Matéo IORI
IG5 Promotion 2024-2025
Département Informatique et gestion

LIRMM
Montpellier

Stage de fin d'études
Département Informatique et gestion
Effectué du 1 avril 2024 au 29 août 2025

Stage de recherche en science des donnés - Détection et visualisation de motifs

Professeur Polytech

Christophe FIORIO

Tuteur entreprise

Arnaud SALLABERRY

Professeur

Table des matières

1. Glossaire

Machine Learning

Branche de l'intelligence artificielle qui permet à des algorithmes d'apprendre à partir de données. Dans mon rapport, je me sers de machine Learning pour faire des regroupements utiles à l'interprétation des contaminations aux PFAS. – Retour au rapport

Unité mixte de recherche

Une UMR (Unité mixte de recherche) est une structure administrative réunissant des chercheurs dans le cadre d'un partenariat entre le CNRS et un laboratoire ou un organisme de recherche, comme l'université par exemple. – Retour au rapport

Docker Compose

Outil permettant de lancer facilement un ensemble de services (comme un backend, une base de données, un frontend). – Retour au rapport

Frontend

Le frontend fait référence à la partie visible et interactive d'une application ou d'un site web, celle avec laquelle les utilisateurs interagissent directement. – Retour au rapport

Backend

Le backend fait référence à la partie serveur d'une application ou d'un site web. Il gère la logique de l'application, les bases de données, l'authentification des utilisateurs, et d'autres fonctions cruciales qui ne sont pas visibles par l'utilisateur final. – Retour au rapport

API

Outil permettant à deux logiciels de communiquer. Le frontend interroge et récupère les données depuis la base de données en passant par l'API. – Retour au rapport

Framework

Un framework est une structure de travail qui fournit un environnement de développement standard pour une application logicielle. Il inclut des bibliothèques de code réutilisables et des outils permettant de simplifier le développement de logiciels. – Retour au rapport

Design pattern

Les design patterns sont des solutions classiques à des problèmes récurrents de la conception de logiciels. Chaque design pattern est une sorte de plan ou de schéma que l'on peut personnaliser afin de résoudre un problème dans notre code. – Retour au rapport

Observable

C'est un exemple de design pattern, qui repose sur un mécanisme de notification : lorsqu'une donnée change, tous les éléments qui y sont liés (par exemple dans une interface utilisateur) sont automatiquement mis à jour. Ce pattern est particulièrement utile pour construire des interfaces interactives et réactives. – Retour au rapport

Écarts interquartiles

Méthode statistique utilisée pour identifier les valeurs aberrantes (outliers). Elle se base sur l'écarts entre le premier quartile et le troisième (Q3 - Q1) de la répartition de données. On ne garde que les valeurs dans l'intervalle [Q1 - 1.5 x (Q3-Q1) ; Q3 + 1.5 x (Q3-Q1)] – Retour au rapport

Clustering

Méthode qui regroupe des données similaires, par zone géographique, ou par attributs semblables par exemples. – Retour au rapport

ETL (Extract, Transform, Load)

Processus en trois étapes permettant de préparer les données : extraction, transformation (nettoyage, formatage) et chargement. – Retour au rapport

Interpolation

Technique mathématique pour estimer des valeurs intermédiaires entre des points connus. Elle peut être utilisée pour améliorer les cartes ou visualiser des zones non mesurées. – Retour au rapport

SHAP (SHapley Additive exPlanations)

Méthode pour expliquer les prédictions d'un modèle de machine Learning en quantifiant l'impact de chaque variable sur la classification finale. – Retour au rapport

Méthode agile

Méthodologie de travail en équipe souple basée sur la collaboration, l'adaptation et des itérations courtes (sprints). – Retour au rapport

Sprints

Périodes courtes (souvent une semaine ou deux) au cours desquelles une équipe réalise un ensemble de tâches bien définies à l'avance. – Retour au rapport

UI (User Interface)

Ensemble des éléments visuels avec lesquels l'utilisateur interagit dans l'interface. Dans mon premier projet, l'UI est essentielle pour permettre aux chercheurs de filtrer, explorer et comprendre les données. – Retour au rapport

Thèse CIFRE

Dispositif qui permet de financer une thèse en entreprise, en collaboration avec un laboratoire public. – Retour au rapport

2. Table des figures

3. Remerciements

Je tiens à exprimer mes sincères remerciements à toutes les personnes qui contribuent à la réussite de mon stage au sein de l'équipe ADVANSE du LIRMM. Leur soutien, leur expertise, leurs encouragements sont précieux et enrichissent grandement mon expérience professionnelle.

Je souhaite tout d'abord remercier mon tuteur sur le projet, M. Alexis GUYOT, pour sa supervision et ses conseils tout au long de mon expérience. Son envie de partager, son professionnalisme et sa disponibilité, malgré son emploi du temps chargé, sont d'une grande aide dans la réalisation de ce stage. Son implication et son sérieux seront une inspiration pour mon projet professionnel.

Je tiens également à exprimer ma reconnaissance envers mon tuteur dans l'équipe ADVANSE M. Arnaud SALLABERRY, permanent de l'équipe. Sa confiance et ses propositions sont des moteurs essentiels dans la réussite des projets.

Mes remerciements vont également à toute l'équipe du département pour leur accueil chaleureux. M. Vincent RAVENEAU et M. Pascal PONCELET contribuent au bon déroulement de ce stage par leur aide et leur gentillesse, en me donnant des conseils et en me proposant des sujets de thèses susceptibles de m'intéresser. Leur esprit d'équipe et leur bienveillance créent un environnement de travail convivial dans lequel je me sens à l'aise.

Je souhaite également exprimer ma gratitude envers Mme. Nadine JACQUET et Mme. Françoise FOURCADIER du département administratif pour leur aide dans les démarches complexes et la planification des déplacements.

Je remercie l'ensemble du personnel de l'entreprise pour leur amabilité et leur coopération. Leur accueil chaleureux et leur ouverture d'esprit rendent mon stage agréable et enrichissant.

Enfin, je tiens à remercier mon professeur Polytech M. Christophe FIORIO ainsi que ma famille pour la relecture de ce rapport.

4. Introduction

Mon stage de cinquième année se déroule au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) situé au nord de la ville. Ce laboratoire est reconnu pour la diversité de ses activités de recherches. Il prône des valeurs d'éthique, de durabilité et d'inclusion à travers différentes initiatives telles que la science ouverte [26], la transition écologique (Green LIRMM) [27] et la promotion de l'égalité et de la parité [28].

Mon objectif était de découvrir le monde de la recherche, et le LIRMM, par sa proximité, ses valeurs et le fait que de nombreux professeurs de Polytech Montpellier y sont chercheurs, s'imposait comme un choix naturel. J'ai été accueilli dans un environnement stimulant et collaboratif, en accord avec mes intérêts pour la science des données.

J'ai eu la chance d'intégrer l'équipe qui m'intéressait après une candidature spontanée et un entretient en présentiel. Cette expérience se révèle extrêmement formatrice et influencera pour sur mes choix professionnels futurs.

Cette équipe, c'est l'équipe ADVANSE (ADVanced Analytics for data SciencE), elle est spécialisée dans l'analyse de données complexes. Placé sous la supervision d'Alexis Guyot (post-doctorant) et d'Arnaud Sallaberry (enseignant-chercheur permanent), j'ai pu contribuer à un projet qui vise à développer des outils d'analyse et de visualisation pour mieux comprendre la pollution environnementale liée aux substances per- et polyfluoroalkylées (PFAS).

Mes travaux ont débuté par la refonte et l'amélioration d'un outil existant permettant le filtrage et la visualisation de données liées aux PFAS. Le projet en cours consiste à utiliser le Machine Learning pour mieux comprendre les contaminations à ces polluants en Europe.

Ce rapport a pour objectif de fournir une vision globale des missions que j'ai menées, de réfléchir sur les résultats obtenus, les choix méthodologiques effectués, ainsi que d'évaluer les compétences sollicitées et développées. Il s'agira également de mettre en lumière les aspects perfectibles et les axes d'amélioration identifiés au cours du stage.

Il est structuré en plusieurs sections : une analyse du contexte avec une présentation du laboratoire, de l'équipe et des enjeux du stage ; un développement détaillé des missions effectuées, suivi d'une analyse critique. En conclusion, je dresserai un bilan personnel de cette expérience, avec des perspectives d'évolution professionnelle. Cette structure permettra de couvrir l'ensemble des aspects significatifs de mon expérience de stage, tout en répondant aux exigences académiques du rapport.

5. Analyse du contexte et des enjeux

Dans cette section, nous allons explorer le contexte ainsi que les enjeux de mon expérience au sein du laboratoire. Nous commencerons par une présentation détaillée du LIRMM, en mettant en lumière son histoire son organisation et son rayonnement. Ensuite, nous nous concentrerons sur l'équipe ADVANSE au sein de laquelle j'ai évolué. Puis nous présenterons le projet auquel j'ai contribué et nous verrons mon rôle dans celui-ci. Enfin, nous analyserons mes conditions de travail, en abordant l'environnement dans lequel j'ai évolué, l'encadrement mis en place et les outils utilisés.

5.1 Présentation du LIRMM

5.1.1 Histoire

Le laboratoire a été créé en 1992, résultant de la fusion de deux entités du CNRS présentes sur le site de Montpellier : le CRIM (Centre de Recherche en Informatique de Montpellier) et le LAMM (Laboratoire d'Automatique et de Microélectronique de Montpellier). Cette initiative visait à regrouper et stimuler les synergies entre les chercheurs en informatique, microélectronique et robotique, sous une seule structure pluridisciplinaire au service de la recherche et de l'innovation. Dès ses débuts, le laboratoire réunit environ 200 membres, chiffre qui doublera en 30 ans pour atteindre aujourd'hui plus de 400 personnes, dont plus de 160 chercheurs et enseignants-chercheurs, une quarantaine de personnel administratif et technique et plus de 150 doctorants et post doctorants [3].

5.1.2 Présentation

Situé sur le campus Saint-Priest à Montpellier, Le LIRMM est dirigé depuis 2023 par Mme. Marianne HUCHARD, Professeur en informatique. C'est une unité mixte de recherche placée sous la double tutelle de l'Université de Montpellier (UM) et du Centre National de la Recherche Scientifique (CNRS) [1] [2]. Il bénéficie également de partenariats avec l'Université Paul-Valéry Montpellier 3 (UPVM), l'Université de Perpignan Via Domitia (UPVD) et l'Institut National de Recherche en Informatique et en Automatique (INRIA). Le laboratoire aspire à être à la pointe de l'innovation dans ses trois domaines, comme le suggère Philippe POIGNET, ancien directeur :

> *"L'objectif du LIRMM est de créer et d'innover dans le domaine de l'informatique, de la robotique et de la microélectronique. De concevoir les nouveaux algorithmes en intelligence artificielle et en analyse de données. De développer de nouveaux logiciels, concevoir de nouveaux robots et algorithmes pour les piloter. Également concevoir de nouveaux circuits intégrés et systèmes embarqués de demain." - Sur le site officiel, consulté en juin 2025.*

Fort de ses 30 ans d'existence, le LIRMM s'est imposé comme un acteur majeur de la recherche scientifique en France et en Europe, notamment grâce à sa capacité à mener des projets interdisciplinaires et à transférer ses innovations vers le monde socio-économique.

