non-null     object
 2   Type de voie                            97 non-null     object
 3   Nombre de voies                         97 non-null     object
 4   Distance capteur / voie                 97 non-null     object
 5   Position capteurs                       97 non-null     object
 6   Presence d'arbres                       97 non-null     object
 7   Feuille d arbres                        97 non-null     object
 8   Morphologie urbaine                     97 non-null     object
 9   Hauteur des batiments                   97 non-null     object
 10  Distance capteur / batiment             97 non-null     object
 11  Particularite                           97 non-null     object
 12  Trafic Routier/ vehicule leger (TMJA)   97 non-null     object
 13  Trafic Routier/ poids lourds (TMJA)     97 non-null     int64
 14  Trafic Routier/  2 roues (TMJA)         97 non-null     int64
 15  emission moyenne annuelle (kg/maille/an)  97 non-null   int64
 16  Donnees Cartographie                    97 non-null     float64
dtypes: float64(1), int64(3), object(13)
memory usage: 13.0+ KB
```

Fichier : Donnees_mesures_PM2_5.xlsx

- 25784 Lignes / 29 colonnes
- Données uniquement quantitatives
- Décrit des données mesurées par différents capteurs

```
RangeIndex: 25784 entries, 0 to 25783
Data columns (total 29 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    25784 non-null  datetime64[ns]
 1   c1      22679 non-null  float64
 2   c3      22817 non-null  float64
 3   c5      22428 non-null  float64
 4   c7      16003 non-null  float64
 5   c8      19543 non-null  float64
 6   c10     22067 non-null  float64
 7   c11     24340 non-null  float64
 8   c13     20490 non-null  float64
 9   c14     23649 non-null  float64
 10  c19     24106 non-null  float64
 11  c16     23750 non-null  float64
 12  c17     23824 non-null  float64
 13  c22     22418 non-null  float64
 14  c23     24959 non-null  float64
 15  c26     13677 non-null  float64
 16  c27     15008 non-null  float64
 17  c28     21866 non-null  float64
 18  c29     22355 non-null  float64
 19  c31     23020 non-null  float64
 20  c87     20360 non-null  float64
 21  c50     23988 non-null  float64
 22  c51     24173 non-null  float64
 23  c64     23808 non-null  float64
 24  c73     23726 non-null  float64
 25  c74     23811 non-null  float64
 26  c75     24919 non-null  float64
 27  c76     20229 non-null  float64
 28  c90     22615 non-null  float64
dtypes: datetime64[ns](1), float64(28)
memory usage: 5.7 MB
```

2.2 - Pré-traitement des données

Table : Donnees_environnement_capteurs

Nous avons procédé à une conversion des données numériques de la colonne “Trafic Routier véhicule léger” car elle n’était pas flottantes à cause d’espaces et de caractères spéciaux présents dans la colonne. On a également uniformisé le type des variables numériques pour éviter les erreurs de calcul python. Nous avons ensuite utilisé un codage ordinal sur les colonnes afin de convertir les données qualitatives en données quantitatives. Un codage ordinal consiste à faire correspondre chaque étiquette unique à une valeur entière.

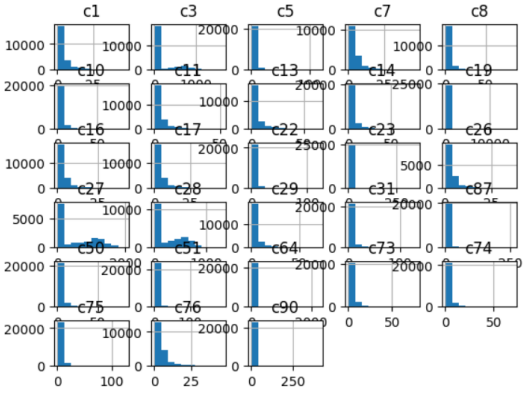
```
RangeIndex: 97 entries, 0 to 96
Data columns (total 17 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Capteurs                              97 non-null     object
 1   Typologie_zone                        97 non-null     float64
 2   Type_voie                             97 non-null     float64
 3   Nombre_voies                          97 non-null     float64
 4   Distance_capteur_voie                 97 non-null     float64
 5   Position_capteurs                     97 non-null     float64
 6   Presence_arbres                       97 non-null     float64
 7   Feuille_arbres                        97 non-null     float64
 8   Morphologie_urbaine                   97 non-null     float64
 9   Hauteur_batiments                     97 non-null     float64
 10  Distance_capteur_batiment             97 non-null     float64
 11  Particularite                         97 non-null     float64
 12  Trafic_Routier_vehicule_leger         97 non-null     float64
 13  Trafic_Routier_poids_lourds           97 non-null     float64
 14  Trafic_Routier_2_roues                97 non-null     float64
 15  emission_moyenne_annuelle_kg_maille_an 97 non-null    float64
 16  Donnees_Cartographie                  97 non-null     float64
dtypes: float64(16), object(1)
memory usage: 13.0+ KB
```

Il nous a alors suffi de normaliser la table en retirant la colonne “Capteurs” et les données étaient alors adéquates pour effectuer l’étude.

