

Rapport Projet Logiciel DAC. GeoLifeClef.

Karmim Yannis, Bourcier Jules

2 juin 2019



Encadrants : Soulier Laure et Zablocki Eloï .



Table des matières

1	Introduction.	3
1.1	Motivation	3
1.2	Présentation de la campagne LifeClef.	3
2	Présentation de la tâche.	4
3	État de l'art.	5
3.1	Modèles Machine Learning de l'édition 2018.	7
3.1.1	Équipe ST.	7
3.1.2	Team FLO.	7
3.1.3	Discussion.	8
3.2	Description des modèles intéressants pour notre tâche.	8
3.2.1	K plus proches voisins.	8
3.2.2	Rayon de voisins.	8
3.2.3	Clustering de groupes d'espèces.	8
3.2.4	Random Forest.	9
3.2.5	XGBoost : eXtreme gradient boosting	9
4	Description, analyse des données et difficultés du problème.	10
4.1	Présentation des données.	10
4.1.1	Données environnementales.	10
4.1.2	Données des observations.	13
4.2	Difficultés de la tâche liée au données.	15
4.2.1	Quelques statistiques	15
4.2.2	Les principaux problèmes liés aux données.	15
4.3	Visualisation	16
4.3.1	Histogramme des espèces.	16
4.3.2	Visualisations des occurrences sur une cartes.	17
4.3.3	Corrélation des variables.	18
4.3.4	Analyse en composantes principales.	19
4.3.5	Visualisations par PCA et T-SNE.	20
5	Apprentissage de nos modèles.	22
5.1	Protocole d'extraction des données et pré-traitements.	22
5.2	Les métriques d'évaluation de notre projet.	23
5.2.1	MRR.	23
5.2.2	Top-30.	23
5.3	Protocole de sélection et d'évaluation des modèles.	24
6	Analyse et discussion des résultats.	25
6.1	Résultats.	25
6.2	Analyse comparative avec les participants 2018.	26
6.3	Idées d'analyses supplémentaires de nos résultats.	27
7	Conclusion et perspectives.	28
7.1	Améliorations.	28
8	Bibliographie.	29

1 Introduction.

1.1 Motivation

En bio-informatique, la prédiction automatique d'espèces que l'on peut observer à un endroit précis est un problème complexe et une tâche utile dans beaucoup de situations. Cela permet tout d'abord de mieux répertorier et classer les espèces en fonction de l'environnement dans lequel on se situe, mais également de développer des systèmes et des logiciels pour des utilisateurs non-experts qui souhaiteraient observer la faune et la flore autour d'eux.

Cela peut améliorer les outils d'identifications d'espèces, et aider les biologistes experts en réduisant la liste possible des espèces observable sachant une localisation.

Enfin cela peut être utile à des fins pédagogiques en apprenant la biologie végétale et animale aux utilisateurs de manière contextuelle. Ces systèmes et méthodes peuvent également servir à la protection des espèces en identifiant les environnements fragiles et les espèces qui s'y trouvent.

1.2 Présentation de la campagne LifeClef.



FIGURE 1 – Logo de la campagne LifeClef.

Dans le cadre du PLDAC nous avons choisi de participer à la campagne d'évaluation LifeClef pour effectuer la tâche sur la prédiction de localisation d'espèces : **GeoLifeClef**.

La campagne CLEF sont des conférences indépendantes évaluées par des chercheurs sur un large éventail de questions dans les domaines du Machine Learning et de la Recherche d'Information, comme le Traitement Automatique du Langage, la reconnaissance d'images ou encore la classification audio.

Quant au projet LifeCLEF il a pour objectif de stimuler la recherche sur l'identification et la prévision des organismes vivants afin de combler le fossé taxonomique et d'améliorer nos connaissances sur la biodiversité.

Il existe plusieurs tâches à l'intérieur de ce projet :

1. **PlantClef** : Identification de plus de 10000 espèces de plantes à partir d'images.
2. **BirdClef** : Détections d'espèces d'oiseau à partir d'enregistrement audio.
3. **GeoLifeClef** : Prédiction de localisation d'espèces (de plantes) à partir de données environnementales.

2 Présentation de la tâche.

Dans cette section nous allons vous présenter la tâche GeoLifeClef. Il s'agit seulement de la seconde édition cette année 2019. Le but du challenge est de prédire une liste d'espèces sachant

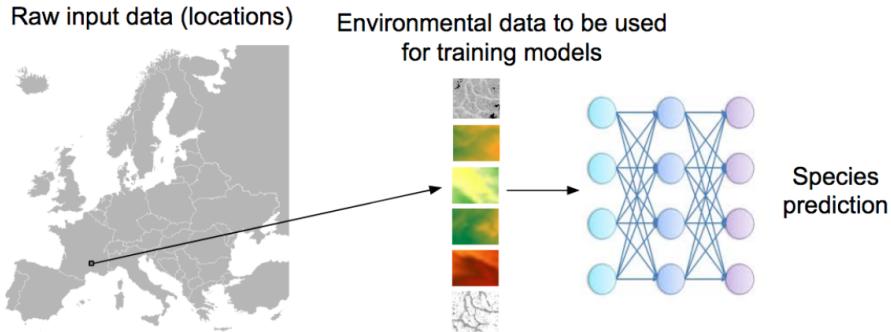


FIGURE 2 – Illustration de la tâche.

une localisation sur une carte de la France.

On possède une grande base de données d'espèces pour entraîner nos modèles, et chaque occurrence est accompagnée, au minimum, de la localisation de l'observation ainsi que l'identifiant de cette espèce.

Cependant il n'est pas possible d'apprendre la distribution des espèces directement des informations de localisation à cause du peu d'occurrences que l'on a pour certaines espèces et du biais d'échantillonnage puisque l'on a peu d'occurrences ou voir même pas du tout à certaines localisations)

Ce qui est fait en écologie c'est de partir de la base d'une représentation d'un environnement. Typiquement un vecteur composé de la température moyenne de la précipitation ainsi que d'autres variables comme le type de sol, la couverture terrestre, la distance à un point d'eau. L'originalité de GeoLifeClef est d'utiliser des images pour représenter directement ces variables environnementales, ce qui ajoute donc un problème d'extraction de données sur ces images. Ces patchs d'images sont récupérées à partir de la localisation (latitude,longitude), ce sont 33 patchs de taille 64*64 pixels qui sont extraites à partir de cartes fournies par les organisateurs du challenge.