5.1.3 Rayonnement

Le laboratoire se distingue aujourd'hui par la création de start-ups issues de ses travaux, telles que *AcuSurgical* [29], spécialisée dans la robotique chirurgicale pour les opérations de la rétine, ou *Algodone* [30] – développée par le directeur de Polytech, M. Lionel TORRES – qui propose des solutions innovantes de traçabilité pour sécuriser les circuits électroniques [9]. Le laboratoire rayonne aussi à l'internationale par ses nombreux projets de recherche financés à l'échelle européenne [31], tels que "Trustworthy AI for CCAM" [32] (intelligence artificielle de confiance pour les véhicules sans conducteur) et "GUARDEN" [33] (Préserver la biodiversité et les services écosystémiques essentiels dans tous les secteurs et à toutes les échelles), tous deux soutenus dans le cadre du programme Horizon Europe [34]. Ces sont des exemples d'excellence scientifique reconnue, avec des publications dans des conférences et revues internationales, attestant de la portée technologique incarnée au LIRMM.

5.1.4 Les principales activités

Le laboratoire est logiquement structuré en trois départements scientifiques : Informatique, Robotique et Microélectronique, couvrant un large spectre de thématiques allant des fondements théoriques aux applications industrielles.

- Les recherches du département Robotique se concentrent sur la robotique médicale et de manipulation, la robotique sous-marine, la perception, l'interaction physique homme-robot ou encore la robotique humanoïde. Ces travaux trouvent des applications directes dans le monde de la médecine, de l'environnement, de l'industrie et aussi du quotidien humain.
- Le département Microélectronique vise à concevoir des systèmes électroniques intégrés toujours plus intelligents en mettant l'accent sur l'efficacité énergétique, la sécurité et la fiabilité. Ses recherches s'inscrivent dans le champ des objets communicants appliqués à l'environnement, y compris les environnements complexes (spatial, radiatif, haute température, vivant).

- Le département informatique couvre un large éventail de recherches, fondamentales et appliquées. Ses travaux portent sur les fondements du calcul, la science des données et l'intelligence artificielle, la bioinformatique, le génie logiciel et l'analyse d'image. Ces recherches entrent dans des domaines extrêmement variés et les domaines d'applications sont infinis. Le département entretient des partenariats avec des hôpitaux, start-ups, entreprises régionales ou grands groupes. Mon responsable « entreprise », M. Arnaud SALLABERRY est le responsable adjoint de ce département.

Le département d'Informatique est composé de 15 équipes [6], celle qui m'a particulièrement intéressé et où j'effectue mon stage est l'équipe ADVANSE. Nous la présentons dans la partie suivante.

5.2 Présentation de l'équipe ADVANSE

L'équipe ADVanced Analytics for data SciencE (ADVANSE) [7] est l'une des équipes de recherche du département d'Informatique du LIRMM. Ses travaux s'inscrivent dans le domaine de l'analyse des grandes bases de données, avec pour objectif l'extraction de nouvelles connaissances à partir de données complexes et hétérogènes. Les activités de recherche de l'équipe ADVANSE se déclinent en trois axes principaux :

- Exploration de données (Data Mining) : développement de méthodes pour l'extraction de savoir, de connaissance, de motifs répétitifs à partir de grandes quantités de données.
- Visualisation analytique : conception d'interfaces visuelles facilitant le raisonnement analytique, permettant aux utilisateurs d'explorer et d'interpréter des données complexes facilement pour soutenir la prise de décision.
- Apprentissage automatique (Machine Learning) : utilisation et élaboration de modèles d'intelligence artificielle pour donner aux ordinateurs la capacité d'apprendre à partir des données. L'accent est mis sur l'explicabilité et l'architecture des modèles, qui sont des enjeux de recherche dans l'ère du temps.

L'équipe ADVANSE adopte une approche interdisciplinaire, combinant des travaux théoriques et expérimentaux, et collabore avec divers partenaires académiques et industriels. Elle est dirigée par le Professeur Pascal PONCELET, avec pour adjointe la Professeure Sandra BRINGAY.

Je vais à présent présenter le projet auquel j'ai activement contribué.

5.3 Le projet DAE-PFAS

Le projet DAE-PFAS, au cœur de mon stage, s'inscrit dans une dynamique de recherche interdisciplinaire mêlant science des données et enjeux environnementaux liés aux PFAS. Une annexe présentant ces composés chimiques et leurs impacts est fournie en fin de rapport [lien]. Sa lecture est recommandée avant de poursuivre, afin de mieux comprendre le contexte de la suite du rapport.

5.3.1 Historique du projet

Le *Forever Pollution Project (FPP)* [17] est une initiative journalistique européenne menée par le journal *Le Monde* publié en 2023. Ce projet représente un travail colossal de fouille et d'homogénéisation des données. Il a permis de cartographier la contamination aux PFAS à travers l'Europe en regroupant des milliers d'informations gouvernementales, industrielles et sanitaires. À partir de ces travaux, le CNRS a initié le projet PFAS Data Hub (PDH), qui a pour but de continuer à faire vivre ces données en les rendant accessible et en les alimentant régulièrement. Ce projet regroupe des data journalistes, des géologues et des chimistes, mais pas de vrai développeur ou analyste de données.

Mes travaux sont entièrement basés sur cette base de données, nous y ferons par la suite référence par les termes « base de données du PDH » ou « données du PDH ».

5.3.2 Contexte du projet

MaDICS [19] (Masses de Données, Informations et Connaissances en Science) est un Groupement de Recherche (GDR) du CNRS créé en 2015, visant à promouvoir les recherches interdisciplinaires en Sciences des Données. Il sert de forum d'échanges et de soutien pour les chercheurs et les acteurs externes (industriels, médias, culturels) confrontés aux enjeux du Big Data. Les activités de MaDICS se structurent autour d'Actions, qui durent entre deux et quatre ans.

Mon stage s'inscrit dans le cadre de l'action MaDICS nommée « Détections d'Anomalies Environnementales » (DAE) [18] [20] portée par Mme. Lylia ABROUK. C'est au sein de cette action qu'a vu le jour le projet DAE-PFAS visant la détection d'anomalies environnementales liées aux PFAS.

5.3.3 Naissance du projet DAE-PFAS

C'est en découvrant le projet PDH du CNRS par l'intermédiaire du chimiste M. Pierre LABADIE, impliqué dans celui-ci, que Mme. Lylia ABROUK a mis en place un projet de recherche interdisciplinaire à fort potentiel : DAE-PFAS. Ce projet est une extension directe

du PDH et lui ajoute une dimension, en le mettant en relation avec des experts en analyses de données. Les ambitions sont maintenant plus grandes, DAE-PFAS vise à exploiter les données du PDH pour faire des analyses sur la contamination liée aux PFAS, en croisant des approches en data science, visualisation et intelligence artificielle. Des outils facilitant l'exploration, l'analyse et la visualisation de ces informations sont donc en cours de développement.

Porté par le CNRS (via le MITI [lien]) et par l'Université de Montpellier dans le cadre du projet ExposUM, DAE-PFAS devrait évoluer dans les prochaines années en un ensemble de sous-projets ciblant différents volets (environnement, santé…).

5.3.4 Mon implication dans la vie de ce projet

J'ai eu la chance et l'honneur de présenter mes travaux lors d'une session du GDR MaDICS à Toulouse fin mai. Mon travail a également été présenté par mon tuteur à la conférence INFORSID à Pau en juin, où il a été très bien accueilli, et le sera à nouveau en juillet à PFIA à Dijon.

5.4 Ma mission dans le cadre de ce projet

Ma mission dans ce projet s'inscrit au croisement entre développement logiciel et recherche. Dans un premier temps, j'ai été chargé de concevoir une nouvelle interface web en m'inspirant du site du PDH avec l'objectif de proposer un outil plus parlant visuellement et plus adapté aux besoins des chercheurs. Ce travail a abouti à une version permettant des filtrages plus poussés et une visualisation plus claire des données. Ce travail est révolu et sera détaillé dans la partie X.X

Dans un second temps, (en cours) ma mission s'oriente vers un axe plus exploratoire du projet : l'analyse des données PFAS via des méthodes de clustering. L'objectif est de mieux comprendre les sites contaminés, en comprenant leurs origines, leurs causes et en dressant des profils de contaminations. Cela implique également la conception de visualisations claires permettant aux experts d'interpréter les résultats générés par nos algorithmes facilement, sans avoir besoin de comprendre la logique intrinsèque de ces mêmes algorithmes. Ce travail est en cours et se développera sur le restant de la durée de mon stage.

Mon travail est resté aligné sur les axes prévus au début du projet : d'une part l'amélioration des outils de sélection des données, et d'autre part l'intégration d'outils d'analyse exploratoire et explicative, dans le but de produire à terme une contribution scientifique valorisable.

5.5 Conditions de travail

Nous allons maintenant passer au détail de mes conditions de travail. Je vais commencer par parler de l'environnement de travail, avant de décrire l'encadrement mis en place puis de finir sur les outils que j'ai beaucoup utilisé lors de ce stage.

5.5.1 Environnement de travail

Mon stage s'est déroulé en présentiel dans le bâtiment 5 du campus Saint-Priest.

Ce campus est un pôle technologique qui a été conçu pour favoriser les synergies entre recherche académique, innovation industrielle et formation. Il regroupe aujourd'hui une grande partie des acteurs montpelliérains de la recherche en sciences et technologies du numérique.

Le bâtiment 5 accueille la quasi-totalité de l'IES (environ 200 personnes), une partie du LIRMM (environ 100 personnes), une antenne du centre INRIA Sophia Antipolis (environ 100 personnes), ainsi que des espaces réservés aux start-ups et à la recherche partenariale avec les entreprises locales.

C'est un environnement propice à la rencontre entre disciplines scientifiques, ce qui favorise les travaux interdisciplinaires — un aspect essentiel pour faire progresser la recherche [21].

5.5.2 Encadrement

L'encadrement durant mon stage s'est inscrit dans une dynamique d'autonomie. Les membres permanents de l'équipe alternant entre enseignement et recherche, leur présence au laboratoire est variable. Mon référent, très sollicité en début de stage, a été régulièrement en déplacement. J'ai dû adopter une démarche proactive : aller chercher l'information auprès des membres de l'équipe plutôt que d'attendre qu'elle me soit apportée. Un suivi hebdomadaire a été mis en place chaque vendredi pour faire le point sur l'avancement des travaux en cours et la prévision des prochains.

5.5.3 Outils Utilisés

J'ai utilisé GitLab pour le versionnement du code, Visual Studio Code comme éditeur, et Discord pour les échanges informels. Pour l'aspect recherche, Google Scholar m'a permis d'identifier les travaux académiques pertinents, et Zotero m'a servi à structurer et gérer la bibliographie de manière rigoureuse.

Il est maintenant temps de rentrer dans le cœur de mon travail, la prochaine partie va détailler les tâches que j'ai réalisé lors de ce stage.

6. Déroulement du stage

Pour structurer l'analyse de mes travaux au cours du stage, je vais commencer par décrire le développement du premier outil. Ensuite, je parlerai de la seconde phase, qui est bien plus exploratoire. Les éléments présentés lors de la seconde partie seront plus abstraits que ceux présentés dans la première.

6.1 Développement et refonte du site du PFAS Data Hub (PDH)

Durant les 2 premiers mois de mon stage, mon travail a consisté à améliorer un outil existant. Je vais le présenter, identifier ses limites et détailler les améliorations implémentées. Je parlerai des problèmes rencontrés et des solutions mises en place avant de présenter les résultats et analyser mon travail avec un regard critique.

