



PROJET LOGICIEL DAC


GeoLifeCLEF



KARMIM Yannis et BOURCIER Jules

Encadrants : SOULIER Laure et ZABLOCKI Eloi

PLAN

- 
1. INTRODUCTION.
 2. ÉTAT DE L'ART.
 3. ANALYSE DES DONNÉES.
 4. VISUALISATIONS
 5. APPRENTISSAGE DE MODÈLES DE MACHINE LEARNING.
 6. RÉSULTATS.
 7. CONCLUSION ET PERSPECTIVES.



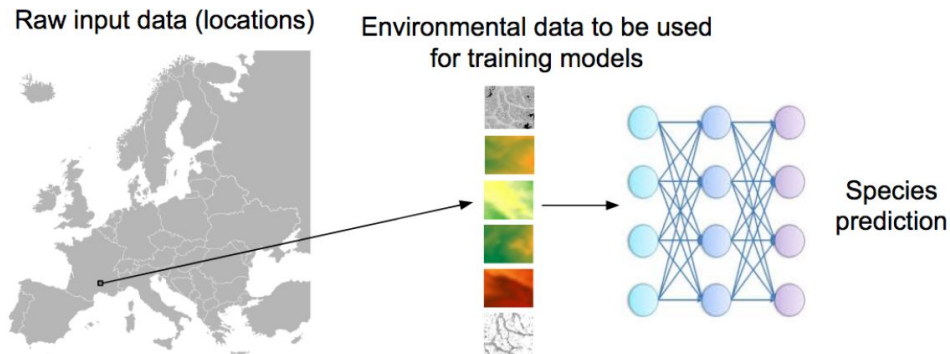
INTRODUCTION

Motivations.

- Participation à la campagne CLEF Pour le challenge GeoLife CLEF
- La tâche consiste à prédire une liste d'espèces sachant une localisation.
- Problème complexe et important en Bio-Informatique.
- Beaucoup d'applications possible.
- Un domaine scientifique → La modélisation de distribution des espèces. (Species distribution Modeling).



Présentation de la tâche.



Contexte :

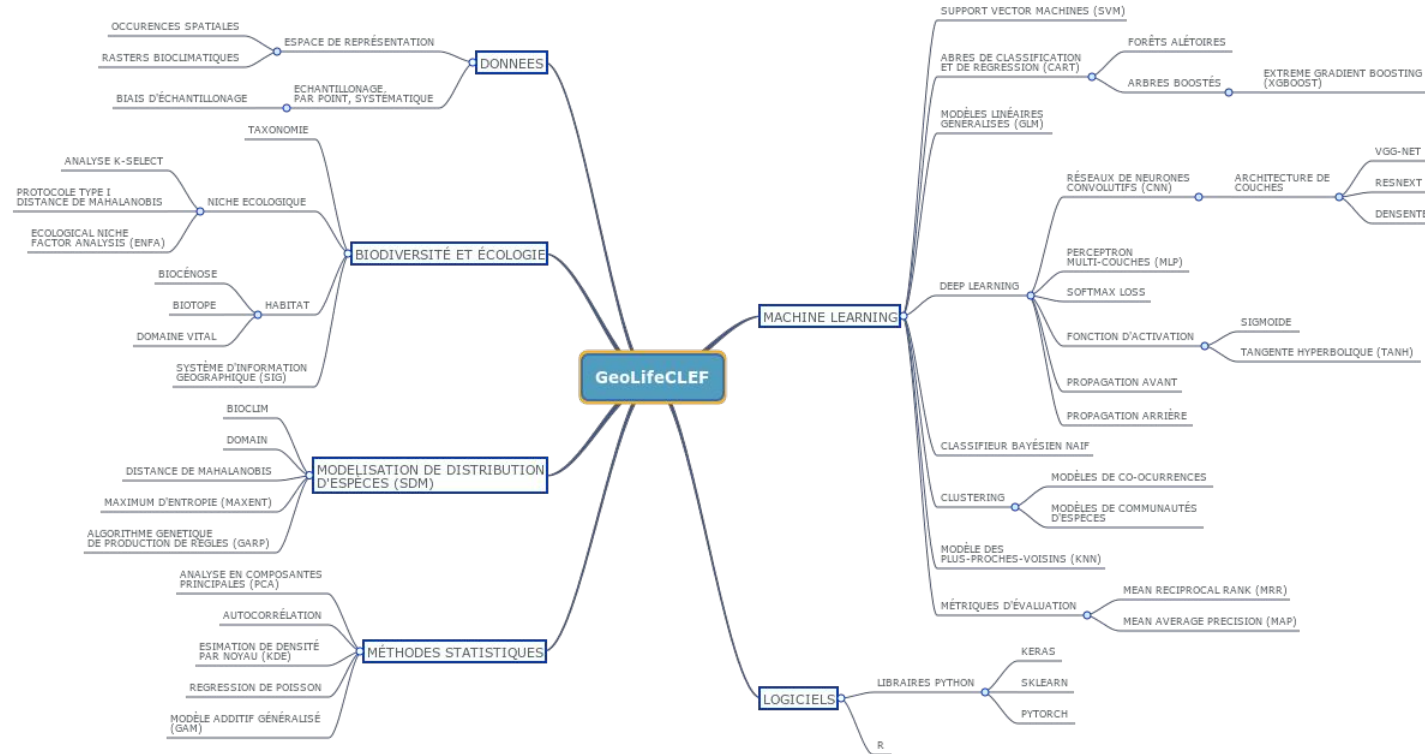
Observation en France métropolitaine →
Beaucoup plus d'observations,
biodiversité variée...

