Analyse des Données de Pollution des Stations de Transport.

GM, ING1
Cy-tech

Zaouche Djaouida

2025

Ce cahier des charges définit le cadre et les objectifs d'un projet d'analyse des données issues du dataset qualite-de-lair-dans-le-reseau-de-transport-francilien.csv, visant à exploiter ces informations pour mieux comprendre la pollution de l'air dans les stations de transport en Île-de-France et identifier des pistes d'amélioration pour une gestion environnementale optimale.

1 Traitement des données

1. Nettoyer efficacement le dataset, qualite-de-lair-dans-le-reseau-de-transport-francilien.csv, en supprimant les lignes vides, les colonnes inutiles et les doublons.
2. Serait-t-il intéressant d'appliquer une ACP à ce dataset ? Justifier votre réponse.
3. Filtrer le dataset en ne conservant que les stations de métro. À partir de ce dataset filtré, construire deux fichiers CSV : train.csv (70 %) et test.csv (30 %).
4. Construire un modèle kmeans pour prédire le niveau de pollution des stations métro.
5. Construire un autre modèle basé sur les k-plus-proches voisins.
6. Èvaluez vos modèles.
7. À partir du dataset, concevez un dashboard permettant de fournir des indicateurs pertinents. Justifiez l'intérêt de ce dashboard et précisez les acteurs auxquels il est destiné.
8. Faites des visualisations et des analyses des données de pollution des lignes de métro avec Qgis. Interpréter les visualisations de Qgis.

2 Modélisation des stations de transport par un graphe

1. Construire un fichier CSV contenant trois colonnes : $'station'_1$, $'station'_2$ et un booléen. Ce dernier sera mis à vrai s'il existe au moins une ligne indiquant un trajet de $'station'_1$ vers $'station'_2$. Utilisez ce fichier pour construire un graphe des connexions entre les stations du métro de Paris. Donnez la taille du fichier CSV. Peut-on réduire cette taille ?
2. Étant données deux stations 1 et 2, comment trouver un chemin partant de la station 1 vers la station 2, qui minimise l'exposition à la pollution tout en respectant une contrainte de temps de trajet minimal ?
3. Étant données deux stations, comment vérifier l'existence d'un chemin entre elles où toutes les stations sur le trajet respectent un seuil de pollution maximal ?
4. Proposez une méthode algorithmique pour détecter l'existence de cycles dans le graphe des stations du Métro de Paris. Justifiez votre choix d'algorithme et calculer sa complexité. Quels sont les intérêts pratiques, s'il y en a, de la détection des cycles pour notre cas d'étude ?

3 Analyse du signal sur graphes pour la modélisation de la pollution des stations de transport

Dans cette partie, nous souhaitons appliquer une analyse spectrale aux données de pollution du métro parisien en utilisant notre dataset. Vous devez aborder l'optimisation des calculs de la matrice laplacienne et vous concentrer sur la propagation spatiale de la pollution dans le réseau, en identifiant les zones les plus exposées ainsi que les éventuels points de concentration. Les résultats permettent de mieux comprendre la propagation de la pollution et d'optimiser les stratégies de filtration et d'aération. de filtration et d'aération.

4 Rendu final

Constituez-vous en groupes de 4 ou 5 élèves. Remettez un rapport détaillé présentant vos solutions, vos choix, vos hypothèse et vos résultats. Ajoutez également un lien GitHub vers votre code.

5 Remarques importantes

1. Vous pouvez enrichir ou modifier le jeu de données, à condition de justifier clairement vos changements. Pour les questions qui semblent trop générales ou incomplètes, n'hésitez pas à ajouter des hypothèses. Chaque modification ou décision doit être justifiée.
2. La partie référence bibliothèque doit ête assez riche pour couvrir les fonctionalités de ce projet.
3. Attention, ne confiez pas ce projet aux IA ; par exemple, l'utilisation de codes fournis par les IA sera sanctionnée.
4. Attention au plagiat et copiage.
5. Votre encadrant contrôlera le taux de plagiat ainsi que ainsi que l'utilisation d'outils d'intelligence artificielle.

6 Référence

1. Chung, Fan R. K. (1997). Spectral Graph Theory. American Mathematical Society. Von Luxburg, Ulrike (2007). "A tutorial on spectral clustering." Statistics and Computing, 17(4), 395–416. Springer.
2. Newman, Mark (2018). Networks. Oxford University Press.
3. Coifman, Ronald R., & Lafon, Stéphane (2006). "Diffusion maps." Applied and Computational Harmonic Analysis, 21(1), 5–30. Elsevier.
4. Broday, David M., & Alpert, Pinhas (2017). "A spectral graph approach to near real-time monitoring of urban air pollution dispersion." Science of the Total Environment, 574, 738–746. Elsevier.