6.1.1 Analyse de l'existant

Le site du PDH constitue aujourd'hui la seule plateforme publique donnant accès à des données relatives à la contamination liée aux PFAS en Europe. Mis en place et maintenu par M. Luc Martinon, data journaliste et consultant au CNRS sur le projet DAE-PFAS. Son but est d'offrir aux chercheurs (notamment chimistes et environnementalistes) un outil leur permettant de visualiser et d'accéder aux données du PDH. Voici un scénario d'utilisation basique pour un chercheur :

1. L'utilisateur filtre les données qu'ils l'intéressent.
2. Pour l'aider à visualiser l'information l'utilisateur à a sa disposition une carte qui se met à jour en direct et une fonctionnalité de clic pour analyser les données d'un point.
3. L'utilisateur exporte les données filtrées pour effectuer des analyses complémentaires.

Ci-dessous une capture d'écran de l'interface de l'outil.

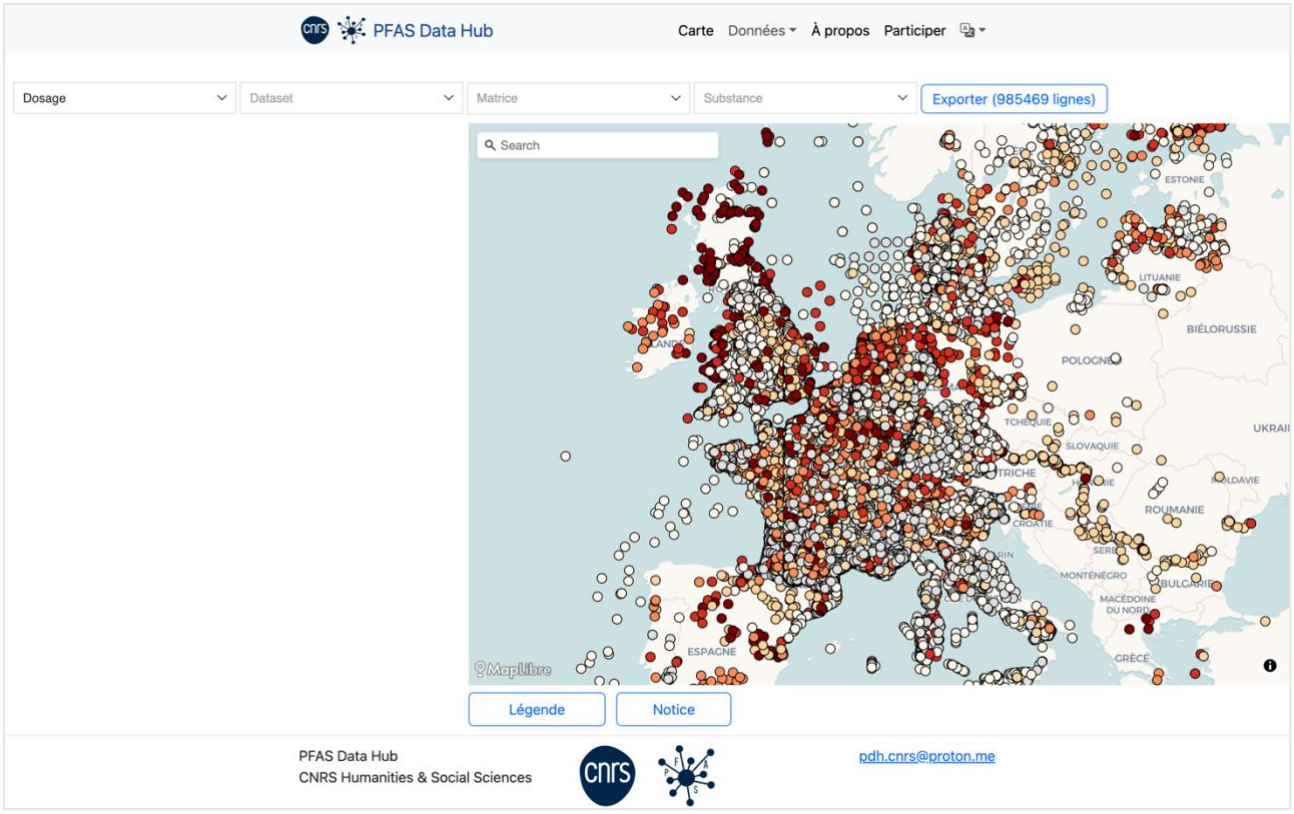


Figure X : Capture d'écran de l'interface du PDH

Les fonctionnalités principales de cet outil, son architecture technique et ses limites sont détaillés dans les prochaines sous-parties.

5.1.1.1 Fonctionnalités principales

Le filtrage constitue une fonctionnalité essentielle pour tout outil visant à afficher un volume important de données. L'interface du PDH, présentée ci-dessus, offre à l'utilisateur la possibilité de filtrer les données selon plusieurs critères :

- Catégorie : Les données peuvent représenter des sites de production de PFAS, des sites qui utilisent des PFAS, des sites de de contamination présumée, ou des mesures effectuées (avec des mesures de concentrations de polluant).
- Dataset : Les données ont été regroupées et homogénéisées dans un outil unique par les membres du projet mais viennent de plusieurs sources, et donc plusieurs 'datasets'.
- Matrice : Nature du lieu de mesure (eau, sol, air...).
- Substance : La famille des PFAS est très large, le site du PDH propose de filtrer par substances (les substances sont des composés chimiques de la famille des PFAS).

Un autre élément important est d'accéder à la donnée. Lors d'un clic sur un point, l'interface du PDH propose un panneau latéral affichant les données complètes :

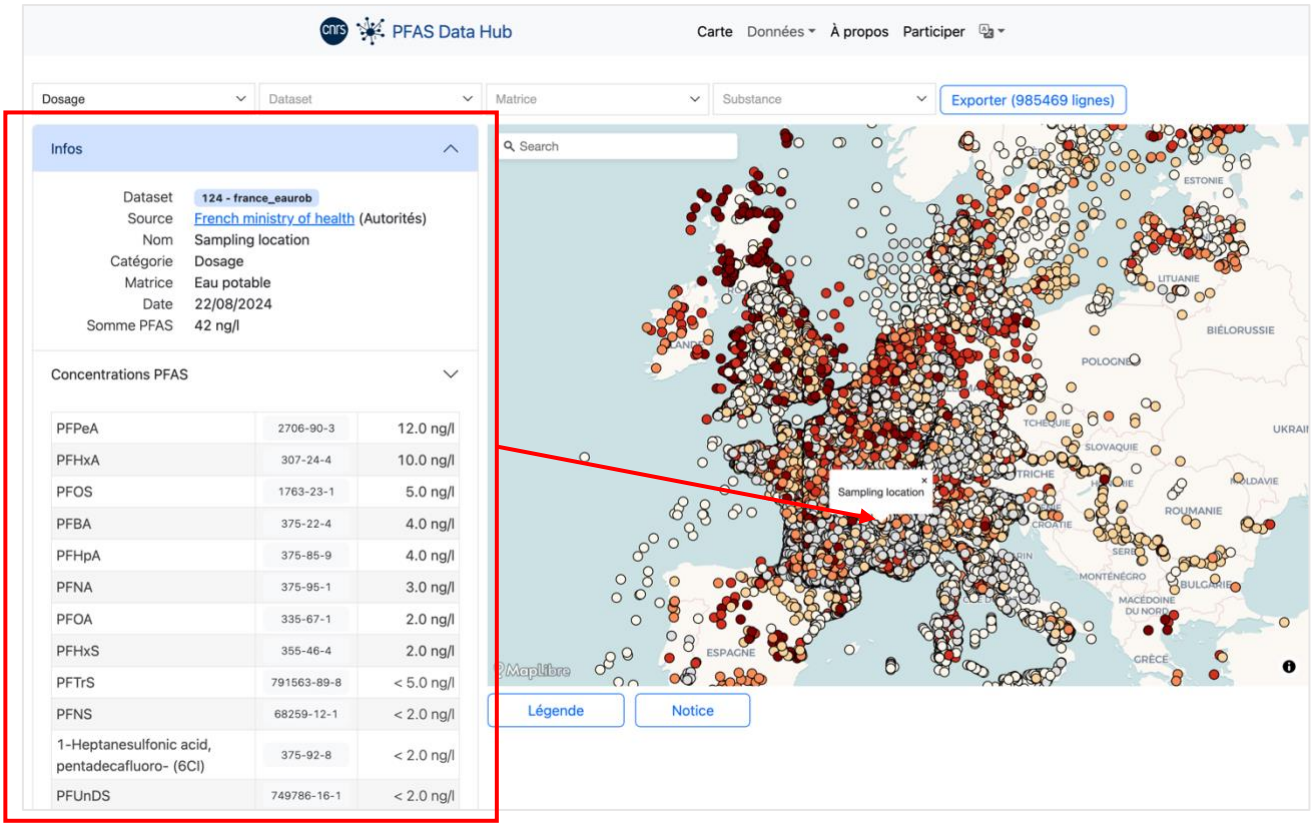


Figure X : Panneau latéral avec le détail des informations du point sélectionné

Il est possible sur la gauche de la figure X de voir les mesures de concentrations réalisées pour le point sélectionné ainsi que d'autres informations.

Les données peuvent être exportées au format CSV.

5.1.1.2 Architecture

Ce projet repose sur une architecture multi services orchestrée via Docker Compose avec les éléments suivants :

Frontend : C'est un site statique principalement composé de pages Markdown. La page carte interactive constitue l'élément le plus dynamique, développée en HTML / JavaScript pur. La carte en elle-même est développée à partir de MapLibre.

Backend/API : FastAPI et DuckDB sont utilisés pour accéder aux données efficacement.

5.1.1.3 Limites

Ce site est le premier à mettre à disposition ce jeu de données du PDH, à été développé et est maintenu par une seule personne : Luc Martinon – qui n'est même pas développeur –. Il est normal qu'il contienne certaines limites.

Premièrement, les fonctionnalités de filtrage sont incomplètes :

- Il est impossible de filtrer par zone géographique ainsi que par date de mesure.
- Les substances sont indépendantes de l'unité dans laquelle elles ont été mesurées. Par exemple, si je sélectionne la substance PFOA – un des PFAS les plus répandu –, mes résultats vont contenir des mesures en ng/L (valeurs entre 0 et 40) comme en ng/Kg (valeurs entre 10 et 5000). Cela fausse les échelles et peut engendrer des erreurs dans les analyses

De plus, avec plus de 1 million de points à afficher, la visualisation du PDH n'est pas claire. Les points se superposent, il est difficile d'identifier des tendances / clusters. La légende de couleur est logarithmique et manque de précision / nuance.

Il est important de noter que pour ces causes, M. Martinon reçoit des demandes de filtrage personnelles par certains chercheurs – qui veulent par exemple isoler une région de l'Europe dans leurs analyses –. Il traite ces demandes au cas par cas et pourrait gagner du temps si l'outil du PDH était plus performant.

Dans les prochaines parties je vais détailler le travail que j'ai réalisé avec l'équipe pour reproduire les fonctionnalités dans un nouvel outil en repartant de zéro, puis présenter les améliorations mises en place.

6.1.2 Reproduction des fonctionnalités

Dans le but d'avoir une version propre au LIRMM et facilement maintenable par les membres de l'équipe, il m'a été demandé de développer l'outil en repartant de zéro en utilisant les technologies suivantes :

- Svelte comme framework pour le frontend.
- PostgreSQL pour la base de données.
- SvelteKit (intégré nativement dans Svelte) pour la liaison du frontend à la base de données.
- Leaflet pour la cartographie car très simple à mettre en place pour des résultats rapides.

