Récapitulatif rencontre 12/09/2023

Le sujet fait abstraction de domaine d'application, il vise à déterminer la présence d'une espèce (ici végétale) dans une zone de l'espace, il convient de définir les rayons des zones (ou peut-être dans les données).

Ce type de projet permet de contourner des problèmes d'observations protocolées à la fois coûteuses et chronophages. Le projet fait appel à la fois aux données récoltées scientifiquement : observations protocolées avec présence / absence et aux données citoyennes obtenues via des applications type PlantNet avec présence uniquement.

La présence d'une espèce dépend de 2 grands domaines : biotique et abiotique. Seule la partie abiotique sera étudiée au début. Il s'agit de la chimie du sol, la topologie, la température etc. La partie biotique concerne la corrélation entre les espèces, la présence d'ancêtre, l'introduction et la conservation d'espèces par l'homme ➔ des données difficiles à mesurer et encore plus à prédire.

Le modèle est amené à évoluer graduellement en précision :

1. Machine Learning basique (« naïf ») éventuellement au hasard, répétition des 50 espèces les plus présentes…
2. A l'aide de fonction type noyau ou K plus proches -> la distance devra être évaluée en tenant compte de la courbure de la terre et influera sur la probabilité de trouver l'espèce.
3. Description des environnements selon leurs caractéristiques (sans trop entrer dans l'unicité de chaque lieu) et possible présence de l'espèce selon les conditions de l'environnement. ~ Deep Learning
4. Deep Learning, à la différence d'un modèle simple à une espèce : modèle complexe ou toutes les variables, les espèces sont prises en compte.

Il faut prêter attention à l'autocorrélation, on ne doit pas cacher une partie des données au modèle mais le tester sur des nouvelles zones ➔ Test Kaggle

Les données sont disponibles en intégralité sur Kaggle (Geolife CLEF 2023) avec un jeu d'entrainement et de test. (Résultats au test : [0 ;0,1] = bug , [0,1 ;0,2] = mauvais mais pas de bug , [0,2 ; +infini[= bon modèle. Il peut être intéressant de tester un modèle naïf pour voir le score obtenu).

Exploration de données :

- Cartes de la répartition des espèces : 1 par espèce ou du moins pour les 50 plus présentes (boucle).
- Cartes des différentes données et essayer de repérer des distributions proches de la répartition des espèces.
- Déterminer les données inutiles.
- Utilisation de SVMs.

- Si présence de dates dans les données : les rendre linéaire pour que 31 décembre précède 1 janvier.