

Visualisation interactive de la pollution de l'air en France

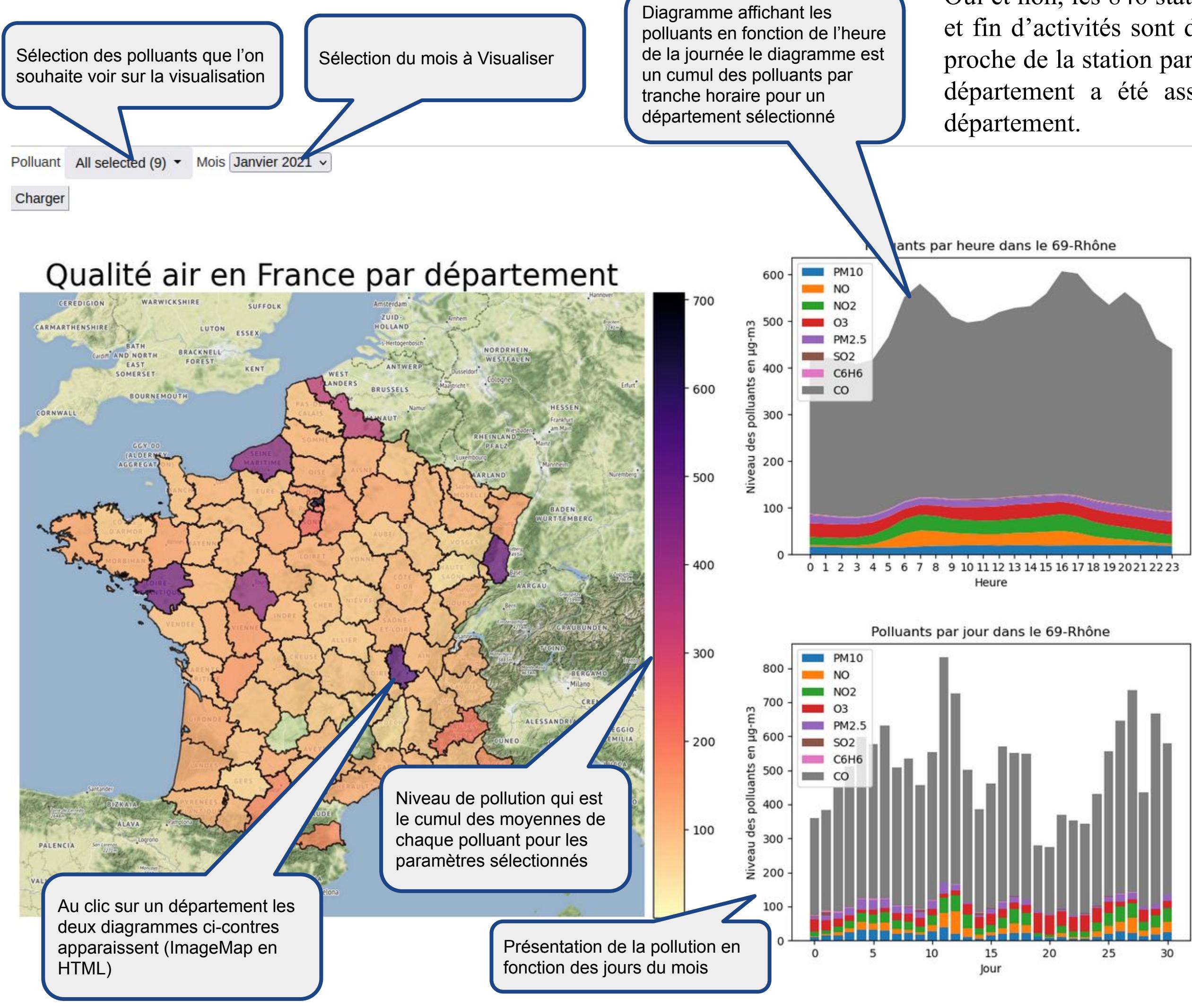
TU Kuan-Hua (M2 Bioinformatique Université Lyon 1)

Année : 2022-2023

Sources de données

Les données sont issues de data.gouv.fr. Il a été choisi le jeu de données : Données « temps réel » de mesure des concentrations de polluants atmosphériques réglementés. La quantité des données est de 7.1 giga-octet. L'AASQA est un ensemble de 18 associations agréées qui ont la responsabilité gouvernementale de la surveillance de la qualité de l'air ambiant. Différentes stations de contrôle de l'air sont dispersées en France et gérées par ces associations. Les 6 polluants atmosphériques principaux sont mesurés: Ozone (O₃), Dioxyde d'azote (NO₂, NOX), Dioxyde de soufre (SO₂), Particules de diamètre inférieur à 10 µm (PM10), Particules de diamètre inférieur à 2,5 µm (PM2.5), Monoxyde de carbone (CO). Les données sont les mesures de chaque sondes pour chaque heure de la journée.