Le système de filtrage repose sur une rune svelte centrale qui contient l'état des filtres. Dans Svelte, une rune est une variable réactive, qui fonctionne un peu comme une "valeur intelligente" : dès qu'on la modifie, tous les éléments qui en dépendent se mettent automatiquement à jour. Elle est basée sur le design pattern [X] observable [X], et permet de propager facilement les changements dans l'interface. Toute modification de filtre dans l'application change l'état de cette rune et entraîne une nouvelle récupération de données. Cette approche utilise l'approche réactive native et très efficace de Svelte.

Une fonction réagit aux changements dans cette rune et lance la récupération des nouveaux points en accord avec les nouveaux filtres. Elle renvoie un objet "response", stocké dans une variable, qui peut avoir 3 états interprétés différemment par l'application :

- Loading : dans ce cas, on affiche un écran de chargement.
- Erreur : on affiche un message d'erreur.
- Data : contient la donnée voulue, on affiche les résultats.

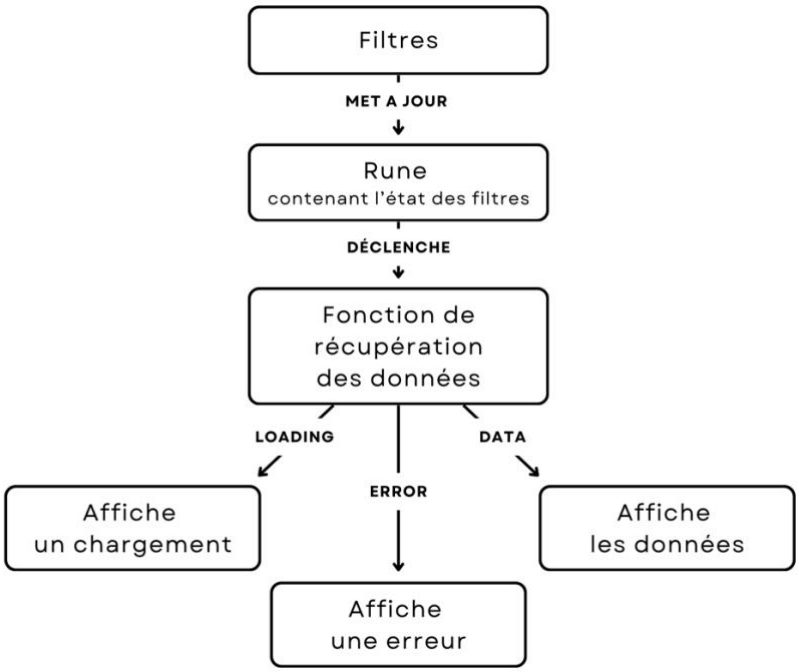


Figure X : Schéma détaillant le fonctionnement du filtrage

Passons maintenant à l'analyse des améliorations apportées, une première partie détaille les améliorations fonctionnelles, une seconde les améliorations visuelles.

6.1.3 Améliorations fonctionnelles

5.1.3.1 Filtrage avancé

- Filtrage par zone géographique

Nous avons mis en place un système de filtrage par zone géographique. L'utilisateur peut ainsi sélectionner un pays, une région ou un département d'un simple clic sur une carte interactive. Il est également possible de définir une zone personnalisée à l'aide d'un outil de sélection libre (lasso), permettant d'effectuer des analyses ciblées sur une ville ou sur une zone ne correspondant à aucun découpage administratif standard.

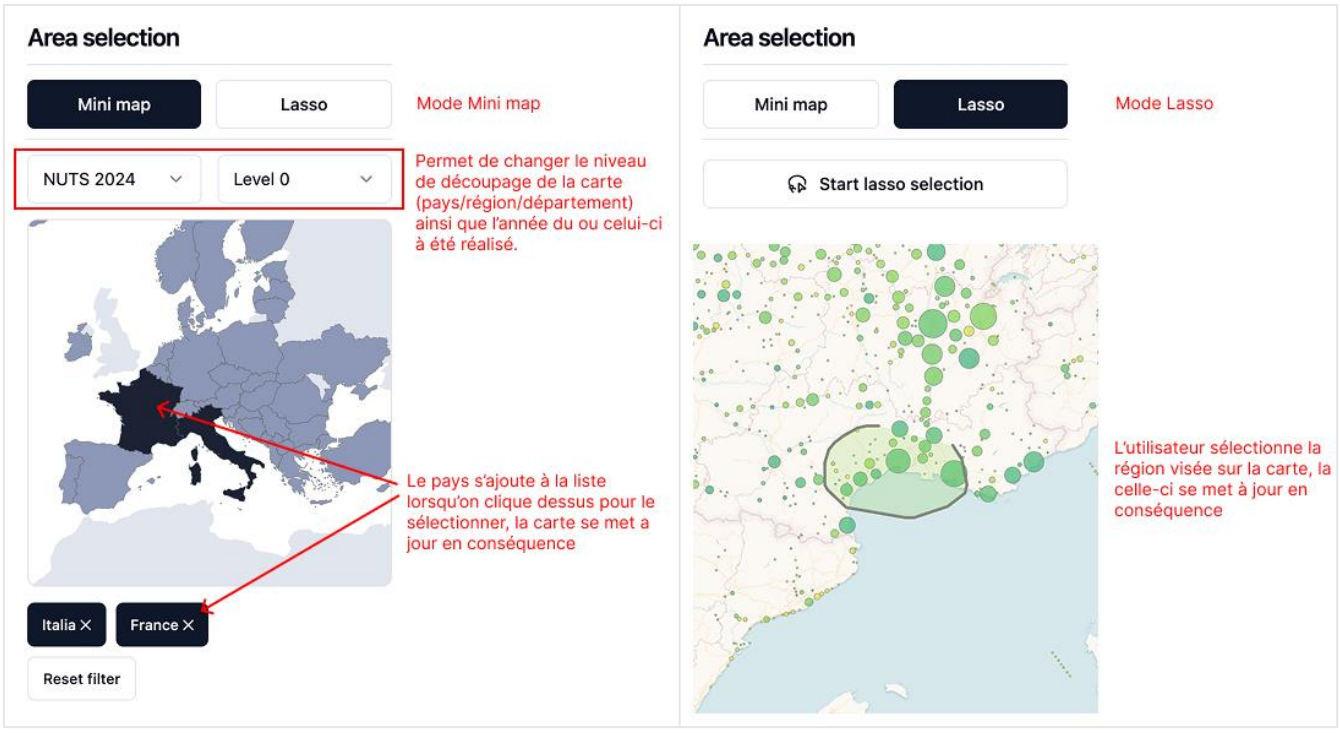


Figure X : Détail du composant de filtrage par zone géographique.

- Filtrage par substance

Un autre filtre essentiel mis en place concerne les substances. Pour chaque substance sélectionnée, l'utilisateur doit également choisir une unité, formant ainsi un couple substance/unité — ce qui permet d'éviter les incohérences évoquées en partie X. Un mode optionnel permet d'exclure les valeurs aberrantes (« outliers »), en appliquant la méthode des écarts interquartiles. Enfin, un curseur (slider) permet de filtrer les concentrations associées à chaque couple, offrant un contrôle précis sur la plage de valeurs à visualiser.

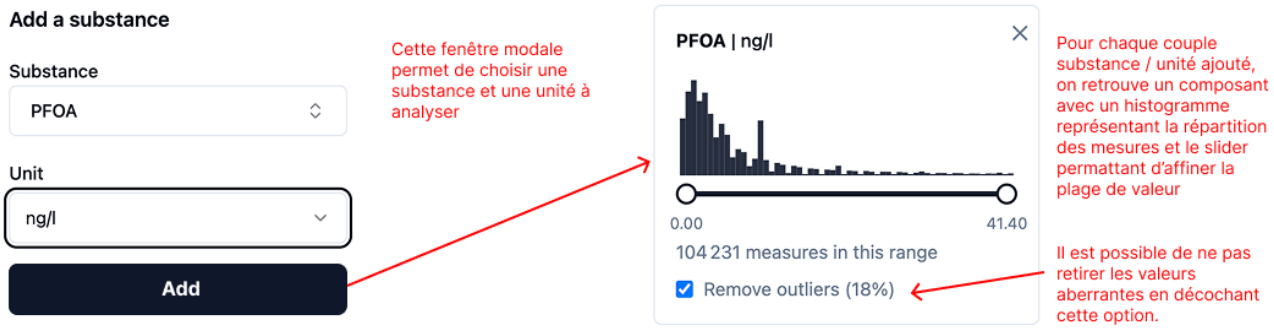
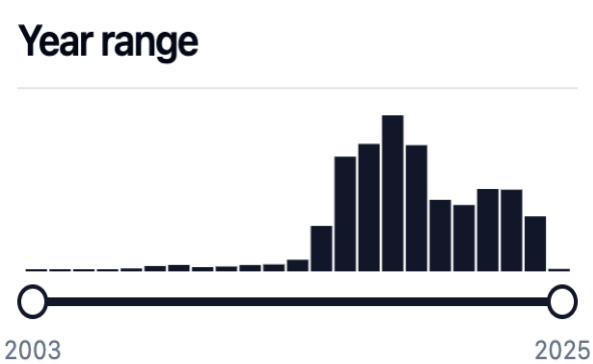


Figure X : Détail du composant de filtrage par substance.

- Filtrage par année de mesure



Il est aussi possible de filtrer par année de mesure, un détail qui n'est pas anodin puisque certaines données trop anciennes peuvent être considérées comme non fiables avec les évolutions des appareils de mesure.

Figure X (a gauche) : Détail du composant de filtrage par plage d'années.

Bien sûr, il est toujours possible de filtrer par matrices, dataset et catégorie comme dans la version initiale.

- Clic

Lorsqu'un utilisateur clique sur un point, nous avons repensé l'affichage des informations pour le rendre plus pertinent. Comme la plupart des points correspondent à des stations de mesure, plusieurs prélèvements peuvent avoir été effectués au même endroit. Nous avons donc choisi de regrouper les résultats par substance, sous forme de répartition (min, max, moyenne, quartiles…).

Cependant, cette approche présente une limite importante : elle ne restitue pas l'évolution temporelle des concentrations, pourtant essentielle dans le cadre du suivi environnemental. En effet, l'une des principales fonctions d'une station est de mesurer les variations dans le temps. Une amélioration envisagée — à intégrer dans les prochaines phases du projet — serait d'ajouter une visualisation temporelle. L'idée serait d'afficher une petite courbe pour chaque ligne permettant d'apercevoir rapidement la tendance de l'évolution de la concentration.

Details

9 points

Substance	Unit	Detections	Mean	Min	Q1	Median	Q3	Max
PFBA	ng/l	6/9	97.00	28.00	35.75	50.00	87.50	323.00
PFBS	ng/l	8/9	21.13	6.00	8.75	12.00	22.75	65.00
PFHxA	ng/l	8/9	11.25	6.00	7.50	11.50	13.25	20.00
PFOA	ng/l	9/9	12.56	5.00	6.00	11.00	16.00	30.00
PFOS	ng/l	8/9	2.06	0.88	1.24	1.90	2.39	4.34
PFPeA	ng/l	5/9	8.80	5.00	6.00	6.00	9.00	18.00

Figure X : Détail du panneau affichant les détails d'une station de mesure.

5.1.3.3 Autres améliorations

La logique d'export des données a également été repensée. Il est désormais possible de choisir les colonnes à inclure dans l'export CSV, ce qui permet de limiter la surcharge d'informations inutiles et de simplifier les analyses.

