Université de Strasbourg

MASTER D'INFORMATIQUE
PARCOURS SDSC

Projet de Données Complexes

ÉVALUATION DE LA QUALITÉ DE L'EAU PAR CLUSTERING ET INTERPRÉTATION SPATIALE

29 mai 2025

Preparé par

Ata Berk KARABAG

Ali Emre SÜSLÜ

Table des matières

Chapitre 1

TP1

1.1 Introduction

L'objectif général de ce projet est d'étudier les relations entre les paramètres physico-chimiques de la qualité de l'eau et la classification des stations de mesure en France, en s'appuyant sur les grilles de qualité définies par le référentiel SEQ-Eau / SANDRE. Ce premier TP vise à explorer les données brutes disponibles, en comprendre la structure, effectuer des prétraitements nécessaires, et sélectionner les variables pertinentes pour la suite de l'étude.

1.2 Données utilisées

Les fichiers mis à disposition comprennent :

- `pc_ARMORICAIN.csv` : mesures physico-chimiques (principal fichier analysé)
- `hb_ARMORICAIN.csv` : mesures hydrobiologiques (non utilisé dans TP1)
- `stations.csv` : métadonnées sur les stations (exploration possible en TP2/TP3)

Le fichier `pc_ARMORICAIN.csv` contient plus d'un million de lignes et 52 colonnes. Les colonnes clés pour cette étape sont :

- `DatePrel` : date du prélèvement
- `LbLongParamètre` : nom du paramètre analysé
- `RsAna` : valeur mesurée

1.3 Prétraitement des données

Les principales étapes de prétraitement ont été :

- Conversion de `DatePrel` en format `datetime`
- Suppression des lignes avec valeurs manquantes sur les champs essentiels : `DatePrel`, `CdParametre`, `RsAna`
- Conversion de `RsAna` en `float` via coercition (gestion des erreurs de type)
- Création d'un sous-ensemble filtré pour les seuls paramètres pertinents

Ces étapes ont permis de préparer un jeu de données cohérent, structuré et prêt pour analyse.

1.4 Sélection des paramètres pertinents (grille SANDRE)

Sur la base du documents, nous avons identifié 15 paramètres physico-chimiques directement exploitables pour une future classification selon le modèle SEQ-Eau :

- Température de l'eau
- Potentiel en hydrogène (**pH**)
- Ammonium (NH_4^+)
- Azote Kjeldahl
- Taux de saturation en oxygène
- Oxygène dissous
- Conductivité à 25°C
- Matières en suspension (MES)
- Demande biochimique en oxygène en 5 jours (DBO_5)
- Nitrites (NO_2^-)
- Nitrates (NO_3^-)
- Orthophosphates (PO_4^{3-})
- Phosphore total
- Carbone organique
- Turbidité (Formazine)

Ces paramètres ont été choisis car ils disposent de seuils de qualité normalisés dans les grilles SEQ-Eau pour les classes 1 à 5.

1.5 Statistiques descriptives

Des statistiques de base (minimum, maximum, moyenne, médiane, écart-type) ont été produites pour chaque paramètre. Les résultats montrent que :

- La couverture des données est excellente : plus de 60 000 observations pour chaque paramètre.
- Certains paramètres (par exemple : Ammonium, Nitrites) présentent une forte asymétrie à droite.
- La distribution de l'oxygène dissous est proche d'une loi normale centrée autour de 9 mg/L.
- Des valeurs aberrantes ont été observées pour plusieurs paramètres, ce qui justifie l'usage de boxplots dans l'analyse visuelle.

1.6 Visualisations exploratoires

Les visualisations suivantes ont été réalisées pour illustrer la distribution et les valeurs extrêmes des paramètres clés :

- **Boxplot pour l'Ammonium** : détection d'une forte présence de valeurs aberrantes, suggérant l'éventuelle nécessité d'une transformation logarithmique.
- **Histogramme pour l'oxygène dissous** : distribution proche d'une loi normale, comme attendu.
- **Boxplots et histogrammes supplémentaires** pour les paramètres suivants : DBO_5, NO_3^-, **pH**, PO_4^{3-}, taux de saturation en oxygène.
- **Visualisation croisée par année** : envisagée pour l'analyse des évolutions temporelles dans les TP ultérieurs.