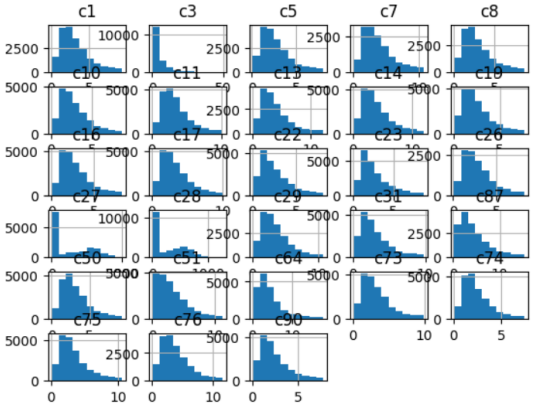
[136]:

	Capteurs	Typologie_zone	Type_voie	Nombre_voies	Distance_capteur_voie	Position_capteurs	Presence_arbres	Feuille_arbres	Morphologie_urbaine	Hauteur_batime
0	c1	-0.517024	0.430795	-1.320749	-1.290219	0.877568	0.808421	-0.094694	-0.40209	-0.239
1	c2	-0.517024	0.430795	-1.320749	-1.290219	-0.573411	0.808421	-0.094694	-0.40209	-0.820
2	c3	-0.517024	-1.934512	-1.320749	-1.290219	-2.508050	0.808421	-0.094694	-0.40209	-0.820
3	c4	-0.517024	-1.934512	-1.320749	-1.290219	-0.573411	0.808421	-0.094694	-0.40209	-0.820
4	c5	-0.517024	0.430795	-1.320749	-1.290219	0.877568	0.808421	-0.094694	-0.40209	-0.239
...	...	...	...	...	...	...	...	...	...	...

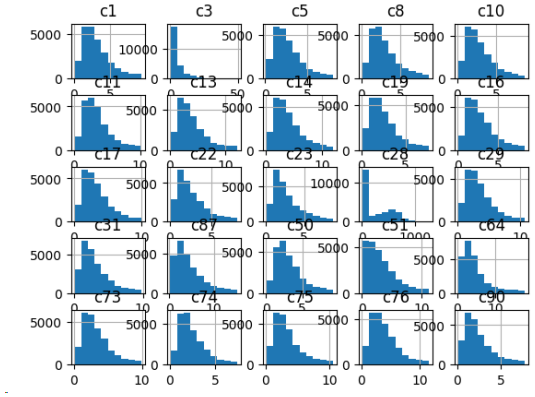
Table : Donnees_mesures_PM2_5



Nous avons observé qu’il y avait un nombre conséquent de données aberrantes dans le jeu de données. Nous pouvons facilement l’observer sur le graphique ci-contre du caractérisé par une dispersion excessive des points.

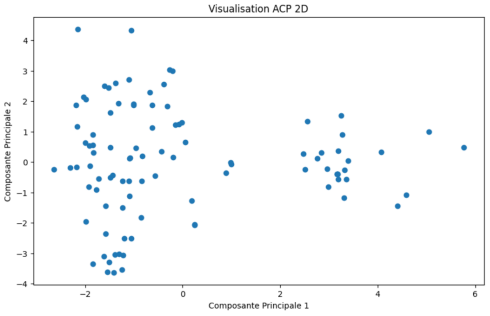


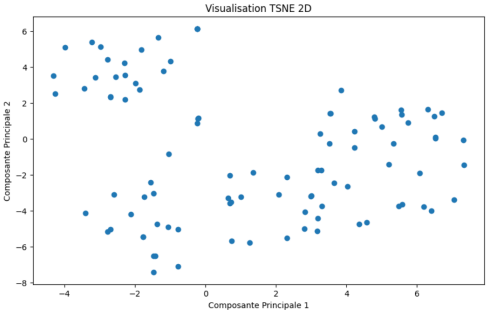
En supprimant les données aberrantes grâce à la méthode interquartile (méthode permettant d'identifier la dispersion des données et de détecter les valeurs aberrantes qui s'éloignent significativement du centre de la distribution) nous obtenons un graphe bien mieux centré et nous avons un meilleur aperçu de la répartition des valeurs.



Nous avons ensuite supprimé les colonnes c7, c26 et c27 qui avaient un taux de valeurs nulles supérieur à 33%. Par la suite, nous avons utilisé la méthode d'Imputation par échantillonnage. Cette méthode à pour but de remplacer les valeurs manquantes par des valeurs tirées au hasard à partir de la distribution existante de la colonne. Cette méthode permet de garder une distribution similaire à celle initiale.

