HAL
open science

Statistical methods for modelling the spatial distribution of plant species from large masses of uncertain occurrences from citizen science programs

Christophe Botella

To cite this version:

Christophe Botella. Statistical methods for modelling the spatial distribution of plant species from large masses of uncertain occurrences from citizen science programs. Statistics [math.ST]. Université Montpellier, 2019. English. NNT : 2019MONTS135 . tel-02519161v3

HAL Id: tel-02519161

https://theses.hal.science/tel-02519161v3

Submitted on 23 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Biostatistique

École doctorale : Information, Structures, Systèmes

Unité de recherche UMR AMAP, Montpellier

Méthodes statistiques pour la modélisation de la distribution spatiale des espèces végétales à partir de grandes masses d'observations incertaines issues de programmes de sciences citoyennes

Présentée par Christophe Botella

Le 8 octobre 2019

Sous la direction de Alexis Joly
et Pascal Monestiez
et François Munoz

Devant le jury composé de

Salmon Joseph, Professeur, Université de Montpellier	Président
Guisan Antoine, Professeur, Université de Lausanne	Rapporteur
Illian Janine, Professeur, Université de Glasgow	Rapportrice
Chadoeuf Joël, Directeur de recherche, UR BioSP, INRA	Examinateur
Bonnet Pierre, Chercheur, UMR AMAP, CIRAD	Invité
Joly Alexis, Chercheur, LIRMM, INRIA	Directeur
Monestiez Pascal, Directeur de Recherche, UR BioSP, INRA	Directeur
Munoz François, Professeur, Université de Grenoble	Directeur

UNIVERSITÉ DE MONTPELLIER

1 Remerciements

Ce manuscrit achève trois ans de thèse, tout pile... Toujours pris 15 minutes de rab pour ne même pas finir mes devoirs, mais j'aurai au moins fini ma thèse dans les temps. Trois années durant lesquelles j'aurai grandit un peu, juste là où il fallait vraiment. Evidément, ça ne s'est pas fait tout seul, alors je voudrais remercier tous ces gens qui ont été décisifs pour aboutir à cette thèse, ont fait plus que leur job ou l'ont fait particulièrement bien, m'ont fait comprendre des nouvelles choses, m'ont aidé ou soutenu, même sans le savoir.

Merci au département de Mathématiques et Informatique Appliquées de L'INRA et au comité de la bourse nationale INRA-INRIA 2016 sans lesquels cette thèse n'aurait pas vu le jour.

Merci à mon cher labo de thèse, l'UMR AMAP qui a été un cadre vivant, diversifié et convivial. Je remercie en particulier Jean-François Molino, qui a soutenu ce projet depuis le début. Un grand merci à Nathalie Hodebert, Miléna Dordevic Giroud et surtout Noémie Cauquil qui avez été toutes les trois super et très cool pour m'accompagner dans l'organisation de ces nombreuses mobilités et plus encore. Un grand merci aussi à Gilles LeMoguedec pour m'avoir aidé à m'intégrer dès le début au labo, pour l'esprit fédérateur de tes cafés et pique-niques. Merci biensûr d'avoir été super compagnon de statistiques, pour ton aide technique, et pour ton exigence de rigueur. Globalement, merci d'avoir été aussi franc et à l'écoute. Merci à Thierry Fourcaud et au comité de direction d'AMAP, notamment pour avoir soutenu financièrement mes deux dernières conférence.

Merci aux autres labo qui m'ont accueilli. Merci aux personnes de l'INRA BioSP pour leur accueil, et en particulier à Sylvie Jouslin pour ton aide bienveillante dans l'organisation des missions, et Joël Chadoeuf. Merci à l'équipe Zenith (étendue) du LIRMM pour son accueil régulier tout au long de ma thèse, en particulier Valentin Leveau, Titouan Lorieul, Maximilien Saervajean pour nos discussions si riches sur l'apprentissage profond qu'elles ne pouvaient se terminer qu'après le travail au bar. Et re-merci à Valentin pour tes commentaires sur la thèse ici présente. Merci aussi à Benjamin Deneu, Lucas Bernigaud-Samatan, Mathilde Negri pour nos intéractions très intéressantes et votre bonne humeur. Merci au LECA, et particulièrement à François Munoz et Wilfried Thuiller pour m'avoir accueilli à trois reprises, où j'ai eu l'occasion de croiser pleins de chercheurs passionnés, à travers présentations, soutenances, paper clubs et j'en passe. Merci à Marc Ohlmann pour ton hospitalité et pour m'avoir guidé à la découverte de la culture Jazz Grenobloise =P, et à Laure Gallien pour ta bonne humeur punchy.

Merci au consortium Floris'Tic qui a soutenu de nombreuses missions, fourni des ressources matérielles et un réseau utile pour mon travail. Merci en particulier à l'équipe de Tela Botanica et qui a suivi et motivé mon travail de thèse. Merci à l'équipe Pl@ntNet pour m'avoir grandement aidé dans mon travail sur les données: Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Sébastien Breton, Hervé Goeau. Merci aussi à Frédéric Andrieu, James Molina et Guilhem De Barros, du Conservatoire Botanique Méditerrannéen pour leur aide dans l'utilisation des données SILENE.

Merci aux personnes qui m'ont formé durant la thèse. Marc Dumas, vous avez été un super prof sur le plan relationnel et vos cours m'ont été réellement utiles pour gérer les nombreuses réunions et prestations orales pendant ma thèse. Merci aussi à Benoite de Saporta pour m'avoir acceuilli dans votre cours. Merci à Frédérik Garcia, Michaël Chelle, Véronique Broussolle, Geneviève Aubin-Houzelstein, pour votre organisation de la formation EDEN qui m'a donné un matériel réflexion et de placement.

Merci aussi aux Groupements de Recherche (GDR) Cisstats pour les rencontres fructueuses

qui ont été faîtes lors des deux rencontres auxquelles j'ai participé, et merci encore à Pascal pour avoir organisé tout ça. Merci aussi au GDR Ecostat, qui a financé deux mobilités durant ma thèse, et en particulier à ses coordinateurs Olivier Gimenez et Stéphane Dray. Les journées du GDR Ecostat était un lieu de rencontres professionnelles et personnelles très riche. Merci au comité d'organisation des campagnes LifeCLEF et des conférences CLEF, qui m'a ouvert vers des horizons différents. Merci à MUSE et à l'Université de Montpellier financé ma mobilité en Ecosse, et à Janine Illian de m'avoir si bien accueilli sur place, une super expérience humaine et scientifique.

Merci à mes encadrants qui m'ont fait confiance dès le début de la construction de ce projet, et m'ont soutenu malgré les difficultés de la distance. Vous m'avez donné une grande liberté, et m'avez fait prendre conscience de la responsabilité qui va avec. Plus particulièrement, merci Alexis d'avoir toujours cherché à comprendre en profondeur mon travail, pointé les faiblesses et avoir considéré mes doutes, m'avoir poussé vers des questions plus ambitieuses, pour m'avoir donné de nombreuses opportunités, pour ta réactivité, et pour avoir toujours été cool au travail. Merci Pascal, pour m'avoir donné toutes ces opportunités de rencontres, de mobilités, de conférences, les avoir supporté financièrement, pour m'avoir guidé dans la compréhension de le monde complexe de la recherche, pour tes sages conseils et ton recul scientifique. Merci François, pour ta réactivité, tes conseils d'écriture et pour ta solidarité sur les calembourgs. Merci Pierre, mon ange gardien, pour ton soutien constant et ta gentillesse. Merci pour avoir été toujours disponible quand j'avais des problèmes et personne à qui m'adresser autour de moi. Tu m'as aussi convaincu de la valeur du sang froid dans la résolution des conflits.

Merci Jean François Molino, Isabelle Chuine, Jean-Jacques Boreux et Nicolas Verzelen pour votre participation à mes comité de suivi et vos conseils. Merci Jean-Noël Bacro, notamment pour m'avoir aidé et guidé dans cette pénible démarche du changement d'école doctorale. Je remercie également Janine Illian et Antoine Guisan de rapporter ma thèse, à Joël Chadoeuf de l'examiner et à Joseph Salomon de présider le beau monde de ce jury.

Merci aux collègues d'AMAP présents du café à l'apéro, de passages plus ou moins prolongé, qui m'avez donné la patate jours après jours: Mélaine Aubry, Anthony LaBrioche (dit frère tuck), François Postic (dit Le tobogan et fidèle comparse d'humour noir), Olivier Martin (prononcé holiviaaaaiiiieuh), François Grand, Claudia, Yohann Soubeyran, Alain Ibrahim, Nicolas Beudez, Antoine Affouard, Hervé Goeau, Rémi Knaff, Mathias Chouet, Sébastien Breton, Adam Mora, Sami Youssef, Jose Carranja Rojas, Geovanni Figueroa Mata, Julien Engel, Guillaume Delaître, Eva Grill, Malo Bourget, Alice Penanhoat, et ceux que j'oublie! Merci tout particulier à mes co-bur, et dresseurs d'AMAPokemons, Dimitri et Sue: Vous êtes géniaux.

Merci Grégory Guirard, mon plexus se souviendra longtemps de tes coups de pieds, qui m'ont appris beaucoup plus qu'à trouver mon second souffle. Merci d'avoir été impitoyable, et d'avoir construit ce groupe génial. Merci aussi à Mégane et Jean-Pierre pour votre bienveillance. Vous avez aussi ravivé un esprit de compétition, un peu enfoui, mais dont j'avais besoin. Merci à mes co-équipiers et copains: Alex, Captain, Anne-Claire, Mathilde, William, Micka, Pauline, Nico, Louis. Qu'on se le dise, Cyril, Yako, Marc merci pour m'avoir ramené dans la voie du vrai travail, m'empéchant de dissiper mon énergie dans l'alimentaire, le futile, le strass et les paillettes. Merci Ludivine, pour ta patience au Seven Wonders, pour tous ces moments de joie innocente et aventures partagées, qui restent intacts. Merci à toi aussi Touille, élément clé de la team Palais. Merci Pierre Candelon pour ta bravoure inspirante dans cette eau froide.

Un merci suprême à mes parents. Vous avez été tout près de moi, à faire grossir ma dette avec une constance admirable, jusqu'aux dernières lignes de cette thèse. Merci de pas m'avoir

cru quand j'ai dit: "Tfaçon, j'ai jamais aimé les maths"... Donc si je me retrouve à pédaler dans la semoule devant ce jury, c'est un peu à cause, mais surtout grâce à vous. Zéro mot peut décrire ma gratitude pour votre amour inconditionnel.

2 Résumé et mots-clefs

L'expertise botanique humaine devient trop rare pour fournir les données de terrain nécessaires à la surveillance de la biodiversité végétale. L'utilisation d'observations botaniques géolocalisées des grands projets de sciences citoyennes, comme Pl@ntNet, ouvre des portes intéressantes pour le suivi temporel de la distribution des espèces de plantes. Pl@ntNet fourni des observations de flore identifiées automatiquement, un score de confiance, et peuvent être ainsi utilisées pour les modèles de distribution des espèces (SDM). Elles devraient permettre de surveiller les plantes envahissantes ou rares, ainsi que les effets des changements globaux sur les espèces, si nous parvenons à (i) prendre en compte de l'incertitude d'identification, (ii) correction les biais d'échantillonnage spatiaux, et (iii) prédire préçisément les espèces à un grain spatial fin.

Nous nous demandons d'abord si nous pouvons estimer des distributions réalistes d'espèces végétales envahissantes sur des occurrences automatiquement identifiées de Pl@ntNet, et quel est l'effet du filtrage avec un seuil de score de confiance. Le filtrage améliore les prédictions lorsque le niveau de confiance augmente jusqu'à ce que la taille de l'échantillon soit limitante. Les distributions prédites sont généralement cohérentes avec les données d'expertes, mais indiquent aussi des zones urbaines d'abondance dues à la culture ornementale et des nouvelles zones de présence.

Ensuite, nous avons étudié la correction du biais d'échantillonnage spatial dans les SDM basés sur des présences seules. Nous avons d'abord analysé mathématiquement le biais lorsque les occurrences d'un groupe cible d'espèces (Target Group Background, TGB) sont utilisées comme points de fond, et comparé ce biais avec celui d'une sélection spatialement uniforme de points de base. Nous montrons alors que le biais de TGB est dû à la variation de l'abondance cumulée des espcèes du groupe cible dans l'espace environnemental, qu'il est difficile de contrôler. Nous pouvons alternativement modéliser conjointement l'effort global d'observation avec les abondances de plusieurs espèces. Nous modélisons l'effort d'observation comme une fonction spatiale étagée définie sur un maillage de cellules géographiques. L'ajout d'espèces massivement observées au modèle réduit alors la variance d'estimation de l'effort d'observation et donc des modèles des autres espèces.

Enfin, nous proposons un nouveau type de SDM basé sur des réseaux neuronaux convolutifs utilisant des images environnementales comme variables d'entrée. Ces modèles peuvent capturer des motifs spatiaux complexes de plusieurs variables environnementales. Nous proposons de partager l'architecture du réseau neuronal entre plusieurs espèces afin d'extraire des prédicteurs communs de haut niveau et de régulariser le modèle. Nos résultats montrent que ce modèle surpasse les SDM existants, et que la performance est améliorée en prédisant simultanément de nombreuses espèces, et sont confirmés par des campagnes d'évaluation coopérative de SDM menées sur des jeux de données indépendants. Cela supporte l'hypothèse selon laquelle il existe des modèles environnementaux communs décrivant la répartition de nombreuses espèces.

Nos résultats supportent l'utilisation des occurrences Pl@ntnet pour la surveillance des invasions végétales. La modélisation conjointe de multiples espèces et de l'effort d'observation est une stratégie prometteuse qui transforme le problème des biais en un problème de variance d'estimation plus facile à contrôler. Cependant, l'effet de certains facteurs, comme le niveau d'anthropisation, sur l'abondance des espèces est difficile à séparer de celui sur l'effort d'observation avec les données d'occurrence. Ceci peut être résolu par une collecte complémentaire protocollée de données. Les méthodes d'apprentissage profond mises au point montrent de bonnes performances et pourraient être utilisées pour déployer des services de prédiction

spatiale des espèces.

Mot-clés: Surveillance de la biodiversité; sciences-citoyennes; espèces exotiques envahissantes ; habitats d'espèces de plantes ; Modèles de Distribution d'Espèces ; Données de présence-seule ; biais d'échantillonnage ; Ecologie statistique ; Réseaux de Neurones Convolutionnels Profonds ; pouvoir prédictif ; recommandation d'espèces

3 Abstract et keywords

Human botanical expertise is becoming too scarce to provide the field data needed to monitor plant biodiversity. The use of geolocated botanical observations from major citizen science projects, such as Pl@ntNet, opens interesting paths for a temporal monitoring of plant species distribution. Pl@ntNet provides automatically identified flora observations, a confidence score, and can thus be used for species distribution models (SDM). They enable to monitor the distribution of invasive or rare plants, as well as the effects of global changes on species, if we can (i) take into account identification uncertainty, (ii) correct for spatial sampling bias, and (iii) predict species abundances accurately at a fine spatial grain.

First, we ask ourselves if we can estimate realistic distributions of invasive plant species on automatically identified occurrences of Pl@ntNet, and what is the effect of filtering with a confidence score threshold. Filtering improves predictions when the confidence level increases until the sample size is limiting. The predicted distributions are generally consistent with expert data, but also indicate urban areas of abundance due to ornamental cultivation and new areas of presence.

Next, we studied the correction of spatial sampling bias in SDMs based on presences only. We first mathematically analyzed the bias when the occurrences of a target group of species (Target Group Background, TGB) are used as background points, and compared this bias with that of a spatially uniform selection of base points. We then show that the bias of TGB is due to the variation in the cumulative abundance of target group species in the environmental space, which is difficult to control. We can alternatively jointly model the global observation effort with the abundances of several species. We model the observation effort as a step spatial function defined on a mesh of geographical cells. The addition of massively observed species to the model then reduces the variance in the estimation of the observation effort and thus on the models of the other species.

Finally, we propose a new type of SDM based on convolutional neural networks using environmental images as input variables. These models can capture complex spatial patterns of several environmental variables. We propose to share the architecture of the neural network between several species in order to extract common high-level predictors and regularize the model. Our results show that this model outperforms existing SDMs, that performance is improved by simultaneously predicting many species, and this is confirmed by two cooperative SDM evaluation campaigns conducted on independent data sets. This supports the hypothesis that there are common environmental models describing the distribution of many species.

Our results support the use of Pl@ntnet occurrences for monitoring plant invasions. Joint modelling of multiple species and observation effort is a promising strategy that transforms the bias problem into a more controllable estimation variance problem. However, the effect of certain factors, such as the level of anthropization, on species abundance is difficult to separate from the effect on observation effort with occurrence data. This can be solved by additional protocolled data collection. The deep learning methods developed show good performance and could be used to deploy spatial species prediction services.

Keywords: biodiversity monitoring ; crowdsourcing ; citizen-sciences; invasive alien species ; plants species habitats ; Species Distribution Models ; Presence-only data ; Sampling bias ; Sampling effort ; Deep Convolutional Neural Networks ; predictive power; species recommendation ; statistical ecology

"*Ça ne prenait pas beaucoup plus de temps! ... Ce qui est difficile c'est la partie pédalo, c'est pas la partie canard.*"
Hubert Bonisseur de La Bath

List of Figures

1 Mapped counts for Pl@ntNet queries and distribution over land cover types . . 17
2 Histograms representing the distribution of points distance to the French roads network (*autoroutes*, *nationales* and *départementales*) for a random subset of 50,000 Pl@ntNet 2018 geolocated queries (Top) and 50,000 points uniformly drawn over the French territory (Bottom) 18
3 Conceptual illustration of fundamental, potential and realized niches of a species in the environmental and geographic spaces 22
4 Definitions and notations 24
5 Illustration of the decomposition of the sampling process for an occurrence data 38
6 TOP30 test accuracy versus geographical distance to plants training occurrences for several runs of GeoLifeCLEF 2019 202

Contents

1 Remerciements **2**

2 Résumé et mots-clefs **5**

3 Abstract et keywords **7**

4 Introduction **11**
4.1 Context 11
4.2 Available spatial data on species for distribution models 12
4.3 Species Distribution Models (SDM): Overview and state-of-the-art of presence-only methods 20
4.3.1 Niche theory 20
4.3.2 Studying the response functions 21
4.3.3 Species distribution models overview 23
4.3.4 MAXENT 29
4.3.5 The unifying framework of point processes models 31
4.4 Boundaries and pitfalls of Species Distribution Models based on massive opportunistic occurrences 36
4.4.1 Biases due to sampling heterogeneity in presence only data 36
4.4.2 Input dimension constraints in the era of deep learning 41
4.5 Questions of the thesis 47

5 Chapter 1:
Species distribution modelling based on the automated identification of citizen observations **50**

6 Chapter 2:
Bias in presence-only niche models related to sampling effort and species niches: lessons for background points selection **62**

7 Chapter 3:
Estimating sampling effort across space from large amounts of species occurrences **81**

8 Chapter 4:
A Deep Learning Approach to Species Distribution Modelling **125**

9 Chapter 5:
A cooperative evaluation campaign for species recommendation algorithms, GeoLifeCLEF **157**

10 Discussion **192**
10.1 Results synthesis 192
10.1.1 Handling identification uncertainty. 192
10.1.2 Spatial sampling bias correction. 193
10.1.3 Multi-species prediction from environmental images using deep learning. 194
10.2 Perspectives 196
10.2.1 Sampling biases 196
10.2.2 Point processes ecological interpretations, related assumptions and limits 200
10.2.3 Studying transferability of complex species distribution models 201
10.2.4 Improving SDM predictions by accounting for species interactions . . . 203

11 Appendices **225**
11.1 Appendix of chapter 2 225
11.2 Appendix of chapter 3 241

4 Introduction

4.1 Context

Climate change is not the only threat to biodiversity. Other effects due to growing human pressure on the land bring much more brutal changes which silently affect the ecosystems all around the world. Two other worldwide anthropogenic factors of peril on wild plants have been identified in Kew (2016): invasive species and land use change.

Invasive plant species continue to appear and spread, brought by increased commercial exchanges and population movement, modifying the ecosystems on their way and causing damages to agriculture and people health (Vitousek et al. (1996)). Invasive plant species are estimated to cost around 1.7 billions pounds to the U.K. every year (Kew, 2016). Recent human-induced land cover change has been shown to globally decrease the vegetation productivity. Indeed, Great proportion of the land cover have been observed to change between 2001 and 2012, especially in mangroves (25%) and tropical coniferous forests (24%), mainly due to conversion of forest to pastures or farmland, inducing a decrease of primary productivity of 2.5% in both biomes (Kew, 2016). The global decrease of forest area worldwide is of about 1.7% between 1990 and 2005, according to Lindquist et al. (2013)). These changes are quite visible, but others are not so easily detected. Agriculture practices are in continuous change. A growing use of pesticides has been observed globally (Oerke, 2006), making life harder for many weed species. In France, change in crop management in the past decades has transformed the habitat of weed plants which affect their regional (Fried et al., 2008) and national (Fried et al., 2010) species diversity. An even more preoccupying phenomenon is biodiversity decrease runaway. Indeed, it has been shown that less diverse local communities are more sensitive to invasions (Kennedy et al., 2002). These phenomenons are ongoing and they have rapid impacts on the ecosystems, especially on plants.

However, the inability of resource managers, scientists, and policy makers to efficiently and effectively prevent, control and react to these phenomenons has already resulted in environmental and economic losses worldwide (Heywood et al. (1995), Born et al. (2005), Tscharntke et al. (2012)). This inability is importantly due to the complexity of ecosystems and a lack of objective, sufficient and regularly updated data that would enable the monitoring of species distributions. Indeed, according to the last ICUN redlist evaluation [1], 13,494 plant species have at least a vulnerable status, and 157 among them are extinct at least in the wild. This evaluation was done for 25,996 plant species having enough data, that is around 7% of all described plant species in the world, according to ThePlantList [2]. Plants are thus among the most critical terrestrial biological groups in terms of lack of knowledge in this regard. Assessment of species status and future vulnerability rely, in particular, on species distribution modelling methods because they provide abundance maps and population response to various types of environments (e.g.:Norris (2004),Thomas et al. (2004)). It is also a primary material to prioritize reserve areas (Ferrier (2002), Loiselle et al. (2003)), evaluate their ability to preserve species habitats given spatial shifts due to climate change (Araújo et al., 2004) or suggest suitable sites for re-introduction (Pearce and Lindenmayer, 1998).

Identifying and preventing species extinction is crucial to conserving biodiversity but often treats only a symptom of a deeper problem of an ecosystem disruption. Ecological research enables to understand the functioning of ecosystems, including the dependency of species to the environment but also their interactions which may enable to anticipate the evolution of the

[1] `https://www.iucnredlist.org/resources/summary-statistics`

[2] `http://www.theplantlist.org/`

ecosystem under perturbations, like removal or addition of some species. It also often relies on statistical species distribution and ecological niche models (Peterson and Soberón, 2012) tools. These models have been used to predict future potential plant species distribution and extinction under climate change (Thomas et al. (2004),Thuiller (2004),Thuiller et al. (2005)), or land use change (Thuiller et al., 2008), invasion paths under climate change (Beerling et al. (1995),Peterson (2003)), or study the invasibility of plant communities (Richardson and Pyšek, 2006).

However, addressing the challenges of plant biodiversity conservation at the scale of the world flora will require to achieve a scale up of taxonomic and spatial coverage and to monitor the effects of current changes on wild populations in order to fill the many gaps of ecological knowledge. This primarily requires to exploit biodiversity data at its maximum. A promising lever for action is to develop species distribution modelling methods that deal appropriately with the worldwide growing available crowd-sourced data.

As awareness of the perils on biodiversity progresses, more and more people get interested to it and eventually invest their time learning about species, or even collect data on the field while wandering in the nature by the means of collaborative naturalist platforms, often in association with citizen-science programs. This new dawn for biodiversity citizen-sciences has largely been favored by the development of online platforms to store, revise, manage, explore and share biological records (e.g. Silvertown et al. (2015)) and mobile devices to collect the data (Graham et al., 2011). These organisations have already contributed to major successes in the conservation of certain species, and this is particularly visible in the case of birds, which have so far benefited from long-term voluntary commitment through extensive citizen science programs (Greenwood, 2007). Plant species are more difficult to identify which has for long restrained the volunteer data collection to a few skilled botanists. Nevertheless, the recent improvement of automated plant species identification from pictures (Wäldchen et al. (2018)) such as in the context of the mobile application Pl@ntNet, brought many amateurs to produce reliable geolocated occurrences of plant species, sometimes even engaging in volunteer field reporting. The huge amount of spatial data collected on thousands of plant species in last years has motivated their use for an automatic regular monitoring of plant species distribution based on citizen observations. As the data is mainly produced around urban and accessible areas, it provides a great opportunity to monitor especially alien invasive plant species, but also species endangered by human activity. Thus, the work of this thesis aim at investigating opportunities and resolving limits in the use of such massive and opportunistic species records produced through crowd-sourcing. More precisely, the species records are sampled without protocol and very heterogeneously in space, entailing biases in species distribution models, which motivates a focus on the development of efficient bias correction techniques.

4.2 Available spatial data on species for distribution models

Spatial data on biodiversity is always limiting to address most ecological questions over large taxonomic, geographic and temporal scales as illustrated by the seven shortfalls of biodiversity knowledge (Hortal et al., 2015). Still, the amount of data available for research on species distribution has dramatically increased in the last decade, as it can be seen on GBIF portal [3]. It is explained by the conjunction of web platforms and mobile technologies (Graham et al., 2011). More precisely it is due to the emergence of large data sharing platforms, including or separated from collaborative data revision platforms (implying non-experts), automatic data annotation and mobile devices that jointly enable the collection of consistent geolocation

[3] www.gbif.org

and pictures. These technologies have led to a diversification of large scale data producers (recrudescence of citizen-scientists and volunteers). We will briefly describe the interests and peculiarities of large scale citizen-sciences and naturalist platforms producing opportunistic presence-only data, and provide some quantitative insights on the Pl@ntNet data. Finally, some comments are given about the geographic environmental data that are used for plant SDM.

Sampling protocol type. It is crucial for the use of biodiversity data with an appropriate type of SDM to characterize the type of sampling protocol that was used to collect it. Firstly, we can differentiate standardized data from nonstandardized data (Miller et al., 2019). On one hand, standardized data has a well defined sampling design and fixed protocol at known sampling location, so that the effect of the observation process is controlled. They are most informative about the species that is measured, but are often costly to collect, constraining, have restricted access and concern restricted areas and/or taxa. Some typical examples of standardized data are abundance, presence-absence (e.g. Violle et al. (2015), Maitner et al. (2018)), counts with standardized effort (Giraud et al. (2016), Sauer et al. (2017)), plants relative cover (like Braun-Blanquet scores Brisse et al. (1995)), camera traps (Ahumada et al., 2011) and probably tele-detection data in a few years (He et al., 2015b). On the other hand, nonstandardized data, as defined in Miller et al. (2019), are "Data not collected under standardized protocol, where sampling locations and sampling effort are often unknown and sampling protocol varies". Thus, the exact observed areas, observation intensity and time, the detection capacity of observers and their reporting behavior are partially or totally unknown. These nonstandardized data are much less informative about the species because of all the uncertainty about the sampling process, but are cheaper and easier to collect. Typical types of partially standardized data are distance sampling, where observers move along a pre-defined transect while observing some species around them (Buckland et al., 2005), and site-occupancy data (MacKenzie and Nichols, 2004), where observers report that they visited a site and if they detected some species. For those types of data, the observation effort along the transect or inside the site, and the capacity of each observer to detect and identify the species variably affect the chances that the observer reports the species when present. Finally, species geolocated **occurrences**, often also referred as **presence-only data** (Pearce and Boyce, 2006), report that a species was present at a point at a certain time. Informations about the sampling behavior of contributors may go with a presence-only dataset. For example, Bradter et al. (2018) interviewed active observers gathering bird occurrences on their reporting behavior and detection capacities to infer some species absences where observers reported other species but not the focal one. Presence-only data is said to be **opportunistic** (Kery et al., 2010) when there is no rule guiding consistently the sampling process across the dataset and individual rules are unknown, i.e. no focus on determined sites or species, and observer may change of reporting behavior along the duration of data collection. It contains less information about the sampling process than previous data types because there is no report of non-detection of a species, and in particular, no report doesn't mean absence. The detection capacities is unknown with opportunistic occurrences, as in site-occupancy data, but we haven't anything either about the area that have been observed except the point where an occurrence was reported, thus we have less information about the observation effort. Many uses of presence-only data in presence-absence SDM by generating pseudo-absences have thus led to estimation biases as it is explained in **sections 5.4**. Still, these have long been the most abundant source of biodiversity data, considering that the number of specimens in museums and herbariums is estimated at around 2-3 billions Krishtalka and Humphrey (2000) and with a significant

part of geolocated and dated ones. Plus, they are currently becoming even more predominant through the exponential development of crowd-sourcing: The Global Biodiversity Information Facility (GBIF) [4] stores around 1,2 billions of geolocated occurrences at the time of writing, 30% of which were added during the 4 last years. Thus, it is crucial to properly account for the sampling process uncertainty in SDM applied to opportunistic occurrences, which is why it is at the center of the present thesis.

Also note that the level of standardization has nothing to do with the expertise or professional status of the data producer. Standardized data may be collected by non-expert volunteers (Sauer et al., 2017), while presence-only data are often collected by experts. For example, the current biggest plant occurrences dataset in France is the SiFlore dataset distributed by the FCBN (Just et al., 2015) which regroups around 20 millions occurrences collected by botanical experts from the regional botanical conservatories. Even though, high level of standardization are often time consuming for data collection, and are difficult to implement in crowdsourcing.

Data quality. Many aspects of data quality are important to consider when one aim at modelling species distributions. Concerns of data quality have been reviewed by (Pipino et al., 2002) who characterized the set of aspects that should be addressed on the quality of the data to use. We extracted the main aspects that should be regarded when dealing with biodiversity data : Accessibility (consider cost of accessing the data, giving credit, restrictions on usage and sharing of the results), appropriate amount for the purpose (typically depend on the sampling protocol and question), completeness (incomplete data require specific methods or partial removal of the data), representation consistency (consider cost of homogenisation of format, and potential loss and incompleteness resulting from the process), timeliness (the extent to which is sufficiently up-to-date for the task), value-added and free-of-error. The error of the data is important to consider for the accuracy of SDM. The data may contain species identification errors, which is an important problem for plants. Indeed, there are around 350,000 accepted plant species recognized by the ThePlantList [5] (accepted names), and many species look closely alike. The discrimination of plant species often suppose to observe reproductive organs which presence depend on the plant phenology. Their analysis require a particular attention which is not easy for an non experimented human eye (Hawthorne and Lawrence, 2013). Also, occurrences having different species label may actually correspond to a same true species because of e.g. species synonyms, or taxonomic referential discrepancies when merging occurrences datasets of various origins (Isaac et al., 2004). This is again typical in the case of plants, whose taxonomy is heterogeneous around the world. This is has been recognized as a major limitation to the renewable production of data on plant biodiversity given the decreasing number of botanical experts worldwide (Paknia et al., 2015). It is a good practice to match initial names to a single reference taxonomic referential, e.g. using the Taxonomic Resolution Service (Boyle et al., 2013). Another source of error is geolocation uncertainty, which remain a problem, even though it has been drastically reduced by the development of naturalist mobile applications. Indeed, acquisition of GPS coordinates varies in space, it is especially difficult to obtain accurate geolocations in covered areas such as forests or inside steep valleys (Frair et al. (2004)). A great part of museum and herbariums occurrences lack geolocation information, or it is inaccurate, or this information is written in text only (Graham et al. (2004),Newbold (2010),Beaman et al. (2004)). In the Pl@ntNet data, geolocation uncertainty is mostly concentrated between 0 and 60 meters, but species richness may exceeds hundreds of species inside the area of uncertainty and the vegetation profile itself may change

[4] `https://www.gbif.org/`

[5] `theplantlist.org`

along such distance. Also note that services exist for automatically georeferencing species occurrences data and computing accuracy attributes from textual information (Guralnick et al. (2006), Hill et al. (2009)).

The era of citizen-sciences programs and collaborative naturalist platforms. Many worldwide crowdsourced opportunistic occurrences data sources became accessible through the internet in the last decade based on either large-scale citizen-science organizations (e.g. Gillings et al. (2019), Sullivan et al. (2009), Affouard et al. (2019)), web based naturalist community platform (e.g. iNaturalist[6], Naturgucker [7]). Great national crowdsourcing platforms also deal with the same scales of data (e.g. Nyegaard (2019), Shah and Coulson (2019)). Indeed, a resurgence of citizen science has occurred in recent decades, even though the concept dates back to at least 1900 with the beginning of the Christmas Bird Count (Silvertown, 2009). For instance, eBird (Sullivan et al., 2009) is now the largest contributor to the GBIF with around 500 millions geolocated occurrences (40% of total) around the world at the time of writing. The data and results produced by bird citizen-sciences projects such organizations had numerous beneficial impacts on conservation management (Greenwood (2007),Schuster et al. (2019)). Besides, contributing to improvements in species conservation is an important motivation of citizens participating in such projects (Hobbs and White, 2012). Those mechanisms induce a positive loop of involvement and results operating with citizen-sciences applied to biodiversity, even if it has been acknowledged that participation remains unevenly spread among socio-economic classes and ethnic groups (Hobbs and White, 2012) and further work is needed to understand the specific barriers to participation, which should include better communication of scientific and conservation outcomes (Novacek, 2008).

Introduction and opportunities of Pl@ntNet data. The overall growing scarcity of professional botanical and taxonomist experts implies a bottleneck in the training of new experts. Despite the many small structures collecting spatialized plant data, the lack of expertise needed to produce high quality species identification, restrains the use of these data for research and conservation, even in collaborative reviewing systems. Then, verification and revisions of the mass of uncertain observations has become a real burden for expert botanists, and automatic plant species identification has become an important need (Gaston and O'Neill, 2004). Pl@ntNet is a citizen-sciences project and a mobile phone application providing a service of automatic plant species identification from pictures, based on deep learning algorithms (Affouard et al., 2019). Some of its goals are to unlock the access to plant identification to a large number of citizens through automated identification, restore human identification expertise, promote plant sciences, and generate renewable data for a use in plant ecology and biodiversity research. In this perspective, the project collaborates since its begins with the network of French botanists Tela Botanica [8] (Botanica, 2019). The application now works on a significant part of the world flora (ThePlantList taxonomic referential [9]), and has also 17 specific botanical taxonomic referentials dedicated to countries or regions of the world covering in total most of America, western Europe, Africa, and many islands. Users of Pl@ntNet ask for species identification by sending a set of pictures, along with their geolocation, the identification engine produces a prediction of species ranked by probability which is sent back to the user, with a illustrative pictures support for each likely species and

[6] `https://www.inaturalist.org/`
[7] `https://naturgucker.de/natur.dll/`
[8] `https://www.tela-botanica.org/`
[9] `theplantlist.org`

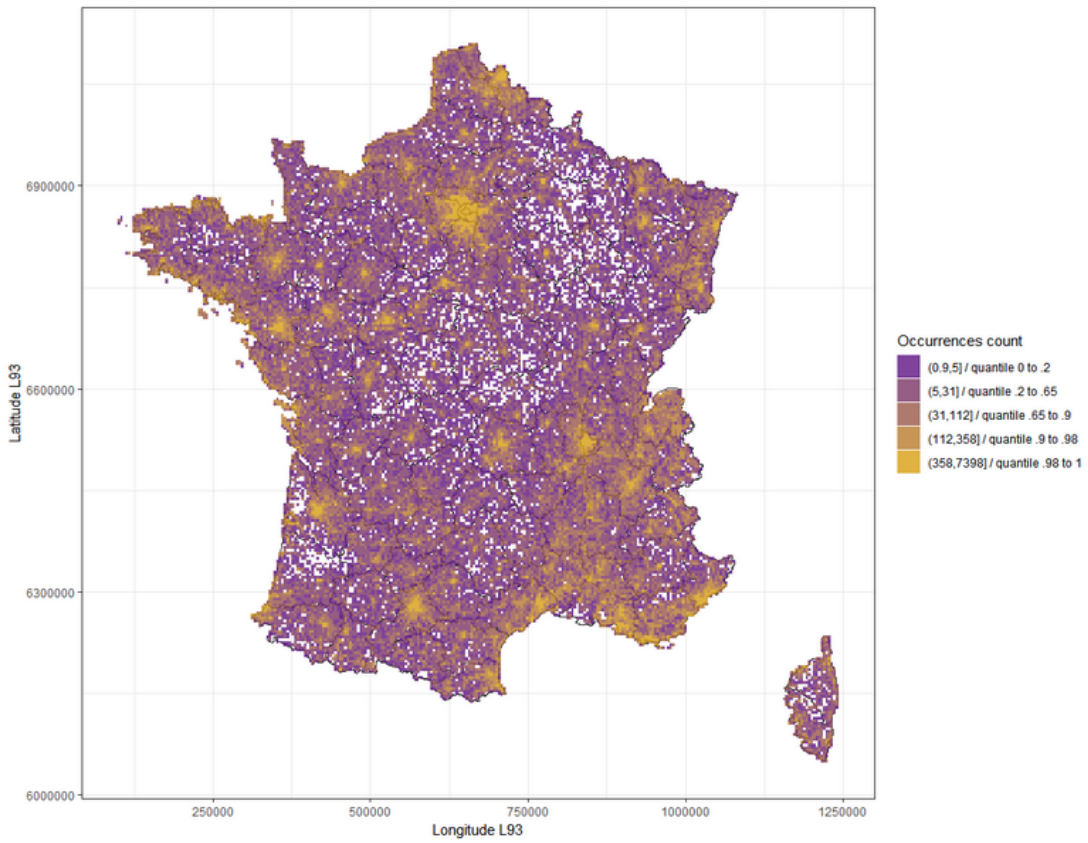
organs. Each time a user ask for identification, it produces a *query* associated to an automatic species ranked list, which is often geolocated. When the user receives the result, he may select the species he thinks is the best identification, and his choice is also stored in Pl@ntNet servers as an *observation*. A confidence score is attributed to the observation depending on a user experience score. Observations pass in real time in a flow seen by all active users who may revise the identification, which will increase the confidence score. This collaborative revision system enables more qualified volunteers to revise uncertain identifications, and to less experimented users to learn from their errors. Once it passes a certain confidence threshold, the observation pictures are added to the database used for training the algorithm, defining an active learning loop. The improvement is not only due to this enrichment of the database but also to innovations in the algorithm design and implementation, integration of complementary plant pictures databases from Encyclopedia Of Life (EOL[10]) and a crawling of web images. These innovations were guided by results of an annual evaluation campaign of algorithm for visual plant species predictions organized in the context of LifeCLEF (Joly et al. (2019)). The Algorithms for plant identification evaluated in the 2014 and 2017 editions of LifeCLEF were compared to identification skills of botanists with various expertise level, from students to the best experts of the French flora. It showed that the algorithms performances have drastically increased in four years. In 2014, most experts botanists were superior to the best algorithm, but in 2017, the best algorithms using deep convolutional neural networks and noisy data from the web, are all better than all experts except the best of them (Bonnet et al. (2016), Bonnet et al. (2018)).

Pl@ntNet capitalizes today around 50 millions geolocated *queries* worldwide, among which 5% have an *observation*. From the beginning of 2019 until the end of July, the application was used 30 millions of times, generating an average of 115,000 geolocated queries per day and 335,578 distinct users have contributed observations or votes. For illustration, Figure 1 represents the geographic distribution of geolocated queries collected during year 2018, restricted to France, and the proportions of land cover categories where they have been collected, compared to the global proportions of land cover categories over the French territory. The IT infrastructure behind Pl@ntNet has been recently reworked (Affouard et al., 2019), and it enables to manage, store, and explore in a reliable and unified way the all data including pictures taken by users, and should support its current exponential growth of use for the years coming.

Thus, the automatically predicted species probability is a good indication of the identification certainty. Geolocated queries may then be used as occurrences for SDM taking this identification probability score into account, as it is done in **Chapter 1**, but they are opportunistic in nature. Users don't have any sampling scheme, we don't know where the user has observed, what where his detection abilities and its species interests at time of the query. This justify the investigation of spatial sampling bias correction in **Chapter 2 and 3**, on the basis of existing work introduced in **section 5.4.1**. Resolving this pitfall would enable a standardized use of Pl@ntNet queries, which offer a unique spatio-temporal coverage of species that are remarkable to most citizens, especially in anthropogenic habitats. In particular, they provide a great opportunity to monitor the emergence and colonization of alien invasive species. Their high geolocation accuracy and control of uncertainty also enables a fine scale study of plant distribution in various environment and their adaptation to very local environmental changes due to human activity.

Environmental data. Spatial environmental data is a complementary very important subject for SDM. A first review was recently published by Mod et al. (2016), who synthesized

[10] `http://www.catalogueoflife.org/portfolio/encyclopedia-life`



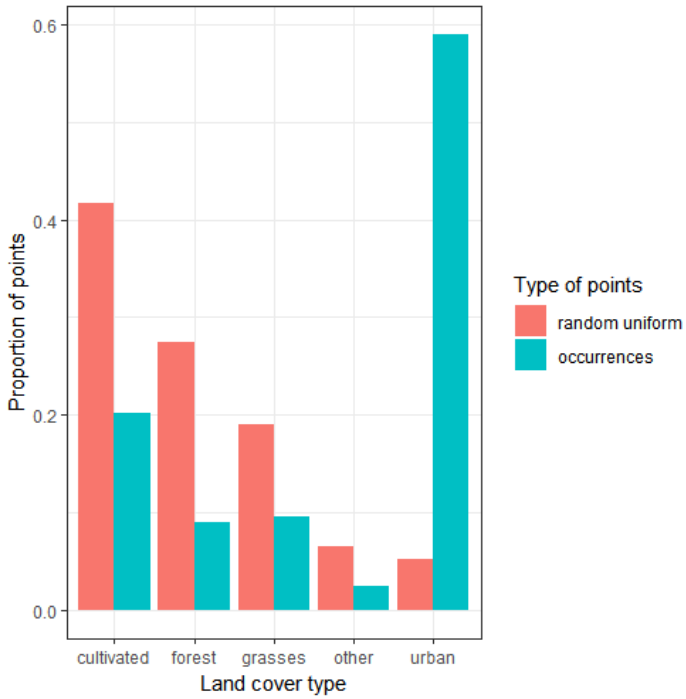


Figure 1: (Top) Mapped counts for Pl@ntNet geolocated queries (1st November 2017 - 1st November 2018) over 4x4km squares. Quantiles bounding the discrete color-scale units have been computed overall non-null square counts. (Bottom) Proportion of simplified land cover categories (aggregation of CORINE land cover 2012) for a random subset of 50,000 Pl@ntNet 2018 geolocated queries (blue) and 50,000 points uniformly drawn over the French territory (pink-orange).

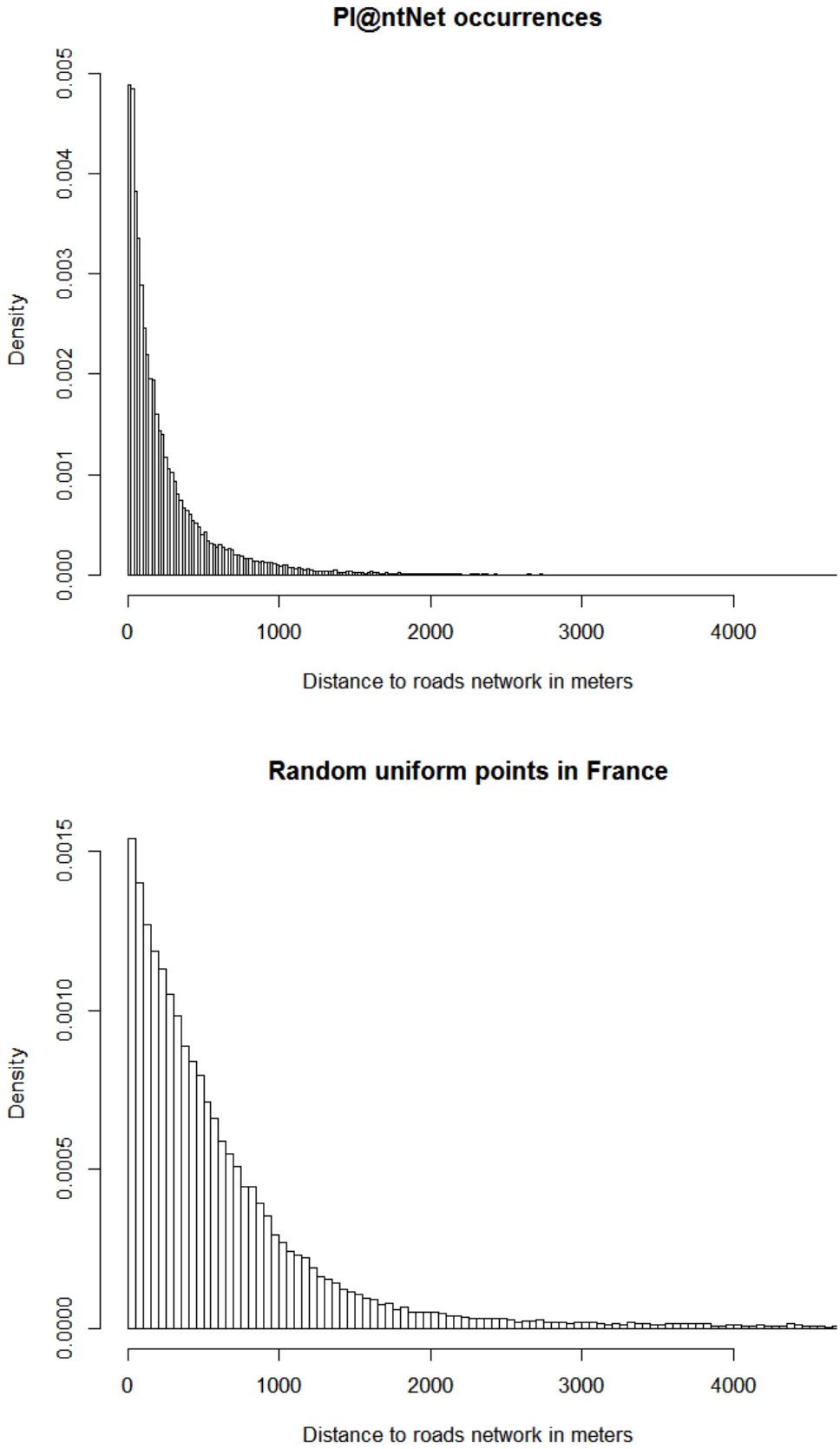


Figure 2: Histograms representing the distribution of points distance to the French roads network (*autoroutes*, *nationales* and *départementales*) for a random subset of 50,000 Pl@ntNet 2018 geolocated queries (Top) and 50,000 points uniformly drawn over the French territory (Bottom)

eco-physiologically based recommendations of environmental variables to consider for SDM. It also put in perspective these recommendations with the available environmental data for researchers and what is actually used in practice. Unfortunately, as noted by (Mod et al., 2016), no study has carried out an itemizing of available geographic environmental information and their specifications. The main specifications, important to consider when planning to use it for SDM, are grain size, extent, uncertainty. Their effects on SDM are reviewed in Moudrỳ and Šímová (2012). Geographic environmental variables were gathered during the PhD (especially suitable for modelling plant distributions and observation pressure) and compiled them in the same format of geographic raster covering France (Botella, 2019). This compilation of datasets is used for SDM applications in **Chapters 1, 3, 4 and 5**.

4.3 Species Distribution Models (SDM): Overview and state-of-the-art of presence-only methods

In this section, we explain the evolution of the theory of species distributions underpinning Species Distribution Models (SDM). This shall help the reader understand the context of SDM variables and models architecture choices and justify the terminology that we use. We define SDM and describe the different categories mainly in terms of the type of observation data they use. We then focus on SDM for presence-only data with the recent unification of those methods based on point processes models, and more particularly Poisson processes that are largely used in the thesis.

4.3.1 Niche theory

The study of species distributions is the fundamental goal of biogeography and thus has a long history (Wallace, 1860). The concept of ecological niche roots in the work of Grinnel in 1917 (Grinnell, 1917) who observed that species presence was restricted along environmental gradients and that these ranges of environments were distinctive enough across species to reflect different ecological properties. He thereby suggested that inclusion in those environmental ranges should reflect species conditions for survival. Besides, interactions with other species also participate to determine the capacity of a species to survive and reproduce in a given place through biotic interactions (competition, facilitation and trophic interactions) as noted by Elton (1927). Hutchinson (1957) proposed a formal definition of the fundamental niche of a species that is still a crucial concept in modern niche theory. He defined the fundamental niche as the hypervolume, in the multi-dimensional space defined by some ecological axes, where the population of the species can persist indefinitely if one excludes any negative biotic interactions with other species. Ecological axes may include environmental gradients unaffected by biological organisms (called *scenopoetic* variables), and resources variables or population variables from other species (called *bionomic* variables). This definition of the niche, that is a characteristic of the species must be distinguished from earlier uses (Grinnell (1917), Elton (1927)) that meant a physical place or a "recess" in the environment.
Hutchinson (1957) already differentiated the concepts of fundamental and realized niche, which restrict the fundamental niche to the conditions where the species can survive indefinitely while competing with other species. As the latter definition can't explain alone the actual species distribution, Pulliam (2000) proposed to include in the realized niche environments made available by stochastic phenomenons of colonization-extinctions, pulling together the original niche theory of Hutchinson (1957), meta-population theory of Hanski (1999) and source-sink theory Pulliam (1988). Indeed, the species may have been maintained outside of its fundamental niche by colonization (active exploration of animal for resources and plants dispersal mechanisms), while it may also be absent in its fundamental niche because of random demographic effects, or removal by a disturbance. As a consequence, the realized niche is not necessarily included in the fundamental niche. The logic of stochastic source-sink dynamical patterns has been interestingly pushed further by the "unified neutral theory of biodiversity and biogeography" proposed by Hubbell (2001). Hubbell studied mathematically the consequences of a surprising hypothesis that is central to its theory: All individuals within the same trophic level have the same probability of reproduction, death and dispersal, regardless of their species identity. In other words, neutral theory assumes, on the contrary of niche theory, that all species coexisting in a given environment, and having the same interactions with other organisms, have the same absolute fitness. Without immigration, the random demographic drift ultimately leads to local extinction or full dominance of a species, but entails no spatio-environmental pattern for the

species distribution. However, when considering immigration, one species becoming abundant in some local communities will favor it's domination in nearby communities through sustained colonization, eventually leading to a spatial pattern of abundance at a certain stage. This observed habitat of the species is then said to be contingent because, e.g. other species were equally likely to have this same distribution. Those radical assumptions underline stochastic phenomena that may lead to wrong conclusions with the analysis of species distributions, and that were neglected because of previous theoretical concepts. A last important extension of the realized niche in its modern definition is made by accounting more broadly for biotic interactions than simple pairwise competitions. Indeed, an environment may be made available to the species by e.g. a facilitation from another species (Brooker et al., 2008), the coexistence induced by intransitive competitive interactions networks (which contain loops) of at least three species (Laird and Schamp, 2006), or more complex interactions phenomenons including negative and positive interactions of at least three species. Even if important steps have been made to better define and show evidence of the mechanisms that separate the real distribution of a species from its fundamental niche, the realized niche itself has not been given a modern formal and consensual definition to my knowledge. Finally, the one that We will retain is the set of environments, inside a certain area with its biotic context and a moment in time, where the species actually lives for a reasonable lifetime.
A last important definition in the modern niche theory is the potential niche, a concept introduced by Jackson and Overpeck (2000). It may be defined as the set of environments where a species could survive for a reasonable lifetime if there was no dispersal constraints, i.e. if the species was at equilibrium in the geographic domain (Araújo and Pearson, 2005). The potential niche include the realized niche. It is the target of methods for predicting species invasions, or future distributions under global changes.
As a summary on the concepts of the niche theory, what we do observe in nature, the current distribution of a species, is shaped by a complex interplay between multiple factors. The fundamental niche of a species defines its elementary requirements for establishment without biotic interactions, disturbances, nor spatial dynamics. It is deviated by the interaction with other species, constrained by a spatio-temporal trajectory of dispersal and an history of stochastic demographic fluctuations, that jointly determine the realized niche. Furthermore, the potential niche is meant by the extension of the realized niche if there was no dispersal constraint, and is of crucial interest for predictive biogeography. See Figure 4 for a summary of definition introduced here.

4.3.2 Studying the response functions

Hutchinson (1957) already noted a limitation in the binary view of niche concept: "*It is supposed that all points in each fundamental niche imply equal probability of persistence of the species, all points outside each niche, zero probability of survival of the relevant species. Ordinarily there will however be an optimal part of the niche with markedly suboptimal conditions near the boundaries*". All factors that negatively affect the population of a species will have an even greater effect if the conditions are less optimal for its development. This already justifies the view of a continuously varying species fitness along environmental gradient (Whittaker, 1967). On top of that, the conjunction of spatial auto-correlation of environmental gradients and dynamic colonization processes tend to induce even more smoothness in the real abundance along an environmental gradient. This effect is known as the Moran theorem in population dynamics (Royama (2012) p.89) and it was analytically studied by Kendall et al. (2000). All together, it led to the view of a continuously varying abundance response to envi-

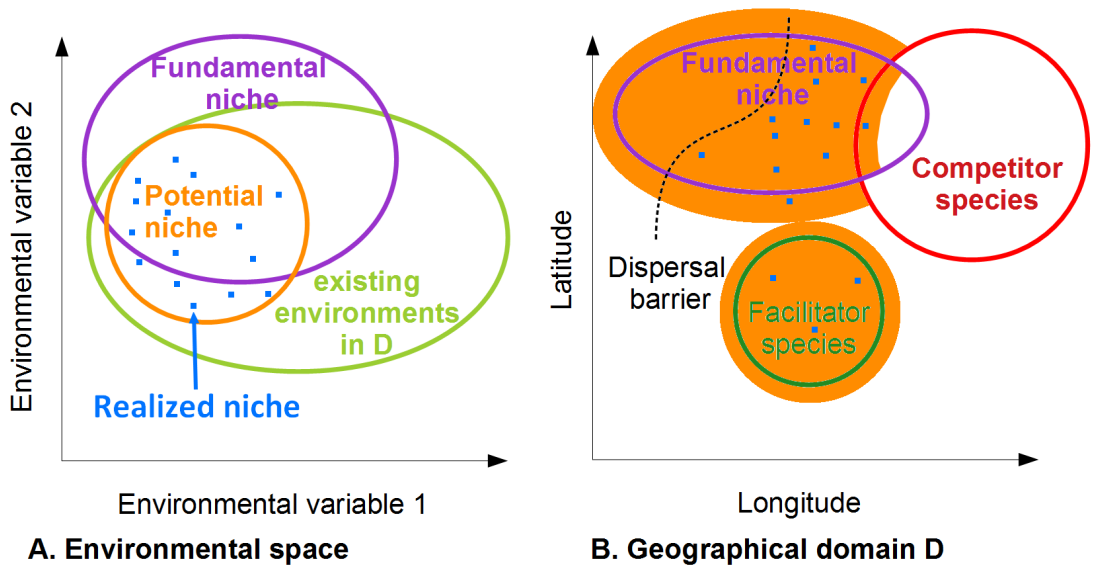


Figure 3: Conceptual illustration of fundamental, potential and realized niches of a species in the environmental and geographic spaces. A. The fundamental niche envelope (purple), potential niche envelope (orange), realized niche (blue squares) and the existing environmental subspace (pale green) in a geographic domain (called D), represented in the space defined by two environmental variables. B. The individuals of the species (blue squares), the projected envelope of its fundamental niche (purple), the presence envelope of a stronger competitor (red), the presence of a facilitator (dark green) and the area of the projected potential niche (orange) in the geographic space. Source-sink mechanisms maintain the presence of the species outside of its fundamental niche, even in the zone of its competitor.

ronmental gradient called the Response Function (RF) in plant ecology. The response function was notably illustrated by the abundance of Californian oaks species along the elevation gradient by Whittaker and Niering (1975). The study of the RF and the processes driving its shape became an important topic in plant ecology Whittaker (1967). In animal biology, the point of view of the Resource Selection Function (RSF) (Austin, 2002) is different: Animals are mobile and seek the most appropriate habitats for their survival and reproduction. They move out to less optimal habitat when the resource is limiting. This point of view led to similar function representation and estimation methods. Because of the fundamental niche borders of Grinnell (1917), the RF/RSF are expected to be tapered, i.e. their value decrease when going to extremities of the gradient values. They are typically assumed to be bell-shaped (Whittaker, 1967). Hypotheses of Gaussian RF for all species, equally spaced and of equal amplitude, with their width restricted by competition have been proposed (Gause (1936), Tilman (1982)). However, a competition with another species aside from the optimal environmental conditions is expected to induce a skewed RF (Austin and Smith, 1990), with slow decrease on the side limited by fitness, and a sharper one on the side where competitor appears. If the optimal environmental range itself is occupied by a stronger competitor, it should lead to a low observed response on the mode of the potential RF, inducing bimodality in the observed RF (Whittaker (1960), Whittaker (1967), Ellenberg and Mueller-Dombois (1974)). See Figure 4 for a summary of definition introduced here.

4.3.3 Species distribution models overview

The present section will provide an overview of the world of Species Distribution Models (SDM). SDM experienced an explosive growth of use in the scientific literature over recent years, and they are especially used by governmental and non-governmental organizations charged with biological resource assessment and conservation (Guisan et al., 2013). It has been facilitated by the development of digital data sets as described in the previous section. In a previous review, Franklin (2010) said "*a Species Distribution Model extrapolates species distribution data in space and time, usually based on a statistical model*". It is an efficient explanation of what SDM does and how it is built. SDMs can use different types of data on species and the environment, statistical models and inference algorithms. Franklin (2010) also mentioned that SDMs have their roots in the study of response functions (ecological gradient analysis, see Whittaker (1960) and Whittaker et al. (1973)), biogeography (Box, 1981), remote sensing and geographic information science.
Note the distinction between SDM and Ecological Niche Models (ENM, Peterson and Soberón (2012)). Both are mostly based on the same statistical methods introduced further, but ENM approximate the fundamental niche of a species in order to extrapolate robust predictions of the species potential distribution in different places or time, typically under environmental change (Soberón and Nakamura, 2009). To hope achieve a robust estimation, the modeler must rely on many hypothesis on the species ecology: he must use environmental variables that are strong determinant of the species environmental requirements (resource or condition gradients, see Guisan and Zimmermann (2000)), avoid dispersal constraints over the area, and generally assume the effects of biotic interactions to be local and average out at the scale of ENM predictions (Eltonian noise hypothesis, see Soberón and Nakamura (2009)). A more straightforward way to estimate species niches is to set up physiological experiments in controlled environmental conditions as recommended by Pulliam (2000). SDM have a different approach. It aims at providing the most accurate continuous maps of species actual distribution in the sampled area and period of time, or the empirical response over certain environmental variables. It needs no prior hypothesis on the species ecology, but doesn't

The quick evolution of the theory underpinning species distributions has induced a lot of ambiguity in its vocabulary and a lot of misuses were propagated in the literature. Here is a summary of the definitions of concepts We use[a].

- **Environmental variable**: A variable defined over the geographical space at a moment in time that quantitatively or qualitatively represent a measure of the abiotic environment. The abiotic environment might be either unaffected (e.g. annual temperature, precipitations, elevation) or affected by living organisms (e.g. soil Ph, land cover, soil organic matter).
- **Fundamental niche** of a species: Given a multi-dimensional coordinate system defined by some ecological axes, it is the hyper-volume (or set of points) of conditions where the population of a species could persist indefinitely if one excludes any competitive biotic interactions with other species. Ecological axes include environmental or biotic variables. This definition is taken from Hutchinson (1957).
- **Realized niche** of a species: Given a multi-dimensional coordinate system of ecological axes, it its the set of points, existing inside a geographic area with its biotic context at a moment in time, where the species actually lives for a reasonable lifetime.
- **Potential niche** of a species: Given a multi-dimensional coordinate system of ecological axes, it its the set of points, existing inside a geographic area with its biotic context at moment in time, where the species would survive for a reasonable lifetime if there was no dispersal constraints. It is always included in the realized niche with set equality when the species is at equilibrium in the area (see Araújo and Pearson (2005)).
- **Response function**: It is a response value linked to the species population (typically abundance, probability of presence, expected intensity or fitness) that is a function of an environmental or biotic variable. We talk of a response surface when the response is a function of several variables.
- **Habitat** : A type of abiotic environment and biotic context empirically characterized to generally contain a given species. It is especially extracted from correlative SDM. The use of this term has been recommended by Kearney (2006) to avoid confusions between elements of the fundamental niche. A species might live in many habitats which are all included in its realized niche.
- **Habitat suitability**: A function that approximates the true abundance of the species, depends on environmental variables and is based on empirical observations. We call it more precisely the "environmental density" when it is based only on environmental variables and estimated as a Poisson process intensity.

I also chose a set of notations for introducing SDM methods. We define $D \subset \mathbb{R}^2$ representing a geographic domain of study. We consider $d \in \mathbb{N}$ environmental variables (sometimes called gradients if continuous) $w = (w_1, ..., w_d)$, and $p \in \mathbb{N}$ environmental features $x := (x_1, \ldots, x_p)$, which are all real functions defined everywhere in D. The environmental features x are actually various functions of the original environmental variables w. They model together the shape of the response surface in the space of w through a parameterized linear combination as expressed in equation 1. For example, x might include w_i and w_i^2 in its components when w_i is continuous, whereas, if w_i is categorical, the categories indicators $(1_{w_i=e})_{e \in Im(w_i)}$ where $Im(w_i)$ is the set of all possible categories. In the context of most SDM, w are step functions over rectangular cells of geographic raster, but they may sometimes be defined explicitly as continuous functions. We name $z := (z_1, \ldots, z_N) \in D^N$ the set of geographical positions of a modelled species occurrences.

[a] Still, inconsistent uses may have been committed in our articles.

Figure 4: Definitions and notations

give any guaranty of the predictions accuracy in non-sampled environments, distant locations and times. SDM have also been called "*habitat suitability models*" (Hirzel et al. (2006), Ray and Burgman (2006)), and "*predictive habitat distribution models*" (Guisan and Zimmermann, 2000) or "*spatially explicit habitat suitability models*" (Rotenberry et al., 2006) when used for prediction in the geographical space. We introduce SDM methods by the angle of their response function class, because it is the core of any SDM, and it should give a clearer view of methods relations and distinctions. We also highlight the differences between methods dealing with presence-absence, site-occupancy data and presences only.

The theoretical concepts of the niche theory paved the way for the first SDM methods. The so-called envelope methods were based on this theory and tried to extract the niche of species from geolocated occurrences only. The first and most simple of envelope methods is BIOCLIM (Busby, 1991), which simply computes a binding multidimensional box of the occurrences in the space of continuous bioclimatic environmental variables. Each marginal response function to a w_i is an interval indicator function $1_{\{w_i \in [a_1^i, a_2^i]\}}$ and the response surface is simply the product of those marginals. A related method called HABITAT (Walker and Cocks, 1991) considers a convex hull. Other famous methods in the same spirit include DOMAIN (Carpenter et al., 1993) and MD (Farber and Kadmon, 2003). Envelope methods determine a subset of environments were the species occurs, thus these methods were the first enabling to predict if a non-prospected geographical area was suitable for the species, which was especially used for predicting ranges of biological invasions (Petitpierre et al. (2017), Barbet-Massin et al. (2018)) and distribution under global change (Midgley et al., 2003).

Around the 90's, there has been a switchover from methods that seek to estimate a niche, in the sense of an hypervolume, to methods that estimate the Response Function in the space of environmental variables.
Firstly, many SDM methods have been used to deal with **presence-absence data**, because presence-absence data are much more informative and less biased than presences-only data. These methods include: Generalized Linear Models (GLM, McCullagh (2019), Thuiller (2003)), Genetic Algorithm for Range Prediction (GARP, Stockwell (1999)), Generalized Additive Models (GAM, Hastie and Tibshirani (1986), Yee and Mitchell (1991)), Multivariate Adaptative Regression Splines (MARS, Friedman (1991), Ward (2007)), Artificial Neural Networks (ANN, Pearson et al. (2002)) and Boosted Regression Trees (BRT, Elith et al. (2008)).

I first give a simple and general form of response function model from which we can easily explain the peculiarities of each method. SDM assume that the relationship between the output variable, taken at a location i, named $y_i \in \mathbb{R}$ (abundance, probability of presence, occupancy probability) and the input environmental variables $w^k \in \mathbb{R}^d$ is determined through a Response Function whose general form is expressed in 1.

$$\begin{aligned} \mathbb{E}(y_i) &= g\left(\sum_{k=1}^{d} f_{\theta^{(k)}}(w_k^i)\right) \\ &= g\left(\sum_{j=1}^{p} \theta_j x_j^i(w^i)\right) \end{aligned} \tag{1}$$

The features x_j were defined in Figure 4 as functions of w. Equation 1 shows that the expected value of the response y_i is a given function g of a parameterized linear combination of the features x^i.
In the following, We synthesize the characteristics of each method with respect to the class of response function, how it is estimated, and how it handles model complexity. Known limits

of the methods are also mentioned.

GLM. In the context of presence-absence data, Generalized Linear Models are generally applied in the form of logistic regression and include a few simple features of the w_i such as quadratic w_i^2 and cubic w_i^3 transformations. Sometimes, product terms $w_i w_j$ are also included in the model. They are fitted with maximum likelihood method, classically implemented in statistical libraries through the Iteratively Weighted Least Squares algorithm (McCullagh (2019), chapter 2.5). This method is implemented in the base package of R. It is straightforward to implement and transparent, because the user chooses himself the features to integrate in the model. Often the Aikaike Information Criterion (Akaike, 1974) is used with such models for features selection, when there are not too many features. In latter case, the L1 penalty (Lasso) is often preferred (Tibshirani, 1996).

GAM. Generalized Additive Models (Hastie and Tibshirani, 1986) were proposed as a *non-parametric* alternative to GLMs by Yee and Mitchell (1991). They model the response surface with $x(w) = (f_1(w_1), ..., f_{m_1}(w_1), f_{m_1+1}(w_2), ...)$ using bases of smooth functions to model the response along each continuous environmental variable. The user chose the basis of functions, called smoothers, like Splines, especially the common choice of B-splines (De Boor, 1972), or LOESS (Locally Estimated Scatterplot Smoothing, Cleveland (1979)). Briefly, a B-spline basis is defined from a finite set of distinct variable values, called knots, and an order of smoothness. The knots define the number of functions in the basis and the union of their supports spanning the variable range, which defines the complexity of the potential response function fitted, while the order defines the degree of "smoothness" of the functions through the order of its defined derivatives (order 1 is piece-wise linear, 2-quadratic- is once differentiable, 3-cubic- is twice differentiable, etc). In most situations, GAM has the advantage to require less parameters for fitting complex response function compare to polynomial features in a GLM, and this method gives thus a more complex model without over-fitting.

MARS. Multivariate Adaptive Regression Splines Friedman (1991) is closely related to GAM but enjoys an automated procedure to define the knots and order of smoothness of the splines basis. It has been proposed later to deal with complex response when d is high and (multiple environmental variables) and one want to estimate terms in x that are functions of complex functions of a few variables, i.e. $x_k = f(w_i, w_j)$ or $x_k = f(w_i, w_j, w_m)$. MARS has an efficient fitting procedure and it is also more comfortable to use than GAM, because it has a statistically sound optimal selection procedure for the knots which allows to estimate more complex models without over-fitting when the sample size is large Wisz et al. (2008).