Ces visualisations ont permis de confirmer que les paramètres étudiés sont bien exploitables dans une logique de classification selon la grille SANDRE, ce qui fera l'objet du TP2.

1.7 Discussion intermédiaire

— Les données sont en grande majorité propres, denses et exploitables sans nécessiter de nettoyage lourd.
— Les paramètres choisis sont conformes aux recommandations officielles définies dans le référentiel SEQ-Eau.
— La granularité temporelle a été conservée, ce qui autorise des analyses chronologiques dans les prochaines étapes.
— L'étape suivante consistera à affecter une **classe de qualité** à chaque mesure, selon les seuils définis par SANDRE, en commençant par le paramètre **Ammonium**.

Chapitre 2

TP2 : Attribution des classes de qualité et analyse exploratoire

2.1 Attribution des classes SANDRE

Nous avons utilisé les seuils officiels fournis par **SEQ-Eau** pour catégoriser chaque observation dans une classe de qualité allant de 1 (très bon état) à 5 (mauvais état). Cela a été réalisé en définissant des **intervalles spécifiques par paramètre**, appliqués à la colonne `RsAna`. Une nouvelle colonne `ClasseQualité` a ainsi été ajoutée au jeu de données.

Ce processus permet d'harmoniser les résultats et d'envisager une **classification supervisée** dans les étapes ultérieures.

2.2 Visualisation de la répartition des classes par station

Nous avons visualisé la répartition des classes de qualité pour plusieurs stations, en utilisant les couleurs normalisées :

- Bleu = classe 1 (très bon état)
- Vert = classe 2
- Jaune = classe 3
- Orange = classe 4
- Rouge = classe 5 (mauvais état)

Ces visualisations illustrent la **diversité des mesures entre stations**, ainsi que les **déséquilibres** fréquents observés sur certains paramètres, notamment l'ammonium.

2.3 Analyse de la complétion des stations

Pour assurer la représentativité des stations dans l'analyse, nous avons étudié le nombre de paramètres disponibles par station. Une **station idéale** devrait comporter des valeurs pour l'ensemble des **15 paramètres sélectionnés**.

Les stations présentant une couverture trop faible ont été identifiées et exclues de l'analyse. Ce filtrage permet d'améliorer la **robustesse des comparaisons** entre stations.

Une carte thermique (*heatmap*) a été produite afin de visualiser la complétion, rendant immédiatement visibles les stations incomplètes.

2.4 Étude des corrélations entre paramètres

Nous avons analysé les **corrélations entre paramètres physico-chimiques**, à partir des moyennes calculées pour chaque station. Une **matrice de corrélation de Pearson** a été générée et visualisée sous forme de carte thermique.

Les résultats ont mis en évidence plusieurs corrélations significatives :

- **Ammonium**, **Azote Kjeldahl** et **Phosphore total** : corrélation positive élevée → indicateurs liés aux rejets azotés.
- **Conductivité**, **Nitrates**, **Orthophosphates** : souvent corrélés entre eux, car associés à la pollution agricole ou urbaine.
- **Oxygène dissous** et **taux de saturation en oxygène** : corrélation très forte, validant la **cohérence interne des mesures**.

Ces corrélations confirment la **pertinence des regroupements thématiques** proposés par le référentiel SEQ-Eau, et orientent les prochaines analyses multivariées (ACP, clustering, etc.).

2.5 Intégration de l'indice I2M2 (mesures hydrobiologiques)

Pour compléter l'analyse physico-chimique, nous avons intégré les données issues du fichier `hb_ARMORICAIN.csv`, contenant les résultats d'indices hydrobiologiques.

En nous basant sur les documents officiels du site *Légifrance* (référence SEQ-Eau / SANDRE), nous avons identifié que l'indice I2M2 (Indice Invertébré Multimétrique) est codé sous le numéro `5910`. Cela nous a permis de filtrer les données pertinentes dans la colonne `CdParametreResultatBiologique`.

