Chriss Osler Santi et Amen Amegnonan

October 18th 2019

Taches 1: Problemes posées dans l'organisation des fichers

- Schema des fichiers

  Nous constatons une différence dans les schémas des fichiers `pond2010.xlsx` et `zoo-*.xlsx` . Cette différence doit être prise en compte lors de l'intégration.

- Colonne La date

  Les chercheurs veulent étudier la migration des planctons suivant les periode de la journée. Cependant dans les données, la colonne date ne mentionne que le jour du prélèvement des échantillons; rendant quasi impossible de savoir à quel période de la journée ces échantillons ont été prélevé.

  Cette remarque est maintenue dans tous les fichers.

- Colonne température

Fichier `pond2010.xlsx`

- Il manque des metat-data pour expliciter certaines colonnes des données.
- Ce fichier comporte une colonne nommée `z` qui n'est pas explicite.
- La colonne `Temp` de ce fichier comporte des données manquantes.

Fichier `zoo-temp.xlsx`

- Dans ce fichier nous retrouvons des données.

	cuni	**chippo**
7th	2.39533333	2.85933333
9th	2.4	2.842

Ces données sont sans réfenrences

Fichier `zoo-temp-main.xlsx`

- La colonne `Temp` comporte des données manquantes, et de plus un asterix apparait dans cette colonne. Le nom du fichier nous indique que probablement, ce fichier est un fichier temporraire et donc sera plutart revue.

Tache 2: Suggestion

Pour un meilleure intégratione et gestion des données nous proposons:

- De garder le format du fichier `pond2010.xlsx` soit: `["Date", "Hour", "Depth", "Temperature", "Density", "ColonyDiameter", "Species"]` avec:
    - **Date**: La date courte dd/mm/yyyy du prélèvement.
    - **Hour**: L'heure du prélèvement.
    - **Depth**: La profondeur en mètre.
    - **Temperature**: La température en C.
    - **density**: Le nombre de plancton par litre.
    - **ColonyDiameter**: Le diametre de la colonie en Cm.
    - **Specie**: Le nom de l'espèce du plancton.
- Dans le cas de manques des heures de prélèvement, nous proposons de les générer.
    - Pour cela nous avons considérer les heures 06:00 - 20:00.

Tache 4: Suggestion de détection de données anormales:

- Statistiquement nous pouvons détecter des anomalies, (specialement des valeurs abérentes) en calculant la variation des données. Pour ce cas, nous supposons que les données suive une distribution normale. et toute données qui varie de 2x la variance, est supposée comme donnée abérante.

- Une deuxième méthode est de représenter les données par un scatter plot et effectuer une régression. Les données abérante dans ce cas sont les plus éloigné (par un facteur), de la ligne de régression.

Tache 5: Suggestion de détection de données manquantes:

- Une représentation graphique (plot, scatter plot), permet de vizulaiser les données, mais aussi de détecter des donner manquantes. En effet la présence de rupture dans la courbe signale, un manque de données.

- Nous pouvons avoir un script aussi, qui scanne les colonnes et retourne les colonnes comportant `N/A` ou `null` ou `empty` selon les metadata.

Tache 6: Suggestion de prévention d'erreurs:

- Nous pouvons prevenir les erreurs en automatisant l'entrée de données, par examples via des feuilles de calculs exels déjà pré-remplies.