- Trop peu d'observations pour chaque espèce + biais d'échantillonnage → Utiliser les variables d'environnements pour l'apprentissage des modèles .
- Particularité de GeoLifeClef → Les variables d'environnements sont représentées par des images.
- Une localisation et 33 images caractérisant l'environnement autour de ce point. (altitude, sécheresse, humidité...). Tenseurs de taille 64x64x33 pour chaque localisation !



ÉTAT DE L'ART

Axes pour notre recherche documentaire.



Quelques références...

. B. Deneu et al (2018) : "Location-based species recommendation using co-occurrences and environment - GeoLifeCLEF 2018 challenge."

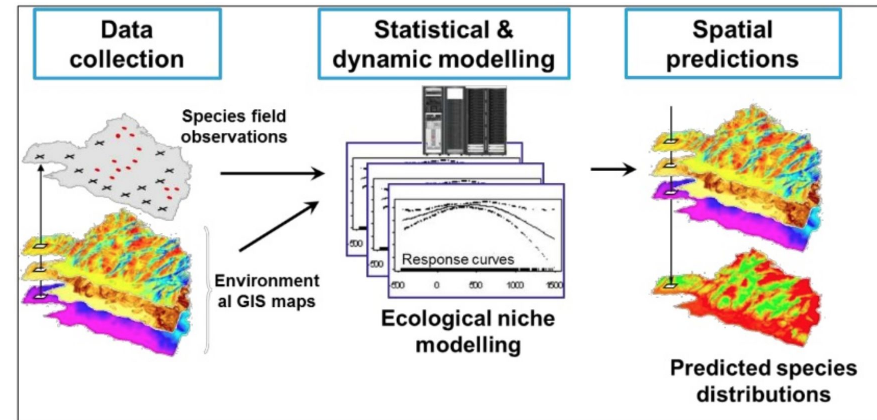
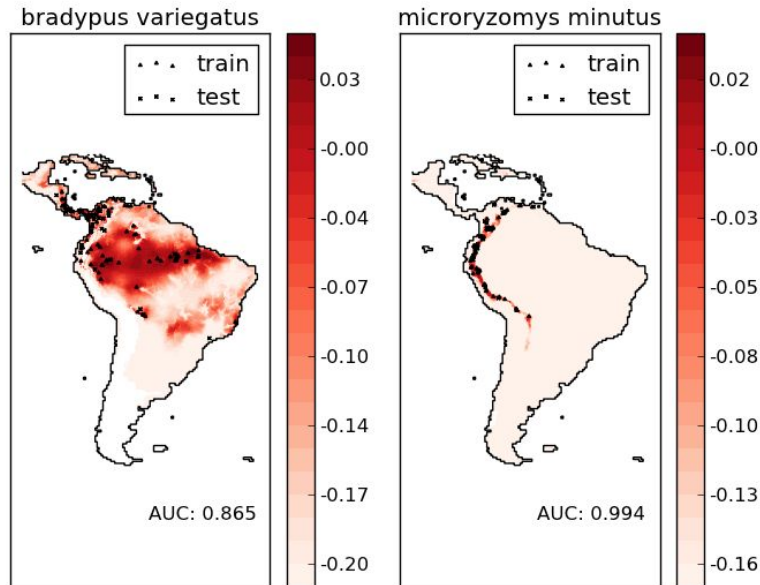
. C. Botella et al (2018) : "A deep learning approach to species distribution modelling."

. SB Kotsiantis et al (2007): "Supervised Machine Learning : A review of classification techniques".

Species distribution modeling

Phillips et al (2006) : “Maximum entropy modeling of species geographic distributions”

Objectif : apprendre une distribution géographique d'espèces via l'environnement.



État de l'art des modèles de l'édition 2018.



- 2 équipes FLO et ST qui ont soumis des modèles intéressants.
- Team FLO première du classement avec un CNN à l'architecture customisé.
Également un modèle de Random Forest.