Une fonction de classification a ensuite été appliquée à la valeur mesurée (`ResIndiceResultatBiolo`) selon les seuils définis par le texte de loi :

- ≥ 17 $\Rightarrow$ Classe 1 (Très bon état)
- $13 \leq$ valeur < 17 $\Rightarrow$ Classe 2 (Bon état)
- $9 \leq$ valeur < 13 $\Rightarrow$ Classe 3 (État moyen)
- $5 \leq$ valeur < 9 $\Rightarrow$ Classe 4 (Mauvais état)
- < 5 $\Rightarrow$ Classe 5 (Très mauvais état)

Cela permet d'obtenir un **second indicateur indépendant** pour évaluer la qualité écologique des stations, qui sera comparé aux résultats physico-chimiques dans les prochaines étapes du projet.

Chapitre 3

TP3 : Agrégation temporelle et classification des stations

L'objectif de ce TP était d'agréger les données de qualité de l'eau issues des stations RCS, puis de réaliser une **classification non supervisée (clustering)** des stations selon leur profil physico-chimique, afin d'interpréter les grandes tendances de qualité dans les bassins étudiés.

3.1 Prétraitement et agrégation des données

Les données initiales comportaient des séries temporelles de mesures (paramètres chimiques) à différentes stations. Pour simplifier le problème, une agrégation temporelle a été nécessaire. Deux stratégies complémentaires ont été appliquées :

- **Agrégation par moyenne annuelle** : permet de représenter un état moyen de la qualité de l'eau.
- **Agrégation par percentile (95^{e})** : méthode utilisée pour capturer les *pics de pollution*, en mettant en avant les valeurs élevées souvent associées à des dépassements de seuils réglementaires.

Dans chaque cas, les valeurs agrégées ont été pivotées pour obtenir une table où :

- chaque **station** correspond à une **ligne**,
- chaque **paramètre physico-chimique** correspond à une **colonne**.

3.2 Clustering avec K-Means

L'algorithme **K-Means** a été appliqué aux deux jeux de données agrégées (moyenne et percentile). Le nombre de clusters a été fixé à $k = 5$, en cohérence avec les 5 classes de qualité de l'eau définies par le SANDRE.

Avant l'apprentissage, les données ont été **normalisées** à l'aide d'un `StandardScaler`, pour éviter que les paramètres à grande échelle ne dominent pas l'analyse.

3.3 Visualisation des clusters

Pour évaluer la qualité de la séparation des clusters et interpréter visuellement les résultats, deux techniques de réduction de dimension ont été utilisées :

- **PCA (Analyse en Composantes Principales)** : méthode linéaire rapide permettant une visualisation globale des axes de variation.
- **t-SNE (t-distributed Stochastic Neighbor Embedding)** : méthode non linéaire mieux adaptée à la préservation des structures locales.

Les visualisations montrent des **regroupements significatifs**, avec des différences notables entre l'approche par moyenne et celle par percentile. Cela met en évidence la **sensibilité du clustering au type d'agrégation** utilisé.

3.4 Analyse des clusters avec l'indice I2M2

Pour interpréter la pertinence environnementale des clusters, nous avons croisé les résultats avec l'indice **I2M2** (indice biologique basé sur les macroinvertébrés).

- Les scores I2M2 ont été extraits depuis `hb_ARMORICAIN.csv` via le code paramètre `5910`.
- Ces scores ont été associés à chaque station, puis transformés en classes de qualité selon les seuils définis par le SANDRE.
- Une **analyse statistique** (moyennes, boxplots) a permis de comparer les clusters selon leur I2M2 moyen.

Les résultats montrent que certains **clusters correspondent à des stations de haute qualité écologique** (I2M2 élevé), tandis que d'autres regroupent des stations plus dégradées. Cette validation par un indicateur biologique renforce la **robustesse écologique du clustering**.

Chapitre 4

TP4 : Visualisation géographique des clusters et lien avec l'environnement

4.1 Création des visualisations cartographiques

L'objectif principal de ce TP est d'évaluer si les classes de qualité obtenues par clustering peuvent être liées à des facteurs géographiques ou environnementaux. Pour cela, nous avons développé des visualisations cartographiques en croisant les résultats du clustering avec les coordonnées géographiques des stations et les données d'occupation du sol (CORINE Land Cover).

Dans un premier temps, les coordonnées des stations ont été extraites depuis le fichier `stations.csv` et associées aux résultats de clustering via une jointure. Une géométrie de type `Point` a ensuite été construite pour chaque station, puis convertie en `GeoDataFrame` en projection Lambert 93 (EPSG :2154), compatible avec les données CORINE.