2.3 - Visualisation des données



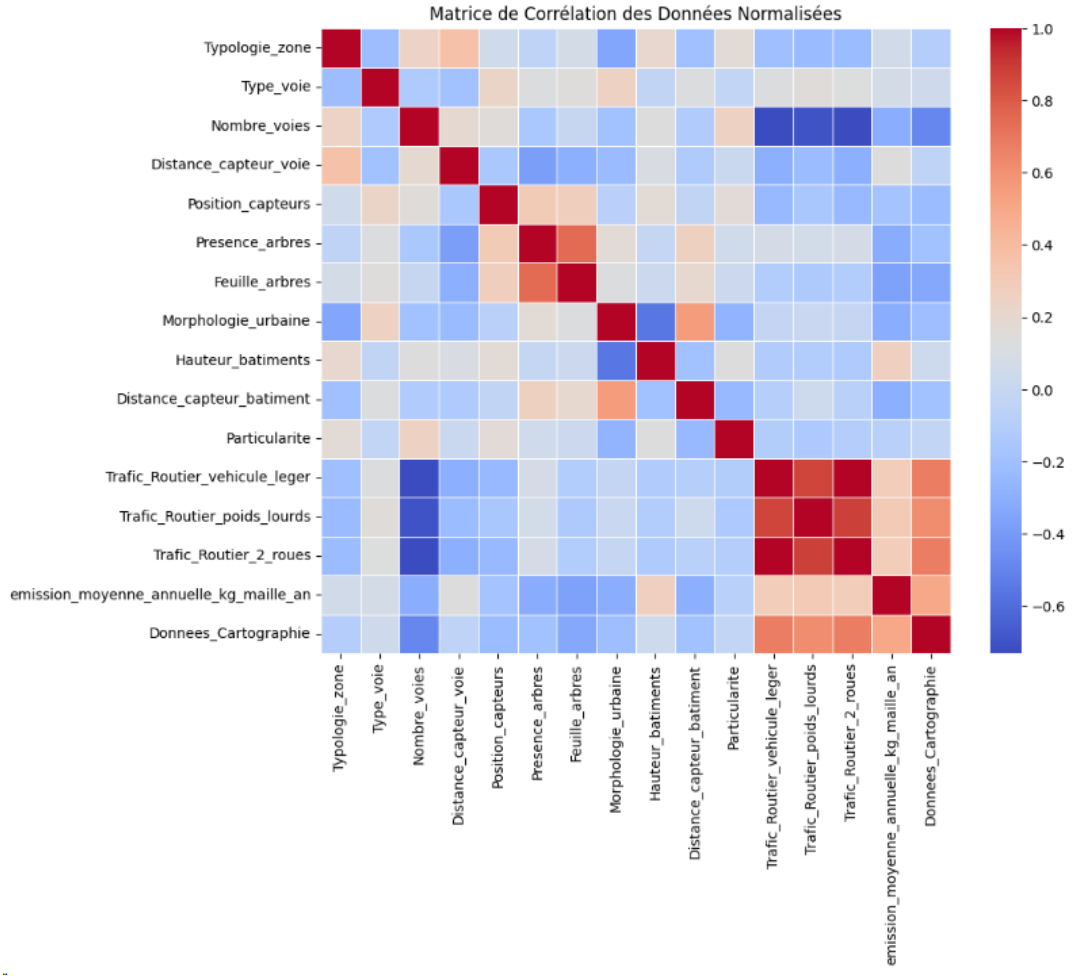


Les deux graphes ci-dessus représentent respectivement les données environnement capteur normées sur lesquels on a effectué une ACP ou un t-SNE. Il est difficile de donner une interprétation uniquement avec une ACP ou une t-SNE. Cependant, visuellement on pourrait faire apparaître deux clusters. Pour l'ACP nous pouvons observer une nette distinction entre la partie de droite et celle de gauche. En ce qui concerne t-SNE

nous pouvons observer une distinction entre les points présents dans la partie supérieure gauche et le reste des points.

2.4 - Recherche des corrélations

Matrice des corrélations :



Variables les plus corrélées positivement :
Trafic_Routier_2_roues / Trafic_Routier_vehicule_leger : 0.999
Tout à fait naturel car les véhicules 2 roues sont par définition des véhicules légers.

Variables les plus corrélées négativement :
Trafic_Routier_2_roues / Nombre_voies : -0.732
Moins il y a de voies sur la route, plus il y a de véhicules 2 roues.

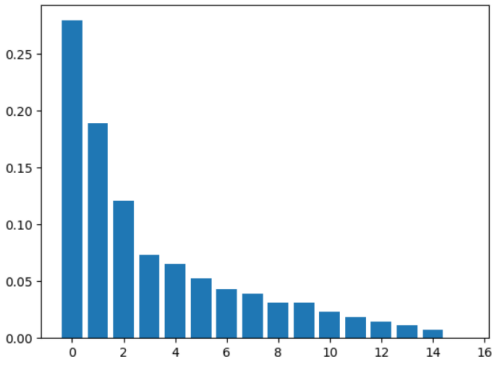
Variables les moins corrélées :

Feuille_arbres / Nombre_voies : 0.004
Le nombre de voies n'influe pas sur la présence de feuilles d'arbres.

Nous pouvons observer que les variables qui au début étaient qualitatives sont en général peu corrélées avec les autres variables. A contrario, les variables numériques ont une forte corrélation entre elles.

2.5 - Analyse exploratoire des données

Après avoir réalisé l'ACP nous pouvons voir un graphique décrivant la part d'inertie de chaque composante.



Part d'inertie : choisir le nombre d'axes de façon à conserver une certaine part de l'inertie totale, cela revient à fixer un seuil de qualité global.