Par ailleurs, l'interface offre la possibilité d'enregistrer directement la carte affichée au format PNG, sans avoir recours à une capture d'écran — une fonctionnalité particulièrement utile pour illustrer des résultats ou réaliser des comparaisons visuelles. Enfin, d'autres améliorations ont été apportées à la navigation sur la carte : centrer la vue sur les données actuellement filtrées, ou zoomer automatiquement sur un cluster en cliquant dessus en sont des exemples.

Ces nouvelles fonctionnalités répondent directement aux limites identifiées et aux besoins exprimés par les utilisateurs du PDH – notamment les demandes récurrentes reçues par M. Luc Martinon.

6.1.4 Améliorations visuelles

Nous avons également ajouté une option de clustering des points, reposant sur l'algorithme F-SAC (*Fast Spatial Agglomerative Clustering*), développé au sein du LIRMM par un ancien doctorant [35]. Cette méthode permet d'agréger les points proches pour alléger l'affichage, tout en conservant une information pertinente, rendant la carte lisible même en présence de plusieurs dizaines voire centaines de milliers de points.

Par ailleurs, la palette de couleurs a été entièrement repensée pour améliorer la perception visuelle des valeurs. Nous avons adopté la palette Viridis, conçue pour être perceptible par les personnes atteintes de daltonisme, tout en assurant une bonne lisibilité en niveaux de gris. Ce choix est soutenu par les recommandations de l'article *Somewhere Over the Rainbow* de Yang LIU et Jeffrey HEER [36].

Ci-dessous une comparaison avant / après de la visualisation des mesures de PFOA (Sans distinction d'unité pour avoir une comparaison fiable).

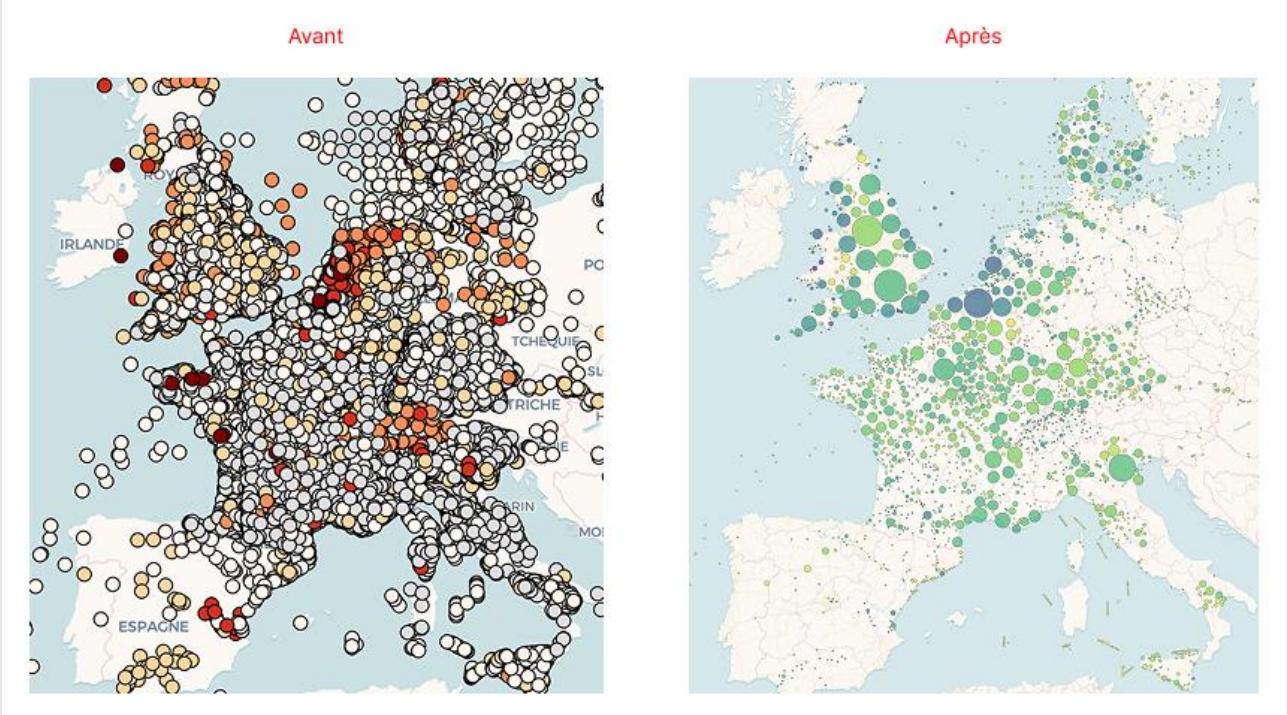


Figure X : Comparaison de la visualisation des mesures liés au PFOA (toutes unités)

Nous pouvons observer une surcharge frappante sur l'image de gauche, des points sont cachés. Sur l'image de droite nous pouvons bien plus facilement observer la répartition des mesures sur le territoire.

Pour améliorer encore l'expérience utilisateur, plusieurs fonds de carte alternatifs ont été intégrés, tels qu'OpenStreetMap, des versions avec ou sans labels, ou encore une version topographique, afin de s'adapter aux besoins d'analyse spécifiques.

Enfin, une annexe dédiée présente l'évolution de l'interface utilisateur tout au long du projet. [lien]

6.1.5 Enjeux techniques et solutions

Plusieurs défis ont jalonné le développement :

- Taille des données : Au-delà de 200 000 points, l'affichage en points classiques devient lourd à supporter dans le navigateur. Pour le mode cluster, c'est le temps d'exécution de l'algorithme F-SAC qui pose problème. Pour pallier cela, on effectue automatiquement une bascule vers un affichage en heatmap – bien moins lourde – lorsqu'il y a plus de 200 000 points à afficher sur l'écran. Pourquoi ne pas afficher tout le temps une heatmap ? Car on perdrait une dimension, la heatmap affiche uniquement la densité de mesure, et pas les valeurs de concentrations.

- Intégration de F-SAC : L'intégration de cet algorithme performant mais complexe a nécessité un travail approfondi. Il a fallu débattre de la version optimale. Une annexe détaille ce travail [lien].
- Performances base de données : Usage de materialized views et d'index spatio-temporels dans PostgreSQL pour accélérer les requêtes. Une annexe détaille ce travail [lien].
- Mise en place d'un ETL (script) pour récupérer les données, les enrichir et les insérer dans la base de données.

6.1.6 Résultats obtenus et validation

Ci dessous une présentation de l'interface finale du prototype mis en place. (Une annexe avec d'autres prises de vues l'explique plus en détail [lien]).

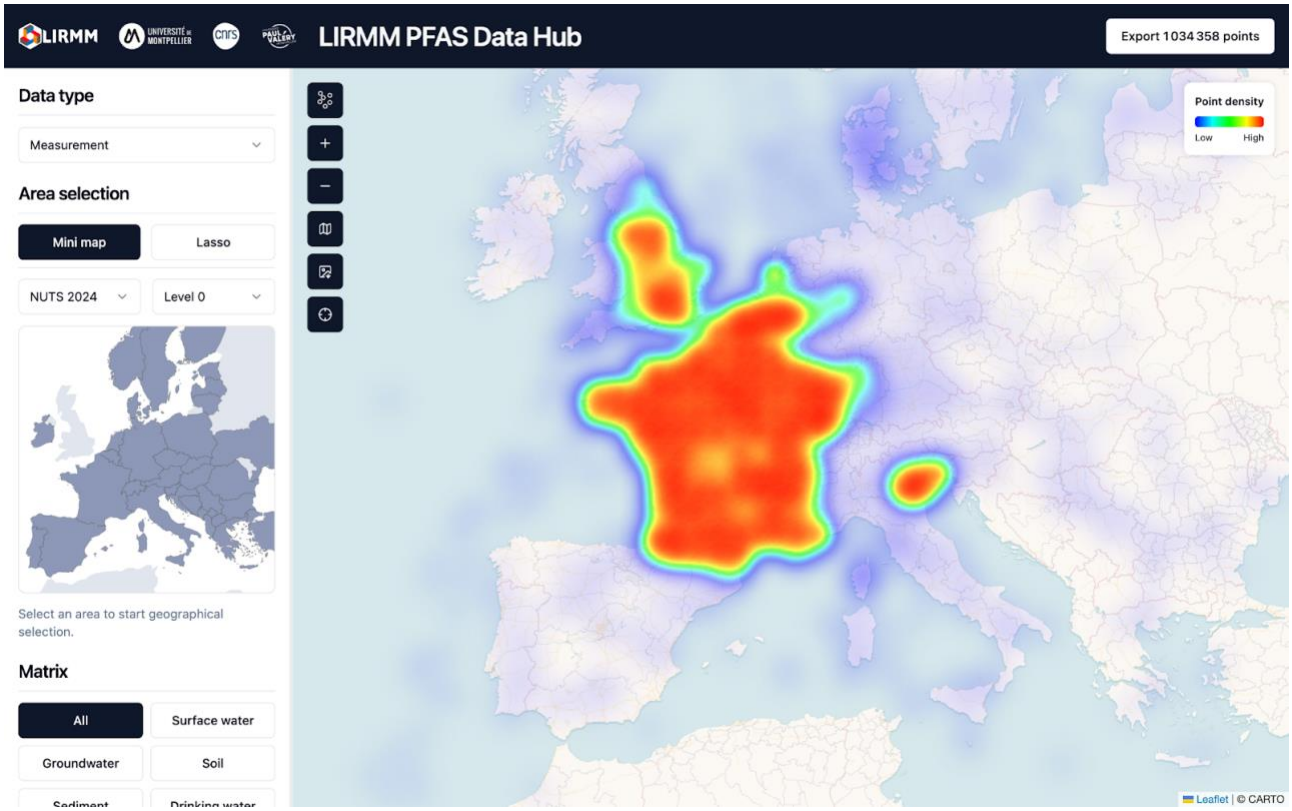


Figure X : Présentation de l'interface

L'interface est moderne, plus intuitive que l'ancienne version et offre aux chercheurs des fonctionnalités plus poussées.

Le prototype a été présenté lors du Symposium (conférence) MaDICS à Toulouse en mai. Les retours ont été très positifs, en particulier de la part de M. Luc Martinon, qui a exprimé un intérêt pour intégrer ces améliorations dans une future version publique du PDH.

Ce projet n'est pas hébergé, une documentation à été créée pour permettre à tout utilisateur d'installer le code, et de récupérer et formatter les données en lançant un script.

6.1.7 Perspectives

- L'interface développée présente un prototype proposant une version améliorée de la version du PDH, à terme, cette version pourrait remplacer l'ancienne. Voici quelques pistes d'améliorations qui pourraient être analysées si tel est le cas :
- Passage à Maplibre (WebGL) pour un affichage plus rapide et capable de supporter davantage de points.
- Hébergement sur serveur hautes performances, pour accélérer les temps de réponse et permettre la consultation à grande échelle.
- Intégration future des résultats de recherche (SHAP, clustering, interpolation) directement dans la plateforme.
- Mise à jour des données automatique en temps réel.

6.1.8 Analyse critique – Outil de filtrage

Analyse critique ici ou dans la partie 6.4 ? ou mélange ? technique ici et mentalilté en 6.4 ?

6.2 Phase exploratoire

La seconde partie de mon stage s'inscrit dans une démarche plus exploratoire, typique d'un vrai travail de recherche. Elle vise à investiguer des pistes pour améliorer la visualisation de certaines métriques d'explicabilité pour des modèles de machine learning. Nous essayerons de les appliquer à nos données en appliquant un modèle de clustering. Ce changement de paradigme — passer du développement à de la recherche — m'a confronté à de nouvelles problématiques, tout en continuant de travailler sur les mêmes données.