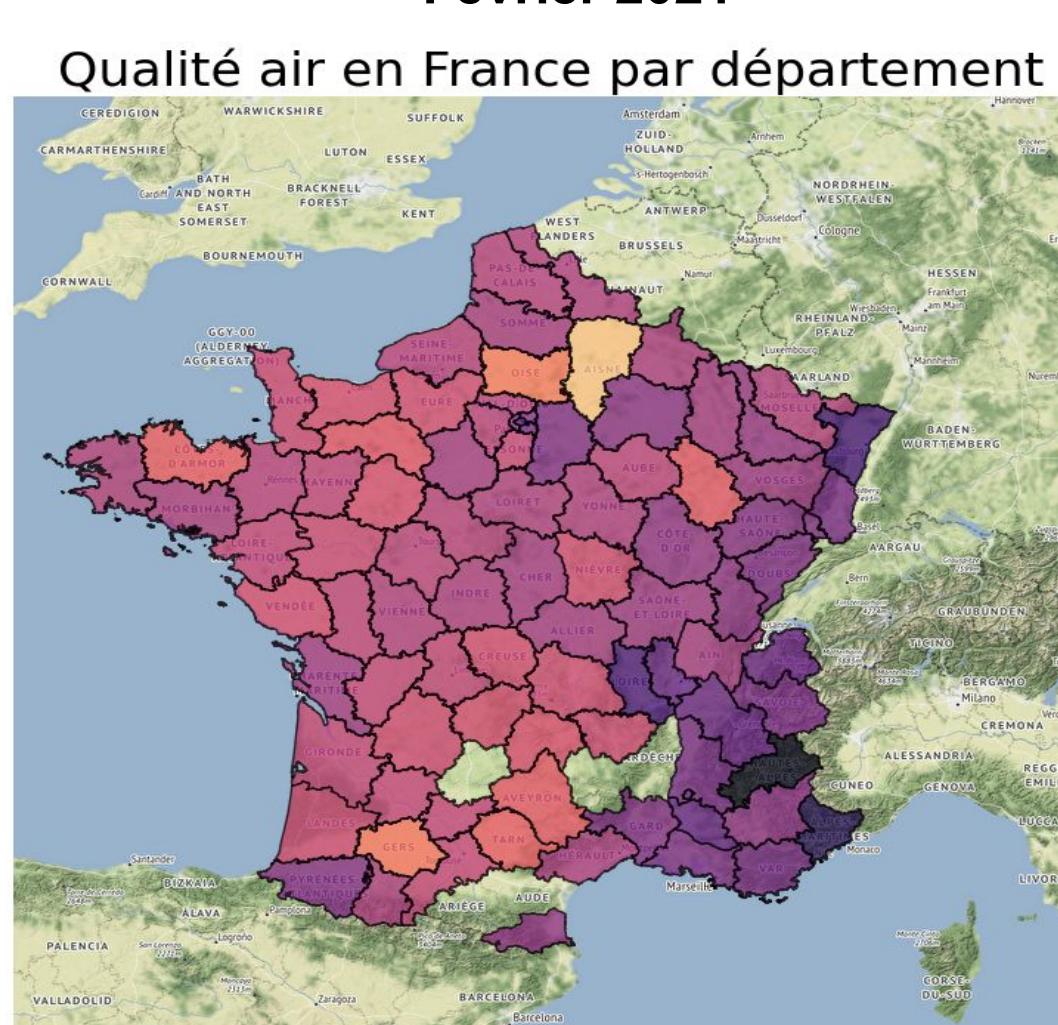
Visualisation interactive de données



Existe-t-il des départements surexposés à la pollution ?