GARP. The Genetic Algorithm for Rule set Prediction was introduced in Stockwell (1999). It is a sophisticated machine learning algorithm that determines an optimal set of rules to predict species presence or absence. A rule is composed of a set of indicators, such as interval indicator $1_{\{w_i \in [a_1, a_2]\}}$ for continuous variables (similarly to BIOCLIM), or atomic indicators $1_{\{w_i = a\}}$ for categorical variables, and a prediction to do (presence or absence) if the product of rule's indicators equals 1. The method optimizes the set of rules through a genetic algorithm that explores the space of all rules using cross-over (exchange of elementary indicators between rules) and mutations (deviation of the interval or value of an elementary rule). Rules predictive performances are evaluated at each step on test data which determine a selection inside the rule population.

ANN. Artificial Neural Networks are a class of models where the "neurons" $x_j(w)$ are parametric non-linear functions of the whole vector w. Neural network models architecture is described more extensively in Chapter **4** but, for example, the simplest ANN with one hidden layer will have neurons $x_j = h(w^T\beta_j)$, where h is a non-linear function. The model is fitted by maximizing the likelihood respectively to the parameters vectors $(\beta_1, ..., \beta_p)$ and respectively to θ. By increasing infinitely the number of neurons p , any function of $w_1, ..., w_d$ can be approximated, which makes this method particularly attractive. However, optimizing ANNs is difficult, because the likelihood has many local optima, and optimizing it too much leads to over-fitting. Optimizing techniques for those models have considerably developed in the last ten years, but were much less efficient at the time of first uses for SDM. Indeed, ANN were early used for SDM, as in Lek and Guégan (1999) or Pearson et al. (2002). Even though, they have shown good prediction performances comparatively to other methods.

BRT. Boosted Regression Trees have been proposed to model species distribution by Elith et al. (2008). The prediction of BRT is based on multiple **regression trees** (decision trees where each leaf predict an output value) predictions. Thus, the combination of all trees contributions may be summarized as a response function that is constant over hyper-rectangles (e.g. $x_j(w) = 1_{\{\cap_k w_k \in]a_1^k, a_2^k[\}}$ where $a_1^k \in \mathbb{R} \cup -\infty$ and $a_2^k \in \mathbb{R} \cup +\infty$) jointly making a partition of the environmental space. The algorithm optimizes the partition (through the number of trees and their own partition) and the function values (θ) to minimize a predictive error. The originality of this method compared to Random Forests (Breiman, 2001) is that trees are added iteratively. At each step, a new regression tree is built with a maximum depth (modeller defined) to minimize the likelihood which has been re-weighted to inflate the importance of data of larger error from previous trees prediction. An important point is that a depth of one doesn't allow variables interaction effects, i.e. non-linear functions of several components of x, a depth of two allow first order interactions effects (between 2 variables), etc. There is a balance to find with the number of trees (more trees imply a more complex response and more over-fit) and the learning rate. This method was originally introduced in Friedman (2001).

These methods are easily adapted to **abundance data**, i.e. counts of individuals in the place, by simply changing the link function g (generally taking the exponential) and the probability distribution of y (generally choosing the Poisson distribution). Abundance data is theoretically more informative than presence-absence data because they provide an additional quantitative information of the suitability across presence sites, which enables to fit a more complex response surface with the same number of samples. However, as presence-absence and abundance data need a constraining sampling protocol, they are costly to collect.

Most SDM methods only account for the environment to predict the response of a species, but SDM using spatial coordinates as input or accounting explicitly for spatial auto-correlation may also have great predictive power (Bahn and McGill, 2007) because the distribution of species show spatial patterns that are often captured through the spatial auto-correlation of environmental variables. However, more recent approaches tried to jointly account for interactions between species, or effects of non-observed environmental variables, in statistical models using the framework of latent variables models (Ovaskainen et al. (2010), Kissling et al. (2012), Pollock et al. (2014)). Several implementations of those Generalized Linear Latent Variables Models (GLLVMs) exist in R. For examples, the package `hmsc` that is based on an

EM algorithm inference (Ovaskainen and Soininen, 2011), while the package `gllvm` [11] is based on variational approximation (Hui et al., 2017). The package `jSDM` [12] implements GLLVMs in a hierarchical bayesian framework through Gibbs sampling and C++ based computations. Those methods may theoretically correct for patterns in the species distribution due to biotic interactions independently of the environment. They thus provide a less biased view of the fundamental niche of the species. They also provide a model based statistical method to investigate directly the species co-variations.
Let's now go to models for the less standardized type of site-occupancy data with imperfect detection, introduced in **section 5.2**. Fundamentally, SDM methods for standardized data (presence-absence, abundance) and site-occupancy data under imperfect detection rely on similar models of the species response function. The key difference is the way to model the reported data conditionally to actual species abundance, or presence. SDM for occupancy-detection data have been introduced by MacKenzie et al. (2002). The elementary data is a site associated with the history of reports or non-detections of the species presence for repeated visits during a period of time where the species is assumed to have been constantly present or absent. The data can be thus considered as a form of presence-absence data degraded by imperfect detection. The occupancy-detection models take into account that the species might not have been detected during a visit in the site where it is present. To define the probabilistic model underlying occupancy-detection methods, we first consider the n sites previously defined, with environmental features x^i for site i. Like before, the probability of presence of the species at i is represented by the response function $g(\theta^T x^i)$. We consider that i is visited T times and we note the detection history of the species by introducing the index of the visit $y_i = y_{i1}, ..., y_{iT} \in \{0,1\}^T$, the likelihood of all those reports is given in equation 2.

$$\begin{array}{rcll} p_\theta(y_i) & = & g(\theta^T x^i) \prod_{t=1}^{T} (d_{it})^{y_{it}} (1-d_{it})^{1-y_{it}} & \text{If } \sum_t y_{it} > 0 \\ & = & 1 - g(\theta^T x^i) & \text{Otherwise} \end{array} \tag{2}$$

Where d_{it} is the probability of detection of the species at i during visit t, which is generally represented as a function of environmental or meteorological variables that affect detection, and observer variables (expertise, interests etc). Then, the global likelihood is simply the product over all sites. This model enables to consistently estimate, as for the presence-absence case, the response function parameters θ when the number of sites increases, and there is repeated visits for each site with a reasonable probability of detection. As the likelihood has a non-standard expression and may not be fitted with standard R libraries like `glm`, `gam` or `gbm`, it may require more sophisticated libraries to fit complex response curve classes on occupancy data with imperfect detection. Rather than estimating the probability of presence, Royle and Nichols (2003) proposed an extension of the model framework to estimate the abundance through occupancy-detection data. Methods have been also developed specifically for other slightly different types of semi-standardized data, e.g. distance sampling data Royle et al. (2004), repeated counts Royle (2004b), or multinomial counts Royle (2004a). The R package `unmarked` [13] was specifically developed to carry out estimation based on those different types of likelihoods for semi-standardized data Fiske et al. (2011). It can easily implement **GLM** like response function models, or **GAM** like models through a user customized design matrix. Other approaches applied to this type of data demand further developments. Similar models may be fitted in a Bayesian framework through the R package `hSDM` [14], with the appreciable

[11] `https://CRAN.R-project.org/package=gllvm`
[12] `https://CRAN.R-project.org/package=jSDM`
[13] `https://CRAN.R-project.org/package=unmarked`
[14] `https://CRAN.R-project.org/package=hSDM`

optional addition of spatially auto-correlated random effects in the models. However, considering that the probability of detecting the species at least once, conditionally to its presence, over all visits $(1 - \prod_t (1 - d_{it}))$ becomes very small. Then, the absence of a species in a site where it has not been detected becomes very uncertain, and thus occupancy-detection data becomes very similar, in term of statistical information, as presence-only data.

This brings us now to the methods for presence-only (PO) data. As the weakest but most abundant and accessible type of data about biodiversity, PO data has been the ground for the development of many SDM methods. We already gave brief overview of the envelope methods (BIOCLIM, etc), the first PO SDM. Those ad hoc methods didn't account for the spatial areas associated to each environmental elementary range. The pitfall of it lies in the imbalance of environments representation in space. We illustrate it with a categorical environmental variable example, but the principle is the same for continuous ones: Consider that a species has no preferences with respect to an environmental binary variable w_i and that $w_i = a_1$ over 90% of the area and $w_i = a_2$ elsewhere. Then, under uniform sampling the species will be reported 9 times more, in expectation, in a_1 than in a_2. Thus, an envelope/profile method applied to a few number of occurrence will estimate that environment a_2 is marginal or even not suitable for the species. PO SDM methods based on pseudo-absences background points were introduced to account for the problem of environmental imbalance. Two types of responses appeared to solve this problem. The first one is to simply apply the discriminative presence-absence statistical methods (GLM, GAM, BRT, etc) by replacing the true absences (that we don't have) by so called pseudo-absences points most often uniformly in space. For example, such methods were applied in the extensive PO SDM methods comparison of Elith et al. (2006). However, as it will explained in section **5.3.5**, this lead to shrunk estimates of the environmental variables effects. On the other hand, environmental density methods essentially estimate the density of occurrences per unit area as a function of the environment. This type of method correct for environmental imbalance by selecting background points uniformly in the study area, which will be contrasted with occurrence points. The first major method of this type is the Ecological Niche Factor Analysis (ENFA, Hirzel et al. (2002)), which estimates the environmental density as a Gaussian in the space of (continuous) environmental variables. In the following subsections, we explain in more detail the MAXENT method, which may also fit gaussian environmental densities but also more complex response functions, and the more general methodology of spatial point processes.

4.3.4 MAXENT

This SDM method is based solely on presences. It is explained quite extensively because it has strong links with the Point processes methodology described, it is used a lot for predicting species distributions in many contexts, and it is at the basis of an important method of sampling bias correction. Also, this method has been used or studied in **Chapters 1, 2, 4 and 5**. MAXENT predictive performance was early evaluated comparatively to other PO-SDM methods of its time on a wide range of test presence-absence data from several taxonomic groups Elith et al. (2006) and demonstrated to have comparatively better (GARP, GAM, GLM, DOMAIN, BIOCLIM), similar (GDM, MARS) or slightly lower (BRT) predictive performances than other methods. Its ease of use and ability to robustly carry out inference on few occurrences Phillips and Dudík (2008) also helped to popularize it. A bibliographic search on the Web of Science database indicates that 22% of articles published between 2008 and june 2019 that were retrieved with topic TS=("species distribution" AND "occurrence") also matched the intersection with TS="MAXENT". It indicates how widely the method is used

for modelling species distributions from occurrence data in recent years.

MAXENT analyses the relation between a collection of geolocated species occurrences $z_{\text{occ}} = (z_1, ..., z_n) \in D$ and environmental features x by first splitting D into a user-defined number n_g of cells (sites) over a regular grid which centers are noted $(g_1, ..., g_{n_g}) \in D^{n_g}$. Then, a binary vector $\delta(z_{\text{occ}}) = (\delta_1, ..., \delta_{n_g})$ is computed with $\delta_i = 1/n_g$ if there is at least an occurrence in site i, or $\delta_i = 0$ otherwise. This vector is called the empirical distribution of the species over sites. Note that the centers g_i, as well as the empirical distribution $\delta(z_{\text{occ}})$, implicitly depends on the defined number of sites n_g. The probability distribution estimated by MAXENT represents for each site the probability that the species would be present there, if it were to be present on only one site. This probability distribution is noted π^{MAX} in equation 3. The probability in each site i is modelled as a function of the environmental features measured on the site $x(g_i)$. In words, equation 3 states that the MAXENT estimate, proposed in Phillips et al. (2004), maximises the entropy of the probability distribution under the constraints that each empirical average of the environmental feature is closer to its expectation under the probability distribution than a given positive regularization constant.

$$\begin{array}{rl} \text{Note} & \Pi = \{\pi = (\pi_1, ..., \pi_{n_g}), \forall i \in [|1, n_g|], \pi_i \geq 0, \sum_{i=1}^{n_g} \pi_i = 1\} \\ \text{and} & \forall k \in [|1, p|], \gamma_k > 0,\ I_k(\pi, \gamma_k) = 1_{\left\{ |\sum_{i=1}^{n_g} x_k(g_i)(\delta_i - \pi_i)| \leq \gamma_k \right\}} \\ \text{Then} & \pi^{MAX} = \underset{\pi \in \Pi, \sum_{k=1}^{p} I_k(\pi, \gamma_k) = p}{\text{argmax}} \quad - \sum_{i=1}^{n_g} \pi_i log(\pi_i) \end{array} \tag{3}$$

The condition meant by the indicator $I_k(\pi)$ is softer than the condition $\sum_{i=1}^{n_g} x_k(g_i)\delta_i = \sum_{i=1}^{n_g} x_k(g_i)\pi_i$. Indeed, we allow the estimated average feature over sites not to be exactly equal to the empirical one, because we know that it is an approximation of the true expectation of the feature. Thus, defining a higher γ_k means a lower confidence in the empirical mean for environmental feature k. It turns out that the solution of equation 3 is of the form $\pi_i^{MAX} = \exp(\beta^T x(g_i)) / \sum_{j=1}^{n_g} \exp(\beta^T x(g_j))$ with $\beta \in \mathbb{R}^p$, which is called a Gibbs distribution. It must be noted that maximum entropy formulation of the problem in equation 3 has an equivalent maximum likelihood formulation which is to minimize $D_{KL}(\exp(\beta^T x)||p) + \sum_{k=1}^{p} \gamma_k |\beta_k|$, i.e. the Kullback-Leibler divergence from the empirical distribution to the latter Gibbs distribution penalized with a Lasso penalty term. The maximum likelihood method regularized by a Lasso penalty has long been known for having good variable selection properties when there are many variables while the data are reduced Tibshirani (1996).

Initially, MAXENT required the modeler to specify the regularization hyper-parameters γ_k. Later, some noticeable new functionalities were added to MAXENT software Phillips and Dudík (2008), including (i) the addition of a new feature class (transformation of the original environmental variables) and (ii) a optimized set of default regularization hyper-parameters γ_k. Firstly, We describe hereafter all the features classes used in MAXENT. Considering some originals environmental variables $w_j \in \mathbb{R}^D$, MAXENT will build a vector x of environmental features based on the following terms:

- linear: w_j.
- quadratic: w_j^2.

- product: $w_i w_j$ with $i \neq j$. If the modeler restrain MAXENT features to linear, quadratic and product, they respectively constrain the estimated distribution to be close to the empirical means, variances and covariances of original environmental variables over occurrences Phillips et al. (2006).

- category indicator: $\sum_{j=1}^{m} 1_{w_1=c_j}$, if for example w_1 is a categorical variable with categories $c_1, ..., c_m$.

- threshold: $1_{w_j>s}$ with many values of $s \in [min(w_j(g_1), ..., w_j(g_{n_g})), max(w_j(g_1), ..., w_j(g_{n_g}))]$.

- forward and backward hinges: $1_{w_j>s}(w_j \breve{} s)/(max(w_j) - s)$ being the forward form and $1_{w_j>s}(w_j \breve{} s)/(max(w_j) - s)$ the backward, with many values of $s \in [min(w_j(g_1), ..., w_j(g_{n_g})), max(w_j(g_1), ..., w_j(g_{n_g}))]$. Without going into details the collections of s for threshold and hinge features depend on the implementation which differ from the initial Java software Phillips and Dudík (2008) to the recent R package `maxnet` Phillips et al. (2017).

Through the fitted features parameters (the β_k), the modeller can visualize the species response function along each original environmental variable w_j. We note $\tilde{x}_j$ the function taking as input value of the original environmental variable w_j and returning the vector of environmental features computed from this value, and $\tilde{\beta}_j$ the parameters associated with those features. Then, ignoring the product terms, the species MAXENT response function along the environmental gradient w_j is proportional to $w \to \exp(\tilde{\beta}_j^T \tilde{x}_j(w))$.

Secondly, according to the website [15] and the statements of Phillips et al. (2017), both the java software Maxent and the R package maxnet currently have a same default regularization scheme. The procedure attributes predefined penalization hyper-parameters that are equals per feature class. This hyper-parameters values are described in Phillips and Dudík (2008). They were determined by a cross-validation procedure evaluating the predictive accuracy of MAXENT tuned with different regularization hyper-parameters per feature class over 226 species covering different taxonomical groups and regions of the world.

To summarize, MAXENT is an SDM method fitted on presence-only data aiming at making reliable prediction of habitat suitability and showed actual better performances than most other PO-SDM of its time Elith et al. (2006). The key strengths of MAXENT is that it enjoys good estimation properties when the relevant quantitative environmental descriptors for defining the model are unknown, because of the multiple features it allows (i). Also, it has good estimation properties even when there are few occurrences, because of its model selection procedure (ii). MAXENT is closely related to point processes recently popularized for the purpose of SDM, which are introduced in the next paragraph.

4.3.5 The unifying framework of point processes models

Point processes are defined as a random collection of points over a given ensemble. They are based on probabilistic models ruling the number and location of the random points. From now on, We always consider the ensemble where points appear to be a subset of $\mathbb{R}^2$ that we

[15] `http://plantecology.syr.edu/fridley/bio793/maxent.html`

call D. Formally, we note Z a random collection of points on D, $\mathcal{B}(D)$ the Borellian tribe over D, and $\#|Z \cap B| \in \mathbb{N}$ the finite cardinal of $Z \cap B$, i.e. the number of points of Z in the sub-space $B \in \mathcal{B}(D)$. Z is a point process if it is locally finite, i.e. there is almost surely no open subspace with infinite number of points, and $\forall B \in \mathcal{B}(D)$, $\#|Z \cap B|$ is a random variable. Because the point process is a collection of points, it makes it quite different in nature from the usual random variables that are dealt with in standard statistical models, may they be integers or real numbers.

The simplest point process is called the Poisson process which is defined by (i) the number of points follows a Poisson distribution, (ii) the points locations are independent and identically distributed. More formally, the Poisson process is characterized by its intensity measure, or alternatively its intensity function, a positive function over $\mathcal{B}(D)$ such that:

Definiton: in-homogeneous Poisson process Let Z be a Poisson process, noted $Z \sim IPP(\lambda)$, where $\lambda \in \mathbb{R}^D$ is the positive intensity function. Then, $\forall B \in \mathcal{B}(D)$, $\#|Z \cap B| \sim P(\Lambda(B)) = P(\int_B \lambda(z)\mu(dz))$, where Λ is the intensity measure of the process and μ the Lebesgue measure on $\mathbb{R}^2$.

In the context of presence-only SDM, the response function being modeled is the intensity of Poisson process. It is homogeneous to the expected number of occurrences per area unit at every point, and it is interpretable as the species expected abundance given the data under exhaustive sampling. Point processes thus explicitely model the number and locations of occurrences in a continuous space under the hypothesis of points independency. The intensity function doesn't depend on the size of the area where points have been sampled.

A first set of remarkable properties of Poisson processes is its mathematical links with other SDM methods. It is first, closely related with MAXENT, because the latter has been shown to be equivalent to a standard Poisson regression Renner and Warton (2013). Briefly, recalling that MAXENT defines sites $g_1, ..., g_{n_g}$ over a regular geographic grid where it summarizes the occurrences of each cell as reported presence or no occurrence, then fitting a L1-penalized Poisson regression over those sites with pseudo-counts 0 or $1/n_g$ is exactly equivalent to fit MAXENT. Then, as n_g increases, the size of sites tend to 0, each occurrence becomes alone in its site (NB: if no point is duplicated, which should be avoided anyway to fit a Poisson process model) and most sites become empty. Asymptotically, the log-linear Poisson regression parameter vector tend to the corresponding log-linear Poisson process log-intensity parameters expect the intercept Renner and Warton (2013). It means that the intensity of the Poisson process is proportional to the Maxent distribution when the grid resolution increases. Thus, the analysis over discrete sites rather than in the continuous space is the key distinction between MAXENT and the associated Poisson process. Besides, the β parameter of our Poisson process intensity have been shown to be the asymptotic limit of the presence-background logistic regression related slope parameters (those that multiply the variables in the linear predictor, not the intercept) as the number of uniformly sampled background points tend to infinity Warton et al. (2010). Indeed, the shrinkage bias on those slope parameters noted by Ward (2007) (section 5.1) when using a finite background sample disappears asymptotically. A more recent result enabled to use in practice the logistic regression as an approximation of the Poisson process. It was brought by Fithian and Hastie (2013) which showed that, under a particular weighting scheme of a finite set of background points in the likelihood, the logistic estimates approximate well the Poisson process parameters.

The fact that presence-only data provide no information about the absence of a species

where no occurrence was recorded has been a trouble for defining a proper SDM methodology for this data for many years Warton et al. (2010). First approaches used to generate pseudo-absences and use them as true absences in PA methods. Ad hoc procedures have been proposed for selection pseudo-absences in this context (Pearce and Boyce (2006), Zarnetske et al. (2007)), but without theoretical justification. No clear methodological recommendations emerged on how to select the pseudo-absence for using PA models. Indeed, different ways of selecting pseudo absences may give different estimates. Plus, it was shown in Ward (2007) (Chapter I) that selecting a finite set of uniformly distributed pseudo-absences induce slope parameters shrinkage bias in the model estimates. Ward et al. (2009) proposed to explicitly model the true state of a pseudo-absence, as presence or true absence, in the logistic regression setting with a binary latent variable. However, it requires a prior knowledge of the global proportion of the "false" pseudo-absences which is seldom known. Warton et al. (2010) proposed instead to model the distribution of presences through Poisson processes use quadrature points, also called background points in reference to MAXENT. To understand what exactly are background points in the context of Poisson processes, we must have a look at the exact log likelihood of a realized Poisson process $Z = (z_1, \ldots, z_N)$, which is written in equation 4.

$$\begin{aligned} & p(z_1, ..., z_N|\theta) &=& \frac{(\int_D \lambda(z)dz)^N}{!N} e^{-\int_D \lambda(z)dz} \prod_{k=1}^{N} \frac{\lambda(z_k)}{\int_D \lambda(z)dz} \\ \Leftrightarrow \quad & p(z_1, ..., z_N|\theta) &\alpha& \exp\left(-\int_D \lambda(z)dz\right) \prod_{k=1}^{N} \lambda(z_k) \\ \Leftrightarrow \quad & log(p(z_1, ..., z_N|\theta)) &\alpha& \sum_{k=1}^{N} log(\lambda(z_k)) - \int_D \lambda(z)dz \end{aligned} \tag{4}$$

Where λ is also implicitly a function of θ. In general, the integral term of the log-likelihood can't be computed exactly or analytically. Even if so, the exact numerical computation would be very costly when we deal with multiple high resolution rasters of environmental variables. We rather use a numerical approximation. In a nutshell, the integral is replaced by a weighted sum of λ computed at some quadrature/background points. They may be drawn uniformly to provide an unbiased estimation of the integral following the Monte Carlo method of integral approximation. This likelihood may then be fitted with generalized linear models libraries using a second controlled approximation proposed by Berman and Turner (1992). This is the method that we use to fit Poisson processes in chapters 2 and 3. More details can be found in the convivial papers Warton et al. (2010) and Renner et al. (2015), or in Appendix 2.3 of the present manuscript. Finally, background points aim at providing a finite representation of the continuous space D through a finite collection of points over which is approximated the expected intensity. Maximizing likelihood then means maximizing a contrast between the intensity of the occurrences and the average intensity over D. Poisson process give thus a clear role to background points and simple procedures have been shown to control the number of points required for a given level of approximation accuracy of the log likelihood Renner et al. (2015).

In my experience, the justification of the log-linear model for Poisson process, while their might be many other non-negative functions to fill this role technically, is often asked and its justification seem to be spread across the literature. We provide hereafter noticeable mathematical modelling assumptions justifying its use. The logarithm is the natural link function of the Poisson regression in generalized linear models McCullagh (2019). Models for counts based on independence lead naturally to a multiplicative effect of each input variable variation on the expected count, which is expressed by the logarithm link as noted in 2.2.3 of the latter

book. Complementarily, the information theoretic result noted by Phillips et al. (2004) is interesting: The log-linear model appears as the maximum entropy solution when setting the constraint of average and expected features equality (see equation 3). It intuitively means that we want the model to fit a rather uniform intensity unless the information contained in the data says differently. As we have seen that the Poisson process is a punctual limit of the Poisson regression model, it is relevant to take the log-linear model for the intensity model in order to preserve those properties.

Poisson processes have other useful properties for the purpose of modelling species distribution that are extracted from Daley and Vere-Jones (2007) and Chiu et al. (2013).

Property 1: Conditional likelihood Let Z be a Poisson process defined over D with $Z \sim IPP(\lambda)$ and $B \in \mathcal{B}$. Then, conditionally to $\#|Z \cap B| = n$, the n points are independently distributed over B with the probability density function $\lambda / \int_B \lambda(z)\mu(dz)$.
This property is useful because it provides an alternative formulation of the Poisson process likelihood that only implies the intensity function proportional density over the domain. We use it for deriving the properties in chapter 2.

Property 2: Superposition Let Z_1 and Z_2 be two Poisson processes defined over D of respective intensity functions λ_1 and λ_2. Then, the superposition of their points collection $Z_1 \cup Z_2$ is also a Poisson process of intensity function $\lambda_1 + \lambda_2$.

Property 3: Thinning Let Z be a Poisson process defined over D of intensity function λ and $s : D \to [0, 1]$ be a measurable function. We build the thinned process such that any point $z \in D$ of Z is kept with probability $s(z)$. Then, this thinned process is also a Poisson process of intensity function $s\lambda$. This property is also known as the Prekopa's theorem.
This property is especially used to model the partial and spatially heterogeneous reporting of species individuals in chapters 2 and 3.

When dealing with several types of occurrences, such as multiple species, a marked point process, where the mark of a point is a category, is the natural object to consider.

Definition: Marked point process. Let $K \subset \mathbb{N}$ be a finite set. Then a marked point process (Z,M) is a point process defined over DxK such that Z is a point process over D and M a collection of elements of K with $\#|Z| = \#|M|$. In the chapter 3, we introduce a marked Poisson process to model the joint distribution of multiple species along with a common observation process.

Some point process diagnostic tools may be useful for analyzing if the model is well specified or detecting structured spatial errors patterns. Poisson processes assume that points are independent, but this hypothesis might be wrong in many situations of ecology because of the spatial dependence due to species dispersal process or the observation process. We must insure, once a Poisson process model is fitted on occurrences, that the hypothesis of points independence given the intensity is reasonable. A procedure for checking a spatial interaction between points has been proposed with the in-homogeneous K-function (Baddeley et al., 2000).

This function is a generalization, applicable to a fitted in-homogeneous Poisson process, of the original K-function Ripley (1977) which computes the ratio of points in a neighborhood of a given distance compared to the process intensity. Besides, there may be spatial patterns in the error of the model fitted intensity due *e.g.* to missing variables. The error may be visualized through the smoothed Pearson residuals Baddeley et al. (2005a). Leverage and influence help to analyse which places or occurrences that contribute the most to the fitted intensity. Pearson residuals, leverage and influence may be visualized with the `spastat` package (Baddeley et al., 2005b).

Spatial point processes have many positive outcomes for modelling different types of data. Apart from punctual occurrences, it is possible to jointly model with a hybrid likelihood: counts of individual over a given area Giraud et al. (2016), presence-absence data Fithian et al. (2015) or presence-absence with imperfect detection / occupancy-detection Koshkina et al. (2017). In those joint models, the species intensity component is shared between all data types. With such setting, the standardized data help to correct for sampling bias in the occurrences data, while the mass of occurrences reduce the variance of the estimates, especially for rare species, and provide broader spatial coverage (Giraud et al. (2016), Dorazio (2014)).

As a last point, we state that through point process models, the modeler is not tied to the constraints of the single Maxent software implementation and he may use the suite of modelling tools of statistical softwares: many smoother types (splines functions were pointed out as a future development for Maxent Phillips and Dudík (2008)), multiple penalties (L2, elastic net), the possibility of weighting the samples to account for various types of data uncertainty, etc. Point processes models have, so to say, a big universe in the probability and statistics theory, but there also exist multiple tools to simulate and infer those models. There are several softwares to fit point processes in R (spatstat, ppmlasso). Poisson processes may also be fitted through generalized linear models libraries like glm/gam/glmm through the Infinitely Weighted Logistic regression Fithian and Hastie (2013) or Poisson regression approximation Berman and Turner (1992) results. One may fit more complex point processes models using more sophisticated optimization libraries, enjoying automatic differentiation and, like ADMB Fournier et al. (2012) or TMB Kristensen et al. (2015) or deep learning specialized librairies like Mxnet Chen et al. (2015), which is used for multiple Poisson regression in **chapter 4**. Bayesian point processes with log-linear model including random effects with complex covariances, or Cox processes may be fitted with the R-INLA library Lindgren et al. (2015).

Points want to break free, don't put them in cells. In summary, point processes models have been popularized for SDM applications with presence only data especially because they emerged as a generalization of the Maxent method. Through their intensity parameter and the stochastic process around it, they enable to model the continuous location of points with parameters that doesn't depend on scale through subjective partitioning of space or loss of information due to removal of occurrences. Even though, we note that this continuous framework often induce a computational burden for inference, but many controlled approximations may achieved to avoid that burden. They identify a clear role of background points, that characterize the sampled environments which is contrasted with occurrence patterns, potentially unlocking optimal selection procedures, but this is yet an under-investigated area to my knowledge. Finally, we gave a brief overview of the many available tools for inference, assumptions checking and analysis of spatial dependence in point processes.

4.4 Boundaries and pitfalls of Species Distribution Models based on massive opportunistic occurrences

This section exposes two general issues addressed in the thesis regarding species distribution modelling based on massive occurrences collected without sampling protocol. Firstly, in section **5.4.1**, we introduce the problem of sampling biases, by explaining how they distort the modelled view of species distribution, and focus on spatial sampling biases, which is the subject of chapters **2 and 3**. Secondly, in section **5.4.2**, we come to limitations concerning model complexity. We give context on the uses of deep learning techniques for SDM and gather recent theoretical, methodological and applications advances of deep learning that may be leveraged to improve fine grain SDM predictions. These concepts are leveraged in Chapter **4**.

4.4.1 Biases due to sampling heterogeneity in presence only data

We have seen that presence-only data is a massive, renewable and wealthy source of data for biodiversity monitoring. However, spatial and temporal sampling biases have been recognized decades ago to be present in many volunteer surveys as explained in Yoccoz et al. (2001). Models based on PO data are probably the most sensitive to this biases as these data are influenced by many types of heterogeneity in the sampling process. In this section, we first explain how sampling intensity bias PO SDM because of variable observation effort across space, time, species or observers, distorting the observed distribution. Then, we focus on existing methods to correct for spatial sampling biases.

Defining opportunistic occurrences and observation effort. The sampling of occurrences data is said to be opportunistic when there is no rule guiding consistently the sampling protocol of the group of observers contributing to a given dataset of occurrences. Equivalently, a sampling rule might have been followed by contributors, but it is not known by the modeller. A counter-example would be, e.g. a dataset where all observers had for mission to prospect homogeneously, i.e. with constant detection probability, a given set of sites is not opportunistic. The term of opportunistic data has been used to qualify such type of presence-only data for the first time by Kery et al. (2010), who applied site-occupancy SDM to opportunistic occurrences coming from citizen sciences datasets to correct mainly for temporal variation in detection probability. They aggregated the reports per season over sites to build the detection histories. In this case, the detection probability modelled in site-occupancy models is dependent on the observation effort. Indeed, in the context of opportunistic occurrences, the probability of collecting the occurrence of a species, conditional on its presence, does not only depend on the observer's ability to detect it while observing. It also depends on the observation intensity towards this site over all observers. This intensity that we call the observation effort is widely variable from an area to another, and from a time to another, in opportunistic datasets, and it will be crucial to study its variation.
To be as general as possible, we define the observation effort as the density of the expected number of views/observations (potentially leading to occurrences) per instant and spatial point. The observation effort is thus a function defined over the product of time and geographic space. Formally, we note $V(u)$ the random variable equals to the number of views of point $u \in U \subset \mathbb{R}^3$ in the geographico-temporal space. we define the observation effort $v(u)$ as $\mathbb{E}(V(u))$. Technically, V is theoretically allowed to be any number because it results of the sum of multiple observers attention, thus several observers may pay attention to a same spatial point at a same instant. If an observer pays attention to a point where a species individual is located, the observer might detect the species and then report it as an occurrence, but the

observation effort is independent of those posterior steps. The definition of the observation effort is used in **chapters 2 and 3** but, as we study its spatial effect, it is integrated over a time period during which the occurrences are collected, and thus becomes a function in the geographical space only. we may call it alternatively sampling effort, when it is assimilated to the product of observation, detection and reporting probabilities (the latter are assumed constant over space sometimes), or when We consider the product of the three instead of the observation effort by itself. The here defined sampling effort is closely related to the concept of recorder's activity of Isaac et al. (2014). We also note that the sampling effort term may be used in the literature with slightly different definitions in other contexts, but We chose this definition to be consistent with those chosen (more or less implicitly, and sometimes with extra assumptions) in the literature of site-occupancy models (e.g.MacKenzie et al. (2002), Kery et al. (2010)) and presence only SDM literature (e.g. Dudík et al. (2006), Phillips et al. (2009), Warton et al. (2013), Fithian et al. (2015), Giraud et al. (2016)).

SDM biases factors under heterogeneous sampling An occurrence gives the only certainty of the presence of a specimen at a given time and place where it was reported. Absence where the species has not been reported remains uncertain, particularly for opportunistic data. Thus, a different distribution of sampling effort in space and time produces a different occurrence density, which distorts our view of the actual distribution of the species.

First, Figure 5 represents a natural model decomposing the drivers of the sampling process. The sampling process of a given observer is decomposed into the probability that a location is observed (observation effort), the probability that it is detected, reported and ultimately identified as the good species. The three events must happen successively for that observer to produce an occurrence. We note $p(u)$ the probability to detect, report and identify correctly an individual located at $u \in \mathbb{R}^3$ (assuming same probability for the m total observers) conditionally to observation of u by the observer. Equation 5 shows that the expected number $N(u)$ of reported occurrences at u may be approximately factorized as the product of $p(u)$ and the sampling effort $v(u) = \mathbb{E}(V(u))$ when $p(u)$ is very small, which is typically the case when dealing with opportunistic sampling. $N(u)$ also becomes a Bernoulli random variable as $p \to -> 0$ and, then, $E(N(u))$ equals the probability to sample the occurrence. This property allows, for example, to simply model observed occurrences as a thinned Poisson process as we do it in Chapter **2** and **3**.

$$\begin{aligned} \mathbb{E}(N(u)) &= \textstyle\sum_{i=1}^{m} i \sum_{j=i}^{m} p(V(u)=j)\binom{j}{i} p(u)^i (1-p(u))^{j-i} \\ &\underset{p=\mathrm{o}(1)}{\approx} \textstyle\sum_{j=1}^{m} p(V(u)=j) jp \\ &= p(u)\, \mathbb{E}(V(u)) \end{aligned} \tag{5}$$

Equation 5 tells us that, under our assumptions, the variation in sampling effort can be decomposed in a multiplicative way into the variation in the probability of observation, detection, reporting and identification. Thus, the density of reported occurrences co-varies with the abundance and with the sampling effort, which will artificially increase the number of occurrences in areas or times that are more visited, where detection is easier or where observers are more prone to contributing. As presence-only SDM methods fit a response surface in the space of environmental variables, an environment with larger average sampling effort will have an over-estimated species response. Evidence of sampling bias have been shown in simulation Leitão et al. (2011). A strong evidence of SDM environmental bias on opportunistic presence only data is given by Syfert et al. (2013). They fit MAXENT on occurrences of ferns in NZ with sampling bias correction, using background points sampled according to a sampling effort

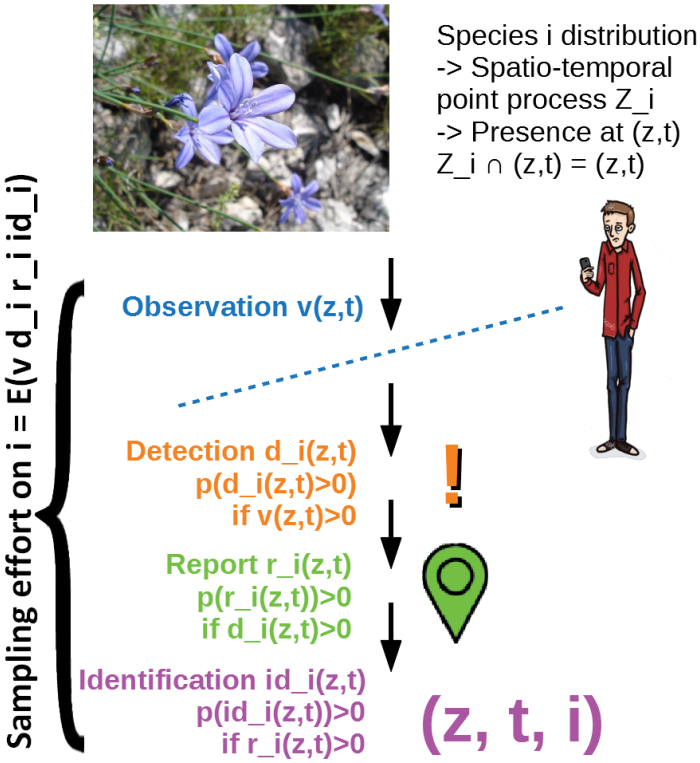


Figure 5: Illustration of the decomposition of the sampling process for an occurrence of *Aphyllanthes monspeliensis* L. (photo credit © Jean Tosti, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=168737`)

grid (computed from all vascular plants occurrences over the region), or no correction (random background). They observed that non-corrected models had much less predictive accuracy on independent presence-absence data than when bias was taken into account.

Variation in the sampling effort may be explain by many factors. The observation effort depends on the terrain accessibility (Warton et al. (2013), Reddy and Dávalos (2003)) and practicability, e.g. roughness or vegetation density. To illustrate uneveness of the observation effort across environments, the bottom graph of Figure 1 shows the distribution of the Pl@ntNet users identification queries across land cover categories compared to the proportion of these categories in space. The proportion of urban occurrences is much higher than urban global percentage cover, which shows how observers activity is concentrated in cities and their surroundings. Also, Figure 2 illustrates the distribution of the distance to roads for the same Pl@ntNet occurrences (above), which is much more concentrated toward roads than the distribution of uniformly drawn points over the territory (below). Then, the detection probability depends on the visibility of the specimens in a given area, and thus on the weather and landscape (Wintle et al., 2005). For plant species, detectability has especially been shown to depend on size, form, flower presence/colour/size (Kéry and Gregg (2003), Slade et al. (2003), Burrows (2004)) and local abundance (Garrard et al., 2013). It is also strongly suspected that the reporting Pl@ntNet probability depends on the interest of the species for the observer and on the propensity of the environment to induce a contributing behaviour, e.g. observers are more likely to conduct surveys when hiking in the wild than when passing over a strip of grass in an industrial area.

If we consider all those factors affecting globally sampling effort, we expect that the later will vary across space, environments, seasons, years, species, and observers. However, SDM do not generally account for all those dimensions, and even less for sampling effort along those dimensions. Problematically, note that when one of the dimensions is integrated out in the model (e.g. static SDM on occurrences collected over several years), and there is interaction of sampling effort along both dimensions, it will induce an irremediable bias. For example, the sampling effort may vary in space and time. For example, Pellet (2008) showed, through site-occupancy models, that butterfly detectability highly varies during a season, and that intervals of high detectability are distinct between species. If we were to model the distribution of those species from presences-only, we should not only account for varying global sampling effort over time in the model, but also for varying spatial distribution of sampling effort over time, because observed areas vary from one season to another. Similar phenomenons appear on many plants because of their phenology: Plants whose vegetative parts disappear during winter are undetectable during this period, and besides most plants are more remarkable when flowering.

Accounting properly for such phenomenons in SDM is really complex. For now, we decided to focus in my thesis on the correction of spatial and environmental variations of sampling effort. Their precise links and effect on presence-only Poisson process models is unravelled in Chapter 2.

Methods for addressing spatial sampling bias in SDM based on opportunistic occurrences. Several methods have been proposed to correct for spatial and environnemental sampling bias. A first type of correction methods manipulate occurrences and background points. A wrong selection of background points in Maxent or Poisson processes, e.g. taking background points distributed over a larger area than the one that has been sampled, or conversely in a smaller area, will entail environmental sampling bias in the model as described in the case of (Phillips, 2008). They state that the background points should represent the area

that has been observed to generate the species occurrences to avoid bias. This is surely an aspect to look at when defining background but this is insufficient, because (i) we can't identify sites that were visited without producing occurrences, and (ii) the level of sampling effort is likely to vary a lot among visited areas. Other authors have proposed to eliminate some occurrences when they are too concentrated in the geographical space (Boria et al. (2014),Varela et al. (2014)) or in the environmental space (Varela et al., 2014) to avoid over-concentration of occurrences in some areas due to high sampling effort. The environmental filtering approach was shown to improve presence-absence discrimination (Varela et al., 2014) on a virtual species. Spatial filtering correction was worst than no correction in certain studies (Varela et al. (2014), Boria et al. (2014)). No guideline is given to select the filtering strength. Plus, the occurrences density reflects not only sampling effort but also the species abundance. Removing blindly occurrences based on their raw density clearly deviates the problem from one type of bias to another. For example, a specialist species over some environmental gradient will have concentrated occurrences along this axis. Then, environmental filtering induce a generalism bias on the response estimate. Another popular approach, called TGB, uses the sites where at least an occurrences have been collected, from a so called Target-Group of species, as background points. It has been introduced and tested on real data by Phillips et al. (2009), after successful simulation results from Dudík et al. (2006). Syfert et al. (2013) independently evaluated the TGB approach compared to uniform background points on two biased ferns occurrences datasets. They found that TGB yielded much better prediction performance on independent presence-absence test data than uniform background, but that the predictions from the most biased training dataset was still inferior, suggesting residual bias. More contrasted results were brought by (Ranc et al., 2016) who showed that TGB correction was efficient for species with wide occurrence area whereas it was worst than no correction for species having a narrow spatial range of occurrence. TGB also received an interesting theoretical critic from Warton et al. (2013) (Figure 2), who remarked that environments with higher species richness would be over corrected by TGB method as they generate more occurrences independently of the sampling effort. Another question is how to define the resolution of "sites" (Phillips et al., 2009), where TG background points are aggregated as one, when dealing with continuously distributed and accurately geolocated species occurrences as in Pl@ntNet.

A second type of correction approach adopt the strategy to explicitly model and estimate sampling effort along with the species intensity. Indeed, Ranc et al. (2016) stated that *"the attention should be focused on a more essential question: how to estimate environmental sampling bias from an opportunistic dataset containing multiple species"*. This is not straightforward as the multiplicative dependence of sampling effort and species abundance in the data lead to a fundamental problem of source separation. The first strategy proposed was to jointly model sampling effort with a single species response (Warton et al., 2013) in the context of Poisson point processes. The sampling effort model is built from ad hoc co-variables assumed to drive its spatial variation, like distance to roads, to cities and to coasts, in the case of this eucalyptus dataset from Queensland, Australia. A successful implementation of this method was done by Stolar and Nielsen (2015). A method was proposed to estimate specifically the sampling effort by Fernández and Nakamura (2015). They model the detection of a target group of species based on a multinomial of observation where the sampling effort is a accessibility factor depending on distances to roads, cities and cities populations, with few parameters. The estimated sampling effort may be used to generate background data and, then, fit a presence only SDM on any species assumed to have been sampled with similar effort than the Target-Group. In the case of jointly modelling sampling effort and species response in a log-linear Poisson process (Warton et al., 2013), the modeler must insure that their respec-

tive co-variables are all distinct and far from co-linearity. In some conditions, this restriction may be relaxed when adding complementary standardized data in the model. Giraud et al. (2016) modelled opportunistic counts of many birds species over several sites jointly with standardized counts (where the sampling effort is known). Opportunistic counts were selected to match sites of standardized counts. However, the latter lacked several species compared to the former, and counts were lower. They set up a fully spatial model (species responses and sampling efforts had a distinct parameter per site). They showed that the known sampling effort of standardized data enabled to identify the whole model, i.e. the opportunistic relative sampling effort in each site, which in turn unlocked the identification of the relative abundance of species unobserved in the standardized data in the same site. Plus, they showed that integrating the opportunistic data significantly reduced the relative abundance estimation variance of rare species compared to the standardized data alone. Fithian et al. (2015) proposed in parallel a close approach. They jointly modelled opportunistic occurrences with a shared sampling effort component among multiple species occurrences intensities, and integrated to this model standardized presence-absence data on some plots. In this case, sampling effort is modelled as a function of geographic variables (as in Warton et al. (2013)) which enable to correct sampling bias of opportunistic occurrences even outside of presence-absence plot. Note that similar modelling approaches based on the same principle were developed for jointly integrating site-occupancy data with opportunistic occurrences (Dorazio (2014), Koshkina et al. (2017)). Finally, the knowledge of reporting behavior of some active observers has been used to infer species absences from opportunistic occurrences of citizen science program (Bradter et al., 2018).

As shown by our bibliographic survey, even if the sampling bias has been highlighted as a major problem of presence only SDM for almost two decades, the published knowledge of its mechanism and interaction with species distribution is empirical and based on a few specific examples. To summarize, two types of approaches have been proposed to correct for spatial and environmental sampling bias. Selecting background points whose distribution is a proxy of the sampling effort, or estimating the sampling effort jointly with species response, either through on a simple tailored model of sampling effort or with complementary standardized data. The former approach is simpler to set up, provides less variables estimates of the species response, but it may induce biases that are not well understood when the background is inappropriate, and their procedure is unclear. The latter has more statistically sound way of correcting bias, but it implies more sophisticated models, to identify a proper sampling model and/or complementary standardized data, and is prone to estimation variance as the number of parameter is increased by the addition of a sampling effort model.

4.4.2 Input dimension constraints in the era of deep learning

Today, many new sources or types of environmental layers become available for SDM at high resolution, especially because of the development of remote sensing technologies and automatic image analysis. Then, the number of potentially relevant variables to include in SDM is high, and the ecological modeller would ideally like to include them all as input and have an algorithm able to extract only the effects that significantly correlate with the studied response. Besides, Lek et al. (1996) early stated that relationships between environmental variables and species response were often poorly captured by linear models, and were often tailored through ad hoc transformations leading to sub-optimal model fitting. However, when adding new variables, the number of potential interactions effects between several variables increases exponentially. For example, a species might have a realized niche that imply a complex combination of many environmental variables values. When accounting for such complex effects in the model,

the modeller is often confronted to the problem of over-fitting. The potential complexity of a class of response functions of a statistical model is often called model expressivity (Raghu et al., 2017). It is easy to build and fit expressive models, but again, it often translates into over-fitted estimates with low out-of-sample predictive power, i.e. low generalization power. Deep and convolutional neural networks along with their state of the art learning procedures are said to be remarkably resistant to over-fitting. Thus, they may be an interesting tool for fitting more complex and robust response functions over more environmental variables in SDM. In the following, we come back to the concepts of generalization power, over-fitting and regularization when increasing the input dimension and model complexity through fundamental statistical learning theory principles. We then introduce the recent theoretical arguments justifying the generalization power of deep and convolutional neural networks and highlight opportunities of applications for SDM. For an introduction of NN and their architectures, the reader may refer to **Chapter 4**.

Generalization power and regularization The statistical learning theory (Vapnik, 2013) aims at answering the problem of learning the most robust approximation of the response function from a finite noisy sample of unknown error distribution. It has been mathematically formalized by the framework of the Probably Approximately Correct (PAC) learning (Valiant (1984), Vapnik (2013)). A concept is said to be learnable by an algorithm if the algorithm may find in polynomial time a function whose predictive error is bounded with any defined probability by some constant. The statistical learning theory of Vapnik (2013) has a different point of view from classical statistical inference. It is based on the idea that the true probabilistic distribution of the output data conditionally on the input is unknown and a robust learning of a predictive function shouldn't be based on any distribution hypothesis. It aims at providing algorithms risk bounds that are independent of it and speed of convergence rate for different learning algorithms. In the following, we note $\mathcal{H}$ the class of functions that may be fitted by a learning algorithm. The statistical learning of a predictive function $f \in \mathcal{H}$ is based on the empirical risk minimization, which generalizes the maximum likelihood method for independent and identically distributed data. An important insight from the work of Vapnik (2013) is given by equation 6.

$$Pr\left(\epsilon^2 \leq \frac{D(\mathcal{H})(\log(2N/D(\mathcal{H})) + 1) + \log(\mu/4)}{N}\right) = 1 - \mu \tag{6}$$

Where D is the VC dimension of $\mathcal{H}$ is the largest integer h such that there exists a sample of size h which is shattered by $\mathcal{H}$. A set of data points are shattered by $\mathcal{H}$ if, for all assignments of labels (resp. values) to those points, there exists a $f \in \mathcal{H}$ such that f makes no errors when evaluated over this set of data points. Equation 6 states that, when the sample size is kept constant, the generalization error increases monotonically with the VC-dimension, i.e. a type of complexity measure of the model class of functions. Note that similar equations have been derived for other measures of complexity such as the Rademacher measure (Gnecco and Sanguineti, 2008). If we consider real input variables the same bounds, as the input dimension increases, the number of variables combinations (product of intervals with same width) increases exponentially. Thus, the number of data required to estimate the output response for each variables combination is also an exponential function of the input dimension, as for the VC dimension of the model class of functions modelling the response. Then, when increasing the input dimension, we increase exponentially the VC dimension if we adapt the model class of functions to account for all variables interactions effects, while our sample size is kept constant. Finally, according to equation 6, it induces a sharp increase in

the generalization error, that is over-fitting. This problem is called the curse of dimensionality in the context of statistical learning (Giraud, 2014).

Then, controlling the complexity of the model class of functions is very important to control over-fitting. Regularization helps to reduce the complexity of a given class of function by setting an a priori preference on certain functions inside the class: Constraining the functions of the class to have certain properties (e.g. function smoothness in the context of non parametric statistics which is the principle of GAM, see citeyee1991generalized, input dimensionality reduction, see Giraud (2014), or restriction of input variables through expert knowledge), quantitatively penalizing certain functions in the loss (e.g. L2/Ridge (REF) or L1/Lasso penalty, Tibshirani (1996)), modifying the loss/error function, using informative prior distributions of parameters for Bayesian inference, or a model optimization algorithm with implicit regularization (Ioffe and Szegedy (2015), Chaudhari and Soatto (2018)). For instance, Maxent regularization scheme allows it to fit predictive SDM from very few occurrences (less than 50, see Phillips and Dudík (2008)) using tens of environmental variables and their transformations. This is because Maxent exploits the variables selection properties of the Lasso regularization scheme (Tibshirani, 1996).

Recent resurgence of deep and convolutional neural networks. Feedforward Neural Network (NN) have been known for long to be highly expressive models. Indeed, (Hornik et al., 1989) showed that, given enough neurons, NN are able to approximate arbitrarily well any measurable function of any number of bounded input variables, and are thus a class of universal approximators. This is true even for a single hidden layer network, and for any non-linear and continuous activation function. Many statistical learning methods also have this property, but not all are equal in their generalization power and regularization ability. Even though the method with highest generalization power should depend on the problem, many empirical and theoretical results appearing during the last decade from many domains have supported the idea that deep NN enjoy more resistance to over-fitting (Poggio et al., 2017), and thus enable to deal with much more high dimensional input data. First deep feedforward NN were learnt more than thirty years ago (Werbos (1974), Parker (1985), Lecun (1985), Rumelhart et al. (1988)), but the last decade has seen many applications showing evidence of their impressive generalization power for speech recognition, visual object identification (Krizhevsky et al., 2012) and object detection compared to other machine learning methods with large and diversified training dataset (LeCun et al., 2015).

This remarkable success asks the question: Why deep NN revealed their empirical efficiency less than a decade ago while they have been initially invented and used more than 30 years ago? Deep NN have indeed been stuck for many years in the bag of models with restricted uptake in the machine learning community after the first introduction of the back-propagation algorithm that enable to optimize their weights (Rumelhart et al., 1988). It is now quite clear that deep learning architectures may only outperform other machine learning methods on real world applications when the sample size is large enough, otherwise the sample doesn't provide a dense enough coverage of the high dimensional input space, to enable any robust estimation of a complex response function. Then, Deep and convolutional NN have been able to show significant performance thanks to the unlocking of three major locks: access to large learning datasets, the democratization of Graphical Processing units (GPU) computing and advances in NN optimization techniques.

Firstly, large labelled dataset started to appear and be freely accessible on the web for important machine learning problems such as image classification and object detection (the main example being Deng et al. (2009)). Such large datasets require a long time to annotate

them and greatly benefited from collaborative efforts.

Secondly, GPU computing softwares for deep learning started to appear (e.g. Krizhevsky et al. (2012)). Indeed, GPU computing enabled to greatly accelerate learning of deep NN on regular computers, enabling any researcher to experiment those methods and engineer them to answer specific problems. Goeau et al. (2017) Thirdly, advances in the theoretical understanding of their optimization process and the development of optimization techniques facilitated and accelerated the optimization of NN and were identified as drivers of improved empirical predictive performances. Two main theoretical issues have been recently finely understood and taken into account in current optimization techniques of deep NN that concern: Escaping the many local minima/saddle points and controlling the gradient value. A wide or deep enough NN is theoretically very expressive and may fit well any set of data, but optimizing its parameters (called weights) by standard gradient descent with the back-propagation algorithm will generally lead to a poor fit due to the many sub-optimal local optima and saddle points of the loss function in the space of weights, where gradient descent will get stuck. mini-Batch stochastic gradient descent algorithm (SGD) and its numerous extensions like ADAM, RMSprop, AdaDelta, or (see Ruder (2016) for a review) were applied with more success. It may indeed quickly vanish or explode during model learning (Hanin, 2018). During the back-propagation step, the weights are updated in the opposite direction of the gradient, i.e. towards the closest local minima. The distance of the minima will depend on the local curvature which is unknown, even though it is approximated adaptatively by some stochastic gradient algorithms. Then, If the learning rate is too low, it will induce a vanishing gradient, because we will get stuck in a close local minima. Conversely if it is too high, weights update might increase the loss compared to the last step, tending to induce a higher gradient norm on the next mini-batch and a loss increase runaway. Furthermore, the control of the learning rate (the global coefficient scaling the update of all weights) is crucial, as the curvature of the loss may vary importantly from one point to another in the space of input variables. An inappropriate random initialization might also induce a wrong gradient regime (Krähenbühl et al., 2015). Internal covariate shift is another important problem. A deep NN has many layers combining the neurons activations from the previous ones. If the distribution of values of the previous layer is biased or has high variance, it may induce strong variations in the activations of the current layer, which will propagate to the next (Ioffe and Szegedy, 2015). This may easily happen with batch stochastic gradient descent algorithms where each descent step computes the predictions and gradient over a very small subset of the data (typically between 32 and 128 data). Initially, the activation function used in NN was the sigmoid function. It has a fast vanishing derivative when the input moves away from 0, progressively forbidding parameters of previous layers to be updated.. The ReLU activation function, $x \to x * 1_{\{x \geq 0\}}$, has been proposed by Nair and Hinton (2010) to help solving this problem. The previous phenomenon can't happen on the positive side because its derivative equals 1 and the activation intensity value is conserved for the next layer. It also accelerates the learning speed of deep NN and was showed to improve deep NN performances on many other machine learning benchmark datasets (e.g. (Zeiler et al., 2013) for speech recognition). Another important ingredient to the optimization of deep NN was Batch-Normalization, which consists of standardizing each activation of the model, while computing the predictions in the forward phase, by the mean and standard deviation of the same activation over the current batch of data (Ioffe and Szegedy, 2015). Using BatchNorm forces each activation to be centered with unit variance and thus prevents them to shift. BatchNorm prevents the vanishing of many neurons, and it allows to increase the learning rate with much more resilience to exploding and vanishing gradient and thus induce a much faster learning. It is also an efficient element of regularization. Empirical

evidence and theoretical elements introduced above have shown that the learning rate policy, the mini-batch size, the type of descent algorithm, parameters initialization, activation function choice and other operations in the optimization procedure greatly impact the ability to learn a deep NN with good generalization power, and with fewer computational time. Thus, they are elements to consider all together to achieve an efficient learning scheme on a new problem.

To summarize, recent advances in datasets availability, GPU oriented deep learning libraries, and optimization techniques have enable the learning of complex deep and convolutional NN models on much larger datasets with a drastically reduced computational cost and time investment. However, it doesn't explain why, fundamentally, deep learning methods have a remarkable resistance to over-fitting and are consequently the most efficient methods nowadays on many machine learning tasks, which is discussed in the next paragraph.

Towards understanding deep and convolutional NN generalization power. An interesting point is that conventional measures of the model class of functions complexity such as the VC dimension highly over-estimate the complexity of deep NN models, and thus their generalization error (Zhang et al., 2016). In other words, the measure of deep NN models complexity through their number of parameters suggest, according to equation 6, that they should over-fit much more than they actually do empirically. This asks why deep NN don't over-fit so easily to the data. Several research branches currently look for answer to this question. In the remainder of this paragraph, we discuss recent state of the art on this topic.

Also, deeper architectures are often said to be resistant to over-fitting to shallower networks, providing better generalization power for complex response function or higher input dimension. However, a concept of function complexity may only be defined based on a set of elementary functional components, but the suitability of basis of functions is specific for representing certain classes functions. As a basic example, a continuous periodic function is exactly fitted by its finite Fourrier serie, while no finite monomial basis $(1, x, x^2, ...)$ can fit it. Conversely, the natural exponential function is quickly approximated with monomial terms of its Taylor expansion while it takes much more terms to achieve the same precision with a Fourrier serie. As it is an ill-defined problem, no general measure of function complexity exists. Plus, the exact response function depend on the problem and is only observed through finite data. Thus, a better way to set the problem is to ask how well a family of approximators (typically a NN model with all possible weights values), may approximate a class of functions that is likely to contain the target function for a wide spectrum of problems. The property of compositionality of NN originally appeared to be a main motivation for hierarchical models of visual cortex because they could be regarded as a pyramid of AND and OR layers, that is a sequence of conjunctions and disjunctions (Riesenhuber and Poggio, 1999). Deep neural networks was pushed by Hinton and Lecun LeCun et al. (1989), arguing that deeper NN architectures provide some exponential expressivity while keeping the number of parameters reasonable. This argument came from results from the complexity theory of boolean circuits (Håstad and Goldmann, 1991) early suggested that Deep NN built with multiple hidden layers were more appropriate for approximating composed functions of many input variables. Mathematically sound results of the same nature were brought on deep NN later. Montufar et al. (2014) provided a discretized view of the notion of NN model expressivity by counting the number of linear regions that can be synthesized by a deep network with rectified linear unit (ReLU) nonlinearities. Very recently, Raghu et al. (2017) showed that a certain measure of expressivity of NN called "trajectory length", and defined by the level of variation of the output along a one dimensional trajectory in the input space, grows exponentially with the depth of the

model. An active field of research in explaining why generalization power of NN comes from their depth is the one merging function approximation theory and PAC learning. It provides a mathematically sound way of defining and characterizing networks architecture efficiency. This branch tries to identify more precise sets of functions that may be approximated by DNNs. Significant results in this branch were summarized in Poggio et al. (2017) review. It is shown that (i) deep networks are especially well designed to approximate response functions that compose "local" constituent functions, e.g. functions with low dimensional input, and in these case avoid the curse of dimensionality, (ii) for such response functions, there is a theoretical guarantee that deep convolutional NN architectures outperform one layered architectures and kernel machines, because convolution is by definition a local function in the previous sense (even without weight sharing, which reduce drastically the family of estimators size). If the elementary polynomials are sparse then deep NN will be even more efficient. A conjecture is also stated about the fact that the multi-class setting of many problems where deep learning is applied favor the success of these architectures, because it forces the extraction of shared functional components between all classes which match the structure of some problem. Further results were brought by (Gribonval et al., 2019) that provided a measure of the speed at which convergence is achieved when increasing the size and depth of the network for several classes of true response functions.

Another important axe of research to explain the generalization power of deep NN is the regularization imposed by the Stochastic Gradient Descent algorithm (Chaudhari and Soatto, 2018). It also worthwhile to note that recent works suggest that generalization puzzle of deep learning can been analysed through an Information theoretic perspective. Notably, Shwartz-Ziv and Tishby (2017) suggests that a deep network can be seen as an information processing pipeline where the information retained about the input progressively reduces as it goes through the network's layers. The so-called information bottleneck phenomenon would naturally prevent the network from learning over-complex representation of the input retaining only useful information to predict the output variable. However, other works suggest that compression might not be the only key procress to provide good generalization as showed through invertible deep neural network function architecture, which preserve all the input information, and still achieve good predictive performance on the ImageNet dataset [16].

Finally, a path for understanding the effect of the regularization operated in deep NN optimization is the study of norms of the model weights along the learning process. A whole body of theory and strong empirical results show evidence that some norm on the model weights might control the excess risk of generalization, i.e. the difference between the test and train errors (Bartlett (1998), Neyshabur et al. (2015), Liang et al. (2017), Barbet-Massin et al. (2018)).

In summary, several important mechanisms of deep NN regularization have been shown through the current merging of approximation and PAC learning theories, the analysis of the implicit regularization of the stochastic gradient descent algorithm and the analysis of weights norm in during the optimization process. However, current research hasn't yet provided a global and firm theoretical framework to understand deep learning generalization power. the state of the art is not yet able to provide clear theoretical support for architecture design guidelines.

Capturing response to spatio-environmental patterns through convolutional NN. We have seen in section **5.3.1-2** that complex realized species distribution patterns of abundance may result from the local structure of the landscape because of source-sink dynamics,

[16] `http://www.image-net.org/`

surrounding biotic context, and environmental changes along time. Then, we may ask if and how SDM could capture patterns of the landscape to better predict species response. Deep convolutional Neural Network (NN) may be a good opportunity to address this challenge. Indeed, Convolutional layers of deep CNN are specifically designed to detect spatial patterns of various scales in n-channel images (Zeiler and Fergus, 2014), while restricting drastically the space of functions compared to other existing approaches. Besides, species distribution models enabling to capture spatial patterns of the environment have never been developed. Thus, CNNs open attractive perspectives for analysing how complex spatial structure of the environment affect patterns of species distribution.

4.5 Questions of the thesis

In this section We take the reader through the research questions of my work of PhD and explain the organisation of the chapters.

As we have seen in section **5.2**, the Pl@ntNet mobile application generates a lot of geolocated species occurrences, which for the most part are only automatically identified by a deep Convolutional NN classifier without any human validation: the queries. An important concern for the Pl@ntNet project was to know how these data could be exploited to contribute to the monitoring of biodiversity. It seems suited to survey invasive alien species. Indeed, most of the concentration of queries is tied to the human population, which overlap with the distribution of most alien invasive species because of their introduction, naturalisation or dissemination mechanisms. However, even though the identification quality was very heterogeneous among species and pictures, we can use the output probability distribution of the queries classifier to approximate identification certainty. Thus, in **Chapter 1** we wondered (i) if we could retrieve realistic invasive species distribution patterns with Maxent applied to the geolocated queries, and (ii) what was the effect of various level of occurrences filtering, based on the classifier score of the most likely species, on the relevance of the final SDM. We selected 7 alien invasive species in the French metropolitan territory, run Maxent (Phillips et al., 2006) SDM on those occurrences, filtered with several level of identification certainty, using uniform background and model based sampling bias correction (Warton et al., 2013) and compared the models predictions to independent national expert reports (FCBN-INPN). We evaluated the True Skill Statistics (TSS) of the predicted response on FCBN presence-absence over 10x10km sites and another metric measuring how well the model recovered the most abundant sites. This study is presented in the form of our first article and is provided in **Chapter 1**.

Dealing with spatio-environmental sampling bias when making SDM from large opportunistic occurrences database was at the center of the questions of the PhD and is addressed in **Chapter 2** and **Chapter 3**. Sampling bias is tied to the environment and affects importantly the distribution of occurrences in the Pl@ntNet data. We noticed that the model based bias correction was unsatisfying in **Chapter 1** because of descriptors associated with the species intensity captured a visible part of the sampling density.

The Target-Group background strategy is a very promising alternative because it avoids the technical difficulties of sampling effort estimation. However, even though the method intuitive idea seem relevant, section 5.4.1 showed that the ways this method is put in practice are heterogeneous, and its results are uncertain (Warton et al. (2013), Ranc et al. (2016)). There is no procedure or justification of practical guidelines to define the composition of the Target-Group of species or the resolution of the sites grid. Taking a step back, the practice of transforming occurrences and background into "at least one occurrence" per site induce a scale dependency that is questionable in the framework of Poisson point processes. We considered

the application of a limit case of TGB, the method that takes all individual Target-Group Occurrences as Background (TGOB). More generally, there is no clear unifying description of the mechanism of sampling bias based on a theoretical background, that could guide the practice of SDM, nor of biases appearing because of TGB correction, even though an effect of spatially varying species richness has been mentioned in Warton et al. (2013). Sampling bias has been empirically shown to depend on the species occurrence area (Ranc et al., 2016), which is a combination of the generalism of the species along many environmental gradients, its global occurrence rate and its global sampling effort. We argue that it is true but quite vague. The bias may be characterized along each environmental gradient, because the SDM is in the end a function of the environmental variable. In **Chapter 2**, we formalized the effect of spatial sampling effort on species occurrences intensity in the environmental space and on Poisson process model inference bias with two types of background points selection scheme: Uniformly drawn Background (UB) points or TGOB. We assumed that species are distributed according to independent Poisson processes, whose realizations are thinned according to the sampling effort function. We especially asked how the Target-Group species distributions affect bias in the Poisson process species intensity estimate, and what it takes to control this bias.

Chapter 2 showed that it is difficult to control for bias when applying the TGOB strategy, due to unknown TG species realized niches. Working on improving the joint modelling of species intensities and sampling effort seems more promising, because integrating explicitly sampling effort in the model enables to recast the initial problem of bias as a problem of estimation variance, which we can hope to solve through the large amount of data available. Joint integration of standardized data with opportunistic presences-only in SDM may greatly help to separate species distribution from observation effort (Giraud et al. (2016), Fithian et al. (2015), Koshkina et al. (2017)), but standardized flora surveys are costly, as the identification expertise becomes scarce. Those data are hard to access or dates back to many years ago, while we are especially motivated by a regular monitoring of alien invasive species distribution. Thus, in **Chapter 3**, we propose a method for jointly estimating observation effort and multiple species responses functions from large amount of opportunistic occurrences only, based on a marked Poisson process model. The observation effort component is shared among species and is a step function defined over a partition of space into cells. Jointly multiple species should improve the estimation of all species response, especially on the margins of their distribution, because we get information on the observation effort everywhere where at least some species live.

Another concern of the PhD was to investigate the opportunities of deep learning methods to improve the spatial prediction of plant species, that could complement picture based identification by developing spatial recommendation systems like in Mac Aodha et al. (2019). We have seen in section **5.3.1-2** that patterns of abundance may result from the local structure of the landscape because of source-sink dynamics, surrounding biotic context, and environmental changes along time. We ask if and how SDM could capture patterns of the landscape to improve prediction of species response. Deep convolutional NN have been shown to be an efficient model architecture to capture spatial patterns that can be modelled by composed function of local low dimensional patterns, as seen in section **5.4.2**. In **Chapter 4**, we investigate if Deep convolutional NN can perform better than state of the art presence-only SDM. Plus, we have seen in section **5.4.2** that multi-class deep NN generalized better than single-class when classes shared input features patterns (Poggio et al., 2017). Thus, **Chapter 4** also evaluates if deep NN, and deep convolutional NN, with more species responses but the same hidden layers architecture generalize better on every species.