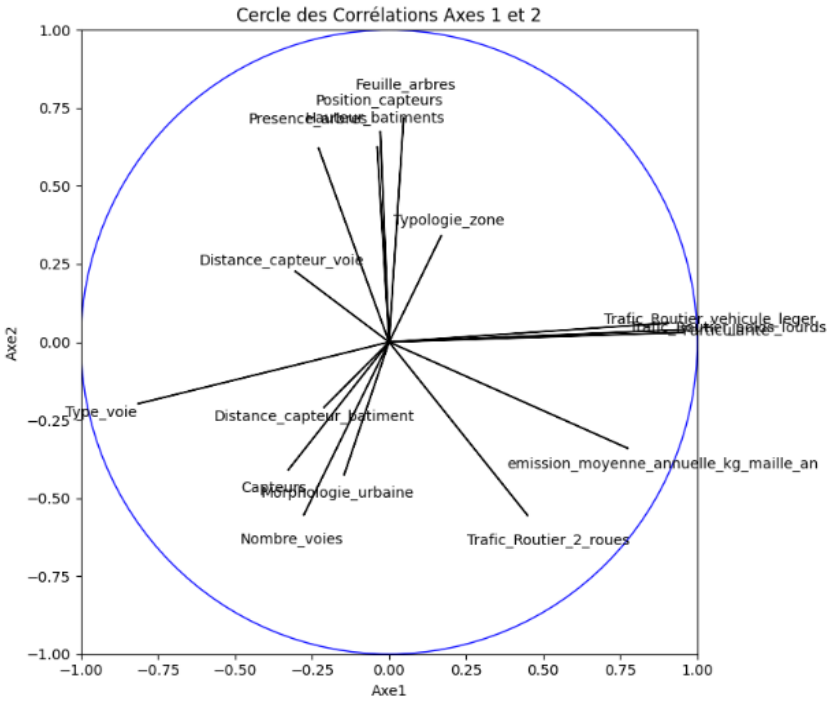
Règle de Kaiser : On retient que les axes associés aux valeurs propres considérées comme les plus "informatives", donc supérieures à leurs moyennes l/p.

Eboulis des valeurs propres : On trace un graphe représentant la décroissance des valeurs propres et on cherche un coude dans le graphe. On retient les axes associées aux valeurs propres situées avant le coude.

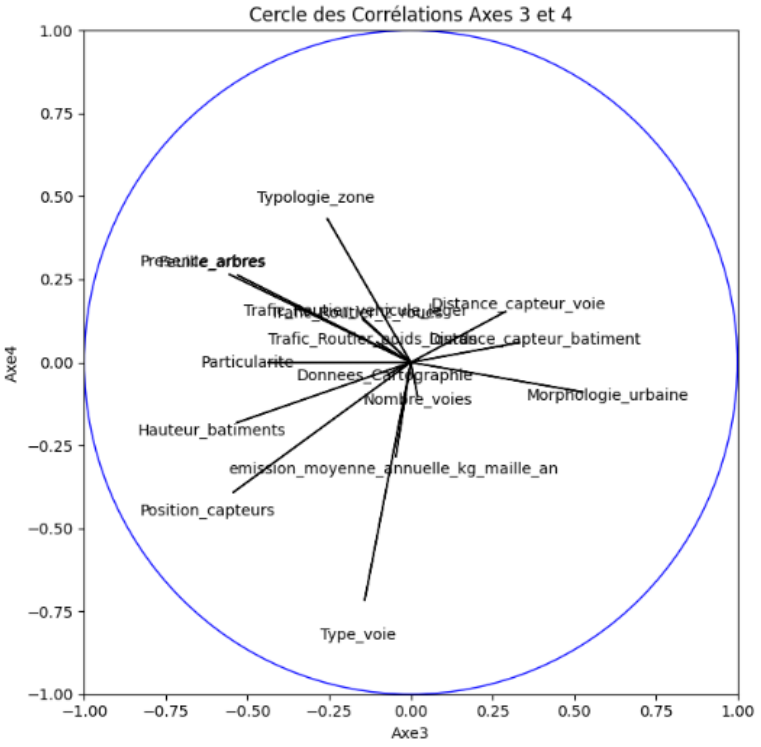
En suivant la définition de ces trois critères, nous pouvons sélectionner pour chacun d' eux le nombre de composantes à garder. Pour la part d'inertie nous sélectionnons 5 composantes, pour la règle de Kaiser 4 composantes et pour l'éboulis des valeurs propres 3 composantes.

En prenant l'ensemble des ces critères, nous avons choisis de garder 4 composantes principales dans la suite de notre étude.

Cercle des corrélations :

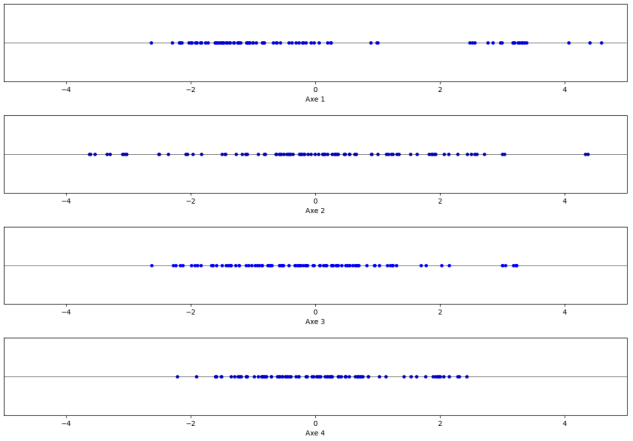


Le cercle des corrélations entre les axes 1 et 2 montre que l'axe 1 est expliqué par le trafic routier véhicule léger et lourd ainsi que la particularité. L'axe 2 quant à lui est expliqué par la variable feuille arbre, la hauteur des bâtiments et la position du capteur .



Le cercle des corrélations entre les axes 3 et 4 montre que l'axe 3 est expliqué par la morphologie urbaine. Quant à l'axe 4, il est expliqué par le type voie.