Pour répondre à cette question il a été décidé de se fixer sur un polluant dans le but de comparer équitablement les régions. En effet comme vu précédemment certains polluants ne sont pas mesurés dans tous les départements. La figure 7 présente le taux de polluants PM10 (Particule fine de 10µm) en Février 2021, Avril 2021 et Juillet 2021. On remarque plus de polluants dans les Alpes en février, sûrement à cause du ski. Plus de pollution dans le nord de la France en avril et plus de pollution en région Parisienne et sur la côte d'Azur en Juillet.

Février 2021



Avril 2021



Juillet 2021

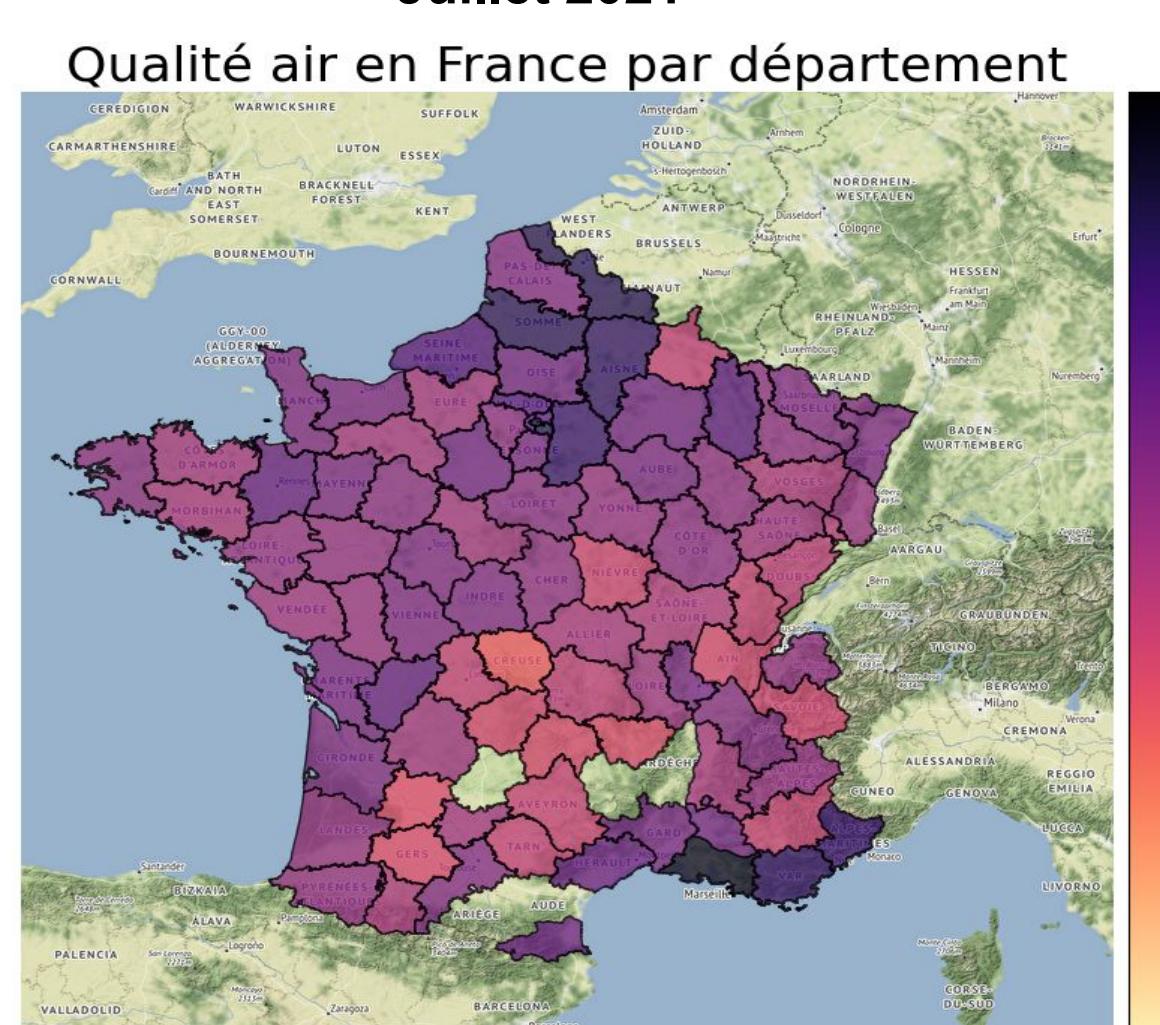
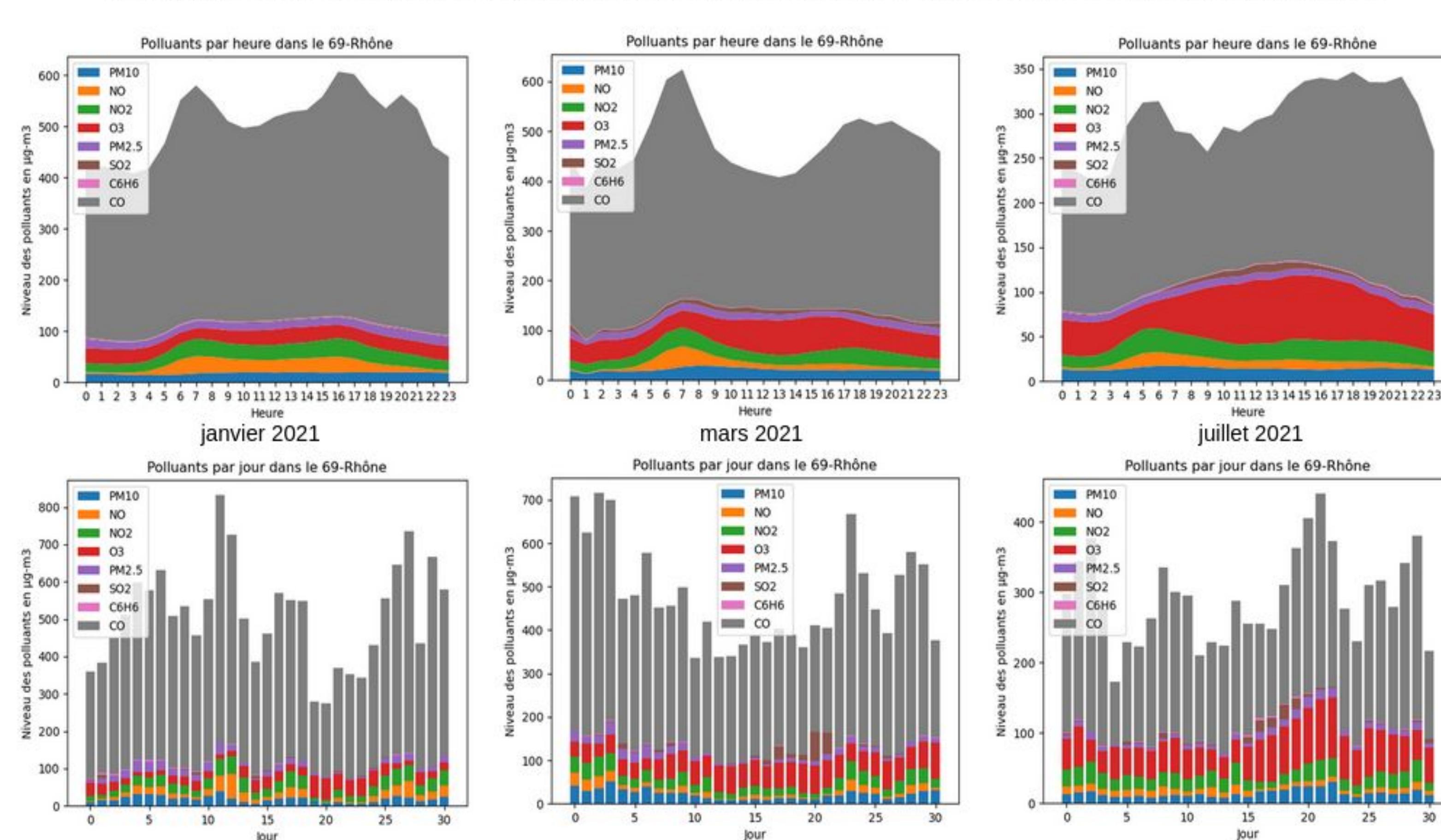