Learning complex NN architectures takes computational time and finding efficient implementations for a given problem open numerous possibilities: e.g. the choice of pre-treatments, input data, architecture, regularization scheme, learning algorithm, etc. Thus, deep learning model improvement is currently an empirical process, where the modeler cannot test every combination and thus refines its choices depending on the effect of local changes. The LifeCLEF evaluation campaign began in 2014 (Joly et al., 2014) and was built in the spirit of providing a place for multiple researchers to work at developing the best algorithms for various task of biodiversity identification. A train dataset is provided to participants who may use it for creating their algorithms. A test dataset is built to evaluate the predictive performances of algorithms on the task, and each participant may submit answers of several algorithms to the evaluation. In the context of this PhD, GeoLifeCLEF has been initiated in 2018 as a task of LifeCLEF specifically dedicated to evaluate fine spatial grain species recommendation algorithms. The aim of the algorithm is to predict the list of species that are the most likely to be observed at a given location. The expected research outcomes of GeoLifeCLEF were of two types: (i) Evaluate new machine learning presence-only SDM algorithms with higher fine spatial grain predictive power that may serve as material for ecological research, (ii) favor new insights on avenues of improvement for next generation predictive models. **Chapter 5** report the main results of the two editions of GeoLifeCLEF coordinated in 2018 and 2019. For the first edition of the task we extracted around 300,000 plant species occurrences from the GBIF [17] in France, among which 70,000 were taken as evaluation data, and provided an image patch of 33 environmental variables for each occurrence to enhance the experimentation of convolutional models. 2019 edition kept the same study area, but provided a much larger training set of species occurrences including the 2018 dataset, more than 2 millions Pl@ntNet queries and around 10 millions of occurrences from other biological groups extracted from the GBIF. The test set improved in quality as we provided 25,000 expert occurrences from the CBNmed [18] with accurate geolocation and a weighted selection scheme to insure uniform distribution in space and more concordance between species representation and their true spatial abundance. We also changed the evaluation metric to account for the fact that many plant species coexist under the scale of geolocation accuracy. 3 participants submitted 33 runs and 5 participants submitted 44 runs to the 2018 and 2019 editions respectively, providing interesting open and perennial benchmark datasets for spatial species recommendation.

[17] `https://www.gbif.org/`
[18] `www.cbnmed.fr//`

5 Chapter 1: Species distribution modelling based on the automated identification of citizen observations

Applications in Plant Sciences

APPLICATION ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Green Digitization: Online Botanical Collections Data Answering Real-World Questions

Species distribution modeling based on the automated identification of citizen observations

Christophe Botella[1,2,3,4], Alexis Joly[1], Pierre Bonnet[3,5,7], Pascal Monestiez[4], and François Munoz[6]

Manuscript received 8 September 2017; revision accepted 2 January 2018.

[1] Institut national de recherche en informatique et en automatique (INRIA) Sophia-Antipolis, ZENITH team, Laboratory of Informatics, Robotics and Microelectronics–Joint Research Unit 5506-CC 477, 161 rue Ada, 34095 Montpellier CEDEX 5, France

[2] Institut National de la Recherche Agronomique (INRA), Joint Research Unit Botanique et modélisation de l'architecture des plantes et des végétations (UMR AMAP), F-34398 Montpellier, France

[3] AMAP, Université de Montpellier, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), French National Center for Scientific Research, INRA, IRD, Montpellier, France

[4] BioSP, INRA, Site Agroparc, 84914 Avignon, France

[5] CIRAD, UMR AMAP, F-34398 Montpellier, France

[6] Université Grenoble Alpes, Laboratoire d'Écologie Alpine, CS 40700, 38058 Grenoble CEDEX, France

[7] Author for correspondence: pierre.bonnet@cirad.fr

Citation: Botella, C., A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. 2018. Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences* 6(2): e1029.

doi:10.1002/aps3.1029

PREMISE OF THE STUDY: A species distribution model computed with automatically identified plant observations was developed and evaluated to contribute to future ecological studies.

METHODS: We used deep learning techniques to automatically identify opportunistic plant observations made by citizens through a popular mobile application. We compared species distribution modeling of invasive alien plants based on these data to inventories made by experts.

RESULTS: The trained models have a reasonable predictive effectiveness for some species, but they are biased by the massive presence of cultivated specimens.

DISCUSSION: The method proposed here allows for fine-grained and regular monitoring of some species of interest based on opportunistic observations. More in-depth investigation of the typology of the observations and the sampling bias should help improve the approach in the future.

KEY WORDS automated species identification; citizen science; crowdsourcing; deep learning; invasive alien species; species distribution modeling.

Identifying organisms is a key step in accessing information related to the ecology of species. Specifically, large-scale monitoring of species distribution dynamics is essential in the context of global change. Such monitoring requires intensive occurrence data, but such data are lacking due to the level of expertise necessary to correctly identify and record living organisms. This is especially true for plants, which are one of the most difficult groups to identify, with more than 350,000 known species on earth. The Rio Conference of 1992 (the Earth Summit, United Nations Conference on Environment and Development [UNCED], Rio de Janeiro, Brazil, 3–14 June 1992 [http://www.un.org/geninfo/bp/enviro.html]) recognized this taxonomic gap as a major obstacle to the global implementation of the Convention on Biological Diversity. Gaston and O'Neill (2004) discussed the potential of using automated identification approaches, typically based on machine learning and multimedia data analysis methods, to produce more intensive occurrence data. They suggested that if the scientific community is able to (1) overcome the production of large training data sets, (2) more precisely identify and evaluate error rates, (3) scale up automated approaches, and (4) detect novel species, it will then be possible to initiate the

 © 2018 Botella et al. *Applications in Plant Sciences* is published by Wiley Periodicals, Inc. on behalf of the Botanical Society of America. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

development of a generic automated species identification system. Such a system should then open important opportunities for studies in biology, ecology, and related fields.

Since Gaston and O'Neill (2004) raised the question, enormous work has been done on the development of automated approaches for plant species identification (Casanova et al., 2009; Yanikoglu et al., 2014; Lee et al., 2015; Champ et al., 2016; Goëau et al., 2016; Joly et al., 2016; Wilf et al., 2016; Wäldchen and Mäder, 2017). Deep learning techniques in particular have been recently shown to achieve impressive recognition performance (Goëau et al., 2017). Some of these results were integrated into effective web or mobile tools and have initiated close interactions between computer scientists and end-users such as ecologists, botanists, educators, land managers, and the general public. One remarkable realization in this domain is the Pl@ntNet mobile application (Affouard et al., 2017). It is used in an eponymous citizen science initiative (SciStarter, available at https://scistarter.com/project/16909-PlntNet) by a growing number of users around the world (more than 6 million downloads since 2013), and tens of thousands of plant pictures are submitted each day. Because a large fraction of this observation stream is geolocalized, it has great potential in terms of biodiversity monitoring and species distribution modeling (SDM). As the use of opportunistic data coming from citizen science initiatives has already been proven by Giraud et al. (2016) to strengthen the estimate of relative bird species abundance, we can expect other potential uses for such data types in a botanical context with Pl@ntNet.

Acquiring a large amount of opportunistic data still occurs at the expense of data quality and reliability, however. Many irrelevant pictures are submitted by the users of the Pl@ntNet application. This includes non-plant pictures, plant pictures of poor quality, or pictures of taxa that are not in the designated checklist (e.g., potted plants, ornamental and horticultural varieties, hybrids). Because the machine learning algorithm is not able to filter all of these pictures, many of them result in false positives (i.e., they are predicted as occurrences of species belonging to the checklist). Indeed, for a species automatically identified from a picture, two problems may induce identification error: (1) there is an intrinsic taxonomic uncertainty given the picture alone (i.e., it does not contain the discriminant visual pattern[s] that would make an expert certain about the exact species identification) or (2) the species was misidentified. Figure 1 illustrates typical examples of identification errors for *Acer monspessulanum* L. In Fig. 1B, one can see that the small symmetrical lobes at the base of the leaf might be confused with those of

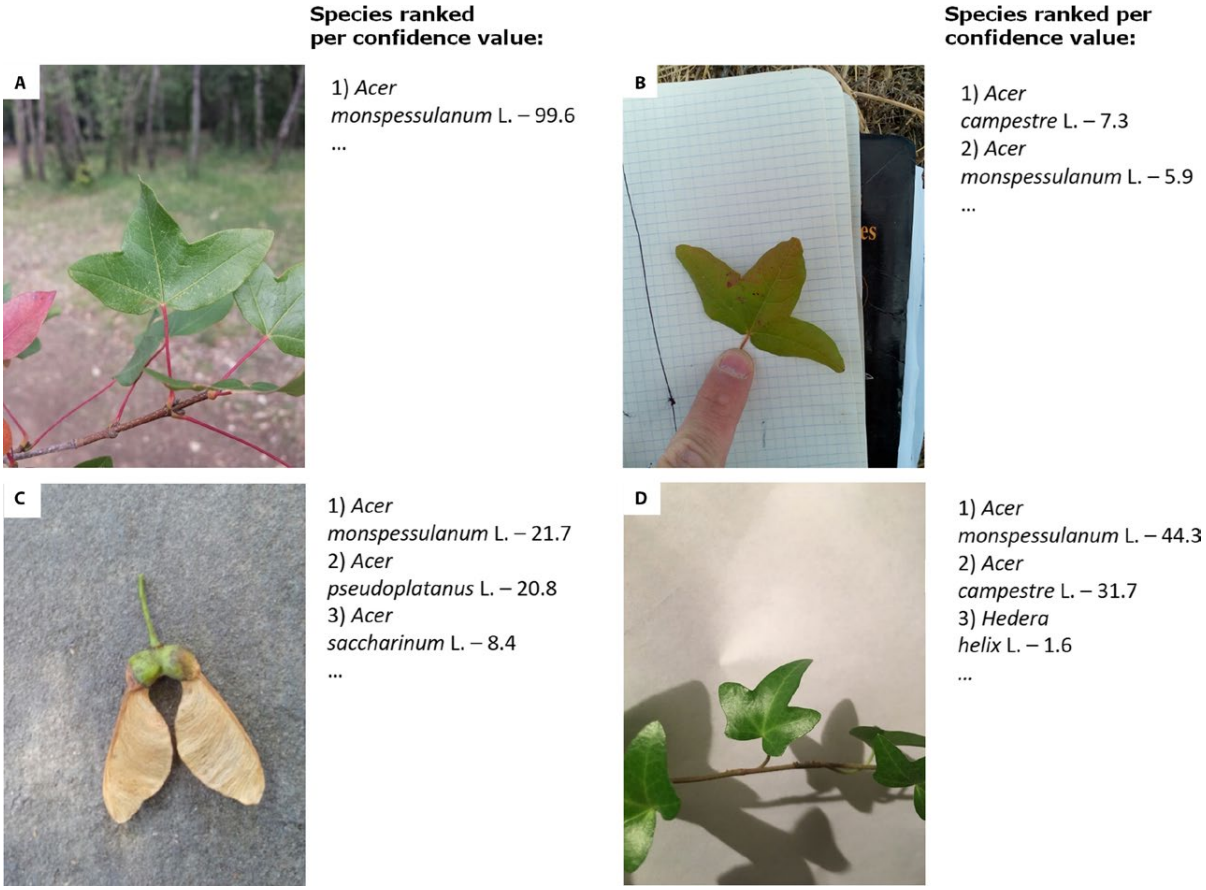


FIGURE 1. Four unvalidated Pl@ntNet plant pictures representing, or identified as, *Acer monspessulanum* and their respective predicted confidence values for the highest ranked species (the sum of scores over all species is always 100). (A) The species is *A. monspessulanum* and is well predicted. (B) The species is *A. monspessulanum*, but the model confounds it with *A. campestre*. (C) The species is *A. monspessulanum* or *A. pseudoplatanus*, but the species cannot be determined with the fruit only; there is an intrinsic taxonomic uncertainty. (D) The species is *Hedera helix* but is predicted as *A. monspessulanum* because this leaf is quite similar, as one can compare with (A).

 © 2018 Botella et al.

a young specimen of *A. campestre* L., which is probably the cause of the model uncertainty. Figure 1C well illustrates the problem of taxonomic uncertainty, as several species cannot be distinguished by the feature recorded in the observer's image where there is high proximity of the confidence values of the first two species. Finally, Fig. 1D shows a leaf of *Hedera helix* L. with three major lobes that have strong visual similarity to those of the *A. monspessulanum* leaf. Manually cleaning such large and noisy data streams is not possible. These problems imply that all species are not equal in their potential for automatic identification. There are several factors that make a species automatically identifiable from a photograph: the scale of the discriminant visual pattern (for example, there are many issues with the Poaceae family because discriminant features are often too small to be easily captured with a photograph), the visual saliency of the pattern compared to other species, and the temporality of the pattern due to the phenology of its organ.

In this article, we explore the possibility of exploiting automatically identified observations, without human validation, for SDM. Specifically, we study the impact of the degree of uncertainty of the retained occurrences when training the popular MAXENT niche modeling approach (Merow et al., 2013). Given the type of Pl@ntNet users, candidate species have to be automatically identifiable by non-expert observers who are often not familiar with the discriminant part of the plant that needs to be photographed. In addition, species that are visually similar in pictures must be avoided, and the chosen species must be well illustrated in the predictive model training database. In addition to these criteria that allow automatic species identification, we must take into account the requirements using SDM on presence-only data to acquire meaningful results. More precisely, the species must have contrasted environmental preferences regarding the study domain, its realized habitat must not be overly constrained by its dispersal capacity or important historical perturbations, and there must be enough observation points regarding the environmental variables considered.

Considering these constraints on species selection, the available data, and the potential use-cases, we applied our protocol to the modeling of the distribution of five species classified in major and moderate categories of invasion by the National Mediterranean Botanical Conservatory of Porquerolles for the southeastern region of France (Conservatoire botanique national méditerranéen de Porquerolles, 2018). Invasive species represent a major economic cost to our society (estimated at nearly €12 billion a year in Europe) and are one of the main threats to biodiversity conservation (Weber and Gut, 2004). The early detection of the appearance of these species is a key element in managing them and reducing the cost of such management. The analysis of Pl@ntNet data can provide a highly valuable response to this problem because the presence of these species is often correlated with that of human activity (and thus to the density of Pl@ntNet data occurrences), and the constant flow of observations enables annual monitoring of species distributions.

METHODS

Automatic species identification and the Pl@ntNet workflow

We first present the workflow of the Pl@ntNet system that yields automatically identified observations. To compute automatic species identification, we use a convolutional neural network (CNN). CNNs have been shown to considerably improve the accuracy of automated plant species identification compared to previous methods (Grinblat et al., 2016; Ghazi et al., 2017; Goëau et al., 2017). More generally, CNNs recently received much attention in the computer vision community because of the impressive performance they can achieve on a large variety of classification tasks. Details of the CNN architecture and of the training procedure we used in this study are provided in Appendix 1. The network was trained in a supervised manner on a set of 332,000 humanly validated plant images belonging to approximately 11,000 species and an additional rejection class (containing non-plant pictures taken by Pl@ntNet users, e.g., faces, animals, manufactured objects). These species cover a large part of the European and North African floras, according to the network of people initially involved in the production and validation of these data (this network was initiated with the Tela Botanica non-governmental organization [http://www.tela-botanica.org] and the network of French-speaking botanists, composed of professionals and amateurs). This data set also includes a few hundred species of common tropical plants from two tropical regions: the Indian Ocean region and tropical Amazonia. Data from these two regions were collected by scientists and engineers from research institutes and universities working on these flora, representatives of the Tela Botanica network in these regions, and Pl@ntNet users. The data validation process was conducted using the IdentiPlante web tool (http://www.tela-botanica.org/appli:identiplante), essentially dedicated to the Tela Botanica community, and was also accessible on the Pl@ntNet Android app. These applications display all botanical records shared by the project members. Logged-in users are able to provide new identifications, post comments, and vote on previous identifications. The revised data are regularly crawled by the visual search engine, which picks up observations considered correctly identified according to a predefined set of rules on the votes and on possible conflicts. These validation tools allow coverage of a growing number of species, from 800 in 2013 up to 11,000 in 2016.

Species distribution modeling using automatically identified Pl@ntNet observations

We performed SDM based on the unvalidated Pl@ntNet observations made in France in 2016. In total, the data represent approximately 2 million observations (most observations have only one image and some have up to five images). Each image $\mathbf{x}$ was passed to the CNN to receive an automated species prediction in the form of a categorical distribution $p(k|\mathbf{x})$ estimating the probability that the image $\mathbf{x}$ is from the k-th species (according to the softmax classification layer of the CNN). For the observations composed of several images, the predictions were simply averaged (i.e., $p(k|x) = 1/n_x \cdot \Sigma p(k|\mathbf{x}_i)$ for an observation $\mathbf{x}$ composed of n_x images $\mathbf{x}_i$). We then kept only the observations for which the most probable species (denoted as k_{max}) belonged to the set of the five potential invasive species considered in our study: *Acer negundo* L., *Carpobrotus edulis* (L.) N. E. Br., *Erigeron karvinskianus* DC., *Opuntia ficus-indica* (L.) Mill., and *Reynoutria japonica* Houtt. The resulting number of occurrences per species and per interval of confidence values $p(k_{max}|x)$ is provided in Fig. 2. For low values of $p(k_{max}|x)$, the level of noise is important (e.g., with several false positives for $p(k_{max}|x) < 30\%$). For the highest values of $p(k_{max}|x)$ (e.g., $p(k_{max}|x) > 95\%$), the level of noise is more reasonable but the number of occurrences is also much lower. Thus, to maximize SDM performance, one could expect a positive trade-off with an intermediate threshold.

 © 2018 Botella et al.

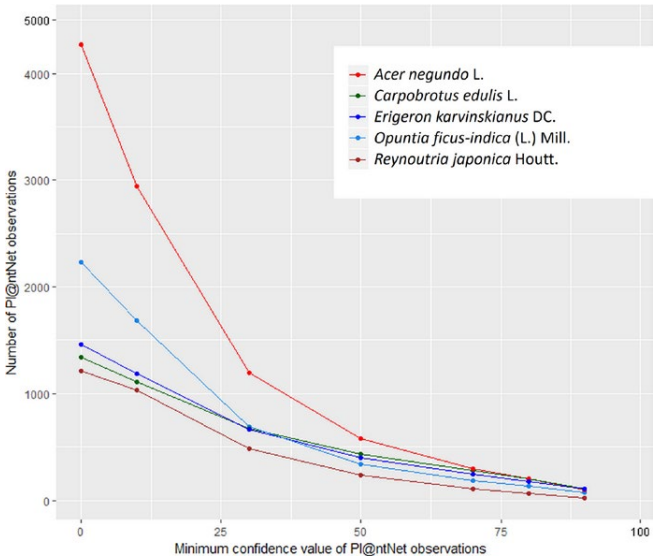


FIGURE 2. The number of Pl@ntNet observations per species and per confidence values $p(k_{max}|x)$.

To validate the species distribution models trained from automatically identified data, we used a second reference data set comprising count data collected and validated by French expert naturalists. This data set, referred to as Inventaire National du Patrimoine Naturel (INPN; https://www.gbif.org/dataset/75956ee6-1a2b-4fa3-b3e8-ccda64ce6c2d; Dutrève and Robert, 2016), comes from the Global Biodiversity Information Facility (https://www.gbif.org/). The underlying occurrences were collected in various contexts, including floras and regional catalogs, specific inventories, field notebooks, and surveys carried out by botanical conservatories. We kept only a subset of these data corresponding to the five invasive species considered in our study. The resulting data set contains 20,810 occurrences (see Table 1 for the detailed numbers per species) aggregated in 3242 quadrat cells of 100 km^2 distributed on a regular grid of 5175 quadrat cells covering the French territory.

Species distribution models were computed via MAXENT (Phillips et al., 2004, 2006), a popular environmental niche modeling method. In particular, we used the implementation of the *maxnet* (Phillips et al., 2017) R package that expands the input environmental variables with several functions (including linear, quadratic, threshold, hinge, and first-order interactions). Because we used presence-only SDM, we used pseudo-absence localities for model parameterization (see Appendix 2 for more details). MAXENT was computed on a set of 29 input environmental variables, including bioclimatic, pedological, topological, hydrographical, and land cover variables from CHELSA Climate data 1.1 (Karger et al., 2017), Consultative Group on International Agricultural Research–Consortium for Spatial Information (CGIAR-CSI) potential evapo-transpiration (ETP) data (Zomer et al., 2007, 2008), ESDBv.2 (Panagos, 2006; Van Liedekerke et al., 2006; Panagos et al., 2012), U.S. Geological Survey Digital Elevation data, the Institut National de l'information Géographique et

TABLE 1. Detailed number of occurrences in the Inventaire National du Patrimoine Naturel (INPN) data set by species.

Species name	No. of observations	No. of 100-km^2 areas
Acer negundo L.	5217	904
Carpobrotus edulis (L.) N. E. Br.	484	114
Erigeron karvinskianus DC.	711	306
Opuntia ficus-indica (L.) Mill.	120	44
Reynoutria japonica Houtt.	14,278	2623

TABLE 2. List and details of the environmental descriptors used in this study.

Name	Description	Nature	Values[a]	Local image
CHBIO_2	Mean monthly temp (max, min)	quanti.	[7.8, 21.0]	Yes
CHBIO_7	Temp. annual range	quanti.	[16.7, 42.0]	Yes
CHBIO_8	Mean temp. of wettest quarter	quanti.	[−14.2, 23.0]	Yes
CHBIO_9	Mean temp. of driest quarter	quanti.	[−17.7, 26.5]	Yes
CHBIO_10	Mean temp. of warmest quarter	quanti.	[−2.8, 26.5]	Yes
CHBIO_11	Mean temp. of coldest quarter	quanti.	[−17.7, 11.8]	Yes
CHBIO_13	Precip. of wettest month	quanti.	[43.0, 285.5]	Yes
CHBIO_14	Precip. of driest month	quanti.	[3.0, 135.6]	Yes
CHBIO_15	Precip. seasonality (CV)	quanti.	[8.2, 26.5]	Yes
CHBIO_18	Precip. of warmest quarter	quanti.	[19.8, 851.7]	Yes
CHBIO_19	Precip. of coldest quarter	quanti.	[60.5, 520.4]	Yes
etp	Potential evapotranspiration	quanti.	[133, 1176]	Yes
alti	Elevation	quanti.	[−188, 4672]	Yes
shade	Shade level	quanti.	[0, 1]	No
slope	Ground slope	quanti.	[0, 13457]	No
dmer	Distance to coastline	quanti.	[|0, 32767|]	No
droute	Distance to roads	quanti.	[|0, 32767|]	No
proxi_eau	<50 m to fresh water	bool.	{0, 1}	Yes
awc_top	Topsoil available water capacity	ordinal	{0, 120, 165, 210}	Yes
bs_top	Base saturation of the topsoil	ordinal	{35, 62, 85}	Yes
cec_top	Topsoil cation exchange capacity	ordinal	{7, 22, 50}	Yes
crusting	Soil crusting class	ordinal	[|0, 5|]	Yes
dgh	Depth to a gleyed horizon	ordinal	{20, 60, 140}	Yes
dimp	Depth to an impermeable layer	ordinal	{60, 100}	Yes
erodi	Soil erodibility class	ordinal	[|0, 5|]	Yes
oc_top	Topsoil organic carbon content	ordinal	{1, 2, 4, 8}	Yes
pd_top	Topsoil packing density	ordinal	{1, 2}	Yes
text	Dominant surface textural class	ordinal	[|0, 5|]	Yes
clc	Ground occupation	categ.	[|1, 48|]	Yes

Note: bool. = Boolean data; categ. = categorical data; CV = coefficient of variation of monthly precipitation; quanti. = quantitative data.

[a]Data presented in curly brackets ({ }) contain the list of all possibles values of the variable, i.e., a discrete ensemble; square brackets ([ ]) indicate the continuous range of values that can take the variable, i.e., a continuous interval; vertical lines indicate the range of integers between the two bounds given, i.e., a discrete interval.

 © 2018 Botella et al.

forestière–Système d'Administration Nationale des Données et Référentiels sur l'Eau (IGN-SANDRE) BD Carthage, CORINE Land Cover 2012 data, and IGN ROUTE500 data. The detailed methodology of how these variables were collected and formatted is described in Appendix 3. The full list of the variables used is presented in Table 2. For each of the considered species, we computed seven models with varying levels of minimal confidence of species occurrences, i.e., different threshold values $p_{min}(k_{max}|x)$ of the categorical probability $p(k_{max}|x)$. We know that the global sampling effort in Pl@ntNet is highly correlated with human population density and the proximity to roads and to the coastline. In our study, the sampling intensity was so high compared to the species abundance that we strongly overestimated the species abundance in cities, on beaches, and on roads. Consequently, we fitted MAXENT models, including variables of urban areas, proximity to roads, and distance to the coastline. In the predicted abundance function, we then kept these variables constant across space to cancel the effect of the sampling effort (see Appendix 2 for more details). This approach has already been proposed and successfully used in the literature of SDMs (Warton et al., 2013; Stolar and Nielsen, 2015). The predictive effectiveness of the models was then assessed using the INPN count data as a validation set. We used two evaluation metrics: (1) the true skills statistics (TSS) equal to the sum of the sensitivity and the specificity minus one (as described in Allouche et al., 2006), and (2) the accuracy on 10% densest quadrats (A10DQ; see Appendix 2 for more details). The TSS is the sum of sensibility and specificity minus one when comparing the SDM predicted presences/absences of a species with the references (the INPN data set). It is a meaningful measure to evaluate the model's ability to detect presences while simultaneously minimizing false positives. It is computed through binarization of SDM continuous prediction based on the threshold that maximizes the TSS. We chose the A10DQ as a complementary metric because it evaluates the accuracy of the models in predicting the quadrats with the highest abundance (INPN count), which is an especially interesting property from the perspective of invasive species management.

RESULTS

Figure 3 displays the evaluation metrics as a function of the confidence threshold $p_{min}(k_{max}|x)$ applied to filter the automatic predictions. We found that the confidence threshold had variable influence depending on the species, but there was an overall trend represented by the average curve (Fig. 3, black solid line). Too-low thresholds did not allow for filtering identification errors sufficiently, thus the model was biased by the presence of too many irrelevant occurrences. A too-high threshold (above 70%) also degraded the model performance (in particular, the accuracy of the quadrat cells with the higher level of counts; see Fig. 3) because the number of retained occurrences in the training set decreased significantly with increasing threshold. Models based on too few occurrences could not provide a relevant prediction of species distribution. With the current Pl@ntNet data, the chosen species, and the variables, a confidence threshold of 70% represented a good compromise for SDM. It filtered identification errors effectively for most species while retaining enough occurrences for model training. The most problematic species was *Reynoutria japonica*: it had very poor TSS for all thresholds (a TSS score of 0 would be a random prediction of presence and absence), indicating that the SDM did not distinguish presence and absence zones very well. This species is the most widespread, which leads to poor SDM performances. Nevertheless, for the best threshold, A10DQ showed that 20% of the densest INPN quadrats were predicted by the model fitted on Pl@ntNet, which is significantly better than a random ranking of quadrats (which would give an average of 10% and a standard deviation of 1.3%). Consequently, the model could capture information on the distribution of *Reynoutria* from the Pl@ntNet data. Conversely, very good results were obtained for both metrics for *Opuntia ficus-indica* and *Carpobrotus edulis*.

Figure 4 further shows the distributions predicted for each species using $p_{min}(k_{max}|x) = 70\%$. For comparison, we also displayed the expert count data of INPN, as well as the specificity and sensitivity of our model measured with that data (at TSS max). Most regions with high INPN counts were reasonably well predicted by

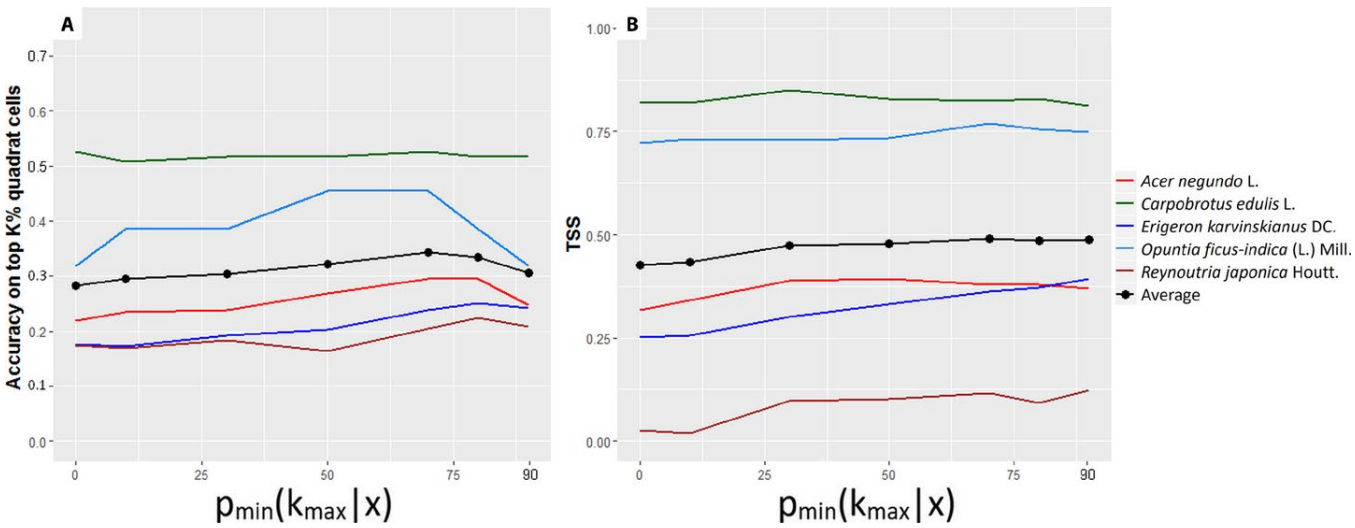


FIGURE 3. Predictive effectiveness of the species distribution models trained on Pl@ntNet data as a function of the confidence threshold value $p_{min}(k_{max}|x)$ showing accuracy on the 10% densest quadrats (A) and true skill statistics (TSS; conversion of prediction value into presence/absence with the threshold that maximizes TSS) (B).

 © 2018 Botella et al.

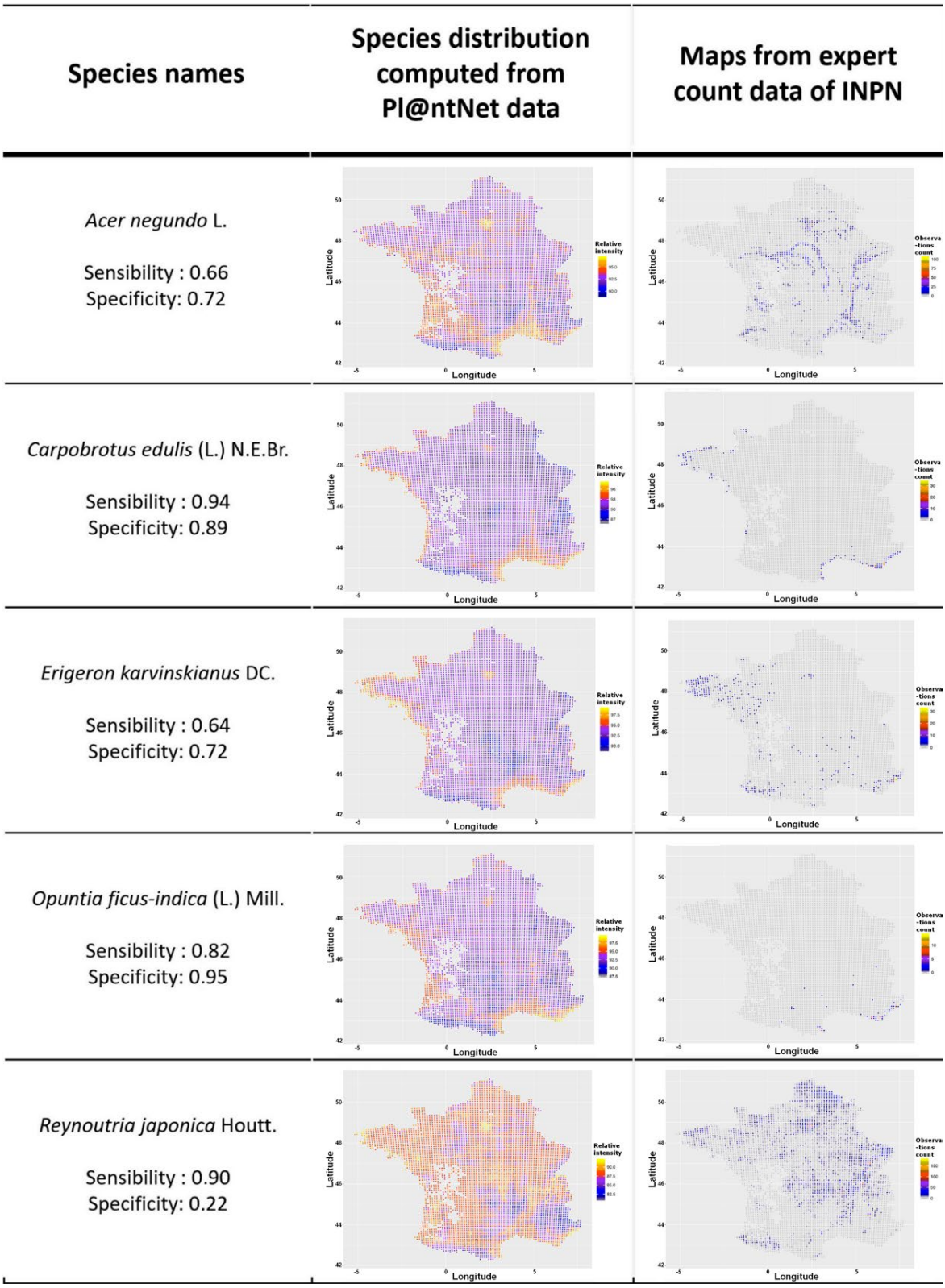


FIGURE 4. Maps of species distribution models computed from Pl@ntNet data (based on $p_{min}(k_{max}|x)$ = 70%) and of expert count data from the Inventaire National du Patrimoine Naturel (INPN). The sensibility and specificity used for the computation of the true skill statistics (for pmin(kmax|x) = 70) is provided for each species.

 © 2018 Botella et al.

the models. Accordingly, sensitivity values were generally accurate for most species. Nevertheless, there were also regions for which the Pl@ntNet model and INPN data disagreed; in these regions the Pl@ntNet model predicted high abundances but there were none or very few occurrences in the INPN data. The strongest disagreement occurred for *Reynoutria japonica*, i.e., the taxon for which the specificity was the lowest. Other false-positive prediction regions included the west coast for *Opuntia ficus-indica* and *Carpobrotus edulis* and the "Golfe du Lion" (arc on the southeast coast) for *O. ficus-indica* and *Erigeron karvinskianus*.

DISCUSSION

Visual inspection of Pl@ntNet observations occurring in such false-positive regions revealed that for the vast majority such observations did not correspond to erroneous identifications ($p_{min}(k_{max}|x) = 70\%$ is a high enough threshold to remove noise efficiently). Rather, they corresponded to real occurrences that can be classified in three main categories (see Fig. 5 for examples of observations belonging to the different categories). The first category can be qualified as cultivated specimens, i.e., specimens planted and/or maintained by humans such as gardening plants, house plants, ornamental plants in city parks, etc. Most occurrences of *Opuntia ficus-indica* on the west coast belonged to this category. A second category of observations could be qualified as casual invasive specimens, i.e., isolated specimens that often flourish close to human construction but that do not form self-replacing populations. Cultivated and casual invasive specimens present in the observations reveal that the species is able to grow in a great diversity of habitats. These specimens should be identified, either to (1) filter them for model learning, (2) evaluate the correlation between species gardening intensity and its abundance in wild surroundings, or (3) learn more complex models that integrate dispersal mechanisms and quantify more precisely the importance of gardening intensity on the species' capacity to colonize a region. To identify cultivated specimens, several options are possible: for example, learning models can be used to identify the context of the picture or the user can be asked to clarify the type of environment where the observation was made, especially when observations appear ambiguous. Apart from the issue of correctly predicting species occurrences in the wild, frequent occurrences of cultivated and casual invasive specimens in a region where there is no presence in the wild can reflect the risk of future invasion in the wild.

A last category of observations can be qualified as newly inventoried invasive specimens, i.e., non-isolated specimens living in natural areas that have yet to be inventoried in the INPN data. Notably, the majority of occurrences of *Carpobrotus edulis* on the west coast belong to this category. Newly inventoried invasive specimens could provide an early warning for territory managers. For

FIGURE 5. Pl@ntNet observations with a species prediction score of more than 70% for plants living in natural conditions or cultivated for ornamental purpose.

 © 2018 Botella et al.

example, we found newly inventoried specimens of *Reynoutria japonica* in the Pl@ntNet data, and we suspect that poor performance of its SDM could reflect a negative bias in the evaluation metrics of this species. Typically, specimens occurring outside of presence areas identified by experts and not categorized as cultivated or casual invasive should be prioritized for expert validation.

In this study, our sampling effort correction approach was based on prior knowledge of sampling intensity in the Pl@ntNet data. We could not evaluate the errors related to the sampling effort bias without complementary systematic survey data. Nevertheless, the INPN data have their own heterogeneity in the spatial distribution of the sampling effort. These data were collected by independent regional conservatories, and variations in sampling by different workforces may have introduced regional heterogeneity. Furthermore, some zones are not surveyed by conservatories, typically cities in most cases, which tends to bias the Pl@ntNet model error in urban areas. The study of global sampling effort bias is crucial for exploiting presence-only data collected without protocol. The spatially heterogeneous sampling effort is especially problematic when it is correlated with environmental variables impacting the species distribution. For example, the sampling effort is correlated with the distance to the coastline, which is also a variable influencing the abundance of *Opuntia ficus-indica*, *Erigeron karvinskianus*, and *Carpobrotus edulis*. Because our bias correction method removes the distance to the coastline effect, it partially removes the ability of the model to capture this effect on the species distribution. When we included these variables in the predicted distribution of the three species (results not presented in this article), we found a much greater predicted abundance gradient toward the coast. However, the maps presented in Fig. 4 show that the model captured a part of the coastal effect through other variables that are correlated with the distance to coastline. The same problem will occur with other invasive species that tend to grow near roads as a result of constant perturbation or dispersal mechanisms. More generally, we note that the presence of invasive species is strongly influenced by human activity. It is also highly correlated with observational intensity in opportunistic presence-only data. Thus, this category of species represents a major methodological challenge for improving SDM based on presence-only data and represents a clear path for future research.

CONCLUSIONS

This study is the first to evaluate the potential of automated identification of opportunistic plant observations for modeling species distributions. The described methodology allowed us to analyze the potential usefulness of the Pl@ntNet data. By comparing SDMs trained on Pl@ntNet unvalidated observations with validated independent count data on a large spatial scale, we found that the data are rich enough to be used for SDM with only a single year of data collection. However, we also showed that distributions reported from Pl@ntNet data do not precisely match those of expert data. The main reasons for these deviations appear to be the presence of cultivated or casual invasive specimens in the data set, the detection of real presence in new areas, and the limits of the sampling bias correction method. Noticing these limits allowed us to underline significant research challenges for SDMs and to provide possible methods to usefully integrate information provided by opportunistic citizen science observations into conservation management.

ACKNOWLEDGMENTS

The authors thank the Inventaire National du Patrimoine Naturel (INPN) and the Fédération française des Conservatoires botaniques nationaux for access to the expert count data used in this study. We also would like to thank the Tela Botanica and Pl@ntNet community who have contributed to produce and revise the data used in this study.

LITERATURE CITED

Affouard, A., H. Goëau, P. Bonnet, J. C. Lombardo, and A. Joly. 2017. Pl@ntnet app in the era of deep learning. *In* 5th International Conference on Learning Representations, 24–26 April 2017, Toulon, France.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223–1232.

Carranza-Rojas, J., H. Goëau, P. Bonnet, E. Mata-Montero, and A. Joly. 2017. Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology* 17: 181.

Casanova, D., J. J. de Mesquita Sá Junior, and O. M. Bruno. 2009. Plant leaf identification using Gabor wavelets. *International Journal of Imaging Systems and Technology* 19: 236–243.

Champ, J., T. Lorieul, P. Bonnet, N. Maghnaoui, C. Sereno, T. Dessup, J. M. Boursiquot, et al. 2016. Categorizing plant images at the variety level: Did you say fine-grained? *Pattern Recognition Letters* 81: 71–79.

Conservatoire botanique national méditerranéen de Porquerolles. 2018. Espèce végétale exotique envahissante (EVEE) [online]. Website http://www.invmed.fr/src/listes/index.php?idma=33 [accessed 31 January 2018].

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *In* Proceedings, Institute of Electrical and Electronics Engineers (IEEE) Computer Society Conference on Computer Vision and Pattern Recognition, 20–25 June 2009, 248–255. IEEE, Piscataway, New Jersey, USA.

Dutrève, B., and S. Robert. 2016. INPN (Inventaire National du Patrimoine Naturel): Données flore des Conservatoires botaniques nationaux (CBN) agrégées par la Fédération des Conservatoires botaniques nationaux (FCBN). Version 1.1. Service du Patrimoine naturel (SPN), Muséum national d'Histoire naturelle, Paris, France. Occurrence Dataset https://doi.org/10.15468/omae84 via GBIF.org [accessed 30 August 2017].

Fithian, W., and T. Hastie. 2013. Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics* 7: 1917.

Gaston, K. J., and M. A. O'Neill. 2004. Automated species identification: Why not? *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 359: 655–667.

Ghazi, M. M., B. Yanikoglu, and E. Aptoula. 2017. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* 235: 228–235.

Giraud, C., C. Calenge, C. Coron, and R. Julliard. 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics* 72: 649–658.

Goëau, H., P. Bonnet, and A. Joly. 2016. Plant identification in an open-world (LifeCLEF 2016). *In* K. Balog, L. Cappellato, N. Ferro, and C. Macdonald [eds.], Working notes of CLEF 2016–Conference and labs of the evaluation forum, 5–8 September 2016, Évora, Portugal. *CEUR Workshop Proceedings* 1609: 428–439.

Goëau, H., P. Bonnet, and A. Joly. 2017. Plant identification based on noisy web data: The amazing performance of deep learning (LifeCLEF 2017). *In* L. Cappellato, N. Ferro, L. Goeuriot, and T. Mandl [eds.], Working notes of CLEF 2017–Conference and labs of the evaluation forum, 11–14 September 2017, Dublin, Ireland. *CEUR Workshop Proceedings* 1866: ceur-ws.org/Vol-1866/invited_paper_9.pdf.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep learning, vol 1. MIT Press, Cambridge, Massachusetts, USA.

 © 2018 Botella et al.

Grinblat, G. L., L. C. Uzal, M. G. Larese, and P. M. Granitto. 2016. Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture* 127: 418–424.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In* Proceedings, IEEE Computer Society Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015, 1026–1034. IEEE, Piscataway, New Jersey, USA.

Ioffe, S., and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In* Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015. *PMLR* 37: 448–456.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *In* Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, Florida, USA, 3–7 November 2014, 675–678. ACM, New York, New York, USA.

Joly, A., P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, et al. 2016. A look inside the Pl@ntNet experience. *Multimedia Systems* 22: 751–766.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, et al. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data* 4: 170122.

Lee, S. H., C. S. Chan, P. Wilkin, and P. Remagnino. 2015. Deep-plant: Plant identification with convolutional neural networks. *In* Proceedings, Institute of Electrical and Electronics Engineers (IEEE) International Conference on Image Processing (ICIP), Macau, China, 16–18 September 2015, 452–456. IEEE, Piscataway, New Jersey, USA.

Merow, C., M. J. Smith, and J. A. Silander. 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* 36: 1058–1069.

Panagos, P. 2006. The European soil database. *GEO: Connexion* 5: 32–33.

Panagos, P., M. Van Liedekerke, A. Jones, and L. Montanarella. 2012. European Soil Data Centre: Response to European policy support and public data requirements. *Land Use Policy* 29: 329–338.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. *In* Proceedings of the Twenty-First International Conference on Machine Learning, New York, New York, USA, 10–14 June 2004, p. 83. ACM Digital Library, New York, New York, USA.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.

Phillips, S. J., R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. 2017. Opening the black box: An open-source release of Maxent. *Ecography* 40: 887–893.

Stolar, J., and S. E. Nielsen. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions* 21: 595–608.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, et al. 2015. Going deeper with convolutions. *In* Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 7–12 June 2015, 1–9. IEEE, Piscataway, New Jersey, USA.

Van Liedekerke, M., A. Jones, and P. Panagos. 2006. ESDBv2 Raster Library: A set of rasters derived from the European Soil Database distribution v2. 0. European Commission and the European Soil Bureau Network, CDROM, EUR, 19945.

Wäldchen, J., and P. Mäder. 2017. Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-016-9206-z.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS One* 8: e79168.

Weber, E., and D. Gut. 2004. Assessing the risk of potentially invasive plant species in central Europe. *Journal for Nature Conservation* 12: 171–179.

Wilf, P., S. Zhang, S. Chikkerur, S. A. Little, S. L. Wing, and T. Serre. 2016. Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences USA* 113: 3305–3310.

Yanikoglu, B., E. Aptoula, and C. Tirkaz. 2014. Automatic plant identification from photographs. *Machine Vision and Applications* 25: 1369–1383.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? *In* Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger [eds.], Proceedings of Neural Information Processing Systems (NIPS 2014), Montréal, Canada, 8–13 December 2014. *Advances in Neural Information Processing Systems* 27: 3320–3328.

Zomer, R. J., D. A. Bossio, A. Trabucco, L. Yuanjie, D. C. Gupta, and V. P. Singh. 2007. Trees and water: Smallholder agroforestry on irrigated lands in Northern India. IWMI Research Report 122. International Water Management Institute (IWMI), Colombo, Sri Lanka.

Zomer, R. J., A. Trabucco, D. A. Bossio, and L. V. Verchot. 2008. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems and Environment* 126: 67–80.

APPENDIX 1. Detailed architecture and training procedure of the convolutional neural network used to compute the automated identifications.

The main strength of convolutional neural network (CNN) technologies comes from their ability to learn discriminant visual features directly from the raw pixels of the images without exponentially increasing the model variables as the dimensionality grows (Goodfellow et al., 2016). This is achieved by stacking multiple convolutional layers, i.e., the core building blocks of a CNN. In general, a convolutional layer takes images as input and produces as output feature maps corresponding to different convolution kernels while looking for different visual patterns.

To get to specific choices in the architecture, we used an extended version of the GoogleNet model (Szegedy et al., 2015) that is a very deep CNN that stacks several so-called inception layers. As in Carranza-Rojas et al. (2017), we extended the base version with batch normalization (Ioffe and Szegedy, 2015), which has been proven to speed up convergence and limit overfitting, and with a parametric rectified linear unit (PReLU) activation function (He et al., 2015) instead of the traditional rectified linear unit (ReLU).

To improve the generalization ability of the network, we used transfer learning, which is a powerful paradigm to overcome the lack of sufficient domain-specific training data. Deep learning models have to be trained on thousands of pictures per class to converge on accurate classification models. It has been shown that the first layers of deep neural networks deal with generic features (Yosinski et al., 2014) so that they are generally usable for other computer vision tasks. Consequently, they can be trained on arbitrary training image data. The last layers contain more or less generic information transferable from one classification task to another. These layers are expected to be more informative for the optimization algorithm than a random initialization of the weights of the network. Therefore, a common practice is to initialize the network by pre-training it on a large available data set and then fine-tune it on the scarcer domain-specific data. Many networks are pre-trained on the generalist data set ImageNet (Deng et al., 2009), which covers a large variety of visual concepts, including animals, vehicles, and manufactured objects. Because the GoogleNet model we used was already pre-trained on this generalist data set, we used the following methodology for fine-tuning it on our data set of 11,000 species (using the Caffe framework [Jia et al., 2014]):

 © 2018 Botella et al.

1. The linear classification layer was replaced by a new one aimed at classifying the new classes (i.e., the 11,000 species). It was initialized with random weights and the learning rate was multiplied by 10 for this layer.
2. The other layers were kept unchanged to initialize the network with the weights learned from ImageNet.
3. The network was trained on the 332,000 plant images of our training set.

A batch size of 16 images was used for each iteration, with a learning rate of 0.0075 with images of 224 × 224 resolution. Simple crop and resize data augmentation was used with the default settings of the Caffe framework.

APPENDIX 2. Description of Pl@ntNet data post-treatments, generation of quadrature points, and experimental procedure. Results were obtained using R.

Filtering of Pl@ntNet geolocated observations: We used the unvalidated observations collected by Pl@ntNet users during the year 2016. We kept only observations for which one of our five species was ranked first according to the identification score. We first selected those whose GPS geolocation falls in the French Metropolitan territory (polygon: getData(country="FRA",level=0), function from package *raster*) excluding Corsica, or are closer than 500 m to the coastline (because of coordinate error). Because observations are very often duplicated due to a repeated submission of the same set of pictures, we kept only one of the identical observations. Unsatisfactory automatic identification of the same specimen allowed the user to take new pictures of the specimen and submit it again. This kind of duplication was removed by the following procedure: for two occurrences closer than 60 sec in time and 100 m in space, we kept the one with highest $p(k_{max}|x)$.

Quadrature points: MAXENT can be interpreted as a non-homogeneous Poisson process model (Fithian and Hastie, 2013). Thus, computing a MAXENT model from observations requires integration of its intensity function over the spatial domain of study D (in this study, the French territory). For this purpose, it approximates the integral with quadrature points, also called "pseudo-absences," that represent the distribution of the environmental descriptors on D. As our domain was wide, and some of our descriptors vary with high spatial frequency (like distance to roads or proximity to fresh water), we used a high number of quadrature points. We generated 101,632 points on a grid with a similar spacing of 0.025 in longitude (approximately 2 km) and latitude (approximtely 2.8 km), and strictly included in the French polygon (see above).

Prediction of model relative abundance for a plot and attribution of quadrature points to plots: With a fitted MAXENT model, we can evaluate its intensity function at every quadrature point via environmental descriptors, which gives a high-resolution map of predicted relative abundance across France. This fine-resolution prediction includes the effect of high-frequency variables. However, to compare model predictions to counts on quadrat cells, we need to upscale our prediction: according to the properties of the inhomogeneous Poisson process, the law of the number of points falling in a quadrat cell is a Poisson law whose parameter is the integral of the intensity function over the quadrat cell. Because the quadrature points are regularly spaced, we can approximate this integral up to a factor (common to every quadrat cell because they have the same area) with the mean of intensity values over quadrature points contained in the quadrat cell. For some cells located mainly above sea or ocean, some did not contain any quadrature points, thus we attributed the closest one while removing it from its original plot. In this way, quadrat cells contained an average of 17.1 quadrature points.

Bias-corrected model prediction: We know that there is sampling bias in the Pl@ntNet observation data. The most important is high sampling effort in cities, close to roads, and near coastlines (because of use during tourist activities). In addition, we know that for the species of interest, distance to roads and cities has no strong link to real abundance. Because we want to remove the artificial importance of those variables in the concentration of observations, one strategy is to integrate the sampling variables in the intensity function, as is now commonly done in such cases (Warton et al., 2013). If there is no perfect linear link between sampling and abundance variables, we will correctly infer our abundance model. Finally, we predict an unbiased relative abundance by setting the sampling variables to a constant value everywhere in space. However, we cannot do this for the distance to coastline because this variable plays a key role in the real abundance of *Carpobrotus edulis*, *Opuntia ficus-indica*, and *Erigeron karvinskianus*.

Evaluation metric: The evaluation metric represents the proportion of the top 10% quadrats in terms of real count that are also in the top 10% in terms of model prediction. However, we have to define the last quadrat cell ranked in the top 10% for counts, which is problematic for some species because of ex aequo cells. That is why we defined the following procedure that is adjusted for each species in the percentage of top cells such that the metrics can be calculated and the percentage is the closest to 10%. It is known as accuracy on the 10% densest quadrats (A10DQ):

$$\frac{N_{p\&c}(i)}{N_c(i)}$$

Where $N_{p\&c}(i)$ is the number of cells that are contained in the $N_c(i)$ higher cells both in terms of count and of model prediction.

Calculation of $N_c(i)$: We order the cells by decreasing the count of **i** and note C_k the count of the k-th cell in this order. As we are interested in the quadrat cells ranked in the highest 10%, if $C_{518} > C_{519}$, we set $N_c(i) = 518$. Otherwise, $C_{518} = C_{519}$ (ex aequo exists for 518th position), then we note *sup* the position of the last cell with count C_{519} and *inf* the position of the first cell with count C_{519}. The chosen rule is to take $N_c(i)$ such that $N_c(i) = \mathrm{Min}(|sup\text{-}518|, |inf\text{-}518|)$.

APPENDIX 3. Detailed methodology of how environmental variables were collected and formatted in our study.

We used data covering the French metropolitan territory, freely available on the web. The environmental descriptors are listed in Table 2. Because the original coordinate systems of the layers used varied among sources, we systematically converted them to WGS84

 © 2018 Botella et al.

using the *rgdal* package in R, which was the reference coordinate system for our observations, quadrature points, and quadrat cells. In the following points, we describe the sources, nature, and eventual transformations of those environmental data:

- CHELSA Climate data 1.1: These are raster data with worldwide coverage and 1-km resolution. A mechanistic climatic model is used to make spatial predictions of monthly mean-max-min temperatures, mean precipitations, and 19 bioclimatic variables that are downscaled with statistical models integrating historical measures of meteorologic stations from 1979 to the present (see Karger et al., 2017). The data are under Creative Commons Attribution 4.0 International License (available at http://chelsa-climate.org/downloads/).
- The ESDB v2, 1kmx1km Raster Library (Panagos, 2006; Van Liedekerke et al., 2006; Panagos et al., 2012): The library contains multiple soil pedological descriptor raster layers covering Eurasia at a resolution of 1 km. We selected 10 descriptors from the library. They represent quantitative physico-chemical quantities of the soil (from the PedoTransfer Rules Database [PTRDB attributes, available at https://esdac.jrc.ec.europa.eu/content/ptrdb-attributes]) that have been deduced from soil classification with expert rules, and their values are aggregated in intervals. As there are few possible intervals by variables (2–6), we integrated them as categorical variables in MAXENT. The data are maintained and distributed freely for scientific use by the European Soil Data Centre at http://eusoils.jrc.ec.europa.eu/content/european-soil-database-v2-raster-library-1kmx1km.
- CORINE Land Cover 2012, version 18.5.1, 12/2016: This is a raster layer describing soil occupation with 48 categories across Europe (25 countries) at a resolution of 100 m. This classification is the result of an interpretation process applied to the earth's surface with high-resolution satellite images. We set this variable as categorical in MAXENT with only 30 relevant categories for our purposes. This database of the European Union is freely accessible online at: http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012.
- CGIAR-CSI ETP data: The Consultative Group on International Agricultural Research–Consortium for Spatial Information (CGIAR-CSI) distributes this worldwide monthly potential evapo-transpiration raster data. It is pulled from a model developed by Antonio Trabucco (Zomer et al., 2007, 2008). Rasters are estimated by the Hargreaves formula using mean monthly surface temperatures and standard deviation from WorldClim 1:4 (http://www.worldclim.org/version1), and radiation on top of atmosphere. The raster is at a 1-km resolution and is freely downloadable for a nonprofit use at http://www.cgiar-csi.org/data/global-aridity-and-pet-database#description.
- U.S. Geological Survey Digital Elevation data: The Shuttle Radar Topography Mission achieved in 2010 by the Endeavour shuttle measured digital elevation at 3 arcs per second resolution over most of the earth's surface. Raw measures have been post-processed by the National Aeronautics and Space Administration and the National Geospatial-Intelligence Agency to correct detection anomalies. This gives a precision measurement of approximately 90 m for this variable. The data are available from the U.S. Geological Survey and are downloadable on the EarthExplorer (https://earthexplorer.usgs.gov/). See https://lta.cr.usgs.gov/SRTMVF for more information.
- BD Carthage v3: BD Carthage is a spatial database holding information on the structure and nature of the French Metropolitan hydrological network. We focus on the geometric segments representing watercourses, polygons representing hydrographic fresh surfaces, and the ocean. The data have been produced by the Institut National de l'information Géographique et forestière (IGN) from an interpretation of the BD Ortho IGN. The database is maintained by SANDRE under free license for non-profit use and is downloadable at: http://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FXX/2014/arcgis/.

For "proxi_eau," i.e., the proximity to fresh water, we used QGIS (https://qgis.org/) to rasterize to a 12.5-m resolution, with a buffer of 50 m, (1) the shapefile COURS_D_EAU.shp and (2) the polygons of SURFACES_HYDROGRAPHIQUES.shp with attribute NATURE="Eau douce permanente". We then created the maximum of the proximity raster derived from COURS_D_EAU.shp and SURFACES_HYDROGRAPHIQUES.shp (so the value of 1 corresponds to an approximate distance of less than 50 m to a watercourse or hydrographic surface of fresh water). For "dmer," i.e., the distance to the ocean, we calculated, using QGIS, the distance raster at a resolution of 12.5 m to polygons with attribute TYPE="Pleine mer" in the shapefile SURFACES_HYDROGRAPHIQUES.shp of BD Carthage up to a distance of 32,767 m for storage format convenience.

- ROUTE500 1.1: This database register classifies road linkages between cities (highways, national roads, and departmental roads) in France in shapefile format, representing approximately 500,000 km of roads. It is produced under free license (all uses) by the IGN. Data are available online at http://osm13.openstreetmap.fr/~cquest/route500/. For deriving the variable "droute," the distance to the main roads networks, we used a similar procedure as for "dmer," calculating the distance raster for all the elements of the shapefile ROUTES.shp (segments).

 © 2018 Botella et al.

6 Chapter 2: Bias in presence-only niche models related to sampling effort and species niches: lessons for background points selection

PLOS ONE

RESEARCH ARTICLE

Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection

Christophe Botella[1,2,3,5☯*], **Alexis Joly**[1☯], **Pascal Monestiez**[5☯], **Pierre Bonnet**[3,4☯], **François Munoz**[6☯]

1 INRIA Sophia-Antipolis - ZENITH team, Montpellier, France, **2** INRAE, UMR AMAP, Montpellier, France, **3** AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France, **4** CIRAD, UMR AMAP, Montpellier, France, **5** INRAE, BioSP, Avignon, France, **6** Université Grenoble Alpes, Laboratoire d'Ecologie Alpine, CS 40700, Grenoble, France

☯ These authors contributed equally to this work.
* christophe.botella@gmail.com

Check for updates

OPEN ACCESS

Citation: Botella C, Joly A, Monestiez P, Bonnet P, Munoz F (2020) Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. PLoS ONE 15(5): e0232078. https://doi.org/10.1371/journal.pone.0232078

Editor: Mirko Di Febbraro, University of Molise, Isernia, ITALY

Received: February 20, 2019

Accepted: April 7, 2020

Published: May 20, 2020

Copyright: © 2020 Botella et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data analyzed in the study fully comes from simulations, which can be reproduced with a script provided on a Github repository (https://github.com/ChrisBotella/UB-and-TGOB). It is referenced in the manuscript text.

Funding: The author(s) received no specific funding for this work other than the unique funding sources from the Funding Information section. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The use of naturalist mobile applications have dramatically increased during last years, and provide huge amounts of accurately geolocated species presences records. Integrating this novel type of data in species distribution models (SDMs) raises specific methodological questions. Presence-only SDM methods require background points, which should be consistent with sampling effort across the environmental space to avoid bias. A standard approach is to use uniformly distributed background points (UB). When multiple species are sampled, another approach is to use a set of occurrences from a Target-Group of species as background points (TGOB). We here investigate estimation biases when applying TGOB and UB to opportunistic naturalist occurrences. We modelled species occurrences and observation process as a thinned Poisson point process, and express asymptotic likelihoods of UB and TGOB as a divergence between environmental densities, in order to characterize biases in species niche estimation. To illustrate our results, we simulated species occurrences with different types of niche (specialist/generalist, typical/marginal), sampling effort and TG species density. We conclude that none of the methods are immune to estimation bias, although the pitfalls are different: For UB, the niche estimate fits tends towards the product of niche and sampling densities. TGOB is unaffected by heterogeneous sampling effort, and even unbiased if the cumulated density of the TG species is constant. If it is concentrated, the estimate deviates from the range of TG density. The user must select the group of species to ensure that they are jointly abundant over the broadest environmental sub-area.

1 Introduction

Species Distribution Models (SDM) ([1]) based on presence-only data are widely used to characterize the ecological niches and distributions of animal and plant species across

Competing interests: The authors have declared that no competing interests exist.

environments and space, for ecological studies and conservation planning. Popular examples of such methods include ENFA ([2]), GARP ([3]), Maxent ([4]) and more recently Bayesian methods ([5, 6]). Large amounts of presence-only data have become available through the digitization of herbarium collections ([7, 8]) and the development of citizen science, and they should improve estimation accuracy in SDM. However, sampling effort is heterogeneous and often depends on environment, yielding estimation biases in SDM ([9]). These biases are not alleviated when increasing occurrence data and require the development of methods acknowledging sampling heterogeneity.

While first presence-only SDM methods like BIOCLIM ([10]) and DOMAIN ([11]) aimed at computing environmental ranges where the species could live, recent methods ([12]) look for more accuracy, and estimate the species density across environment. This density is proportional to the species expected abundance regarding only the environment. To estimate this species environmental density, such methods use a set of "background" or "pseudo-absences" points (or "quadrature" points in literature on Poisson process models, see [12], which should reflect the sampling intensity across the environmental space. Background points are usually drawn uniformly over the region, assuming a uniform sampling of the focal species distribution (default option in Maxent). However, this assumption is inadequate in most cases. Indeed, the occurrences are mostly collected without a strict sampling protocol. People visit more certain places than others, e.g. because they are closer from where they live, easier to access, biologically interesting, or aesthetically attractive. This geographic bias translates into an environmental bias, i.e. the global sampling effort that is induced by the sum of observers covaries with the environment. For instance, Fig 1 shows the that distribution of opportunistic observations of the mobile app Pl@ntNet in 2017 ([13]) is higher in lower-elevation areas. For a species specialized to mountain ecosystems, small populations at lower elevation could be over-sampled. When inferring an SDM with a uniform background, species occupancy at higher elevation would be under-estimated and the estimated niche would thus be biased toward lower elevation.

Presence-only data has evolved in availability and format. Indeed, thanks to large scale citizen-sciences programs like iNaturalist (https://www.inaturalist.org/), eBird (https://ebird.org/home), Pl@ntNet (https://plantnet.org/) or Naturgucker (https://www.naturgucker.de/), spreading the use of smartphone applications for reporting naturalist observations ([14]), presence-only data become massive in developed countries and geolocation of individual specimens becomes more accurate. In the past, most presence only data came from experts collections: Natural museums, naturalist surveys, conservatories data or environmental agencies. Observations of species presences were often aggregated to a prospection site geolocation, which spatial coverage is unknown and varies between sites. The Target-Group Background method (TGB) was proposed by [15] to correct for sampling bias in presence-only niche models in this context. It proposes to define background points as the sites where there has been at least one presence among a Target-Group of species. Today, almost each species presence reported from a mobile phone has its own geolocation and to aggregate them a posteriori in sites asks specific methodological questions. A simpler, and slightly different method is to integrate all species occurrences from the Target-Group as background. Of course, this procedure has strong links with the original TGB approach, but while TGB requires sampling effort to be homogeneous between sites to work properly, as noticed by [16] (page 429), the other method might better correct for a varying sampling effort because the concentration of occurrences from all TG species sounds more proportional to the prospection pressure in the area.

In this study, we propose a new theoretical investigation of specific advantages and biases of this approach, that we will call Target-Group Occurrences Background (TGOB) in the following. A basic problem is that the density of occurrences in the TG might be a poor

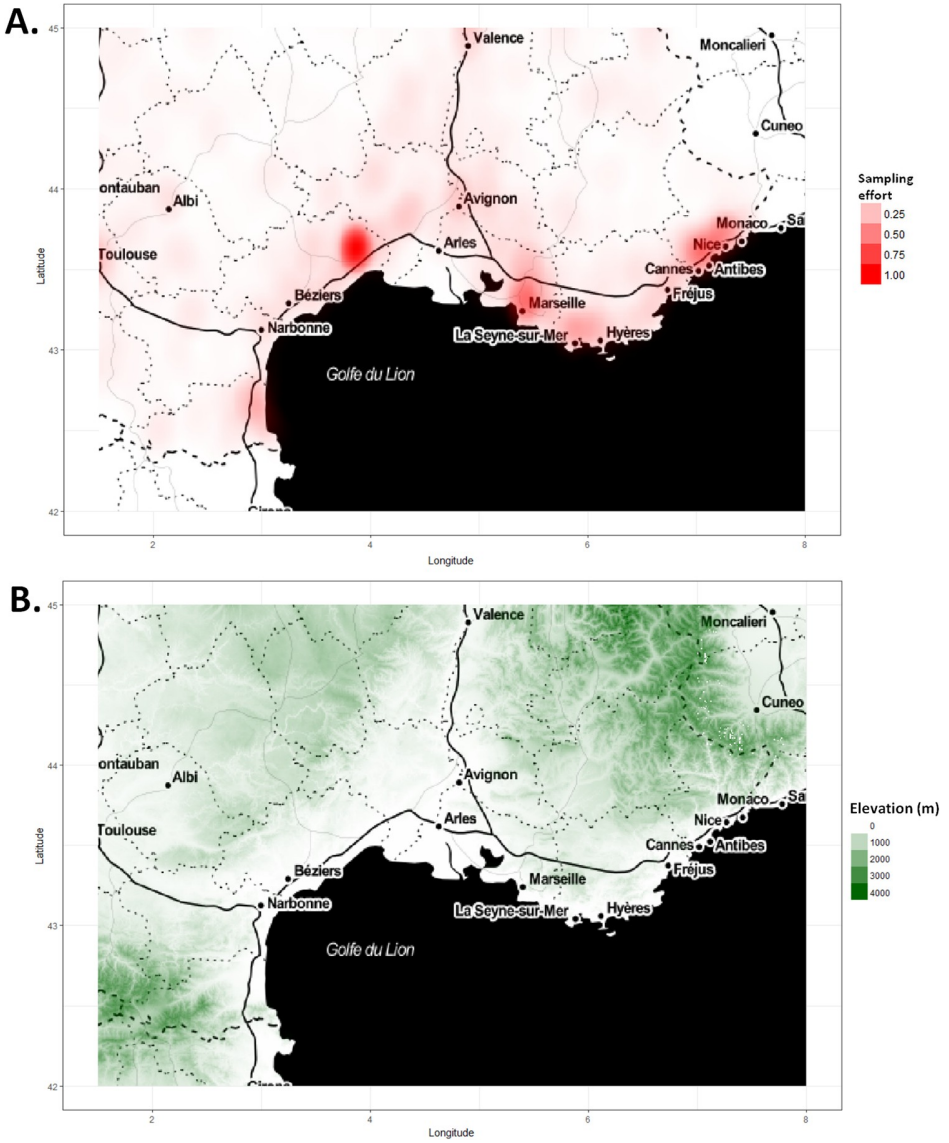


Fig 1. Elevation versus sampling effort in the French mediterranean region. A. An illustration of what might look like the sampling probability (or sampling effort function) over the French mediterranean region. This function is based on a kernel density estimate fitted on all the plant identifications queries sent to the Pl@ntNet mobile application system during 2016 and 2017. B. Ground elevation in meters over the French Mediterranean region. This data is extracted from the SRTM 2010 elevation database with resolution 3 arc-seconds (≈ 90 meters), see the U.S. Geological Survey website (https://lta.cr.usgs.gov/SRTMVF).

https://doi.org/10.1371/journal.pone.0232078.g001

approximation of the real sampling effort, because it does not only reflect sampling effort but also the varying species densities and ecological preferences of species in the TG. Thus, using Target-Group occurrences background may entail new estimation biases in SDM. However, there is no comprehensive perspective on the conditions leading to such bias. Here we address which properties of sampling effort and which ecological characteristics of species in TG can entail biases in (i) an analysis with uniform background points, and (ii) an analysis with Target-Group occurrences background.