. B. Deneu et al (2018) : "Location-based species recommendation using co-occurrences and environment - GeoLifeCLEF 2018 challenge."
- Team ST ont aussi des modèles CNN mais avec une architecture prédéfini DenseNet, VGG...Moins bons scores que des modèles "classiques" comme XGboost.

Présentation des modèles intéressants pour la tâche

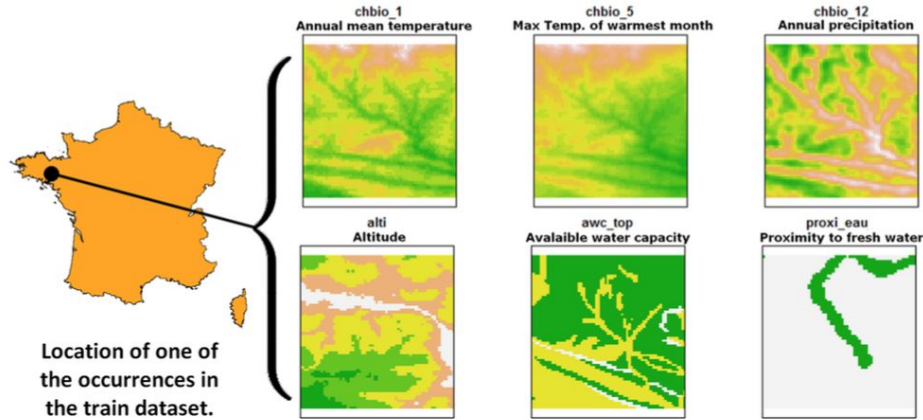


- Random Forest. → Méthode de bagging sur des classifieurs simples, moyenne pondérés de ces classifieurs afin de prendre de meilleures décisions.
L.Breiman. (2001) : "Random Forests."
- Boosted Trees (XGBoost) → Agrégation séquentielle de classifieur simple, afin de minimiser une fonction de coût (généralement coût exponentielle). . L. Mason et al (2018) : "Boosting Algorithms as Gradient Descent."
- CNN. → Réseaux convolutifs, adaptés aux problèmes de classification d'images.
Y. LeCun et al. (2010) : "Convolutional networks and applications in vision"
- Modèle de plus proche voisins : k-NN, rayons de voisins



ANALYSE DE NOS DONNÉES

Analyse des données environnementales



Avantages: Utiliser des images plutôt que des mesures → Informations continues autour d'un point.

Le cas particulier de la variable 'clc':
Variable catégorielle → 48 valeurs possibles.
Ne peut pas être codée comme une variable réelle.

Color maps representing 6 slices of the environmental array of the point.

N.B. : The geographic extent of each map results directly from the source environmental data resolution and is not necessarily identical from one map to another.

Sources 19 premières variables climatiques proviennent de "Chelsea Climate".
10 variables ordinales proviennent de "ESDB Soil pedology data".
4 dernières variables "CGIAR-CSI evapotranspiration data", "USGC Elevation Data", "BD Carthage Hydrological data".

Karger, D.N., et al. (2016): "Climatologies at high resolution for the earth's land surface areas."

Analyse des données des observations



Les données de tests ne sont pas fournis.

Les différentes sources de données: Observations du Global Biodiversity Information Facility (GBIF). → Peu précises
Observations avec l'application PlantNet sur smartphone → Plus fiable !

Une structure commune : Id de l'espèce, localisation, nom taxonomique, point de localisation (latitude, longitude).

Les difficultés liés aux données



- Le cas de la variable 'clc' → Dilemme : ajouter 48 dimensions à nos entrées ou supprimer cette variable?
- PlantNetTrusted → 35% de points superposés. Beaucoup de données superposées
- Problème de ranking, beaucoup de classes (1364), points superposés, et pb de discrimination d'espèces entre elles (en pratique de nombreuses espèces sont observable au même endroit).

Formalisation de notre problème.



Problème de classification multi-classes en apprentissage supervisé, à partir d'images.

$\mathcal{D} = \{\mathbf{x}^i, y^i\}, \mathbf{x}^i \in \mathcal{X}$ Représente notre ensemble d'apprentissage. Et $\mathcal{X} : \mathbb{R} \times \mathbb{R} \times \mathcal{T}$, et $\mathcal{T} \in \mathbb{R}^{64 \times 64 \times 33}$

$\mathbf{x}^i = (lat, long, t)$, avec $t \in \mathcal{T}$ le tenseur d'environnement.