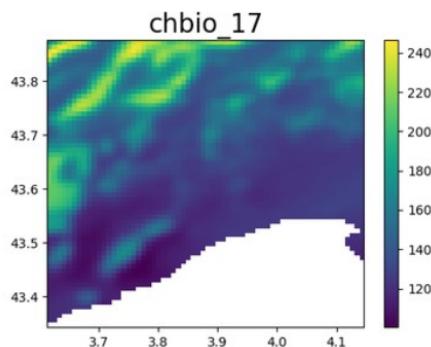


FIGURE 3 – Exemple d'une des images de variables d'environnement pour une localisation.

Chaque patch représentant une valeur de variable de l'environnement (température moyenne annuelle, max température du mois le plus chaud , précipitations annuel, l'altitude , la proximité à de l'eau fraîche, la capacité d'eau disponible...). Nous détaillerons dans la **section 4** les données.

On se ramène donc à un problème de machine learning de classification multi-classes, puisque chaque espèce différente représente une classe. Chaque image d'une variable d'environnement représente une dimension dans nos données.

3 État de l'art.

Dans cette section nous allons détailler notre recherche documentaire pour ce projet afin de montrer l'état de l'art existant sur ce problème de prédiction de localisation d'espèces.

Cela va nous servir à bien comprendre les enjeux et difficultés de notre tâche, mais également nous orienter vers des modèles intéressant pour notre projet.

La particularité de notre projet est qu'il s'agit d'un challenge proposé par la campagne d'évaluation annuelle CLEF. On dispose donc déjà de tout les working notes et références des années précédentes que les organisateurs mettent en ligne à ce lien : <http://ceur-ws.org/Vol-2125/> [1][2][3][4].

Les articles des participants des années passées ont donc été le fondement de notre travail de recherche, puisque ils ont été confrontés aux mêmes problématiques et difficultés que nous. De plus, il est souvent conseillé de s'appuyer sur des modèles et méthodes déjà performantes et existantes pour ces challenges.

Dans la plupart des working notes des participants, ils citaient également des sources primaires sur la théorie des modèles qu'ils utilisaient, par recherche par rebond on a pu accéder à plusieurs sources intéressantes.

Notre recherche bibliographique a donc surtout été essentielle pour la compréhension de ces modèles, puisque beaucoup des techniques utilisées dans le challenge sont des concept nouveaux pour nous.

Il nous a fallu alors comprendre les enjeux de notre projet, ainsi qu'à quelle domaine de l'apprentissage statistique cela se référait.

Pour cela on a trouvé pertinent de construire une carte heuristique afin de guider nos recherches supplémentaires.

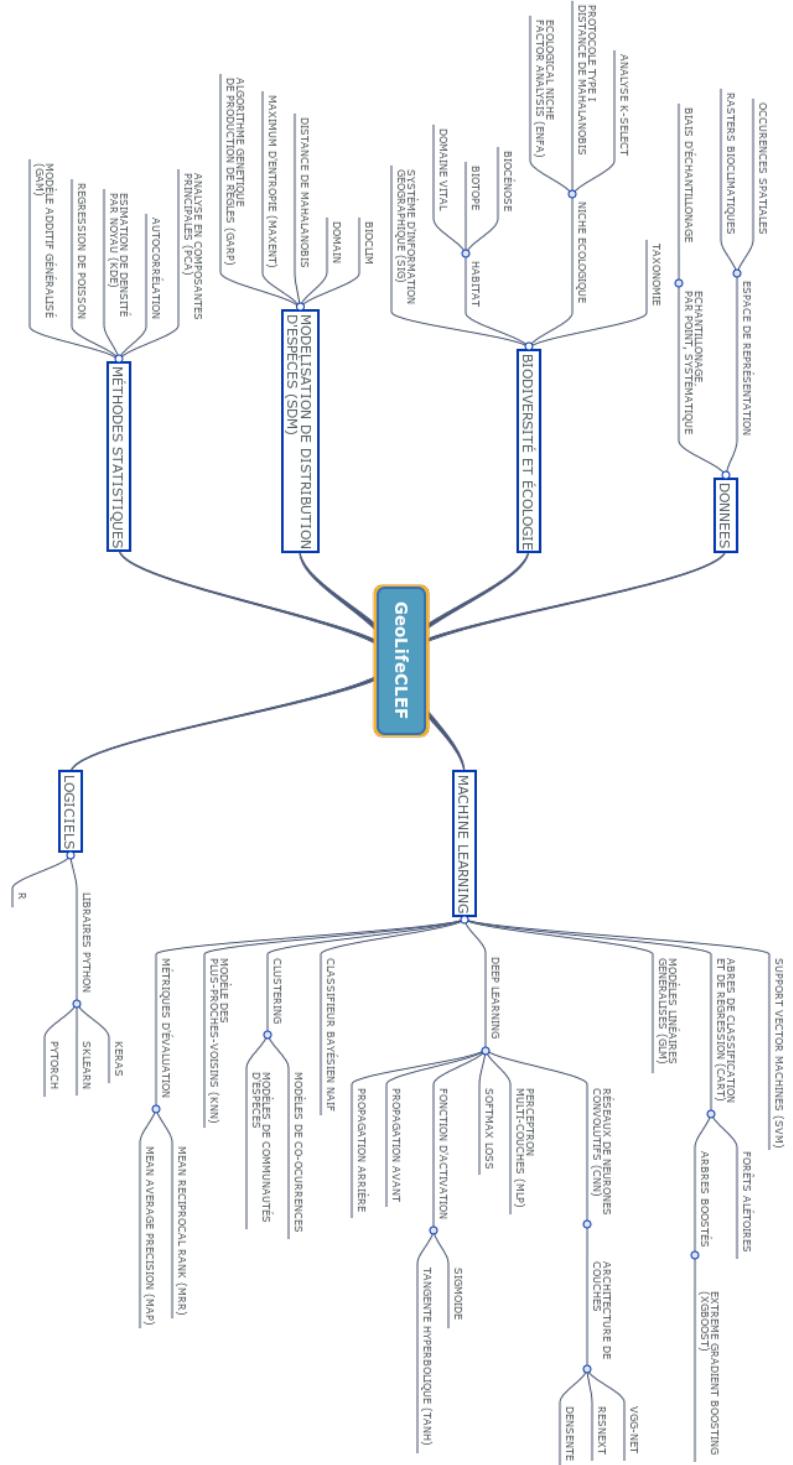


FIGURE 4 – Carte heuristique pour le projet GeoLifeClef.

