Les données et leur exploration :

Source de données :

Notre source de données est unique et très complète, il s'agit d'une collection de fichiers csv et de fichiers tif disponible sur Seafile Repository - Plantnet. Cette collection est une agrégation de plusieurs sources avec d'une part les relevés de présences / absences réalisés par des experts sur le terrain ayant vérifié toutes les espèces de plantes, d'une autre part les relevés de présences uniquement fournis par participation citoyenne via l'application PlantNet, enfin les images satellites et les séries temporelles pour 6 satellites.

Les données de présences / absences sont extrêmement coûteuses en termes d'argent et de temps à récolter, c'est pour cela qu'un modèle prédictif performant serait intéressant pour palier à ce problème. Les données de présences uniquement ne sont pas particulièrement couteuses mais très incomplètes : une personne se baladant ne peut pas vérifier l'intégralité d'une zone pour répertorier toutes les espèces, de la même façon il est possible que seules les espèces « notables » attirent l'attention des utilisateurs de PlantNet, discriminant les espèces par leur rareté ou leur beauté.

Description des données :

Présences / absences :

« enviroTab_pa_Train.csv » : 5949 lignes 48 colonnes. Contient pour chaque PatchId (Id de zone) les données environnementales associées : année, jour de l'année, coordonnées gps, 19 variables bio, 9 variables de composition du sol, 11 variables sur la population et impact humain.

« Presence_absence_train.csv » : 85326 lignes 15 colonnes. Contient une observation de SpeciesId (Id de l'espèce) à un PatchId et un timeSerieId (Id de série temporelle) avec les coordonnées géographiques , le jeu de donnée d'origine, les personnes à l'origine de l'observation, la date d'observation.

« train_pa_spAround_1km.csv » : 58040 lignes 4 colonnes. Contient pour chaque patchID, speciesId et jour de l'année le nombre d'observation de l'espèce.

Présences uniquement :

« Présences_only_train.csv » : 4 908 319 lignes 15 colonnes. Contient un patchID, un speciesId, un timeSerieID associés à des coordonnées gps, une date et la source de la donnée ainsi que son auteur.

Baseline_runs :

4 fichiers csv : « run_constant_baseline.csv », « run_enviro_RF.csv », « run_maxent_on_pa.csv » et « run_spatial_RF.csv » qui sont des fichiers exemple de soumission Kaggle avec une colonne Id et une colonne Predicted qui est une liste d'espèces. 22405 lignes et 2 colonnes.

For_submission :

Dossier contenant les 3 jeux de données de test.

« enviroTab_pa_test.csv » : 22405 lignes et globalement les mêmes colonnes que « enviroTab_pa_train.csv » avec un ID et sans le speciesId.

« test_blind.csv » : 22405 lignes et globalement les mêmes colonnes que « Présence_absence_train.csv » avec un ID et sans l'espèce

EnvironnementalRasters :

5 sous dossiers « Climate », « Elévation », « HumanFootprint », « LandCover » et « Soilgrids » qui contiennent des documents .tif respectivement sur :

- le climat mensuel (de 2000 à 2019) et le bioclimatique Average de 1981 à 2010
- L'élévation (altitude)
- L'empreinte humaine (la lumière, la densité de population, présence de route, chemins de fer…)
- La couverture terrestre (présence d'eau, utilisation de la terre, ...)
- La composition du sol (sable en g/kg, pH de l'eau, …)

SatelliteImage :

2 dossiers zippé de 17Go d'images jpeg. Le premier est intitulé « patchs_nir.zip » pour les images infrarouges, le second est intitulé « patchs_rgb.zip » pour le spectre de lumière visible.

SatelliteTimeSeries :

6 fichiers csv regroupes en 1 fichiers csv zippé intitulés « time_series_ » « blue », «green », « red », « nir », « swir1 » et « swir2 ». Ils contiennent des valeurs associées à un TimeSeriesID pour chaque mois depuis Janvier 2000 jusqu'à Avril 2020. Ces valeurs correspondent à la bande spectrale en question (bleu, rouge, vert, infrarouge, …)

Visualisation des données :

Notre analyse porte sur les données de présences absences principalement et sur les données de présence uniquement.

Une partie de l'analyse existe sous forme de carte car nos données sont majoritairement liées à des coordonnées et qu'elles prennent plus de sens sous cette forme qui aide à la compréhension et l'interprétation. Les cartes sont celles de la répartition des 50 espèces les plus présentes, une série de cartes par variable environnementale échelonnées selon leurs valeurs ainsi qu'une carte de répartition des observations présences uniquement.

L'autre partie de l'analyse est contenue dans un rapport html réalisé via RStudio, il comporte une ACP sur les données environnementales avec un habillage par variable ainsi que son interprétation. Ce rapport comporte aussi des graphiques sur la répartition temporelle des observations. L'ACP nous permet d'établir des liens entre les variables et la réduction de dimensions est envisageable pour construire des modèles de prédiction.

Ces documents sont disponibles en annexes du rapport.

Nettoyage et sélection des données :

Les données étant relativement propres, il n'y a pas eu beaucoup de traitement en amont. Dans le jeu de données des variables environnementales d'entraînement, certaines variables possédaient des NA alors pour les modèles les utilisant nous avons supprimé ces lignes.