Poisson process are useful models for presence-only SDM because they enable a clear probabilistic model and inference procedure for estimating the species environmental density. We consider Poisson process models with log-linear intensity function, which includes the most popular Maxent model ([17]). Starting from a model of species occurrences based on a thinned Poisson process where the thinning intensity is heterogeneous in space and represents the sampling effort, we first exhibited the induced Poisson process in the environmental space and showed how its intensity factorizes into the species intensity and the sampling effort averaged over space for any environment. We then re-expressed the expected density estimator as a divergence depending on focal species density, TG species density and observation density. We assessed how estimation biases arise when these densities are environmentally heterogeneous. We simulated basic cases where estimation biases are expected, for different types of sampling effort, varying niche types of the focal species (specialist vs generalist, typical vs marginal optimum), and three levels of niche breadth in TG species. We show that using background points drawn from the sampling effort proportional density is asymptotically unbiased, and show two types of bias related to alternative ways of defining background points: (i) a bias due to a mismatch of background points with actual sampling effort in the uniform background selection scheme, (ii) a bias due to ecological preferences of TG species, but irrespective of sampling heterogeneity, in TGOB.

To our knowledge, this is the first study bringing such theoretical insights to characterize sampling-related biases in presence-only SDM. Our results should help SDM users anticipate those biases, and decide whether they can use uniform, TGO backgrounds, or orientate them towards other methods and complementary data. Guidelines are provided for building the TG. It should guide good practices for performing more reliable presence only habitat models.

In **section 2**, the model of species distribution and observation is described, we introduce the form of the point process intensity in the environmental space and the observation intensity factor. In section 3, the simulation and inference settings are described. In section 4, detailed results are provided and finally, in section 5, they are discussed in order to provide guidelines for modelers.

2 Model of species observations

We introduce here a probabilistic model controlling the random generation of species located occurrences. It is a two step process where (i) species individuals locations are distributed according to a Poisson point process (see section 2.2), (ii) the individuals are partially observed through a random thinning operation (section 2.3). Section 2.3 also introduces an intermediary result, showing how the expected density of occurrences in the environmental space factorizes with an observation density factor that will be crucial to determine the bias of species density estimation. Before anything else, section 2.1 introduces some notations used all along the article, and the reader may find all notations are summarized and explained in Table 1.

2.1 Notations

We define a measured two dimensional space $(D, \mathcal{L}(D), \mu)$, where $\mathcal{L}(D)$ is the Lebesgue σ-algebra over D, a bounded subset of $\mathbb{R}^2$, and μ is the Lebesgue measure on $\mathbb{R}^2$, which can be understood as the standard measure of area. Individuals of a species are represented by points distributed over D, and only a part of them is reported by observers. Over this domain we consider an environmental variable that is represented by a measurable function $x : D \to \mathbb{R}$, continuous almost everywhere and bounded. We note $\mathrm{Im}(x) = \{w \in \mathbb{R}, \exists z \in D,\ x \text{ is continuous at z and } x(z) = w\}$. Then, $\forall W \subset \mathbb{R}$, we note $x^{-1}(W) = \{z \in D, x(z) \in W\}$. We deal here with a single environmental variable x for clarity,

Table 1. Notations summary: Mathematical notation, name, definition and meaning in our model. *Almost everywhere.

Notation	Name	Formal definition	Role in model
D	Geographic domain	$D \subset \mathbb{R}^2$ bounded	Represent the study area
x	Environmental variable	$D \to \mathbb{R}$ continuous a.e.* and bounded	Enviro. variable measured over D ex: anual precipitations
λ	Species intensity	$\lambda : \mathbb{R} \to \mathbb{R}^+$ continuous a.e.* and bounded on any bounded subset	Expected species abundance per space unit
f	Species density	$f : \mathbb{R} \to \mathbb{R}^+, f := \frac{\lambda}{\int_{\mathbb{R}} \lambda d\mu}$	Density derived from λ over $\mathbb{R}$
s	Sampling effort	s: $D \to [0, 1]$ continuous	Locally represents the probability to report a species individual
$\bar{s}$	Observation intensity	$\bar{s} : \mathbb{R} \to [0, 1]$, Expressed in Eq 1	Avg. sampling effort on areas of D where $x = w$
s_x	Observation density	$s_x : \mathbb{R} \to \mathbb{R}^+, s_x := \frac{\bar{s}}{\int_{\mathbb{R}} \bar{s} d\mu}$	Density derived from $\bar{s}$ over $\mathbb{R}$. Controls UB bias, see Eq 2
a	Cumulated Target-Group species density	$a : \mathbb{R} \to \mathbb{R}^+, a := \frac{\sum_{i=1}^{N} \lambda^i}{\int_{\mathbb{R}} (\sum_{i=1}^{N} \lambda^i) d\mu}$	Controls TGOB bias see Eq 5

https://doi.org/10.1371/journal.pone.0232078.t001

but the results can be extended to more variables with the same method. We also define μ_x, the geographic area where x takes a certain range of values: For all subset of environment value $W \in \mathcal{L}(\mathbb{R})$, $\mu_x(W) = \mu\{x^{-1}(W)\} = \int_{x^{-1}(W)} 1 d\mu$, where $\mathcal{L}(\mathbb{R})$ is the Lebesgue σ-algebra over $\mathbb{R}$. The almost continuity of x means that $\mu_x(\mathrm{Im}(x)) = \mu(D)$, i.e. the spatial area over which x is continuous equals the area of D, or said differently, the area of all points of discontinuity of x taken together is null. This hypothesis allows us to deal either with a continuously varying variable (e.g. defined by a mathematical function over space), or a locally discontinuous one, typically like raster environmental data (see for example [18] for a review on commonly used environmental variables in plants SDM), and even a mixture of both. For example, x could be the elevation variable illustrated by Fig 1. Thus, this hypothesis makes our analysis quite general regarding x.

2.2 Distribution model

Species individuals are represented by the random set Z of their positions in D. We assume Z is distributed according to an inhomogeneous Poisson process over D with intensity function $\lambda o x : D \to \mathbb{R}^+$, where o is functions composition. The intensity λ depends on the environmental variable x. We assume it is continuous almost everywhere on $\mathbb{R}$, has bounded values on any bounded subset of $\mathbb{R}$ and note: $Z \sim IPP(\lambda o x(.))$. Poisson process have indeed been proposed and used as natural probabilistic models for the distribution of species individuals in space ([12, 16]). The intensity represents the punctual limit of the expected species abundance per space unit. We note, $\forall w \in \mathbb{R}$, $f(w) = \frac{\lambda(w)}{\int_{\mathbb{R}} \lambda(u) du}$, a formal definition of the ecological concept of the species response function to variable x ([19, 20]). It can be seen as the probability density function of the random environmental variable $x(z)$ of any individual random location z inside a virtual geographic space where all possible environmental values of x are equally represented in terms of area (this is not necessarily the case in D). In short, we call f the species density. The inhomogeneous Poisson process model proposed here represents a broad class of presence-only SDM including the popular Maxent model, even though Maxent further uses a L1 penalty for model selection. This regularization was not integrated in the study as it doesn't change the incidence of sampling bias.

2.3 Observation model and observation density along the environmental gradient

We use a probabilistic model of observation in order to study the effect of heterogeneous sampling effort on bias. It is similar to the models used in [4, 15, 16, 21]. We consider a continuous **sampling effort** function s: $D \to [0, 1]$. For any point $z \in D$ where an individual of some species is located, the probability to report it is $s(z)$. Note that s is not a probability density over D. There is, of course, no occurrences apart from true locations of individuals. Under this model, the thinning property of inhomogeneous Poisson process ([22]), called Prekopa's theorem, states that reported presences of the species Z_r are distributed according to $Z_r \sim IPP(s(.)\lambda \circ x(.))$. To understand more clearly sampling bias on estimated niche, we propose to look rather at the environmental space rather than the geographic space. Indeed, we are especially interested in the bias of the estimated species density, which is a function of the environmental variables. However, estimation bias will depend on the sampling effort, which is defined over the geographic space but may be transposed to the environmental space. Our first and intermediary result (proved in Text A of S1 Appendix) is that the distribution of the observed species individuals in the environmental space $\mathbb{R}$ also follows a general Poisson process ([22, 23]) whose measure is, for any $W \in \mathbb{R}$, $\int_W \lambda \bar{s} d\mu_x$ and intensity $\lambda\bar{s}$. Where $\bar{s}$ is defined by Eq 1. This intensity function $\lambda(w)\bar{s}(w)$ in environment w represents the expected number of occurrences on any spatial unit where the environment is constant and equal to w, given the underlying shape of the sampling effort s. We show that it is the product of the species intensity λ and the average of the sampling effort $\bar{s}$ across all areas of D with the given environment. This factorization appears because the species intensity is a function of x.

$$\forall w \in \mathbb{R}, \bar{s}(w) = \begin{cases} \lim_{\delta \to 0} \frac{\int_{x^{-1}\left(\left[w-\frac{\delta}{2}, w+\frac{\delta}{2}\right]\right)} s d\mu}{\mu_x\left(\left[w-\frac{\delta}{2}, w+\frac{\delta}{2}\right]\right)} & \text{if } w \in \mathrm{Im}(x) \\ 0 & \text{otherwise, by convention.} \end{cases} \quad (1)$$

We note s_x the environmental density associated to $\bar{s}$ on $\mathbb{R}$, called the **observation density**: $\forall w \in \mathbb{R}, \quad s_x(w) = \frac{\bar{s}(w)}{\int_{\mathbb{R}} \bar{s} d\mu}$. In other words, s_x is the probability density of $x(z)$ when z is randomly drawn over D according to the proportional density of the sampling effort ($s/\int_D s d\mu$). For example, if the environment where observers spend the most time per area unit is $x = w$, then $s_x(w)$ will be the maximum of s_x. The results section will tell precisely how s_x induce bias with the uniform background scheme.

3 Simulation and inference setting

To clarify and illustrate the practical consequences of the mathematical results presented in section **4**, we carry out a simulation experiment exhibiting the estimation biases in various scenarios. In the following, **UB** denotes the estimation of a Poisson Point Process model with uniform background, and **TGOB** the Target-Group occurrences background alternative. We simulate large samples of observed points of a focal species under contrasted scenarios of focal species density and observation density shapes. We also generate a large set of alternatively uniform or Target-Group background points, with various shapes of species cumulated density for the latter. We carry out the species density model estimation from the given focal species observed points and background points. We finally plot the estimated density, approximating the expected estimation, against the true one and the observation density along the enviromnental variable axis. For UB, we also plot the focal species occurrences, that is the theoretically expected density estimate, while for TGOB we plot the TG species cumulated

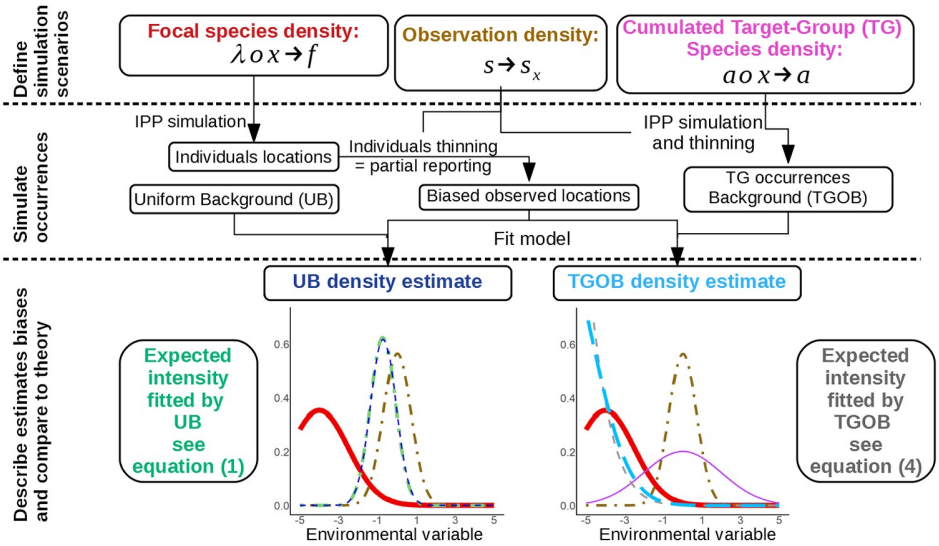


Fig 2. Illustration of the simulation experiment procedure used in this paper to evaluate species density estimation bias under various scenarios. This flowchart shows the role of every component (i.e. the focal species intensity f, the observation density s_x, and the cumulated TG species density a) in the simulation of occurrences, the density estimation with TGOB and UB, and the illustrative comparison of the estimates with the theoretical expectations respectively exhibited by Eqs 2 and 5.

https://doi.org/10.1371/journal.pone.0232078.g002

density shape and the theoretically expected density estimate. This experimental procedure is summarized in diagram of Fig 2. This part presents each step of the simulation scheme and technical settings.

3.1 Environmental variable

We consider a square spatial domain $D = [-5, 5]^2$ where the environmental variable x is a linear gradient from west to east, such that $x(z) = z_1$. In this setting, μ_x is equal to the restriction of the $\mathbb{R}$-Lebesgue measure to $\mathrm{Im}(x) = [-5, 5]$, *i.e.* each x value has the same spatial extent, and thus the estimate will not be better in most represented values. Illustrations of the density derived from μ_x, $\mathrm{Im}(x)$, an observation density and species density (see further) are provided in S1 Fig.

3.2 Focal species

The species density f is the probability density function of the environmental value of a specimen random location. We model it with a Gaussian function, which is a standard assumption related to the representation of species distribution over environmental gradients ([19, 24]). We give some insights about the underlying model assumptions in **Text B of** S1 Appendix. The mean of f is called μ_0, it is the environmental optimum of the species, and we take $\mu_0 \in \{-1, -4\}$ (typical vs marginal). Besides, σ_0 is the standard deviation, or the niche breadth of the species, and we take $\sigma_0 \in \{0.6, 1.5\}$, for a specialist or generalist species. We thus simulate 4 virtual species. f is illustrated in each graph of Fig 3.

3.3 Types of observation density

We want to estimate the density of the focal species from reported points. We examine how the bias in estimated intensity is related to s_x, the observation density in $\mathrm{Im}(x)$. We define several shapes for s_x in $\mathrm{Im}(x)$, which is illustrated with the yellow curve in each graph of Fig 3:

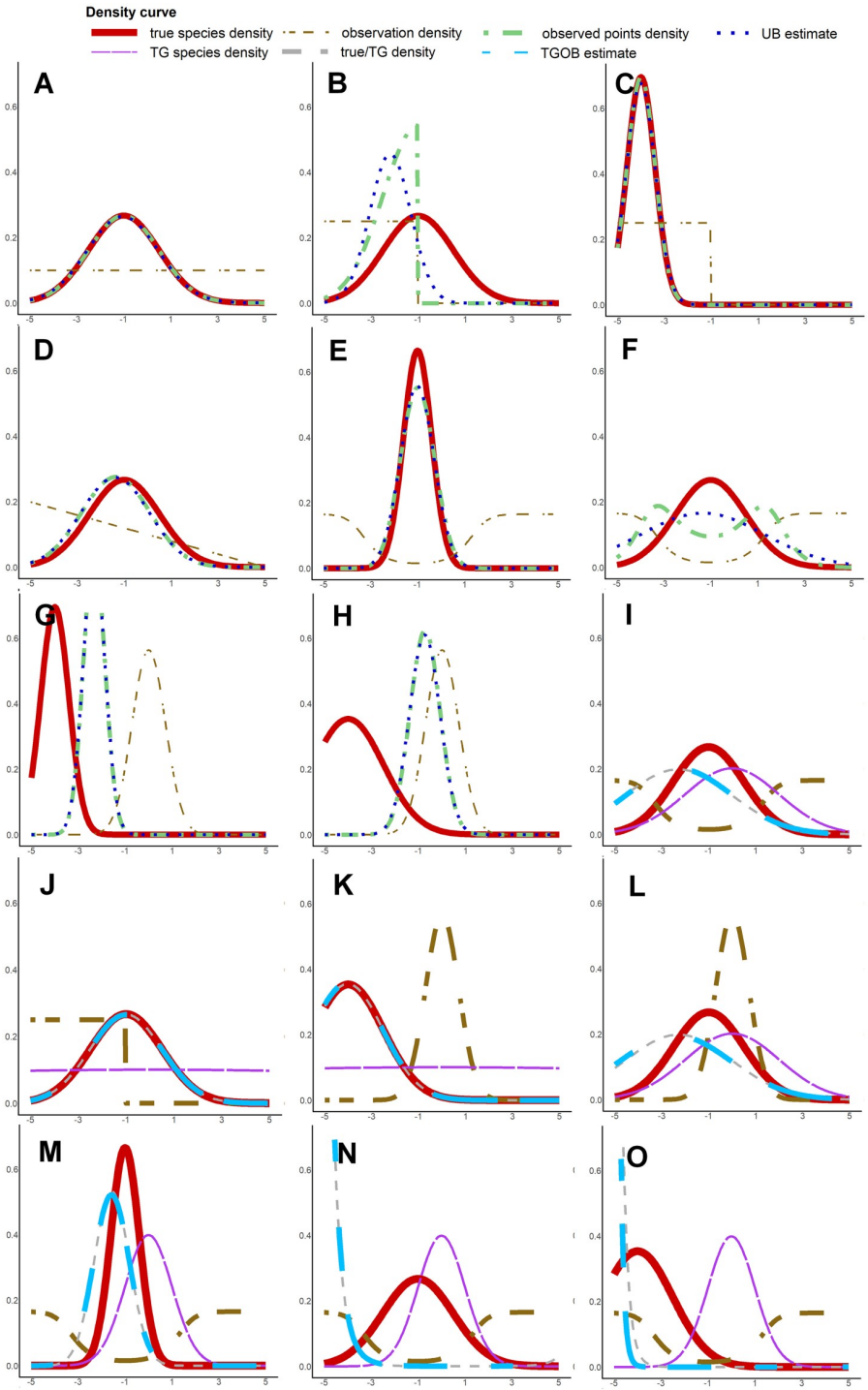


Fig 3. Plot of the estimated niche density with UB (A-H) and TGOB (I-O) methods for a selection of simulation situations. The different curves are: The focal species intensity function (f), observation density (s_x), observed points density ($\lambda_0\, s_x$, in UB graphs), Target-Group species density (a, in TGOB graphs), ratio density of species over target group (λ_0/a, in TGOB graphs), UB and TGOB estimators of species density from simulated points. A-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *CST*. B-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *CUT*. C-$\mu_0 = -4$; $\sigma_0 = 0.6$; *obs* = *CUT*. D-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *LIN*. E-$\mu_0 = -1$; $\sigma_0 = 0.6$; *obs* = *HOL*. F-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *HOL*. G-$\mu_0 = -4$; $\sigma_0 = 0.6$; *obs* = *GS*. H-$\mu_0 = -4$; $\sigma_0 = 1.5$; *obs* = *GS*. I-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *HOL*. J-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *CUT*. K-$\mu_0 = -4$; $\sigma_0 = 1.5$; *obs* = *GS*. L-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *GS*. M-$\mu_0 = -1$; $\sigma_0 = 0.6$; *obs* = *HOL*. N-$\mu_0 = -1$; $\sigma_0 = 1.5$; *obs* = *HOL*. A-$\mu_0 = -4$; $\sigma_0 = 0.6$; *obs* = *HOL*.

https://doi.org/10.1371/journal.pone.0232078.g003

1. Constant (CST), representing unbiased, constant observation over the domain. See graph A.
2. $(1/10) - (x/50)$, *i.e.* linearly decreasing from west to east (LIN). See graph D.
3. $\frac{1_{x\in[-5,0]}}{5}$, constant observation on the lower part of the domain (CUT). See graph B.
4. $\frac{\log(1+(x+1)^2)}{\int_{[-5,5]}\log(1+(w+1)^2)dw}$, with depleted observation density around -1 (HOL). See graph E.
5. A standard normal distribution (NOR). See graph G.

Note that s_x is determined through the definition of the sampling effort s which is in the spatial domain. We set the sampling effort to be constant along the second dimension of space (latitude) in our simulation setting, which enforces $s_x \propto s$ and we thus control the shape of s_w through the shape of s over the longitude.

3.4 Target group of species

TGOB method uses occurrences from a set of species called the Target Group (TG) as background points in the inference setting (see methods implementation below). We thus simulate the TG occurrences background by generating occurrences of N independent species, constituting the TG, through their observed intensities. For species i, its local observed intensity takes values $\lambda^i(x(z))\bar{s}(x(z)), \forall z \in D$ (assuming constant detection in space), and regrouping occurrences of all TG species is equivalent to drawing points with a global intensity $C_a a(x(z))\bar{s}(x(z)) = \sum_{i=1}^N \lambda^i(x(z))\bar{s}(x(z))$, where $a(x(z)) := \sum_{i=1}^N \lambda^i(x(z))/C_a$ is called the TG species cumulated density and $C_a := = \int_{\mathbb{R}}(\sum_{i=1}^N \lambda^i)d\mu$ is its normalisation constant. As it is shown further, a will determine the bias of TGOB. Thus, we do not define each TG species density individually in the simulation, but rather test 3 shapes of a. It enables to visualize clearly its effect on TGOB bias: (i) FLAT: A Gaussian density of mean 0 and standard deviation 20 ($\approx$ constant), (ii) THICK: A Gaussian density of mean 0 and standard deviation 2 and (iii) THIN: A Gaussian density of mean 0 and standard deviation 1. They are represented in, respectively, graphs J, I and M of Fig 3.

3.5 Simulating observation points

Statistical theory insures that the density estimate will converge towards its expectation when increasing the size of the sample. Then, for all simulations, we generate a very large sample of points (occurrences and background) so that the estimate approximates well this expectation, insuring that the estimation error is completely due to bias and not the randomness of the sample. To generate points according to a Poisson process of intensity function f on $\mathrm{Im}(x)$, we first determine an upper bound B of f on $\mathrm{Im}(x)$. Then, we repeat (i) Draw a point $z \sim U(D)$, (ii) Draw a variable $y \sim U([0, B])$, (iii) We accept z if $y <= f(x(z))$ and (iv) If 20000 points are accepted, finish the procedure, otherwise go back to (i). This algorithm is applied to the focal species observed points, target group observed points and background points (see next section). 20000 points were enough for convergence of all estimates in UB and TGOB.

3.6 Computation of models and software

In the UB method, we estimate the model parameters with the standard maximum likelihood approach. We use the Poisson process likelihood approximation of [25], which transform the original likelihood to a Poisson regression likelihood, using background points. We draw the background points uniformly in the spatial domain D. Details on the construction of

approximation, the weighting of points and the reparametrization of μ_0 and σ_0 are presented in **Text C of** S1 Appendix. As the objective function is a particular case of Generalized Linear Model likelihood, we fit the parameters using the standard R package **glm**. For *TGOB* method, the procedure is the same except that the background points are independently drawn from the density $sa/\int_D sad\mu$ rather than uniformly on D.

4 Results

We present results on estimation biases for UB and TGOB methods based on both a mathematical analysis and simulation. Our main results are formal Eqs (2) and (5) which express the target of the density estimate in the environmental space as a function of the true focal species density f, the observation density s_x (for UB) or the cumulated TG species density a (for TGOB) given the generative model described in section **2**. Estimation bias then depends on the instanciation of f and s_x for UB, or of f and a for TGOB. We qualitatively describe the bias, i.e. the estimated density deviation compared to the true one, that will appear depending on the shape of the dependent densities: The observation density (for UB in sections **4.2, 4.4, 4.5, 4.6** and for TGOB in **4.8**), the focal species density (for UB in **4.3**, and for TGOB in **4.9, 4.10**) and the Target-Group species density for TGOB (**4.8, 4.9**). This qualitative description are based on interpretation of Eqs 2, 3, 4, 5 and 6. This qualitative description of bias is numerically illustrated with several simulated scenarios. Graphs of all simulated scenarios are represented in S2 Fig for UB, and S3, S4 and S5 Figs for TGOB. R scripts for running the simulations and generating the graphs can be found in at https://github.com/ChrisBotella/UB-and-TGOB. Results are presented here for a single environmental variable. In the case of several environmental variables $x_1, \ldots, x_p$, the Kullback-Leibler (KL) divergence used in the following equations is simply applied to densities over the multidimensional space, with adapted definitions for $s_{x_1,\ldots,x_p}$ and $\mu_{x_1,\ldots,x_p}$. For simplifying notations, we will possibly mean, by the notation of a function, a product or a quotient of functions, the density associated with it on its definition space, and this in all that follows. For example, fs_x refers to the proportional density function $fs_x/\int_{\mathrm{Im}(x)} fs_x\, d\mu_x$ over $\mathrm{Im}(x)$.

4.1 UB: Niche estimate minimizes KL divergence from observed density

We show in **Text D of** S1 Appendix that the expectation of the parameters estimates of the UB method is:

$$\mathbb{E}(\hat{\theta}_{UB}) = \mathrm{argmin}_\theta\ \mathcal{D}_{\mathrm{KL}}^{\mu_x}(fs_x||f_\theta) \tag{2}$$

Eq 2 means that the estimated species density $f_{\hat{\theta}_{UB}}$ will fit the observed environmental density fs_x as close as possible within the parametrization constraints in term of the KL Divergence with measure μ_x (μ_x-almost everywhere). For example, in our simulation model, $f_{\hat{\theta}_{UB}}$ is Gaussian, so it cannot fit perfectly to fs_x which is non-Gaussian (see graph **B** of Fig 3), but achieves the best Gaussian approximation. However, in the case where s_x and f are two Gaussian densities with distinct means and variances, fs_x will also be Gaussian [26]. Thus, $f_{\hat{\theta}_{UB}}$ will exactly converge to fs_x (see graph **H** of Fig 3). However, it has a different mean and variance from f, so that the UB estimate is biased. A Complementary explanation about the significance of μ_x for the KL-Divergence, and its consequences are given in **Text E of** S1 Appendix.

4.2 UB: Bias is small for small variations of observation density over the species niche

UB bias is tightly linked to the concentration of the observation density in the environmental space but this concept of concentration is hard to define. Still, as a density get less concentrated it get closer to a uniform density, and its variation get close to zero everywhere. Thus, we study the effects of variations of s_x and f on bias, we propose an explanation of the bias behavior observed in simulation through a simple analysis based on the density functions derivatives. For this purpose, both density functions are assumed to be differentiable over Im(x), which is true in the simulation setting, except in the case of observation type **CUT**. Eq 3 shows that when s_x varies little, the observed points density $s_x f$, which is fitted by the UB estimate, will get close to the true species density f.

$$\lim_{\max|\partial s_x/\partial x|\to 0} \frac{\partial f s_x}{\partial x} = \lim_{\max|\partial s_x/\partial x|\to 0} \frac{\left(\frac{\partial f}{\partial x} s_x + \frac{\partial s_x}{\partial x} f\right)}{\int_{\mathbb{R}} f\, s_x\, d\mu_x} = \frac{\partial f}{\partial x} \tag{3}$$

Fig 3A confirms that UB is not biased when observation density is constant: The species true density f (red curve) is equal to the observed point density $s_x f$ (green curve), which is perfectly fitted by the UB estimated density (blue curve). Even for graph D, the gap between true and estimated density is very small. This behavior is explained by Eq 3: If linearly decreasing observation density varies slowly, i.e. $max|\partial s_x/\partial x$ is close to zero, the derivative of the target $\partial f_{\theta_0} s_x \approx f_{\theta_0}/\partial x$ is close to the derivative of the species true density, implying that the estimate will fit this density. In addition, in environments where species specimens are rare, very low observation density doesn't affect the global estimate. Type *CUT* illustrates this: There is almost no bias for $\mu_0 = -4$ (graph C of Fig 3), as the observed species density (green curve) is very close to the true species density (red curve). We note as a side remark that the differentiability of s_x over Im(x) is not necessary. It depends on complex conditions on x and s. As a counter example, continuity of s_x doesn't even have a standard sense if x is defined by a geographic raster. Indeed, Im(x) is then discrete set of x values taken over the raster cells, and $\bar{s}$ is only defined on these values which don't include any continuum of real numbers. The differentiability is only assumed here to analyse the effects of s_x variations in a simplified context.

4.3 UB: Smaller bias for more specialist species

The comparison of the graphs **G** (specialist) to **H** (generalist) in Fig 3 shows that the bias on niche optimum and breadth estimates is stronger for the generalist species. Indeed, we deduce from Eq 4 that fs_x approaches s_x as the variation of f over Im(x) decreases.

$$\lim_{\max|\partial f/\partial x|\to 0} \frac{\partial f s_x}{\partial x} = \lim_{\max|\partial f/\partial x|\to 0} \frac{\left(\frac{\partial f}{\partial x} s_x + \frac{\partial s_x}{\partial x} f\right)}{\int_{\mathbb{R}} f s_x d\mu_x} = \frac{\partial s_x}{\partial x} \tag{4}$$

We can thus say that for a generalist species, the variation speed of s_x is high compared to the one of f, and UB estimate will fit more the observation density than the species density.

4.4 UB: Over-estimated specialization when sampling effort is concentrated

When the observation density is highly concentrated in a restricted range of the environment, as with the type *GS*, UB estimates that the species is more specialized than it is actually (see graphs G and H of Fig 3). The estimated niche variance is then lower than expected.

4.5 UB: Strong deviations from optimum

Graphs **B** and **H** in Fig 3 show that, when the observation density is concentrated far from the optimum of the species density, we get a strongly deviated estimated optimum. This might be very misleading for ecological analysis. Estimation of graph **H** suggests that the species is the most abundant in a range where it is actually cryptic.

4.6 UB: Sampling marginal specimens means over-estimating generalism

Graph **F** of Fig 3 shows that when the observation is more intense in the margin of the species niche, UB over-estimates the niche breadth of the species. This case represent observers having more interest in reporting a species out of its typical environment.

4.7 TGOB: Integrating samples from a Target Group of species

Firstly, using the same analytical approach as previously, we show in **Text F** of S1 Appendix that drawing directly background points from the sampling effort proportional density $s/\int_D s(z)dz$ give unbiased species intensity estimate. This answers an open question of [4] who introduced this theoretical method (called `ApproxFactorBiasOut` in the article). Unfortunately, we rarely have directly access to a true sample from the sampling effort distribution. An interesting alternative is to use Target-Group species occurrences as background points (TGOB), i.e. making the hypothesis that those occurrences are approximately drawn from the sampling effort proportional density. We will investigate biases occurring with this method and a necessary and sufficient condition on Target-Group species to avoid them under our modeling hypothesis. In the following, we introduce an equation showing the displaced target of the TGOB estimator. It shows how the cumulated TG species density, especially when it is concentrated in restricted environments, can bias the estimated focal species density. We have a target group of N species whose individuals are distributed independently according to the species model described above, and reported from the same area D with the probability of observation s (same as the species of interest), giving for each of them a set of observation locations $(Z^i)_{i\in[|1,n_{tg}|]}$. $\forall i \in [|1, N|],\ Z^i \sim \mathcal{IPP}(s\,\lambda^i \circ x)$. We assume a constant detection probability of individuals across space for any species conditionally to observation. Then, the global set of Target Group observations locations is $Z^{tg} := \cup_{i\in[|1,N|]} Z^i \sim \mathcal{IPP}(s\,a \circ x)$, where $\forall z \in D, a(x(z)) := \sum_{i=1}^{N} \lambda^i(x(z)))$ is the cumulated TG species intensity. The expected estimate of TGOB is:

$$\mathbb{E}(\hat{\theta}_{TGOB}) = \text{argmin}_\theta\ D_{\text{KL}}^{\mu_x}(fs_x||f_\theta s_x a) \tag{5}$$

The proof is given in **Text G of** S1 Appendix. If $\forall w \in \text{Im}(x)$, $a(w) > 0$, we can set $f_\theta := f/a$ to cancel the divergence. Eq 5 means the TGOB estimate is expected to fit to density f/a, which is independent of the observation density, but depends on the cumulated TG species density. This result leads to the following consequences described in sections **4.8, 4.9** and **4.10**.

4.8 TGOB: If *a* is constant, TGOB is unbiased

We can see that when a is constant, $s_x\ a\ \alpha\ s_x$. Thus, the background points are distributed according to the sampling effort, and TGOB yields an unbiased estimation as `ApproxFactorBiasOut`. This is true whatever is the observation density. We illustrate it in two cases of Fig 3: $\mu_0 = -1; \sigma_0 = 1.5$; *CUT* with graph J and $\mu_0 = -4$; $\sigma_0 = 1.5$; *GS* with graph **K**. Here the TGOB estimator approaches almost perfectly the true species density,

correcting well for unbalanced observation density in both cases, while in those same cases UB gives a strongly biased estimate. Furthermore, even with non constant a, the different types of observation density never affect TGOB. The bias is only due to the Target Group species density. For example, graphs **I** and **L** of Fig 3 show that TGOB estimator do not change in two very different observation density situations, *HOL* and *GS*, but with the same species density $\{\mu_0 = -1, \sigma_0 = 1.5\}$ and TG.

4.9 TGOB: The estimate deviates from a peaky Target Group species density

The more the Target Group species density (a) is concentrated in some range of x, the more our niche estimate will be located outside of this range. It may entail an over estimation of niche breadth, a bias in optimum, or even an hyper-concentration on the borders. To show this, we can analyse the effect of the variation speed of a and f, by again assuming that they are differentiable over $\mathrm{Im}(x)$ and examining the derivative of f/a:

$$\frac{\partial f/a}{\partial x} = \frac{1}{a}\left(\frac{\partial f}{\partial x} - \frac{f}{a}\frac{\partial a}{\partial x}\right) \tag{6}$$

If a gets high in a neighborhood v of $\mathrm{Im}(x)$, we will have $f/a \to 0$ on v, and $\frac{\partial f/a}{\partial x}$ tends to 0 as well. Our estimate then becomes flat and low on v as it fits to f/a. In parallel, a is low outside of v because it must integrate to 1. Therefore, in $\mathrm{Im}(x)\backslash v$, we will have $f/a \to +\infty$, and its derivative becomes important with the same sign as $-\frac{\partial a}{\partial x}$. In summary, as a concentrates in a neighborhood v, our TGOB estimate becomes flat and low on v, while it increases outside of v, with bigger slopes where a varies. This expulsion phenomenon entails bias in optimum and variance estimation. Thus, the magnitude of bias depends on the concentration of a, but also on the marginality of the optimum of the focal species (μ_0) compared to the one of the Target-Group. Indeed, the graphs **I** and **M** of Fig 3 show that when the species optimum is close to the one of the TG density (typical species), the niche breadth is over-estimated. There is also a small deviation in optimum because the focal species is not centered around the TG optimum. In other words, the focal species density overlaying with the cumulated TG species density is deviated outside in the estimate. On the contrary, when the species optimum is far from the cumulated TG species density optimum (marginal species, see graph **O** of Fig 3), or when the cumulated TG density is just more concentrated (compare graph **N** to **I** in Fig 3), the situation is worse. The estimate cancels on the range of the cumulated TG species density, while it gets hyper-concentrated outside. In summary, the more the Target Group of species has a global environmental preference and the focal species is marginal, the more its niche estimate will be dispersed, or expelled, out of this environment.

4.1 TGOB: Stronger bias for generalist species

When comparing graph **M** to **N** in Fig 3, we see that TGOB is more biased on generalist species. For a generalist species, the estimate is more expelled from the TG species density volume. Thus, generalism of the focal species increases bias in both UB and TGOB, but the cause of bias differs, respectively, the heterogeneity of observation density and the TG global density. As UB fits the product of f and s_x, TGOB does the same with the product of f and $1/a$, and the latter varies in $-\frac{\partial a}{\partial x}\frac{f}{a^2}$ because the variation of f is small.

5 Discussion

In this study, we have explained two types of bias related to the way to define background points: the **sampling selection bias** in UB and the the **TG definition bias** in TGOB. The former case concerns the way background points reflect sampling heterogeneity, while the latter case concerns the influence of ecological preferences in TG species.

Concerning UB, our results confirm some empirical results in Maxent literature. The niche estimate will fit to the product of the focal species and observation densities. A major consequence is that bias is stronger for generalist species. Bias is also strong when the sampling effort is concentrated towards places representing a restricted range of environmental values, which happens when observers have specific preferences towards these restricted conditions. This will overestimate species specialization. Conversely, observing a species more intensively at the margin of its niche leads to overestimate niche breadth.

If the Target-Group is well selected, the method Target-Group occurrences background does account for varying sampling effort. A well selected Target-Group means that the sum of Target-Group species intensities is constant across environments. However, it is biased when this cumulated intensity of TG species varies in the environmental space, e.g. when there is some systematic environmental preference among TG species. In this case, the magnitude of bias will depend on the concentration of the TG density (depending on the TG species), the generalism of the focal species, and the marginality of its niche compared to the TG density. As the TG species density gets more concentrated compared to the focal species niche, the niche breadth will be over-estimated, and ultimately focal species density will strongly deviate from TG density. If TG species density approaches 0 faster than the species of interest in some environmental range, TGOB estimator should dramatically increase there, overriding variations elsewhere. Including the focal species in the Target-Group should partly prevent the niche expulsion effect because at least background points from the focal species will cover its niche. Also, the ecological niche of the focal species plays an important role. A generalist species is more affected by bias, as well as species with marginal niche compared to the TG density. On the contrary, when applied to a non-marginal focal species, TGOB will overestimate the niche breadth, or from another point of view, the effect of corresponding covariates will be reduced. This covariate effect cancellation will be all the stronger with Maxent ([27]) because of its Lasso regularisation. We recommend to carefully chose Target Group of species so as to insure, at least, that there are TG occurrences in the widest environmental subspace associated with the study domain. It will insure that at least one of the TG species is present in any kind of environments. Generalist species over each environmental variable should be included if possible to overall decrease the variation of the cumulated TG species density. The modeler must avoid using TGOB if presences of the focal species reach marginal environments compared to the whole Target-Group distribution.

Alternatives methods to TGOB and UB to account for sampling bias in presence only SDMs may be more suited in certain situations. [28] proposed to model sampling effort with distinct environmental variables from the species intensity (e.g. distance to roads or to cities). Thus it removes species intensity bias due to the covariation of sampling effort and species intensity covariates. However, often some covariates influence both sampling and species density. Still, our results support this approach if the sampling effort variation along its dedicated covariates is stronger than the species intensity variation (Eq 4), and the species intensity variation along its covariate is stronger than the sampling effort variation (Eq 3). Besides, for modelers who can access complementary systematic survey data, integrated models combining occurrences and presence-absence data have been developed in [16] and [29] with the same

goal. In the same spirit, models combining presence-background with site-occupancy data ([30]) may be another efficient way to account for sampling bias.

We underline that our results directly concern a vast class of presence-only SDM called Poisson process models ([12]) whose intensity function is strictly positive. Indeed, modelers may use different variables transformations as predictors (GAM [31], MARS [32]), or learn those transformations automatically, like with deep neural networks ([33]). Qualitatively speaking, bias behaviors extend to L1 penalized Poisson process methods like Maxent ([34]) and to other related SDMs methods (whose predictive function is based on covariates) when using pseudo-absences, e.g. GARP ([3]), ENFA ([2]), or BRT ([1]). Models integrating interactions effects between species, called joint SDMs ([35]), should be similarly affected by described biases, as species interactions are assumed independent of the environment, but a specific investigation on biases of such methods would be important in view of the recent attention they are receiving in ecology. We notice that potential biases of the studied methods are not restricted to the ones presented here, and the modeler must be careful to other sources of errors. For example, other authors recently studied how the interaction of environmental variables resolution and niche breadth induce bias ([36]). Besides, model errors might not be due to biases, but rather to estimation variance which is also investigated in the SDM literature ([37, 38, 39]). A limitation of this study is that we did not study some other proposed sampling bias correction methods, such as occurrence thinning procedures, in spatial ([40, 41]) or environmental ([42]) domains. As occurrences thinning increases the entropy of the observed points density, it brings its own bias which should be investigated more closely. Such procedures could be studied through the formalism that we are developing.

TGOB is exactly equivalent to TGB, proposed by [15], if each TG site (defined either by the environmental rasters or the spatial aggregation of the occurrences) contain only one occurrence. However, it may differ significantly when many occurrences are aggregated on sites. If so, TGB will be biased by a varying prospection intensity between sites and varying TG density, while TGOB may be biased only by the latter factor. In this context, the strengths of TGOB would be leveraged by the search for a criterion to select the best Target-Group of species, which guarantees a low variation of the cumulated TG species density in the environment. The difficulty is that such criterion must be computable from the sets of occurrences of species eligible for the Target-Group. This is an open problem and an area for future work, leading to a clear and reliable background points selection method applicable by SDMs end users.

Supporting information

S1 Appendix. Texts and mathematical proofs.
(PDF)

S1 Fig. Illustrations of μ_x, f and s_x along x values. An example species density with the standard normal distribution (red curve), the density derived from μ_x chosen uniform over [−5, 5] for the simulation study (black curve), and the observation density s_x of type *LIN* (gold curve).
(PNG)

S2 Fig. Illustrations of all simulation results for UB. Plotted true species density (f), observation density (s_x), observed points density (fs_x) and UB estimate of species density in the environmental space. Each situation of the simulation study is represented.
(PNG)

S3 Fig. Illustrations of all simulation results for TGOB with FLAT TG species density. Plotted true species density (f), observation density (s_x), **flat** Target Group species density (a), ratio density of species over target group (f/a) and TGOB estimate of species density in the

environmental space. Each situation of the simulation is represented.
(PNG)

S4 Fig. Illustrations of all simulation results for TGOB with THICK TG species density. Plotted true species density (f), observation density (s_x), **thick** Target Group species density (a), ratio density of species over target group (f/a) and TGOB estimate of species density in the environmental space. Each situation of the simulation is represented.
(PNG)

S5 Fig. Illustrations of all simulation results for TGOB with THIN TG species density. Plotted true species density (λ_0), observation density (s_x), **thin** Target Group species density (a), ratio density of species over target group (λ_0/a) and TGOB estimate of species density in the environmental space. Each situation of the simulation is represented.
(PNG)

Author Contributions

Conceptualization: Christophe Botella, Alexis Joly, Pascal Monestiez, Pierre Bonnet, François Munoz.

Formal analysis: Christophe Botella.

Investigation: Christophe Botella, Pascal Monestiez.

Methodology: Christophe Botella, Alexis Joly, Pascal Monestiez, Pierre Bonnet, François Munoz.

Project administration: Pascal Monestiez, Pierre Bonnet, François Munoz.

Software: Christophe Botella.

Supervision: Alexis Joly, Pascal Monestiez, Pierre Bonnet, François Munoz.

Validation: Christophe Botella, Alexis Joly, Pascal Monestiez, Pierre Bonnet, François Munoz.

Visualization: Christophe Botella.

Writing – original draft: Christophe Botella.

Writing – review & editing: Christophe Botella, Alexis Joly, Pascal Monestiez, Pierre Bonnet, François Munoz.

References

1. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics. 2009; 40:677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159
2. Hirzel AH, Hausser J, Chessel D, Perrin N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology. 2002; 83(7):2027–2036. https://doi.org/10.1890/0012-9658(2002)083%5B2027:ENFAHT%5D2.0.CO;2
3. Stockwell D. The GARP modelling system: problems and solutions to automated spatial prediction. International journal of geographical information science. 1999; 13(2):143–158. https://doi.org/10.1080/136588199241391
4. Dudík M, Phillips SJ, Schapire RE. Correcting sample selection bias in maximum entropy density estimation. In: Advances in neural information processing systems; 2006. p. 323–330.
5. Divino F, Golini N, Lasinio GJ, Penttinen A. Bayesian logistic regression for presence-only data. Stochastic environmental research and risk assessment. 2015; 29(6):1721–1736. https://doi.org/10.1007/s00477-015-1064-y

6. Tonini F, Divino F, Lasinio GJ, Hochmair HH, Scheffrahn RH. Predicting the geographical distribution of two invasive termite species from occurrence data. Environmental entomology. 2014; 43(5):1135–1144. https://doi.org/10.1603/EN13312 PMID: 25198370
7. Newbold T. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography. 2010; 34(1):3–22. https://doi.org/10.1177/0309133309355630
8. Meineke EK, Davis CC, Davies TJ. The unrealized potential of herbaria for global change biology. Ecological Monographs. 2018; 88(4):505–525. https://doi.org/10.1002/ecm.1307
9. Ruete A. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodivers Data J. 2015; p. e5361. https://doi.org/10.3897/BDJ.3.e5361 PMID: 26312050
10. Nix H, Busby J. BIOCLIM, a bioclimatic analysis and prediction system. Division of Water and Land Resources: Canberra. 1986.
11. Carpenter G, Gillison A, Winter J. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity & Conservation. 1993; 2(6):667–680. https://doi.org/10.1007/BF00051966
12. Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, et al. Point process models for presence-only analysis. Methods in Ecology and Evolution. 2015; 6(4):366–379. https://doi.org/10.1111/2041-210X.12352
13. Joly A, Bonnet P, Goëau H, Barbe J, Selmi S, Champ J, et al. A look inside the Pl@ ntNet experience. Multimedia Systems. 2016; 22(6):751–766. https://doi.org/10.1007/s00530-015-0462-9
14. Graham EA, Henderson S, Schloss A. Using mobile phones to engage citizen scientists in research. Eos, Transactions American Geophysical Union. 2011; 92(38):313–315. https://doi.org/10.1029/2011EO380002
15. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications. 2009; 19(1):181–197. https://doi.org/10.1890/07-2153.1 PMID: 19323182
16. Fithian W, Elith J, Hastie T, Keith DA. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution. 2015; 6(4):424–438. https://doi.org/10.1111/2041-210X.12242 PMID: 27840673
17. Renner IW, Warton DI. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics. 2013; 69(1):274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x PMID: 23379623
18. Mod HK, Scherrer D, Luoto M, Guisan A. What we use is not what we know: environmental predictors in plant distribution models. Journal of Vegetation Science. 2016; 27(6):1308–1322. https://doi.org/10.1111/jvs.12444
19. Whittaker RH. Gradient analysis of vegetation. Biological reviews. 1967; 42(2):207–264. https://doi.org/10.1111/j.1469-185x.1967.tb01419.x PMID: 4859903
20. Whittaker RH, Niering WA. Vegetation of the Santa Catalina Mountains, Arizona. V. Biomass, production, and diversity along the elevation gradient. Ecology. 1975; 56(4):771–790. https://doi.org/10.2307/1936291
21. Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA. Point pattern modelling for degraded presence-only data over large regions. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2011; 60(5):757–776. https://doi.org/10.1111/j.1467-9876.2011.00769.x
22. Chiu SN, Stoyan D, Kendall WS, Mecke J. Stochastic geometry and its applications. John Wiley & Sons; 2013.
23. Haenggi M. Stochastic geometry for wireless networks, Cambridge Uni; 2013.
24. Franklin J. Mapping species distributions: spatial inference and prediction. Cambridge University Press; 2010.
25. Berman M, Turner TR. Approximating point process likelihoods with GLIM. Applied Statistics. 1992; p. 31–38. https://doi.org/10.2307/2347614
26. Bromiley P. Products and convolutions of Gaussian probability density functions. Tina-Vision Memo. 2003; 3(4):1.
27. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecological modelling. 2006; 190(3):231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026
28. Warton DI, Renner IW, Ramp D. Model-based control of observer bias for the analysis of presence-only data in ecology. PloS one. 2013; 8(11):e79168. https://doi.org/10.1371/journal.pone.0079168 PMID: 24260167

29. Coron C, Calenge C, Giraud C, Julliard R. Estimation of species relative abundances and habitat preferences using opportunistic data. arXiv preprint arXiv:170608281. 2017.

30. Koshkina V, Wang Y, Gordon A, Dorazio RM, White M, Stone L. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods in Ecology and Evolution. 2017; 8(4):420–430. https://doi.org/10.1111/2041-210X.12738

31. Guisan A, Edwards TC Jr, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling. 2002; 157(2-3):89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

32. Friedman JH. Multivariate adaptive regression splines. The annals of statistics. 1991; p. 1–67. https://doi.org/10.1214/aos/1176347963

33. Botella C, Joly A, Bonnet P, Monestiez P, Munoz F. A deep learning approach to species distribution modelling. Multimedia Tools and Applications. 2018.

34. Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution modeling. In: Proceedings of the twenty-first international conference on Machine learning. ACM; 2004. p. 83.

35. Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, et al. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution. 2014; 5(5):397–406. https://doi.org/10.1111/2041-210X.12180

36. Connor T, Hull V, Viña A, Shortridge A, Tang Y, Zhang J, et al. Effects of grain size and niche breadth on species distribution modeling. Ecography. 2017.

37. Wisz MS, Hijmans R, Li J, Peterson AT, Graham C, Guisan A, et al. Effects of sample size on the performance of species distribution models. Diversity and distributions. 2008; 14(5):763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

38. Proosdij AS, Sosef MS, Wieringa JJ, Raes N. Minimum required number of specimen records to develop accurate species distribution models. Ecography. 2016; 39(6):542–552. https://doi.org/10.1111/ecog.01509

39. Soultan A, Safi K. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. PloS one. 2017; 12(11):e0187906. https://doi.org/10.1371/journal.pone.0187906 PMID: 29131827

40. Fourcade Y, Engler JO, Rödder D, Secondi J. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PloS one. 2014; 9(5):e97122. https://doi.org/10.1371/journal.pone.0097122 PMID: 24818607

41. Boria RA, Olson LE, Goodman SM, Anderson RP. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling. 2014; 275:73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

42. Varela S, Anderson RP, García-Valdés R, Fernández-González F. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography. 2013; 37:1084–1091.

7 Chapter 3: Estimating sampling effort across space from large amounts of species occurrences

Jointly estimating spatial sampling effort and habitat suitability for many species from opportunistic occurrences

Get spatial sampling from species occurrences

Christophe Botella[1,2,3,4,¤], Alexis Joly[1,‡], Pierre Bonnet[3,5,‡], François Munoz[6,¶], and Pascal Monestiez[4,¶]

[1]INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.

[2]INRAE, UMR AMAP, F-34398 Montpellier, France.

[3]Univ Montpellier, UMR AMAP, Montpellier, France.

[4]INRAE, BioSP, Site Agroparc, 84914 Avignon, France.

[5]CIRAD, UMR AMAP, F-34398 Montpellier, France.

[6]Université Grenoble Alpes, 621 avenue Centrale, 38400 Saint-Martin-d'Hères, France.

¤Those authors contributed equally to the paper.

‡Those authors contributed equally to the paper.

¶Those authors contributed equally to the paper.

Number of words (including references, tables and figure legends): 6869

Corresponding author: Christophe Botella; **christophe.botella@gmail.com**

Abstract

• Building reliable Species Distribution Models (SDMs) from species occurrence information collected without protocol requires correctly acknowledging the spatial distribution of observation, *i.e.* the sampling effort. In most cases sampling effort is unknown but entails bias. Jointly estimating parameters of this effort and of species densities should allow controlling sampling biases. We propose a method and guidelines to jointly estimate the spatial variation of sampling effort and multiple species densities from massive presences only.

• We define a spatial lattice and estimate (i) the relative sampling effort as a step function across lattice cells, and (ii) multiple species densities as functions of environmental variables. A marked Poisson process models simultaneously multiple species occurrences along with a common factor representing sampling effort. We evaluate estimation performance and robustness to variation inside lattice cells on realistic simulated datasets: We define sampling effort derived from real occurrences, simulate species occurrences, perform estimation, and evaluate it. We also illustrate the method on a real dataset of around 500,000 occurrences from 300 plant species in France, stemming from a large-scale citizen science observatory (Pl@ntNet).

• We show that sampling effort is correctly estimated, in expectation, when the true sampling effort is constant inside lattice cells. We observed bias when the covariation of sampling with covariates inside cells is strong, otherwise the method is robust to sampling variations inside cells. Running the model on real occurrences of 300 plant species provided a relative sampling effort map covering 40% of the French territory, and its spatial variations reached a factor of several thousands. We also show the density estimated for an exotic invasive plant is consistent with prior knowledge and predicts invaded areas that are unknown or likely to be invaded in the future.

• This is the first method estimating sampling effort as an explicit spatial function from multiple species occurrences. It has good scalability and can take advantage of the distributions of most observed species to better infer sampling effort and other species densities. For large opportunistic occurrences datasets, like in citizen-sciences projects, it should be useful to correct for sampling bias and study spatial variations of sampling effort.

Keywords: presence only data; sampling effort; citizen-science; biodiversity monitoring; species distribution modeling; niche models; multi-species models; Poisson point process; sampling bias ; opportunistic data.

1 Introduction

Studying biodiversity dynamics and defining appropriate conservation strategies require characterizing and analyzing species distributions in space and time. Worldwide citizen science projects and naturalist networks provide massive species occurrence data, from numerous active contributors, and thus convey valuable insights into biodiversity patterns. In order to perform ecologically meaningful Species Distribution Models (SDMs, Elith and Leathwick [2009]) with such data, characterizing how observers report presences, *i.e.* their sampling

effort, is crucial. However, both a spatial variation in sampling effort and species ecological niches shape observed species distributions, and an appropriate method is required to disentangle their influences.

Digitized, geolocated species presence records, called occurrences, have first been compiled from digitized expert collections, mainly field naturalist surveys and records in natural history museums (Soberón and Peterson [2004]). Species occurrences have now become massively available through worldwide citizen-science programs or naturalist community platforms (e.g., iNaturalist, e-Bird, Pl@ntNet, Naturgucker, see Chandler et al. [2017]), thanks to new digital tools and smartphone applications (Teacher et al. [2013]). For example, eBird currently shares around 500 millions valid geolocated species occurrences worldwide on GBIF[1]. In addition, automatic identification from images or sound samples (Joly et al. [2018]), and collaborative review of observations have enhanced identification quality of such data. However, these occurrence data are mostly reported without a planned sampling protocol. For example, in the case of Pl@ntNet and iNaturalist, contributors generally submit for identification, to an automatic system and/or a community of members, some specimens that are sampled non-randomly in space and seem remarkable, atypical or new to them. Such sampling, often called "opportunistic" (Kery et al. [2010]), depends on the specific behaviour and reporting choices of contributors. It globally leads to spatially heterogeneous sampling effort. Our objective here is to characterize how spatial sampling heterogeneity affects reported species distributions and the ecological niche inferred from such data. The sampling effort is defined as an intensity function measuring the number of visits during which observers can report a specimen occurrence at a given point. It is sometimes also called "observation effort" (Calenge et al. [2015]). The sampling effort defined in this way does not depend on species detectability or reporting interest (Fithian et al. [2015] and Giraud et al. [2016]).

[1]https://www.gbif.org/

Estimating the spatial variation in sampling effort in a set of species occurrences is crucial for many purposes. The unknown spatial variation in sampling effort can be correlated to an environmental factor and yields biases in SDM results (Botella et al. [2020]). Thus, it is crucial to acknowledge sampling effort in SDM and several approaches have been proposed to tackle this problem. Sampling effort may be approximated from some available information on the sampling scheme. Calenge et al. [2015] thus used the number of driven kilometers reported by agents as a proxy of the relative sampling effort in the process of collecting dead animals occurrences along roads. Alternatively, one can represent sampling effort by the distribution of background points used for inference of environmental density in SDM. Phillips et al. [2009] thus proposed the Target-Group Background (TGB) procedure, where sites with at least one observation among a Target-Group of species provides a proxy of sampling effort. Bradter et al. [2018] proposed using information about the prospecting behaviours and the detection skills of very active reporters to infer true absences of species, and then use the information in a joint model with presence-only data. De Solan et al. [2019] provided another approach to estimate sampling effort in multi-species presence-only SDMs, by using the presence-only data to estimate sampling effort. Finally, Warton et al. [2013] proposed to jointly model the sampling effort along with a single species abundance. Sampling effort is modeled with a set of carefully chosen dedicated variables assumed to be its main drivers. Once the joint model is fitted, the sampling effort can be predicted in space separately from species abundance. Explicitely or implicitely, introduced approaches require using additional information or specific assumptions on sampling effort or species additionnally to the occurrences data themselves to get unbiased estimation.

In this work, we propose a new SDM method for presence-only data, requiring less prior knowledge on the sampling process. It is based on a spatial smoothness assumption on the sampling

effort over units of a spatial mesh, and it jointly estimates the sampling effort with multiple species environmental densities from occurrences data. Indeed, we are doomed to a spatially averaged estimation of sampling effort because we don't exactly know where observers have been. In spatial statistics, bases of spatially smooth functions, called smoothers, are often used to estimate response surfaces in a computationally efficient way when the number of samples is large [Johannesson and Cressie, 2004]. The response surface typically represents non-observed spatially smooth predictors. In our approach, we use a cell-wise constant function to model sampling effort, which can be expressed as a linear combination of spatial B-splines of order 0 [Eilers and Marx, 1996]. We formally demonstrate that the method can alleviate biases on sampling effort and species niches estimates, while allowing computational efficiency for large occurrences datasets. We show that the method works with a realistic size dataset and a realistic profile of sampling effort, derived from real a occurrences density. We also study the method robustness to the crucial assumption of sampling effort constancy inside cells by varying the degree of spatial variation speed and curvature of the sampling effort. We finally illustrate the method outputs on a large dataset of opportunistic plant occurrences automatically identified from pictures coming from a citizen sciences observatory called Pl@ntNet. The use of these data for modeling the distribution of remarkable species like exotic invasive species is promising (Botella et al. [2018]). As an example, we comment the estimated density of an exotic invasive plant species in France.

2 Material and methods

2.1 A spatial model for the sampling effort

We jointly model multiple species occurrences as independent marked Poisson point processes. The density of each species occurrence process is the product of the sampling effort and of the given species density, representing its abundance, which is a function of environmental variables. It corresponds to the presence-only SDM framework introduced in Renner et al. [2015]. The diagram of Figure **??** illustrates the principle and elements of the method and defines the components of the statistical model. Apart from the sampling effort component, our model can be seen as a multi-species version of the model proposed by Warton et al. [2013]. In addition, our model is equivalent to Fithian et al. [2015] if the presence-absence term is removed from their log-likelihood. Our purpose is to devise a model suited to presence-only data, which are more frequently available today. In the following, we present the probabilistic model of occurrences underlying the proposed estimation method, highlight the assumptions,and explain when and how it theoretically improves SDM estimation compared to bias correction in a single species model. We then expose the conditions for proper use of the method, from which guidelines are derived.

Species occurrence processes and density functions. We denote D the two dimensional geographic domain where occurrences have been collected. We consider N species included in the model. The model assumes that the individuals of any species i are distributed over D according to a Poisson process of intensity function λ_i. λ_i is assumed to be a log-linear function of environmental variables defined all over D. We note by the vector $x^i(z) = (x^i_1(z), ..., x^i_{p_i}(z))$ the environmental features of species i at point z, where p_i is the number of features. A

feature is the result of a function applied to an environmental variable. Different features may be derived from a same variable: For instance, the identity and quadratic features together model a gaussian response to a variable. $\beta^i = (\beta^i_1, ..., \beta^i_{p_i})$ are the parameters associated with the features such that the species intensity is written $\lambda_i(z) = exp(\alpha_i + \sum_{k=1}^{p_i} \beta^i_k x^i_k(z))$. For the model estimation, we can only estimate the relative species intensity (*i.e.* its density) across space and across species, not the absolute abundance, so we assume that $\alpha_1 = 0$ by convention. Note that the environmental features vector x^i depends on the species. This formulation of the species density model is probably the most popular when modeling species distribution with point processes (Phillips et al. [2006], Chakraborty et al. [2011], Warton et al. [2013], Dorazio [2014], Renner et al. [2015], Fithian et al. [2015], Koshkina et al. [2017]). It is flexible, because many non-linear transformation of a same initial environmental variable can be integrated to a species features vector.

Sampling effort and occurrences report. We defined earlier the sampling effort as the function over space equal to the number of passages of all observers at a point over a time period potentially. This function can vary spatially at very high resolution, but it makes sense to model it by a random function whose parameter is a smooth spatial intensity. we also assume that the reporting probability is constant in space, time and across observers. More precisely, we assume that the sampling effort at point $z \in D$, noted $s(z)$ models the probability of observing a spatial point z. Then, if an individual is present at z, it is detected and reported with probability R_i, which implies overall that the individual is sampled with probability $R_i\ s(z) \in [0, 1]$. It means that we assume that the probability of sampling species i individuals vary proportionally to s across space, but may be globally higher or lower than any other species. The distribution of observed species occurrences then follows a thinned Poisson process, *i.e.*, a Poisson process of intensity $z \to R_i s(z) \lambda_i(z)$ (Chiu et al. [2013]). We expect

the number of occurrences to be proportional to species abundance while keeping sampling effort constant.

Sampling effort model. We model s as a cell-wise constant function. s is assumed to be constant within the units of a spatial mesh defined over D. This central assumption of our model makes sense if the sampling effort is known to vary reasonably slowly across space, at the scale of mesh cells. In the following experiments, we chose a lattice with square cells for simplicity, but any other type of partition of D could be examined with the same modeling framework. The sampling effort is a factor in the intensity function as shown in equation 2 of Fig. 1. We set, for any point $z \in D$, $s(z) = \exp(\sum_{j\in[|1,C|]} \gamma_j 1_{z\in c_j})$ where $(c_j)_{j\in[|1,C|]}$ are the cells of the mesh verifying $\cup_{j\in[|1,C|]} c_j = D$ and $\cap_{j\in[|1,C|]} c_j = \emptyset$, and $\gamma = (\gamma_1, ...\gamma_C)$ are the sampling effort model parameters to estimate. There is a parameter in $\mathbb{R}$ for each unit of the spatial mesh. Therefore, the sampling effort is defined as a categorical effect associated to the cell identifier. We can only estimate the relative sampling effort across space, and thus we assume by convention that $\gamma_1 = 0$.

Model identifiability and estimability. It is impossible to identify absolute values of R_i, the sampling effort s and the species density λ_i from presence only data. We can only estimate sampling effort and species density up to a constant factor (see Fithian and Hastie [2013], Hastie and Fithian [2013]). R_i is confounded with the intercept of s and λ_i. It is important that our model design enables good parameters estimability [Jacquez and Greif, 1985]. As shown in **Appendix B**, if the features basis composed of the environmental features and sampling cells indicator functions are too much collinear, we will get high covariances between the sampling effort and the species densities parameters estimates. That is, the true densities of sampling effort and species densities will be somehow mixed together in our estimates.

Similarly to the model of Dorazio [2014], we should also control the condition number of the model observed Fisher information matrix (explicitly given in **Appendix A**). We recall that the condition number is the ratio of highest and lowest eigenvalues of this matrix. It must be low and ideally close to one. We advise to select sampling cells and environmental features such that the condition number is inferior to 10^6. Higher resolution sampling cells tend to increase the condition number, and if the sampling cells are nested in the cells of a raster environmental feature, then the model is simply non-identifiable.

Model design guidelines We provide some conditions and recommendations for proper use of the method:

1. There should be at least several tens of occurrences (all species included) per sampling cells included in the model. Otherwise discard the cells, the occurrences within, and do not include any background points over these cells. Alternatively, the size of cells can be increased to meet the condition. Scarce cells integrate as many background points as other cells, but would be useless computational burden to the model and a potential source of variance: The information gain on the sampling effort parameter in a cell is equal to the total number of occurrences in this cell (see **Appendix A**). As the sampling effort in those cells is very uncertain, they don't contribute to reduce the variance on the species parameters.

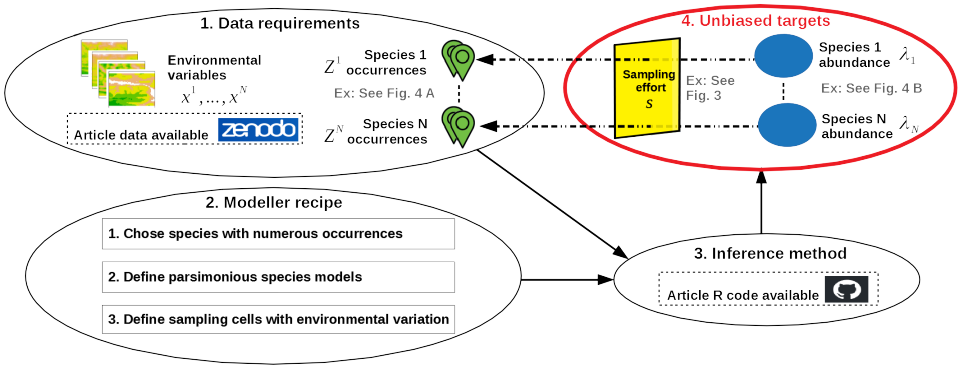
2. There should be at least several tens of occurrences for each environmental feature of each species.

3. For each environmental feature, the standard deviation of this feature over all occurrences divided by the standard deviation over background points should not be too small, at least 1/3 in practice. This is a proxy of the spread of the global occurrence intensity along

the feature gradient. It is a good coverage indication for this feature. The estimation of its parameter with a certain confidence will require all the more occurrences as this indicator is low.

4. Regarding the choice of cell sizes, an optimal compromise should exist, but we have no definite procedure to reach it in practice yet. Three main limits can prevent good estimation when the sampling mesh reach a too high resolution: the estimation variance (see the first point above), the identifiability (discussed earlier) and the memory limitation (number background points required). Conversely, designing too large cells entails more variation of sampling effort inside cells, which tends to favor estimation bias (see section of the results and **Appendix B**, paragraph 2). In practice, a cross-validation scheme should be run for each tested cell area. Decreasing the size of cells can very quickly increase estimation variance of the species parameters, as shown for a simulation example in paragraph 4 of **Appendix C**.

5. It is important to include some highly observed species in the model if available, especially if they have a wide distribution over the territory. Further, an environmental variable should be from the model of a species, if it is known to be generalist along this gradient, so that to (i) reduce the estimation variance for all others species density parameters associated with this gradient as shown in a paragraphs 2 and 3 of **Appendix C**, and (ii) drastically reduce the estimation bias. Generalist species provide a reference for sampling effort along the environmental gradient for the model. Globally, the modeler should include environmental variables parsimoniously to avoid high covariances between sampling effort and species densities estimates.



- **Species distribution model**

The environmental intensity of species i is

$$\begin{array}{rcl} \lambda_i \;:\; \mathbb{R}^{p_i} & \to & \mathbb{R}^+ \\ x^i & \to & \exp(\alpha_i + \sum_{k=1}^{p_i} \beta_k^i x_k^i) \end{array} \tag{1}$$

Where $x^i : D \to \mathbb{R}^{p_i}$ are the environmental variables considered for species i. The intensity represents the species expected abundance in some environment. We assume that the locations of species i individuals Z_i are distributed according to the In-homogeneous Poisson process: $Z_i \sim IPP(\lambda_i \circ x)$.