$y^i \in \mathcal{Y}$ Avec y^i l'id d'une espèce. $|\mathcal{Y}| = 1364$ et $|\mathcal{D}| = 237k$

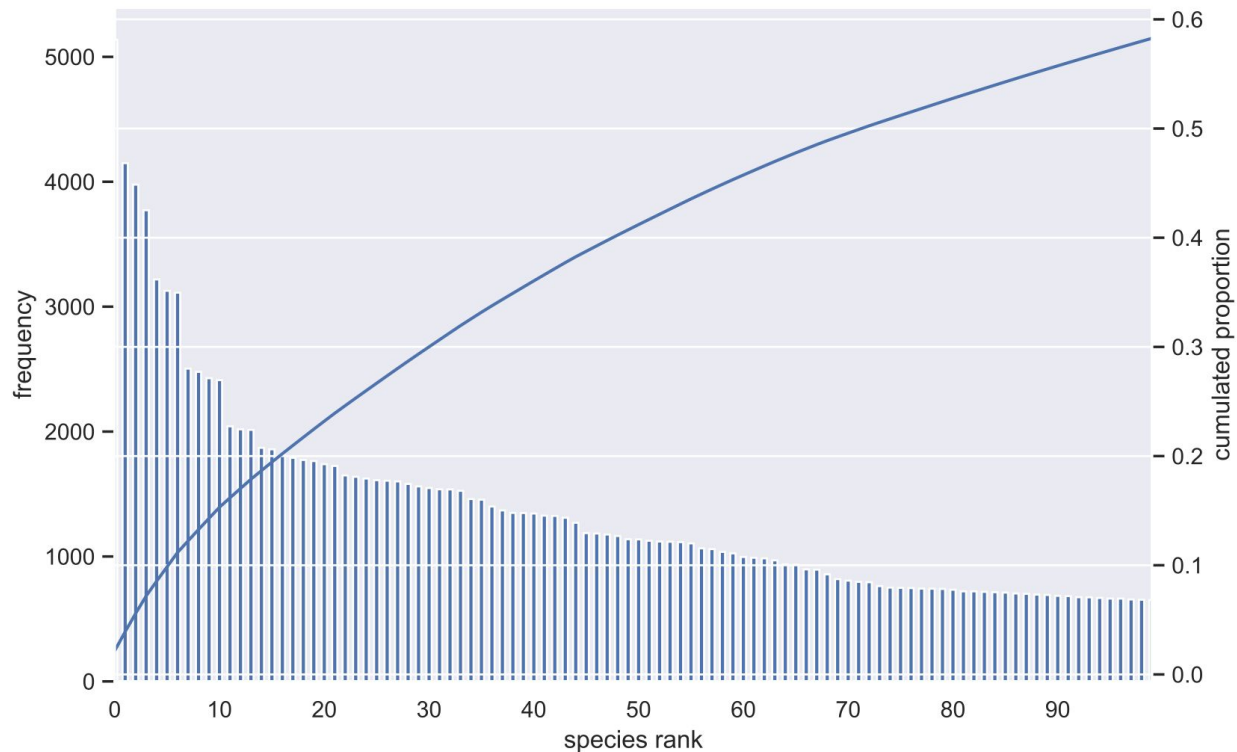
Le but est de trouver : $f : \mathcal{X} \rightarrow \mathcal{Y}^{100}$ avec $f(\mathbf{x}^i) = \hat{y} = (\hat{y}_1, \dots, \hat{y}_{100})$ avec f^* tel que : $f^* = \operatorname{argmin} \sum_D l(f(\mathbf{x}^i), y^i)$

l une fonction de coût.