3.1 Modèles Machine Learning de l'édition 2018.

Sur les 22 groupes ayant participé au challenge de 2018, seuls 3 ont rendu des résultats.

- Team ST
- Team FLO
- Team SSN

Dans cette partie nous vous présentons quelques modèles intéressants des équipes de l'édition 2018.

3.1.1 Équipe ST.

Boosted Trees. L'équipe ST a développé des modèles à base de plus proches voisins ainsi qu'un modèle XGBoost (eXtreme Gradient Boosting).

Avec cette méthode il affirme qu'il est possible de prédire une variable cible basé sur les caractéristiques des données d'apprentissage.

Le boosting est une technique utilisée en apprentissage statistique qui a pour but de combiner des classificateurs simple séquentiellement, à quoi on attribue un certain poids, afin d'optimiser les performances de prédiction.

Le gradient boosting a pour particularité d'utiliser le gradient de la fonction de perte pour calculer les poids des classificateurs. Les classificateurs ici sont généralement des arbres binaires de régressions CART de profondeur faible.

Le boosting est une technique robuste au sur-apprentissage.

Réseaux de neurones. La team ST a utilisé un réseau de neurones convolutif pour le challenge, il s'agit en effet de modèle très prometteur pour les modèles de distribution d'espèces (voir [5]). Cependant en utilisant des CNN aux architectures prédéfinis, comme VGG ou DenseNet, n'a pas vraiment porté ces fruits et sont moins performants que les performances obtenues sont assez médiocres et moins bonnes que des méthodes d'ensemble les random forests ou Xgboost. [3]

3.1.2 Team FLO.

Random Forest. En plus d'un modèle de plus proche voisin similaire à la team ST, la team FLO a soumis un résultat en utilisant une Random Forest directement implémenté de *Scikit-learn*. Ce modèle est intéressant car il fournit souvent de bons résultats. Ils utilisent uniquement la position géographique en entrée, sans prendre en compte les rasters d'environnement. Pour chaque occurrence du fichier de test il renvoie un top-100 des espèces les plus probables sachant la localisation. Les hyper-paramètres optimaux trouvés après cross-validation sont une forêt de 50 arbres de profondeur 8.

Réseaux de neurones.

CNN. Contrairement à la team ST, les réseaux de neurones de la team FLO ont été bien plus efficaces et ont remportés le challenge de l'édition 2018.

Les modèles de Machine Learning précédent nécessitaient de réduire les tenseurs 64*64*33 pixels à un vecteur de taille 33. Avec les réseaux de neurones convolutifs il est possible désormais d'apprendre directement sur l'ensemble des patchs.

Ils ont utilisés une fonction de coût soft-max contrairement à ce qui est utilisé en [5] (régression de Poisson). Ce qui est plus pratique également pour donner un classement des espèces les plus

vraisemblablement observable à un point donné.

La team FLO a utilisée une architecture customisée contrairement aux autres équipes. L'architecture peut être consultées dans leur working notes (cf [2])

3.1.3 Discussion.

Les CNN sont prometteurs pour cette tâche mais nécessite un travail de modification de l'architecture du réseau. Si ce n'est pas le cas, des modèles tels que XGBoost ou Random Forest peuvent dépasser les CNN.

XGBoost est un modèle très répandu en SDM, et il semble pouvoir donner de bonnes performances pour cette tâche.

La team SSN n'a pas eu de bons résultats comparés aux deux autres équipes. Ils ont essayé une approche sur les groupes taxonomiques des plantes en essayant de regrouper les espèces par famille, sous genre... Mais cela n'a pas été très concluant[4].

3.2 Description des modèles intéressants pour notre tâche.

Dans cette partie nous allons faire une courte présentation théorique des différents modèles intéressant pour le projet GeoLifeClef.

3.2.1 K plus proches voisins.

Simple modèle de K-nn : pour un point en test, le modèle calcule la distribution de probabilité des classes parmi les voisins, et retourne les classes par probabilités décroissantes.

Paramètres :

- Métrique utilisée, 'euclidienne' par exemple.
- Poids des voisins dans le calcul de la distribution de probabilités :
 - 'uniform' : c'est à dire que chaque voisin a un poids de 1.
 - 'distance' : le poids d'un voisin est l'inverse de la distance ; dans ce cas, les points proches pèsent plus que ceux éloignés.

Ces deux paramètres sont importants, et il faut les optimiser par une random search ou grid search.

3.2.2 Rayon de voisins.

Variante de Knn où ce n'est plus le nombre de voisins en paramètre mais le rayon du cercle autour du point.

Paramètres : métrique utilisée et poids des voisins comme pour K-nn.

Problème La prédiction est coûteuse, mais il n'y a pas d'étape d'apprentissage.

3.2.3 Clustering de groupes d'espèces.

Cela consiste à créer des clusters de groupes d'espèces.

1. Pour chaque label, construire un vecteur caractéristique agrégant les exemples en apprentissage correspondant à cette classe (par une moyenne ou une médiane par exemple). On obtient ainsi un espace non labélisé où l'on peut faire du clustering
2. Faire un clustering sur ce nouvel ensemble vecteurs, avec K-means (hard ou pondéré), un modèle de mixture de gaussiennes, ou un modèle de clustering spectral en dernier

choix, afin de construire des groupes d'espèces qu'on observe dans des environnements similaires.

3. Pour chaque cluster, on fait correspondre un modèle sur les exemples en apprentissage correspondant aux labels de ce cluster (par ex, un Knn ou random forest).
4. Pour un exemple en test, on l'affecte, en hard ou soft, aux clusters créés, et on prédit la distribution de probabilités des labels en pondérant par l'affectation à un cluster, la prédiction du modèle sur ce cluster.

On peut faire une version hard avec K-means puis un Knn sur ce cluster d'affectation.

Paramètres : Nombre de clusters, modèles utilisés sur les clusters.

3.2.4 Random Forest.

Une forêt aléatoire construit un ensemble d'arbres de décision sur des sous ensembles d'exemples et de paramètres pris aux hasard. La distribution de probabilités pour un point est prédite en moyennant les prédictions de chacun des arbres.