Figure 7 - Visualisation de la pollution au PM10 en Février 2021, Avril 2021 et Juillet 2021

La pollution de l'air est-elle la même à chaque moment de la journée ou du mois ?

La figure 9 nous montre la répartition des pollutions dans le département du Rhône par heure et par jour. On constate ici une pollution au milieu de la journée, en effet la pollution est liée à l'heure de travail, les activités des citoyens. La figure 10 nous montre la pollution pour différents polluants comme (1) O₃, (2) NO, NO₂, (3) CO, (4) PM2.5 et PM10, (5) C6H6 et (6) SO₂ pour le département du Rhône. Pour un même mois, il est présenté la pollution à l'heure et à la journée. On constate une pollution en moyenne plus importante le matin et le soir. Certains polluants sont plus présents en moyenne aux heures de trajet pour le travail.

Comparaison des niveaux des polluants pour le département 69 en janvier, mars et en juillet 2021



Conclusion et Perspectives

Il a été proposé une application pour la visualisation des polluants atmosphériques. Cette application propose la visualisation et l'exploration de plus de 34 millions de mesures sur 2 ans. Cette application est fonctionnelle et propose à l'utilisateur la sélection de polluants et la visualisation du cumul des polluants par heure et par jour. Il serait intéressant d'implémenter cette application avec un Framework comme Leaflet et des web services renvoyant les données en JSON.

Problématiques et Objectifs de la Visualisation

Les études récentes tendent à montrer un lien fort entre cancer du poumons et particules fines dans l'air. La qualité de l'air est importante pour tous. Le jeu de données du gouvernement est une opportunité de répondre à quelques problématiques : (1) Mesure-t-on la qualité de l'air de manière identique sur tout le territoire Français ? (2) La pollution est-elle la même à différents moments de la journée ou du mois ? (3) Existe-t-il des départements surexposés à la pollution ? La visualisation proposée présente une vue globale de la France puis par un clic sur un département il est présenté une vue spécifique au département cliqué.

Mesure-t-on la qualité de l'air de manière identique sur tout le territoire Français ?

Oui et non, les 846 stations sont répartis sur plusieurs départements, leurs dates de début et fin d'activités sont différentes. Le travail effectué a été de retrouver la ville la plus proche de la station par ses coordonnées spatiales. Une fois que la ville a été trouvée, le département a été associé à la station. Ainsi pour chaque mesure on obtient le département.

