Propager les Incertitudes associées au Traitement Automatique des données collectées par les CApteurs en écologie grâce à la modélisation statistique (PITACA)

La thèse en résumé : L'utilisation de capteurs en écologie, notamment photographiques et acoustiques, permet un suivi des populations animales et végétales partout et à tout moment. Le traitement automatisé par apprentissage profond permet d'extraire rapidement l'information souhaitée de ces capteurs (identifier une espèce ou déterminer l'identité ou le comportement d'un individu). Cette information est toutefois traitée comme une observation in situ sans erreurs. Or si on ignore ces erreurs, le risque est grand de produire une inférence écologique erronée, avec des biais dans l'estimation de la taille des populations ou des interactions entre espèces. Dans ce projet de thèse, nous développerons des méthodes statistiques pour quantifier et propager l'incertitude associée au traitement automatique des données issues de capteurs, et nous valoriserons ces méthodes sur deux cas d'étude sur des espèces de mammifères en interaction avec des activités humaines. Ce projet a obtenu le soutien financier du CNRS à travers le programme 80|Prime.

L'équipe encadrante : Les deux personnes qui co-encadreront la thèse, Marie-Pierre Etienne (IRMAR, https://marieetienne.github.io/) et Olivier Gimenez (CEFE, https://oliviergimenez.github.io/) travaillent en modélisation statistique, en traitant notamment des données de capteurs et sur les modèles de distribution des espèces. Ces deux chercheurs évoluent dans un milieu scientifique dynamique et le doctorant ou la doctorante pourra selon le type de questions solliciter des collègues ayant une solide expérience des capteurs (Simon Chamaillé-Jammes, CEFE) ou des méthodes d'apprentissage automatique (Vincent Miele, LBBE et Mathieu Emily, IRMAR). Le projet de thèse sera aussi l'occasion de tirer profit des liens existants avec des partenaires non académiques qui sont en charge de nombreux programmes de collectes de données impliquant des capteurs et qui sont des utilisateurs potentiels des méthodes développées dans ce projet. Il s'agit de l'Office Français de la Biodiversité (https://ofb.gouv.fr/) qui est en charge du suivi national de nombreuses espèces animales, et du bureau d'étude TerrOïko (https://www.terroiko.fr/) qui développe des approches de suivi de la biodiversité grâce au deep learning.

Le profil que nous recherchons : Nous recherchons une personne avec un profil en statistique intéressée par les applications en écologie, ou une personne au profil en écologie avec un intérêt fort pour la modélisation statistique. Les objectifs de la thèse seront à moduler selon le profil de la personne recrutée. Nous envisageons un déroulé sur une moitié du temps à Rennes (avec Marie-Pierre Etienne) et l'autre moitié à Montpellier (avec Olivier Gimenez) à mettre en place en accord avec la personne recrutée. Le profil de la personne recrutée déterminera également le choix de l'école doctorale de rattachement. Des séjours seront prévus et financés pour la personne en thèse dans les équipes des partenaires. Conscients de la difficulté de partager son temps entre deux lieux, nous veillerons à accompagner et soutenir la personne recrutée pour l'accès à un logement sur le site qui ne sera pas sa résidence principale.

État de l'art et objectifs de la thèse

L'utilisation de capteurs révolutionne la recherche en écologie en permettant le suivi des populations animales et végétales partout et à tout moment (Lahoz-Monfort & Magrath 2021). Les capteurs photographiques et acoustiques en particulier permettent de mesurer passivement les interactions entre espèces de mammifères (des relations entre des prédateurs et leurs proies par exemple), et quantifier les interactions de ces espèces avec les activités humaines (les risques de collisions avec les voitures ou les trains par exemple). Ces capteurs génèrent des bases de données de grandes dimensions et utilisent des méthodes d'apprentissage profond (« deep learning » ou DL) pour automatiser le traitement des images et des sons qui en sont issus. Les méthodes de DL connaissent un succès grandissant en écologie (Miele, Dray & Gimenez 2021). Les données ainsi traitées par les algorithmes de DL sont ensuite utilisées pour nourrir des modèles statistiques qui permettent de répondre à des questions écologiques. C'est le cas par exemple en écologie des populations avec l'estimation des effectifs des populations ou en écologie des communautés avec l'inférence sur les interactions entre espèces.

Le problème que nous avons identifié est que ces sorties sont systématiquement traitées comme des

données brutes, au même titre que des observations qui seraient faites par les écologues, et non pour ce qu'elles sont, c'est-à-dire des sorties de modèles, entachées d'erreurs de classification des individus et/ou des espèces (Gimenez et al. 2021). Or si on ignore ces erreurs, le risque est grand de produire une inférence écologique erronée, avec la sur-estimation des effectifs ou des interactions entre espèces (Chambert et al. 2018) ou pour le moins de sous-estimer la variabilité attachée à nos résultats. De plus les algorithmes de DL fournissent un degré de confiance dans leur classification qui pourrait dont être utilement exploitée pour prendre en compte une partie de l'erreur.