Projection des capteurs dans les plans principaux

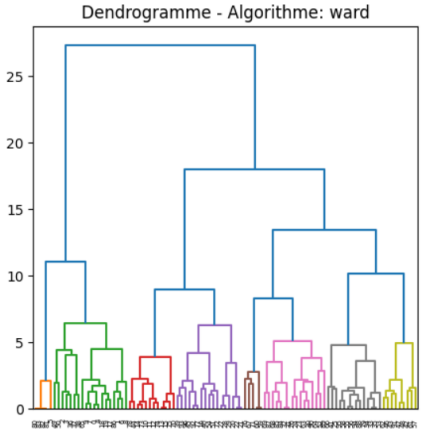


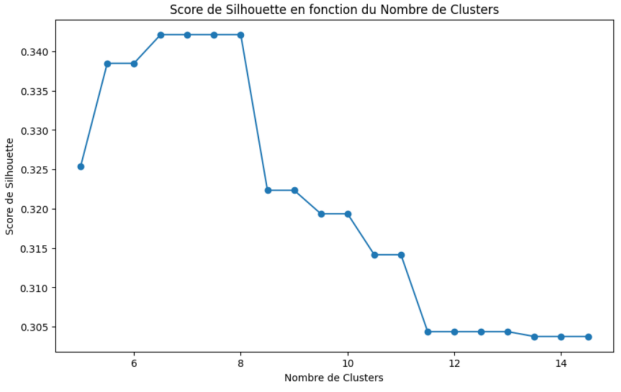
Nous pouvons observer avec la projection des capteurs sur les différents plans principaux que les données sont mieux représentés sur les axes 1 et 2 (données plus étalées) que sur les axes 3 et 4 (données plus rapprochées de 0)

2.6 - Clustering des données

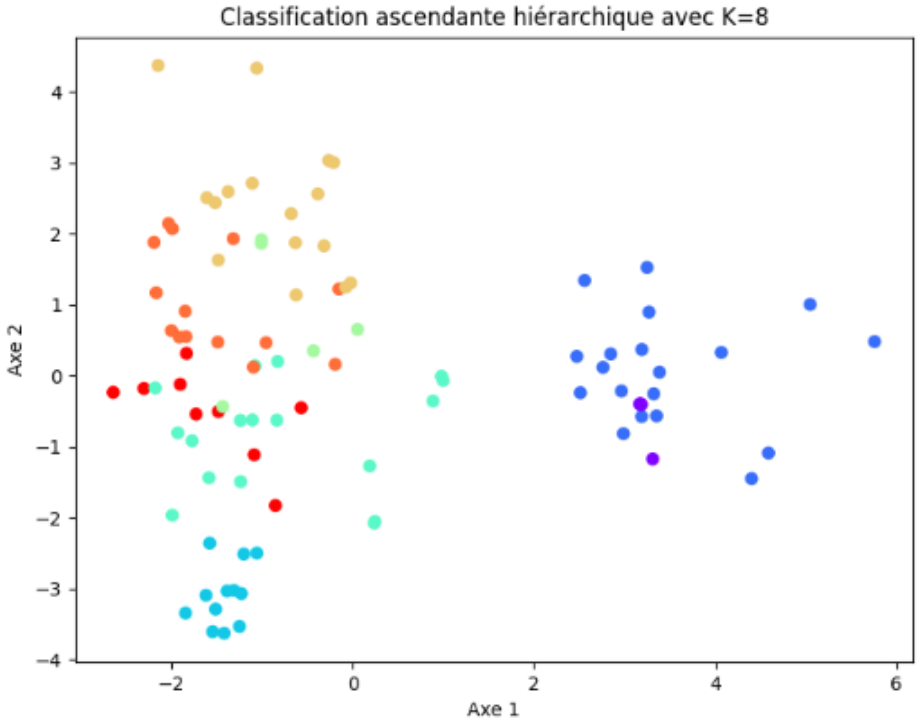
a. Classification ascendante hiérarchique

Pour ce clustering, nous avons choisi d'utiliser la méthode ward. Cette méthode est celle qui rend le dendrogramme le plus équilibré entre les distances des regroupements de cluster. Nous avons ensuite appliqué le score de silhouette pour trouver la valeur de t la plus optimale pour couper le dendrogramme et récupérer le nombre de clusters le plus performant sur ce score. Suite au score de silhouette, nous avons fait le choix de prendre pour valeur de t 8 ce qui ne permet d'avoir 8 clusters.