6.2.1 Qu'est ce que SHAP ?

SHAP, pour *SHapley Additive exPlanations*, est une méthode d'explicabilité largement utilisée dans le domaine du machine learning. Son objectif est d'indiquer, pour chaque prédiction d'un modèle, dans quelle mesure chaque variable a influencé le résultat obtenu.

Concrètement, SHAP calcule des valeurs pour chaque point de données : pour une instance donnée, chaque variable de celle-ci reçoit une valeur SHAP qui mesure sa contribution à la prédiction. Une valeur SHAP positive signifie que la variable a poussé la prédiction vers le haut, une valeur négative qu'elle l'a tirée vers le bas.

Prenons un exemple : si l'on entraîne un modèle pour prédire un salaire à partir de différentes informations (âge, niveau d'études, sexe, etc.), alors une valeur SHAP de +1200 pour la variable "niveau d'études" signifie que cette variable a contribué à augmenter le salaire prédit de 1200 € pour cette personne. À l'inverse, une valeur SHAP de -500 pour la variable "âge" indiquerait que l'âge a contribué à diminuer le salaire prédit de 500 € pour la personne.

Ce principe s'étend également aux modèles qui produisent des probabilités de classification. Si le modèle prédit la probabilité qu'une personne appartienne à une classe (par exemple : « cette contamination est due à l'utilisation de mousse à incendie », oui ou non), alors les SHAP values indiquent si chaque feature a renforcé ou affaibli la probabilité d'appartenir à cette classe.

Dans notre cas : nous cherchons à expliquer les résultats d'un modèle de *clustering*, c'est-à-dire un modèle qui regroupe les données selon leurs similitudes. Grâce à SHAP, il est possible d'obtenir, pour chaque point et pour chaque cluster, une valeur SHAP par feature, indiquant si cette feature a poussé le point à être classé dans ce cluster.

Ainsi, si notre modèle génère 4 clusters, alors chaque mesure (ou "point") aura 4 ensembles de valeurs SHAP — un par cluster — associées à chaque variable.

6.2.2 Motivations

L'un des grands défis de l'intelligence artificielle moderne est ce qu'on appelle l'explicabilité : autrement dit, la capacité à comprendre comment un modèle a pris une décision. Les modèles de machine learning, en particulier les plus puissants, sont souvent comparables à des « boîtes noires » : ils peuvent donner des résultats très précis, mais sans qu'on sache clairement *pourquoi* ni *comment*. Or, dans des domaines sensibles comme la santé, l'environnement ou la sécurité, il ne suffit pas d'avoir une prédiction correcte — il faut aussi pouvoir *l'expliquer*. Comprendre les facteurs qui ont mené à une décision permet non seulement d'avoir confiance dans le modèle, mais aussi de détecter d'éventuelles erreurs ou biais. En d'autres termes, si l'on veut pouvoir s'appuyer sur l'IA pour prendre des décisions importantes, il faut être capable de quantifier et analyser son raisonnement, comme on le ferait pour un être humain.

Les méthodes comme SHAP sont devenues incontournables pour répondre à cet enjeu, car elles offrent des explications locales et compréhensibles sur l'impact des variables. Néanmoins, un problème persiste : lorsque l'on travaille avec un grand nombre de variables, les visualisations produites par SHAP deviennent vite illisibles et difficiles à exploiter. Cela pose un problème d'accessibilité : seuls des experts familiers de la méthode parviennent à interpréter correctement les résultats.

Les graphes classiques générés par SHAP, comme le *beeswarm plot* (voir Figure ci-dessous), sont riches en information mais denses et peu intuitifs. Ils demandent un effort cognitif important. Il est possible d'en tirer un nombre d'informations énorme, mais la plupart ne sont pas directement implicites : les shap values sont explicites par exemple, car elles sont représentées sur l'axe des abscisses, par contre la répartition est analytique, il faut analyser la figure pour en tirer des conclusions (la liste est non exhaustive). Nous avons identifié trois types d'informations particulièrement riches en enseignements, mais difficiles à lire lorsqu'elles sont toutes superposées sur un même graphique. Les isoler dans des visualisations distinctes permettrait de réduire les risques de mauvaise interprétation et, surtout, de rendre l'analyse plus accessible aux non-spécialistes, leur offrant ainsi la possibilité de tirer des conclusions approfondies sur leur propre jeu de données.

L'enjeu est donc de proposer une manière plus claire et plus universelle de représenter ces résultats, pour qu'ils puissent être analysés plus rapidement et compris par des chercheurs ou des acteurs métiers non spécialistes. Il s'agit aussi de faire ressortir facilement les variables clés qui structurent les clusters, en facilitant ainsi la formulation d'hypothèses scientifiques.

En résumé, notre motivation est double :

Scientifique, en contribuant à la communauté XAI avec une méthode de visualisation adaptée à des cas complexes comme le clustering sur des données multi-dimensionnelles.

Pratique, en facilitant l'interprétation des résultats de SHAP au sein de notre projet, dans un contexte mêlant environnement, santé publique et analyse de données.

6.2.3 Méthodologie

Dans cette seconde partie plus orientée recherche, j'ai beaucoup appris grâce à mon tuteur, qui m'a initié à une vraie démarche scientifique. Il m'a transmis une méthode structurée pour réfléchir, poser une problématique pertinente, et avancer pas à pas avec l'objectif, à terme, de pouvoir rédiger une publication scientifique si nos résultats le permettent.

La démarche suivie ressemble à celle utilisée dans la recherche académique, avec différentes étapes :

- Commencer par poser une vraie problématique, claire et ciblée.
- Expliquer pourquoi elle est intéressante, en montrant les enjeux derrière.
- Explorer ce qui a déjà été fait, en construisant un état de l'art à partir de recherches sur Google Scholar, en testant différents mots-clés pour affiner les résultats.
- Vérifier qu'il y a bien un « trou » dans la littérature, c'est-à-dire un besoin non couvert ou mal traité.
- Imaginer une proposition ou une solution originale.
- Tester cette idée, valider ce qui fonctionne ou pas.
- Puis évaluer les résultats pour voir si notre approche apporte vraiment quelque chose.
- Enfin, résumer les conclusions et envisager une rédaction plus formelle.

Cette méthode m'a permis de structurer mon raisonnement et de prendre du recul sur ce que je faisais. Elle m'a aussi montré qu'un bon projet de recherche, ce n'est pas forcément partir d'une idée géniale, mais savoir poser les bonnes questions, tester, itérer… et bien présenter le tout.

6.2.4 Résultats intermédiaires

Ce travail est donc en cours. Pour le moment j'ai simplement essayé d'appliquer des méthodes des clustering sur nos données => conclusions pas fou

7. Analyse du travail réalisé

7.1 Mon apport au laboratoire

Bien que le projet DAE-PFAS soit un projet collaboratif de long terme, appelé à évoluer avec ou sans ma participation, j'estime avoir bien contribué à son avancement durant ma période de stage. Grâce à mes bases solides en développement travaillées à Polytech, j'ai pu prendre en main les besoins techniques du projet et proposer des solutions efficaces dans un délai court. J'ai développé un outil fonctionnel et documenté. J'ai veillé à appliquer les bonnes pratiques de développement, tant sur le plan de l'organisation du code que de la clarté de l'interface, afin de garantir la maintenabilité de l'outil. Celui-ci pourra servir de socle pour les prochaines phases du projet, notamment lors d'analyses plus poussées… Enfin, les travaux exploratoires menés pendant le stage, constitueront une base de référence utile pour les futurs stagiaires ou chercheurs impliqués. Ils fourniront des premières pistes d'analyse, des prototypes et un cadre technique sur lequel s'appuyer.

7.2 Compétences développées

Au cours de ce stage, j'ai pu de renforcer mon autonomie, mes compétences en développement ainsi qu'élargir mes connaissances en recherche scientifique, notamment dans un cadre de travail interdisciplinaire.

Sur le plan technique, j'ai tout d'abord découvert un nouveau framework : Svelte. Cela m'a amené à revisiter la documentation et à approfondir certains concepts fondamentaux du développement web. Cette remise à niveau m'a permis de revoir les bases. Avec l'expérience je commence à vraiment comprendre les enjeux liés aux frameworks web, en termes de performance, d'expérience utilisateur et de maintenabilité du code. J'ai également travaillé avec le langage Python sur plusieurs problématiques, j'ai créé des visualisations, des scripts pour modifier des données… Ce langage, que je continuerai à utiliser dans la suite du projet, m'a permis de consolider ma pratique en matière de traitement de données. Par ailleurs, j'ai développé des compétences en Machine Learning, notamment à l'occasion de la session MaDICS à Toulouse fin mai. Cette thématique, qui est centrale la seconde phase de mon stage, m'a particulièrement intéressé par sa richesse et son potentiel.

En parallèle, mon immersion dans un environnement de recherche interdisciplinaire m'a permis d'acquérir des compétences spécifiques à la démarche scientifique. Les réunions hebdomadaires ont été l'occasion de confronter les idées, de discuter des orientations à suivre et de trouver collectivement des solutions à des problématiques complexes. J'ai assisté à des échanges rassemblant des chercheurs aux profils très variés (experts en visualisation, en base de données, ingénieurs …), ceux-ci m'ont sensibilisé à l'importance de

l'écoute, de la pédagogie et de la collaboration dans un projet de recherche. Cette capacité à travailler ensemble, malgré les différences de spécialités, m'a impressionné.

J'ai également appris la « méthodologie de recherche ». On peut retrouver des similitudes dans la manière de penser, de s'organiser quels que soient les problématiques scientifiques abordées. Avoir pratiqué cette méthodologie m'aidera à appréhender avec plus de recul les futures problématiques auxquelles je serai confronté. Faire de la recherche en m'appuyant sur des articles scientifiques m'a conduit à utiliser des outils que je n'avais jamais utilisé en dehors des cours, comme Google Scholar pour la recherche d'articles et Zotero pour la bibliographie.

Enfin, ce stage m'a permis de progresser en autonomie. L'encadrement, bien que toujours disponible en cas de besoin, était relativement souple. Il m'a donc fallu quelques jours pour prendre mes marques, organiser mon travail, poser mes questions au bon moment et gagner en assurance. Cette liberté à été déstabilisante au début mais je m'y suis habitué, j'ai pu développer à terme un rythme de travail efficace.

7.3 Méthodologie de travail

Tout au long du stage, l'organisation de travail à été efficace, s'inspirant des principes agiles. Chaque vendredi, une réunion avec l'équipe permettait de donner du feedback, de faire le point sur les avancées de la semaine écoulée et de définir précisément les objectifs pour la semaine suivante. Ces réunions agissaient comme des sprints, favorisant un rythme régulier et une bonne visibilité sur l'avancement du projet. Pour appuyer cette dynamique, j'ai tenu à jour un fichier *Weekly-digest* (suivi hebdomadaire) recensant les tâches accomplies, les difficultés rencontrées et les objectifs à venir sous forme de to-do list. Ce document a facilité les échanges et permis d'éviter les hésitations ou les pertes de temps. En complément, des points rapides étaient organisés à la demande avec mon tuteur, soit en présentiel à son bureau, soit à distance via Discord, assurant un suivi réactif et fluide.

