

M2 SDSC

UE sciences des données appliquées

Projet – 17-25/11/2022 – F. Le Ber, X. Dolques

Vous avez vu dans les semaines précédentes les caractéristiques de données concernant la qualité de l'eau des cours d'eau français. Pour ce projet, vous devrez mobiliser ces données et d'autres à votre disposition pour établir une analyse de ces données permettant de répondre à une question.

But du projet

Nous mettons à votre disposition un jeu de données contenant un ensemble de stations du réseau de contrôle et surveillance (RCS) des eaux douces de surface. Pour ces stations nous fournissons un ensemble de prélèvements biologiques présentés sous la forme de listes faunistiques et floristiques. Chaque liste est datée et contient un ensemble de taxons et leur dénombrement. Nous fournissons par ailleurs des données issues de prélèvements physico-chimiques. Ces données concernent les paramètres physico-chimiques les plus importants et les mieux renseignés, garantissant ainsi une quantité minimale de données.

Le but du projet est de proposer une réponse à la question « Quel lien peut-on établir entre la physico-chimie de l'eau et son état biologique ? ». Il s'agira dans un premier temps d'établir quels taxons sont influencés par certains paramètres physico-chimiques, puis dans un second temps d'affiner cette analyse en mobilisant les données géographiques : les taxons n'étant pas répartis uniformément sur le territoire, ou les données temporelles : l'impact sur la biologie n'étant pas forcément immédiat il peut être intéressant d'étudier l'évolution des données dans le temps.

Il est à noter que ces questions sont à l'heure actuelle des problèmes ouverts et qu'il n'existe par conséquent pas de réponse unique et absolue. Ainsi l'évaluation du projet ne portera pas tant sur un résultat que sur la cohérence de la démarche qui y a mené.

Rendu

Nous vous demandons de fournir à l'issue de ce projet un rapport décrivant votre démarche pour répondre à la question.

Il devra notamment décrire :

- **Le problème posé.** Le sujet est volontairement ouvert, il est nécessaire de bien cadrer la question que l'on se pose, d'identifier les limites imposées et celles que vous décidez d'ajouter.
- **Les données mobilisées.** Vous avez la possibilité d'enrichir les données fournies par d'autres données disponibles en accès libre. Lorsque vous intégrez à votre travail un nouveau jeu de données, vous êtes amenés à procéder à certaines opérations d'analyse et de transformation de ces données qu'il vous faudra décrire.

- **Les outils utilisés.** Vous avez le choix d'utiliser les outils et techniques qui vous semblent les plus pertinents. Nous vous conseillons pour l'analyse des outils logiciels tels que R ou Weka qui contiennent les implémentations de nombreuses techniques d'analyse issues de l'état de l'art mais vous êtes libres d'utiliser les outils que vous souhaitez en fonction des besoins et de votre maîtrise.
- **Les résultats obtenus** sous la forme qui vous apparaît la plus pertinente. Ces résultats devront amener à une discussion sur la démarche employée et ses limites. Pour mener ce projet à bien vous serez amenés à composer avec de nombreuses contraintes, notamment les données fournies, le temps qui vous est imparti et vos connaissances.

Organisation

Ce travail sera effectué en binôme, et le rapport devra être rendu avant le 20/12/2022 18h sur moodle.