Stage M2 : Modélisation statistique et apprentissage profond pour les données ponctuelles massives de biodiversité

Contexte

Suivre la distribution spatiale et temporelle des espèces est crucial pour comprendre les réponses aux changements globaux (climatique, d'utilisation des sols, mondialisation des transferts d'espèces) de la biodiversité et définir les priorités d'action en conservation. La cartographie spatio-temporelle des espèces passe par la modélisation statistique d'une variable de population (e.g. fréquence d'occurrence, probabilité de présence, abondance) en fonction de descripteurs de l'environnement spatialisés et susceptibles de varier dans le temps.

Les modèles de distribution d'espèces basés sur l'apprentissage profond (deepSDMs, Deneu et al., 2021) permettent de capitaliser efficacement sur les masses de données d'observation de biodiversité disponibles aujourd'hui et sur les descripteurs environnementaux complexes (e.g. images et séries temporelles satellites). L'apprentissage des réseaux de neurones profonds a été largement éprouvé pour des tâches de classification complexes. C'est le cas des deepSDMs, souvent entraînés à prédire l'espèce observée à un endroit et une date donnée parmi un grand nombre d'espèces possibles. Cependant, le modèle de classification classique (régression logistique multinomiale) est limité pour représenter la distribution des espèces. En particulier, ce modèle ne peut pas différencier la taille des communautés d'espèces locales et l'incertitude sur leur composition. De plus, il est difficile de relier rigoureusement les modèles de classification à l'ensemble des données de biodiversité existantes et complémentaires (e.g. présence ponctuelle d'une espèce sans information sur l'absence des autres, inventaire de présence-absence, comptages d'abondance locale).

Les processus ponctuels (e.g. processus de Poisson), qui modélisent de manière probabiliste la distribution de points sur des espaces continus, sont des modèles plus généraux qui comblent ces manques et unifient la modélisation des données de biodiversité (Miller et al., 2019). Cette classe de modèles semble idéale pour permettre aux deepSDMs de capturer les variations d'abondance et de richesse spécifiques à fine résolution spatiale. Cependant, les réseaux de neurones profonds entraînés en régression (incluant les processus ponctuels) montrent des problèmes de généralisation dûs à leur dynamique d'apprentissage et de convergence (Stewart et al., 2023). Nous proposons donc (i) de caractériser les problèmes d'apprentissage de réseaux de neurones profonds dans le cadre des processus ponctuels, (ii) d'évaluer dans quelle mesure la vision de la régression comme une problème de classification (par intervalles) permet de résoudre ces problèmes.

Mission du stage

Le stage vise à tester empiriquement des hypothèses théoriques sur les propriétés de l'apprentissage profond dans le contexte de la modélisation spatio-temporelle des communautés d'espèces.
Il s'agira notamment de définir une stratégie d'approximation de la vraisemblance d'un processus ponctuel en tenant compte de la complexité intrinsèque des réseaux de neurones profonds, de comparer la prédiction de probabilité relative entre espèces (probabilité d'observer une espèce parmi l'ensemble des espèces modélisées) en un point donné entre

une vraisemblance de classification (cross-entropy) contre une vraisemblance de processus ponctuel de Poisson à architecture de réseau de neurones constante. Cela permettra de comparer les deux estimateurs entraînés avec l'algorithme Stochastic Batch Gradient Descent, au centre du succès de l'apprentissage profond, et éventuellement leur sensibilité aux hyper-paramètres: initialisation, learning rate, taille de batch, régularisation. Dans le cas du processus ponctuel, il faudra également valider l'estimateur de l'espérance du nombre de points (potentiellement problématique avec l'apprentissage par mini-batchs). L'expérience pourra être menée en simulation et sur des jeux de données existants (GeoLifeCLEF 2023) ou à venir (GeoLifeCLEF 2024).
Ce travail déterminera la possibilité de généraliser l'apprentissage des deepSDMs aux processus ponctuels. Dans le cas contraire, le stagiaire évaluera en théorie et empiriquement le potentiel d'une vraisemblance de classification basée sur la discrétisation en intervalles de comptage du signal de points par espèce. Cela consiste à caractériser analytiquement l'erreur d'approximation minimale de l'espérance du comptage et de son intervalle de confiance.

Références:

- Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, *17*(4), e1008856.
- Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22-37.
- Stewart, L., Bach, F., Berthet, Q., & Vert, J. P. (2023, April). Regression as Classification: Influence of Task Formulation on Neural Network Features. In International Conference on Artificial Intelligence and Statistics (pp. 11563-11582). PMLR.

Compétences :
- Python
- Machine learning
- Statistiques,
- Des connaissances en écologie et biodiversité seront un plus

Durée : 6 mois
Début : flexible (à partir de février ou après)
Organisme de rattachement : Université de Montpellier
Laboratoire de rattachement : LIRMM
Équipes : Advanse et Zenith
Gratification : 4,05 euros de l'heure
Encadrement :

- Maximilien Servajean (servajean@lirmm.fr)
- Christophe Botella (christophe.botella@inria.fr)
- Alexis Joly (alexis.joly@inria.fr)

Possibilité de poursuite en thèse.