- **Sampling effort model**

The sampling effort $s : D \to [0,1]$ represents the probability that a punctual location is observed, which doesn't depend on the species (it doesn't include the detection probability, or reporting interest). We approximate it by a step function over a mesh making a partition of D: $s(z) = \exp(\sum_{j\in[|1,C|]} \gamma_j 1_{z\in c_j})$, where $(c_j)_{j\in[|1,C|]}$ are the cells of the mesh. We chose a regular mesh in the following experiments for simplicity, but it is not a requirement.

- **Full model**

A species individual located at $z \in D$ is reported with probability $R_i\, s(z)$ where $R_i \in [0,1]$ is the constant probability of detecting and reporting i. Then, with the thinning property of Poisson processes, the joint probability distribution of the model is:

$$(Z_1, ..., Z_N) \sim \otimes_{i=1}^{N} IPP(s\; R_i\; \lambda_i) \tag{2}$$

- **Main assumptions**

1. Individuals locations are independent given the environment.
2. The probability of detection and reporting of any species is constant in space, time and across observers.
3. The reporting of two individuals at distinct point locations are independent random variables.
4. The proportion of sampled individuals is small everywhere.
5. The sampling effort is constant per sampling cell.

Figure 1: Method workflow summary and statistical model

2.2 Inference

We summarize here the procedure for inferring parameter values from multi-species occurrences data, and the detailed procedure is given in **Appendix D**. A log-linear Poisson processes is fitted over multiple species, but with a shared term in their linear predictor, *i.e.* the log-sampling effort for a given spatial mesh of cells (regular mesh of squares in the following applications). The procedure basically minimizes the sum of negative log-likelihoods of each species' Poisson process. Thus the objective function that is maximised is similar to the one of Fithian et al. [2015], except that the presence-absence term is removed. We use a convergent approximation of the Poisson process likelihood, whose integral term is heavy to compute, by a Poisson regression likelihood [Berman and Turner, 1992], and we use spatially uniformly distributed background points to achieve the estimation of the integral term [Warton et al., 2010]. The implementation is done with the `glmnet` library for `R` (`https://www.r-project.org/`), in a similar way to Renner et al. [2015], except that it is extended to the multi-species case. `glmnet` handles sparse matrices and is very efficient in terms of memory and computational load, given the structure of the model design matrix. The R code for reproducing the results and fitting the model is provided in a Github repository: `https://github.com/ChrisBotella/SamplingEffort`.

2.3 Simulation study

We simulated virtual occurrence datasets to assess the reliability of inferences. The R code to reproduce this simulation study, *i.e.* to generate sampling effort rasters, simulate species occurrences, fit the model and run analysis over all scenarios, is provided in Github repository: `https://github.com/ChrisBotella/SamplingEffort`.

Real study area. French Mediterranean region over the Longitude/Latitude extent $[1.5, 8] \times [41, 45]$ (defining D in our model).

Virtual species densities. We simulated $no = (n_1, ..., n_50)$ occurrences of 50 virtual species, based on occurrence numbers of the 50 most represented plant species in the Pl@ntNet queries dataset (`https://zenodo.org/record/2634137#.XpNmqZngphE`) over D, so that $min(no) = 1502, max(no) = 5002$ and $\sum_i n_i/50 \approx 2206$. All the virtual species densities (λ_i for species i in our model) were defined as Gaussian functions of the same single environmental variable (two cases considered: elevation/**alti** or annual precipitations/**chbio_12**, see **Appendix E**). The mean of the Gaussian density was drawn uniformly inside the quantiles 0.1 and 0.9 of the environmental variable range of values, while the standard deviation was drawn according to a gamma distribution of shape parameter 3 and scale parameter 50. The two environmental variables were chosen because they are both strongly linked to the simulated sampling effort (described further). In addition, the resolution of **alti** variable (around 90 meters) was much finer than **chbio_12** (around 1km), and thus **alti** varies more strongly inside the sampling cells of our model. Therefore, the effect of **alti** was expected to be less well estimated with our method.

Realistic sampling effort. We simulated a realistic spatial distribution of sampling effort while controlling the smoothness of its variation. The sampling effort ($s : D \to \mathbb{R}$ in our model) was derived from the spatial distribution of all automatically identified plant observations in the Pl@ntNet queries data. We applied an exponential quadratic kernel density estimator function to the counts of those occurrences per small square cells (resolution=0.002 in longitude and latitude) over D. This yields a smoothed raster over D with same resolution. We experimented 4 values for the bandwidth parameter $H = \{20, 50, 80, 100\}$ respectively in

cells units, *i.e.* $3.2, 8, 12.8, 16$ kilometers in longitude, or $4.4, 11, 17.6, 22$ kilometers in latitude. Thus we computed everywhere the local sampling effort value as a weighted local average of the surrounding counts with a weight decreasing with distance. The contribution of a count at distance n cells was proportional to $exp(-n^2/H)$, where H is the bandwidth introduced above. For instance, for $H = 20$, the weight decreased of 80% at 3.8km in longitude. At highest bandwidth $H = 100$, we mainly represented the large-scale populations and coastline effects, while at lowest bandwidth $H = 20$ we saw the influence of important rivers and roads connecting cities on sampling effort. Additionally, we devised a fifth profile of sampling effort that was constant with the spatial cells of the models (defined below), and equal to the average of counts within cells. This profile called **H=+Inf**, was used as a reference, and enabled us to check the performance of the method under the best model specification, and to characterize the error only due to estimation variance. The sampling effort sharply decreased at low values for both environmental gradients, **alti** and **chbio_12**, which motivates correction for sampling bias in both simulation scenarios.

Occurrences simulation. For a given species i with spatial intensity $\lambda_i \circ x$, and for a given sampling effort surface s, we draw independently n_i occurrences according to the conditional Poisson process of intensity $s\lambda_i \circ x : D \to \mathbb{R}^+$ through a simple acceptation-rejection algorithm. This procedure was consistent with our model of distribution and observation as described in the Box of Figure 1.

Model fitting. We fitted the proposed method model over the 50 species with a spatial mesh of square cells with $(0.1, 0.1)$ dimensions in (longitude,latitude), or approximately $(8, 11)$ in kilometers. Thus, except for the case where the simulated sampling effort was cell-wise constant, the fitted model was deliberately mis-specified. Indeed, the simulated sampling

effort varied strongly inside cells for the lowest bandwidth $H = 20$, and much more softly for the highest $H = 100$. We also filtered sampling cells and occurrences. We only kept sampling cells where there was at least 50 occurrences overall. The background points were drawn uniformly over cells as explained in **Appendix D**. We drew background points until there was at least 10 per sampling cells. The model was fitted for the following 10 simulation scenarios: Two environmental variables (elevation and precipitations) and 5 sampling effort profiles (including the 4 levels of smoothness and the cell-wise constant sampling effort).

Evaluation of performance. We used two metrics to evaluate the estimation performance of the sampling effort:

1. The coefficient of determination between the simulated sampling effort and its estimation over the points of a fine regular spatial grid across D (around 200 meters resolution).

2. The coefficient of determination between the simulated sampling effort averaged per sampling cell and its estimation over the same points. In other words, this metric computes the correlation with the best possible approximation of the true sampling effort and is necessarily superior to the first.

We also evaluated the estimation performance of species i density parameters as the coefficient of determination between λ_i and its estimate $\hat{\lambda}_i$ across uniformly distributed values of x in the range $[\min\{x(z), z \in D\}, \max\{x(z), z \in D\}]$. We computed the metric over the environmental gradient x rather than over the geographic space D, to avoid biasing evaluation toward the most represented environmental values.

2.4 Illustration with real data

We illustrate the estimation of the relative sampling effort and species density for plant species occurrences from the Pl@ntNet citizen-science project[2]. Geo-located occurrences are collected by citizens using a mobile application (Joly et al. [2016]). They are automatically identified by the Pl@ntNet engine. Details on the current identification system and the database infrastructure are provided in Affouard et al. [2017]. The R code for extracting occurrences and environmental data, and fitting the model is provided in a Github repository: `https://github.com/ChrisBotella/SamplingEffort`.

Species occurrences. The **Pl@ntNet queries 2017-2018 in France** provided the occurrence dataset, which is described and freely downloadable at Botella et al. [2019] (`http://doi.org/10.5281/zenodo.2634137`). Species presence records were collected in France from beginning of 2017 to October 2018 with the Pl@ntNet mobile application. The user takes one or several pictures of a plant specimen organs (e.g. leaf, flower, fruit, or bark). Pictures are then sent to the Pl@ntNet API to carry out automatic identification of the species producing a distribution of probabilities over species. The identification certainty score is the highest of these probabilities. The global Pl@ntNet identification system is described in Affouard et al. [2017] (although the identification engine has regularly evolved since). We first filtered species occurrences whose identification certainty score (field `FirstResPLv2Score`) was above 0.85. Then, we kept only the **300** species with highest number of occurrences. The list of species is provided in the table `speciesTable.csv` of the Github repository [3]. We kept only occurrences no missing values for the selected environmental variables (described below). We then defined a regular spatial grid of squares of 4km side over the French metropolitan

[2]https://plantnet.org/en/
[3]`https://github.com/ChrisBotella/SamplingEffort`

territory including Corsica, and restricted it to squares whose center was inside the territory or closer than 4km from the border or coast. We kept only the squares that had more than 5 occurrences and thus excluded all the occurrences within other squares. We ended up with a set of 475,138 occurrences, distributed over 15,556 spatial squares covering around 40% of the French territory. These squares are colored on the map of Fig. 3. To illustrate the method output regarding species densities, we analysed the fitted density of *Phytolacca americana* L., an exotic invasive plant species in France called, using external available data. We especially referred to the FCBN (National Botanical Conservatories Federation) occurrences which are geographically summarized at : `http://siflore.fcbn.fr/?cd_ref=&r=metro`. It is a national expert dataset and independent of Pl@ntNet.

Environmental data. We selected a set of 9 environmental variables to model the environmental density of species included in the model. The critical point here is that we need a parsimonious number of variables related to the niche of many plant species. Following the recommendations of Mod et al. [2016] on environmental variables for modeling macro ecological species niches, we included mean and annual variation of temperature, annual precipitations, potential evapo-transpiration, elevation, slope, available soil water capacity, a soil pH proxy and a simplified plant habitat type descriptor. The variables are presented in Table 1 of **Appendix E**. These environmental data come from multiple sources [Karger et al., 2016, Panagos, 2006, Panagos et al., 2012, Van Liedekerke et al., 2006, Zomer et al., 2007, 2008]. The local values were extracted from the geographic rasters described and downloadable at Botella [2019] [4].

Species density model. For continuous environmental gradients, the distribution of plant species is often modelled with a Gaussian density function. This choice can be justified be-

[4] `http://doi.org/10.5281/zenodo.2635501`

cause it is a maximum entropy probability density function for which we control independently the mean and variance, two useful criteria for describing species environmental densities. We chose the Gaussian distribution model for the continuous environmental variables: `chbio_1`, `chbio_5`, `chbio_12-etp`, `etp`, `alti` and `slope`. We combined annual rainfall `chbio_12` and potential evapotranspiration `etp` into `chbio_12-etp`, called the water balance, which is commonly used in plants SDM (Mod et al. [2016]). We included categorical pedologic variables representing physico-chemical properties categories. There were 48 categories in the original CORINE Land Cover 2012 classification. To avoid inflating the number of parameters for land cover effects, we defined a Simplified Habitat Typology (`spht`) with 5 types: `cultivated`, `forest`, `grasslands`, `urban` and `other`. Each type included primary CORINE Land Cover 2012 categories as shown in Table 2 of the **Appendix E**. We included an interaction effect between water balance and slope. When this effect is strong, a different slope translates in a different water balance optimum. Thus a plant establishing on steeper slope could compensate for water run-off. To summarize, equation 3 shows the R formula of the linear predictor of any species density, with 19 features terms computed from the environmental variables of Table 1 of **Appendix E**. It thus yields 19 parameters for the density of each species, including the intercept, plus $15,556 - 1$ parameters of observation in sampling cells, yielding $21,255$ parameters in total, for $475,138$ occurrences.

$$\begin{aligned} \sim\ & 1 + \texttt{etp} + \texttt{I(etp}^2\texttt{)} + \texttt{I(chbio_12-etp)} + \texttt{I((chbio_12-etp)}^2\texttt{)} + \texttt{chbio_1} + \texttt{I(chbio_1}^2\texttt{)} \\ & + \texttt{chbio_5} + \texttt{I(chbio_5}^2\texttt{)} + \texttt{alti} + \texttt{I(alti}^2\texttt{)} + \texttt{slope} + \texttt{I(slope}^2\texttt{)} \\ & + \texttt{awc_top} + \texttt{I(awc_top}^2\texttt{)} + \texttt{bs_top} + \texttt{I(bs_top}^2\texttt{)} + \texttt{spht} + \texttt{slope:I(chbio_12-etp)} \end{aligned} \tag{3}$$

Background points. We uniformly drew a fixed number of points per sampling cell as described in **Appendix D**. It avoids the problems of total uniform sampling, *i.e.* cells with no background points. We draw 6 points per sampling cells to account for environmental heterogeneity within cells, which makes 93,336 background points duplicated for each species, that is 28,000,800 background points in total. The model design matrix was then of dimensions (28,475,938 ; 21,255) in the likelihood optimization process. A standard `R` numeric matrix of this dimensions would require around 3,7 To of RAM memory. However, our design matrix was sparse, including only $2 * (p_i + 1) + 1 = 39$ non-null values per line for most lines in our example, so that storage cost was decreased with a factor 500 with the `R` sparse matrix format (see library `Matrix`). The model could be fitted with `R-glmnet` on a machine with 196 Go of RAM.

3 Results

We firstly assessed the reliability of our joint model estimation method over around 100,000 simulated occurrences of 50 virtual species, for 10 simulation scenarios covering the different values of two factors, the environmental gradient (2 types, **alti** and **chbio_12**) and the simulated sampling effort (5 types). We summarized the performance metrics of parameter estimation in Figure 2 for the sampling effort (graph A.) and the 50 species densities (graph B.). Secondly, we illustrated the method on 300 plant species using around 500,000 occurrences in France. Specifically, we examined estimation results for an exotic invasive plant, *Phytolacca americana* L.

Simulation: Very good fit when the simulated sampling effort is cellwise constant. As shown in Figure 2, the estimate of sampling effort had a R^2 of 0.97 (for **alti** and **chbio_12**)

when cellwise sampling effort was constant (**H=+Inf**), while the average R^2 for the species environmental densities was 0.95 for **alti** and 0.88 for **chbio_12**. It shows that the method recovers unbiased niches and sampling effort estimates under good model specifications, and that it is almost unaffected by estimation variance for this size of sample and parameterization. However, the crucial assumption of sampling effort homogeneity within cells, which enables inference with our model, is of course wrong in the reality, and we need to assess the effect of realistic violations of the assumption on estimation performance, as shown below.

Simulation: Smoother is better. Red and black curves of Figure 2 (A.) show that the approximation of the sampling effort was better when the sampling effort was smoother, for both environmental variables. While the red curve represents the fit to the raw sampling effort, the black curve represents the fit to the sampling effort averaged per cell and is always above (*i.e.* the Best Cellwise Constant Approximation (BCCA) of the true sampling effort that can be estimated by the model in the ideal case). The model estimating sampling effort could not fit the variations of sampling effort inside cells, which were integrated as error by the red curve. As H increased, the true sampling effort spatial variation became softer and it was thus closer to be constant inside model cells. This should reduce the gap between the red and the black curve, if the model estimate converges towards the BCCA. It is surprising though that for **x:alti H:-20** the gap was much smaller than for **x:alti H:-50**, but still the R^2 computed with true sampling effort was 0.0044 while it was two times smaller for the average per cell (0.01, unshown).

Bias under joint variation of sampling effort and environmental variable within cells. It is most unlikely that the high error of **x:alti H:-20** was due to estimation variance, as the fit is almost perfect for the cellwise constant effort. The error was most likely due to an

estimation bias when the model of sampling effort cannot fit the variations of occurrence density within cells. Specifically, bias could appear if sampling effort strongly covaried with the environmental feature *within* cells, at least in a restricted range of the gradient. We observed empirically and described such bias in sampling effort profile (3) of the complementary simulation experiment in **Appendix F**, with a visualisation of species and sampling effort density estimates. To simplify, in a single species case, the model is optimized so that the variation of $s_\gamma(z)\lambda_\beta \circ x(z)$ across space (product of the sampling effort and the species density estimates) fits the variation of observed occurrence density $s(z)\lambda \circ x(z)$. However, the best approximation of this product of densities is not necessarily the product of the best approximations per density, namely the BCCA of s and $\lambda \circ x$ itself. We can characterize more accurately this phenomenon in the multi species case with a re-expression and analysis of the asymptotic model negative log-likelihood given in equation (1) of **Appendix A**. By re-expressing the equation with a single environmental variable $x \in \mathrm{Im}(x)$, we obtained the equation 4. For large samples, fitting the model is equivalent to minimizing the right term of equation 4, where the terms $\mathrm{Err}^{W_j}_{s,\lambda^i}(s, s_\gamma)$ and $\mathrm{Err}^{W_j}_{s,\lambda^i}(\lambda^i, \lambda^i_{\beta_i})$ can be seen as logarithmic density errors over the range of environment W_j for the sampling effort and the species i density, respectively. Those errors are spatially weighted by occurrence density of species i, $s\ \lambda^i \circ x$, and its number of occurrences n_i. If sampling effort s is badly approximated by the sampling mesh, *i.e.* by the BCCA, and if s shows a strong and consistent co-variation with x within cells, then $\mathrm{Err}^{W}_{s,\lambda^i}(s, s_\gamma)$ can show monotonic variation along the environmental gradient. The effect can be counterbalanced by an opposite variation profile in the error terms of the species densities, which can be achieved by adjusting their parameters to minimize the overall error. Such lack of robustness of the sampling mesh to environmentally structured variations within cells is a consequence of the latent lack of identifiability of the model. On the contrary, if the sampling effort variation

within cells is independent from that of environmental variables, no bias is caused whatever the strength of sampling effort variation. This problem is related to the problem of spatial confounding in spatial statistics Hodges and Reich [2010], or to interlinked biases between covariates and purely spatial effects in generalized linear mixed models.

$$
\begin{aligned}
\{\hat{\gamma}, \hat{\beta}_1, ..., \hat{\beta}_N\} &= \underset{\gamma,\beta_1,...,\beta_N}{\operatorname{argmin}} \sum_{j=1}^{B} \sum_{i=1}^{N} n_i \left(\mathrm{Err}^{W_j}_{s,\lambda^i_{\beta^*_i}}[s, s_\gamma] + \mathrm{Err}^{W_j}_{s,\lambda^i_{\beta^*_i}}[\lambda^i_{\beta^*_i}, \lambda^i_{\beta_i}] \right) \mu(x^{-1}(W_j)) \\
\text{Where} \quad & (W_j)_{j\in[|1,B|]} \text{ is a partition of } \mathrm{Im}(x) \text{ into small intervals} \\
\text{and} \quad & \forall f, g \in \mathbb{R}+^D \text{ densities over } D \\
\mathrm{Err}^{W}_{s,\lambda}[f, g] &:= \frac{\int_{x^{-1}(W)} s(z)\lambda\circ x(z)(\log(f)-\log(g))dz}{\mu(x^{-1}(W))}
\end{aligned}
\tag{4}
$$

Note that in equation 4, we consider that all densities integrate to 1 over D.

Simulation: Estimation of species densities improves for smoother sampling effort. Figure 2 -B. shows that species responses were on average well estimated in most scenarios, even when sampling effort estimation was worst. In the scenario **x:alti H:20**, the average R^2 of the 50 species densities was around 0.85. In fact, as shown by the asymmetry of density plots in all scenario, most species had good fit with similar performance, while a few other ones had significantly worse fit. As for the sampling effort estimation, estimation quality notably increased with H. Therefore, the robustness issue with sampling effort variation within cells translated into a bias in species estimates. In addition, some species had consistently bad estimation with R^2 below 0.50 even for $H = +\mathrm{Inf}$. It could be the consequence of a simulated niche optimum being in scarcely sampled areas and/or of a lack of occurrences. Species density estimation was overall less good in case **chbio_12** compared to case **alti**, even with good model specification (0.88 for **x:chbio_12 H:+Inf** on average compared to 0.95 for

x:alti H:+Inf on average, see B. of Figure 2) where the sampling effort density estimation is almost perfect. It implies that lower performance is not due to estimation bias, but to estimation variance, due to responses that are harder to estimate given the sampling effort and occurrences. The lower estimation quality with chbio_12 was thus not intrinsically due to the variable itself, but a consequence of species niches (randomnly defined, see the protocol section) that are harder to estimate. It also highlights that even though species estimation can be unbiased, its precision necessarily depends on the overall intensity of sampling, *i.e.* we need a sufficient number of points everywhere (all species confounded) in environmental space to insure homogeneity in the estimation quality across species, as highlighted in section **model design guidelines**.

Illustration: Evaluating sampling heterogeneity in Pl@ntNet data. Fig. 3 shows the estimated log-relative sampling effort of 300 plants species reported in Pl@ntNet between 2017 and 2018. The maximal variations of estimated sampling effort across the territory are of a factor 1000, e.g. Biarritz (a very touristic city) compared to remote sites in the natural reserve of Camargues. We can interpret from the multiplicative model that any species was observed 1000 times more in certain places just because of the frequency of visits and independently of their actual abundance.

Illustration: *Phytolacca americana* L. distribution. The fitted model also provided environmental densities of 300 plant species. We predicted the log relative density for any species i, at any point z where the species variables x_i were all known, by computing its specific linear predictor $\sum_{k=1}^{p_i} \hat{\beta}_k^i x_i^k(z)$. We projected the decimal logarithm of *Phytolacca americana* L. density estimation across all France in graph B of Fig. 4. It is consistent with the knowledge of *Phytolacca* habitat as described in Dumas [2011]: It is cultivated as

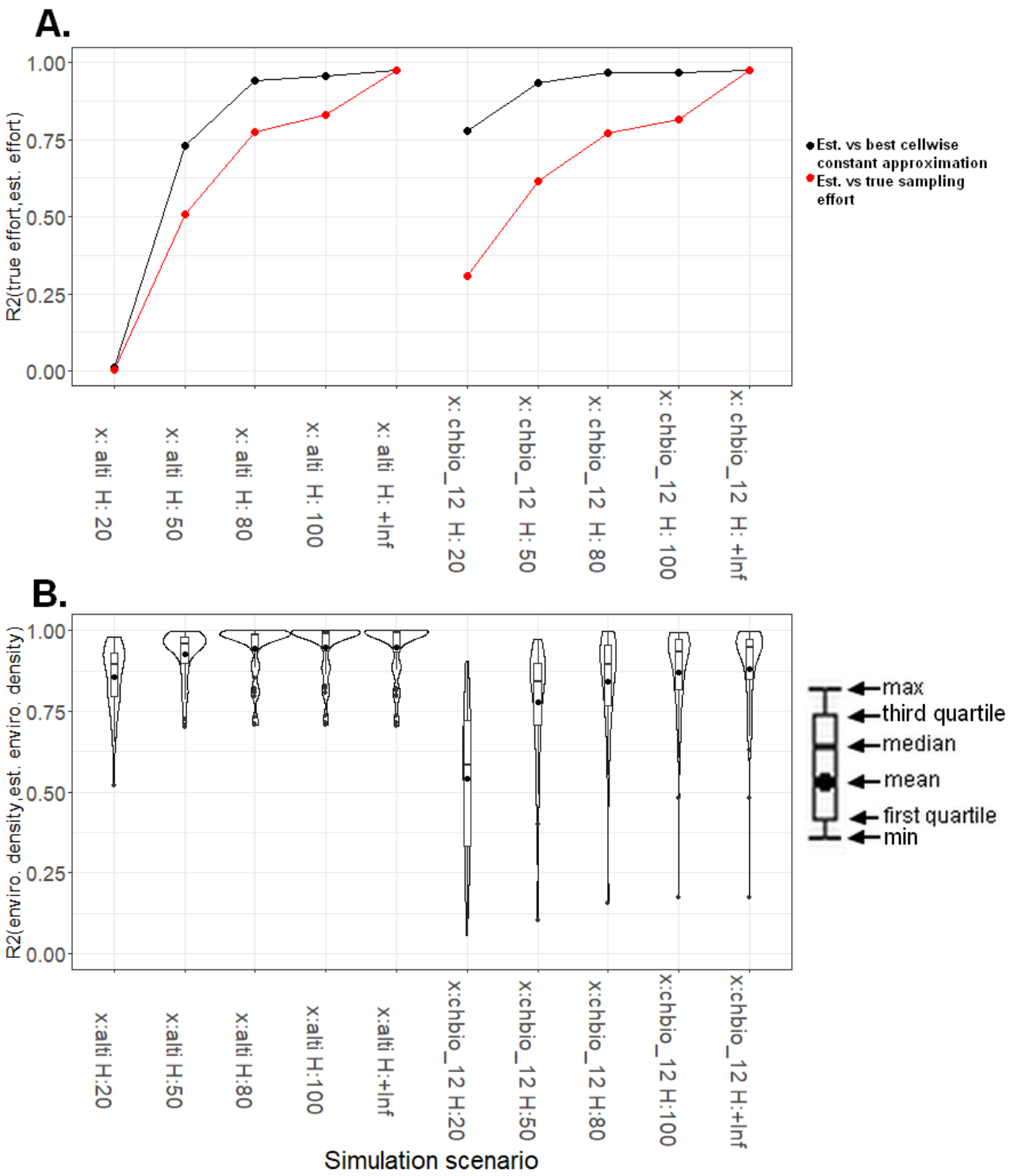


Figure 2: R^2 between generative and estimated model components in the 10 simulation scenarios for the Sampling effort (A.) and the species environmental densities (B.). In (A.) the R^2 is computed between the simulated sampling effort density (raw in red or averaged per estimation cell in black) and the estimated density over the geographic space. Regarding the evaluation of the species densities estimates, the same metric is computed between the true and the estimated densities over the environmental gradient and for the 50 species, that for each scenario. In (B.) we summarize the 50 species metrics values through the boxplot overlayed on a density plot.

ornamental all over France - one of the reasons of its introduction - and often establishes on disturbed soils in the surroundings. In rural areas, it prefers managed forests with acidic and sandy soils. It is also found along rivers bordered with trees, as predicted by the model along the Rhone and the Garonne. The northern France is not favorable to it. The model recovers true hot spots even in scarcely sampled areas. Indeed, the model predicts that the species is abundant in several scarcely sampled departments, like the Indre (36), Aude (11), Charente (16) and the Gers (32). FCBN records from 2000, which can be seen at `http://siflore.fcbn.fr/?cd_ref=113418&r=metro`, confirm that the species is indeed widely present in Indre (36). Conversely, there are very few reports in the INPN data for Aude (11), Charente (16) or Gers (32), although presence records exist (Dumas [2011] and Pl@ntNet occurrences). Those regions have indeed been under-prospected by conservatories experts in the last 20 years. Thus the current *Phytolacca americana* abundance stayed either undetected by convervatories sampling or it is a recent invasion.

We also see that the predicted density of *Phytolacca* within cities is very high. It asks the question of whether (i) it is really due to *Phytolacca*'s higher abundance inside cities or (ii) to residual estimation bias. Hypothesis (ii) is supported by the fact the effect of the urban category of `spht` variable was the highest of all categories for most species, even those avoiding such habitats (e.g. *Vicia faba* L. had lowest urban effect of all but was predicted to be 2 times more intense in cities). The fact that many species of the list were partly gardened may favor such bias. However, effect of (i) is probably also strong because *Phytolacca* is an invasive species used as ornamental or medicinal plant in many gardens, and its urban inflation of factor 24 is much higher than for *Vicia faba*. It emphasizes that the effect of gardening species abundance on urban inflation is strong.

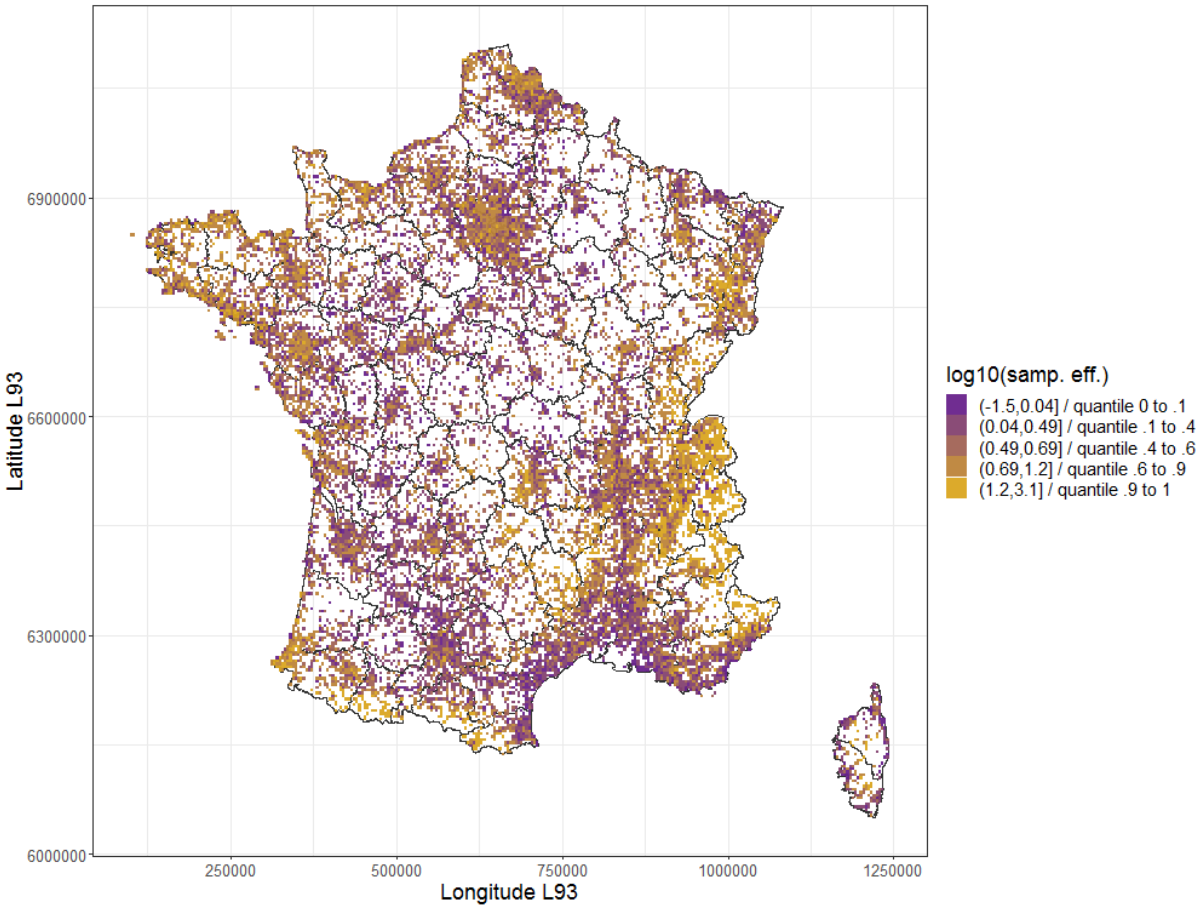


Figure 3: Relative sampling effort estimated from Pl@ntNet occurrences in France. The model was fitted on 475,316 occurrences of 300 plant species in France reported between 2017 and 2018 through the Pl@ntNet application. We represent the logarithm in base 10 of the estimated sampling effort to more clearly show the broad orders of variation. The white cells are those with too few occurrences, which were discarded in the analyses.

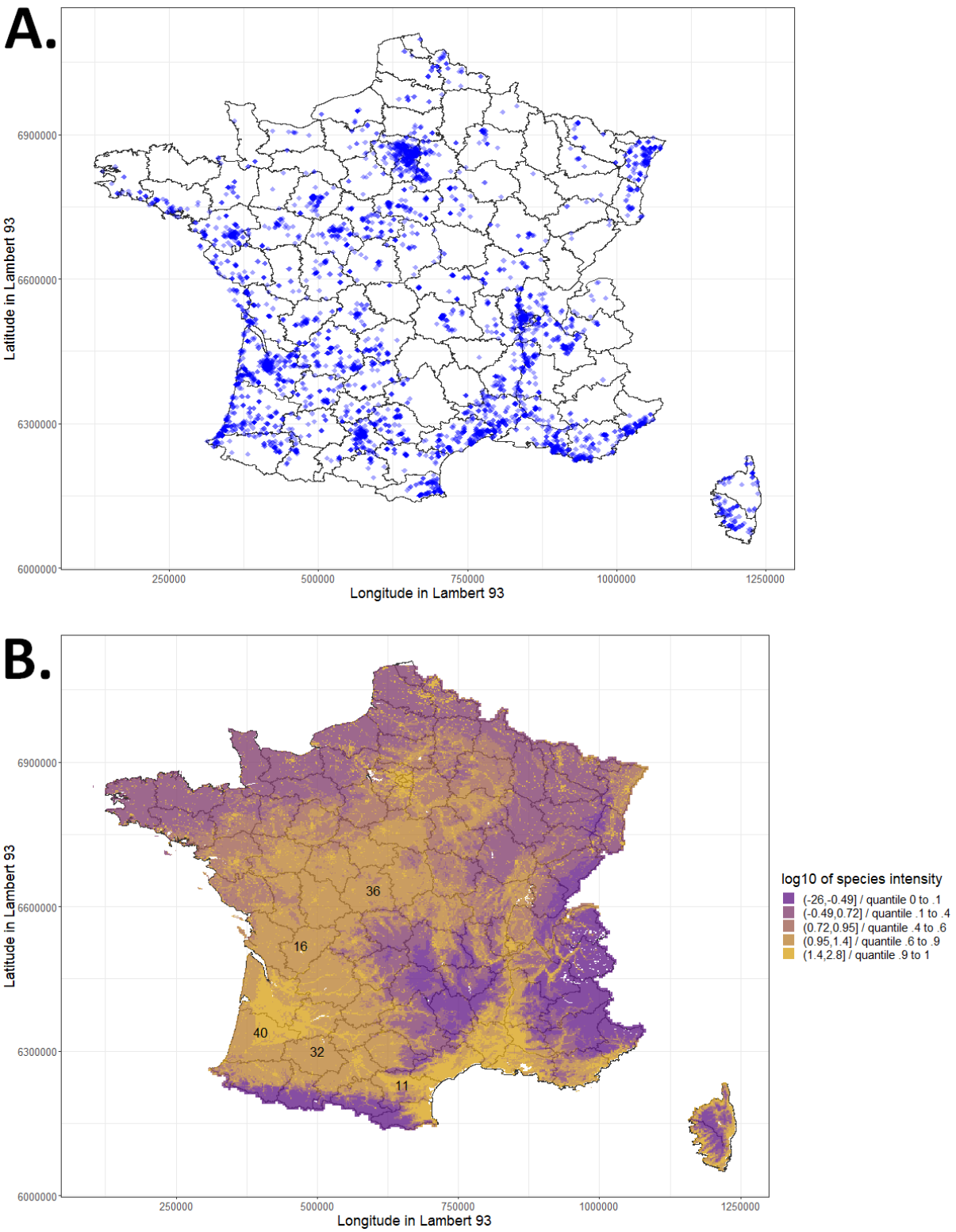


Figure 4: Raw occurrences and estimated density of *Phytolacca americana* L. from Pl@ntNet data. A. 5,176 occurrences of *Phytolacca americana* L. collected through Pl@ntNet users with automatic identification over the 2017-2018 period. B. Decimal logarithm of predicted relative density of *Phytolacca americana* L. across France estimated from the occurrences with the proposed method. The discrete gradient of colors represent quantiles intervals ranges. The model corrects for the over-concentration of occurrences in northern cities due to sampling effort.

4 Discussion

We propose a method to jointly estimate environmental densities of many species, along with a spatial function representing a common sampling effort. By modeling the sampling effort as a step-wise constant function over a spatial mesh, the methods offers flexible estimation of the sampling effort without a priori constraints or knowledge on the spatial determinants. For modelling species densities, the method can be seen as a multi-species extension of the Warton et al. [2013], or a particular case of the method of Fithian et al. [2015] for presence only data. However, our method differs from the FactorBiasOut method based on Target-Group Background (TGB Phillips et al. [2009]). FactorBiasOut is unbiased if the cumulative intensity of TG species is constant across environments [Botella et al., 2020], and in this case it should yield the same estimation as our method. Nevertheless, selecting a Target-Group fulfilling this condition is challenging.

We have shown that our method provides unbiased estimation of species relative densities and sampling effort if the later is constant within the cells of a spatial mesh. Although this condition is crucial to disentangle species and sampling densities, we have also shown that the method is robust to reasonable variation of sampling effort within cells, and even to stronger variation unrelated to environmental drivers of species densities. However, if the sampling effort covaries with an environmental driver within cells, bias in the sampling effort and species estimates is likely to appear. We also demonstrated that information on sampling effort gained from the most observed species helps to better estimate the niche of less observed species with our method. Our method is devised for analysing large volume of occurrences. In the illustration here, we could successfully analyze around 500,000 opportunistic occurrences from 300 plant species distributed all over France, with a total of 20,000 parameters. The

illustration also indicated that sampling effort varied by a factor of several thousands across spatial cells, supporting its ability to handle opportunistic datasets with high variation in sampling effort. The analysis of *Phytolacca americana* L. suggested a broader distribution and potential areas yet undetected based on published knowledge and data. Nevertheless, predictions out of the training area must be critically and carefully examined, as they can present different environmental conditions and be subject to extrapolation errors.

Pooling occurrences to control estimation variance. A major advantage of our method is to require less prior constraints or knowledge on sampling effort, but such flexibility comes at some cost. Our method becomes quickly data hungry when decreasing the spatial area of sampling cells. In the illustration provided here, pooling the occurrences of many species was needed to allow reasonable estimation of the relative sampling effort on $15,556$ cells of area $16km^2$. Our method is not suited when the concentration of occurrences per sampling cell is too low, as for herbarium datasets including few samples collected over large areas with much heterogeneous sampling effort. In such cases, the FactorBiasOut method (Phillips et al. [2009]) should be more reliable because it does not require many degrees of freedom to model sampling effort. However, our methods proved efficient when the global number of occurrences is high and the average number of occurrences per species is high (the minimum per species being reasonable, see our guidelines), because pulling information from all species allow improving the estimation of sampling effort. It is suited for opportunistic datasets where some species are highly observed, for instance with citizen-science or naturalist programs. We have also shown that the highly observed species are all the more useful as they are widely distributed. Estimation variance on a species density can thus be drastically reduced compared to the single species model (**Appendix C**, paragraph 1). An even more efficient way to reduce estimation variance and bias is to withdraw some environmental variables that should not play on a

given species density (**Appendix C**, paragraphs 2-3). This species could allow improving estimation of relative sampling effort over gradients of these environmental variables.

How to efficiently design a spatial mesh for sampling effort? An optimal design of the spatial mesh should ensure that the real sampling effort is constant within every cell. Otherwise, estimation biases can arise (see **Appendix B**, paragraph 2). Although a thinner partition of space should allow more constant sampling effort within cells, a coarser partition should be preferred in order to minimize (i) estimation variance of all parameters, and (ii) parameters co-variances, especially between sampling and species densities, because a more variable environment within cells should allow better distinguishing its influence against a constant sampling effort. A cross-validation scheme combined with a density evaluation metric (see e.g. Tsybakov [2009]) should guide the design of the spatial mesh with homogeneous cell sized, but a more efficient sampling design can integrate heterogeneous cells sizes and shapes. Therefore, optimizing the design of the sampling mesh remains an promising way for improving the method.

How to manage variation in species detection probability? Several assumptions regarding the detection probabilities may deviate from reality. First, the sampling effort is assumed identical across species, but our model can allow varying detection probability across species (R_is), which is still not distinguishable from species global abundance. It means that we assume the detection probability density to vary similarly across space for all species. This assumption is not specific to our method (seeFithian et al. [2015]). However, biases can appear if species detection probability varies differently in space from one species to another. For instance, some species might be looked for only in specific areas and such sampling peculiarity can induce bias in the estimation of the species densities. We also make the assumption

that for each modelled species, the detection probability is identical across observers. Species included in the model should be selected to respect this assumption. It also concerns the problem of heterogeneous identification skills in the case of citizen-sciences data. A rule of thumb is to only include in the model species that are all well identified by all observers, and to ensure that most observers have enough skills to identify species less frequent species. Lastly, we assume the expected number of occurrences to be proportional to the local abundance of the species and the sampling effort. If for instance, observers report at maximum only one individual from the local population, it may impact the estimation of our model. But it will depend on the number of observers that prospect and their distribution. If this number of observers is high (everywhere) and their probability of detection of specimens is globally low, then estimates of our model should not change drastically. However, if the number of observers is low everywhere, and their probability of detection is high, then we expect that the estimation of the environmental density by our model will be shrinked. The assumption seems somehow consistent with the citizen science context, but otherwise, occurrences thinning strategies may be useful to avoid bias (Boria et al. [2014], Fourcade et al. [2014], Varela et al. [2013]) and could be applied for our method.

Scalability. Our method can handle occurrence datasets including many species, occurrences over large geographic and environmental scales as shown by our illustration on the Pl@ntNet data over France. The fit of this model required computations on a very large design matrix of dimensions 29 millions rows by 22 thousand columns. However, thanks to the sampling effort cellwise constant model, lines only had 39 non null columns. This structure is exploited by the sparse matrix format of the `R-glmnet` (Friedman et al. [2010]) R package, minimizing memory use. We could thus use 28,000,800 background points for the likelihood approximation. The memory load increases only linearly with the number of sampling cells

(number of background points $\propto$ number of cells). Memory limitation on standard computers can still appear with more species (several thousands) and/or higher resolution of environmental variables (e.g. derived from satellite imagery) and/or larger domain, because the number of required background points is roughly proportional to the number of species and to the maximum number of environmental raster cells. Optimizing the selection of background points and using a batch gradient descent algorithm should allow handling much larger dimensions.

Bias under covariation of sampling effort and environmental features within cells. With a cellwise constant model of sampling effort and a coarser grain of cells compared to environmental variation, spatial confusion of sampling effort and species densities should be prevented (Hodges and Reich [2010]) . However, the simulation experiment showed that the model can confound the influence of environment and sampling effort on species density. Based on theoretical arguments and simulation experiments (**Appendix F** scenario 2), we showed that there is an approximately linear deviation in the average of log-sampling effort estimate along an environmental feature if the true sampling effort strongly and monotonically covaries within cells with this feature environmental. A similar bias due to heterogeneity in sampling effort related to land-cover and elevation variables within cells is suspected for the illustration on real occurrences. Bias appears when the sampling mesh does not allow to capture strong variation in sampling effort along an environmental gradient playing on species density. Therefore, we recommend not to include an environmental variable if the sampling effort varies very quickly over its most represented range of values in space.

Perspectives to improve joint estimation of sampling effort and species abundances. More generally, joint modeling approaches from presence only are vulnerable to wrong model specification regarding spatial covariation in sampling effort and species density

(discussed in **Appendix B**). There is currently no blindly reliable solution for sampling bias correction with presence only data. An approach that has been recently emphasized is to integrate more standardized data (e.g. presence-absence, counts, occupancy detection, distance sampling) which provide information on species abundance that is not, or less, affected by an unknown sampling process. Such integration is possible within the Poisson process framework as, e.g., with presence-absence data (Giraud et al. [2016] and Fithian et al. [2015]), abundance counts with imperfect detection (Dorazio [2014]), and site occupancy with imperfect detection (Koshkina et al. [2017]). Including standardized data on a small subset of species can be enough to drastically improve joint estimation of sampling effort and species densities (Giraud et al. [2016]). If no standardized dataset is available, one possible bypass is to build complementary site-occupancy (Louvrier et al. [2018]) or absence (Bradter et al. [2018]) data from the opportunistic presences only, using external knowledge to determine a priori the sampled area or constrain detection probabilities. Data integration methods have thus gained momentum Miller et al. [2019], but not all standardized data can equally well improve estimation. For example, there is common observation bias of opportunistic datasets such as Pl@ntNet in urban areas. Overcoming the bias would require integrating standardized species surveys both within and outside cities, because surveys conversely undersampling urban areas would be useless to resolve model confusion. However, it is difficult to find standardized data covering as well areas within and outside cities. It brings us to the question of data collection orientation or selection based on a specific goal of model improvement. Such problem might be cast as the optimization of some function of parameters variances or co-variances over all possible sampling effort distributions given some constraints on the global effort. Similar questions have been defined and found answers in the literature of optimal design (Pukelsheim [2006]). For instance, in the context of our method, our results imply that the estimation would be

much improved only if we could orientate the sampling effort, during data collection, to insure that it constant inside some geographic areas with heterogeneous environment. Those areas should then be used as sampling cells. More broadly, data collection orientation has strong implications for better integrating citizen-science projects in biodiversity monitoring, but there is still little and only recent work on it Reich et al. [2018].

Synthesis. We have shown that our method can estimate sampling effort in geographic space with much less prior assumptions than previous methods. We have also proposed and discussed several possible extensions allowing using the method in a broad range of situations. It is especially suited to analyze observations of many species by many citizens at large spatial scale, and should decrease biases in species distribution estimates. We thus expect that the approach will be useful to recover information of sampling effort from purely opportunistic occurrences, enabling post-analysis of sampling effort variation in citizen science programs and guiding strategies for further data collection. In addition, insofar as citizen science occurrences are generally collected continuously, our method should allow regular monitoring of many taxa and support on-time and adapted conservation and management strategies.

5 Data accessibility

In the aim to follow the FAIR principles, datasets and source code used in this manuscript are provided at the follow urls :

- Species occurrences data may be freely downloaded at `http://doi.org/10.5281/zenodo.2634137`.
- Environmental rasters may be freely downloaded at `http://doi.org/10.5281/zenodo.`

2635501.

- The R code for running simulations and real data illustration, as well as the list of modelled species are provided on the manuscript dedicated Github repository : https://github.com/ChrisBotella/SamplingEffort.

References

Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., and Joly, A. (2017). Pl@ntnet app in the era of deep learning. In *ICLR: International Conference on Learning Representations*. hal-01629195, v1.

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38.

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–77.

Botella, C. (2019). A compilation of environmental geographic rasters for sdm covering france (version 1) [data set]. Zenodo. `http://doi.org/10.5281/zenodo.2635501`.

Botella, C., Bonnet, P., Joly, A., Lombardo, J.-C., and Affouard, A. (2019). Pl@ntnet queries 2017-2018 in france. Zenodo. `http://doi.org/10.5281/zenodo.2634137`.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (2018). Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences*, 6(2):e1029.

Botella, C., Joly, A., Monestiez, P., Bonnet, P., and Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. In *PLOS One*.

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A., and Snäll, T. (2018). Can oppor-

tunistically collected citizen science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution*, 9(7):1667–1678.

Calenge, C., Chadoeuf, J., Giraud, C., Huet, S., Julliard, R., Monestiez, P., Piffady, J., Pinaud, D., and Ruette, S. (2015). The spatial distribution of mustelidae in france. *PloS one*, 10(3):e0121689.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):757–776.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., et al. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, 213:280–294.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

De Solan, T., Renner, I., Cheylan, M., Geniez, P., and Barnagaud, J.-Y. (2019). Opportunistic records reveal mediterranean reptiles' scale-dependent responses to anthropogenic land use. *Ecography*, 42(3):608–620.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484.

Dumas, Y. (2011). Que savons-nous du raisin d'amérique (phytolacca americana), espèce exotique envahissante ? synthèse bibliographique. *Rendez-vous techniques ONF, 2011*, pages 48–57.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS one*, 9(5):e97122.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Giraud, C., Calenge, C., Coron, C., and Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, 72(2):649–658.

Hastie, T. and Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227.

Johannesson, G. and Cressie, N. (2004). Finding large-scale spatial trends in massive, global, environmental datasets. *Environmetrics: The official journal of the International Environmetrics Society*, 15(1):1–44.

Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., Dufour-Kowalski, S., Affouard, A., Carré, J., Molino, J.-F., et al. (2016). A look inside the pl@ ntnet experience. *Multimedia Systems*, 22(6):751–766.

Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.-P., Planqué, R., and Müller, H. (2018). Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 247–266. Springer.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N., Linder, H. P., and Kessler, M. (2016). Climatologies at high resolution for the earth's land surface areas. *arXiv preprint arXiv:1607.00217*.

Kery, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Haefliger, G., and Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, 24(5):1388–1397.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430.

Louvrier, J., Duchamp, C., Lauret, V., Marboutin, E., Cubaynes, S., Choquet, R., Miquel, C., and Gimenez, O. (2018). Mapping and explaining wolf recolonization in france using dynamic occupancy models and opportunistic data. *Ecography*, 41(4):647–660.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10:22–37.

Mod, H. K., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. *Journal of Vegetation Science*, 27(6):1308–1322.

Panagos, P. (2006). The european soil database. *GEO: connexion*, 5(7):32–33.

Panagos, P., Van Liedekerke, M., Jones, A., and Montanarella, L. (2012). European soil data centre: Response to european policy support and public data requirements. *Land Use Policy*, 29(2):329–338.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.

Pukelsheim, F. (2006). *Optimal design of experiments*. Society for Industrial and Applied Mathematics.

Reich, B. J., Pacifici, K., and Stallings, J. W. (2018). Integrating auxiliary data in opti-

mal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, 9(6):1626–1637.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

Soberón, J. and Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):689–698.

Teacher, A. G., Griffiths, D. J., Hodgson, D. J., and Inger, R. (2013). Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecology and Evolution*, 3(16):5268–5278.

Tsybakov, A. (2009). Introduction to nonparametric estimation. In *Springer Series in Statistics, ISBN 978-0-387-79051-0*. Springer-Verlag New York.

Van Liedekerke, M., Jones, A., and Panagos, P. (2006). Esdbv2 raster library-a set of rasters derived from the european soil database distribution v2. 0. *European Commission and the European Soil Bureau Network, CDROM, EUR*, 19945.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2013). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37:1084–1091.

Warton, D. I., Renner, I. W., and Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168.

Warton, D. I., Shepherd, L. C., et al. (2010). Poisson point process models solve the "pseudo-

absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402.

Zomer, R. J., Bossio, D. A., Trabucco, A., Yuanjie, L., Gupta, D. C., and Singh, V. P. (2007). *Trees and water: smallholder agroforestry on irrigated lands in Northern India*, volume 122. IWMI.

Zomer, R. J., Trabucco, A., Bossio, D. A., and Verchot, L. V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, ecosystems & environment*, 126(1):67–80.

8 Chapter 4: A Deep Learning Approach to Species Distribution Modelling

Chapter 10
A Deep Learning Approach to Species Distribution Modelling

Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz

Abstract Species distribution models (SDM) are widely used for ecological research and conservation purposes. Given a set of species occurrence, the aim is to infer its spatial distribution over a given territory. Because of the limited number of occurrences of specimens, this is usually achieved through environmental niche modeling approaches, i.e. by predicting the distribution in the geographic space on the basis of a mathematical representation of their known distribution in environmental space (= realized ecological niche). The environment is in most cases represented by climate data (such as temperature, and precipitation), but other variables such as soil type or land cover can also be used. In this paper, we propose a deep learning approach to the problem in order to improve the predictive effectiveness. Non-linear prediction models have been of interest for SDM for more

C. Botella (✉)
INRIA Sophia-Antipolis - ZENITH Team, LIRMM - UMR 5506 - CC 477, Montpellier, France

INRA, UMR AMAP, Montpellier, France

AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France

BioSP, INRA, Site Agroparc, Avignon, France

A. Joly
Inria ZENITH Team, Montpellier, France
e-mail: alexis.joly@inria.fr

P. Bonnet
CIRAD, UMR AMAP, Montpellier, France

AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France
e-mail: pierre.bonnet@cirad.fr

P. Monestiez
BioSP, INRA, Site Agroparc, Avignon, France
e-mail: pascal.monestiez@inra.fr

F. Munoz
Université Grenoble Alpes, Saint-Martin-d'Hères, France
e-mail: francois.munoz@cirad.fr

© Springer International Publishing AG, part of Springer Nature 2018

A. Joly et al. (eds.), *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, Multimedia Systems and Applications,
https://doi.org/10.1007/978-3-319-76445-0_10

than a decade but our study is the first one bringing empirical evidence that deep, convolutional and multilabel models might participate to resolve the limitations of SDM. Indeed, the main challenge is that the realized ecological niche is often very different from the theoretical fundamental niche, due to environment perturbation history, species propagation constraints and biotic interactions. Thus, the realized abundance in the environmental feature space can have a very irregular shape that can be difficult to capture with classical models. Deep neural networks on the other side, have been shown to be able to learn complex non-linear transformations in a wide variety of domains. Moreover, spatial patterns in environmental variables often contains useful information for species distribution but are usually not considered in classical models. Our study shows empirically how convolutional neural networks efficiently use this information and improve prediction performance.

10.1 Introduction

10.1.1 Context on Species Distribution Models

Species distribution models (SDM) have become increasingly important in the last few decades for the study of biodiversity, macro ecology, community ecology and the ecology of conservation. An accurate knowledge of the spatial distribution of species is actually of crucial importance for many concrete scenarios including the landscape management, the preservation of rare and/or endangered species, the surveillance of alien invasive species, the measurement of human impact or climate change on species, etc. Concretely, the goal of SDM is to infer the spatial distribution of a given species based on a set of geo-localized occurrences of that species (collected by naturalists, field ecologists, nature observers, citizen sciences project, etc.). However, it is usually not possible to learn that distribution directly from the spatial positions of the input occurrences. The two major problems are the limited number of occurrences and the bias of the sampling effort compared to the real underlying distribution. In a real-world dataset, the raw spatial distribution of the observations is actually highly correlated to the preference and habits of the observers and not only to the spatial distribution of the species. Another difficulty is that in most cases, we only have access to presence data but not to absence data. In other words, occurrences inform that a species was observed at a given location but never that it was not observed at a given location. Consequently, a region without any observed specimen in the data remains highly uncertain. Some specimens could live there but were not observed, or no specimen live there but this information is not recorded. Finally, knowing abundance in space doesn't give information about the ecological determinants of species presence.

For all these reasons, SDM is usually achieved through *environmental niche modeling* approaches, i.e. by predicting the distribution in the geographic space on the basis of a representation in the environmental space. This environmental space is in most cases represented by climate data (such as temperature, and precipitation), but also by other variables such as soil type, land cover, distance to water, etc. Then,

the objective is to learn a function that takes the environmental feature vector of a given location as input and outputs an estimate of the abundance of the species. The main underlying hypothesis is that the abundance function is related to the *fundamental ecological niche* of the species, in the sense of Hutchinson (see [1]). That means that in theory, a given species is likely to live in a single privileged ecological niche, characterized by an unimodal distribution in the environmental space. However, in reality, the abundance function is expected to be more complex. Many phenomena can actually affect the distribution of the species relative to its so called *abiotic* preferences. For instance, environment perturbations, or geographical constraints, or interactions with other living organisms (including humans) might have encourage specimens of that species to live in a different environment. As a consequence, the *realized ecological niche* of a species can be much more diverse and complex than its hypothetical fundamental niche.

10.1.2 Interest of Deep and Convolutional Neural Networks for SDM

Notations When talking about environmental input data, there could be confusions between their different possible formats. Without precision given, x will represent a general input environmental variable which can have any format. When a distinction is made, x will represent a vector, while an array is always noted X. To avoid confusions on notations for the different index kinds, we note the spatial **site** index as superscript on the input variable (x^k or X^k for k^{th} site) and the component index as subscript (so x_j^k for the jth component of kth site vector $x_k \in \mathbb{R}^p$, or for the array $X^k \in \mathscr{M}_{d,e,p}(\mathbb{R})$, $X^k_{.,j,.}$ is the jth matrix slice taken on its second dimension). When we denote an input associated with a precise **point location** taken in a continuous spatial domain, the point z is noted as argument: $x(z)$.

Classical SDM approaches postulate that the relationship between output and environmental variables is relatively simple, typically of the form:

$$g(\mathbb{E}[y|x]) = \sum_j f_j(x_j) + \sum_{j,j'} h_{j,j'}(x_j, x_{j'}) \tag{10.1}$$

where y is the response variable targeted, a presence indicator or an abundance in our case, the x_j's are components of a vector of environmental variables given as input for our model, f_j are real monovariate functions of it, $h_{j,j'}$ are bivariate real functions representing pairwise interactions effects between inputs, and g is a link function that makes sure $\mathbb{E}[y|x]$ lies in the space of our response variable y. State-of-the-art classification or regression models used for SDM in this way include GAM [2], MARS [3] or MAXENT [4, 5]. Thanks to f_j, we can isolate and understand the effect of the environmental factor x_j on the response. Often, pairwise effects form of $h_{j,j'}$ is restricted to products, like it is the case in the very popular model MAXENT. It facilitates the interpretation and limits the dimensionality of model

parameters. However, it sets a strong prior constraint without a clear theoretical founding as the explanatory factors of a species presence can be related to complex environmental patterns.

To overcome this limitation, deep feedforward neural networks (NN) [6] are good candidates, because their architecture favor high order interactions effects between the input variables, without constraining too much their functional form thanks to the depth of their architecture. To date, deep NN have shown very successful applications, in particular image classification [7]. Until now, to our knowledge, only one-layered-NN's have been tested in the context of SDM (e.g. in [8] or [9]). If they are able to capture a large panel of multivariate functions when they have a large number of neurons, their optimization is difficult, and deep NN have been shown empirically to improve optimization and performance (see section 6.4.1 in [6]). However, NN overfit seriously when dealing with small datasets, which is the case here (≈5000 data), for this reason we need to find a way to regularize those models in a relevant way. An idea that is often used in SDM (see for example [10]) and beyond is to mutualize the heavy parametric part of the model for many species responses in order to reduce the space of parameters with highest likelihood. To put it another way, a NN that shares last hidden layer neurons for the responses of many species imposes a clear constraint: the parameters must construct high level ecological concepts which will explain as much as possible the abundance of all species. These high-level descriptors, whose number is controlled, should be seen as environmental variables that synthesize the most relevant information in the initial variables.

Another limitation of models described by Eq. (10.1) is that they don't capture spatial autocorrelation of species distribution, nor the information of spatial patterns described by environmental variables which can impact species presence. In the case of image recognition, where the explanatory data is an image, the variables, the pixels, are spatially correlated, as are the environmental variables used in the species distribution models. Moreover, the different channels of an image, RGB, can not be considered as being independent of the others because they are conditioned by the nature of the photographed object. We can see the environmental variables of a natural landscape in the same way as the channels of an image, noting that climatic, soil, topological or land use factors have strong correlations with others, they are basically not independent of each other. Some can be explained by common mechanisms as is the case with the different climatic variables, but some also act directly on others, as is the case for soil and climatic conditions on land use in agriculture, or the topology on the climate. These different descriptors can be linked by the concept of ecological environment. Thus, the heuristic that guides our approach is that the ecological niche of a species can be more effectively associated with high level ecological descriptors that combine non linearly the environmental variables on one hand, and the identification of multidimensional spatial patterns of images of environmental descriptors on the other hand. Convolutional neural networks (CNN, see [11]) applied to multi-dimensional spatial rasters of environmental variables can theoretically capture those, which makes them of particular interest.

10.1.3 Contribution

This work is the first attempt in applying deep feedforward neural networks and convolutional neural networks in particular to species distribution modeling. It introduces and evaluates several architectures based on a probabilistic modeling suited for regression on count data, the Poisson regression. Indeed, species occurrences are often spatially degraded in publicly available datasets so that it is statistically and computationally more relevant to aggregate them into counts. In particular, our experiments are based on the count data of the National Inventory for Nature Protection (INPN[1]), for 50 plant species over the metropolitan French territory along with various environmental data. Our models are compared to MAXENT, which is among the most used classical model in ecology. Our results first show how mutualizing model features for many species prevent deep NN to overfit and finally allow them to reach a better predictive performance than the MAXENT baseline. Then, our results show that convolutional neural networks performed even better than classical deep feedforward networks. This shows that spatially extended environmental patterns contain relevant extra information compared to their punctual values, and that species generally have a highly autocorrelated distribution in space. Overall, an important outcome of our study is to show that a restricted number of adequately transformed environmental variables can be used to predict the distribution of a huge number of species. We believe the study of the high-level environmental descriptors learned by the deep NNs could help to better understand the co-abundance of different species, and would be of great interest for ecologists.

10.2 A Deep Learning Model for SDM

10.2.1 A Large-Scale Poisson Count Model

In this part, we introduce the statistical model which we assume generates the observed data. Our data are species observations without sampling protocol and spatially aggregated on large spatial quadrat cells of 10×10 km. Thus, it is relevant to see them as counts.

To introduce our proposed model, we first need to clarify the distinction between the notion of "obsvered abundance" and "probability of presence". Abundance is a number of specimens relatively to an area. In this work, we model species *observed abundance* rather than *probability of presence* because we work with presence only data and without any information about the sampling process. Using presence-absence models, such as logistic regression, could be possible but it would

[1] http://https://inpn.mnhn.fr/.

require to arbitrarily generate absence data. And it has been shown that doing so can highly affect estimation and give biased estimates of total population [12]. Working with observed abundance doesn't bias the estimation as long as the space if homogeneously observed and we don't look for absolute abundance, but rather relative abundance in space.

The observed abundance, i.e. the number of specimens of a plant species found in a spatial area, is very often modeled by a Poisson distribution in ecology: when a large number of seeds are spread in the domain, each being independent and having the same probability of growing and being seen by someone, the number of observed specimens in the domain will behave very closely to a Poisson distribution. Furthermore, many recent SDM models, especially MAXENT as we will see later, are based on inhomogeneous Poisson point processes (IPP) to model the distribution of species specimens in an heterogeneous environment. However, when geolocated observations are aggregated in spatial quadrats ($\approx 10 \times 10$ km each in our case), observations must be interpreted as count per quadrats. If we consider K quadrats named $(s_1, \ldots, s_K)$ (we will call them sites from now), with empty intersection, and we consider observed specimens are distributed according to $\mathscr{IPP}(\lambda)$, where λ is a positive function defined on $\mathbb{R}^p$ and integrable over our study domain D (where x is known everywhere), we obtain the following equation:

$$\forall k \in [|1, K|], N(s_k) \sim \mathscr{P}\left(\int_{s_k} \lambda(x(z))dz\right) \tag{10.2}$$

Now, in a parametric context, for the estimation of the parameters of λ, we need to evaluate the integral by computing a weighted sum of λ values taken at quadrature points representing all the potential variation of λ. As our variables x are constant by spatial patches, we need to compute λ on every point with a unique value of x inside s_k, and to do this for every $k \in [|1, K|]$. This can be very computationally and memory expensive. For example, if we take a point per square km (common resolution for environmental variables), it would represent 518,100 points of vector, or patch, input to extract from environmental data and to handle in the learning process. At the same time, environmental variables are very autocorrelated in space, so the gain in estimation quality can be small compared to taking a single point per site. Thus, for simplicity, we preferred to make the assumption, albeit coarse, that the environmental variables are constant on each site and we take the central point to represent it. Under this assumption, we justify by the following property the Poisson regression for estimating the intensity of an IPP.

Property The inhomogeneous Poisson process estimate is equivalent to a Poisson regression estimate with the hypothesis that $x(z)$ is constant in any given site of the domain.

Proof We note $z_1, \ldots, z_N \in D$ the N species observations points, K the number of disjoints sites making a partition of D, and assumed to have an equal area. We write the likelihood of $z_1, \ldots, z_N$ according to the inhomogeneous poisson process of intensity function $\lambda \in (\mathbb{R}^+)^D$:

$$p(z_1, \ldots, z_N|\lambda) = p(N|\lambda) \prod_{i=1}^{N} p(z_i|\lambda)$$

$$= \frac{(\int_D \lambda)^N}{N!} \exp\left(-\int_D \lambda\right) \prod_{i=1}^{N} \frac{\lambda(x(z_i))}{\int_D \lambda}$$

$$= \frac{\exp\left(-\int_D \lambda\right)}{N!} \prod_{i=1}^{N} \lambda(x(z_i))$$

We transform the likelihood with the logarithm for calculations commodity:

$$\log(p(z_1, \ldots, z_N|\lambda)) = \sum_{i=1}^{N} \log\left(\lambda(x(z_i))\right) - \int_D \lambda - \log(N!)$$

We leave the $N!$ term, as it has no impact on the optimisation of the likelihood with respect to the parameters of λ. Now, $\int_D \lambda$ simplifies to a sum, as $x(z)$ is constant inside each site of D :

$$\sum_{i=1}^{N} \log\left(\lambda(x(z_i))\right) - \int_D \lambda = \sum_{i=1}^{N} \log\left(\lambda(x(z_i))\right) - \sum_{k \in \text{Sites}} \frac{|D|}{K} \lambda(x^k)$$

$$= \sum_{k \in \text{Sites}} n_k \log\left(\lambda(x^k)\right) - \frac{|D|}{K} \lambda(x^k)$$

Where n_k is the number of species occurrences that fall in site k. We can aggregate the occurrences that are in a same site because x is the same for them. We can now factorize $|D|/K$ on the whole sum, which brings us, up to the factor, to the poisson regression likelihood with pseudo-counts $Kn_k/|D|$.

$$= \frac{|D|}{D} \sum_{k \in \text{Sites}} \frac{Dn_k}{|D|} \log\left(\lambda(x^k)\right) - \lambda(x^k)$$

So maximizing this log-likelihood is exactly equivalent to maximizing the initial Poisson process likelihood. □

Proof uses the re-expression of the IPP likelihood, inspired from [13], as that of the associated Poisson regression. In the following parts, we always consider that, for a given species, the number y of specimens observed in a site of environmental input x is as follows:

$$y \sim \mathscr{P}(\lambda_{m,\theta}(x)) \tag{10.3}$$

Where m is a model architecture with parameters θ.

From Eq. (10.3), we can write the likelihood of counts on K different sites $(x_1, \ldots, x_K)$ for N independently distributed species with abundance functions $\lambda_{m_1,\theta_1}, \ldots, \lambda_{m_N,\theta_N} \in (\mathbb{R}^+)^{\mathbb{R}^p}$, respectively determined by models $(m_i)_{i\in[|1,N|]}$ and parameters $(\theta_i)_{i\in[|1,N|]}$:

$$p\left((y_k^i)_{i\in[|1,N|],k\in[|1,K|]}|(\lambda_{m_i,\theta_i})_{i\in[|1,N|]}\right) = \prod_{i=1}^{N}\prod_{k=1}^{K} \frac{(\lambda_{m_i,\theta_i}(x_k))^{y_k^i}}{y_k^i!} \exp(-\lambda_{m_i,\theta_i}(x_k))$$

Which gives, when eliminating $\log(y_k^i)!$ terms (which are constant relatively to models parameters), the following negative log-likelihood :

$$\mathscr{L}\left((y_k^i)_{i\in[|1,N|],k\in[|1,K|]}|(\lambda_{m_i,\theta_i})_{i\in[|1,N|]}\right) := \sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_{m_i,\theta_i}(x_k) - y_k^i \log(\lambda_{m_i,\theta_i}(x_k)) \tag{10.4}$$

Following the principle of maximum likelihood, for fitting a model architecture, we minimize the objective function given in Eq. (10.4) relatively to parameters θ.