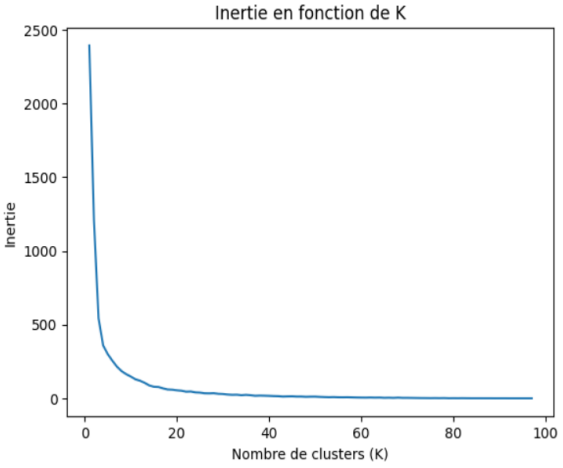


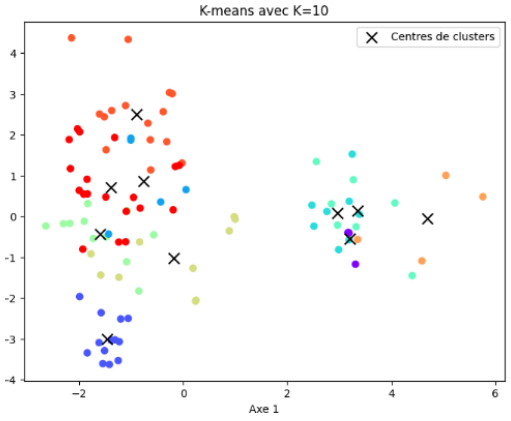


Graphe des données clusterisées avec t-SNE et CAH



b. Classification K-means



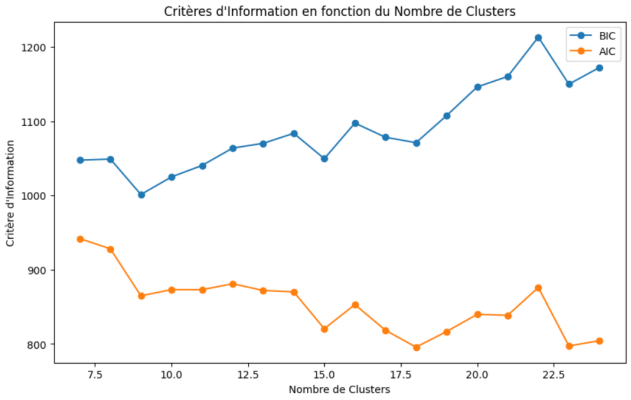


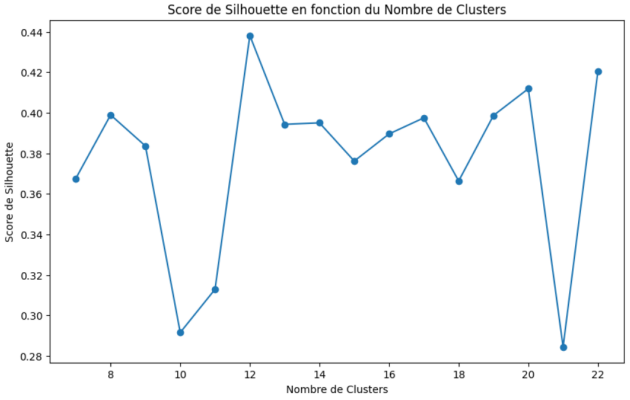
Pour le clustering en utilisant la méthode K-means nous avons utilisé la somme des carrés intra-classe pour déterminer le nombre optimal de clusters. L'idée est de choisir le nombre de clusters de telle sorte que l'inertie intra-cluster soit minimale tout en évitant d'avoir un nombre excessif de clusters. L'inertie est calculée comme la somme des carrés des

distances entre chaque point et le centre de son cluster assigné. Plus l'inertie est faible, plus les points à l'intérieur de chaque cluster sont proches les uns des autres.

Pour choisir le nombre optimal de clusters, nous avons tracé la courbe d'inertie en fonction du nombre de clusters et cherché le point où l'inertie cesse de diminuer de manière significative (coude de la courbe). Nous sommes arrivés à la conclusion qui fallait prendre 10 clusters (K=10) pour cette méthode.

c. Modèle de mélange de Gaussiennes

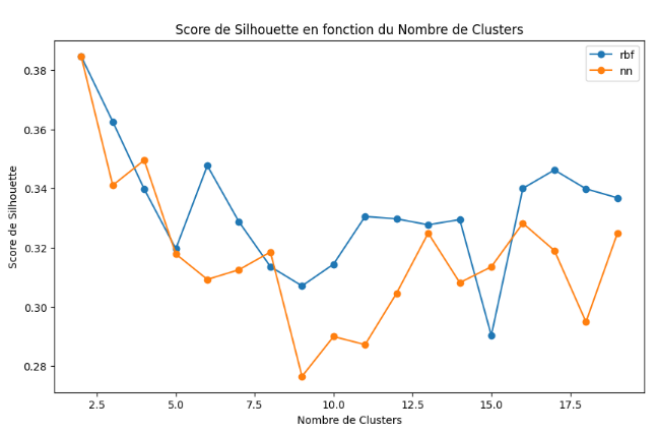


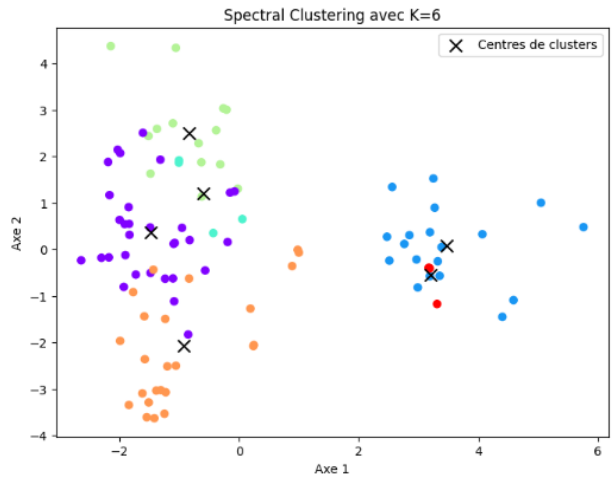


Pour le clustering utilisant la méthode de mélange de Gaussiennes, le nombre de clusters à définir est moins évident. Nous avons dans un premier temps utilisé les critères d'informations de Bayésien (BIC) et de Akaike (AIC). L'AIC évalue la qualité d'un modèle en considérant à la fois son ajustement aux données et sa complexité. Il favorise les modèles précis avec un nombre de paramètres modéré. Le BIC est une mesure similaire à l'AIC, mais il pénalise davantage les modèles complexes. Il privilégie des modèles plus simples tout en maintenant un bon ajustement aux données.