Ce modèle permet de diminuer la variance souvent élevée des arbres de décision afin de mieux généraliser les prédictions.

Paramètres : nombre de classifieurs, profondeur maximum, minimum d'exemples pour séparer un noeud.

Pour régulariser on peut contrôler le nombre d'arbre, entre 50 et 1000 par exemples, et la profondeur maximum entre 3 et 10.

Variation : Extra-trees random forest, où l'attribut dans le noeud à chaque niveau de construction est choisi au hasard. Permet de diminuer encore la variance en pouvant cependant augmenter le biais.

3.2.5 XGBoost : eXtreme gradient boosting

XGBoost est un modèle de boosting de gradient amélioré (tree pruning, fonction de loss personnalisé...). Il combine des classifieurs faibles (petits arbres de décision) pour construire un classifieur fort. XGBoost est considéré comme un modèle état-de-l'art en machine learning particulièrement adapté à la modélisation de distributions d'espèces.

Paramètres : Profondeur max, nombre d'étapes de boosting, learning rate, fonction de coût à minimiser.

4 Description, analyse des données et difficultés du problème.

4.1 Présentation des données.

Chaque participant obtient des données d'apprentissage et de test des occurrences des espèces géo-localisées ainsi que de variables d'environnement.

Les participants ont reçu une série d'apprentissage et une série de test d'occurrences géo-localisées d'espèces. Les deux étaient d'abord composés d'un fichier .csv contenant les coordonnées spatiales des occurrences, les valeurs ponctuelles des variables environnementales à l'endroit de l'occurrence et, pour le tableau de l'apprentissage, le nom de l'espèce et son identification. Ces observations sont issues de données d'utilisateur de l'application pour smartphone Pl@ntNet qui prédit automatiquement des espèces photographiées. Deuxièmement, chaque ligne du tableau (train et test) renvoie à une image à 33 canaux contenant le tenseur environnemental extrait à cet endroit.

Seules les données observées sur le territoire de la France métropolitaine ont été gardées. Le label qu'on doit prédire dans le test est dans le (field species GLC19SpId). Toutes les espèces différentes sont associées à leurs noms scientifiques qui sont donnés dans le champ scName.

4.1.1 Données environnementales.

Chaque occurrence pour un environnement est caractérisé par 33 images de 64 * 64 pixels. Ces variables d'environnement ont été construites à partir de sources diverses. On a donc pour chaque occurrences des tenseurs de taille 64 * 64 * 33 pixels.

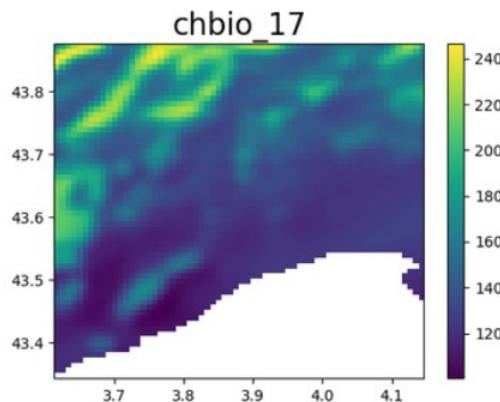


FIGURE 5 – chbio_17 décrit le taux de précipitation de la période la plus sèche de l'année

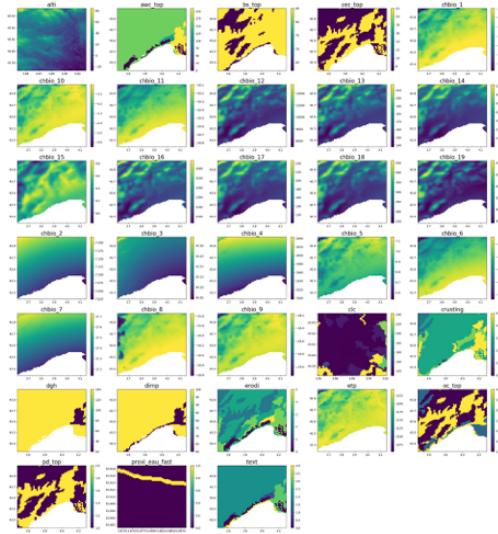


FIGURE 6 – Exemple de 33 images décrivant un environnement à une localisation précise.

Il est à noter que les images ne sont pas des photographies, mais des représentations de valeurs liés à une caractéristique d'environnement autour d'un point (latitude,longitude).

Il y a 4 types de variables :

Les variables quantitatives *Nature : quanti* (*Figure 7*).

Les variables ordinaires *Nature : ordinal* (*Figure 7*) qui peuvent être traitées dans nos classifieurs comme une variable réelle donc cela ne pose pas problème.

Les variables booléennes *Nature : boolean* (*Figure 7*). Elles peuvent également être encoder comme une variable réelle facilement.

Les variables catégorielles *Nature : Categorical* (*Figure 7*). Il s'agit de la variable clc, cela va poser un problème dans nos données puisqu'il s'agit d'une description catégorique de l'environnement, c'est à dire comme "homme ou femme" par exemple.

Il n'existe pas vraiment d'ordre comme les variables ordinaires. La solution est donc soit de faire du *0-1 encoding* mais cela va énormément augmenter nos dimensions puisqu'il y a plus de 50 valeurs différentes possibles, soit de ne pas prendre en compte cette variable dans notre apprentissage.

4.1.2 Données des observations.

Ensemble de test. L'ensemble de test du challenge n'est pas fourni de base, il n'est publié qu'en mars. Pour chaque espèce présentes dans le test il y a au moins une observation dans l'apprentissage.

Une observation d'espèce dans le test est distante d'au moins 100 m des toutes observations de cette espèce dans l'ensemble d'entraînement pour éviter que ce soit trop facile à prédire.

Dans l'ensemble de notre projet puisque nous n'avons pas soumis de résultat au challenge, nos évaluations et tests ont été faits en divisant nos données d'apprentissage.

Ensembles de données d'occurrences. Tous les datasets ont subis un filtrage taxonomique et géographique. Cela garanti que les noms taxonomiques correspondent au *Taxref v12 referential* et que les points sont tous situés sur le territoire, avec une imprécision de 30 mètres maximum.

Pl@ntNet complete dataset.