VISUALISATIONS

Frequency by specie rank

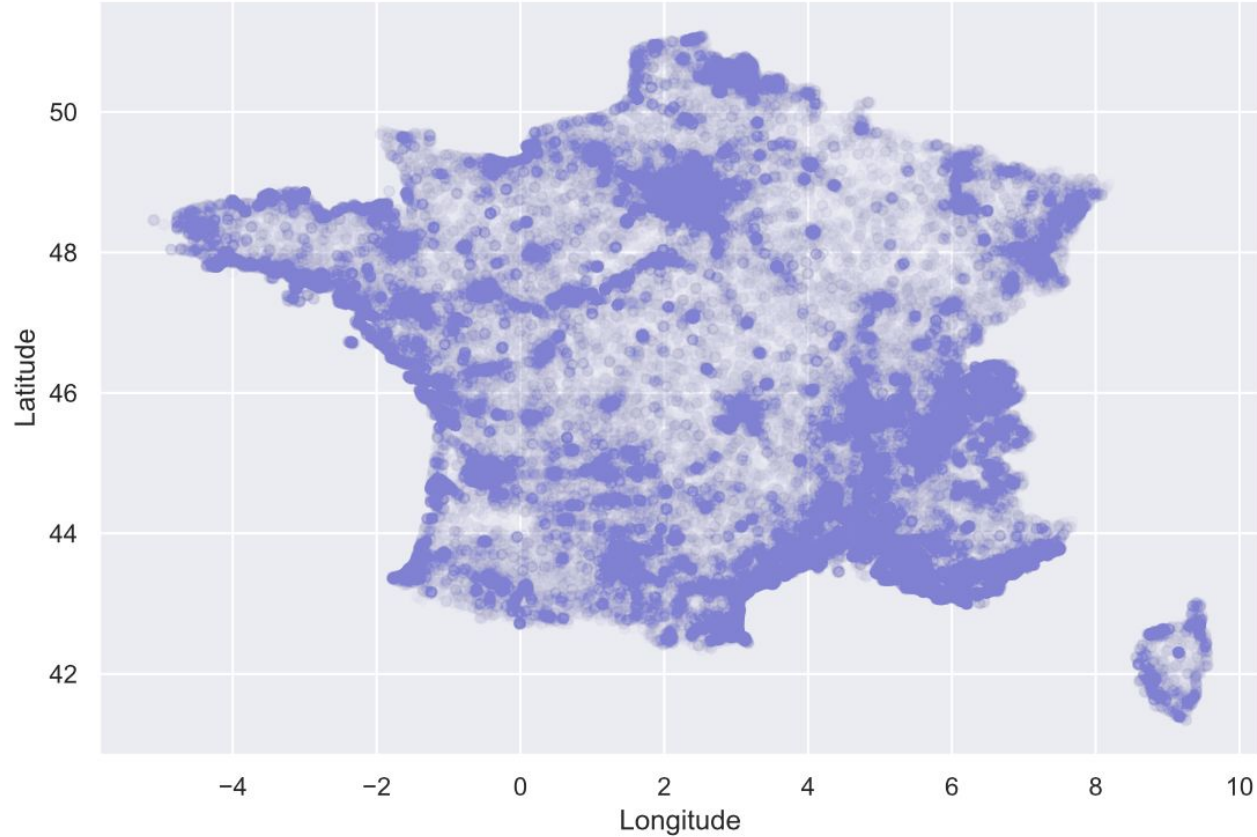


Nous observons une Loi de Zipf : la fréquence est inv. prop^o au rang.

Les espèces rares (moins de 10 occ) représentent 26% des espèces.

71% des classes ont moins de 100 occ.

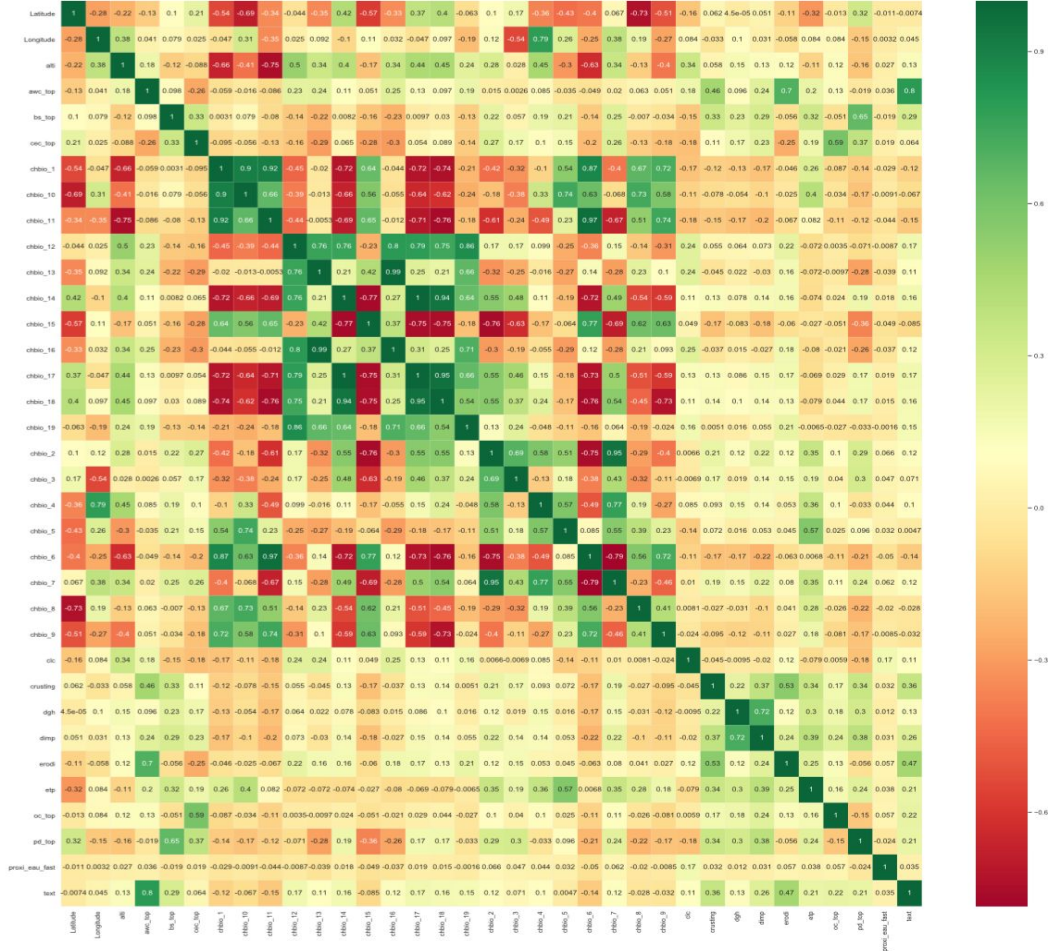
Observations by location



Hétérogénéité de la répartition géographique des observations.