10.2.2 Links with MAXENT

For our experiment, we want to compare our proposed models to a state of the art method commonly used in ecology. We explain in the following why and how we can compare the chosen reference, MAXENT, with our models.

MAXENT [4, 5] is a popular SDM method and related software for estimating relative abundance as a function of environmental variables from presence only data points. This method has proved to be one of the most efficient in prediction [14], while guaranteeing a good interpretability thanks to the simple elementary form of its features and its variable selection procedure. The form of the relative abundance function belongs to the class described in Eq. (10.1). More specifically:

$$\log\left(\lambda_{MAX,\theta}(x)\right) = \alpha + \sum_{j=1}^{p}\sum_{s=1}^{S} f_j^s(x_{(j)}) + \sum_{j<j'} \beta_{j,j'} x_j x_j' \tag{10.5}$$

where $x_{(j)}$ is the jth component of vector x. The link function is a logarithm, and variables interactions effects are product interactions. If x_j is a quantitative variable the functions $(f_s)_{s\in[|1,S|]}$ belongs to four categories: linear, quadratic, threshold and hinge. One can get details on the hinges functions used in MAXENT in [15]. If x_j is categorical, then f_j takes a different value for every category, with one zero category.

It has been shown that MAXENT method is equivalent to the estimation of an IPP intensity function with a specific form and a weighted L1 penalty on its variables [16]. Let's call $\lambda_{MAX,\theta}(x)$ the intensity predicted by MAXENT with parameters θ at x. Last property says that on any given dataset, $\hat{\theta}$ estimated from a Poisson regression (aggregating observations as counts per site) is the same as the one of the IPP (each observation is an individual point, even when there are several at a same site). In our experiments, we ran MAXENT using the `maxnet` package in R [17], with the default regularization, and giving to the function :

1. A positive point per observation of the species.
2. A pseudo-absence point per site.

MAXENT returns only the parameters of the $(f_j^s)_{s,j}$ and the $(\beta_{j,j'})_{j<j'}$, but not the intercept α, as it is meant to only estimate the absolute abundance. We don't aim at estimating absolute abundance either, however, we need the intercept to measure interesting performance metrics across all the compared models. To resolve this, for each species, we fitted the following model using the `glm` package in R as a second step:

$$y \sim \mathscr{P}\left(\exp(\alpha + \log(p))\right)$$

Where α is our targeted intercept, p is the relative intensity prediction given by MAXENT at the given site, and y is the observed number of specimens at this site.

10.2.3 SDM Based on a Fully-Connected NN Model

We give in the following a brief description of the general structure of fully-connected NN models, and how we decline it in our tested deep model architecture.

10.2.3.1 General Introduction of Fully-Connected NN Models

A deep NN is a multi-layered model able to learn complex non-linear relationship between an input data, which in our case will be a vector $x \in \mathbb{R}^p$ of environmental variables that is assumed to represent a spatial site, and output variables $y_1, \ldots, y_N$, which in our case is species counts in the spatial site. The classic so called **fully-connected** NN model is composed of one or more **hidden layer(s)**, and each layer is

composed of one or more **neuron(s)**. We note $n(l, m)$ the number of neurons of layer l in model architecture m. m parameters are stored in θ. In the first layer, each neuron is the result of a parametric linear combination of the elements of x, which is then transformed by an **activation function** a. So for a NN m, $a_m^{1,j}(x, \theta) := a(x^T \theta_j^1)$ is called **the activation** of jth neuron of the first hidden layer of m when it is applied to x. Thus, on the lth layer with $l > 1$, the activation of the jth neuron is $a((\theta_j^l)^T a_m^{l-1,\cdot})$. Now, we understand that the neuron is the unit that potentially combines every variables in x, and, its activation inducing a non-linearity to the parametric combination, it can be understood as a particular basis function in the p dimensional space of x. Thus, the model is able to combine as many basis functions as there are neurons in each layer, and the basis functions become more and more complex when going to further layers. Finally, these operations makes m theoretically able to closely fit a broad range of functions of x.

Learning of model parameters is done through optimization (minimization by convention) of an objective function that depends on the prediction goal. Optimization method for NN parameters θ is based on stochastic gradient descent algorithms, however, the loss function gradient is approximated by the back-propagation algorithm [18].

Learning a NN model lead to a lot of technical difficulties that have been progressively dealt with during last decade, and through many different techniques. We present some that have been of particular interest in our study. A first point is that there are several types of activation functions, the first one introduced being the sigmoid function. However, the extinction of its gradient when $x^T \theta_j^1$ is small or big, has presented a serious problem for parameters optimization in the past. More recently, the introduction of the ReLU [19] activation function helped made an important step forward in NNs optimization. A second point is that when we train a NN model, simultaneous changes of all the parameters lead to important change in the distribution (across the dataset) of each activation of the model. This phenomenon is called internal covariate shift, and perturbs learning importantly. Batch-Normalization [20] is a technique that significantly reduces internal covariate shift and help to regularize our model as well. It consists of a parameterized centering and reduction of pre-activations. This facilitates optimization and enables to raise the learning rate leading to a quicker convergence. At the same time, it has a regularization effect because the centering and reduction of a neuron activation is linked to the mini-batch statistics. The mini-batch selection being stochastic at every iteration, a neuron activation is stochastic itself, and the model will not rely on it when it has no good effect on prediction.

10.2.3.2 Models Architecture in This Study

For a given species i, When we know the model parameter θ, we can predict the parameter of the Poisson distribution of the random response variable $y_i \in \mathbb{N}$, i.e. the count of species i, conditionally on its corresponding input x, with the formula :

$$\lambda_{m,\theta}(x) = \exp(\gamma_i^T a_m^{N_h,\cdot}(x, \theta)) \tag{10.6}$$

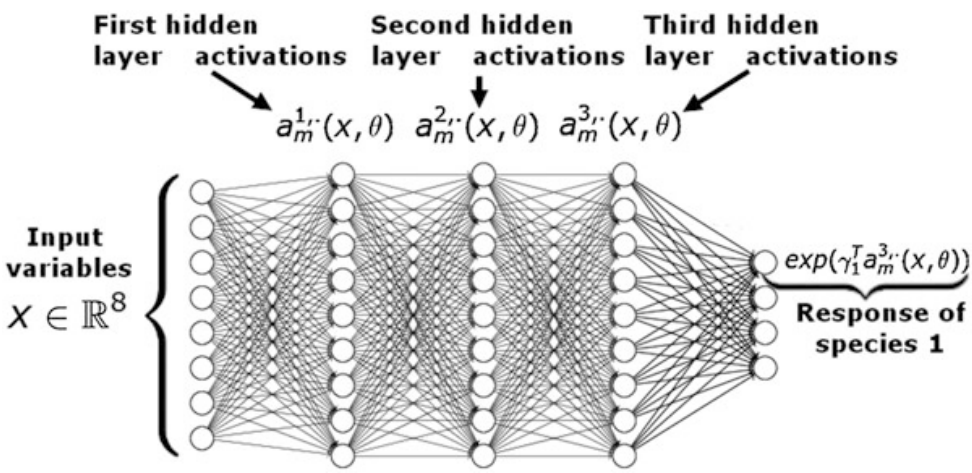


Fig. 10.1 A schematic representation of fully-connected NN architecture. Except writings, image comes from Michael®Nielsen[2].

For this work, we chose the logarithm as link function g mentioned in **1.2**. It is the conventional link function for the generalized linear model with Poisson family law, and is coherent with MAXENT. $\gamma_i \in \mathbb{R}^{n(N_h,m)}$ is included in θ. It does the linear combinations of last layer neurons activations for the specific response i. If we set $n(N_h, m) := 200$ as we do in the following experiments, there are only 200 parameters to learn per individual species, while there are a lot more in the shared part of the model that builds $a_m^{N_h,\cdot}(x,\theta)$. Now for model fitting, we follow the method of the maximum likelihood, **the objective function** will be a negative-loglikelihood, but it could otherwise be some other prediction error function. Note that we will rather use the term **loss function** than negative loglikelihood for simplicity. We chose **the ReLU as activation function**, because it showed empirically less optimization problems and a quicker convergence. Plus, we empirically noticed the gain in optimization speed and less complications with the learning rate initialization when using Batch-Normalization. For this reason, Batch-Normalization is applied to every pre-activation (before applying the ReLU) to every class of NN model in this paper, even with CNNs. We give a general representation of the class of NN models used in this work in Fig. 10.1.

10.2.4 SDM Based on a Convolutional NN Model

A convolutional NN (CNN) can be seen as a extension of NN that are particularly suited to deal with certain kind of input data with very large dimensions. They are of particular interest in modeling species distribution, because they are able to capture the effect of spatial environmental patterns. Again, we will firstly describe the general form of CNN before going to our modeling choices.

[2]http://neuralnetworksanddeeplearning.com/chap6.html

10.2.4.1 General Introduction of CNN Models

CNN is a form of neural network introduced in [11]. It aims to efficiently apply NN to input data of large size (typically 2D or 3D arrays, like images) where elements are spatially auto-correlated. For example, using a fully-connected neural network with 200 neurons on an input RGB image of dimensions $256 \times 256 \times 3$ would imply around $4 * 10^7$ parameters only for the first layer, which is already too heavy computationally to optimize on a standard computer these days. Rather than applying a weight to every pixel of an input array, CNN will apply a **parametric discrete convolution**, based on a kernel of reasonable size (3/3/p or 5/5/p are common for N/N/p input arrays) on the input arrays to get an intermediate feature map (2D). The convolution is applied with a moving windows as illustrated in Fig. 10.2b. Noting $\mathbf{X} \in \mathscr{M}_{d,d,p}$ an input array, we simplify notations in all that follows by writing $\mathscr{CV}(X, k_\gamma(c))$ the resulting feature map from applying the convolution with (c, c, p) kernel of parameters $\gamma \in \mathbb{R}^{c^2 p}$. If the convolution is applied directly on $\mathbf{X}$, the sliding window will pass its center over every $X_{i,j,.}$ from the up-left to the bottom-right corner and produce a feature map with a smaller size than the input because $c > 1$. The **zero-padding operation** removes this effect by adding $(c - 1)/2$ layers of 0 on every side of the array. After a convolution, there can be a Batch-Normalization and an activation function is generally applied to each pixel of the features maps. Then, there is a synthesizing step made by the **pooling** operation. Pooling aggregates groups of cells in a feature map in order to reduce its size and introduce invariance to local translations and distortions. After having composed these operations several times, when the size of feature maps is reasonably small (typically reaching 1 pixel), a **flattening** operation is applied to transform the 3D array containing all the feature maps into a vector. This features vector will then be given as input to a fully-connected layer as we described in last part. The global concept underlying convolution layers operations is that first layers act as low level interpretations of the signal, leading to activations for salient or textural patterns. Last layers, on their side, are able to detect more complex patterns, like eyes or ears in the case of a face picture. Those high levels features have much greater sense regarding predictions we want to make. Plus, they are of much smaller dimension than the input data, which is more manageable for a fully-connected layer.

10.2.4.2 Constitution of a CNN Model for SDM

The idea which pushes the use of CNN models for SDM is that complex spatial patterns like a water network, a valley, etc., can affect importantly the species abundance. This kind of pattern can't be really deducted for punctual values of environmental variables. Thus, we have chosen to build a SDM model which takes as input an array with a map of values for each environmental variable that is used in the other models. This way, we will be able to conclude if there is extra relevant

information in environmental variables spatial patterns to predict better species distribution. In Fig. 10.2a, we show for a single site a subsample of environmental variables maps taken as input by our CNN model. To provide some more detail about the model architecture, the input array X is systematically padded such that the feature map resulting from the convolution is of same size as 2 first dimensions of the input ($(c-1)/2$ cells of 0 after on the sides of the 2 dimensions). To illustrate that, our padding policy is the same as the one illustrated in the example given in Fig. 10.2b. However, notice that the kernel size can differ and the third dimension size of input array will be the number of input variables or feature maps. For an example of For the reasons described in **2.3**, **we applied a Batch-Normalization** to each feature map (same normalization for every pixels of a map) before the activation, which **is still a ReLU**. For the pooling operation, we chose the **average pooling** which seems intuitively more relevant to evaluate an abundance (=concentration). The different kinds of operations and their succession in our CNN model are illustrated in Fig. 10.2c.

10.3 Data and Methods

10.3.1 Observations Data of INPN

This paper is based on a reference dataset composed of count data collected and validated by French expert naturalists. This dataset, referred as INPN[3] for "national inventory of natural heritage" [21], comes from the GBIF portal.[4] It provides access to occurrences data collected in various contexts including Flora and regional catalogs, specific inventories, field note books, and prospections carried out by the botanical conservatories. In total, the INPN data available on the GBIF contains 20,999,334 occurrences, covering 7626 species from which we selected 1000 species.

The assets of this data are the quality of their taxonomic identification (provided by an expert network), their volume and geographic coverage. Its main limitation, however, is that the geolocation of the occurrences was degraded (for plant protection concerns). More precisely, all geolocations were aggregated to the closest central point of a spatial grid composed of 100 km^2 quadrat cells (i.e. sites of 10×10 km). Thus, the number of observations of a species falling in a site gives a count.

In total, our study is based on 5181 sites, which are split in 4781 training sites for fitting models, and 400 test sites for validating and comparing models predictions.

[3] https://inpn.mnhn.fr.

[4] https://www.gbif.org/.

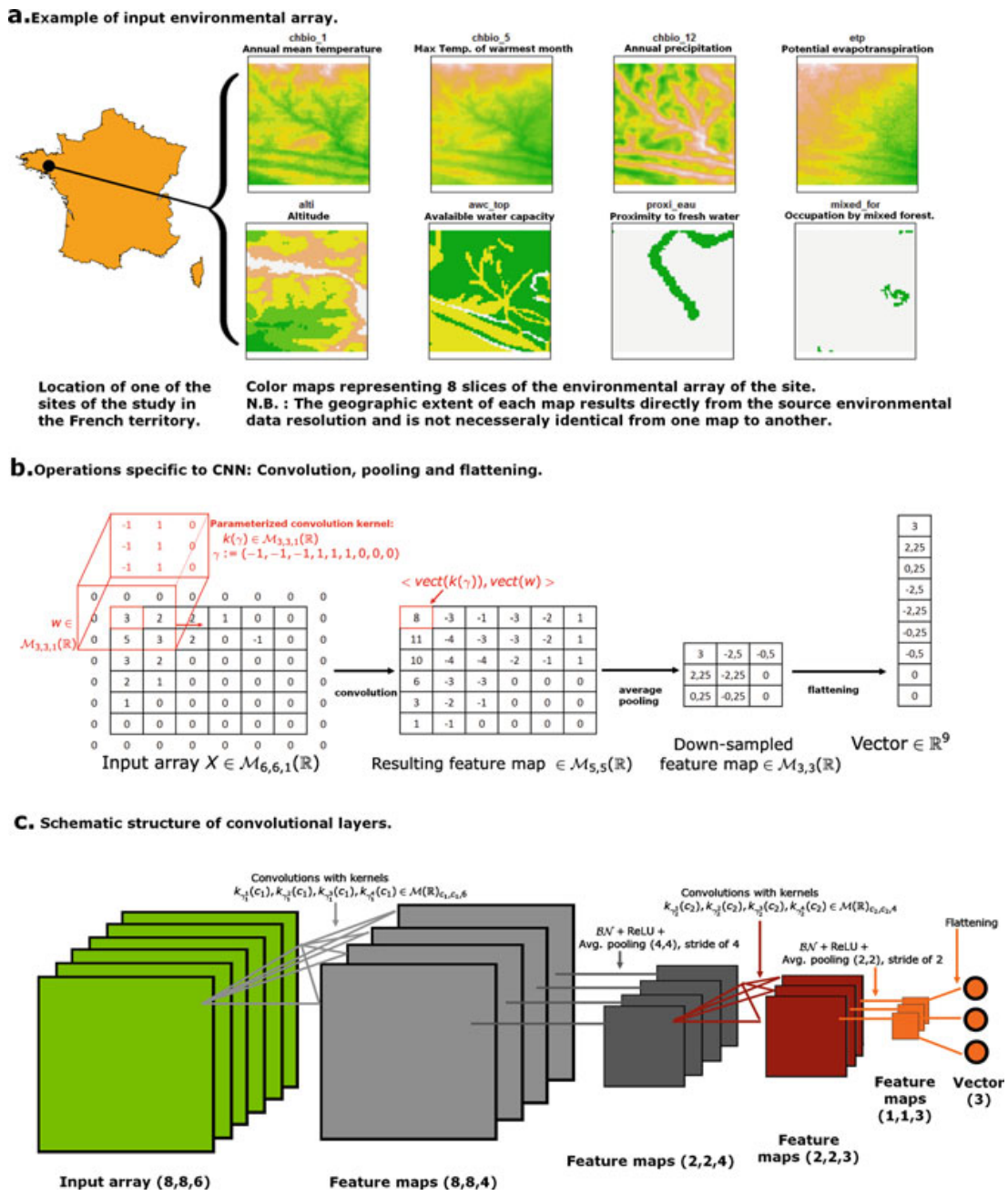


Fig. 10.2 (**a**) Examples of input environmental data (**b**) for convolution, pooling and flattening process in our (**c**) Convolutional Neural Network architecture

10.3.2 Species Selection

For the genericity of our results and to make sure they are not biased by the choice of a particular category of species, we have chosen to work with a high number of randomly chosen species. From the 7626 initial species, we selected species with more than 300 observations. We selected amongst those a random subset of 1000 species to constitute an ensemble E_{1000}. Then, we randomly selected 200 species amongst E_{1000} to constitute E_{200}, and finally randomly selected 50 in E_{200} which

gave E_{50}. E_{50} being the main dataset used to compare our model to the baselines, we provide in Fig. 10.1 the list of species composing it. The full dataset with species of E_{1000} contains 6,134,016 observations in total (see Table 10.1 for the detailed informations per species).

10.3.3 Environnemental Data

In the following, we denote by p the number of environmental descriptors. For this study, we gathered and compiled different sources of environmental data into $p = 46$ geographic rasters containing the pixel values of environmental descriptors presented in Table 10.2 with several resolutions, nature of values, but having a common cover all over the metropolitan French territory. We chose some typical environmental descriptors for modeling plant distribution that we believe carry relevant information both as punctual and spatial representation. They can be classified as bioclimatic, topological, pedologic hydrographic and land cover descriptors. In the following, we briefly describe the sources, production method, and resolution of initial data, and the contingent specific post-process for reproducibility.

10.3.3.1 Climatic Descriptors: Chelsea Climate Data 1.1

Those are raster data with worldwide coverage and 1 km resolution. A mechanistical climatic model is used to make spatial predictions of monthly mean-max-min temperatures, mean precipitations and 19 bioclimatic variables, which are downscaled with statistical models integrating historical measures of meteorologic stations from 1979 to today. The exact method is explained in the reference papers [22] and [23]. The data is under Creative Commons Attribution 4.0 International License and downloadable at (http://chelsa-climate.org/downloads/).

10.3.3.2 Potential Evapotranspiration: CGIAR-CSI ETP Data

The CGIAR-CSI distributes this worldwide monthly potential-evapotranspiration raster data. It is pulled from a model developed by Antonio Trabucco [24, 25]. Those are estimated by the Hargreaves formula, using mean monthly surface temperatures and standard deviation from WorldClim 1:4 (http://www.worldclim.org/), and radiation on top of atmosphere. The raster is at a 1km resolution, and is freely downloadable for a nonprofit use at:

http://www.cgiar-csi.org/data/global-aridity-and-pet-database#description

Table 10.1 List of species in E_{50} with the total number of observations and prevalence in the full database

Taxon name	Total # obs.	Prevalence
Alisma plantago-aquatica L.	15,324	56.3
Alopecurus geniculatus L.	5703	31.5
Antennaria carpatica (Wahlenb.) Bluff & Fingerh	1780	4.0
Anthrisen sylvestris (L.) Hoffm.	27,381	64.9
Astragalus hypoglottis L.	1901	5.7
Berteroa incana (L.) DC.	3966	11.2
Biscutella brevicaulis Jord.	450	1.0
Campanula spicata L.	544	1.7
Carduus vivariensis Jord.	1577	7.4
Carex ericctorum Pollich	538	1.8
Carlina acanthifolia All.	6214	10.6
Centranthus augustifolius (Mill.) DC.	2755	5.9
Cladanthus mixtus (L.) Chevall.	637	5.3
Coronilla coronata L.	325	0.9
Cynoglossum creticum Mill.	1470	9.2
Cytisus villosus Pourr.	562	1.0
Dianthus pyrenaicus Pourr.	392	0.8
Epilobium alpestre (Jacq.) Krocker	1197	3.5
Euphorbia dendroide L.	747	0.5
Festuca cinerea Vill.	3795	5.3
Galium lucidum All.	3204	11.7
Galium timeroyi Jord.	1362	6.6
Helictotrichon sedenense (Clarion ex DC.) Holub	8498	5.4
Hieracium lawsonii Vill.	629	3.2
Hieracium praecox Sch.Bip.	998	4.7
Iris lutescens Lam.	2537	6.6
Juncus trifidus L.	3570	3.9
Lathyrus niger (L.) Bernh.	2474	13.8
Myrtus communis L.	2054	1.9
Meconopsis cambrica (L.) Vig.	1291	3.8
Oxalis corniculata L.	5628	37.5
Oxytropis fetida (Vill.) DC.	315	1.0
Persicaria vivipara (L.) Rouse Decraene	11,122	5.9
Phleum alpinurn L.	7267	6.3
Potamogeton coloratus Hornem.	813	5.5
Potentilla pusilla Host	655	1.7
Primula latifolia Lapeyr.	1268	1.8
Psilurus incurvus (Gouan) Schinz & Thell.	597	4.2
Ranunculus parnassifolius L.	371	1.0
Ranunculus repens L.	76,346	83.0
Reseda lutea L.	16,756	49.0

(continued)

Table 10.1 (continued)

Taxon name	Total # obs.	Prevalence
Rorippa pyrenaica (All.) Rchb.	2169	9.2
Rubus ulmifolius Schott	14,523	35.5
Thalictrum aquilegifolium L.	2855	8.8
Thalictrum alpinum L.	581	1.0
Trifolium micranthum Viv.	767	8.0
Valerianella rimosa Bast.	1518	13.8
Vicia onobrychioides L.	1602	6.3
Viola lactea Sm.	520	4.7
Visearia vulgaris Bernh.	781	3.2

10.3.3.3 Pedologic Descriptors: The ESDB v2: 1 km × 1 km Raster Library

The library contains multiple soil pedology descriptor raster layers covering Eurasia at a resolution of 1 km. We selected 11 descriptors from the library. More precisely, those variables have ordinal format, representing physico-chemical properties of the soil, and come from the PTRDB. The PTRDB variables have been directly derived from the initial soil classification of the Soil Geographical Data Base of Europe (SGDBE) using expert rules. SGDBE was a spatial relational data base relating spatial units to a diverse pedological attributes of categorical nature, which is not useful for our purpose. For more details, see [26, 27] and [28]. The data is maintained and distributed freely for scientific use by the European Soil Data Centre (ESDAC) at http://eusoils.jrc.ec.europa.eu/content/european-soil-database-v2-raster.

10.3.3.4 Altitude: USGS Digital Elevation Data

The Shuttle Radar Topography Mission achieved in 2010 by Endeavour shuttle managed to measure digital elevation at three arc second resolution over most of the earth surface. Raw measures have been post-processed by NASA and NGA in order to correct detection anomalies. The data is available from the U.S. Geological Survey, and downloadable on the Earthexplorer (https://earthexplorer.usgs.gov/). One can refer to https://lta.cr.usgs.gov/SRTMVF for more informations.

10.3.3.5 Hydrographic Descriptor: BD Carthage v3

BD Carthage is a spatial relational database holding many informations on the structure and nature of the french metropolitan hydrological network. For the purpose of plants ecological niche, we focus on the geometric segments representing

Table 10.2 Table of 46 environmental variables used in this study

Name	Description	Nature	Values	Resolution
CHBIO_1	Annual mean temperature	quanti.	[−10.6, 18.4]	30
CHBIO_2	Mean of monthly max(temp)-min(temp)	quanti.	[7.8,21.0]	30
CHBIO_3	Isothermality (100*chbio_2/chbio_7)	quanti.	[41.2,60.0]	30
CHBIO_4	Temperature seasonality (std. dev.*100)	quanti.	[302,778]	30
CHBIO_5	Max temperature of warmest month	quanti.	[36.4,6.2]	30
CHBIO_6	Min temperature of coldest month	quanti.	[−28.2, 5.3]	30
CHBIO_7	Temperature annual range (5–6)	quanti.	[16.7,42.0]	30
CHBIO_8	Mean temperature of wettest quarter	quanti.	[−14.2, 23.0]	30
CHBIO_9	Mean temperature of driest quarter	quanti.	[−17.7, 26.5]	30
CHBIO_10	Mean temperature of warmest quarter	quanti.	[−2.8, 26.5]	30
CHBIO_11	Mean temperature of coldest quarter	quanti.	[−17.7, 11.8]	30
CHBIO_12	Annual precipitation	quanti.	[318,2543]	30
CHBIO_13	Precipitation of wettest month	quanti.	[43.0,285.5]	30
CHBIO_14	Precipitation of driest month	quanti.	[3.0,135.6]	30
CHBIO_15	Precipitation seasonality (Coef. of Var.)	quanti.	[8.2,26.5]	30
CHBIO_16	Precipitation of wettest quarter	quanti.	[121,855]	30
CHBIO_17	Precipitation of driest quarter	quanti.	[20,421]	30
CHBIO_18	Precipitation of warmest quarter	quanti.	[19.8,851.7]	30
CHBIO_19	Precipitation of coldest quarter	quanti.	[60.5,520.4]	30
etp	Potential evapotranspiration transpiration	quanti.	[133,1176]	30
alti	Elevation	quanti.	[−188, 4672]	3
awc_top	Topsoil available water capacity	ordinal	{0, 120, 165, 210}	30
bs_top	Base saturation of the topsoil	ordinal	{35, 62, 85}	30
cec_top	Topsoil cation exchange capacity	ordinal	{7, 22, 50}	30
crusting	Soil crusting class	ordinal	[|0, 5|]	
dgh	Depth to a gleyed horizon	ordinal	{20, 60, 140}	30
dimp	Depth to an impermeable layer	ordinal	{60, 100}	30
erodi	Soil erodibility class	ordinal	[|0, 5|]	30
oc_top	Topsoil organic carbon content	ordinal	{1, 2, 4, 8}	30
pd_top	Topsoil packing density	ordinal	{1, 2}	30
text	Dominant surface textural class	ordinal	[|0,5|]	30
proxi_eau	<50 meters to fresh water	bool.	{0, 1}	30
arti	Artificial area: clc ∈ {1, 10}	bool.	{0, 1}	30
semi_arti	Semi-artificial area: clc ∈ {2, 3, 4, 6}	bool.	{0, 1}	30
arable	Arable land: clc ∈ {21, 22}	bool.	{0, 1}	30
pasture	Pasture land: clc ∈ {18}	bool.	{0, 1}	30
brl_for	Broad-leaved forest: clc ∈ {23}	bool.	{0, 1}	30
coni_for	Coniferous forest: clc ∈ {24}	bool.	{0, 1}	30
mixed_for	Mixed forest: clc ∈ {25}	bool.	{0, 1}	30
nat_grass	Natural grasslands: clc ∈ {26}	bool.	{0, 1}	30
moors	Moors: clc ∈ {27}	bool.	{0, 1}	30

(continued)

Table 10.2 (continued)

Name	Description	Nature	Values	Resolution
sclero	Sclerophyllous vegetation: clc $\in$ {28}	bool.	{0, 1}	30
transi_wood	Transitional woodland-shrub: clc $\in$ {29}	bool.	{0, 1}	30
no_veg	No or few vegetation: clc $\in$ {31, 32}	bool.	{0, 1}	30
coastal_area	Coastal area: clc $\in$ {37, 38, 39, 42, 30}	bool.	{0, 1}	30
ocean	Ocean surface: clc $\in$ {44}	bool.	{0, 1}	30

watercourses, and polygons representing hydrographic fresh surfaces. The data has been produced by the *Institut National de l'information Géographique et forestière* (IGN) from an interpretation of the BD Ortho IGN. It is maintained by the SANDRE under free license for non-profit use and downloadable at:
http://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FX
From this shapefile, we derived a raster containing the binary value of variable `proxi_eau`, i.e. proximity to fresh water, all over France. We used qgis to rasterize to a 12.5 m resolution, with a buffer of 50 m, the shapefile `COURS_D_EAU.shp` on one hand, and the polygons of `*SURFACES_HYDROGRAPHIQUES.shp` with attribute NATURE="Eau douce permanente" on the other hand. We then created the maximum raster of the previous ones (So the value of 1 correspond to an approximate distance of less than 50 m to a watercourse or hydrographic surface of fresh water).

10.3.3.6 Land Cover: Corine Land Cover 2012, Version 18.5.1, 12/2016

It is a raster layer describing soil occupation with 48 categories across Europe (25 countries) at a resolution of 100 m. This classification is the result of an interpretation process from earth surface high resolution satellite images. This data base of the European Union is freely accessible online for all use at http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012 and commonly used for the purpose of plant distribution modeling. For a need of meaningfull variables at our scale and reduced memory consumption, we reduced the number of categories to 14 following mainly the procedure of They eliminate some categories of few interest, too rare or inaccurate, and groups categories that are associated with similar plant communities. In addition, we introduce a category "Semi artificial surfaces", which regroups perturbed natural areas, interesting for the study of alien invasive species. We keep the category "Sea and ocean" from the Corine Land Cover classification because it can be an important contextual variable for the convolutional neural network model. The final categories groups are detailed in Table 10.2. for each of the retain categories, we created a raster of the same resolution as the original one, where the value 1 means the pixel belongs to the category, or the value is 0 otherwise.

10.3.3.7 Environmental Variables Extraction and Format

When creating the p global GeoTIIF rasters, as the original coordinate system of the layer vary among sources, we change it if necessary to WGS84 using `rgdal` package on R, which is the coordinate system INPN occurrences databases. As explained previously, for computational reasons considering the scale, and simplicity, we chose to represent each site by a single geographic point, and chose the center of the site. We are going to compare two types of models. For a site k, the first takes as input a vector of p elements which values are those of the environmental variables taken at the geolocation of the center of the site k, while the other takes p rasters of size (d,d) cropped (with package `raster`) from the global raster of each environmental descriptors and centered at the center of k. If we denote $res_{\text{lon},j}$ the spatial resolution in longitude of global raster of the $j_t h$ environmental descriptor, and $res_{\text{lat},j}$ its resolution in latitude, the spatial extent of $X^k_{.,.,j}$ is $(d.res_{\text{lat},j} \times d.res_{\text{lon},j})$. As a consequence, the extents are heterogeneous across environmental descriptors. In this study, we experimented the method with $d = 64$, so the input data items X^k learned by our convolutional model is of dimension $64 \times 64 \times 46$.

10.3.4 Detailed Models Architectures and Learning Protocol

MAXENT is learned independently on every species of E_{50}. Similarly, we fit a classic loglinear model to give a naive reference. Then, two architectures of NN are tested, one with a single hidden layer (SNN), one with six hidden layers (DNN). Those models take a vector of environmental variables x^k as input. As introduced previously, we want to evaluate if training a multi-response NN model, i.e. a NN predicting several species from a single $a_m^{N_h(m)}(x, \theta)$, can prevent overfitting. One architecture of CNN is tested, which takes as input an array X^k. Hereafter, we described more precisely the architecture of those models.

10.3.4.1 Baseline Models

- **LGL** Considering a site k, and its environmental variables vector x^k, the output function λ_{LGL} of the loglinear model parametrized by $\beta \in \mathbb{R}^p$ is simply the exponential of a scalar product between x^k and β :

$$\lambda_{LGL}(x^k, \beta) = \exp\left(\beta^T x^k\right)$$

As LGL has no hidden layer, we learned a multi-response model, which is equivalent to fitting the 50 mono-response models independently.

- **MAXENT**.

10.3.4.2 Proposed Models Based on NN

- **SNN** has only 1 hidden layer ($N_h = 1$) with 200 neurons ($|a^1_{SNN}| = 200$) all batch-normalized and the activation function is ReLU. As the architecture is not deep, it makes a control example to evaluate when stacking more layers. SNN is tested in 3 multi-response versions, on E_{50}, E_{200} or E_{1000}.
- **DNN** is a deep feedforward network with $N_h = 6$ hidden layers and $n(l, \text{DNN}) = 200, \forall l \in [|1, 6|]$. Every pre-activation is Batch-normalized and has a ReLU activation. DNN is tested in 4 versions, the mono-response case fitted independently on each species of E_{50} like MAXENT and LGL, and the multi-response fitted on E_{50}, E_{200} or E_{1000}.
- **CNN** is composed of two hidden convolutional layers and one last layer fully connected with 200 neurons, exactly similar to previous ones. The first layer is composed of 64 convolution filters of kernel size (3, 3) and 1 line of 0 padding. The resulting feature maps are batch-normalized (same normalization for every pixels of a feature map) and transformed with a Relu. Then, an average pooling with a (8, 8) kernel and (8, 8) stride is applied. The second layer is composed of 128 convolution filters of kernel size (5, 5) and 2 lines of padding, plus Batch-Normalization and ReLU. After, that a second average pooling with a (8, 8) kernel and (8, 8) kernel and (8, 8) stride reduces size of the 128 feature maps to one pixel. Those are collected in a vector by a flattening operation preceding the fully connected layer. This architecture is not very deep. However, considered the restricted number of samples, a deep CNN would be very prone to over fitting. CNN is tested in multi-responses versions on E_{50}, E_{200} and E_{1000}.

10.3.4.3 Models Optimization

Our experiments were conducted using the `R` framework (version 3.3.2), on a Windows 10 machine with 2 CPUs with 2.60 GHz and 4 cores each, and one GPU NVIDIA Quadro M1000M. `mxnet` [29] is a convenient C++ library for learning deep NN models and is deployed as an R package. It integrates a high level symbolic language for quickly building customized models and loss functions, and automatically distributes calculations under CPUs or GPUs.

We fit the MAXENT model for every species of E_{50} with the recently released R package `maxnet` [17] and the vector input variables.

The LGL model was fitted with the package `mxnet`. The loss being convex, we used a simple **gradient descent algorithm** and stopped when the gradient norm was close to 0. The learning took around 2 min.

SNN, DNN and CNN models are fitted with the package `mxnet`: All model parameters were initialized with a uniform distribution $U(-0.03, 0.03)$, then we applied a **stochastic gradient descent algorithm with a momentum** of 0.9, a batch-size of 50 (batch samples are randomly chosen at each iteration), and an initial learning rate of 10^{-8}. The choice of initial learning rate was critical for a good optimization behavior. A too big learning rate can lead to training loss divergence,

whereas when it is too small, learning can be very slow. We stopped when the average slope of the training mean loss had an absolute difference to 0 on the last 100 epochs inferior to 10^{-3}. The learning took approximately 5 min for SNN, 10 min for DNN, and 5 h for CNN (independently of the version).

10.3.5 Evaluation Metrics

Predictions are made for every species of E_{50} and several model performance metrics are calculated for each species and for two disjoints and randomly sampled subsets of sites: A train set (4781 sites) which is used for fitting all models and a test set (400 sites) which aims at testing models generalization capacities. Then, train and test metrics are averaged over the 50 species. The performance metrics are described in the following.

10.3.5.1 Mean Loss

Mean loss, just named loss in the following, is an important metric to consider because it is relevant regarding our ecological model and it is the objective function that is minimized during model training. The Mean loss of model m on species i and on sites $1, \ldots, K$ is:

$$\text{Loss}(m, i, \{1, \ldots, K\}) = \frac{1}{K} \sum_{k=1}^{K} \lambda_{m,\theta_i}(x_k) - y_k^i \log(\lambda_{m,\theta_i}(x_k))$$

In Table 10.3, the loss is averaged over species of E_{50}. Thus, in the case of a mono-response model, we averaged the metric over the 50 independently learned models. In the multi-response case, we averaged the metric over each species response of the same model.

10.3.5.2 Root Mean Square Error (Rmse)

The root mean square error is a general error measure, which, in contrary to the previous one, is independent of the statistical model:

$$\text{Rmse}(m, i, \{1, \ldots, L\}) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(y_k^i - \lambda_{m,\theta_i}(x_k)\right)^2}$$

In Table 10.3, the average of the **Rmse** is computed over species of E_{50}. Mono-response models are treated as explained previously.

10.3.5.3 Accuracy on 10% Densest Quadrats (A10%DQ)

It represents the proportion of sites which are in the top 10% of all sites in term of both real count and model prediction. This is a meaningful metric for many concrete scenarios where the regions of a territory have to be prioritized in terms of decision or actions related to the ecology of species. However, we have to define the last site ranked in the top 10% for real counts, which is problematic for some species, because of ex-aequo sites. That is why we defined the following procedure which adjust for each species the percentage of top cells, such that the metrics can be calculated and the percentage is the closest to 10%. Denoting y the vector of real counts over sites and $\hat{y}$ the model prediction:

$$A10\%DQ(\hat{y}, y) := \frac{N_{p\&c}(\hat{y}, y)}{N_c(y)} \tag{10.7}$$

Where $N_{p\&c}(\hat{y}, y)$ is the number of sites that are contained in the $N_c(y)$ highest values of both y and $\hat{y}$.

Calculation of $N_c(y)$: We order the sites by decreasing values of y and note C_k the value of the kth site in this order. Noting $d := \text{round}(\dim(y)/10) = \text{round}(\dim(\hat{y})/10)$, as we are interested in the sites ranked in the 10% highest, if $C_d > C_{d+1}$ we simply set $N_c(y) = d$. Otherwise, if $C_d = C_{d+1}$ (ex-aequo exist for d^{th} position), we note **Sup** the position of the last site with value C_{d+1} and **Inf** the position of the first site with count C_d. The chosen rule is to take $N_c(y)$ such that $N_c(y) = \min(|\mathbf{Sup} - d|, |\mathbf{Inf} - d|)$.

10.4 Results

In the first part we describe and comment the main results obtained from performance metrics. Then, we illustrate and discuss qualitatively the behavior of models from the comparison of their predictions maps to real counts on some species.

10.4.1 Quantitative Results Analysis

Table 10.3 provides the results obtained for all the evaluated models according to the three evaluation metrics. The four main conclusions that we can derive from that results are that (1) performances of LGL and mono-response DNN are lower than the one of MAXENT for all metrics, (2) multi-response DNN outperforms SNN in every version and for all metrics, (3) multi-response DNN outperforms MAXENT in test Rmse in every version, (4) CNN outperforms all the other models, in every versions (CNN50, 200, 1000), and for all metrics.

According to these results, MAXENT shows the best performance amongst mono-response models. The low performance of the baseline LGL model is mostly due to underfitting. Actually, the evaluation metrics are not better on the training set than the test set. Its simple linear architecture is not able to exploit the complex relationships between environmental variables and observed abundance. DNN shows poor results as well in the mono-response version, but for another reason. We can see that its average training loss is very close to the minimum, which shows that the model is overfitting, i.e. it adjusts too much its parameters to predict exactly the training data, loosing its generalization capacity on test data.

However, for multi-responses versions, DNN performance increases importantly. DNN50 shows better results than MAXENT for the test Loss and test Rmse, while DNN200 and DNN1000 only show better Rmse. To go deeper, we notice that average and standard deviation of test rmse across E_{50} species goes down from DNN1 to DNN1000, showing that model becomes less sensitive to species data. Still, test loss and A10%DQ decrease, so there seems to be a performance trade-off between the different metrics as a side effect of the number of responses.

Whatever is the number of responses for SNN, the model is under-fitting and its performance are stable, without any big change between SNN50, 200, and 1 K. This model doesn't get improvement from the use of training data on a larger number of species. Furthermore, its performance is always lower than DNN's, which shows that stacking hidden layers improves the model capacity to extract relevant features from the environmental data, keeping all others factors constant.

The superiority of the CNN whatever the metric is a new and important result for species distribution modeling community. Something also important to notice, as for DNN, is the improvement of its performance for te.Loss and te.Rmse when the number of species in output increases. Those results suggest that the multi-response regularization is efficient when the model is complex (DNN) or the input dimensionality is important (CNN) but has no interest for simple models and small dimension input (SNN). There should be an optimal compromise to find between model complexity, in term of number of hidden layers and neurons, and the number of species set as responses.

For the best model CNN1000, it is interesting to see if the performance obtained on E_{50} could be generalized at a larger taxonomic scale. Therefore, we computed the results of the CNN1000 on the 1000 plant species used in output. Metrics values are :

- Test Loss = -1.275463 (minimum = -1.95)
- Test Rmse = 2.579596
- Test A10%DQ = 0.58

These additional results show that the average performance of CNN1000 on E_{1000} remains close from the one on E_{50}. Furthermore, one can notice the stability of performance across species. Actually, the test Rmse is lower than 3 for 710 of the 1000 species. That means that the learned environmental features are able to explain the distribution of a wide variety of species. According to the fact that French flora is compound of more than 6000 plant species, the potential of improvement of CNN

Table 10.3 Train and test performance metrics averaged over all species of E_{50} for all tested models

# species in output	Archi.	Loss on E_{50}		Rmse on E_{50}		A10%DQ on E_{50}	
		tr.(min:-1.90)	te.(min:-1.56)	$tr.$	te.	$tr.$	te.
1	MAX	−1.43	−0.862	2.24	3.18	0.641	0.548
	LGL	−1.11	−0.737	3.28	3.98	0.498	0.473
	DNN	−1.62	−0.677	3.00	3.52	0.741	0.504
50	SNN	−1.14	−0.710	3.14	3.05	0.494	0.460
	DNN	−1.45	−0.927	2.94	2.61	0.576	0.519
	CNN	−1.82	−0.991	1.18	2.38	0.846	0.607
200	SNN	−1.09	−0.690	3.25	3.03	0.479	0.447
	DNN	−1.32	−0.790	5.16	2.51	0.558	0.448
	CNN	−1.59	−1.070	2.04	2.34	0.650	0.594
1K	SNN	−1.13	−0.724	3.27	3.03	0.480	0.455
	DNN	−1.38	−0.804	3.86	2.50	0.534	0.467
	CNN	−1.70	−1.09	1.51	2.20	0.736	0.604

For the single response class, the metric is averaged over the models learnt on each species

predictions based on the use of this volume of species could be really important and one of the first at the country level (which is costly in terms of time with classical approaches).

We can go a bit deeper in the understanding of model performances in terms of species types. Figure 10.3 provides for CNN1000 and MAXENT the test Rmse as a function of the species percentage of presence sites. It first illustrates the fact that all SDMs are negatively affected by an higher percentage of presence sites, even the best, which is a known issue amongst species distribution modelers. Actually, the two models have quite similar results for species with high percentage of presence sites. Moreover, CNN1000 is better for most species compared to Maxent, and especially for species with low percentage of presence sites. For those species, we also notice that CNN's variance of Rmse is much smaller than MAXENT: there is no hard failing for CNN.

10.4.2 Qualitative Results Analysis

As metrics are only summaries, visualization of predictions on maps can be useful to make a clearer idea of the magnitude and nature of models errors. We took a particular species with a spatially restricted distribution in France, *Festuca cinerea*, in order to illustrate some models behavior that we have found to be consistent across this kind of species in E_{50}. The maps of real counts and several models predictions for this species are shown on Fig. 10.4. As we can note on map A of, *Festuca cinerea* was only observed in the south east part of the French territory. When we compare the different models prediction, CNN1000 (B) is the closest to

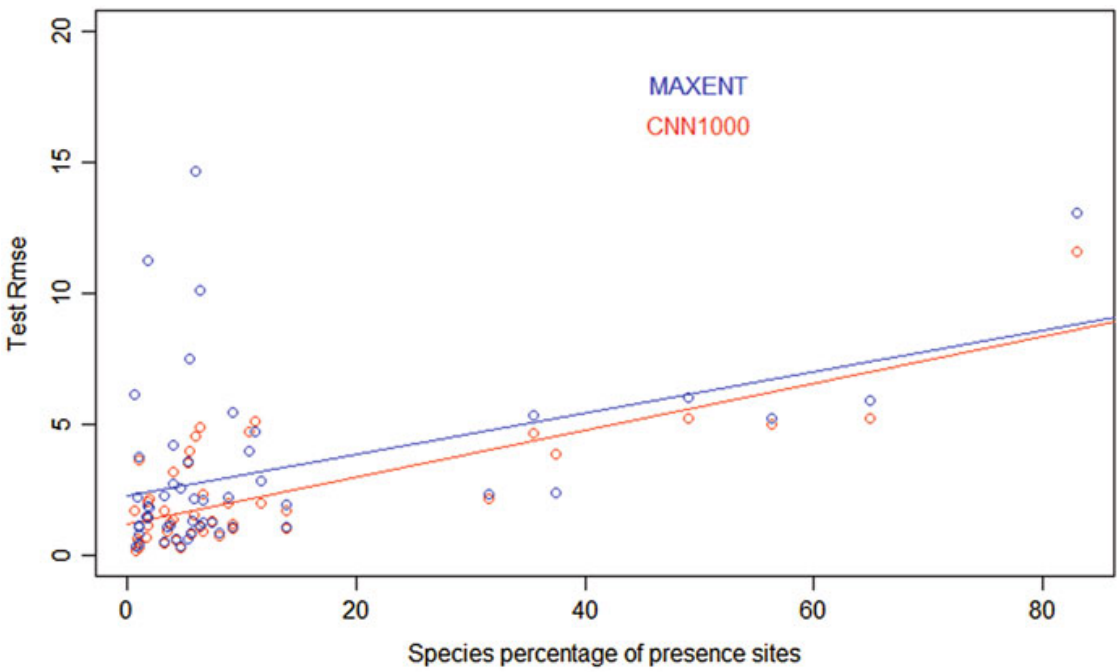


Fig. 10.3 Test Rmse plotted versus percentage of presence sites for every species of E_{50}, with linear regression curve, in blue with Maxent model, in red with CNN1000

real counts though DNN50 (C) and MAXENT (E) are not far. Clearly, DNN1000 (E) and LGL (F) are the models that over estimate the most the species presence over the territory. Another thing relative to DNN behavior can be noticed regarding Fig. 10.4. DNN1000 has less peaky punctual predictions than DNN50, it looks weathered. This behavior is consistent across species and could explain that the A10%DQ metric is weak for DNN1000 (and DNN200) compared to DNN50: A contraction of predicted abundance values toward the mean will imply less risk on prediction errors but predictions on high abundance sites will be less distinguished from others.

Good results provided in Table 10.3 can hide bad behavior of the models for certain species. Indeed, when we analyze, on Fig. 10.5, the distribution predicted by Maxent and CNN1000 for widespread species, such as *Anthriscus sylvestris* (L.) and *Ranunculus repens* L., we can notice a strong divergence with the INPN data. These two species, with the most important number of observation and percentage of presence sites in our experiment (see Table 10.1), are also the less well predicted by all models. For both species, MAXENT shows very smooth variations of predictions in space, which is sharply different from their real distribution. If CNN1000 seems to better fit to the presence area, it has still a lot of errors.

As last interesting remark, we note that a global maps analysis, on more species than the ones illustrated here, shows a consistent stronger false positive ratio for models under-fitting the data or with too much regularization (high number of responses in output).

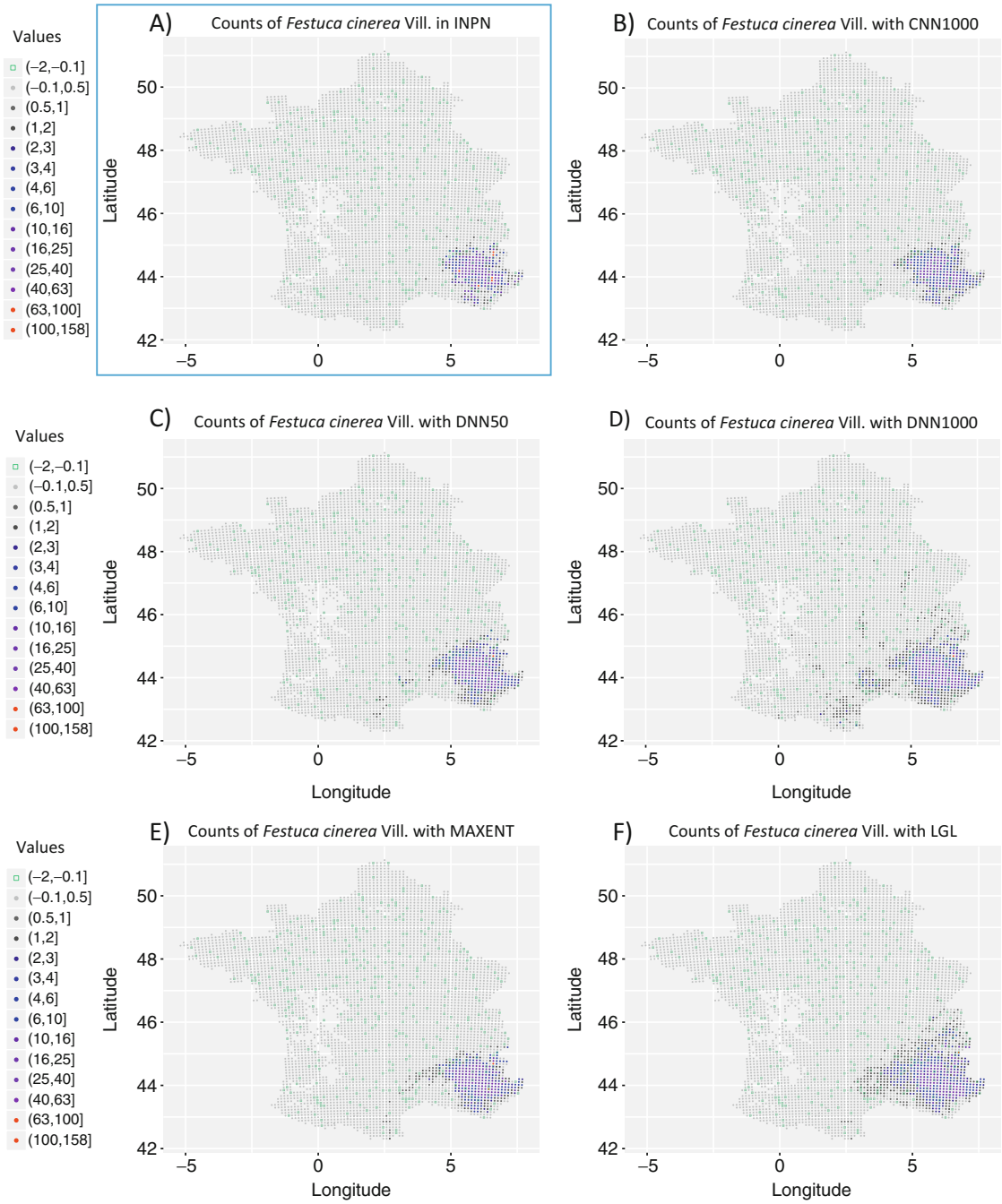


Fig. 10.4 Real count of *Festuca cinerea* Vill. and prediction for five different models. Test sites are framed into green squares. (**a**) Number of observations in INPN dataset, and geographic distribution predicted with (**b**) CNN1000, (**c**) DNN50, (**d**) DNN1000, (**e**) Maxent, (**f**) LGL

10.5 Discussion

The performance increase with multi-responses models shows that multi-responses architecture are an efficient regularization scheme for NNs in SDM. It could be interesting to evaluate the performance impact of going multi-response on rare species where data rare limited. We have systematically noticed false predicted

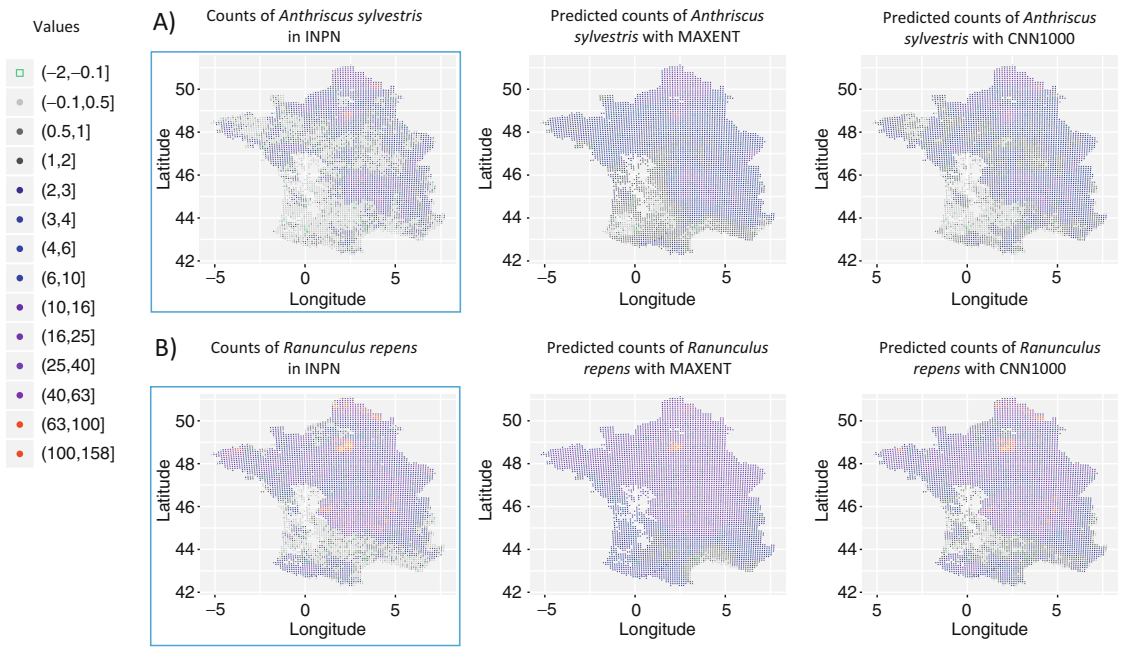


Fig. 10.5 (**a**) Species occurrences in INPN dataset, and geographic distribution predicted with Maxent and CNN1000 for *Anthriscus sylvestris* (L.) Hoffm., (**b**) Species occurrences in INPN dataset, and geographic distribution predicted with Maxent and CNN1000 for *Ranunculus repens* L

presence for species that are not in the Mediterranean region. It could be due to a high representativity of species from this region in France. In the multi-response modeling, the Mediterranean species could favor prediction in this area through neurons activations rather than other areas where few species are present, inducing bias. Thus, the distributions complementarity between selected species could be an interesting subject for further research.

Even if our study presents promising results, there are still some open problems. A first one is related to the bias in the sampling process that is not taken into account in the model. Indeed, even if the estimation of bias in the learning process is difficult, this could strongly improve our results. Bias can be related to the facts that (1) some regions and difficult environments are clearly less inventoried than others (this can be seen with "empty region" in South western part of the country in Figs. 10.4 and 10.5); (2) some regions are much more inventoried than others, according to the human capacities of the National botanical conservatories, which have very different sizes ; (3) some common and less attractive species for naturalists are not recorded, even if they are present in prospected areas, which is a bias due to the use of opportunistic observations rather than exhaustive count data.

In the NN models learning, there is still work to be done on quick automated procedure for tuning optimization hyper-parameters, especially the initial learning rate, and we are looking for a more suited stopping rule. On the other hand, in the case of models of species distributions, we can imagine to minimize the number of not null connections in the network, to make it more interpretable, and introduce an L1-type penalty on the network parameters. This is a potential important perspective of future works.

One imperfection in our modeling approach that induces biased distribution estimate is that the representation (vector or array of environmental variables) of a site is extracted from its geographic center. MAXENT, SNN and DNN models typically only integrate the central value of the environmental variables on each site, omitting the variability within the site. Instead of that, an unbiased data generation would sample for each site many representations uniformly in its spatial domain and in number proportional to its area. This way, it would provide richer information about sites and at the same time prevent NN model over-fitting by producing more data samples.

A deeper analysis of the behavior of the models according to the ecological preferences of the species could be of a strong interest for the ecological community. This study could allow to see dependences of the models to particular spatial patterns and/or environmental variables. Plus, it would be interesting to check if NN perform better when the species environmental niche is in the intersection of variables values that are far from their typical ranges into the study domain, which is something that MAXENT cannot fit.

Another interesting perspective for this work is the fact that, new detailed fine-scale environmental data become freely available with the development of the open data movement, in particular thanks to advances in remote sensing methods. Nevertheless, as long as we only have access to spatially degraded observations data at kilometer scales like here, it is difficult to consistently estimate the effect of variables that vary at high frequency in space. For example, the informative link between species abundance and land cover, proximity to fresh water or proximity to roads, is very blurred and almost lost. To overcome this difficulty, there is much hope in the high flow of finely geolocated species observations produced by citizen sciences programs for plant biodiversity monitoring like **Tela Botanica**,[5] **iNaturalist**,[6] **Naturgucker**[7] or **Pl@ntNet**.[8] From what we can see on the **GBIF**,[9] the first three already have high resolution and large cover observation capacity: they have accumulated around three hundred thousand finely geolocated plant species observations just in France during last decade. Citizen programs in biodiversity sciences are currently developing worldwide. We expect them to reach similar volumes of observations to the sum of national museums, herbaria and conservatories in the next few years, while still maintaining a large flow of observations for the future. With good methods for dealing with sampling bias, those fine precision and large spatial scale data will make a perfect context for reaching the full potential of deep learning SDM methods. Thus, NN methods could be a significant tool to explore biodiversity data and extract new ecological knowledge in the future.

[5] http://www.tela-botanica.org/site:accueil.

[6] https://www.inaturalist.org/.

[7] http://naturgucker.de/enjoynature.net.

[8] https://plantnet.org/en/.

[9] https://www.gbif.org/.

10.6 Conclusion

This study is the first one evaluating the potential of the deep learning approach for species distributions modeling. It shows that DNN and CNN models trained on 50 plant species of French flora clearly overcomes classical approaches, such as Maxent and LGL, used in ecological studies. This result is promising for future ecological studies developed in collaboration with naturalists expert. Actually, many ecological studies are based on models that do not take into account spatial patterns in environmental variables. In this paper, we show for a random set of 50 plant species of the French flora, that CNN and DNN, when learned as multi-species output models, are able to automatically learn non-linear transformations of input environmental features that are very relevant for every species without having to think a priori about variables correlation or selection. Plus, CNN can capture extra information contained in spatial patterns of environmental variables in order to surpass other classical approaches and even DNN. We also did show that the models trained on higher number of species in output (from 50 to 1000) stabilize predictions across species or even improve them globally, according to the results that we got for several metrics used to evaluate them. This is probably one the most important outcome of our study. It opens new opportunities for the development of ecological studies based on the use of CNN and DNN (e.g. the study of communities). However, deeper investigations regarding specific conditions for models efficiency, or the limits of interpretability NN predictions should be conducted to build richer ecological models.

References

1. Hutchinson, G. (1957). Concluding remarks. Cold spring harbor symposium on quantitative biology. 22, 415–427.
2. Hastie, T. & Tibshirani, R. (1986). Generalized Additive Models. Statistical Science, 1(3), 297–318.
3. Friedman, J. (1991). Multivariate adaptive regression splines. The annals of statistics, 1–67.
4. Phillips, S., Dudik, M., Schapire, R. (2004). A maximum entropy approach to species distribution modeling. Proceedings of the twenty-first international conference on Machine learning, 83.
5. Phillips, S., Anderson, R. & Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. Ecological modelling, 190(3), 231–259.
6. Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep Learning. MIT Press.
7. Krizhevsky, A., Sutskever, I. & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097–1105.
8. Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. Ecological modelling, 90(1), 39–52.
9. Thuiller, W. (2003). BIOMOD–optimizing predictions of species distributions and projecting potential future shifts under global change. Global change biology, 9(10), 1353–1362.
10. Leathwick, J.R. Elith, J. & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological modelling, 199(2), 188–196.

11. LeCun, Y. & others. (1989). Generalization and network design strategies. Connectionism in perspective, 143–155.
12. Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. (2009). Presence-only data and the EM algorithm. Biometrics, 65(2), 554–563.
13. Berman, M., & Turner, T. R. (1992). Approximating point process likelihoods with GLIM. Applied Statistics, 31–38.
14. P Anderson, R., Dudk, M., Ferrier, S., Guisan, A., J Hijmans, R., Huettmann, F., ... & A Loiselle, B. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29(2), 129–151.
15. Phillips, S. & Dudik, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31(2), 161–175.
16. Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. The annals of applied statistics.7,4,1917.
17. Phillips, S. Anderson, R., Dudik, M. Schapire, R. & Blair, M. (2017). Opening the black box: an open-source release of Maxent. Ecography.
18. Rumelhart, D., Hinton, G. & Williams, R. and others (1988). Learning representations by back-propagating errors. Cognitive modeling, 5(3).
19. Nair, V. & Hinton, G. (2010). Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10), 807–814.
20. Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. 448–456.
21. Dutrève, B. & Robert, S. (2016). INPN - Données flore des CBN agrégées par la FCBN. Version 1.1. SPN - Service du Patrimoine naturel, Muséum national d'Histoire naturelle, Paris. Occurrence Dataset https://doi.org/10.15468/omae84 accessed via GBIF.org on 2017-08-30.
22. Karger, D. N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W. & Kessler, M. (2016). Climatologies at high resolution for the earth's land surface areas. arXiv preprint arXiv:1607.00217.
23. Karger, D. N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W. & Kessler, M. (2016). CHELSEA climatologies at high resolution for the earth's land surface areas (Version 1.1).
24. Zomer, R., Bossio, D., Trabucco, A., Yuanjie, L., Gupta, D. & Singh, V. (2007). Trees and water: smallholder agroforestry on irrigated lands in Northern India.
25. Zomer, R., Trabucco, A., Bossio, D. & Verchot, L. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, ecosystems & environment, 126(1), 67–80.
26. Panagos, P. (2006). The European soil database. GEO: connexion, 5(7), 32–33.
27. Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L. (2012). European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy, 29(2),329–338.
28. Van Liedekerke, M. Jones, A. & Panagos, P. (2006). ESDBv2 Raster Library-a set of rasters derived from the European Soil Database distribution v2. 0. European Commission and the European Soil Bureau Network, CDROM, EUR, 19945.
29. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., ... & Zhang, Z. (2015). Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274.

9 Chapter 5: A cooperative evaluation campaign for species recommendation algorithms, GeoLifeCLEF

Overview of GeoLifeCLEF 2018: location-based species recommendation

Christophe Botella[1,2], Pierre Bonnet[3], Franois Munoz[4], Pascal Monestiez[5], and Alexis Joly[1]

[1] Inria, LIRMM, Montpellier, France
[2] INRA, UMR AMAP, France
[3] CIRAD, UMR AMAP, Montpellier, France
[4] Universit Grenoble-Alpes, Grenoble, France
[5] BioSP, INRA, Avignon, France

Abstract. The GeoLifeCLEF challenge provides a testbed for the system-oriented evaluation of a geographic species recommendation service. The aim is to investigate location-based recommendation approaches in the context of large scale spatialized environmental data. This paper presents an overview of the resources and assessments of the GeoLifeCLEF task 2018, summarizes the approaches employed by the participating groups, and provides an analysis of the main evaluation results.

Keywords: LifeCLEF, biodiversity, big data, environmental data, visual data, species recommendation, evaluation, benchmark

CEUR-WS.org/Vol-2125/invited_paper_8.pdf

1 Introduction

Automatically predicting the list of species that are the most likely to be observed at a given location is useful for many scenarios in biodiversity informatics. First of all, it could improve species identification processes and tools by reducing the list of candidate species that are observable at a given location (be they automated, semi-automated or based on classical field guides or flora). More generally, it could facilitate biodiversity inventories through the development of location-based recommendation services (typically on mobile phones) as well as the involvement of non-expert nature observers. Last but not least, it might serve educational purposes thanks to biodiversity discovery applications providing functionalities such as contextualized educational pathways.

The aim of the challenge is to predict the list of species that are the most likely to be observed at a given location. Therefore, we provided a large training set of species occurrences, each occurrence being associated to a multi-channel image characterizing the local environment. Indeed, it is usually not possible to learn a species distribution model directly from spatial positions because of the limited number of occurrences and the sampling bias. What is usually done in ecology is to predict the distribution on the basis of a representation in the environmental

space, typically a feature vector composed of climatic variables (average temperature at that location, precipitation, etc.) and other variables such as soil type, land cover, distance to water, etc. The originality of GeoLifeCLEF is to generalize such niche modeling approach to the use of an image-based environmental representation space. Instead of learning a model from environmental feature vectors, participants may learn a model from k-dimensional image patches, each patch representing the value of an environmental variable in the neighborhood of the occurrence (see Figure 1 below for an illustration). From a machine learning point of view, the challenge will thus be treatable as a multi-channel image classification task.

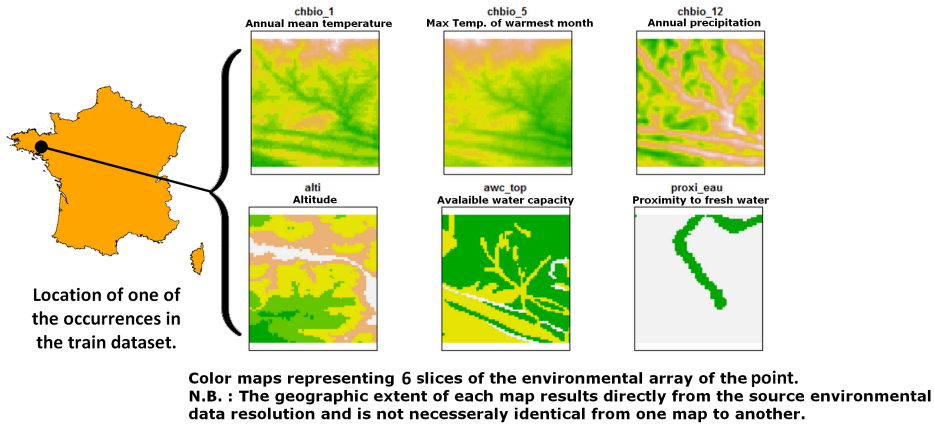


Fig. 1. Example of 6 channels from the environmental tensor of an occurrence. Each channel is an environmental heatmap, i.e. a matrix representing the values of an environmental variable in a square spatial area centered at the occurrence location.

2 Dataset

The participants were provided with a train and test set of species geolocated occurrences. Both were first composed of a **.csv** file with the occurrences spatial coordinates, the punctual values of environmental variables at the occurrence location, and, for the train table, the species name and identifier. Secondly, each row of the table (train and test) referred to a 33-channel image containing the environmental tensor extracted at that location.

2.1 Species occurrences

Occurrences data were extracted from the Global Biodiversity Information Facility platform (GBIF [6]). To achieve precise species prediction from a geolocation, the geolocations in question must be as precise as possible. However, a high number of occurrences from the GBIF have a spatially degraded geolocation for conservation reasons. Thus, we have chosen source datasets with undegraded geolocations in France, which are :

1. Carnet en ligne from Tela Botanica.
2. Cartographie des Leguminosae (Fabaceae) en France from Tela Botanica.
3. Naturgucker dataset.
4. iNaturalist Research-grade Observations.

Only observations falling in the metropolitan French territory were kept so as to focus on a region for which we had an easy access to rich and homogeneous environmental descriptors for the whole dataset. Occurrences with uncertain names, as notified by the GBIF, were removed. The full dataset is finally composed of 291,392 occurrences. The labels to be predicted within the challenge are the species identifier (field **species_glc_id**). There are 3,336 species identifiers in total, and their associated taxonomic names are provided by the field **espece_retenue_bdtfx** (bdtfx referential 4.1). Due to some unreferenced heterogeneity in the data collection protocol (naturalists checklists, conversion of site name to geolocation, etc), some geographical points accumulate several occurrences. Indeed, there are in total 75,668 distinct geolocations (with a maximum of 527 points in one geolocation). All occurrences geolocations are represented in Figure 2. It reveals the bias in the spatial distribution of the occurrences.