(Fichier **PL_complete.csv**) Le dataset Pl@ntNet complete est contient 2,377,610 observations incertaines. Le champ *FirstResPLv2Score* donne la confiance du score de prédiction automatique des espèces identifiées. Ce dataset est très hétérogène dans la qualité des identifications. Le champ *accuracy* donne l'incertitude en mètres calculée principalement par les smartphones.

Dates : début 2017 - novembre 2018.

Pl@ntNet trusted dataset.

(Fichier **PL_trusted.csv**) Le fichier sur lequel nous avons travailler. Un filtre de confidence dans l'identification a été appliqué au Pl@ntNet complete dataset.

Les occurrences gardées sont seulement celles ayant une probabilité d'identification pour la première espèce supérieure à 0.98. Ce score a été déterminé par des experts pour donner un degré raisonnable de vraisemblance dans l'identification. Ça a enlevé 90% des occurrences. Ce set de 237,087 occurrences couvrant 1,364 espèces avec une géolocalisation et une identification précise n'a jamais été utilisé précédemment.

Dates : début 2017 - novembre 2018.

GeoLifeClef 2018 dataset.

(Fichier **GLC_2018.csv**). Ce set contient 281,952 occurrences couvrant 3,231 espèces. Avec ce dataset, les occurrences sont souvent agrégées au même point géographique, ce qui dénote des géolocalisations dégradées. Le champ *coordinateuncertaintyinmeters*, quand présent, informe sur l'incertitude de localisation.

Dates : inconnues

NoPlant dataset.

(Fichier **noPlant.csv**) contient 5,771,510 occurrences couvrant 23,893 taxons. Aucune de ces espèces n'apparaîtra dans l'ensemble de test. Mais ce dataset complémentaire peut être utilisé pour améliorer la puissance prédictive des modèles en utilisant les corrélations fortes entre espèces de plantes et les autres taxons.

Structure commune du dataset.

Toutes les occurrences de train sont données dans des fichiers CSV séparées par ";" avec une

en-tête. Les trois champs d'intérêts sont les suivants

glc19SpId : L'identifiant de référence GLC19 pour les noms d'espèces. La correspondance entre les identifiants et les noms scientifiques d'un base taxonomique, Taxref, est données dans un fichier annexe.

Longitude : longitude décimale dans le système de coordonnées WGS84.

Latitude : latitude décimale dans le système de coordonnées WGS84.

4.2 Difficultés de la tâche liée au données.

4.2.1 Quelques statistiques

Le dataset Pl@ntNet Trusted - les observations dont la probabilité d'identification de l'espèce est supérieur à 98%. Il y a 237,087 occurrences, couvrant 1364 classes, et 35% de points superposés (84,524).

GLC_2018 dataset : Le dataset de l'an dernier de bases de données naturalistes (géolocalisations dégradées ou incertaines) 281,952 occurrences, couvrant 1,364 classes, et 81% de points superposés.

4.2.2 Les principaux problèmes liés aux données.

Pour Pl@ntNet Trusted, on voit que 84,524 points, 35% du total, sont superposés entre eux. La difficulté de GeoLifeCLEF est principalement dans la superposition des points ce qui rend difficile la discrimination entre les classes, car certaines sont associées au mêmes descriptions.

Et pour GLC 2018, c'était pire, 81 % de points étaient superposés. Ceci était principalement du à l'"arrondi" peu précis des données GPS. Les organisateurs sont arrivés à la conclusion que les scores relativement faibles des modèles étaient causés par les données peu discriminantes, en plus de la difficulté qui est celle d'un problème d'ordonnancement de classes, avec beaucoup de classes, et des données sous forme d'images...

Il y a aussi le problème que les classes sont très déséquilibrées et certaines sont sous représentées par rapport à d'autres et cela peut créer un déséquilibre dans notre apprentissage.

4.3 Visualisation

4.3.1 Histogramme des espèces.

Nous avons tracer l'histogramme des fréquences des espèces dans l'ensemble des données. Les 100 espèces les plus fréquentes sont triées par rang de gauche à droite :

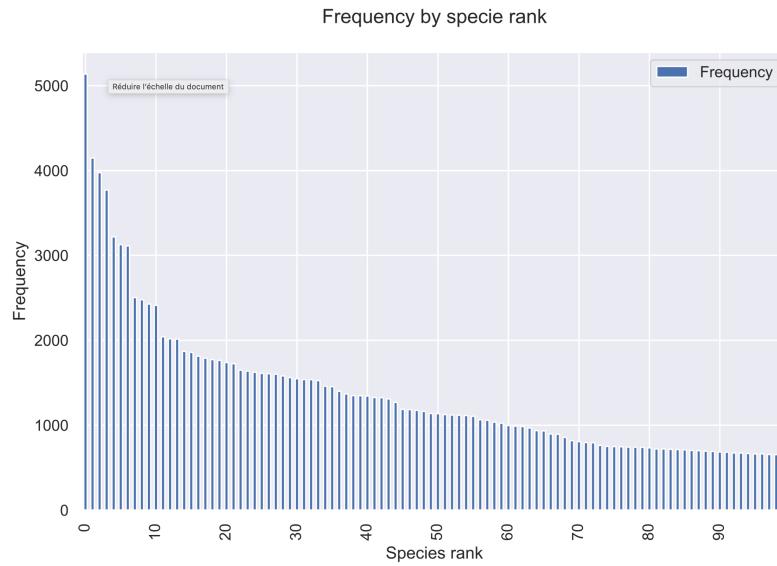


FIGURE 8 – Histogrammes des espèces.

Nous observons l'application de la loi de Zipf sur les fréquences des espèces. La fréquence d'une espèce est inversement proportionnelle à son rang et les plus fréquentes représentent la majorité des observations. De plus un grand nombre d'espèces sont très peu observées. Il y a par exemple 354 espèces qui ont moins de 10 observations dans l'ensemble de données soit 26% des espèces et 971 qui en ont moins de 100 soit 71% des classes. Pour pré-traiter les données, plus loin, nous filtrerons pour supprimer les espèces très peu observées.

4.3.2 Visualisations des occurrences sur une carte.

Visualisons les occurrences des observations en fonction de leur localisation sous forme de latitudes-longitudes.