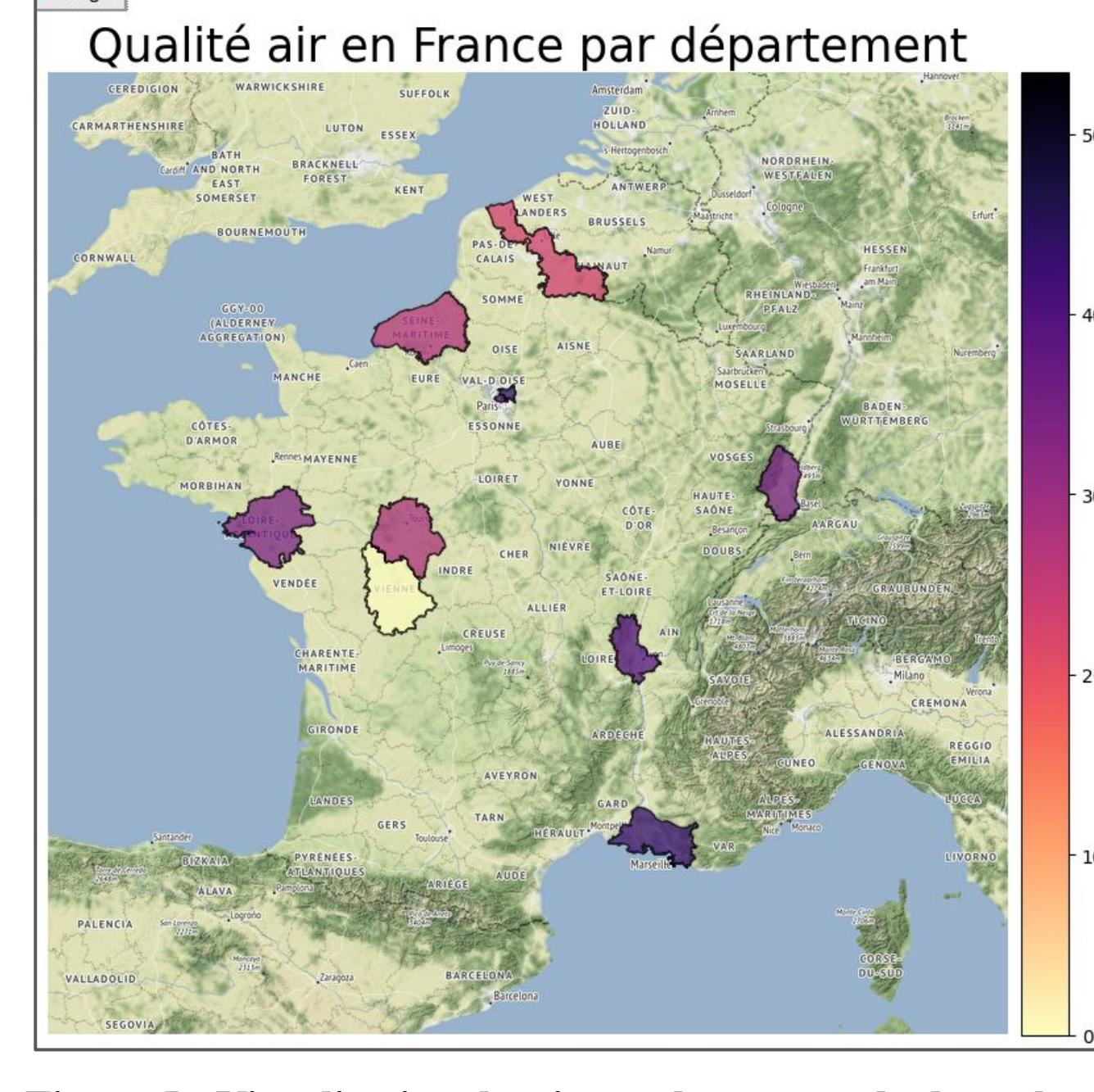


Figure 5 - Visualisation du niveau de monoxyde de carbone (CO) dans l'air en Janvier 2021

On constate que certains départements n'ont pas de mesure à une date donnée. En figure 4, on a l'impression que les départements en violet sont plus pollués que les autres. La figure 5 nous montre que seuls quelques départements sont mesurés pour le dioxyde de carbone. La figure 4 nous donne donc une fausse impression de département plus pollué alors que ce sont des départements contrôlés sur le dioxyde de carbone.

Intégration des données

D'un côté 688 fichiers de mesures sont fournis par l'Open Data pour 23 mois de données. Ce sont des mesures au format CSV donnant pour chaque jour les mesures des différentes stations par heure. Talend est un ETL (Extract Transform and Load) utilisé pour effectuer l'intégration de données, il se base sur du Java. Il a été implémenté 4 pipelines dans Talend : (1) L'intégration des mesures, (2) l'intégration des stations, (3) l'intégration des villes et département, (4) Le rapprochement des stations au ville pour établir leur département. La Fig. 1 présente un exemple de pipeline pour le rapprochement des stations aux régions.

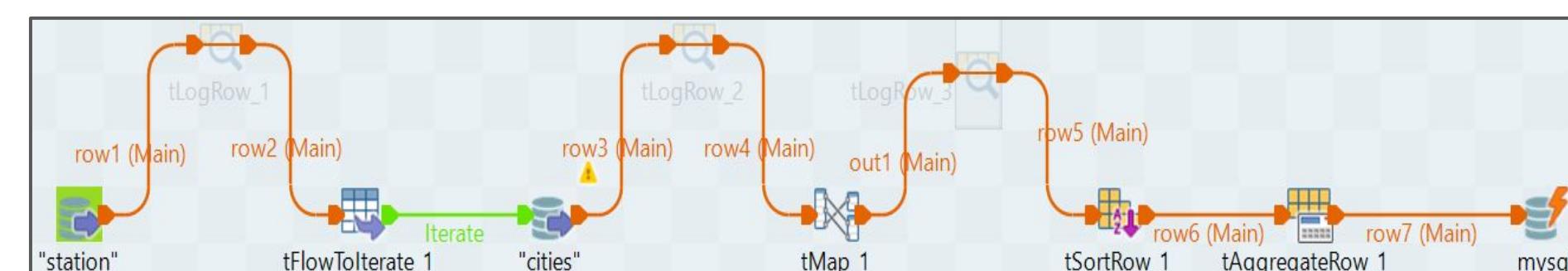


Figure 1 - Pipeline Talend pour le rapprochement des stations aux départements

Base de données

Une base de données MySQL est réalisée pour les requêtes des quantités des polluants dans les différents départements. Une table mesure contient tous les détails de mesures des polluants, une ligne de cette table est la mesure d'un polluant par une station à une heure donnée. Une table station est créée, elle est associée avec les données des Département ville et région. Un MCD de la base de données est réalisé (cf. Fig. 2).