7.4 Analyse critique

7.4.1 Mentalité

L'intégration au sein de l'équipe n'a pas été immédiate. Dès le premier jour, le travail a commencé sans réel temps dédié à l'accueil ou à la présentation des membres de l'équipe. Cela m'a généré un léger sentiment d'isolement, il me manquait un cadre bien défini. Je me sentais un peu comme lâché dans le grand bain, sans trop savoir comment faire ce qu'on me demandait ni à qui demander de l'aide. Avec le recul, c'était totalement normal. J'ai juste été habitué à quelque chose de plus « doux » avec le stage en entreprise de l'année dernière, ou tout a pris plus de temps. Il y avait un nombre incalculable de protocole à respecter avant de pouvoir commencer à travailler. J'aurais pu faire davantage d'efforts dès

le début pour aller vers les autres et m'intégrer plus activement. Ce point constitue une première piste d'amélioration pour mes futures expériences : prendre l'initiative dès le début dans le but d'être acteur pour établir des liens et faciliter la communication au sein de l'équipe.

7.4.2 Technique

J'ai ressenti des lacunes techniques lors de l'implémentation du système de clustering. Le code de base étant déjà écrit, je pensais initialement pouvoir l'adapter rapidement. En pratique, l'adaptation à mon cas d'usage a été complexe : il m'a fallu beaucoup de temps pour comprendre les subtilités de l'algorithme, ajuster les paramètres, et obtenir des résultats pertinents. En expérimentant, j'ai eu tendance à me détourner de l'objectif initial, cherchant avant tout à obtenir un résultat "visuel" satisfaisant, au détriment de la finalité scientifique recherchée. Ce constat m'a permis de prendre conscience de l'importance de toujours garder en tête les objectifs du projet, même en cas de difficulté technique.

La seconde phase du stage, axée sur des outils de Machine Learning, je m'attends à devoir surmonter une nouvelle phase d'adaptation. Les bases acquises à Polytech sont correctes pour des projets encadrés, mais le niveau d'exigence scientifique attendu ici est plus élevé, notamment en matière de choix de modèles, ou d'interprétabilité des résultats. Cela représente un défi stimulant, pour m'adapter je vais devoir m'organiser pour monter en compétence rapidement le moment venu.

La prochaine partie conclue le rapport, elle détaille le devenir de mes travaux, je vais y dresser un bilan et une réflexion sur mon avenir professionnel.

8. Conclusions et perspectives

8.1 Devenir du projet

Le projet DAE-PFAS dispose encore d'environ un an dans sa configuration actuelle. Dans la continuité d'un travail amorcé depuis un an déjà, au cours duquel les PFAS ont été explorés sous différents angles. Ce temps d'exploration a permis de dégager des axes de recherche plus ciblés, révélant la complexité et l'ampleur du sujet. En conséquence, il est prévu que DAE-PFAS évolue en un ensemble de sous-projets thématiques, chacun concentré sur un domaine spécifique, tels que la santé, ou l'environnement par exemple.

8.2 Devenir de ma contribution

Concernant mon travail, le prototype que j'ai développé pourrait à terme être intégré au site du PFAS Data Hub. Cette intégration nécessitera surement une phase de validation et d'ajustements, en collaboration avec un expert du domaine dans l'idéal. Le travail effectué dans la seconde phase de mon stage ayant été plus exploratoire, les travaux engagés resteront en suspens à la fin de mon passage. Ces éléments, bien documentés, servirons de point de départ pour un nouveau stagiaire prévu l'année prochaine.

8.3 Bilan personnel

Ce stage au LIRMM m'a offert une expérience enrichissante et formatrice, tant sur le plan technique que personnel. J'ai eu l'opportunité de découvrir le fonctionnement d'un laboratoire de recherche, avec une grande diversité de profils et une ambiance de travail agréable. Cela m'a permis de m'introduire de la meilleure des manières au monde de la recherche scientifique, un univers stimulant et riche en idées.

Sur le plan des compétences, ce stage m'a permis de consolider des acquis essentiels pour la suite de mon parcours, qu'il s'agisse du développement logiciel, de la gestion autonome d'un projet ou encore de la collaboration dans un cadre interdisciplinaire. J'ai également découvert un domaine qui m'a particulièrement attiré : la visualisation de données. Réfléchir à la manière de représenter clairement des résultats complexes – cartes, graphiques, interfaces – a été une facette à la fois technique et créative du stage que j'ai beaucoup appréciée. Elle fait écho à mon intérêt pour l'UI (Interface Utilisateur) développé lors de projets à Polytech, et ouvre peut-être une piste à explorer plus profondément à l'avenir.

8.4 Évolution de mon projet professionnel

Avant ce stage, je m'interrogeais sur la possibilité de poursuivre mes études par une thèse, ou de m'orienter directement vers le monde de l'entreprise. Ces quelques mois au LIRMM m'ont permis de clarifier ce point.

Aujourd'hui, je ne me projette pas dans une thèse à court terme. Ce choix est lié à la fois à des envies personnelles, à un besoin de stabilité, mais aussi au fait que les sujets que j'ai trouvé dans mes recherches et qu'on m'a proposés pendant mon stage ne m'ont pas pleinement convaincu. J'ai tout de même passé un entretien pour une thèse CIFRE en lien avec AcuSurgical — une entreprise cofondée par l'ancien directeur du LIRMM évoquée plus tôt dans le rapport — ce qui m'a permis d'explorer cette voie. Toutefois, je pense qu'il serait prématuré de m'y engager dès maintenant, notamment en raison du niveau académique élevé que ce type de parcours exige.

Pour autant, je n'exclus pas de revenir vers une thèse à moyen terme. J'ai rencontré plusieurs doctorants ayant choisi de débuter une CIFRE après quelques années de CDI, une

trajectoire qui pourrait me correspondre davantage. Je compte d'ailleurs poser la question de l'éligibilité aux dispositifs CIFRE lors de mes prochains entretiens d'embauche, car cela pourrait influencer mon choix d'entreprise.

Dans l'immédiat, mon objectif est de trouver un poste en CDI dans une entreprise implantée dans la région de Montpellier. Je resterai attentif aux opportunités qui pourraient combiner activité en entreprise et recherche appliquée. Notamment dans une équipe de R&D.

9. Webographie

[1] Page d'accueil, LIRMM. [En ligne]. Disponible sur : LIRMM

[2] Fiche entreprise, LIRMM – Robotics Place. [En ligne]. Disponible sur : Robotics-place

[3] Le LIRMM fête ses 30 ans de recherche scientifique. Université de Montpellier. [En ligne]. Disponible sur : LIRMM-30ans

[4] Laboratoire LIRMM – Unité de recherche. Université de Montpellier. [En ligne]. Disponible sur : LIRMM-UMR

[5] Wikipédia – Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier. [En ligne]. Disponible sur : LIRMM-Wiki

[6] LIRMM – Équipes de recherche. [En ligne]. Disponible sur : LIRMM-Equipes

[7] Équipe ADVANSE, LIRMM. [En ligne]. Disponible sur : LIRMM-ADVANSE

[9] Université de Montpellier – 30 ans d'innovation avec le LIRMM. [En ligne]. Disponible sur : LIRMM-Innov

[10] UNEP – Persistent Organic Pollutants and PFAS. [En ligne]. Disponible sur : UNEP-PFAS

[11] ANSES – PFAS : substances chimiques persistantes. [En ligne]. Disponible sur : ANSES-PFAS

[12] Santé publique France – Imprégnation de la population française par les composés perfluorés (Esteban 2014–2016). [En ligne]. Disponible sur : Santé-publique-France

[13] ECHA – Perfluoroalkyl chemicals (PFAS). [En ligne]. Disponible sur : ECHA-PFAS

[14] Vie publique – Loi du 27 février 2025 sur les PFAS. [En ligne]. Disponible sur : Loi-PFAS

[15] Ministère de la Transition Écologique – Plan d'action PFAS 2023-2027. [En ligne]. Disponible sur : Plan-action

[16] France 3 Régions – PFAS : recommandations sanitaires autour de l'usine Arkema. [En ligne]. Disponible sur : Article

[17] Forever Pollution Project. [En ligne]. Disponible sur : FPP

[18] MADICS – Événement sur la visualisation des données. [En ligne]. Disponible sur : MaDICS-DAE

[19] MADICS – Accueil. [En ligne]. Disponible sur : MaDICS

[20] DFAE – Data for Environmental Applications. [En ligne]. Disponible sur : DFAE-DAE

[21] Pôle STICS – Université de Montpellier / Opération Campus. [En ligne]. Disponible sur : Pole-STICS

[22] SHAP Documentation. SHAP. [En ligne]. Disponible sur : SHAP

[23] SESAME – Plateforme MathNum, INRAE. [En ligne]. Disponible sur : SESAME

[24] PFAS – Rejets industriels en France. [En ligne]. Disponible sur : PFAS-rejets

[25] Union Européenne – Réglementation sur les emballages alimentaires contenant des PFAS. [En ligne]. Disponible sur : EU-Emballages-alimentaires

[26] LIRMM – Science ouverte et publications. [En ligne]. Disponible sur : LIRMM-Science-ouverte

[27] LIRMM – Engagements en matière de développement durable (Green LIRMM). [En ligne]. Disponible sur : Green-LIRMM

[28] LIRMM – Égalité et parité. [En ligne]. Disponible sur : LIRMM-égalité-parité

[29] AcuSurgical – Spin-off du LIRMM spécialisée en chirurgie robotique. [En ligne]. Disponible sur : AcuSurgical

[30] Algodone – Start-up issue du LIRMM dans la microélectronique sécurisée. [En ligne]. Disponible sur : Algodone

[31] LIRMM – Projets européens collaboratifs. [En ligne]. Disponible sur : LIRMM-Projets-EU

[32] Projet AI4CCAM – Intelligence artificielle pour les systèmes coopératifs et autonomes. [En ligne]. Disponible sur : AI4CCAM

[33] GUARDEN – Projet européen sur la surveillance environnementale. [En ligne]. Disponible sur : GUARDEN

[34] Horizon Europe – Projets financés par le programme Horizon. [En ligne]. Disponible sur : Horizon-EU

[35] Disponible sur : F-SAC

[36] Disponible sur : Over-the-rainbow

10. Annexes

10.1 Présentation de l'outil en détail

10.2 Intégration de F-SAC

Au départ, je pensais que l'algorithme F-SAC ne nécessitait qu'un seul paramètre : le rayon initial permettant de vérifier si un point est suffisamment proche d'un autre pour être agrégé au même cluster.

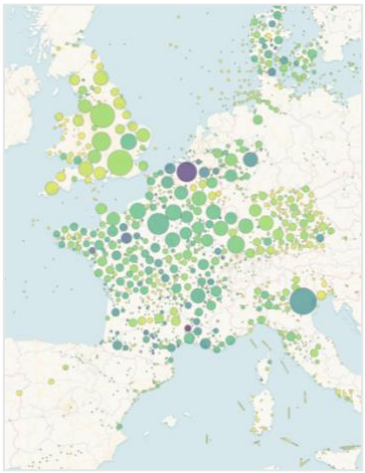
Cependant, face à la densité très élevée des points, tous les échantillons tendaient à être regroupés en un seul cluster, même en utilisant des rayons très faibles (de l'ordre de 10^{-63} m). Après analyse, le problème provenait de la fonction interne de l'algorithme qui fusionnait deux clusters sans tenir compte du rayon initial : elle recalculait automatiquement un rayon de cluster basé uniquement sur la racine carrée du nombre de points contenus. Par exemple, un cluster de deux points se voyait attribuer un rayon de √2 ≈ 1,41, ce qui correspond à environ 70 km en projection cartographique sur l'Europe. Cette distance, bien trop grande, expliquait les regroupements trop rapides.