Comme le choix du nombre de clusters uniquement avec ces critères est compliqué, nous avons utilisé le score de sihlouette. C'est ainsi que nous avons choisi comme nombre de clusters 12 pour le modèle de mélange de gaussiennes.

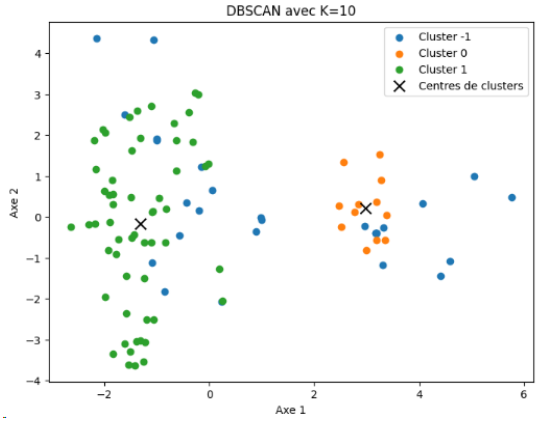
d. Spectral Clustering





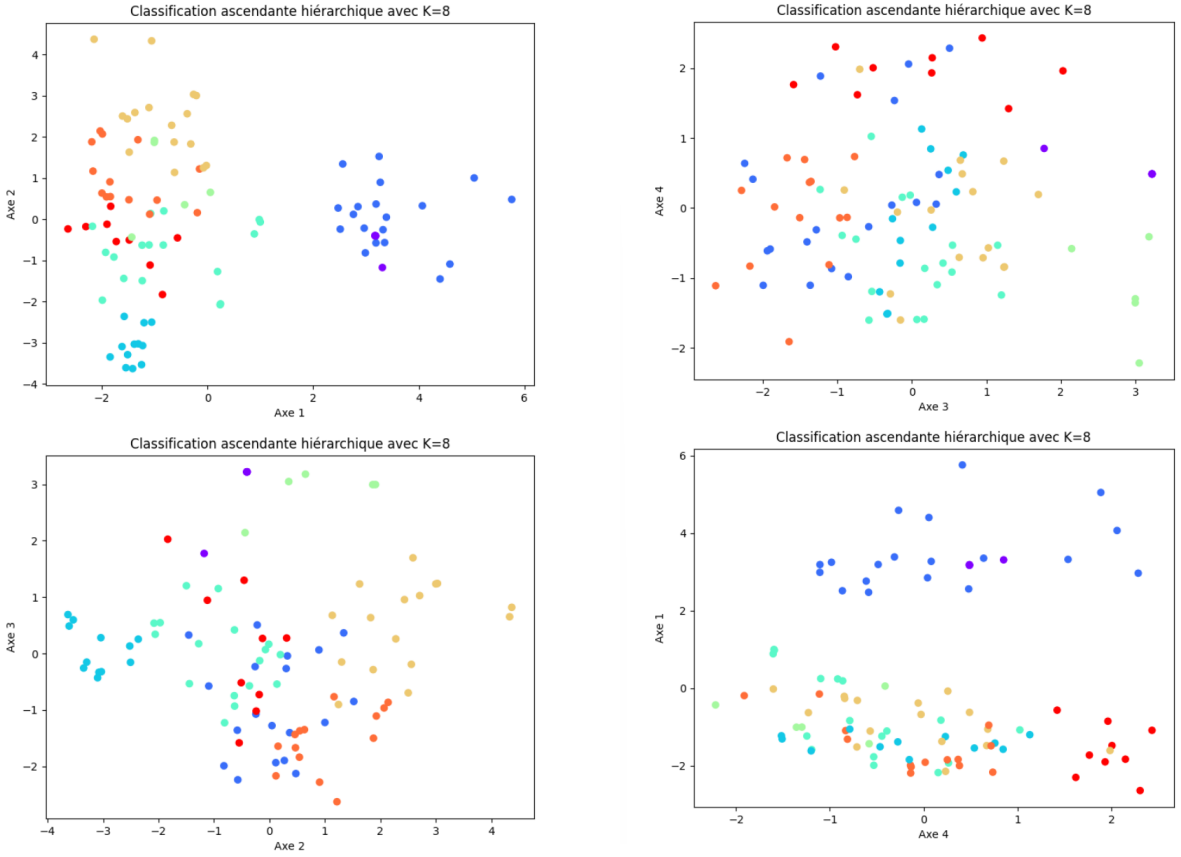
Pour le clustering utilisant la méthode de clustering spectral nous sommes partie sur une nombre de 6 clusters en utilisant la méthode ‘rbf’ qui permet de construire la matrice d'affinité à l'aide d'un noyau de fonction à base radiale. Comme pour les précédentes méthodes de clustering nous avons utilisé le score de silhouette pour faire ce choix.