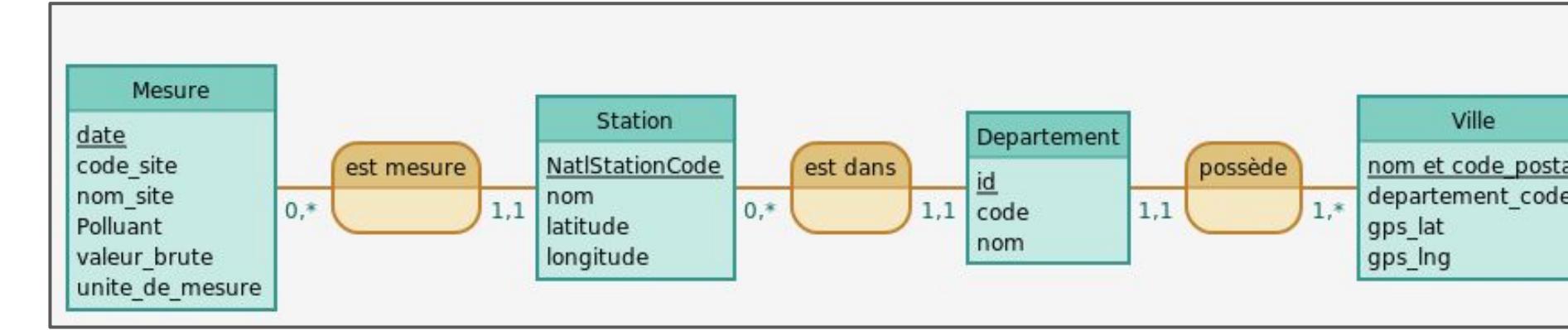


Figure 2 - Modèle Conceptuel de Données de la base de données

Structure de l'application

L'application est en 3-tiers (Client, serveur et base de données). Le client est en HTML, un peu de CSS et JS. Le serveur est en Python Flask et la base de données en MySQL. Les diagrammes sont générés avec Matplotlib et GeoPandas pour la cartographie.

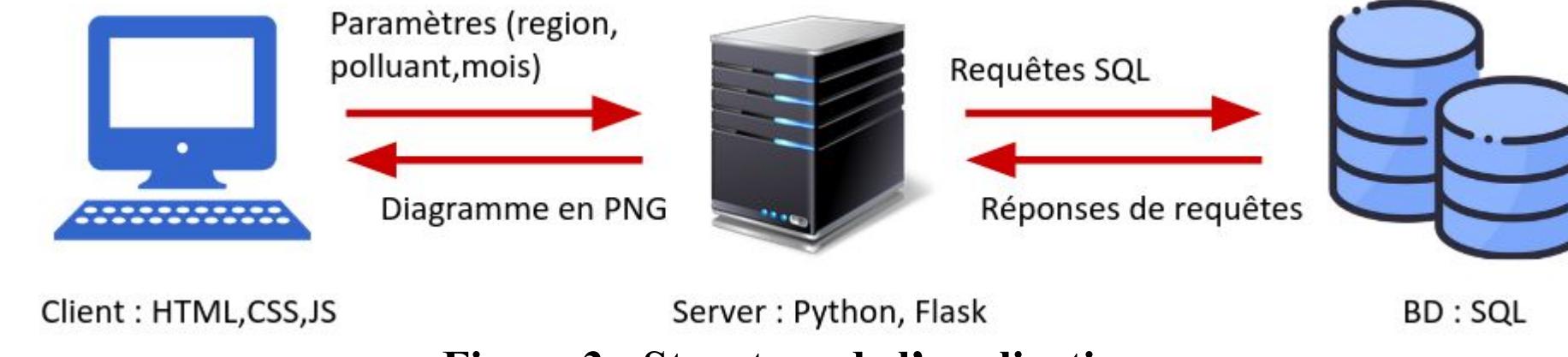


Figure 3 - Structure de l'application

Interactions

L'application permet l'interaction pour appliquer des filtres de recherche. Il est possible de cumuler différents polluants par le sélecteur multiple et de sélectionner un mois précis. Au clic sur le département le diagramme affichant le détail de la pollution par heure et par mois apparaît. La figure 6 présente différents moyens d'interaction. Le click sur l'image s'effectue par des ImageMap HTML. Un clic sur le bouton treemap permet la visualisation des polluants par région et département. Cette visualisation change en fonction des filtres sélectionnés. La figure 8 est une visualisation de treemap sur une répartition des polluants par rapport aux régions et aux départements.