2.2 Environmental data

Each occurrence is characterized by 33 local environmental images of 64x64 pixels. These environmental images were constructed from various open datasets and include 19 bioclimatic quantitative variables at 1km resolution from Chelsea Climate [6], 10 pedological ordinal variables at 1km resolution from ESDB soil pedology data [11,12,16], one land cover categorical descriptor at 100 meters resolution from Corine Land Cover 2012 soil occupation data (version 18.5.1, 12/2016), one potential evapo-transpiration quantitative variable at 1km resolution from CGIAR-CSI evapotranspiration data ([18,19]), one elevation quantitative variable at 90 meters resolution from USGS Elevation data (Data available from the U.S. Geological Survey and downloadable on the Earthexplorer[7]) and one indicator of fresh water proximity at 12,5m resolution from the BD Carthage hydrologic data. As each of those variables are stored in large raster covering the French geographical territory. For any occurrence, we crop a 64×64 pixels window centered on the occurrence geolocation from the raster of each environmental variable. This way, we make the $64 \times 64 \times 33$ environmental tensor

[6] https://www.gbif.org/
[7] (https://earthexplorer.usgs.gov/)

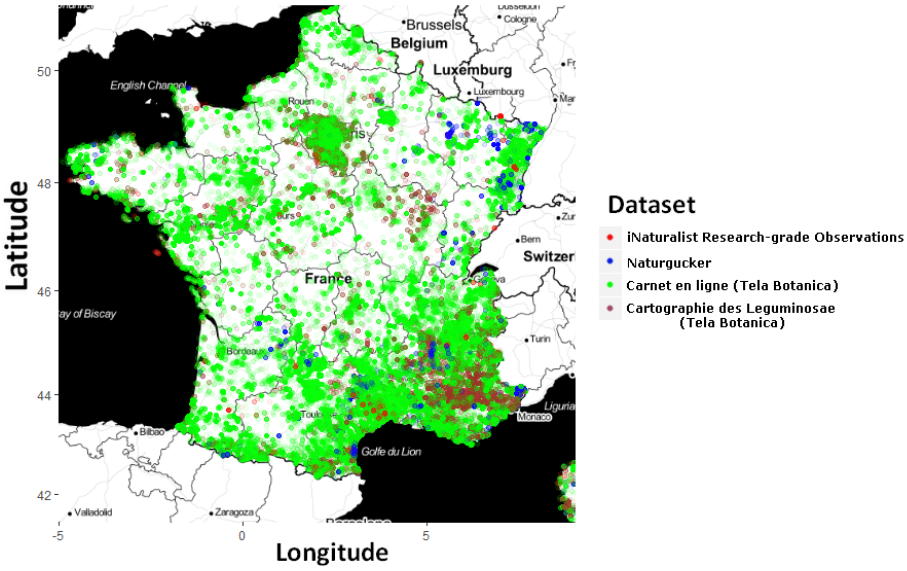


Fig. 2. Occurrences geolocations in GeoLifeCLEF 2018 and their source dataset over the metropolitan territory.

associated with this occurrence. Besides, the punctual environmental values associated with an occurrence, are simply the extracted cell's values from the rasters at the occurrence geolocation.

2.3 Train and test sets

The total of 291,392 occurrences were randomly split into a training set (218,543) and a test set (72,849) with the constraints that :

- For each species in the test set, there is at least one observation of it in the training set.
- An observation of a species in the test set is distant of more than 100 meters from all observations of this species in the train set to avoid major reporting dependencies.

Thus, the final train set contained all of the 3,336 species, while the test set contained 3,209 species.

3 Task Description

For every occurrence of the test set, participants must supply a list of 100 species maximum, ranked without ex-aequo. The used evaluation metric is the Mean

Reciprocal Rank (MRR). The MRR is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer. The MRR is the average of the reciprocal ranks for the whole test set:

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

where Q is the total number of query occurrences x_q in the test set and $rank_q$ is the rank of the correct species $y(x_q)$ in the ranked list of species predicted by the evaluated method for the occurrence x_q.

4 Participants and methods

22 research groups registered to the GeoLifeCLEF challenge 2018. Among this large raw audience, 3 research groups finally succeeded in submitting run files. Details of the used methods and evaluated systems are synthesized below and further developed in the working notes of the participants ([2], [15] and [8]). Table 1 reports the results achieved by each run as well as a brief synthesis on the methods used in each of them. Complementary, the following paragraphs give a few more details about the methods and the overall strategy employed by each participant.

FLO team, France, 10 runs, [2]: FLO developed four prediction models, (i) one convolutional neural network trained on environmental tensors (FLO_3). The CNN implemented a customized architecture. It also treated the categorical land cover descriptor independantly from quantitative variables for the primary layers. Activation's of both variables types where then fused in deeper layers. (ii) one neural network (FLO_2) trained on species occurrences falling at the closest spatial point and two other models only based on the spatial occurrences of species: (iii) a closest-location classifier (FLO_1) and (iv) a random forest fitted on the spatial coordinates (FLO_4). Other runs correspond to late fusions of that base models, either by simply averaging either the output probabilities (FLO_5,FLO_6,FLO_7,FLO_8), or ranks with the Borda method (FLO_9,FLO_10).

ST team, Germany, 16 runs, [14]: ST experimented two main types of models, convolutional neural networks on environmental tensors with different data augmentations like rotation and flip of images (ST_1, ST_3, ST_11, ST_14, ST_15, ST_18, ST_19) and Boosted Trees (XGBoost) on vectors of environmental variables concatenated with spatial positions (ST_6, ST_9, ST_10, ST_12, ST_13, ST_16, ST_17). They also proposed a nearest-neighbor classifier based on the environmental variables of occurrences (ST_5), and two species cluster models

(ST_17,ST_8) where groups of species are constituted by the similarity of the environmental variables where they occur. For analysis purposes, ST_2 corresponds to a random predictor and ST_7 to a constant predictor returning always the 100 most frequent species (ranked by decreasing value of their frequency in the training set).

SSN, India, 4 runs, [10]: SSN attempted to learn a CNN-LSTM hybrid model, based on a ResNext architecture [17] extended with an LSTM layer [3] aimed at predicting the plant categories at 5 different levels of the taxonomy (class, then order, then family, then genus and finally species). The four runs are derived from this model.

5 Results

We report in Figure 3 and Table 1 the main results achieved by the 33 submitted runs as well as some synthetic information about the used methods and variables for each run. The main conclusions we can draw from that results are the following:

Convolutional Neural Networks outperformed boosted trees: Boosted trees are known to provide state-of-the-art performance for environmental modelling. They are actually used in a wide variety of ecological studies [4,1,7,9]. Our evaluation, however, demonstrate that they can be consistently outperformed by convolutional neural networks trained on environmental data tensors. The best submitted run that does not result from a fusion of different models (FLO_3), is actually a convolutional neural network trained on the environmental patches. It achieved a MRR of 0.043 whereas the best boosted tree (ST_16) achieved a MRR of 0.035. As another evidence of the better performance of the CNN model, the six best runs of the challenge result from the combination of it with the other models of the Floris'Tic team. Now, it is important to notice that the CNN models trained by the ST team (ST_1, ST_3, ST_11, ST_14, ST_15, ST_18, ST_19) and SSN team did not obtain good performance at all (often worse than the constant predictor based on the class prior distribution), which could be due to a mismatch of species identifiers, as noticed by the participant. For team ST, results can't be interpreted directly as a failure of the methods. The ranking of runs in the test set was not consistent with validation results and the learning process can be improved according to [15]. This illustrates the difficulty of designing and fitting deep neural networks on new problems without former references in the literature. Lastly, the approaches trying to adapt existing complex CNN architectures that are popular in the image domain (such as VGG [13], DenseNet [5], ResNEXT [17] and LSTM [3]) were not successfull. High difference of performances in CNN learned with homemade architectures ($FLO_6, FLO_3, FLO_8, FLO_5, FLO_9, FLO_10$ compared to ST_3, ST_1) could underline the importance of architecture choices.

Purely spatial models are not so bad: the random forest model of the FLO team, fitted on spatial coordinates solely (FLO_4), achieved a fair MRR of 0.0329, close to the performance of the boosted trees of the ST team (that were trained on environmental & spatial data). Purely spatial models are usually not used for species distribution modelling because of the heterogeneity of the observations density across different regions. Indeed, the spatial distribution of the observed specimens is often more correlated with the geographic preferences of the observers than with the abundance of the observed species. However the goal of GeoLifeClef is to predict the most likely species to observe given the real presence of a plant. Thus, the heterogeneity of the sampling effort should induce less bias than in ecological studies.

It is likely that the Convolutional Neural Network already captured the spatial information: The best run of the whole challenge (**FLO_6**) results from the combination of the best environmental model (CNN **FLO_3**) and the best spatial model (Random forest **FLO_4**). However, it is noticeable that the improvement of the fused run compared to the CNN alone is extremely tight (+ 0.0005), and actually not statistically significant. In other words, it seems that the information learned by the spatial model was already captured by the CNN. Besides, CNN uses the whole environmental tensor as input and is better than the XGBoost methods which used only the average of each environmental matrix as input. So it is likely that CNN captured more information than the average of the environmental image. It might be some patterns associated with a particular area, or more generic environmental patterns (a wet valley, etc.).

The learning of species communities patterns has potential: We first state that species have marked spatial patterns. Indeed, predicting the nearest species in space (FLO_1) or in the environmental space (ST_5) is much more efficient than simply listing species per global abundance (ST_7), which corresponds to a uniform prior on spatial distribution of each species. Second, methods that allow interactions between species abundance, either by building and predicting group of species that have similar environmental preferences (ST_17), or learning the association between species that co-occur in a close surrounding (FLO_2) perform better than simple nearest-neighbor approaches. However, these approaches are still limitating as, for example, FLO_2 only used the closest point as input information about surrounding species. Besides, even though the good performance of ST_17, there was very few groups of more than 1 species in their algorithm, which leaves small chances to predict non-common species while they represent the majority of species.

A significant margin of progress but still very promising results: even if the best MRR scores appear to be very low at a first glance, it is important to relativize them with regard to the nature of the task. Many species (tens to hundred) are actually living at the same location so that achieving very high MRR scores is not possible. The MRR score is useful to compare the methods between each others but it should not be interpreted as for a classical information retrieval task. In the test set itself, several species are often observed at exactly the same location. So that there is a max bound on the achievable MRR equal to

0.56. The best run (FLO_3) is still far from this max bound (MRR=0.043) but it is much better than the random or the prior distribution based MRR. Concretely, it retrieves the right species in the top-10 results in 25% of the cases, or in the top-100 in 49% of the cases (over $3,336$ species in the training set), which means that it is not so bad at predicting the set of species that might be observed at that location.

Table 1: Methods and results of runs submitted to GeoLifeCLEF2018.

rank	run-name	score	algorithm type	variables	top1	top10	top30
	Perfect	0.5593			0.4214	0.854	0.97
1	FLO_6	0.0435	fusion: CNN+Random Forest	Env.tensor, Geoloc	0.0116	0.100	0.25
2	FLO_3	0.0430	CNN	Env.tensor	0.0110	0.098	0.25
3	FLO_8	0.0423	fusion: CNN + Random Forest + NN	Env.tensor, Geoloc, Neighbors	0.0111	0.097	0.24
4	FLO_5	0.0422	fusion: CNN + NN	Env.tensor, Neighbors	0.0112	0.097	0.24
5	FLO_9	0.0388	CNN + Random forest + NN	Env.tensor, Geoloc, Neighbors	0.0099	0.088	0.22
6	FLO_10	0.0365	CNN + nearest-neighbors (spatial) + Random Forest	Env.tensor, Neighbors, Geoloc	0.0116	0.074	0.20
7	ST_16	0.0358	XGBoost (1 model/species)	Env.tensor, Geoloc	0.0111	0.077	0.18
8	ST_13	0.0352	XGBoost	Env.tensor, Geoloc	0.0114	0.075	0.18
9	ST_10	0.0348	XGBoost	Env.tensor, Geoloc	0.0113	0.073	0.18
10	ST_9	0.0344	XGBoost	Env.tensor, Geoloc	0.0108	0.073	0.18
11	ST_12	0.0343	XGBoost	Env.tensor, Geoloc	0.0110	0.072	0.18
12	ST_6	0.0338	XGBoost (1 model/species)	Env.tensor, Geoloc	0.0104	0.072	0.18
13	FLO_4	0.0329	Random forest	Geoloc	0.0098	0.071	0.17
14	FLO_7	0.0327	NN + Random forest	Neighbors, Geoloc	0.0093	0.071	0.18
15	ST_17	0.0326	Environmental clustering of species	Env.tensor, Geoloc	0.0106	0.068	0.16
16	FLO_2	0.0274	NN	Neighbors	0.0079	0.057	0.15
17	ST_5	0.0271	nearest-neighbors (environment, spatial)	Neighbors	0.0098	0.051	0.14
18	ST_8	0.0220	Environmental clustering of species	Env.tensor, Geoloc	0.0087	0.043	0.08
19	FLO_1	0.0199	nearest-neighbors (spatial)	Geoloc	0.0077	0.042	0.09
20	ST_3	0.0153	CNN	Env.tensor	0.0047	0.029	0.08
21	ST_1	0.0153	CNN	Env.tensor	0.0047	0.029	0.08
22	ST_14	0.0144	CNN	Env.tensor	0.0040	0.030	0.07
23	ST_7	0.0134	Global species frequency	occurrences counts	0.0026	0.033	0.08
24	ST_15	0.0103	CNN (vgg-like)	Env.tensor	0.0046	0.019	0.03
25	ST_19	0.0099	standard DenseNet121 CNN	Env.tensor	0.0021	0.024	0.05
26	ST_11	0.0096	CNN (vgg-like)	Env.tensor	0.0022	0.022	0.05
27	ST_18	0.0096	ensemble CNN (vgg-like)	Env.tensor	0.0019	0.023	0.05
28	ST_4	0.0085	XGBoost	Env.tensor, spatial	0.0021	0.022	0.04
29	SSN_3	0.0030	CNN(Resnext)-LSTM	Env.tensor, Taxonomy	0.0006	0.006	0.02
30	SSN_4	0.0016	CNN(Resnext)-LSTM	Env.tensor, Taxonomy	0.0000	0.005	0.01
31	ST_2	0.0016	Random list		0.0003	0.003	0.01
32	SSN_2	0.0013	CNN(Resnext)-LSTM	Env.tensor, Taxonomy	0.0001	0.002	0.01
33	SSN_1	0.0004	CNN(Resnext)-LSTM	Env.tensor, Taxonomy	0.0001	0.002	0.00

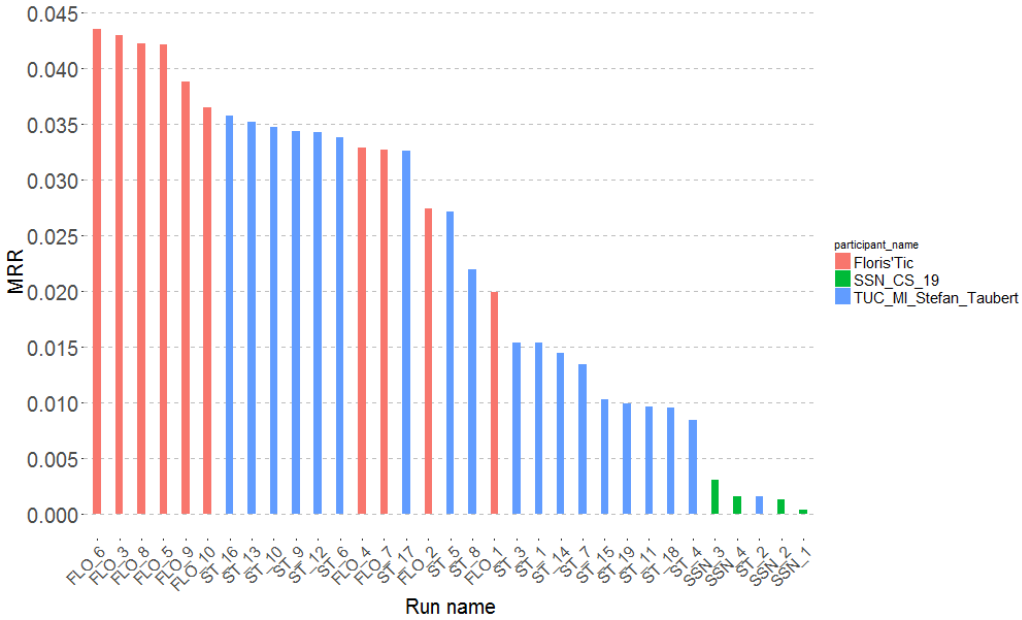


Fig. 3. MRR scores per submitted run and participant.

6 Complementary Analysis

Spatial heterogeneity of model performances: We computed the MRR restricted to occurrences that fall in spatial quadrats of 10×10 km all over the French territory. We projected this on a map in Figure 4. The global performances of the methods hide spatial heterogeneity, as shown in the map. Indeed, Paris is the best predicted area, then the Mediterranean region and the Alpes. Then other regions like the Loire, the Pyrenees and the Atlantic coastline. One could think this is due to the larger number of points available in these areas, but this is not exactly true. Complementary analysis showed that the importantly sampled areas had a more stable MRR but not higher in average. Thus, improving models predictions should pass by finding reasons of varying regional performances, in the hope to bring a solution.

Rare species are not unpredictable: For each species and method, we calculated the MRR over the occurrences of this species in the test set. We ordered species per decreasing global occurrences count in the test set in order to compare the performances of each method along the gradient from common to rare species. The raw graphs were difficult to analyse because the MRR varies

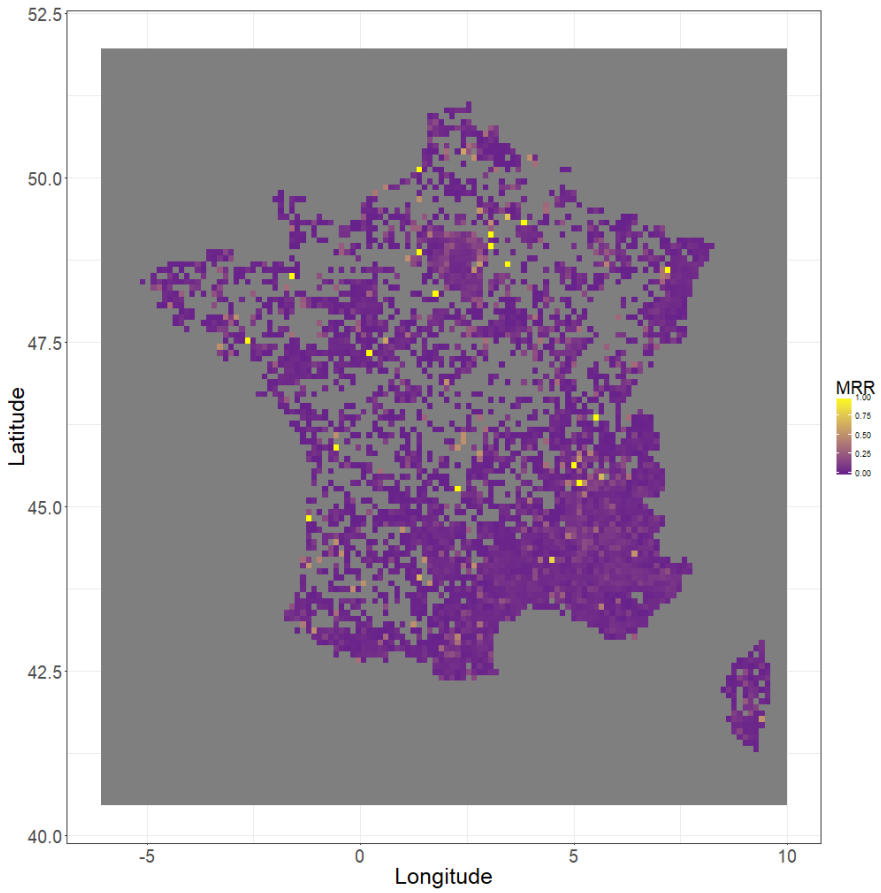


Fig. 4. MRR per 10×10km square spatial quadrat for FLO_3 over the study region.

importantly for rare species, as there are very few occurrences. Thus, we operate a smoothing along the scarcity gradient. For each species we took the median of the MRR over the 40 species of closest rank on this scarcity gradient. Figures 5 and 6 show the result for FLO_3 (environmental CNN), ST_16 (XGBoost), FLO_4 (spatial Random Forest) and ST_7 (Global frequency of species). One can see that ST_7 early cancels along the scarcity gradient. This is because more than 50% of the species over which the median is calculated have a null MRR, which correctly represents the tendancy we want to observe. First, it seems that non-common species have marked spatial preferences because FLO_4 is much better when getting scarcer than ST_7. Second, the progression of predictions of FLO_3 and ST_16 compared to FLO_4 for rare species (in the long tail) suggests that those species mainly have marked environmental preferences that is not easy to capture with a spatial model which doesn't have access to this information. The CNN is very good at predicting non-common species, which may be a bit surprising as (i) its predictions should be smooth in space according to the width of some environmental images (64x64km for climatic and pedological variables) and the chosen architecture and (ii) rare species often have a restricted niche.

7 Conclusion

We have analyzed the results of the 3 participants of GeoLifeCLEF 2018. CNN models learnt on environmental tensors revealed to be the most performing method, however challenging to operate. According to those results, they are more efficient than Boosted Trees a state of the art method in species distribution modeling. This might be because they may detect particular area or environmental patterns as they access to the full surrounding environment data, but that remain to be proved. Spatial and species association methods have shown reasonably good results, but there is room for improvement, especially for the use of interdependence. The complementary analysis revealed that all methods had the same areas of unreliability. Furthermore, the integration of environmental variables seems to be very beneficial to the prediction of non-common species. The task of finding the species found at a precise location is difficult because many species co-exist at very small spatial scales (under the meter). The accuracy of current geolocation devices doesn't even allow to indicate with this precision the point where the specimen was observed. Thus, in the future, the evaluation process shouldn't penalize predictions of other species that have been observed in such a close surrounding regarding the precision of the reported geolocation.

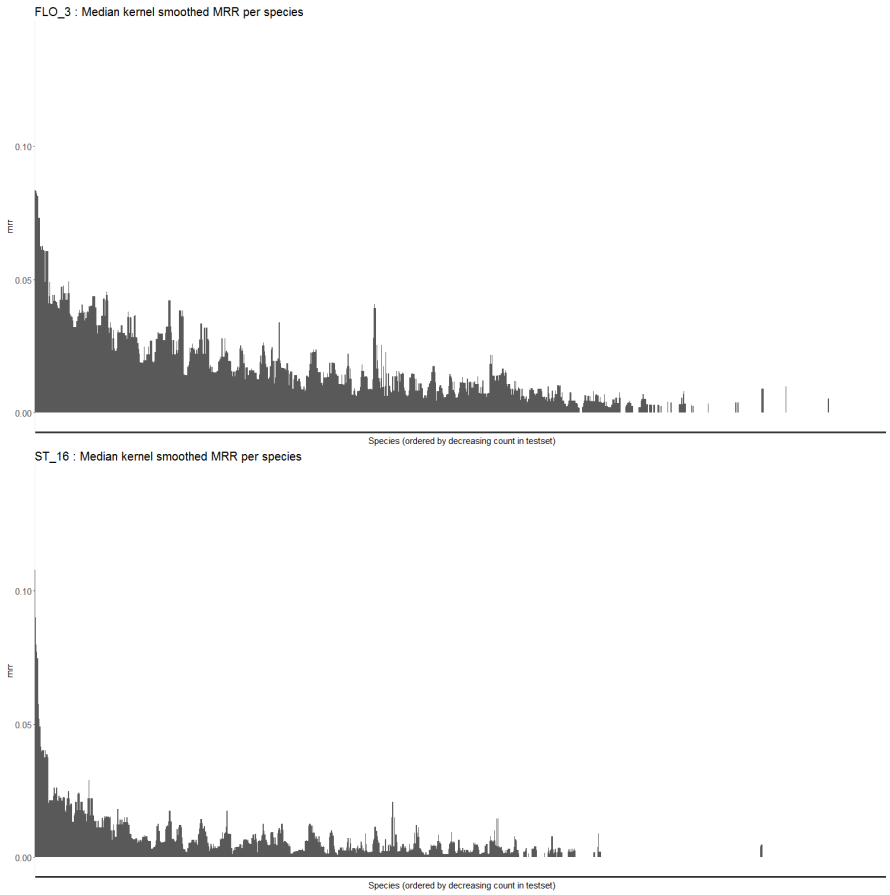


Fig. 5. Smoothed MRR per species for FLO_3 and ST_16. Species are ordered by number of occurrences in the test set. Each species MRR is smoothed by taking the median over the MRR the 40 species of closest rank along the scarcity gradient.

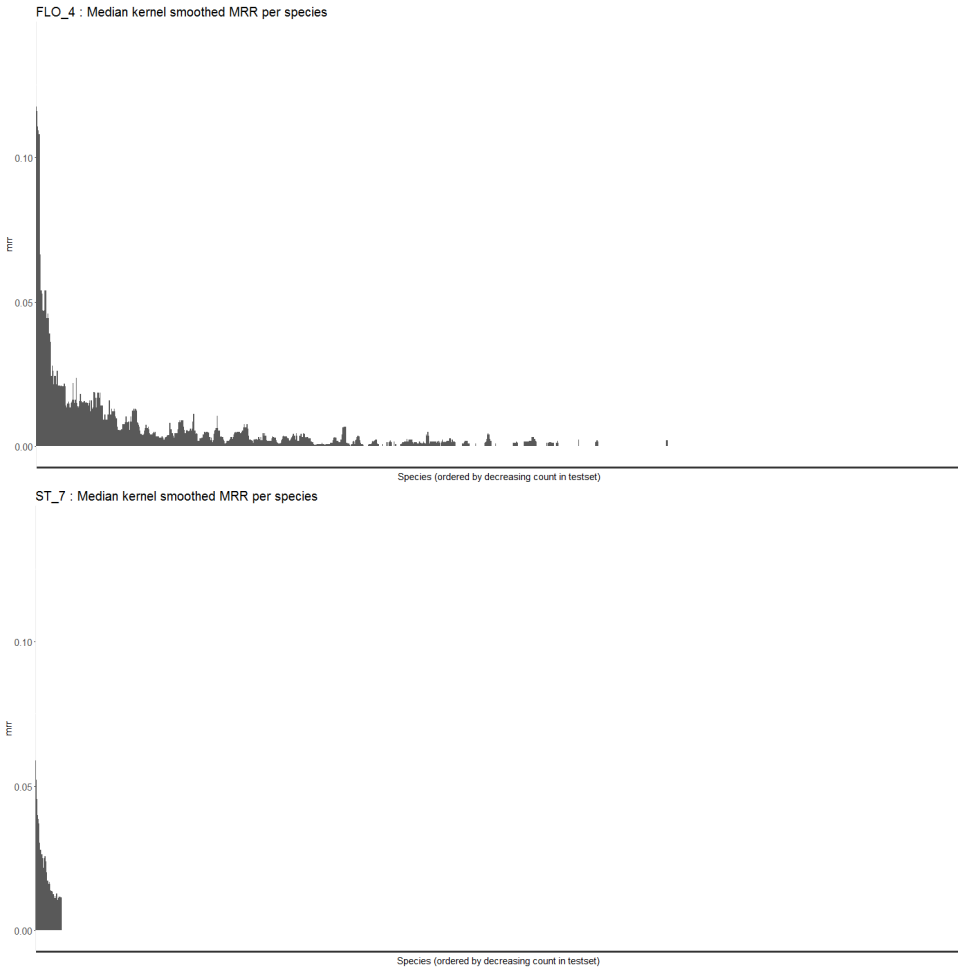


Fig. 6. Smoothed MRR per species for FLO_4 and ST_7. Species are ordered by number of occurrences in the test set. Each species MRR is smoothed by taking the median over the MRR the 40 species of closest rank along the scarcity gradient.

References

1. De'Ath, G.: Boosted trees for ecological modeling and prediction. Ecology 88(1), 243–251 (2007)
2. Deneu, B., Servajean, M., Botella, C., Joly, A.: Location-based species recommendation using co-occurrences and environment - geolifeclef 2018 challenge. In: CLEF working notes 2018 (2018)
3. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm (1999)
4. Guisan, A., Thuiller, W., Zimmermann, N.E.: Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press (2017)
5. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. vol. 1, p. 3 (2017)
6. Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N., Linder, H.P., Kessler, M.: Climatologies at high resolution for the earth's land surface areas. arXiv preprint arXiv:1607.00217 (2016)
7. Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., Weiss, D.J., Golding, N., Ruktanonchai, C.W., Gething, P.W., Cohn, E., et al.: Mapping global environmental suitability for zika virus. Elife 5 (2016)
8. Moudhgalya, N.B., Sundar, S., Divi, S., Mirunalini, P., Aravindan Bose, C.: Hierarchically embedded taxonomy with clnn to predict species based on spatial features. In: CLEF working notes 2018 (2018)
9. Moyes, C.L., Shearer, F.M., Huang, Z., Wiebe, A., Gibson, H.S., Nijman, V., Mohd-Azlan, J., Brodie, J.F., Malaivijitnond, S., Linkie, M., et al.: Predicting the geographical distributions of the macaque hosts and mosquito vectors of plasmodium knowlesi malaria in forested and non-forested areas. Parasites & vectors 9(1), 242 (2016)
10. Nithish B Moudhgalya, Sharan Sundar, S.D.M.P., Bose, C.A.: Hierarchically embedded taxonomy with clnn to predict species based on spatial features. In: CLEF working notes 2018 (2018)
11. Panagos, P.: The european soil database. GEO: connexion 5(7), 32–33 (2006)
12. Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L.: European soil data centre: Response to european policy support and public data requirements. Land Use Policy 29(2), 329–338 (2012)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
14. Stefan Taubert, Max Mauermann, S.K.D.K., Eibl, M.: Species prediction based on environmental variables using machine learning techniques. In: CLEF working notes 2018 (2018)
15. Taubert, S., Mauermann, M., Kahl, S., Kowerko, D., Eibl, M.: Species prediction based on environmental variables using machine learning techniques. In: CLEF working notes 2018 (2018)
16. Van Liedekerke, M., Jones, A., Panagos, P.: Esdbv2 raster library-a set of rasters derived from the european soil database distribution v2. 0. European Commission and the European Soil Bureau Network, CDROM, EUR 19945 (2006)
17. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 5987–5995. IEEE (2017)

18. Zomer, R.J., Bossio, D.A., Trabucco, A., Yuanjie, L., Gupta, D.C., Singh, V.P.: Trees and water: smallholder agroforestry on irrigated lands in Northern India, vol. 122. IWMI (2007)
19. Zomer, R.J., Trabucco, A., Bossio, D.A., Verchot, L.V.: Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, ecosystems & environment 126(1), 67–80 (2008)

Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences

Botella Christophe[1,2,3], Servajean Maximilien[5], Bonnet Pierre[3,4], Joly Alexis[1]

[1] INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.
[2] INRA, UMR AMAP, F-34398 Montpellier, France.
[3] AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France.
[4] CIRAD, UMR AMAP, F-34398 Montpellier, France.
[5] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, Montpellier, France

Abstract. The GeoLifeCLEF challenge aim to evaluate location-based species recommendation algorithms through open and perennial datasets in a reproducible way. It offers a ground for large-scale geographic species prediction using cross-kingdom occurrences and spatialized environmental data. The main novelty of the 2019 campaign over the previous one is the availability of new occurrence datasets: (i) automatically identified plant occurrences coming from the popular Pl@ntnet platform and (ii) animal occurrences coming from the GBIF platform. This paper presents an overview of the resources and assessment of the GeoLifeCLEF 2019 task, synthesizes the approaches used by the participating groups and analyzes the main evaluation results. We highlight new successful approaches relevant for community modeling like models learning to predict occurrences from many biological groups and methods weighting occurrences based on species infrequency.

Keywords: LifeCLEF, biodiversity, environmental data, species recommendation, evaluation, benchmark, Species Distribution Models, methods comparison, presence-only data, model performance, prediction, predictive power

1 Introduction

The automatic prediction of the species most likely to be observed at a given location is an important issue for many areas such as biodiversity conservation, land management or environmental education. First, it could improve species identification processes and tools by reducing the list of candidate species observable at a given site (whether automated, semi-automatic or based on traditional

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

CEUR-WS.org/Vol-2380/paper_257.pdf

field guides or flora). More generally, it could facilitate biodiversity inventories and compliance with regulatory obligations for the environmental integration of development projects. Finally, it could be used for educational purposes through biodiversity discovery applications offering functionalities such as contextualized educational pathways.

In the context of LifeCLEF evaluation campaign 2019 [6], the objective of the GeoLifeCLEF challenge is to evaluate the state of the art of species prediction methods over the long term and with a view to reproducibility. To achieve this, the challenge freely provides researchers with large-scale, documented and accessible data sets over the long term. Concretely, the aim of the challenge is to predict the list of species that are the most likely to be observed at a given location. Therefore, we provide a large training set of species occurrences and a set of environmental rasters that characterize the environment in a quantitative and qualitative way at any position in the territory. Indeed, it is usually not possible to learn a species distribution models directly from spatial positions because of the limited number of occurrences and the sampling bias. What is usually done in ecology is to predict the distribution of species based on a representation in environmental space, typically a characteristic vector composed of climatic variables (mean temperature at that location, precipitation, etc.) and other variables such as soil type, land cover, distance to water, etc. GeoLifeCLEF's originality is to encourage the extension of this approach to learning a more complex representation space that takes into account various input data such as environmental descriptors, their spatial structure and the known biotic context. Therefore, we provide tools to facilitate the extraction of environmental tensors that can be easily used as input data to models such as convolutional neural networks.

In 2019, the provided data was significantly enriched and several methodological improvements have been made. In more details, the new features introduced are as follows:

1. Pl@ntNet occurrences: to increase the amount of plant occurrences in the training set, we completed the publicly available data from the GBIF[6] with user-generated observations of the Pl@ntNet mobile application [1]. These data are clearly noisier and more biased than conventional occurrence data but they can be filtered by the confidence level of the taxonomic automatic classifier used in the app and they have the advantage of being produced in huge quantities.
2. Occurrences of other kingdoms: to investigate how knowledge of the presence of non-plants organisms can help predict the presence of plants species, we provided a large training set of occurrences from other kingdoms coming from the GBIF platform.
3. A better quality test set: to ensure the reliability of our evaluation, the occurrence data of the test set were restricted to expert data with the highest species identification certainty and high geographical accuracy (lower than 50 m). Last but not least, the test occurrences were sampled in order to avoid, as

[6] `https://www.gbif.org/`

much as possible, biases of spatial coverage and in the species representation. By this way, it contributes to give relatively more importance to rare species and scarce areas.

In the following sections, we describe in more details the data produced and the evaluation methodology used. We then present the results of the evaluation and the analysis of these results.

2 Dataset

2.1 Train occurrences

Pl@ntNet raw data. (`PL_complete`) This data is directly pulled from [4]. Pl@ntNet[7] is a smartphone app using machine learning to identify plant species from pictures submitted by a broad public of users. For each submission, also called a query, the Pl@ntNet algorithm answers a distribution of probability values across the targeted taxonomic referential. If the users allows it, the query's geolocation is also stored. In the provided training data, we used all accurately geolocated queries (with maximum 30 meters uncertainty) in France from the beginning of 2017 to the end of October 2018. Each geolocated occurrence is labelled with the species of higher identification probability. This dataset is thus very heterogeneous in species identification quality, due to the high variability of the image quality submitted by users. The confidence score is provided to GeoLifeCLEF participants as specific field in this dataset, who can use it to account for identification uncertainty in their models. This data set contains 2,377,610 occurrences covering 3,906 plant species.

Pl@ntNet filtered data. (`PL_filtered`) We proposed a filtered version of the previous dataset based on species identification quality. We only kept the occurrences for which the first species probability value was above 0.98. This score has been determined by expert to give a reasonable degree of identification confidence. This set of 237,087 occurrences covers 1,364 species.

GeoLifeClef 2018. (GBIF) Train and test occurrences datasets from the previous year edition [5] were merged to feed the current challenge. Those plants occurrences were extracted from the Global Biodiversity Information Facility [8]. This set of occurrences is around ten times smaller than the Pl@ntNet dataset, as shown in Figure 1. Within this dataset, occurrences are often aggregated on a same geographic point, which denotes uncertain or degraded geolocation. However, the geolocation certainty field is often missing. It contains $281,952$ occurrences covering $3,231$ plant species.

[7] https://plantnet.org
[8] `https://www.gbif.org/`

Occurrences of other kingdoms. (GBIF) This data source is made of species that are not plants, but may interact somehow with plants (e.g. trophic, pollination, symbiosis, use of plant as habitat or shelter), and are thus likely to carry interesting correlations with plant species presences. None of those species are in the list of species to predict in the test set (which are only plant species). Those occurrences have also been extracted from the GBIF; based on the following filters: { Basis of record: Human, Location : include coordinates, Country or area : France }. We extracted occurrences from 7 non-plant taxonomic groups:

- Chordata/ Aves (8,000,000).
- Chordata/ Mammalia (1,300,000)
- Chordata/ Amphibia (300,000)
- Chordata/ Reptilia (200,000)
- Arthropoda/ Insecta (3,250,000)
- Arthropoda/ Arachnida (70,000)
- Fungi/ Basidiomycota (50,000)

It contains 10,618,839 occurrences in total covering 23,893 taxa.

Taxonomic and geographic filters applied to all datasets. Because scientists do not name species by the same way in all regions of the world, many official lists of species names, called referentials, co-exist. There are no exact matching between them (in particular because of the new scientific knowledge acquired during the period between the creation of two separate lists) except those suggested by the scientific latin names themselves. In our case, the distinct data sources don't use the same referentials. Furthermore, distinct species names might be considered as redundant (synonyms) in some referentials. GBIF uses its own referential made from several taxonomic referentials, and GBIF occurrences may not be at the species taxonomic level, but at sub-species, or genus, etc. Pl@ntNet data includes occurrences from several plants taxonomic referentials (like The Plant List[9], GRIN[10], the French National plant list, etc.).
Thus, for attributing species identifiers in GeoLifeCLEF, it was important to first match all occurrences names to a single taxonomic referential adapted for the French Flora. We chose to use Taxref v12 [11] referential. We only kept names matching Taxref v12 according to an exact matching algorithm (R script provided on Github [12]). Some true species might have been lost due to distinct spelling between the GBIF taxonomy and Taxref.
We only kept points falling inside the French territory (Polygon from GADM[13]) or inside a 30 meters buffer zone, to account for geolocation uncertainty. Finally, occurrences were randomly shuffled to avoid any bias introduced by their order of use.

[9] http://www.theplantlist.org/
[10] https://www.ars-grin.gov/
[11] https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en
[12] https://github.com/maximiliense/GLC19/blob/master/GITHUB_taxonomic_and_spatial_filtering.R
[13] https://gadm.org/

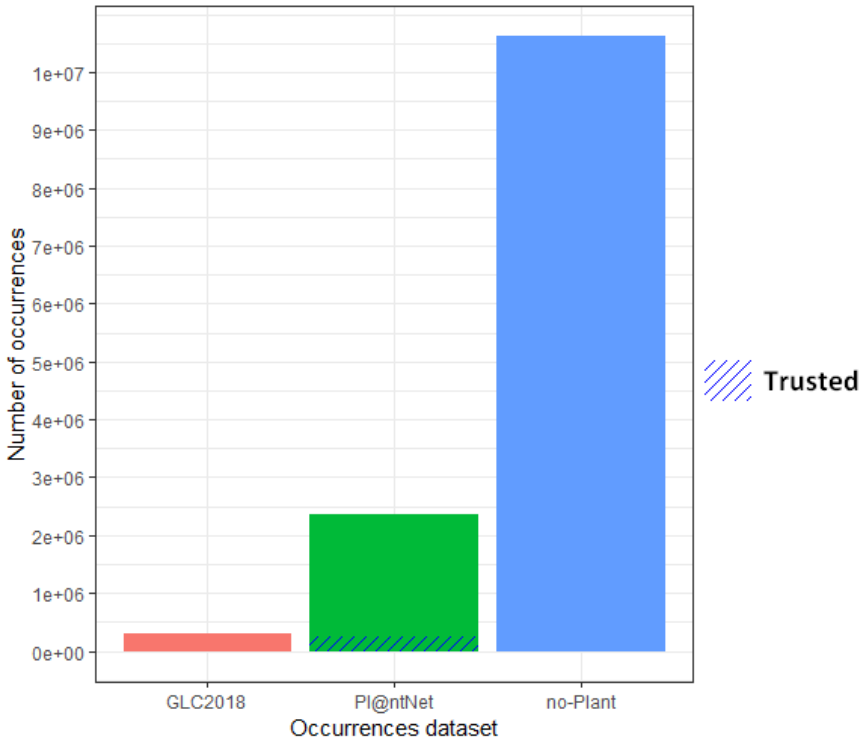


Fig. 1. Number of occurrences per training dataset. Trusted occurrences were determined from Pl@ntNet species identification engine certainty score.

2.2 Environmental data

Geographic rasters. The geographic and environmental data proposed to participants are a compilation of geographic rasters. The variables represented are often used for the purpose of species distribution modelling, especially for plants. The nature of values stored in the rasters are quantitative (bioclimatic, topological, hydrographical and evapo-transpiration variables), ordinal (pedological variables) or categorical (land cover). The rasters are extracted from the data repository of Botella [3], where readers can find a detailed description.

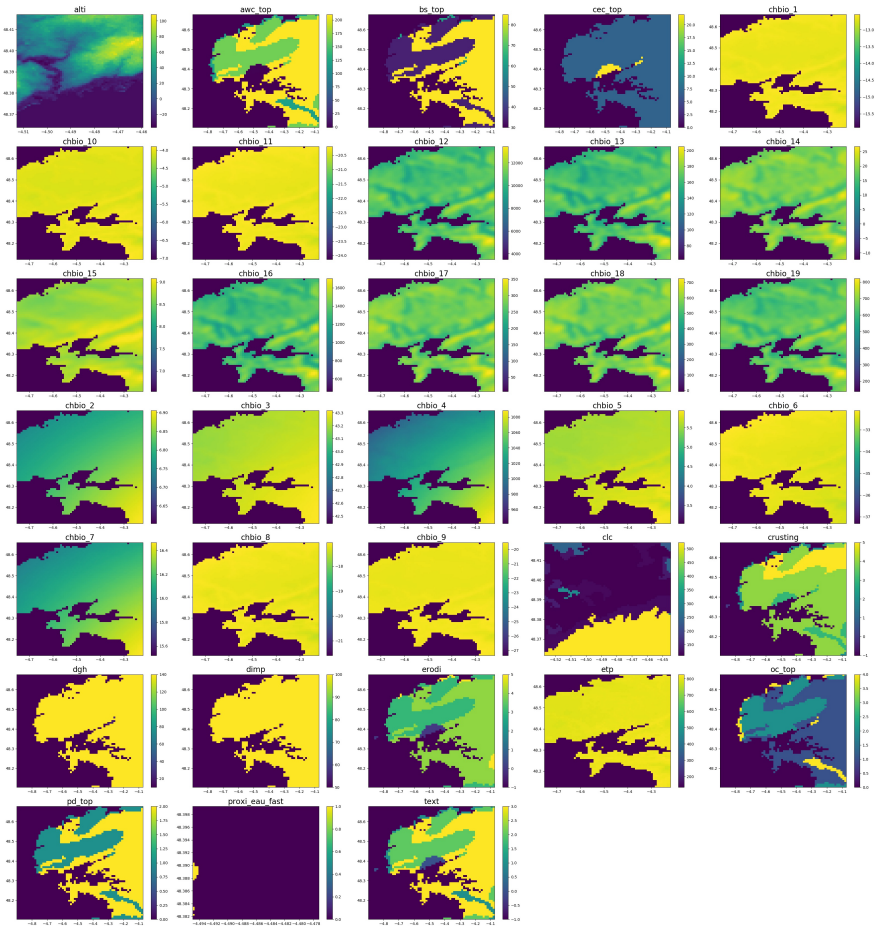


Fig. 2. Patch extracted at the city of Brest, France.

Tensors extraction. To facilitate the learning of representations taking into account the spatial structure of the environment, we provided a Python toolbox[14] allowing to extract local environmental tensors from any position in the rasters. By default, it extracts for each raster a $64x64$ pixels patch centered on the target position and aggregate the patches from all rasters in the form of a tensor of size nx64x64 where n is the number rasters.

2.3 Test data

We have chosen an independent and unpublished source dataset of occurrences for the test set. It is extracted from the *SILENE* database maintained by the *Conservatoire Botanique Mediterranéen* [15]. Those observations come from various providers including the conservatory himself, but also national parks, botanical associations or impact study consultants. We removed species (i) that were not present in the train set, (ii) vulnerable species according to the SINP referential *"espèces sensibles"* [16], (iii) and species that are at least *vulnerable* according to the IUCN red list [17]. This dataset has a high degree of identification certainty because only botanical experts contribute to it. Its geolocation certainty is under 50 meters. We used random weighted selection scheme to draw 25,000 test occurrences among the 700,000 of the initial set noted S. We compute, for each occurrence s_i in S a weight w_i:

$$w_i = 1/(n_i \times r_i)$$

Where r_i is the number of species in the neighborhood of s_i defined by a circle of radius d. n_i is the total number of occurrences in the neighborhood. We define the spatial scale $d = 2$ kilometers. With these weights and the following algorithm, we guaranty that (i) test occurrences are uniformly distributed in the geographic space at scale 2d, (ii) there is as many occurrences of each present species on neighborhoods of radius $2d$. We then draw the test occurrences from S without replacement, through the following algorithm:

- Initialize the bag of test occurrences $S' := S$ and the test set $T = \emptyset$.
- Randomly draw an occurrence in S', say i.
- Draw a scalar $z \sim U(0, max(w_1, ..., w_|S|))$.
- If $z < w_i$, remove i from S' and add it to T, otherwise leave it in S'.
- Stop if $|T| = 25000$, otherwise we go back to step (1).

3 Task description

For every occurrence of the test set, the evaluated systems must return a list of 50 species maximum, ranked without ex-aequo. The main evaluation metric

[14] https://github.com/maximiliense/GLC19
[15] `http://flore.silene.eu/index.php?cont=accueil`
[16] `http://www.naturefrance.fr/languedoc-roussillon/referentiel-des-donnees-sensibles`
[17] `https://uicn.fr/liste-rouge-flore/`

used is the top 30 accuracy (TOP30). We provide its expression hereafter:

$$\text{TOP30} : \frac{1}{Q} \sum_{q=1}^{Q} 1_{\text{rank}_q \leq 30}$$

where Q is the total number of query occurrences x_q in the test set and $rank_q$ is the rank of the correct species $y(x_q)$ in the ranked list of species predicted by the evaluated method for the occurrence x_q.
A secondary metric is the Mean Reciprocal Rank (MRR), a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries ordered by probability. The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer. We provide its expression hereafter :

$$\text{MRR} : \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q}$$

The MRR was used as main metric during last year edition. We compute it this year, in order to enable comparisons between two campaigns.

4 Participants and methods

61 participants registered to the challenge through the online platform, among which 5 participants managed to submit runs in times. A total of 44 runs were submitted. All participants runs methods are characterized by their types of model architecture, the occurrences and input data they used in table 6. In the following paragraph, we describe in more details the methodology of each team.

LIRMM, Inria, Univ. Paul Valery, Univ. Montpellier, France, 4 runs, [10] : This team used a single deep convolutional neural network architecture derived in four models. All models take as input the default environmental tensors extracted by the provided python toolbox (see section 2.1), with a one-hot encoding transformation for each category of the land cover variables (`clc`), inducing 77 layers images in the input of the model. The chosen architecture was an Inception V3 ([13]). Models were trained as classifiers, using a softmax output and a cross-entropy loss (also known as multinomial logistic regression). Model of run 27006 was trained on all occurrences of `PL_complete` and `glc18` datasets, while models 27004 used `PL_complete` with identification score ≥ 0.7, and 27005 used `PL_complete` with identification score ≥ 0.98 (filtered dataset). Furthermore, runs 27004 and 27005 were only trained on a subset of the occurrences: a sample of around 30K occurrences was drawn according to the same selection procedure as for the test set. Thus, all those models predicted only plant species. On the contrary, model 27007 was trained on all occurrences datasets including `PL_complete`, `glc18` and also `noPlants`. This one was trained to predict plant species and many animal species.

SaraSi, EcoSols, UMR 1222 INRA - Montpellier SupAgro, France, 5 runs, [12] : This team used mainly two types of models: a convolutional neural network (CNN) based on the environmental tensors in the same spirit as LIRMM (27086, 27087, 27088) with a customized architecture, and a deep neural network using only a vector of co-occurrences of non-plants taxa as input (27089, 27082). The CNN model architecture separates the feature extraction depending on the type of variables that is deal with. Indeed, it apply distinct convolutional layers to the three categories of environmental patches (continuous, ordinal and categorical). The extracted features are concatenated and used as input in a series of fully-connected layers. A noticeable technique of "categories embedding" was used for the categorical and ordinal patches. It transforms the one-hot encoded patches in a lower number of continuous valued matrices. Also, they addressed the class imbalance of the training set by optimizing a weighted cross-entropy loss so that occurrences of more abundant species were less numerous. They trained this model on the *PL_complete* dataset (27086) and on a reduced version of this dataset to test set species (27088). the run 27087 was like 27086 but trained longer. For the other approach they implemented a customized version of the Continuous Bag of Words model [8]. The input is a set of identifiers of the non-plant "super-taxa" occurring in the neighborhood. An embedding vector associated to the set of "super-taxa" is learned. A "super-taxa" is an aggregation of many species assumed to share a same type of interaction with plants. They were determined through experts knowledge.

SSN_CSE, SSN College of Engineering of Chennai, and VIT University of Vellore, India, 12 run, [7] : This team tackles the challenge with classical machine learning techniques. They relied on three datasets : (i) spatial position of the occurrences only, (ii) spatial position and punctual environmental vector at the position of the occurrence, (iii) spatial position and vector of the average value of the environmental variables within a 16x16 pixels square centered on the occurrence. As a baseline, the authors first propose a probabilistic model where the probability of a species depends on its frequency in the whole training set (Const. prior). In addition, the authors relied on three categories of models. They first used random forest with spatial coordinates only as input (27102), and boosted trees (XGBoost: 26997, 26996, 27O13, 27012, 26988) and artificial neural network (27069, 27070, 27064, 27067) for using either spatial positions, environmental vectors or both. For one neural network, the authors split the features in 5 groups and trained a neural network per group for which predictions are then combined to form a single model.

Atodiresein, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania, 20 runs [2] : This team based their runs on standard machine learning algorithms: nearest neighbors (K-NN), random forests (Rand. For.), boosted trees (XGBoost) and deep neural networks (ANN). Those algorithms were applied to either the *PL_complete* or *PL_trusted* datasets. They used either the

spatial coordinates or the environmental punctual values of a selection of 29 environmental variables, or the concatenation of coordinates and variables. All combinations of algorithms, occurrences data and input data were evaluated on a validation set and the best of them were submitted. They also carried ensemble predictions from those models (runs 26969, 26970, 26958, 27062, 26960, 26971, 26961, 26964, 26968). A partial explanation of the low performances of their runs is that they only answered a short list of species (maximum 5) for each test occurrences, which lowers down performances a lot, especially for the top30 metric.

Lot_of_Lof, Inra, France, 3 runs, [9] : This team used occurrences density estimation based on log-linear spatial in-homogeneous Poisson point processes (PPP). They used a restricted set of environmental variables to model the distribution of occurrences based on expert knowledge: `etp, alti, chbio_5, chbio_12, awc_top, bs_top, slope` and aggregated `clc` in 5 land covers categories. They built their models with the 141 test species having the most occurrences in the *PL_trusted* dataset. Run 27124 is the standard PPP, while runs 27123 and 27063 apply different corrections for spatial sampling bias.

5 Results and discussion

The TOP30 and MRR evaluation scores achieved by all submitted runs are provided in Figures 3 and 4 (numerical values of the TOP30 are also replicated in the third column of Table 6). As a complementary analysis, Figure 5 displays the average TOP30 accuracy obtained for each species in the test set as a function of the number of occurrences of this species in the test set.

These results contributes to drive the following findings:

The occurrences of the other kingdoms significantly improve plants prediction. This can be observed from the comparison of run 27007 and run 27006 of the LIRMM team which are all things equal except the use of the occurrences of other kingdoms. The TOP30 increases from 0.136 to 0.177, which represents an improvement of 30%. The use of the occurrences of the other kingdoms is therefore the main cause of the best performances obtained by this team with regard to the SaraSi team. From the ecological point of view, this suggests that the biotic interactions (competition, predation, facilitation) between plant species and other biological groups play a very important role in determining the distribution of the species. From a deep learning point of view, it means that the convolutional neural network is able to transfer a consistent knowledge from the domain of the other kingdoms to the plant domain. An architecture that aim at predicting so many species through mutual neurons (as run 27007) might be a more efficient design for learning those relationships than using the co-occurrences as input data (as did runs 27089, 27082). It would be interesting to investigate this by comparing the latter strategy with a model taking both environmental patches and co-occurrences as input.

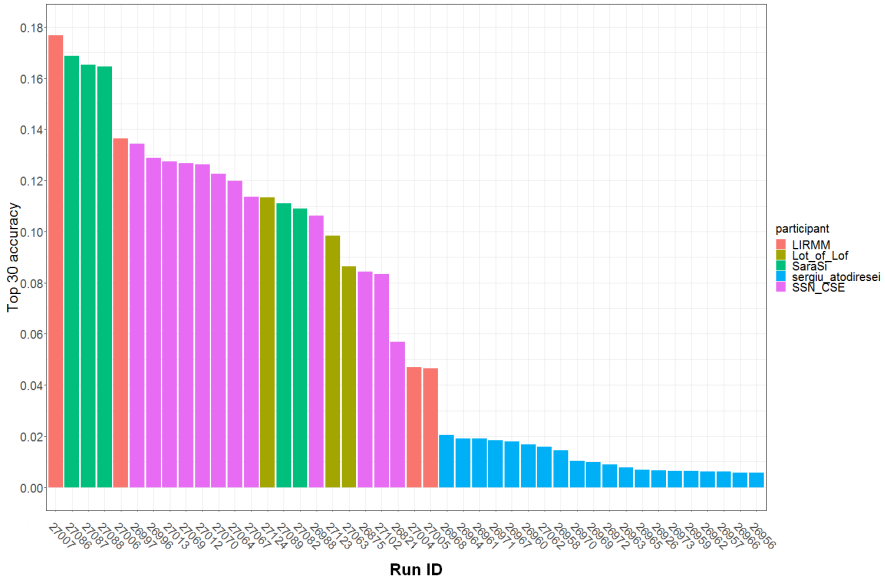


Fig. 3. Average Top30 accuracy per run and participant. It was computed over the 25,000 test occurrences. This was the official ranking metric for the task.

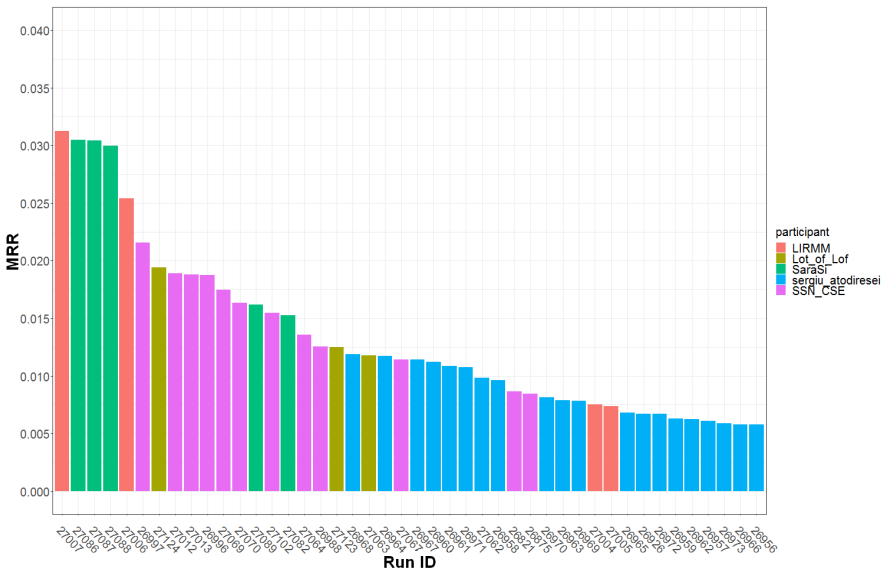


Fig. 4. Mean Reciprocal Rank per run and participant. It was computed over the 25,000 test occurrences.

Weighting the loss by species is better for predicting rare species. The CNN models learnt by the SaraSi team were based on a weighted cross-entropy loss penalizing the classes with more samples as a way to compensate class imbalance. Interestingly, it can be seen in Figure 5 that this significantly increased the ability of the model 27086 to predict the species having few occurrences compared to the winner CNN (run 27007) from LIRMM. Run 27086 is better than 27007 for more than 80% of the species. LIRMM team gave equal weights to all occurrences in the loss for training model 27007. It also shows how the most represented species hide the performances on the majority of species, which rarely occur. Giving more balanced weights across species is certainly important to achieve more robust predictions because the observation preferences across species vary a lot from one biodiversity dataset to another, as it is the case here between Pl@ntNet, the GBIF and SILENE.

The more complex the model, the better the prediction. The analysis of the column "model" of Table 6 suggests that, at least models using environmental inputs, can be ranked according to their performance as: (i) Convolutional Neural Network (CNN), (ii) Boosted trees (XGboost), (iii) Deep Neural Network (ANN), (iv) Poisson point processes, (v) K-Nearest Neighbors. This clearly shows a gradient from the models that integrate the most complex input data (CNN having the most complex with many channels of environmental images) and the most flexible architectures (CNN, XGBoost and ANN can fit very complex functions of their input data), to the models that are the most constrained by their input data (environmental vectors only) and with simple architectures (log-linear model of PPP, no optimized parameters for K-NN). This shows that the size of the available datasets and the complexity of the problem give a real advantage to complex statistical learning methods. More specifically, once again CNN results far exceeded those of the other methods which reinforces the results obtained in the last edition of the challenge. The CNN are likely to extract complex features of spatio-environmental patterns in their highest level neurons which are more suited to describe species habitats than environmental variables designed by experts. They may also captures spatial configurations of habitats that favor certain dispersion mechanisms, e.g. source-think colonization, or detect signatures of particular trophic assemblages.

The training of CNN can fail. Although the best models were based on CNNs, not all CNNs obtained so good results. Indeed, some runs based on CNNs were even worst than the prior ranking of species according to their global abundance (see 27004 $\leq$ 26821). Furthermore, non-submitted CNN models mentioned in a participant working note did perform less in validation than simpler approaches (see [7] 3.4). Model design (architecture, selection of environmental channels, management of categorical variables), regularization (optimization algorithm, use of dropout, learning rate and stopping rule policy), training data (especially size, see runs 27004 and 27005) and occurrence weighting scheme de-

termine jointly the implementation success.

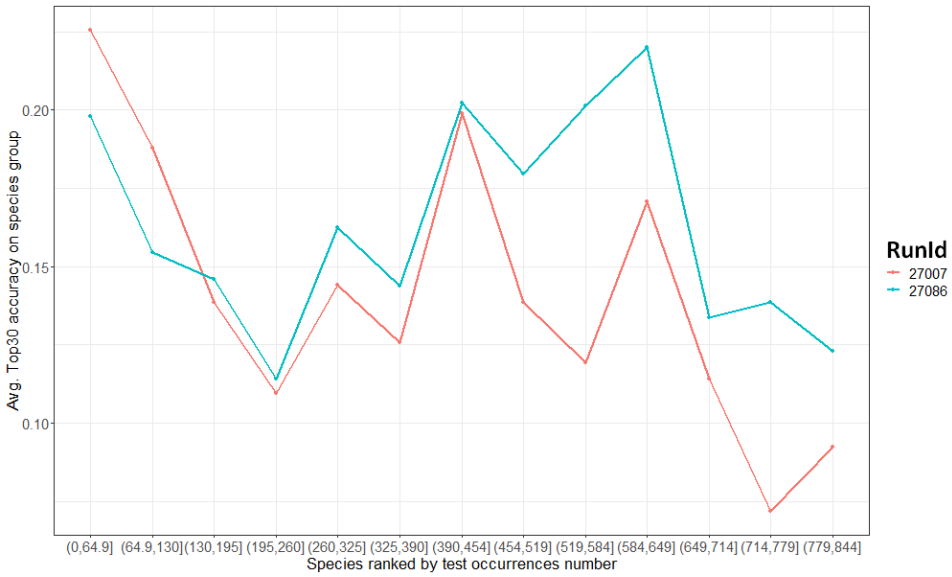


Fig. 5. Top30 accuracy averaged per species abundance class for the two best CNN models. Species were ranked by decreasing number of occurrences in the test set and then aggregated in 14 classes of abundances. For run 27086, each occurrence is weighted inversely proportional to the abundance of its species in the loss function.

Results of the MRR show that performances were globally lower than last year. Indeed, last year average MRR of the ten best runs was 0.039 while it is 0.024 this year. This large global performance gap is probably due to the difficulty of the test set, given that last year dataset was included in the training data. We note that the test set was not identically distributed, firstly because it was located on the Mediterranean region only, but also because the occurrences were sampled to avoid spatial and species biases. We know that all models predict less well rare species and under-sampled areas. Thus, this drop in overall performance supports the idea that the new test set has succeeded in giving greater importance to rare species and sub-sampled areas.
In absolute terms, the best run gives the good answer 20% of the times in its top-30. Thus, roughly speaking, even the best model gives generally a large majority of wrong species in its top-30 list. To give an order of comparison, the database Sophy [11] contains more than 35,000 exhaustive plant species inventories on plots generally not exceeding $400m^2$, and covers a wide range of environments in France. According to it, the species diversity in such plots is 25 in average

and rarely exceeds 70. There is thus large room for improvement in automated predictions.

6 Conclusion and perspectives

We now come back on the main outcomes of this task and discuss its perspectives.

LIRMM best CNN successfully integrated many non-plants species occurrences in their models predictions to better extract spatio-environmental patterns that more robustly predict plants species. It suggests that the global biotic assemblage highly determine the plant assemblage through underlying species interactions, and the multi-species prediction proved again to be a good deep learning strategy to account for it. This is the main new outcome of this year's edition. However, there should be significant room for improvement in the implementation of this approach. Indeed, LIRMM indicated that the winning model training couldn't be finished for time constraints reasons. Furthermore, light and customized models architectures accounting for the different variables natures seem more adapted to the problem than heavily parameterized state-of-the-art image classification architectures. Indeed, SaraSi customized CNN architecture has performed better than the related LIRMM Inception V3 CNN with the same output. Merging the strengths of both strategies promises good improvements in the future.

A rich source of information that remains unexploited for this task is the high resolution satellite images data. For example, today, 50 cm resolution satellite images are freely available for research all over the french territory through the National Institute of Geography (IGN) [18]. Including such images as input in the current models would inform them about very local land cover type and thus give much finer resolution prediction, if one can efficiently handle the size of this data.

The philosophy of the evaluation was to favor models that are more robust to biases in the training data, especially the imbalance of species representation and the heterogeneous spatial coverage, both consequences of the reporting process heterogeneity. We can say that it is a success concerning species imbalance representation. Indeed, SaraSi achieved remarkably stable performances even for rare species through a per class weighting scheme in the cost function. A next step would be to account for spatial sampling heterogeneity, as we have seen that all methods still struggle a lot with scarcely reported areas.

Regarding the evaluation process on this problem globally, we put an effort this year in the quality of the occurrences identification, and corrected for the species imbalance bias and heterogeneous spatial coverage (due to the reporting heterogeneity). Our new evaluation strategy was quite discriminant across the methods, and lowered globally the computed results. In absolute terms, we have also seen that even the best model tends to rank a lot of relevant species (i.e. probably absent from the surroundings) before the good one. The problem of spatial prediction of plant species lists is objectively far from being solved. Still,

[18] `https://geoservices.ign.fr/documentation/geoservices/`

with the new areas of improvements that the task results pointed out, we are optimistic about the future methodological advances on the problem of location based species prediction.

Rk	runId	top30	participant	model archi.	occurrences	covariates	supp. info.
1	27007	0.1769	LIRMM	CNN	all	enviro. tensors	–
2	27086	0.1687	SaraSi	CNN	complete	enviro. tensors	–
3	27087	0.1653	SaraSi	CNN	complete	enviro. tensors	27286 trained longer
4	27088	0.1646	SaraSi	CNN	complete $\cap$ test	enviro. tensors	–
5	27006	0.1364	LIRMM	CNN	complete + glc18	enviro. tensors	–
6	26997	0.1342	SSN_CSE	XGboost	filtered	enviro.+coord.	16 pixels avg.
7	26996	0.1288	SSN_CSE	XGboost	filtered	enviro.\ clc	–
8	27013	0.1273	SSN_CSE	XGboost	filtered	enviro.+coord.	max depth=3
9	27069	0.1268	SSN_CSE	ANN	filtered	enviro. selec.	16 pixels avg.
10	27012	0.1263	SSN_CSE	XGboost	filtered	enviro.+coord.	16 pixel. avg. max depth=3
11	27070	0.1227	SSN_CSE	ANN	filtered	enviro. selec.	–
12	27064	0.1198	SSN_CSE	$(\text{ANN})^5$	filtered	enviro.+coord.	$\neq$ covariates
13	27067	0.1135	SSN_CSE	ANN	filtered	enviro.+coord.	$\neq$ covariates 16 pixel. avg.
14	27124	0.1135	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
15	27089	0.1110	SaraSi	NN	all plants	co-occurrences	large embed.
16	27082	0.1090	SaraSi	NN	all plants	co-occurrences	small embed.
17	26988	0.1063	SSN_CSE	XGBoost	filtered	coord.	–
18	27123	0.0984	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
19	27063	0.0864	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
20	26875	0.0844	SSN_CSE	ANN	filtered	coord.	–
21	27102	0.0834	SSN_CSE	Rand. For.	filtered	coord.	–
22	26821	0.0570	SSN_CSE	Const. prior	filtered	–	–
23	27004	0.0470	LIRMM	CNN	complete score$\geq$0.7	enviro. tensors	test-like sampling
24	27005	0.0465	LIRMM	CNN	filtered	”	”
25	26968	0.0205	atodiresei	$(\text{Rand.For.})^2$	filtered + complete	(enviro.\bio_2 ,text,clc) +coord.	–
26	26964	0.0191	atodiresei	$(\text{Rand.For.})^2$	filtered + complete	(enviro.\bio_2 ,text,clc)	–
27	26961	0.0190	atodiresei	$(\text{1-NN})^2$	filtered + complete	(enviro.\bio_2 ,text,clc) +coord.	–
28	26971	0.0184	atodiresei	1-NN $\times$ RF	filtered + complete	(enviro.\bio_2 ,text,clc) +coord.	–

Rk	runId	top30	participant	model archi.	occurrences	covariates	supp. info.
29	26967	0.0180	atodiresei	Rand. For.	filtered + complete	coord.	–
30	26960	0.0168	atodiresei	3-NN× 5-NN	filtered + complete	coord.	–
31	27062	0.0159	atodiresei	1-NN × RF × ANN × ANN × XGB	filtered	(enviro.\bio_2 ,text,clc)	–
32	26958	0.0146	atodiresei	3-NN× 5-NN	filtered + complete	(enviro.\bio_2 ,text,clc)	–
33	26970	0.0102	atodiresei	1-NN × RF	complete	(enviro.\bio_2 ,text,clc) +coord.	–
34	26969	0.0099	atodiresei	1-NN × RF	filtered	(enviro.\bio_2 ,text,clc) +coord.	
35	26972	0.0089	atodiresei	ANN	filtered + complete	(enviro.\bio_2 ,text,clc)	–
36	26963	0.0079	atodiresei	Rand. For.	complete	(enviro.\bio_2 ,text,clc)	–
39	26973	0.0064	atodiresei	XGBoost	filtered	(enviro.\bio_2 ,text,clc) +coord.	–
40	26959	0.0063	atodiresei	1-NN	complete	coord.	–
41	26962	0.0062	atodiresei	Rand. For.	filtered + complete	coord.	–
42	26957	0.0061	atodiresei	1-NN	complete	(enviro.\bio_2 ,text,clc)	–
43	26966	0.0058	atodiresei	Rand. For.	complete	coord.	–
44	26956	0.0058	atodiresei	1-NN	complete	(enviro.\bio_2 ,text,clc)	–

Table 1. Results and summarized methodology description of all runs submitted to GeoLifeCLEF 2019. Symbols and abbreviations: $A + B$ means that variables/data B was added to A. $A \backslash B$ means that variables/data B where removed from A. $complete \cap test$ means that only test species occurrences from the complete dataset were used. Products ($\times$) and exponent notations in column "model archi." decompose an ensemble methods with its different models. Occurrences: complete=$PL_complete$,filtered=$PL_filtered$, all plants=$PL_complete$ + $PL_filtered$ + $glc18$, all=$PL_complete$ + $PL_filtered$ + $glc18 + nonPlants$. Covariates in model input: "enviro. tensors"=environmental tensors with spatial neighborhood", "enviro."=punctual values of environmental variables, "coord."= spatial coordinates.