e. DBSCAN

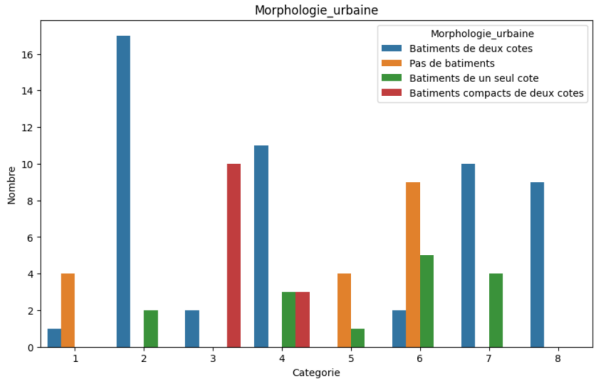


En ce qui concerne la méthode de clustering DBSCAN, nous avons utilisé comme valeurs eps=1.5 et min_samples=5. Ces deux valeurs ont été trouvées grâce à l’utilisation du score de silhouette en faisant varier l’ensemble des paramètres.

2.7 Catégorisation des capteurs de pollution



Nous avons décidé d'utiliser la méthode de classification ascendante hiérarchique sur les données traitées par ACP pour catégoriser les capteurs de pollution. Ce graphe décrit 8 catégories de capteurs de pollution. (on l'a projeté sur les 4 axes principaux pour se rendre compte) Pour déterminer une liste d'environnements-types avec les principales caractéristiques des catégories, on a décidé d'afficher un graphe pour chaque variable afin d'en faire ressortir les particularités de chacun des clusters. Voir un exemple de ces graphes ci-dessous :



Catégorie 1 :

- Nombre de voies : Boulevard à 3 voies de circulation et circulation bus
- Distance capteur voie : supérieur à 4 mètres
- Position du capteur : Au dessus de toit
- Trafic routier léger : environ 18000

Catégorie 2 :

- Typologie zone : centre ville
- Type voie : Rue droite
- Nombre de voies : Boulevard à 3 voies de circulation et circulation bus
- Distance capteur voie : 1 à 2 mètres
- Trafic routier léger : grande quantité supérieur à 30000

Catégorie 3 :

- Nombre voies : Rues très étroite
- Distance Capteur Voie : inf à 1m
- Présence arbre : non
- Feuille arbres : Non
- Morphologie urbaine : bâtiments compacts de deux côtés
- Hauteur Bâtiments : R+3
- Distance capteur bâtiments : 0m
- Particularité : Rue canyon
- Trafic léger <= 6000
- Aucun Poids Lourd
- Aucun 2 roues
- 3500 kg maille/an

Catégorie 4 :

- Type voie : rue droite
- Distance capteur voie : inf à 1 mètre
- Hauteur bâtiments : R+3

Catégorie 5 :

- Typologie zone : Quartiers Périphériques
- Présence arbres : Non
- Feuille arbres : Non
- Aucun Poids Lourd
- Aucun 2 roues
- Trafic léger <= 6000

Catégorie 6 :

- Typologie zone : Quartiers Périphériques
- Présence arbre : rangée arbres
- Hauteur bâtiments : Non
- Distance capteur bâtiment : Non ou sup 4m

Catégorie 7 :

- Typologie zone : zones résidentielles
- Position capteur : Parking 4 roues et arbre
- Présence arbre : rangée arbres

Catégorie 8 :

- Typologie zone : zones résidentielles
- Type de voie : Intersection en T
- Morphologie Urbaine : Bâtiments de deux côtés

Par exemple, on voit que la catégorie 3 regroupe les capteurs se trouvant dans des petites ruelles étroites. Alors que la catégorie 8 regroupe des capteurs situés dans des zones résidentielles d'où la forte présence d'intersections et de bâtiments des 2 côtés. En outre, on peut voir que la catégorie 1 décrit des capteurs placés dans des zones à forte circulation (3 voies de circulations donc les capteurs sont plus éloignés des voies, etc…)

2.8 - Pertinence de la catégorisation

Pour finir nous avons décidé de lier les mesures de capteurs au groupe auquel il appartient. Pour ce faire nous avons utilisé les labels fournis par le clustering et nous avons assimilé chaque mesure au cluster dont fait partie le capteur. Notre idée était de regarder les statistiques descriptives de chacun des clusters afin d'observer si nous avions une différence observable entre les clusters.

Vérifions donc si notre catégorisation est efficace en comparant les capteurs c11, c13 et c14 qui appartiennent à la catégorie 3 et dont nous avons les données qu'ils ont mesurées. Ils devraient avoir des données similaires comme ils sont catégorisés en tant que capteurs de petites ruelles.

Statistiques descriptives des capteurs :

	c11	c13	c14	c19
count	25784.000000	25784.000000	25784.000000	25784.000000
mean	3.486884	3.923727	3.949062	2.552340
std	2.025223	2.461266	2.432666	1.565017
min	0.080000	0.087000	0.073000	0.084000
25%	1.981000	2.096000	2.104000	1.378000
50%	3.065500	3.329000	3.384000	2.193000
75%	4.450250	5.090250	5.166250	3.334000
max	10.201000	12.137000	11.901000	7.630000

Comme vous pouvez l'observer sur l'image ci-dessus, Les 3 capteurs de la catégorie 3 ont des statistiques relativement similaires, tandis qu'un 4e capteur aléatoire (c19) a des statiques sensiblement différentes des 3 autres. Nous pouvons donc en conclure que cette catégorisation semble cohérente.