Biais d'échantillonnage.

Correlation heatmap

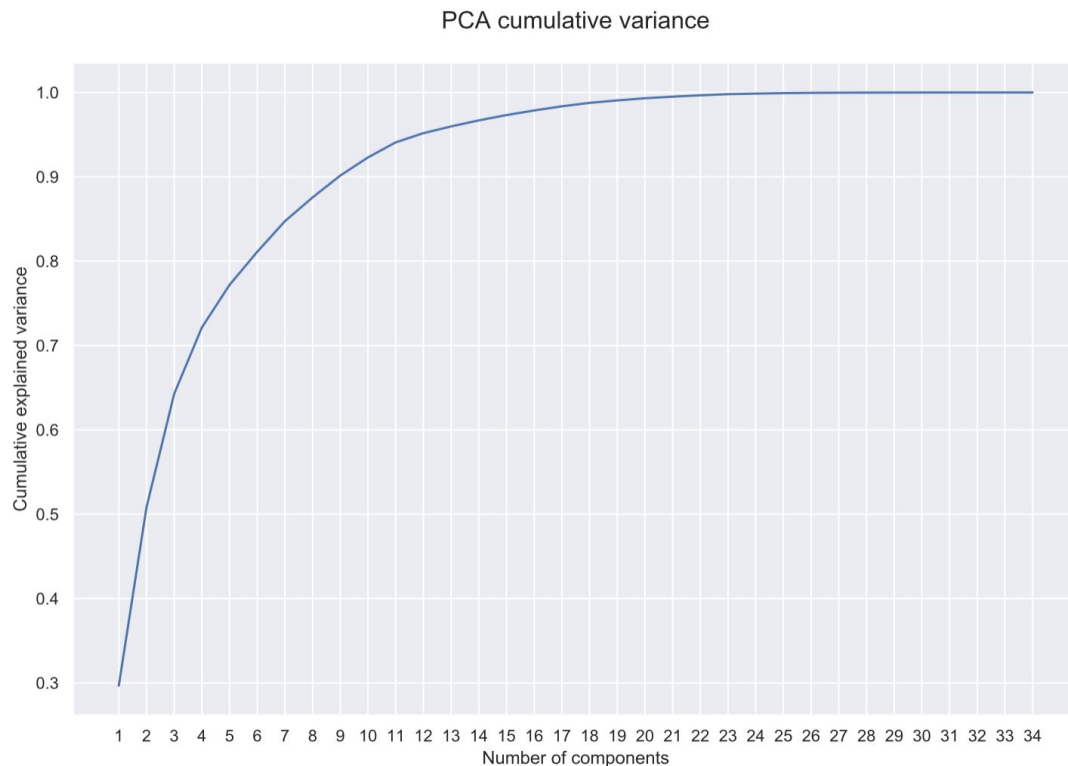


Un certain nombre de **fortes corrélations positives**, redondance dans nos données.

Nous pouvons réduire la dimensionnalité sans perdre trop d'information.

Solution : analyse en composantes principales

Analyse en composantes principales



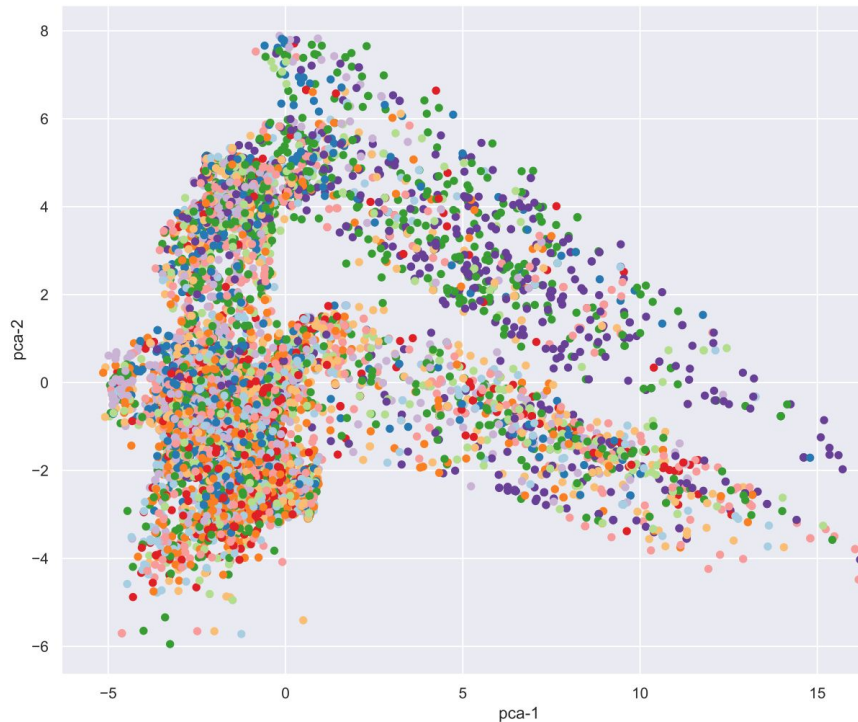
PCA effectuée sur les données standardisées.

