M2 SDSC

UE Entreposage et protection des données

Projet – 3/10/2024 – F. Le Ber, X. Dolques

1 But du projet

Nous avons vu lors des séances précédentes comment récupérer et traiter les données de description de l'état physicochimique des stations de mesure issues du site naiades. Sur ce même site, vous trouverez des calculs d'indices sur l'état biologique des stations de mesure que vous devrez considérer. Vous vous intéresserez particulièrement à l'Indice Invertébré Multimétrique (I2M2). Nous vous fournissons des fichiers de données, un pour les données physico-chimiques, un pour les données hydrobiologiques et un fichiers décrivant les stations des prélèvements hydrobiologiques.

Les données hydrobiologiques vont de 2007 à 2022 et ne contiennent que les résultats du calcul de l'I2M2.

Les données physico-chimiques vont de 2005 à 2022 et contiennent les résultats des mesures des 16 paramètres physico-chimiques les plus mesurés.

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 ? ». Le sujet étant large nous vous proposons quelques pistes sur lesquelles vous engager :

- **Analyse de séquence temporelle pour un paramètre** : Les paramètres physico-chimiques évoluent dans le temps, on peut donc voir sur une station donnée une séquence temporelle par paramètre. On estime que l'état biologique à un instant T n'est pas le résultat d'un état physico-chimique au même moment mais de l'évolution de celui ci durant les mois voire les années précédant T. Choisissez 1 paramètre parmi les plus présents dans les prélèvements (par ex. nitrates dont on sait qu'il a un fort impact sur la biologie) et faites un clustering des stations ayant un état biologique proche (donc au moins 2 clusterings : un clustering des stations en bon état et un clustering des stations en mauvais état). Le but étant de caractériser le comportement de ce paramètre lorsque l'état biologique se dégrade ou reste bon.
- **Prédiction de note d'indice** : L'état biologique est dépendant d'un certain nombre de pressions parmi lesquelles on trouve la physico-chimie du milieu. Est-ce que pour autant les paramètres physico-chimiques sont suffisants pour obtenir la valeur d'un indice biologique ? Pour cela construisez un modèle de régression permettant de prédire une note biologique à partir de paramètres physico-chimiques.

Dans tous les cas, votre projet doit prendre en compte les dimensions temporelle et spatiale :

- Le temps : quand on observe la nature on est confronté à des évènements exceptionnels, des cycles saisonniers et à ce qu'un évènement n'ait pas toujours d'effet immédiat, c'est pourquoi il est important de ne pas se fier à un seul prélèvement mais plutôt à un intervalle de temps (aggrégé ou non) ; de faire attention aux dates (ne pas comparer des valeurs de février à des valeurs d'août) et de prévoir un délai quand il s'agit de voir un effet d'une pression sur un milieu.
- L'espace : le fonctionnement des cours d'eau n'est pas le même selon qu'il se trouve en Alsace ou dans les Alpes. C'est pour cela que les experts du domaine, plutôt que de travailler sur la France

entière, considèrent souvent des zones géographiques possédant des caractéristiques communes, notamment les hydroécorégions.

Pour compléter votre analyse, des données référentielles pourront être nécessaires. En voici une liste non exhaustive :

- États biologiques de l'I2M2 : vous trouverez la grille de conversion d'une note d'I2M2 en classe d'état biologique dans le Tableau 52, page 61 du JO n° 0198 du 28/08/2015 . Cette traduction en classes est complexe et dépend de la situation géographique, il faudra sans doute proposer une alternative simplifiée.
- Les données du Sandre : le site https://www.sandre.eaufrance.fr/ regroupe les données référentielles sur l'eau. Vous y trouverez notamment des donnés géographiques comme les hydroécorégions et les stations de mesure. Avec un SIG (par ex. QGIS) vous pourrez alors sélectionner les stations qui vous intéressent par zone géographique

Le projet consiste donc en grande partie dans la préparation de données avec des choix à motiver selon différentes contraintes (dont le temps de travail qui est limité). L'évaluation du projet ne portera pas tant sur un résultat que sur la cohérence de la démarche qui y a mené.

2 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 êtes libres d'utiliser les outils que vous souhaitez en fonction des besoins et de votre maîtrise. Mais gardez à l'esprit que l'évaluateur·rice ne les maîtrise pas forcément et que votre rendu devra permettre de comprendre sans avoir besoin d'exécuter votre projet. Par ex. si vous utilisez un notebook en python, vous devrez fournir un rapport PDF qui ne devra pas nécessiter d'ouvrir le notebook pour être compréhensible.
- **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 accessibles, le temps qui vous est imparti et vos connaissances.

3 Organisation

Ce travail sera effectué en binôme, et le rapport devra être rendu avant le 20/12/2024 18h sur moodle.