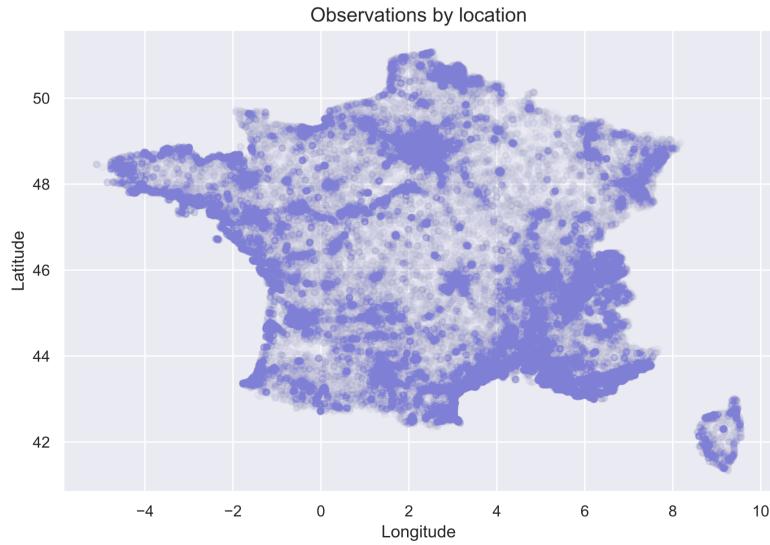


FIGURE 9 – Affichages des occurrences des espèces sur la carte de France.

Les observations ne sont pas uniformément réparties sur le territoire de la France mais se concentrent dans certaines zones, on observe bien ici le biais d'échantillonage discuté précédemment dans le cadre de la modélisation de distribution d'espèces.

4.3.3 Corrélation des variables.

Un certain nombre de variables d'environnements sont corrélées entre elles comme nous pouvons le voir sur cette matrice présentant les corrélations entre paires.

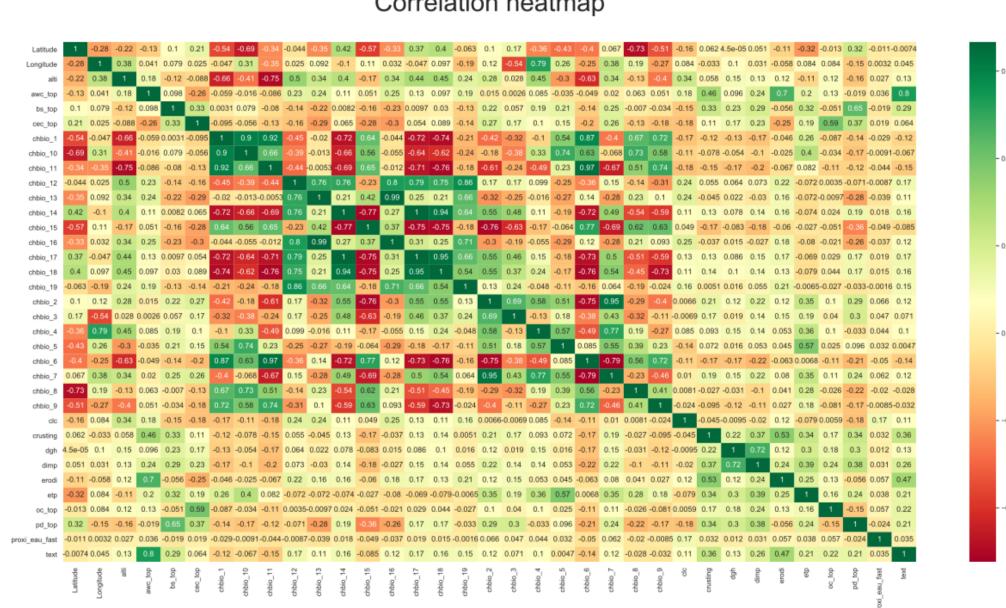


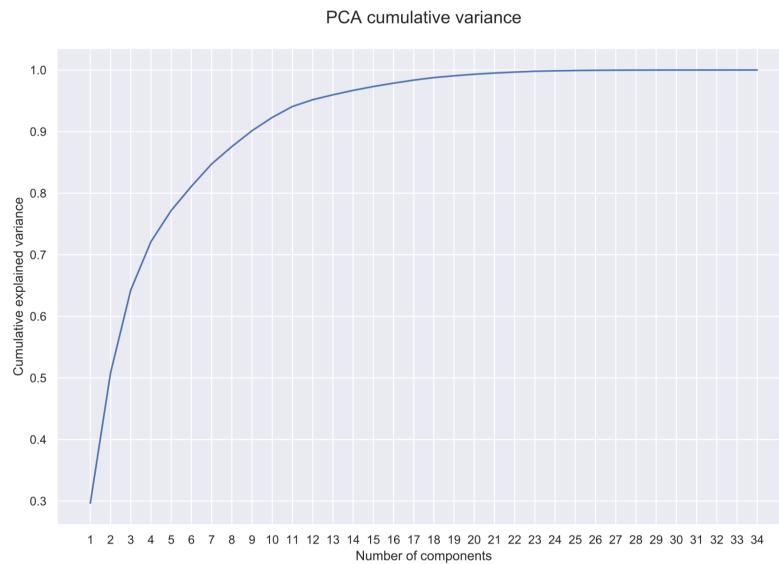
FIGURE 10 – Corrélations des dimensions des variables d'environnement.

Il y a certaines corrélations positives fortes donc il y a de la redondance dans nos données. Sachant cela nous savons que nous pouvons réduire les dimensions de nos entrées sans perdre trop d'informations.

4.3.4 Analyse en composantes principales.

Nous expérimentons une analyse en composantes principales pour réduire les dimensions ainsi que pour visualiser des points.