Réduction significative du nombre de dimensions.

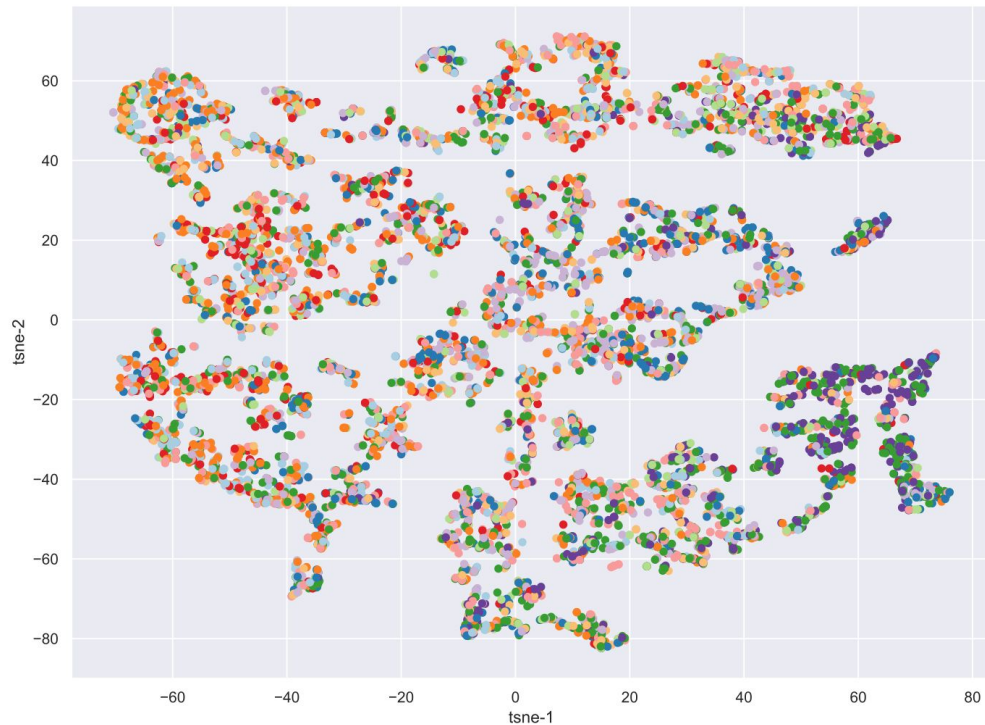
99% de la variance sur 21 axes.

Visualisation PCA et T-SNE

PCA in two dimensions



T-SNE visualization



Données ne semblent pas être facilement séparables.

APPRENTISSAGE DE MODÈLES DE MACHINE LEARNING

Pré-traitement des données

1. Extraction des variables environnementales depuis les rasters.

| | Latitude | Longitude | alti | awc_top | bs_top | cec_top | ... | erodi | etp | oc_top | pd_top | proxi_eau_fast | text | glc19SpId |
|---|----------|-----------|------------|-----------|-----------|-----------|-----|----------|-------------|----------|----------|----------------|---------|-----------|
| 0 | 43.95195 | 2.118889 | 215.917969 | 146.71875 | 85.000000 | 8.171875 | ... | 3.921875 | 1216.015625 | 1.078125 | 2.000000 | 0.0 | 1.75000 | 0 30021 |
| 1 | 45.10639 | -0.592500 | 31.777344 | 148.12500 | 66.250000 | 16.375000 | ... | 4.375000 | 1139.238281 | 2.875000 | 1.000000 | 0.0 | 1.62500 | 1 31997 |
| 2 | 48.38958 | -4.534861 | 57.539062 | 85.90625 | 57.890625 | 12.093750 | ... | 1.468750 | 720.234375 | 1.703125 | 1.359375 | 0.0 | 0.96875 | 2 31385 |