Pour remédier à cela, j'ai défini une nouvelle fonction de calcul du rayon d'un cluster :

$$r = (\log10(\mathrm{n}+1) \times 0{,}97 + \mathrm{n} \times 0{,}03) \times 65000$$

Cette formule combine une croissance logarithmique avec une composante racine carrée plus douce, permettant d'obtenir un rayon qui évolue progressivement avec le nombre de points, sans devenir excessif.

Une fois ce problème résolu, un autre point est apparu : le rayon des clusters ne s'ajustait pas selon le niveau de zoom. En d'autres termes, même en zoomant sur une zone très localisée, les clusters gardaient leur taille d'affichage d'origine, ce qui nuisait à la lisibilité de la carte.

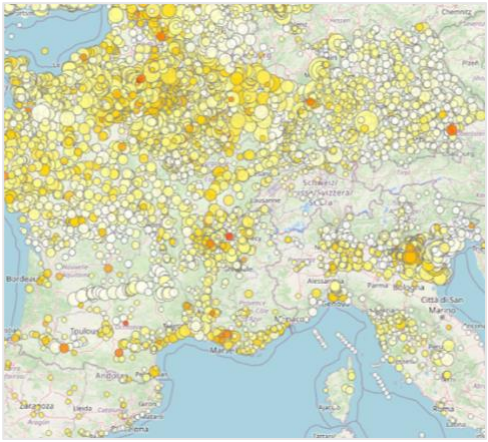


J'ai donc modifié la fonction précédente en introduisant une adaptation au niveau de zoom :

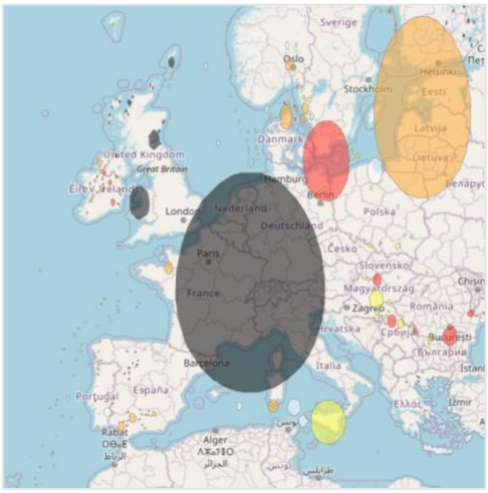
$$r = (\log10(\mathrm{n}+1) \times 0{,}97 + \mathrm{n} \times 0{,}03) \times 65000/zoom^{1.4}$$

Cette version permet d'ajuster dynamiquement la taille des clusters pour conserver une représentation claire et proportionnelle à l'échelle de la carte. Le résultat obtenu offre une visualisation bien plus lisible et adaptée à l'exploration des données.

Ci-dessous, quelques rendus intermédiaires illustrant des problématiques non détaillées ici, mais intéressantes à observer visuellement :



Exemple 1 — Trop de points proches les uns des autres : Les clusters ne se regroupent pas, même lorsqu'ils se chevauchent visuellement, ce qui nuit à la lisibilité globale.



Exemple 2 — Déformation des clusters : Il a fallu adapter l'affichage en tenant compte de la projection cartographique, pour éviter des distorsions importantes (clusters étirés ou « en forme d'œuf »), particulièrement visibles aux hautes latitudes.

Retour au rapport

10.3 Annexe performance base de données

Les 3 prochaines requêtes SQL sont importantes dans le chargement des données lors du lancement de mon application. Telles quelles, elles étaient longues à se charger et l'utilisateur devait attendre de longues secondes que l'outil charge. Nous avons créé des materialized views pour stocker le résultat de chacune de ces requêtes. Cela permet d'accéder aux données sans temps de calcul. Cela n'alourdi pas la base car chaque view stocke au maximum 300 lignes pour 5 colonnes maximum.

```sql
-- Cette requête renvoie tous les datasets représentés dans le jeu de donnée
-- et leur count (pour classer par ordre d'importance).
SELECT dataset_id, dataset_name, COUNT(*) AS count
FROM pfas_points
GROUP BY dataset_id, dataset_name
ORDER BY count DESC;
-- 1,3 sec d'execution

-- Récupération des mêmes données depuis la materialized view.
select * from mv_datasets_counts_for_init ;
-- 40 ms d'execution
```

Récupération de 118 datasets en divisant le temps de calcul par 32

```sql
-- Cette requête renvoie toutes les matrices représentées dans le jeu de donnée
-- et leur count (pour classer par ordre d'importance)
SELECT dataset_id, dataset_name, COUNT(*) AS count
FROM pfas_points
GROUP BY dataset_id, dataset_name
ORDER BY count DESC;
-- 1 sec d'execution

-- Récupération des mêmes données depuis la materialized view.
select * from mv_matrix_counts_for_init ;
-- 40 ms d'execution
```

Récupération de 13 datasets en divisant le temps de calcul par 25

```sql
-- Cette requête renvoie toutes les substances représentées (+ leurs unités) dans le
-- jeu de donnée et leur count (pour classer par ordre d'importance)
SELECT substance, COUNT(*) AS count, ARRAY_AGG(DISTINCT unit) AS units
FROM pfas_substances
GROUP BY substance
ORDER BY count DESC;
-- 13 secondes d'execution

-- Récupération des mêmes données depuis la materialized view.
select * from mv_substance_counts_for_init ;
-- 50 ms d'execution
```

Récupération de 291 datasets en divisant le temps de calcul par 260

Ces 3 materialized views ont permis de gagner 15 secondes de chargement au lancement de l'app.

J'ai aussi mis en place un index sur une de mes tables, il a permis de gagner du temps sur la récupération de points selon leur zone géographique. Moins il y a de points à récupérer, plus c'est efficace. Tout simplement car l'ordinateur ne parcours plus toutes les données, il filtre rapidement vers le pays voulu.

```sql
SELECT p.id, p.lat, p.lon, p.year, p.pfas_sum
FROM pfas_points p
JOIN pfas_nuts n ON p.id = n.point_id
WHERE n.version = 'nuts2016'
AND n.level = 0
AND n.nuts_id = 'FR';
-- FRANCE (660 000 données) avec index 3.4 sec - sans index 3.6 sec
-- BELGIQUE (150 000 données) avec index 1 sec - sans index 1 sec 5
-- ITALIE (50 000 données) avec index 0.4 sec - sans index 1 sec
-- ESPAGNE (1500 données) avec index 0.1 sec - sans index 1 sec
```

10x plus rapide pour les pays avec peu de données

Retour rapport

Présentation des PFAS et des problématiques liées à leur présence dans l'environnement.

Le développement suivant fait office d'annexe TE&DS (Transition Écologique & Développement Soutenable).

Que sont les PFAS ?

Les PFAS (substances per- et polyfluoroalkylées) forment une vaste famille de plusieurs millions de composés chimiques synthétiques. Leur particularité réside dans la présence de liaisons carbone-fluor, extrêmement stables, ce qui leur confère des propriétés recherchées dans l'industrie : résistance à l'eau, aux graisses, aux hautes températures.

C'est pourquoi on les retrouve depuis les années 1950 dans de nombreux produits du quotidien : emballages alimentaires, textiles imperméables, revêtements antiadhésifs (comme le Téflon), mousses anti-incendie, cosmétiques, etc. Mais cette même stabilité chimique fait aussi des PFAS des polluants dits "éternels", car ils se dégradent très lentement dans l'environnement.

- Les Nations Unies déclarent : « Les substances per- et polyfluoroalkylées (PFAS) sont des produits chimiques toxiques, artificiels et dangereux, qui ont des effets nocifs sur l'environnement et sur notre santé. » [10].

Pourquoi en retrouve-t-on dans l'environnement ?

Les PFAS sont relâchés dans l'environnement tout au long de leur cycle de vie : production industrielle, usage domestique, incinération ou enfouissement des déchets, rejets dans les eaux usées. Ces composés sont mobiles : ils voyagent dans l'air, l'eau, les sols, peuvent parcourir de longues distances, mais surtout, ils ne se dégradent pas !

> *« Le PFOS (sulfonate de perfluorooctane) et le PFOA (acide perfluorooctanoïque), dont les usages ont été très fortement restreints au niveau international, respectivement depuis 2009 et 2020, sont encore fréquemment mesurés dans l'environnement. » [10]*

Des investigations comme le Forever Pollution Project [17] ont révélé des niveaux inquiétants de pollution à proximité d'usines, d'incinérateurs ou de sites militaires, avec parfois des teneurs dépassant les seuils sanitaires recommandés de plusieurs centaines de fois.

- Environ 23 000 sites pollués ont été recensés en Europe [17].

Quelles en sont les conséquences ?

Sur l'environnement

Les PFAS contaminent les eaux de surface, les nappes phréatiques et les sols. Ils s'accumulent dans les plantes, les poissons et les animaux, et perturbent les écosystèmes.

Sur les cultures et l'alimentation

En s'infiltrant dans les sols agricoles, les PFAS peuvent passer dans les cultures. Certaines zones en France ont vu leurs produits interdits à la consommation, comme des œufs ou des légumes [16].

Sur la santé humaine

Les PFAS s'accumulent dans le corps humain et sont suspectés de provoquer divers effets : troubles hormonaux, cancers, baisse de la fertilité, affaiblissement du système immunitaire [12]. Certains sont classés comme cancérogènes probables (PFOA, PFOS).

- En France, 100 % des échantillons sanguins testés contiennent des PFAS [12].

Quelles mesures sont prises ?

En Europe

Les PFOS et les PFOA sont restreint par des régulations en Europe depuis 15 et 5 ans respectivement. Depuis 2023, via le règlement REACH, L'Union européenne a restreint d'autres PFAS et à établit une liste de « PFAS of very high concern » [13]. À partit d'août 2026, l'utilisation des PFAS dans les emballages alimentaires sera interdite dans l'UE. [25][14]

En France

Un plan d'action national a été lancé en 2023 avec plusieurs axes : renforcer la surveillance, identifier les sites pollués, réduire les émissions industrielles, informer le public [15]. Une loi adoptée en février 2025 interdira les cosmétiques, les vêtements, les chaussures et les farts pour les skis qui en contiennent à partir de 2026. En 2030, tous les textiles contenant des PFAS seront interdits. Un contrôle de l'eau potable, une carte des sites émetteurs de PFAS et une taxe pollueur-payeur sont aussi prévus [14].

Depuis 2024, les industriels doivent déclarer leurs rejets de PFAS dans les cours d'eau [24].

Renvoi vers le texte

Résumé

Ce stage, réalisé au sein du LIRMM à Montpellier, s'inscrit dans le domaine de la science des données appliquée à l'environnement. J'ai participé au développement d'un outil de visualisation de données sur les polluants PFAS, en collaborant avec des chercheurs de plusieurs domaines. Le stage m'a permis de mobiliser des compétences en développement web, en visualisation interactive, ainsi qu'en intelligence artificielle. J'ai également mené une phase exploratoire sur la représentation visuelle des résultats d'algorithmes de clustering pour faciliter leur interprétation par les utilisateurs.

Mots-clés

Visualisation de données, Développement web, PFAS, Intelligence artificielle, XAI, UX, Clustering, Analyse exploratoire.

Abstract

This internship, carried out at LIRMM in Montpellier, focused on data science for environmental research. I contributed to the development of a visualization tool for PFAS pollutant data, working alongside researchers from various scientific domains. The project allowed me to apply skills in web development, interactive visualization, and artificial intelligence. I also explored ways to visually represent clustering results to help users interpret them more easily.

Keywords

Data visualization, Web development, PFAS, Artificial intelligence, UX, Clustering, Exploratory analysis.