Sujet Filière Métier Recheche, 2024-2025

Prédire la présence d'éspèces avec les species distribution models : revue de l'état de l'art et application au marin

Encadrant : Jérémy Fix (CentraleSupélec, LORIA)

1 Contexte

Dans le domaine marin, les données collectées sont bien souvent éparses et la connaissances des espèces présentes est limitée. Lorsqu'il est question de construire des indicateurs de biodiversité, pour quantifier l'état de santé d'un écosystème, cette vision parcellaire de l'environnement est limitant. Les données sont éparses pour plusieurs raisons mais notamment parce que les observations sont coûteuses à réaliser. Les données sont parfois collectées en un point, par des balises fixes ou parfois à l'aide de balises accrochées à un navire et donc qui ne fournit des observations que le long de la trajectoire du bateau. Une autre spécificité des données marines est que les données collectées sont bien souvent des données de présence seule, sans que ce ne soit répertoriée l'absence d'une espèce.

Dans le cadre du projet ANR Challenge IA-Biodiv, CentraleSupélec participe au consortium SmartBiodiv aux côtés de Georgia Tech, du LIEC, du LOV et du LOCEAN. L'objectif de ce projet est de développer des méthodes d'intelligence artificielle pour d'une part densifier les données dans le temps et dans l'espace et d'autre part pour expliciter des indicateurs de biodiversité de ces écosystèmes marins.

En écologie, les modèles de distribution d'espèces (SDM) sont des outils qui permettent de prédire la présence d'espèces à partir de données environnementales en utilisant des données de présence seule[1]. Nous proposons dans ce sujet de réaliser un état de l'art des modèles de distribution d'éspèces et de les appliquer sur des données marines qui sont utilisées dans le cadre du projet SmartBiodiv.

2 Définition du problème

Un certain nombre d'études portent sur l'application des modèles de distribution d'espèces pour des données terrestres [2] et mettent en avant en particulier l'approche MaxEnt [3, 4]comme l'une des plus performantes. On se propose dans ce travail d'appliquer ces techniques sur des données marines. Les données à disposition sont des données d'observation de présence seule, et parfois de présence/absence. Dans le cadre de ce travail, on se propose de n'utiliser que les données de présence pour l'apprentissage mais, lorsqu'elles sont disponibles, les données d'absence peuvent permettre de quantifier la capacité d'un modèle à prédire la présence d'une espèce. Sans données d'absence lors de l'entrainement, les modèles SDM estiment l'adéquation environnementale d'un site plutôt que réellement la présence de l'espèce, différents facteurs pouvant empêcher une espèce d'occuper un site qui lui serait adéquat.

L'état de l'art devra permettre d'identifier quelques techniques de SDM les plus prometteuses avant de les appliquer sur les données marines. Dans le consortium SmartBiodiv, nous disposons au minimum de deux bases de données :

— la base Jedi avec environ 500k observations de zooplankton gelatineux (e.g. méduses)
— une base de données construites à partir de la base GBIF

Chacune de ces bases d'observations d'espèces est étendue par des données environnementales (e.g. température, salinité, chlorophylle, courants, nutriments, ...) qui sont fournies par le service Copernicus.

Pour des besoins de comparaison, concernant l'approche MaxEnt, on pourra également utiliser les données d'eau douce de [4] pour la prédiction de présence d'un poisson carnivore dans les rivières Australiennes.

3 Attendus

Les livrables attendus pour ce projet sont un mini rapport construit à partir de l'état de l'art et détaillant la ou les techniques qui auront été testées et consignant également les résultats obtenus.

Un certain nombre d'outils en écologie sont disponibles en R. Si les étudiants sont à l'aise avec ce langage, ces codes pourront servir de base de comparaison. Mais dans l'idéal, l'attendu sont des implémentations python que nous pourrions réutiliser dans le cadre du projet SmartBiodiv. Deux étudiants en thèse seraient particulièrement intéressées par ces implémentations.

Références

[1] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners : A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '21, page 329–348, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Roozbeh Valavi, Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, and Jane Elith. Predictive performance of presence-only species distribution models : a benchmark study with reproducible code. *Ecological Monographs*, 92(1) :e01486, 2022.

[3] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3) :231–259, 2006.

[4] Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. A statistical explanation of maxent for ecologists. *Diversity and Distributions*, 17(1) :43–57, 2011.