2. Filtrage des espèces rares : Fréquence supérieure à 10. 26% des classes éliminées, 99% de données conservées

4. Encodage one-hot des variables catégoriques

5. Analyse en composante principales : 21 axes expliquants 99% de la variance

| | pca-1 | pca-2 | pca-3 | pca-4 | pca-5 | pca-6 | ... | pca-16 | pca-17 | pca-18 | pca-19 | pca-20 | pca-21 |
|---|-----------|-----------|-----------|----------|-----------|-----------|-----|-----------|-----------|-----------|-----------|-----------|-----------|
| 0 | -1.945958 | -1.756125 | -0.917585 | 0.421551 | -0.985261 | 1.504026 | ... | -0.546972 | 0.599075 | -0.203069 | 0.108532 | -0.439735 | -0.071614 |
| 1 | -2.198327 | -0.602683 | 0.776839 | 1.448802 | 1.233409 | 1.114648 | ... | 0.009172 | 0.113226 | 0.493028 | -0.056814 | 0.776326 | 0.033171 |
| 2 | 3.220271 | -1.165626 | 2.136813 | 3.930181 | 3.645509 | -2.023236 | ... | -0.305992 | -0.025754 | 0.359020 | -0.403396 | 0.105869 | -0.045677 |

Modèles évalués

Besoin de classifieurs capables de modéliser des probabilités.

Baselines : modèle aléatoire, modèle fréquence et modèle de centroïde le plus proche.

Modèles de voisinage évaluants la distribution des classes parmi les voisins. K plus proches voisins, modèle de rayon de voisins, modèle vectoriel (1-nn)

Bagging d'arbres de décision : **Random forests** et la variante **extra-trees**

Boosting de gradient : **XGBoost**

SVM à noyau linéaire et **régression logistique** entraînés par descente de gradient stochastique.

Analyse discriminante linéaire (LDA)

Naive bayes



Sélection et évaluation de modèles

Métriques d'évaluations :

Top30 : métrique du challenge. Score 1 si l'espèce observée est parmi les 30 premières prédictions, 0 sinon.

MRR : Rang réciproque moyen du label observé dans les 100 premières prédictions.

Protocole utilisé :

1. Séparation entre apprentissage et test (60% - 40%). 141k occ. en apprentissage et 94k en test.
2. Recherche aléatoire d'hyper-paramètres par validation croisée 5-folds.
3. Évaluation globale du modèle avec les scores en Top30 et MRR





RÉSULTATS

Résultats des scores

Le modèle Xgboost est le plus performant des modèles testés, suivit par la random forest.

Les méthodes d'ensembles dominent le tableau à l'exception de du modèle LDA.

Les classifieurs linéaires (LDA, SVM, reg. logistique) ont de bonnes performances.

Modèles de voisins ne passent pas à l'échelle en mémoire.

| Scores en test | Top30 | MRR | Accuracy | Paramètres |
|--------------------------------|---------|----------|----------|---|
| Xgboost | 0.43210 | 0.103806 | 0.03420 | {'num_boost_round':100, 'eta': 0.0623421, 'max_depth':6, 'colsample_bytree': 0.9} |
| Random forest | 0.37500 | 0.09196 | 0.029100 | {'bootstrap': False, 'class_weight': None, 'criterion': 'gini', 'max_depth': 5, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 150} |
| Analyse discriminante linéaire | 0.37480 | 0.08007 | 0.02740 | {'solver': 'svd', 'shrinkage': 'auto'} |
| Extra trees | 0.36502 | 0.08799 | 0.02730 | {'bootstrap': False, 'criterion': 'gini', 'max_depth': 3, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'class_weight': None} |
| Support vector machine | 0.34902 | 0.08514 | 0.02734 | {'class_weight': 'balanced', 'alpha': 0.1925744} |
| Régression logistique | 0.34775 | 0.08542 | 0.02721 | {'class_weight': None, 'alpha': 0.1268961} |
| Naive bayes | 0.30004 | 0.05514 | 0.01353 | {'var_smoothing': 1e-09} |
| Modèle fréquence | 0.29600 | 0.06453 | 0.02185 | / |
| Plus proche centroïde | 0.12330 | 0.02320 | 0.00600 | {'metric': 'cosine'} |
| Modèle aléatoire | 0.03660 | 0.00623 | 0.00115 | / |

CONCLUSION



Tâches difficiles, données peu fiables. Beaucoup de données d'images.

Idées d'analyses sur nos résultats :

Analyser les scores sur des régions dans le territoire → Identifier les zones les mieux observées...

Matrice de confusion sur la prédiction des espèces → Quelles sont les espèces souvent confondues entre elles.

Améliorations possibles :

Utiliser plus de données quitte à ce qu'on ait un niveau de confiance moins précis et comparer les résultats.

CNN architecture customisée, fusionner CNN et modèles classiques comme XGBoost, Random Forest.

Données proviennent de photographies d'espèces identifiées : inclure en plus les photos des observations en entrée pourrait mener à de meilleures prédictions.