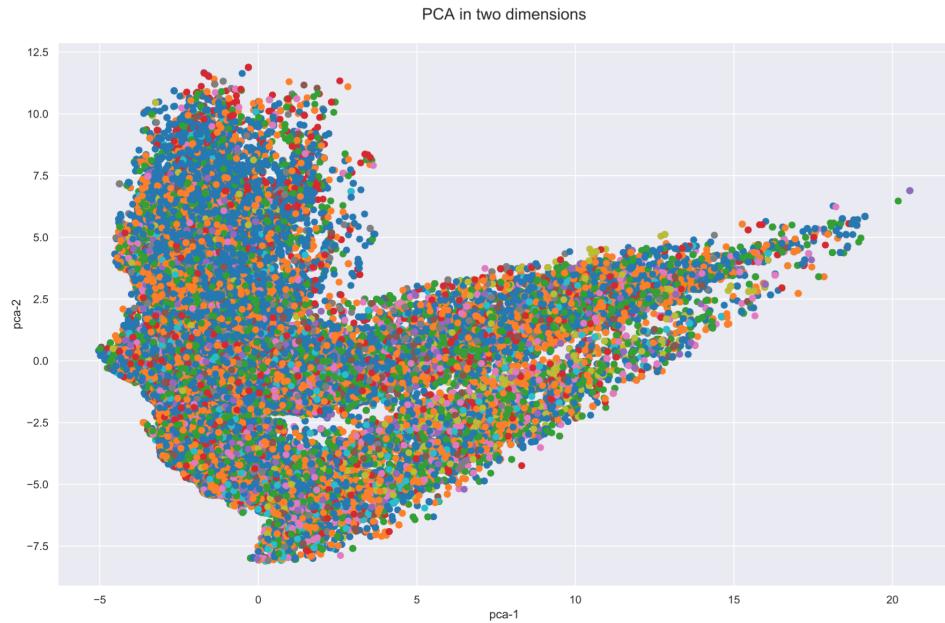
Une standardisation est appliquée au données et une PCA est effectuée sur les données standardisées. La figure suivante montre la variance cumulative expliquée en fonctions du nombre de composantes principales.



Nous observons qu'il y possible de réduire significativement le nombre de dimensions sans grande perte de variance. En effet on converse 99% de la variance avec 21 composantes principales.

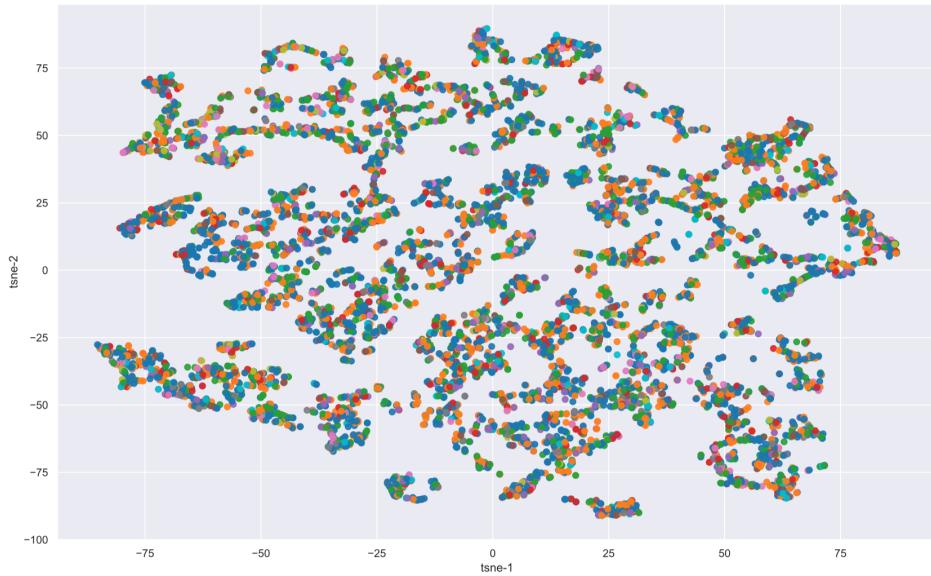
4.3.5 Visualisations par PCA et T-SNE.

Nous avons tenté de visualiser les données en deux dimensions. Dans le figure suivante nous projetons les points sur les deux premières composantes principales de la PCA.



Cette visualisation n'est pas très informative. Pour tenter d'obtenir une meilleure visualisation nous appliquons l'algorithme T-SNE sur un échantillon aléatoire de 10000 points transformés par PCA.

T-SNE visualization



Au vu du grand nombre de classes (1364) il n'est pas possible de visualiser la classe d'un point via sa couleur. Nous voyons certains clusters de points qui se distinguent. Ces clusters peuvent être du à une similarité entre espèces dans l'espace environnemental. Néanmoins au vue du nombre de points et de classe il n'est pas possible de faire une analyse plus poussée.

5.2 Les métriques d'évaluation de notre projet.

5.2.1 MRR.

Pour chaque occurrences du test les participants doivent classer par top 100 sans ex-aequo.

$$MRR = \frac{1}{Q} \sum_{q=1}^Q \frac{1}{rank_q} \quad (1)$$

Avec Q le nombre total d'occurrence x_q dans l'ensemble de test et $rank_q$ est le rang correct de l'emplacement de l'espèce $y(x_q)$ dans la liste des espèces évalué par la méthode pour l'occurrence x_q .

La MMR est une mesure statistique permettant d'évaluer les processus qui renvoie une liste de valeurs possibles en réponse à une requête et ordonnés par la plus grande probabilité de vraisemblance.

La MRR a été utilisé dans GeoLifeClef 2018, cependant étant donné la sévérité de cette métrique elle a été remplacé par le Top-30 dans l'édition 2019. On a tout de même évaluer nos modèles avec la MRR afin de nous comparer aux résultats des années précédentes.

5.2.2 Top-30.

Dans GeoLifeClef 2019 le Top30 a été la nouvelle métrique adopté, en raison de sa plus grande souplesse. Il s'agit tout simplement de donner un score de 1 si le label du test est dans les 30 premières espèces à être prédit par le classifieur, et 0 sinon.

Les métriques que nous utilisons sont le Top30, le MRR et l'accuracy. Le challenge GeoLifeCLEF 2019 est basé sur la métrique Top30 donc c'est cette métrique que nous cherchons à maximiser lors de la sélection de modèles. Le MRR est calculé afin d'obtenir une comparaison avec les résultats du challenge de l'année dernière, qui utilisait le MRR. L'accuracy est calculée à titre d'observation.

Pour chacun des modèles nous retournons une liste de 100 espèces ordonnées par probabilité d'apparition décroissante, ceci pour chaque observation.

5.3 Protocole de sélection et d'évaluation des modèles.

Puisque nous ne disposons pas des données de test, nous effectuons une séparation aléatoire entre apprentissage et test avec 60% des données en apprentissage et 40% en test.

1. Sélection d'un sous-ensemble de données pour la validation croisée :

Les étapes 2 et 3 du protocole de sélection et d'évaluation des modèles sont effectuées sur un sous-ensemble des données. En effet une recherche d'hyper-paramètres par validation croisée sur toutes les données est trop coûteuse. Ce sous-ensemble doit être suffisamment petit pour faire de la validation croisée en temps raisonnable, et suffisamment grand pour être représentatif. Nous choisissons un sous-ensemble de 10000 occurrences, et stratifié, c'est-à-dire que la distribution des classes est conservée dans le sous-ensemble.