Dans ce projet, nous développerons des solutions visant à intégrer les erreurs de classification dans le traitement statistique des données écologiques. Il ne s'agit pas d'un travail en DL, mais bien d'un projet de recherche à l'interface de la statistique et de l'écologie.

Plutôt que travailler avec une classification dure, qui ne rend pas compte de l'incertitude de classement et de ses conséquences sur l'incertitude associée aux paramètres d'intérêt écologique, **l'objectif méthodologique** consistera à exploiter l'ensemble de l'information disponible dans la sortie des algorithmes de DL, et à développer des méthodes pour propager l'incertitude de classement dans les sorties des modèles écologiques. **L'objectif en écologie** sera d'étudier la structure des communautés grâce aux modèles joints de co-occurrence (ou de distribution) d'espèces. Nous considérerons deux cas d'étude sur des espèces de mammifères en interactions avec des activités humaines : l'un avec des capteurs d'images (pièges photographiques) et la question de la prédation du lynx sur les ongulés et les interactions avec la chasse, et l'autre avec des capteurs acoustiques avec la question de l'impact des infrastructures routières sur la répartition spatiale des chauves-souris.

Méthodologies

Volet 1. Une première idée consiste à développer une approche de Monte Carlo dans laquelle pour chaque entrée du jeu de données, on tire l'étiquette au hasard selon la loi de probabilité fournie par les sorties du DL. N répétitions de cette procédure d'association aléatoire produisent N jeux de données différents qui sont ensuite analysés classiquement avec des modèles de co-occurrence. La diversité des paramètres obtenus pour ces N jeux de données approchent la variabilité d'estimation que l'on devrait constater si on prenait en compte l'incertitude. L'intérêt de cette approche est sa simplicité de mise en œuvre qui permettra une implémentation rapide et qui permet également d'illustrer les conséquences attendues de la non prise en compte de l'incertitude.

Volet 2. Une deuxième approche, moins gourmande d'un point de vue computationnel mais plus complexe du point de vue de la modélisation et des méthodes d'inférence consiste à modifier les modèles écologiques pour qu'ils s'adaptent à la nature des sorties du DL. On peut tout d'abord envisager une approche hiérarchique qui permettra de rendre compte de la présence de faux positifs tout autant que de faux négatifs. Les probabilités de faux positifs et faux négatifs seront renseignées par les sorties du DL (Tabak et al. 2020). Le cadre d'estimation de tels modèles hiérarchiques est aujourd'hui assez bien balisé, notamment dans un cadre bayésien, et on dispose d'outils efficaces d'estimation (e.g. de Valpine et al. 2017) connus des écologues. Une autre piste consiste à modifier en profondeur les modèles écologiques pour qu'ils acceptent comme données d'entrée non pas une information sur la présence (possiblement bruitée) d'une espèce mais une loi de probabilité sur la présence de chacune des espèces considérées.

Volet 3. Enfin au-delà de la manière pertinente d'adapter les modèles de co-occurrence pour qu'ils puissent s'accommoder de l'incertitude des sorties de DL, nous aborderons la question de la régularisation des modèles utilisés. En effet, avec la démocratisation des méthodes de DL, et l'enrichissement des bases de données d'apprentissage, la tendance naturelle consiste à considérer un nombre de classes de plus en plus important conduisant à des modèles de co-occurrence avec un grand nombre de paramètres dont

probablement peu sont réellement informatifs. On pourra donc transférer les idées de régularisation utilisée dans les problèmes de grandes dimensions, notamment dans le cadre de l'écologie microbienne pour assurer une parcimonie aux modèles considérés.

Références citées

- Chambert et al. (2018). Two-species occupancy modelling accounting for species misidentification and non-detection. Methods Ecol Evol. 9: 1468-1477.

- Gimenez et al. (2021). Trade-off between deep learning for species identification and inference about predator-prey co-occurrence: Reproducible R workflow integrating models in computer vision and ecological statistics. Soumis à Computo. https://arxiv.org/abs/2108.11509.
- Lahoz-Monfort & Magrath (2021). Comprehensive Overview of Technologies for Species and Habitat Monitoring and Conservation, BioScience, 71: 1038–1062.
- Miele, Dray & Gimenez (2021). Images, écologie et deep learning. Regards sur la biodiversité, Société Française d'Écologie et d'Évolution. https://hal.archives-ouvertes.fr/hal-03142486.
- Tabak et al. (2020). Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: MLWIC2. Ecol Evol. 10: 10374-10383.
- de Valpine et al. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE, Journal of Computational and Graphical Statistics, 26:403-413.

Si vous êtes intéressés, n'hésitez pas à contacter
"Olivier Gimenez" <olivier.gimenez@cefe.cnrs.fr>
"Marie-Pierre ETIENNE" <marie-pierre.etienne@agrocampus-ouest.fr>

La thèse est financée par le CNRS via le programme 80|Prime.