Une carte statique a été produite en superposant :

— la couche d'occupation du sol issue du fichier `CLC12_FR_RGF.shp`, affichée en gris clair ;

— les stations, colorées selon leur **cluster d'appartenance**, en utilisant une palette catégorielle.

4.2 Choix du clustering visualisé

Pour cette première cartographie, nous avons choisi d'afficher les résultats du clustering basé sur les **valeurs moyennes** agrégées des paramètres physico-chimiques. Ce choix se justifie par la stabilité de cette agrégation vis-à-vis des valeurs extrêmes, et par son interprétation plus simple dans un contexte de comparaison spatiale.

4.3 Interprétation spatiale des résultats

L'analyse visuelle de la carte montre que les stations appartenant à un même cluster sont géographiquement regroupées. Certaines régions géographiques sont do-

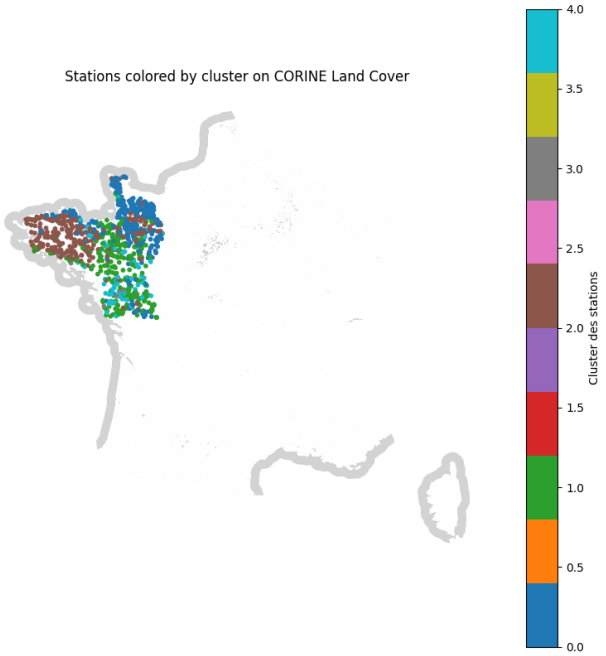

FIGURE 4.1 – Visualization de mean aggregation clustering

minées par un seul cluster, tandis que d'autres présentent une cohabitation de plusieurs classes.

Cette résultat suggère une influence possible de l'environnement local sur la qualité de l'eau. En effet, si les clusters étaient distribués aléatoirement sur le territoire, on pourrait considérer que les facteurs environnementaux ont peu d'impact. Or ici, on observe une **forte structuration spatiale**.

Cependant, cette observation ne permet pas de conclure de manière définitive. Plusieurs limitations doivent être prises en compte :

— L'algorithme K-Means n'utilise aucune information géographique : les regroupements géographiques ne sont donc pas forcés, mais peuvent résulter d'un biais latent (structure des données, effets régionaux, etc.).
— Nous n'avons pas encore analysé précisément les **types d'occupation du sol** ou les **bassins versants** associés à chaque station.
— L'effet de la densité des stations (très variable selon les régions) peut aussi introduire une fausse impression de regroupement.

Conclusion

Ce projet a permis de mettre en œuvre une chaîne complète d'analyse de données environnementales appliquée à la qualité de l'eau. En partant de fichiers bruts hétérogènes, nous avons réalisé un prétraitement structuré, une sélection de variables pertinentes selon le référentiel SEQ-Eau er SANDRE, une classification des mesures, et une exploration par clustering non supervisé.

L'intégration de l'indice biologique I2M2 a enrichi l'analyse en apportant un point de comparaison indépendant aux résultats physico-chimiques. Les visualisations cartographiques ont quant à elles permis de mieux comprendre la répartition spatiale des clusters et d'envisager des hypothèses sur l'influence de l'environnement.

Les résultats obtenus montrent une bonne cohérence entre les différentes sources de données et approches analytiques, et constituent une base solide pour des travaux futurs, notamment en intégrant plus finement les données d'occupation du sol ou les pressions anthropiques.

Ce travail illustre l'intérêt des méthodes de science des données dans une perspective environnementale, combinant rigueur technique et capacité d'interprétation territoriale.