2. Évaluation à première vue du modèle. :

Nous choisissons des paramètres pour un modèle et nous l'évaluons sur quatre séparations apprentissage-test aléatoires, pour voir s'il peut passer à l'échelle sur toute les données et avoir une première impression des scores et du temps de calcul pris. De cette manière nous avons exclu plusieurs modèles qui ne nous ont pas paru intéressant (tel que naïve Bayes, ou l'analyse discriminante linéaire (LDA)).

3. Recherche aléatoire d'hyper-paramètres par validation croisée. :

Pour trouver une combinaison d'hyper-paramètres optimale nous effectuons une recherche aléatoire sur une distribution de paramètres définie. Nous effectuons 30 tirages aléatoires et sélectionnons le tirage qui maximise le score de Top30 en validation croisée. Une recherche sur grille exhaustive n'était pas envisageable car trop coûteuse.

4. Évaluation globale du modèle. :

Nous entraînons notre modèle sur l'ensemble d'apprentissage global et l'évaluons sur l'ensemble de test mis de côté. Les scores en Top30 et MRR sont calculés et constituent l'évaluation finale du modèle.

6.3 Idées d'analyses supplémentaires de nos résultats.

Faute de temps sur cette dernière partie nous n'avons pas mis en place les idées suivantes, cependant nous voulions quand même souligner quelques analyses qui nous semble pertinentes. Premièrement nous aurions pu comparer nos résultats en sélectionnant des sous-régions de la carte de la France afin de voir quels régions sont favorisés au niveau du nombres d'occurrences et de la qualité de la prévision.

Deuxièmement nous aurions pu également faire une matrice de confusion afin de voir quelles sont les espèces souvent confondues entre elles lors de la prévision, afin de mieux comprendre les similarités entre ces espèces.

7 Conclusion et perspectives.

En conclusion de ce projet, nous pouvons affirmer qu'il s'agit d'une tâche qui nous a semblé particulièrement difficile. Bien qu'on dispose de beaucoup de données d'observations, auquel on ajoute les images caractérisant les environnements, prédire la bonne espèce observée parmi de nombreuses espèces possibles est une tâche d'apprentissage non triviale. La partie pré-traitement a été essentielle pour tirer le maximum d'information et avoir des modèles supérieures au modèle fréquentiel. Les nombreux points superposés, le niveau de confiance des observations, et la date d'observation des espèces qui peuvent être assez espacées dans le temps sont certains des obstacles à la qualité de nos prédictions.

Cependant, les modèles de l'état de l'art prometteurs tels que XGBoost ou Random Forest ont fourni des résultats relativement satisfaisants, en tout cas les meilleurs que nous ayons obtenus. Il serait intéressant d'implémenter un CNN et d'observer ou non sa supériorité pour cette tâche.

7.1 Améliorations.

Les CNN ont présentés des résultats très prometteurs dans le domaine de la prédiction de localisation des espèces [5]. Tout d'abord parce qu'ils permettent de prendre en compte toute l'information disponible que fournissent les images représentant les variables d'environnements, contrairement aux modèles que nous avons implémentés jusqu'à présent. Les CNN est les modèles de deep learning en général sont une piste de recherche à creuser dans le domaine. De plus, pour améliorer la prédiction des espèces sachant la localisation, l'autre challenge PlantCLEF de LifeCLEF avait pour but quand à lui d'utiliser les photographies des plantes prises par les utilisateurs avec l'application PlantNet. Ces données sont complémentaires aux notre et il serait sûrement possible d'obtenir de meilleures prédictions en incluant aussi les images dans un modèle d'apprentissage de prédictions d'espèces.

8 Bibliographie.

- [1] C. Botella, P. Bonnet, F. Munoz, P. Monestiez, A. Joly, "Overview of GeoLifeCLEF 2018 : location-based species recommendation" *In : CLEF working notes 2018* (2018)
- [2] B. Deneu, M. Servajean, C. Botella, A. Joly : "Location-based species recommendation using co-occurrences and environment - GeoLifeCLEF 2018 challenge." *In : CLEF working notes 2018* (2018)
- [3] S. Taubert, M. Mauermann, S.K.D.K., M. Eibl : "Species prediction based on environmental variables using machine learning techniques." *In : CLEF working notes 2018* (2018)
- [4] N.B. Moudhgalya, S. Sundar, S. Divi, P. Mirunalini, C. Aravindan Bose : "Hierarchically embedded taxonomy with CLNN to predict species based on spatial features". *In : CLEF working notes 2018* (2018)
- [5] C. Botella, A. Joly, P.B.P.M., F. Munoz : "A deep learning approach to species distribution modelling." *Multimedia Technologies for Environmental & Biodiversity Informatics* (2018)
- [6] J. Brown, Y. Anne. (2015). "Shifting ranges and conservation challenges for lemurs in the face of climate change", *Ecology and Evolution* 5(6) (2015)
- [7] J.-C. Svenning, C. Fløjgaard, K. A. Marske, D. Nogués-Bravo, S. Normand : "Applications of species distribution modeling to paleobiology", *Quaternary Science Reviews*, Volume 30
- [8] C. Merow, M. J. Smith, J. A. Silander Jr : "A practical guide to MaxEnt for modeling species' distributions : what it does, and why inputs and settings matter", *Ecography*, Volume 36
- [9] Q. Huang, C. H. Fleminga, B. Robba, A. Lothspeicha, M. Songera : "How different are species distribution model predictions?—Application of a T new measure of dissimilarity and level of significance to giant panda *Ailuropoda melanoleuca*", *Ecological Informatics* 46 (2018)
- [10] S. S. Bucak, P. Kumar Mallapragada, R. Jin, A. K. Jain : "Efficient multi-label ranking for multi-class learning : application to object recognition", *Proceedings / IEEE International Conference on Computer Vision* (2009)
- [11] S. Marsland : "Machine learning - an algorithmic perspective", *second edition, Taylor & Francis Group* (2015)
- [12] S. Thiria, Y. Lechevallier, O. Gascuel, S. Canu : "Statistiques et méthodes neuronales", *Dunod* (1997)