Figure 6 - Sélection de Filtres pour la visualisation d'information



Figure 8 - Visualisation de treemap de la répartition des polluants sur les différentes régions et les différents départements

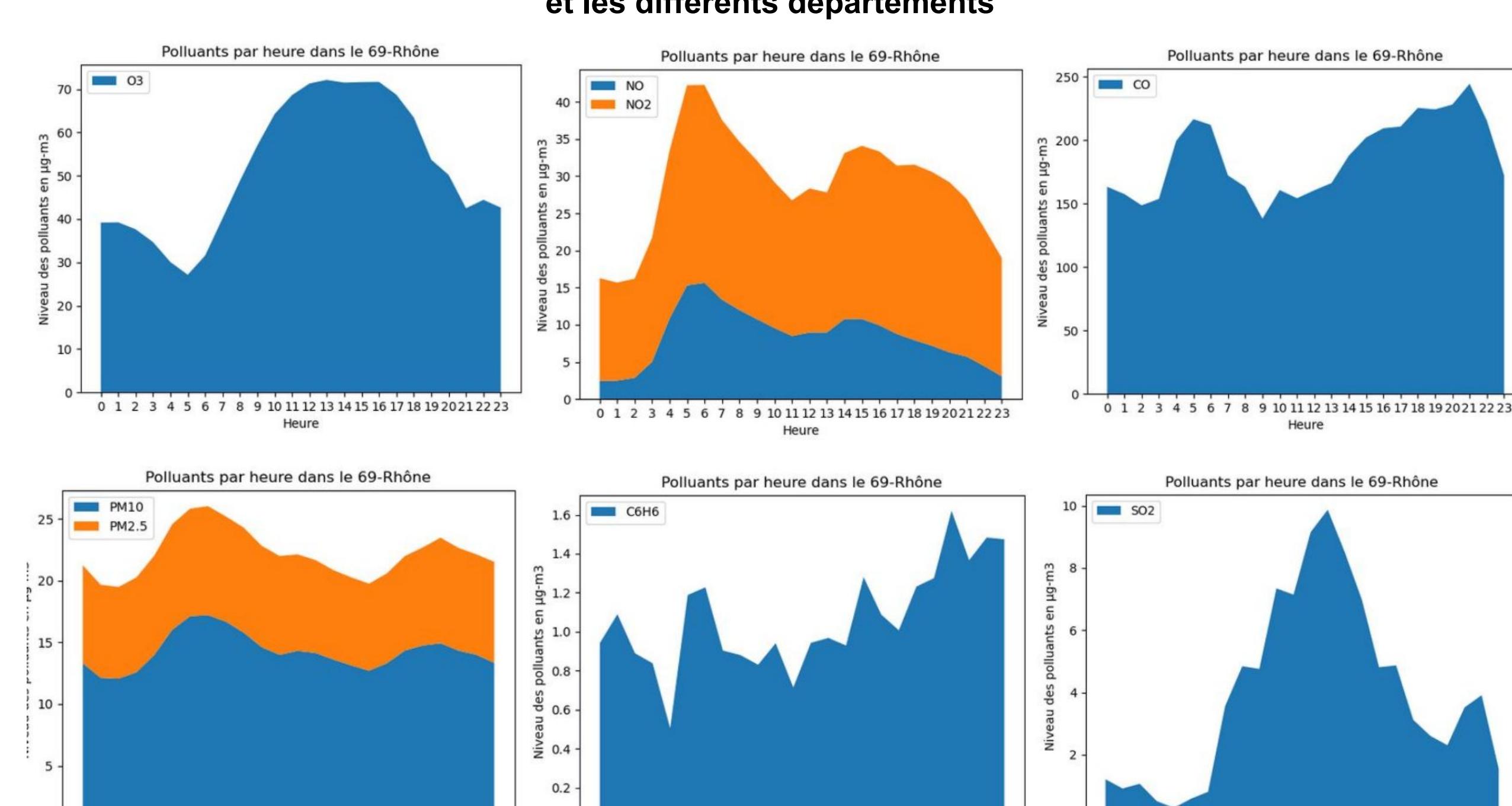


Figure 10 - Visualisation de différents groupes des polluants en Juillet 2021 dans le Rhône