Bibliography

[1] Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.C., Joly, A.: Pl@ ntnet app in the era of deep learning. In: ICLR 2017-Workshop Track-5th International Conference on Learning Representations. pp. 1–6 (2017)

[2] Atodiresei, Costel-Sergiu, I.A.: Location-based species recommendation - geolifeclef 2019 challenge. proceedings of CLEF 2019 (2019)

[3] Botella, C.: A compilation of environmental geographic rasters for sdm covering france (version 1) [data set]. Zenodo (2019), http://doi.org/10.5281/zenodo.2635501

[4] Botella, C., Bonnet, P., Joly, A., Lombardo, J.C., Affouard, A.: Pl@ntnet queries 2017-2018 in france. Zenodo (2019), http://doi.org/10.5281/zenodo.2634137

[5] Botella, C., Bonnet, P., Munoz, F., Monestiez, P., Joly, A.: Overview of geolifeclef 2018: location-based species recommendation. In: CLEF 2018 (2018)

[6] Joly, A., Goëau, H., Botella, C., Kahl, S., Poupard, M., Servajean, M., Glotin, H., Bonnet, P., Vellinga, W.P., Planqué, R., Schlüter, J., Stöter, F.R., Müller, H.: Lifeclef 2019: Biodiversity identification and prediction challenges. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 275–282. Springer International Publishing, Cham (2019)

[7] Krishna, Nanda, K.P.K.R.M.P.A.C.J.S.: Species recommendation using machine learning - geolifeclef 2019. proceedings of CLEF 2019 (2019)

[8] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[9] Monestiez, Pascal, B.C.: Location-based species recommendation - geolifeclef 2019 challenge. proceedings of CLEF 2019 (2019)

[10] Negri, Mathilde, S.M.J.A.: Plant prediction from cnn model trained with other kingdom species (geolifeclef 2019: Lirmm team). proceedings of CLEF 2019 (2019)

[11] Ruffray, P., B.H.G.r.G.H.M.: "sophy", une banque de données phytosociologiques; son intérêt pour la conservation de la nature. Actes du colloque "Plantes sauvages et menacées de France: bilan et protection", Brest, 8-10 octobre 1987 pp. 129–150 (1989)

[12] Si-Moussi, Sara, G.E.H.M.D.T.T.W.: Species recommendation using environment and biotic associations. proceedings of CLEF 2019 (2019)

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)

10 Discussion

10.1 Results synthesis

10.1.1 Handling identification uncertainty.

In **chapter 1**, we have introduced a new kind of data for SDM: Automatically identified and geolocated plant pictures coming from Pl@ntNet are taken as species occurrences. We have exposed the nature of major identifications biases: varying confusion level between species, and intrinsic identification uncertainty in the data because the pictures often don't contain appropriate visual information. Both may lead to biases in SDM and it is thus important to filter the occurrences based on a proxy of identification certainty. We have applied MAXENT, a standard presence-only SDM method, to unvalidated Pl@ntNet occurrences of 2016 for 7 alien invasive species. We accounted for spatial sampling bias in MAXENT based on a sampling effort model. We applied this method on several subsets of the whole occurrences filtered with increasing thresholds of confidence value (identification engine probability output). We have shown that (i) for most species and all applied thresholds, models yielded consistent results with models applied on independent expert data, (ii) the average effect of the identification certainty score threshold was positive for low thresholds and negative for high thresholds, (iii) the model performance was variable across species but models were almost exactly ranked according to the species prevalences (because generalist species are known to be more sensitive to biases, e.g. sampling biases, see next section). The optimal threshold to apply may vary depending on the species, the number of occurrences, and we recommend to select it through cross-validation. This experimentation also showed that the models fitted on the Pl@ntNet data identified true species occurrence areas corresponding to absences in the expert data. Indeed, many true specimens were cultivated, point out that Pl@ntNet data could be used to identify invasive species cultivation patterns across territories. However, those specimens bias importantly the distribution model because the species may be cultivated in much more environments than it may survive in the wild, thus they should be removed in order to properly study the distribution of the species in the nature. Other where casual invasive or newly inventoried invasive specimens. Their detection could provide early warnings for territories managers. Separating the occurrences of cultivated plants from the wild ones is an important problem to resolve prior to use the Pl@ntNet data in the modelling of invasive species distribution in the wild. It could be done in a probabilistic way through the recognition of the cultivation context in the visual content of the plant pictures, and by combining the geolocation with high resolution land cover maps as the one recently produced by the CESBIO from the sentinel-2 satellite images[19].

We note that, since experiment of **Chapter 1** on the 2016 data, the Pl@ntNet identification engine greatly improved as it followed the best implementations of the annual PlantCLEF challenges. It was compared to human expert identification two times (Bonnet et al. (2016), Bonnet et al. (2018)). The best algorithms reached the performances of the best expert the second time. This performance increase especially concerns western Europe flora. Also, the data has increased with exponential growth (factor at least 2) every year since, and the combination of both factor must have greatly improved the potential for detection and modelling of alien invasive species in the wild. An application of the Pl@ntNet identification engine of january 2019 on all queries from the begin of 2017 to october 2018 gave birth to a new dataset of 2.2 millions of opportunistic occurrences over France, with a geolocation accuracy above 30

[19] `http://www.cesbio.ups-tlse.fr/multitemp/?p=11778`

meters, and associated with identification confidence values, that is freely accessible at Botella et al. (2019).

10.1.2 Spatial sampling bias correction.

In **Chapter 1**, spatial sampling effort was explicitly modelled with three covariates: Distance to roads, distance to the coastlines and presence of cities. Those covariates were then set constant in prediction, as recommended by Warton et al. (2013). This bias correction had a visible effect on prediction maps, and the bias model especially captured a strong negative effect of distance to road, and effects of cities. However, some important limitations have to be highlighted in this methodology: (i) An effect of distance to coastline has been captured, but we can still observe an increasing density towards coastline in the corrected species intensity maps of *Carpobrotus edulis*, *Erigeron Karvinskianus* and *Opuntia ficus-indica*. Thus, the species intensity covariates can be linearly combined to reproduce a gradient that is similar to distance to coastlines. Then, we are not really able to isolate the effect of distance to coastline on occurrences intensity with this model. (ii) Model specification of the distance to roads and coastlines effects is arbitrary and questionable. (iii) The true abundance of several of our species is known to be closely linked with the covariates of sampling effort and the contribution of abundance and sampling effort on the occurrences intensity may not be separated, which is a fundamental problem of this correction approach.

This leads us to bias correction methods investigated in **chapters 2 and 3** that have the potential to solve problems (ii) and (iii). **Chapter 2** investigated biases arising by the use of multiple species opportunistic occurrences as a proxy of the sampling effort distribution. We carried out a theoretical study of the biases arising with 2 methods of background points selection in SDM based on Poisson point processes for presence-only data. The first is the standard spatially random uniform background selection, while the second use occurrences of a target group of species as background points. For this method, we show that the estimated species intensity fits the product of the true species intensity and the sampling effort, thus bias is particularly strong when the sampling effort is concentrated along environmental gradients. The second method uses occurrences from a set of species, the Target-Group, as background. We have shown that, under constant detection probability, it fits the ratio of the focal species density divided by the sum of Target-Group species intensities (TG species density). Thus, this second method is unaffected by observation bias, but it is affected by variation of the TG species density: The fitted relative intensity will be under-estimated for environmental conditions where the sum of TG species is higher, and conversely it will drastically be over-estimated in environments where the TG species density is low. Reducing the expected bias when applying the Target-Group strategy would require to minimize the variation of whole sum of TG species abundance along every environmental gradients through an appropriate selection of those species.

Our results also prove the conjecture of Dudík et al. (2006), a distribution of background points drawn from the sampling effort spatial density provides an unbiased estimate. The results formalize and extend the first critics of the Target-Group strategy in Warton et al. (2013), which took the example of bias due to varying level of species richness in two environments. Expressing the bias as a function of TG species density generalizes the concept of species richness bias of Warton et al. (2013), in the sense that even with a constant species richness, some species might be very rare in an environment rather than in another.

TGOB procedure directly takes points as background, whereas the TGB procedure usually uses TG sites, i.e. spatial areas where at least a TG occurrence has been reported Phillips

et al. (2009). Thus, TGOB is simpler because we don't have to define a grid of sites for occurrences. Also, the procedure by site is not consistent with the Poisson process framework where the spatial variation of occurrences concentration carries the information of the process local intensity. As a consequence, the bias of TGB depends on both the sampling effort and the TG species distributions in a complex way, contrarily to TGOB which bias doesn't depend on sampling effort. A limit is that we didn't defined rigorously the concept of density concentration. A formal definition of density concentration would enable to study analytically the effect of concentration on bias. For example, it might be useful in order to provide a bias metric bound depending on the level of concentration.

We have seen in **Chapter 2** that we can't eliminate TGOB bias without further knowledge about the TG species distributions. Then, the other strategy of jointly modelling species intensities along with the sampling effort might be improved to solve some limitations identified in **Chapter 1**. Indeed, we need to deal with complex and unknown environmental factors affecting sampling effort on a large spatial scale, and with some species that depend on the same environmental factors.

We thus studied the estimation of sampling effort from presence only data that without relying on much prior hypothesis on environmental drivers of sampling effort. In this perspective, **chapter 3** introduced a Poisson process SDM framework based on the joint modelling of many species intensities along with of the sampling effort. The sampling effort model is a stepwise constant function over a partition of space defined by the modeler, which can typically be a regular spatial grid of squares. The identification of the species response model parameters from the sampling effort model parameters is enabled by the variation of occurrences concentration along the environmental gradients inside the sampling cells. The multi-species model enables to infer relative sampling effort in space, even when a single species is absent. We have shown in simulation that the method works well when the species response model is well specified, i.e. the relevant environmental gradients are chosen and the response function shape may be well approximated, and the sampling effort cells limits correctly reflect variations of the sampling effort. We illustrated our method with an application over around 500,000 occurrences from 300 plant species collected through Pl@ntNet over the 2017-2018 period in France. However, the estimation of sampling effort and species intensities is biased when the sampling effort varies systematically along an environmental gradient inside designed sampling cells.

10.1.3 Multi-species prediction from environmental images using deep learning.

In **Chapter 4**, we investigated the prediction performance of deep learning modelling approaches in the context of SDM. We developed a Poisson regression model whose intensity parameter was modelled by either a deep neural network (DNN), or a convolutional neural network (CNN), predicting counts of species occurrences collected over sites of equal area. We compared these models with the state of the art SDM method MAXENT for randomly chosen species of the french Flora through metrics computed on counts over test sites. We also implemented a multi-species shared network: The final hidden layer of the neural network model is shared and used as input for the linear predictor of all response functions. Multi-response versions of the deep and convolutional networks models were implemented with 50, 200 and 1000 species. It showed that multi-responses neural networks always outperformed single-response models, in particular MAXENT, and were all the more efficient in predicting the test 50 species as they were trained to predict more species. It also shows that many species share common signatures in their response functions, which are extracted robustly inside non-linear environmental features through multi-response models. These shared environmental features suggest that the model recovered patterns of species communities and potentially sampling

effort patterns structured in the environmental space. Also, CNN was always better than DNN, showing that this model architecture has important generalization ability for predicting species distribution, even in distant sites (test sites were at 10km away from train sites). It suggests that either the CNN were able to capture the spatial information of the patch of environmental variables given as input, or it captured transferable spatial patterns of the environment.

Implementing, fitting and testing complex models such as deep neural networks require time and computational resources because of the many possible model designs, optimization techniques and problems, e.g. exploding or vanishing gradient. Regarding any machine learning problem, it is often impossible to compare all promising implementations for a single research team, which motivates collaborative evaluation to foster new species distribution models approaches and engineer more efficient prediction algorithms.

In **chapter 5**, we gave an overview of the two first editions of GeoLifeCLEF international evaluation campaign. It was initiated in 2018 in the context of the CLEF evaluation campaign. It is designed to evaluate location-based species prediction algorithms, which are strongly related to SDM. We here summarize the results of the two editions. It was first shown that Convolutional Neural Networks based on environmental variables image patches (3D-tensors where the third dimension is the type of environmental variable) (Deneu et al. (2018), Negri (2019), Si-Moussi (2019)) had the best results. They were better than other machine learning methods used by the participants : Boosted Trees, Neural Networks, Random Forest, Nearest Neighbors, clustering methods, which were learnt on various type of input (environmental variables, spatial coordinates or species co-occurrences). Even though, the CNN implementations varied from existing architectures taken from image recognition tasks to sophisticated architectures customized for the task (Deneu et al., 2019). Secondly, multi-classes predictions including non-plant species (Negri (2019), 34 000 species in total) showed to largely improve the predictive performance, suggesting that the model could capture more relevant environmental patterns for plants through the added animals distribution. Finally, models accounting for environmental variables improved performances on rare species compared to models based only on spatial coordinates. Environmental CNN were better than other models for predicting rare species, especially when compensating occurrences number imbalance through train loss weighting (Si-Moussi, 2019).

The superiority of environmental CNN models was confirmed on three independent test datasets (Chapter 4, Chapter 5 - GLC18 and Chapter 5 - GLC19) with four different implementations (Chapter 4, Deneu et al. (2018), Negri (2019), Si-Moussi (2019)). However, deep CNN may also deeply fail to generalize. Indeed, some CNN implementations failed in test predictions (Moudhgalya et al. (2018), Taubert et al. (2018), Negri (2019), Krishna (2019)) often because learning complex architectures without overfitting requires specific optimization and regularization techniques (batch stochastic gradient descent, Dropout, batch-normalization and the joint effect of multiple-response/multi-label classification modelling and architecture bottleneck (Chapter 4)), time and computational resources for model tuning. Indeed, it shows that the particular resistance of deep learning to the curse of dimensionality is relative and conditioned on the alchemy of many techniques and practices whose combination is time consuming to find. But these effort were not done in vain, as recent models implementations have shown consistently good performances across space and species. They could be used to develop prediction services for spatial species recommendation on the French territory and help on site-identification, automatically alarm on potential identification errors in the data, provide biodiversity context information for educational purposes. However, more work on the standardization of model regularization would be needed to propose a package or a clearer

methodological guidelines.

10.2 Perspectives

10.2.1 Sampling biases

Guidelines to chose the proper bias correction method. The questions that naturally arise from our work is: Now, how to concretely reduce at maximum sampling bias when faced with the estimation of some species distribution and a dataset of opportunistic occurrences from multiple species?

What we have come to is that the bias induced by the TGOB method is unknown in the reality without information on the TG species real abundances. However, if the modeler have such information for some species, he must constitute the TG so that the sum of their abundance is approximately constant over all environmental gradients. Then, the TGOB method may be the best alternative, because it avoids the increased estimation variance problem due to joint model of the sampling effort and species intensities.

The method proposed in **Chapter 3** should be most suited when we have many species with a large number of occurrences, which together cover most of the study domain. Those species occurrences are expected to contribute together to a good sampling effort estimation everywhere in the spatial domain, which reduces in turn the estimation variance on scarcer species models. We note that the modeller may compare several grid sizes by using standard cross validation and appropriate error metric for density estimators, e.g., the averaged negative log-likelihood (proportional in expectation to the KL-Divergence from the density) or the density cross-validation criterion as introduced in Tsybakov (2009).

The modeler may also follow the approach of Warton et al. (2013) if he is confident about a sampling effort model and has enough occurrences of the focal species. We advise to check the first principal angle between the family of environmental variables vector along a large set of uniformly drawned spatial positions in the domain, and the family of sampling effort variables evaluated at the same position. This angle should be maximised to minimize the confusion between sampling effort and species response. This method may be directly compared with the previous one through the above mentioned cross-validation metrics restricted to the focal species occurrences.

In some cases, the modeler may convert at least a part of its opportunistic occurrences from multiple species into site-occupancy data. Indeed, if the focal species has been observed at least once on a site, all other visits of the site, proven by other species occurrences, may be used as a non-detection data. This method was already proposed in Kery et al. (2010). It pulls more information from the data when the detection and reporting probabilities of observers for the species are not too small. This method would benefit to be applied only to observers that were indeed looking for the focal species, and not indifferent to it, which yields a similar approach than generating absences based on observer reporting behavior information (see Bradter et al. (2018)), but morever account for imperfect detection.

In other cases, we advocate that the methods mentioned above have high risks of biases or errors, and thus the modeler should rather consider integrating complementary standardized data.

Resolving biases by integrating different data types and orientating data collection. Poisson process model enable to compute a joint likelihood of e.g. presence-only, presence-absence and site-occupancy data (Dorazio (2014), Fithian et al. (2015), Koshkina

et al. (2017)) with a shared model component for the species intensity. The more standardized data helps resolve identifiability issues in the less standardized data. Models and methods integrating several types of data have recently become more popular because has many small projects have begin to share their data collected with different sampling protocols (Miller et al., 2019). However, data integration may not be sufficient because the standardized data must be in the right place and concern the right species to be usefull to the model. Many large scale naturalist community platforms have all the ingredients to organize an orientated scheme of somehow standardized data collection. Indeed, inside the large community of Pl@ntNet users, there is a restricted core of very active contributors that would be motivated towards producing greater quality data contribution. Stohlgren and Schnase (2006) already proposed a conceptual iterative sampling design for continuously monitoring invasive species with a constraint of sampling a small portion of the territory. Starting from a likely biased presence-only model of the species distribution fitted on opportunistic occurrences, they propose to determine from this model some areas to sample presence-absence data determined from the presence-only model. Then they propose to fit a more accurate presence-absence model on the newly collected data. It would help next data collection phase, and so on. Today we could automatize and transpose this design in a model based continuously updated sampling design. We could build a statistically sound framework for priorizing new standardized data collection areas. Indeed, using models integrating opportunistic occurrences and standardized data and model based optimal sampling theory (Jacquez and Greif, 1985), we may determine a priorization of sites to collect the new data that optimally improve the estimability of model parameters. Then, proposition of most useful standardized reports could be proposed to experienced and active Pl@ntNet users around their area. We may alternatively envision an active learning scheme, were a model based short list of species representative of the site is proposed to the observer, which confirm of infirm their presence. This kind of automatic active learning system based on distribution models and a network of volunteer contributors is actually not new and was already implemented in the bird watching community (Kelling et al., 2012). It may now be the time for plants to be brought to light too, which they sorely need.

Improving joint estimation of sampling effort and species intensities. A fundamental problem of the joint estimation of **Chapter 3** is that, as we are trying to separate two density signals from one. We separate the sampling effort and the species intensity from the occurrence intensity of each species, with the additional difficulty that occurrence intensity translates in a finite number of points in the data. Even with an infinite number of points, many equivalent mathematical solutions are possible. As we add some assumptions about the signal, the model becomes identifiable, but if the models of species responses or sampling effort is wrongly specified, it may lead to important biases. More precisely, the estimates asymptotically minimizes a weighted sum of the divergences on each species occurrence densities $D_{KL}(\lambda^i(x(.))s(.)||\lambda^i_{\theta_i}(x(.))s_\gamma(.))$. A crucial problem is that our estimates of sampling effort $s_{\hat{\gamma}}$ and species intensities $\lambda^i_{\hat{\theta}_i}$ may not converge to the best approximation of the truth, i.e. the λ^is and s, because either the model of $\lambda^i_{\theta_i}$ or s_γ is not suited. Then, it may exist couples (s_b, λ^i_b) that are very biased compared to their respective truths, but whose product is much closer to the occurrences density $\lambda^i(x(.))s(.)$.

We have observed such behavior in simulations when the sampling effort varied continuously inside sampling cells and systematically along an environmental gradient. For example, in **Chapter 3**, all species had a very high categorical effect for the urban land cover (the intensity was multiplied by a factor three at least compared to all other categories), even for some species that are almost absent from cities and not cultivated. The most likely explanation

is that the true sampling effort intensity in cities has not been captured by urban sampling cells, but instead by the urban categorical effect of all species, because the spatial variation of occurrences concentration was better explained by a change of land cover than a change of sampling cells. The challenge is then to induce in our model a realistic constraint that prevent such undesirable effect. We here propose some ideas to resolve this estimability issue.

A first idea comes from the fact that the sum of all species abundances (per unit of space) is upper bounded. The idea is to penalize locally the sum of all species abundances, i.e. adding a penalization term $\zeta \int_D \sum_i \tilde{\lambda}^i_{\theta_i}(z)dz$ in the negative log-likelihood, where $\tilde{\lambda}^i_{\theta_i} := \lambda^i_{\theta_i}/p_i$ is a re-weighting of the species fitted intensity where p_i is the species global detection and report probability (averaged over seasons). In fact, the fitted species intensity is only approximately proportional to the species abundance, so the sum of intensities does not vary in space like the sum of abundances. This sum of intensity will be much more affected by the variation in intensity of a very remarkable species than by a discrete species, even if the latter is globally more abundant. Thus, penalizing the sum of intensities would not be equivalent to penalizing the sum of real abundances and could induce biases. On the other hand, if the intensity of each species is divided by the probability of detection and reporting of the species in the sum, this sum becomes proportional to the sum of species abundances. Then, our penalty should avoid the model from transferring the sampling intensity into all species locally. This is likely to limit the transfer of sampling effort intensity inside every species response. To avoid other biases, we should include a large set of species representative of the local flora. A way to estimate the species probabilities of detection and reporting is proposed in the following paragraphs.

Another possible realistic constraint would use the knowledge of the generalism of some species over certain environmental gradient to help the model converge to the good estimation. If any species is known to be widely distributed over a certain environmental gradient, or indifferent to the categories of a categorical variable, the corresponding parameters must be "locked" to 0 in the model (e.g. by giving a large penalty hyperparameter to the parameter, or removing the effect from the model). This way, the locked species occurrence density variation will greatly help the model to estimate the sampling effort variation along variation of the locked environmental gradient.

Other improvements of the methods should ideally include deep learning dimensionality reduction techniques of the species intensities environmental features. Indeed, their are numerous likely useful environmental features and they induce both estimation variance and decrease estimability of all species responses and the sampling effort. In **Chapter 4**, we have used deep convolutional NN to extract a small number of environmental features summarizing the most important information of the input for species responses prediction. Such high level environmental features could be used as parsimonious input variables of the species responses in the method of **Chapter 3**. This dimensionality reduction through model architecture bottleneck design might also be directly tested to separate sampling effort and species responses in a deep NN model. E.g. we could dedicate a neuron in the last layer to log-sampling effort, by forcing an equal and constant associated parameter values in the linear predictors of all species.

Accounting for temporal heterogeneity in species detection due to phenology. Plant species don't flower at the same periods of the year, some even pass most of the winter with underground organs and are thus undetectable during this period. It has been shown that observers detection and reporting probabilities are clearly affected by the period of the year for all plants (Burrows, 2004) and by plants flowers and size and other morphological

traits that depends on the species phenology (Kéry and Gregg, 2003). Being able to estimate species detection and reporting probabilities for each period of the year has already been done with opportunistic occurrences of butterfly (Kery et al., 2010) and has two main interests: (i) To be able to compare estimated species intensities to each others, which may be used to get estimation of all species absolute abundance based on one species absolute abundance, and further (ii) to apply any method based on global plant species abundance, or on the species richness, e.g. the regularization technique proposed in the last paragraph. We can moreover use the knowledge that plants are static and most have an at least annual cycle, under temperate climate. Then, if an individual has been observed at some time in the year it is very likely that he has been here all along the year. Then, we may use an occurrence as a proof of presence and generate non-detection data for all the year with all other visits of the site that didn't led to an occurrence of this species, but of another. Then, we may model and fit the probability of detection and reporting, in order to estimate its varying values across seasons, based on a conditional likelihood on the species presence. We emphasize that this data is more informative than site-occupancy data because we have no uncertainty about the presence of the species on the site and may directly estimate our probability. Still, these data may be integrated in a model combining it with site-occupancy data. The main problem here is that we don't know the abundance of the species when it is present, which should affect the detection ability of observers. The proposed idea could then be extended to account for this, e.g. following the lines of N-mixture models Royle (2004b) which account for this abundance effect on detection probability in the context of site-occupancy models. The link of temporal heterogeneity in detection and plants phenology could also be pushed further to help detect spatial variation in species flowering through spatio-temporal variations in their detectability.

Accounting for mis-identification. In presence-only data, mis-identification may globally induce smoother species responses because the real species occurrences are diluted by a contamination of other species that doesn't have the same distribution. When confusions are approximately symmetrical among a group of species, their estimated intensities will become more similar. However, to my knowledge, works accounting for mis-identification in SDM are very rare (but see Guilbault et al. (2019)). In the case of Pl@ntNet, we noticed that species that have the most illustrations in the identification engine tend to get in general higher confidence scores compared to species that have scarcer illustrations. Then, a rare species that looks similar to a well illustrated species will be often missed, and we will lack geolocated data to estimate its distribution. SDM accounting for mis-identification may thus be especially useful to obtain more accurate distributions estimations of rare species. The Pl@ntNet system generates *observations* with high identification certainty, which could be used for estimating a matrix of confusion probabilities between many species. Also an online platform, called ThePlantGame[20] is specifically dedicated to resolve uncertain automatic identifications by attributing identification task to a group human identifiers automatically selected among many volunteers for their estimated capacity to discriminate the good classification among the set of likely species. The system establish final identification through a majority vote. The system actively learns to better qualify the identifiers capacities along time based on their answers Servajean et al. (2017). However, it is not easy to get true out-of-sample data informing about the algorithm species true confusions because most of the valid identification data is soon used for next round training of the identification engine. A practice to avoid it could be to build an extensive validation dataset with an explicit license restricting its specific use without permissions, to prevent this data to be used in future algorithms training

[20] `theplantgame.com`

data. Otherwise, we could use the whole distributions of output probability on species from the algorithm output as a proxy to likely confusions and use it for an appropriate weighting of SDM likelihood. As an illustration, the contribution to the location based classification log-likelihood of an occurrence collected at z that has probability 0.8 to be species A and 0.2 to be species B would be the term $0.8log(p(a|z)) + 0.2log(p(b|z))$.

10.2.2 Point processes ecological interpretations, related assumptions and limits

In this section, We exhibit a simple probabilistic model of plant seedlings and survival to understand under which assumptions the Poisson process intensity may be directly linked to the fundamental niche of a species. Then, we discuss the limitations of Poisson processes when points are interacting in the context of dispersal or biotic interactions, and we point out examples of other point processes to account for such phenomena. Consider that N seeds are distributed uniformly and independently in a geographical area D. Each seed belongs to the same species and we assume that they have the same genotype, and thus the same environmental requirements or fundamental niche. The area has a heterogeneous environment, and is not equally suited for the seeds to grow and produce a mature individual. We define by $p(z)$ the probability that a seed produces a mature individual if it is at location $z \in D$. When the number of seeds N tends to infinity, the set of mature individuals are distributed according to an in-homogeneous Poisson process of intensity function $z \to Np(z)/|D|$. The process intensity is thus proportional to the survival probability p. If each mature individual has the same distribution of probability for its number of children, independently of the environment, then the population growth rate, or absolute fitness, is also proportional to p. In this case, recalling that the fundamental niche was defined as the environmental conditions where the population growth rate is strictly superior to one, its projection in D is inside the region where p is superior to an unknown threshold. In reality, the number of children is likely to depend on the suitability of the environment for reproduction. Also, the uniform seeds rain assumption is obviously wrong in nature, as it primarily depends on the distribution of parent individuals. Thus, unfortunately for the real life, the intensity may not be simply related to the fundamental niche. Poisson process models also encounter limits by their inability to deal with points interactions. Starting back from our conceptual model. The distribution of seeds from the parents can't always be exactly modelled by an in-homogeneous Poisson process. For example, a problematic deviation from Poisson processes is that seeds will be naturally clustered in space. Indeed, as long as parents will disperse seeds in areas that are not completely overlapping, the children points depend on their parent position. Still, we note that this phenomenon may however be modelled with a process built on Poisson processes called the Neymann-Scott process, for which approximate inference methods have been proposed and applied to study the distribution of tropical trees (Waagepetersen (2007), Shen et al. (2009)). Another limitation of the described model is that it doesn't integrate the interactions of the individual with other species that plays at fine spatial grain. These interactions (facilitation, competition, predation) with a set of locally growing species multiplies by a positive factor (<1 if unfavorable, >1 otherwise) the probability of survival of the focal species. Simultaneously, the focal species participates to modify the survival probability of all other species growing locally. Thus biotic interactions introduce another level of dependency which is not, this time, between the points of a given species process, but between points of distinct species. For example, Illian et al. (2009) modelled the influence of trees locations over reseeders locations in a small plot plant community with species diversity. More precisely, they explicitly modelled the effect of individuals locations from 19 resprouters species over the spatial intensity of individuals from 5 reseeders species, using Poisson process conditional likelihood. Another

kind of point process accounting for interaction between points is the Cox process: It as a point process whose intensity function is the exponential of a gaussian process. A multivariate Cox process has recently been used to jointly model spatial point patterns of multiple species distributions (Waagepetersen et al., 2016) which may be due to positive biotic interaction. However modelling explicitly pairwise species spatial interaction is much more challenging with this kind of model, and has never been done to my knowledge. Recently, Schnoerr et al. (2016) showed the spatio-temporal cox process may be used as a mean to estimate efficiently parameters of stochastic reaction-diffusion processes. This might be used in the future for the inference of a more mechanistic process of species dispersal from occurrence data.

Thus, the intensity estimated by a Poisson process model should not be rigorously interpreted to be proportional to the probability of survival alone, but as the product of colonization intensity, survival probability, locally reweighted by a global biotic interaction factor due to the biotic context. This intensity is also consistant with the estimate of the Resource Selection Function in the context of animal ecology (Aarts et al. (2012), McDonald et al. (2013)).

We can conclude that because inhomogeneous Poisson point processes have a clear probabilistic basis, they enable a clearer interpretation of the intensity that is estimated even though this framework is, alike other presence-only SDM methods, limiting for sophisticated ecological models because it doesn't account properly for reproduction, colonization and species interactions processes. Putting apart the difficulty of inference and implementations for sophisticated point processes, it must be highlighted that important care should be taken regarding the estimability of parameters in models including several complex ecological processes. Non-identifiability of the model, or confusion of effects may appear in a similar way to what we have shown for disentangling species abundance and observation effort.

10.2.3 Studying transferability of complex species distribution models

Transferability of habitat features learnt by deep NN SDM. As we have seen in **section 5.3.2-3**, the actual frontier between Species Distribution (or habitat suitability) and Ecological Niche Models (ENM) is blurred and porous. They use fundamentally similar statistical tools, while their essential differences consists in different degree of input prior knowledge and strength of about assumptions about species ecology (Peterson and Soberón, 2012). Ecological Niche Models clearly aim at transferability of predictions in space (Randin et al., 2006), time and under global changes of the climate, land use or cover. Complex statistical models such as machine learning SDM approaches are often avoided because of their propensity to fit contingent or too indirect explanatory variables effects on the species response with poor generalization power. Indeed, often the variables selection for model input is done without prior knowledge of real ecological requirements of the species.

However, we would point out that precisely incorporating variables whose effect on the species is not directly known and using complex automatic learning methods, can sometimes, on the contrary, reveal new habitat features that are transferable and important for the species ecology. Also, a known effect of some habitat features may not be properly taken into account inside a hand-made species response function model. This may be detected by the ability of a model to give good predictions when evaluated in distant areas.

For instance, a complementary analysis of the GeoLifeCLEF 2019 was carried out to visualize the effect of the distance to training occurrences on out-of-sample predictive accuracy for the evaluated algorithms. It is displayed in Figure 6. The two best methods are based on deep convolutional NN (runs 27007 and 27086), and we see that, even if they seem to overfit around the training data (because their test performance is relatively small at short distance), the performance increases until around 3-4 kilometers, while simpler models like Maxent (run

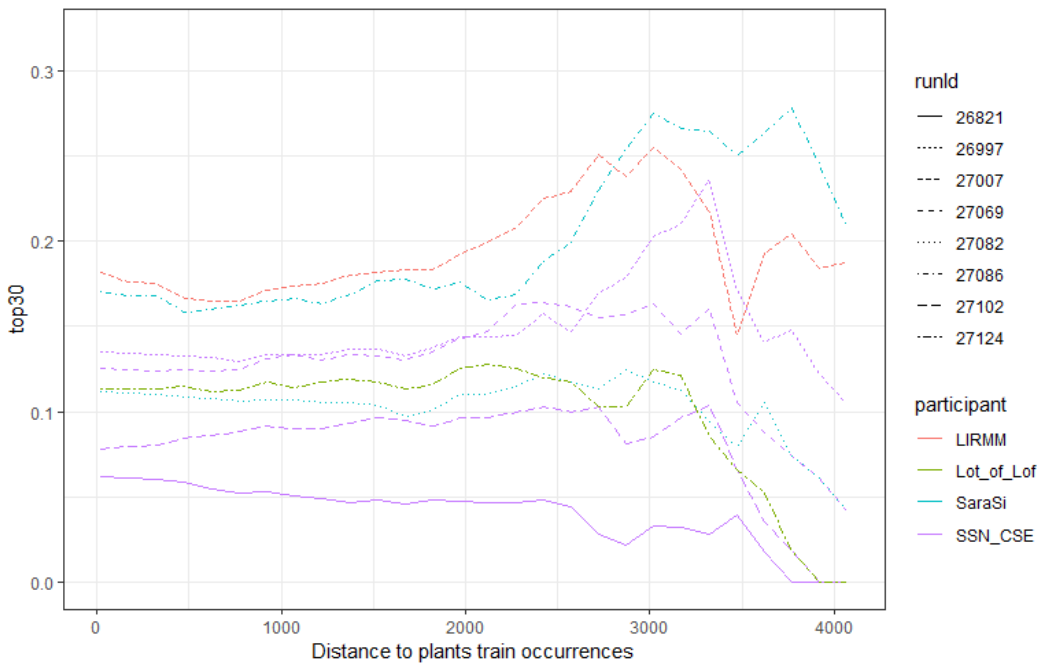


Figure 6: Smoothed TOP30 test accuracy versus geographical distance (in meters) to plants training occurrences for several runs of GeoLifeCLEF 2019. 27007 is the winner environmental CNN; 27086 is another environmental CNN implementation ranked 2nd on the task; 26997 is XGBoost, a method related to BRT; 27069 is a deep NN based on environmental features; 27124 is Maxent; 27102 is a Random Forest with only spatial coordinates as input variables; 26821 is a constant prediction based on whole species abundances in the training set.

27124) have their maximum predictive power at short distances and it decreases further. It suggests that convolutional NN captured some more spatially transferable habitat features than simpler models.

Learning deep NN leads to the construction of activation functions, which can be seen as synthetic variables of the information contained in the images of environmental variables. Indeed, these activation functions, later called habitat features, summarize all the information of the explanatory variables useful to predict the composition of species of a place. They can be thus interpreted as attributes for qualifying the ecological niches of each species. In this sense, machine learning may also be used as a tool in the process of discovering new transferable features for improving ENM.

Still, the knowledge discovery process is not straightforward, it would require to (i) empirically determine models having a high transferability potential with an appropriate evaluation, (ii) disentange the features providing transferable power from those that don't, and (iii) characterize in an human communicable way new transferable features, that also contributes to interpret them ecologically to enrich ENM.

Identifying transferable habitat features. This could be done by statistically analysing the link between the activations of high-level environmental features with the out-of-sample model performance on test data, typically taken at distant places.

Interpretability of transferable habitat features and species niches. Interpretation of deep SDM can be decomposed in two aspects: (i) understanding what habitat features

mean, (ii) and understanding which habitat features describe which species.

Firstly, it is important to understand the ecological meaning of transferable habitat features, both with respect to variables and species, in order to be able to interpret the model as a whole. Methods for interpreting what patterns are used by convolutional NN models features and how each neurons of a model correlate with external interpretable concepts have seen important advances with medical applications (see for example Selvaraju et al. (2017) and Graziani et al. (2018)) in last years. For convolutional NN based on environmental image patches, we could empirically characterize the relationships between habitat features activation and some expert defined concepts of the images: mean values, variance, textures, geometric patterns, and their logical combinations. The effect of modifications of the input data with various transformations could also be studied.

Secondly, a complementary analysis is to characterize the distribution of species in the space of neurons values, i.e. to understand which environmental concept is useful to describe which species niche and conversely to what type of species may correspond a neuron activation. For instance, **Chapter 4** showed that 1000 species responses could be well predicted with 200 neurons (habitat features), which suggests some mutualization of the neurons activation between the species. Some examples of questions emerging from this statement would be: Does species with atypical niches have specific neurons? Does species with typical restricted niche share neurons with others species? Does generalist species mobilize more neurons? On the theory side, is there some parsimony rule in the way deep NN models match neurons and output?

10.2.4 Improving SDM predictions by accounting for species interactions

Most SDM models used in the works introduced here predict species responses conditionally to the environment. They are not able to use the information of a given species presence to better predict other species. However, this information may be very relevant as suggested by the Eltonian conception of the species niche Elton (1927). Even the description of the environment itself is partial or erroneous, and other present species may inform about unobserved environmental factors.

The role of biotic interactions in determining the species distribution lacks documentation and has been identified as a limitation for reliable prediction (Davis et al., 1998). Even though, there is evidence that using variables of interacting species presence in SDM may improve their predictions (Heikkinen et al., 2007). We have introduced and evaluated a class of method that implicitly integrate potential species interactions through a joint modelling of multiple species responses based on the environment. This type of model predict an expected species composition based on the environment $p(y_1, ..., y_N|x)$. It uses a shared neural network between all predicted species until the last hidden layer (Chapter 4), and this type of model was since evaluated in several studies (Deneu et al. (2018), Negri (2019), Deneu et al. (2019), Si-Moussi (2019)). Because the number of neurons is restricted compared to the number of species (architecture bottleneck), it allows to extract environmental features that are linearly combined to activate groups of species. Thus, we expect this method to extract environmental patterns of species communities. This small number of features learnt to predict many species have a fundamentally similar role as the archetypes species in the response functions mixture model of Dunstan et al. (2011). However, this type of model can't predict species responses conditionally to the presences and absences of other species in the surrounding, e.g. $p(y_1|y_2, x)$. Indeed, we would like to condition the focal species likelihood on other species because it provides extra information on unobserved local environment and potential biotic interactions. It has been enabled by the joint SDM framework (Ovaskainen et al. (2010), Pollock et al. (2014))

where multiple species responses are modelled through multivariate generalized linear models including latent variables. In jSDM, the species responses are conditionally independent given the latent variables z, i.e. $p(y_1, y_2|x, z) = p(y_1|x, z)p(y_2|x, z)$. Then, equation 7 shows that, contrarily to the first approach, a species distribution is in general not independent of another species conditionally on the environment because a dependence is introduced through the unobserved variable.

$$p(y_1, y_2|x) = \int p(y_1, y_2|x, z)p(z)dz = \int p(y_1|x, z)p(y_2|x, z)dz \neq p(y_1|x)p(y_2|x) \tag{7}$$

We mention that a method for jointly learning multiple species responses to the environment, through a neural network, and latent variables effects, as previously, has been proposed by Chen et al. (2016).

However, it is generally hard to successfully optimize generalized linear models with a high number of latent variables. Thus, this framework is limiting to account for complex effects of the many environmental variables and other species on a focal species. Other approaches directly infer expected response of the focal species conditionnally on the environment and other species observed responses, i.e. $p(y_1|y_2, x)$. For example, a neural network combining features learnt from convolutional layers applied to environmental patches and other features learnt from species co-occurrences has been proposed and fitted by Deneu et al. (2019). This model was learnt on the GeoLifeCLEF 2018 dataset and performed better than the best run of this task on the test data.

References

Aarts, G., Fieberg, J., and Matthiopoulos, J. (2012). Comparative interpretation of count presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, 3:177–187.

Affouard, A., Lombardo, J.-C., Goeau, H., Bonnet, P., and Joly, A. (2019). Pl@ntnet. In *https://hal.archives-ouvertes.fr/hal-02096020/*.

Ahumada, J. A., Silva, C. E., Gajapersad, K., Hallam, C., Hurtado, J., Martin, E., McWilliam, A., Mugerwa, B., O'Brien, T., Rovero, F., et al. (2011). Community structure and diversity of tropical forest mammals: data from a global camera trap network. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578):2703–2711.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of applied ecology*, 43(6):1223–1232.

Araújo, M. B., Cabeza, M., Thuiller, W., Hannah, L., and Williams, P. H. (2004). Would climate change drive species out of reserves? an assessment of existing reserve-selection methods. *Global change biology*, 10(9):1618–1626.

Araújo, M. B. and Pearson, R. G. (2005). Equilibrium of species' distributions with climate. *Ecography*, 28(5):693–695.

Atodiresei, Costel-Sergiu, I. A. (2019). Location-based species recommendation - geolifeclef 2019 challenge. *proceedings of CLEF 2019*.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2-3):101–118.

Austin, M. and Smith, T. (1990). A new model for the continuum concept. In *Progress in theoretical vegetation science*, pages 35–47. Springer.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005a). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666.

Baddeley, A., Turner, R., Waagepetersen, R., Berthelsen, K. K., Schuhmacher, D., Qi, A. A., Beale, C., Biggerstaff, B., Biv, R., Bonneu, F., Burgos, J., Chen, J. B., Chin, Y. C., La, M. D., Diggle, P. J., Eglen, S., Gault, A., Genton, M., Grabarnik, P., Graf, C., Franklin, J., Hahn, U., Hansen, M. B., Hazelton, M., Heikkinen, J., Hornik, K., Ihaka, R., John-ch, R., Laake, J., Mateu, J., Mccullagh, P., Mi, X. C., Moller, J., Nielsen, L. S., Parilov, E., Picka, J., Reiter, M., Ripley, B. D., Rowlingson, B., Rudge, J., Sarkka, A., Schladitz, K., Scott, B. T., m, I., Spiess, M., Stevenson, M., Surovy, P., Turlach, B., Burgel, A. V., Wang, H., and Wong, S. (2005b). spatstat: Spatial point pattern analysis, modelfitting and simulation. r package version.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350.

Bahn, V. and McGill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, 16(6):733–742.

Barbet-Massin, M., Rome, Q., Villemant, C., and Courchamp, F. (2018). Can species distribution models really predict the expansion of invasive species? *PloS one*, 13(3):e0193085.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536.

Beaman, R., Wieczorek, J., and Blum, S. (2004). Determining space from place for natural history collections. *D-lib Magazine*, 10(5):9.

Beerling, D. J., Huntley, B., and Bailey, J. P. (1995). Climate and the distribution of fallopia japonica: use of an introduced species to test the predictive capacity of response surfaces. *Journal of Vegetation Science*, 6(2):269–282.

Belloni, A. and Chernozhukov, V. (2010). Post-l1-penalized estimators in high-dimensional linear regression models. Technical report, cemmap working paper.

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38.

Bonnet, P., Goëau, H., Hang, S. T., Lasseck, M., Šulc, M., Malécot, V., Jauzein, P., Melet, J.-C., You, C., and Joly, A. (2018). Plant identification: experts vs. machines in the era of deep learning. In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 131–149. Springer.

Bonnet, P., Joly, A., Goëau, H., Champ, J., Vignau, C., Molino, J.-F., Barthélémy, D., and Boujemaa, N. (2016). Plant identification: man vs. machine. *Multimedia Tools and Applications*, 75(3):1647–1665.

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–77.

Born, W., Rauschmayer, F., and Bräuer, I. (2005). Economic evaluation of biological invasions—a survey. *Ecological Economics*, 55(3):321–336.

Botanica, T. (2019). Carnet en ligne, occurrences dataset. https://doi.org/10.15468/rydcn2 accessed via GBIF.org on 2019-07-28.

Botella, C. (2019). A compilation of environmental geographic rasters for sdm covering france (version 1) [data set]. Zenodo. `http://doi.org/10.5281/zenodo.2635501`.

Botella, C., Bonnet, P., Joly, A., Lombardo, J.-C., and Affouard, A. (2019). Pl@ntnet queries 2017-2018 in france. Zenodo. `http://doi.org/10.5281/zenodo.2634137`.

Box, E. (1981). Macroclimate and plant forms: an introduction to predictive modelling in phytogeography. dr. w. *Junk, The Hague, The Netherlands*.

Boyle, B., Hopkins, N., Lu, Z., Garay, J. A. R., Mozzherin, D., Rees, T., Matasci, N., Narro, M. L., Piel, W. H., Mckay, S. J., et al. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC bioinformatics*, 14(1):16.

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A., and Snäll, T. (2018). Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution*, 9(7):1667–1678.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brisse, H., De Ruffray, P., Grandjouan, G., and Hoff, M. (1995). European vegetation survey. the phytosociological database "sophy". part 1. calibration of indicator plants. part 2. socio-ecological classification of the relevés. In *In 4th International Workshop 'European Vegetation Survey' (IAVS)*, volume 53, pages 177–223.

Bromiley, P. (2003). Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, 3(4):1.

Brooker, R. W., Maestre, F. T., Callaway, R. M., Lortie, C. L., Cavieres, L. A., Kunstler, G., Liancourt, P., Tielbörger, K., Travis, J. M., Anthelme, F., et al. (2008). Facilitation in plant communities: the past, the present, and the future. *Journal of ecology*, 96(1):18–34.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (2005). Distance sampling. *Encyclopedia of biostatistics*, 2.

Burrows, G. (2004). The importance of seasonality in the timing of flora surveys in the south and central western slopes of new south wales. *Cunninghamia*, 8(4):514–520.

Busby, J. (1991). Bioclim-a bioclimate analysis and prediction system. *Plant protection quarterly (Australia)*.

Calenge, C., Chadoeuf, J., Giraud, C., Huet, S., Julliard, R., Monestiez, P., Piffady, J., Pinaud, D., and Ruette, S. (2015). The spatial distribution of mustelidae in france. *PloS one*, 10(3):e0121689.

Carpenter, G., Gillison, A., and Winter, J. (1993). Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity & Conservation*, 2(6):667–680.

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., and Joly, A. (2017). Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology*, 17(1):181.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):757–776.

Champ, J., Lorieul, T., Bonnet, P., Maghnaoui, N., Sereno, C., Dessup, T., Boursiquot, J.-M., Audeguin, L., Lacombe, T., and Joly, A. (2016). Categorizing plant images at the variety level: Did you say fine-grained? *Pattern Recognition Letters*, 81:71–79.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., et al. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, 213:280–294.

Chaudhari, P. and Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE.

Chen, D., Xue, Y., Chen, S., Fink, D., and Gomes, C. (2016). Deep multi-species embedding. *arXiv preprint arXiv:1609.09353*.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. (2015). Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., and Liu, J. (2017). Effects of grain size and niche breadth on species distribution modeling. *Ecography*.

Coron, C., Calenge, C., Giraud, C., and Julliard, R. (2017). Estimation of species relative abundances and habitat preferences using opportunistic data. *arXiv preprint arXiv:1706.08281*.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Davis, A. J., Jenkinson, L. S., Lawton, J. H., Shorrocks, B., and Wood, S. (1998). Making mistakes when predicting shifts in species range in response to global warming. *Nature*, 391(6669):783.

De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation theory*, 6(1):50–62.

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251.

Deneu, B., Servajean, M., Botella, C., and Joly, A. (2018). Location-based species recommendation using co-occurrences and environment- geolifeclef 2018 challenge. In *CLEF working notes 2018*.

Deneu, B., Servajean, M., Botella, C., and Joly, A. (2019). Evaluation of deep species distribution models using environment and co-occurrences. In *CLEF 2018 Best of Labs*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484.

Dudík, M., Phillips, S. J., and Schapire, R. E. (2006). Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330.

Dumas, Y. (2011). Que savons-nous du raisin d'amérique (phytolacca americana), espèce exotique envahissante ? synthèse bibliographique. *Rendez-vous techniques ONF, 2011*, pages 48–57.

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, 222(4):955–963.

Dutrève, B. and Robert, S. (2016). Inpn (inventaire national du patrimoine naturel): Données flore des conservatoires botaniques nationaux (cbn) agrégées par la fédération des conservatoires botaniques nationaux (fcbn). Version 1.1. Service du Patrimoine naturel (SPN), Muséum national d'Histoire naturelle, Paris, France. Occurrence Dataset https://doi.org/10.15468/ omae84 via GBIF.org [accessed 30 August 2017].

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, pages 129–151.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813.

Ellenberg, D. and Mueller-Dombois, D. (1974). *Aims and methods of vegetation ecology*. Wiley New York.

Elton, C. S. (1927). *Animal ecology*. University of Chicago Press.

Farber, O. and Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance. *Ecological modelling*, 160(1-2):115–130.

Fernández, D. and Nakamura, M. (2015). Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, 299:147–155.

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic biology*, 51(2):331–363.

Fiske, I., Chandler, R., et al. (2011). Unmarked: an r package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of statistical software*, 43(10):1–23.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS one*, 9(5):e97122.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.

Frair, J. L., Nielsen, S. E., Merrill, E. H., Lele, S. R., Boyce, M. S., Munro, R. H., Stenhouse, G. B., and Beyer, H. L. (2004). Removing gps collar bias in habitat selection studies. *Journal of Applied Ecology*, 41(2):201–212.

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Fried, G., Norton, L. R., and Reboud, X. (2008). Environmental and management factors determining weed species composition and diversity in france. *Agriculture, ecosystems & environment*, 128(1-2):68–76.

Fried, G., Petit, S., and Reboud, X. (2010). A specialist-generalist classification of the arable flora and its response to changes in agricultural practices. *BMC ecology*, 10(1):20.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Garrard, G. E., McCarthy, M. A., Williams, N. S., Bekessy, S. A., and Wintle, B. A. (2013). A general model of detectability using species traits. *Methods in Ecology and Evolution*, 4(1):45–52.

Gaston, K. J. and O'Neill, M. A. (2004). Automated species identification: why not? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):655–667.

Gause, G. (1936). The struggle for existence. *Soil Science*, 41:159.

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm. In *IET*.

Ghazi, M. M., Yanikoglu, B., and Aptoula, E. (2017). Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*, 235:228–235.

Gillings, S., Balmer, D. E., Caffrey, B. J., Downie, I. S., Gibbons, D. W., Lack, P. C., Reid, J. B., Sharrock, J. T. R., Swann, R. L., and Fuller, R. J. (2019). Breeding and wintering bird distributions in britain and ireland from citizen science bird atlases. *Global Ecology and Biogeography*.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Giraud, C., Calenge, C., Coron, C., and Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, 72(2):649–658.

Gnecco, G. and Sanguineti, M. (2008). Approximation error bounds via rademacher's complexity. *Applied Mathematical Sciences*, 2(4):153–176.

Goeau, H., Bonnet, P., and Joly, A. (2017). Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). In *Working notes of CLEF 2017–Conference and labs of the evaluation forum, 11–14 September 2017, Dublin, Ireland.* CEUR Workshop Proceedings 1866: `ceur-ws.org/Vol-1866/invited_paper_9.pdf`.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning.* MIT press.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., and Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in ecology & evolution*, 19(9):497–503.

Graham, E. A., Henderson, S., and Schloss, A. (2011). Using mobile phones to engage citizen scientists in research. *Eos, Transactions American Geophysical Union*, 92(38):313–315.

Graziani, M., Andrearczyk, V., and Müller, H. (2018). Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132. Springer.

Greenwood, J. J. (2007). Citizens, science and bird conservation. *Journal of Ornithology*, 148(1):77–124.

Gribonval, R., Kutyniok, G., Nielsen, M., and Voigtlaender, F. (2019). Approximation spaces of deep neural networks. *arXiv preprint arXiv:1905.01208.*

Grinblat, G. L., Uzal, L. C., Larese, M. G., and Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, 127:418–424.

Grinnell, J. (1917). Field tests of theories concerning distributional control. *The American Naturalist*, 51(602):115–128.

Guilbault, E., Renner, I., Mahony, M., and Beh, E. (2019). Classification of unlabelled observations in species distribution modelling using point process models. *bioRxiv*, page 651125.

Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3):89–100.

Guisan, A., Thuiller, W., and Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R.* Cambridge University Press.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., et al. (2013). Predicting species distributions for conservation decisions. *Ecology letters*, 16(12):1424–1435.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135(2):147–186.

Guralnick, R. P., Wieczorek, J., Beaman, R., Hijmans, R. J., Group, B. W., et al. (2006). Biogeomancer: automated georeferencing to map the world's biodiversity data. *PLoS biology*, 4(11):e381.

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582–591.

Hanski, I. (1999). *Metapopulation ecology*. Oxford University Press.

Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, 1(2):113–129.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hawthorne, W. and Lawrence, A. (2013). Plant identification: creating user-friendly field guides for biodiversity management. *Routledge, 288 Pages, ISBN 9781844070794*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M.-N., Schmidtlein, S., Turner, W., Wegmann, M., and Pettorelli, N. (2015b). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, 1(1):4–18.

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., and Körber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, 16(6):754–763.

Heywood, V. H., Watson, R. T., et al. (1995). *Global biodiversity assessment*, volume 1140. Cambridge University Press Cambridge.

Hill, A. W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V., and Remsen, D. (2009). Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC bioinformatics*, 10(14):S3.

Hirzel, A. H., Hausser, J., Chessel, D., and Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, 83(7):2027–2036.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., and Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological modelling*, 199(2):142–152.

Hobbs, S. J. and White, P. C. (2012). Motivations and barriers in relation to community participation in biodiversity recording. *Journal for Nature Conservation*, 20(6):364–373.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., and Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46:523–549.

Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography (MPB-32)*. Princeton University Press.

Hui, F. K., Warton, D. I., Ormerod, J. T., Haapaniemi, V., and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26(1):35–43.

Hutchinson, G. E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding remarks*, 22:415–427.

Illian, J. B., Møller, J., and Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, 16(3):389–405.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

Isaac, N. J., Mallet, J., and Mace, G. M. (2004). Taxonomic inflation: its influence on macroecology and conservation. *Trends in ecology & evolution*, 19(9):464–469.

Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P., and Roy, D. B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10):1052–1060.

Jackson, S. T. and Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late quaternary. *Paleobiology*, 26(S4):194–220.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.

Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., Dufour-Kowalski, S., Affouard, A., Carré, J., Molino, J.-F., et al. (2016). A look inside the pl@ ntnet experience. *Multimedia Systems*, 22(6):751–766.

Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.-P., Planqué, R., and Müller, H. (2018). Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 247–266. Springer.

Joly, A., Goëau, H., Botella, C., Kahl, S., Poupard, M., Servajean, M., Glotin, H., Bonnet, P., Vellinga, W.-P., Planqué, R., et al. (2019). Lifeclef 2019: Biodiversity identification and prediction challenges. In *European Conference on Information Retrieval*, pages 275–282. Springer.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Planqué, R., Rauber, A., Fisher, R., and Müller, H. (2014). Lifeclef 2014: Multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

Just, A., Gourvil, J., Millet, J., Boullet, V., Milon, T., Mandon, I., and Dutrève, B. (2015). Siflore, a dataset of geographical distribution of vascular plants covering five centuries of knowledge in france: Results of a collaborative project coordinated by the federation of the national botanical conservatories. *PhytoKeys*, 56:47.

Kearney, M. (2006). Habitat, environment and niche: what are we modelling? *Oikos*, 115(1):186–191.

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.-K., Yu, J., Damoulas, T., and Gomes, C. (2012). ebird: A human/computer learning network for biodiversity conservation and research. In *Twenty-Fourth IAAI Conference*.

Kendall, B. E., Bjørnstad, O. N., Bascompte, J., Keitt, T. H., and Fagan, W. F. (2000). Dispersal, environmental correlation, and spatial synchrony in population dynamics. *The American Naturalist*, 155(5):628–636.

Kennedy, T. A., Naeem, S., Howe, K. M., Knops, J. M., Tilman, D., and Reich, P. (2002). Biodiversity as a barrier to ecological invasion. *Nature*, 417(6889):636.

Kéry, M. and Gregg, K. B. (2003). Effects of life-state on detectability in a demographic study of the terrestrial orchid cleistes bifaria. *Journal of Ecology*, 91(2):265–273.

Kery, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Haefliger, G., and Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, 24(5):1388–1397.

Kew, R. B. G. (2016). The state of the world's plants report–2016. *Royal Botanic Gardens, Kew*.

Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J., Montoya, J. M., Römermann, C., Schiffers, K., Schurr, F. M., et al. (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, 39(12):2163–2178.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430.

Krähenbühl, P., Doersch, C., Donahue, J., and Darrell, T. (2015). Data-dependent initializations of convolutional neural networks. *arXiv preprint arXiv:1511.06856*.

Krishna, Nanda, K. P. K. R. M. P. A. C. J. S. (2019). Species recommendation using machine learning - geolifeclef 2019. *proceedings of CLEF 2019*.

Krishtalka, L. and Humphrey, P. S. (2000). Can natural history museums capture the future? *BioScience*, 50(7):611–617.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B. (2015). Template model builder tmb. *J. Stat. Softw*, 70:1–21.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Laird, R. A. and Schamp, B. S. (2006). Competitive intransitivity promotes species coexistence. *The American Naturalist*, 168(2):182–193.

Leathwick, J., Elith, J., and Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling*, 199(2):188–196.

Lecun, Y. (1985). Une procedure d'apprentissage pour reseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks). In *Proceedings of Cognitiva 85, Paris, France*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

LeCun, Y. et al. (1989). Generalization and network design strategies. *Connectionism in perspective*, pages 143–155.

Lee, S. H., Chan, C. S., Wilkin, P., and Remagnino, P. (2015). Deep-plant: Plant identification with convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 452–456. IEEE.

Leitão, P. J., Moreira, F., and Osborne, P. E. (2011). Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern portugal. *International Journal of Geographical Information Science*, 25(3):439–454.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., and Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological modelling*, 90(1):39–52.

Lek, S. and Guégan, J.-F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological modelling*, 120(2-3):65–73.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. (2017). Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*.

Lindgren, F., Rue, H., et al. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25.

Lindquist, E. J., D'Annunzio, R., Gerrand, A., MacDicken, K., Achard, F., Beuchle, R., Brink, A., Eva, H., Mayaux, P., San-Miguel-Ayanz, J., et al. (2013). *Changement d'utilisation des terres forestieres mondiales 1990 2005*. FAO/CCR.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G., and Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation biology*, 17(6):1591–1600.

Mac Aodha, O., Cole, E., and Perona, P. (2019). Presence-only geographical priors for fine-grained image classification. *arXiv preprint arXiv:1906.05272*.

MacKenzie, D. I. and Nichols, J. D. (2004). Occupancy as a surrogate for abundance estimation. *Animal biodiversity and conservation*, 27(1):461–467.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Maitner, B. S., Boyle, B., Casler, N., Condit, R., Donoghue, J., Durán, S. M., Guaderrama, D., Hinchliff, C. E., Jørgensen, P. M., Kraft, N. J., et al. (2018). The bien r package: A tool to access the botanical information and ecology network (bien) database. *Methods in Ecology and Evolution*, 9(2):373–379.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

McDonald, L., Manly, B., Huettmann, F., and Thogmartin, W. (2013). Location-only and use-availability data: analysis methods converge. *Journal of Animal Ecology*, 82(6):1120–1124.

Merow, C., Smith, M. J., and Silander Jr, J. A. (2013). A practical guide to maxent for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069.

Messina, J. P., Kraemer, M. U., Brady, O. J., Pigott, D. M., Shearer, F. M., Weiss, D. J., Golding, N., Ruktanonchai, C. W., Gething, P. W., Cohn, E., et al. (2016). Mapping global environmental suitability for zika virus. *Elife*, 5.

Midgley, G., Hannah, L., Millar, D., Thuiller, W., and Booth, A. (2003). Developing regional and species-level assessments of climate change impacts on biodiversity in the cape floristic region. *Biological Conservation*, 112(1-2):87–97.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10:22–37.

Mod, H. K., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. *Journal of Vegetation Science*, 27(6):1308–1322.

Monestiez, Pascal, B. C. (2019). Location-based species recommendation - geolifeclef 2019 challenge. *proceedings of CLEF 2019*.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932.

Moudhgalya, N. B., Sundar, Sharan Divi, S., Mirunalini, P., and Bose, A. C. (2018). Hierarchically embedded taxonomy with clnn to predict species based on spatial features. In *CLEF working notes 2018*.

Moudrỳ, V. and Šímová, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*, 26(11):2083–2095.

Moyes, C. L., Shearer, F. M., Huang, Z., Wiebe, A., Gibson, H. S., Nijman, V., Mohd-Azlan, J., Brodie, J. F., Malaivijitnond, S., Linkie, M., et al. (2016). Predicting the geographical distributions of the macaque hosts and mosquito vectors of plasmodium knowlesi malaria in forested and non-forested areas. *Parasites & vectors*, 9(1):242.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Negri, Mathilde, S. M. J. A. (2019). Plant prediction from cnn model trained with other kingdom species (geolifeclef 2019: Lirmm team). *proceedings of CLEF 2019*.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, 34(1):3–22.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401.

Norris, K. (2004). Managing threatened species: the ecological toolbox, evolutionary theory and declining-population paradigm. *Journal of Applied Ecology*, 41(3):413–426.

Novacek, M. J. (2008). Engaging the public in biodiversity issues. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11571–11578.

Nyegaard, T. (2019). Dof - observations from the danish ornithological society. danish ornithological society. Occurrence dataset https://doi.org/10.15468/tww7cj accessed via GBIF.org on 2019-07-29.

Oerke, E.-C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, 144(1):31–43.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9):2514–2521.

Ovaskainen, O. and Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92(2):289–295.

Paknia, O., Sh, H. R., and Koch, A. (2015). Lack of well-maintained natural history collections and taxonomists in megadiverse developing countries hampers global biodiversity exploration. *Organisms Diversity & Evolution, 15(3), 619-629*.

Parker, D. (1985). Learning-logic. *Report TR-47*.

Pearce, J. and Lindenmayer, D. (1998). Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (lichenostomus melanops cassidix) in southeastern australia. *Restoration ecology*, 6(3):238–243.

Pearce, J. L. and Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3):405–412.

Pearson, R. G., Dawson, T. P., Berry, P. M., and Harrison, P. (2002). Species: a spatial evaluation of climate impact on the envelope of species. *Ecological modelling*, 154(3):289–300.

Pellet, J. (2008). Seasonal variation in detectability of butterflies surveyed with pollard walks. *Journal of Insect Conservation*, 12(2):155–162.

Peterson, A. T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *The quarterly review of biology*, 78(4):419–433.

Peterson, A. T. and Soberón, J. (2012). Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação*, 10(2):102–107.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., and Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, 26(3):275–287.

Phillips, S. J. (2008). Transferability, sample selection bias and background data in presence-only modelling: a response to peterson et al.(2007). *Ecography*, 31(2):272–278.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., and Blair, M. E. (2017). Opening the black box: an open-source release of maxent. *Ecography*.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259.

Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.

Phillips, S. J., Dudík, M., and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406.

Proosdij, A. S., Sosef, M. S., Wieringa, J. J., and Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39(6):542–552.

Pukelsheim, F. (2006). *Optimal design of experiments*. Society for Industrial and Applied Mathematics.

Pulliam, H. R. (1988). Sources, sinks, and population regulation. *The American Naturalist*, 132(5):652–661.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology letters*, 3(4):349–361.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Dickstein, J. S. (2017). On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org.

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., and Maiorano, L. (2016). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*.

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., and Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of biogeography*, 33(10):1689–1703.

Ray, N. and Burgman, M. A. (2006). Subjective uncertainties in habitat suitability maps. *Ecological modelling*, 195(3-4):172–186.

Reddy, S. and Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in africa. *Journal of Biogeography*, 30(11):1719–1727.

Reich, B. J., Pacifici, K., and Stallings, J. W. (2018). Integrating auxiliary data in optimal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, 9(6):1626–1637.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

Renner, I. W. and Warton, D. I. (2013). Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281.

Richardson, D. M. and Pyšek, P. (2006). Plant invasions: merging the concepts of species invasiveness and community invasibility. *Progress in physical geography*, 30(3):409–431.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192.

Rotenberry, J. T., Preston, K. L., and Knick, S. T. (2006). Gis-based niche modeling for mapping species'habitat. *Ecology*, 87(6):1458–1464.

Royama, T. (2012). *Analytical population dynamics*, volume 10. Springer Science & Business Media.

Royle, J. A. (2004a). Generalized estimators of avian abundance from count survey data. *Animal Biodiversity and Conservation*, 27(1):375–386.

Royle, J. A. (2004b). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A., Dawson, D. K., and Bates, S. (2004). Modeling abundance effects in distance sampling. *Ecology*, 85(6):1591–1597.

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84(3):777–790.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747.

Ruete, A. (2015). Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodivers Data J*, page e5361.

Ruffray, P., B. H. G. r. G. H. M. (1989). "sophy", une banque de données phyto-sociologiques; son intérêt pour la conservation de la nature. *Actes du colloque "Plantes sauvages et menacées de France: bilan et protection", Brest, 8-10 octobre 1987*, pages 129–150.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Sauer, J. R., Pardieck, K. L., Ziolkowski Jr, D. J., Smith, A. C., Hudson, M.-A. R., Rodriguez, V., Berlanga, H., Niven, D. K., and Link, W. A. (2017). The first 50 years of the north american breeding bird survey. *The Condor: Ornithological Applications*, 119(3):576–593.

Schnoerr, D., Grima, R., and Sanguinetti, G. (2016). Cox process representation and inference for stochastic reaction–diffusion processes. *Nature communications*, 7:11729.

Schuster, R., Wilson, S., Rodewald, A. D., Arcese, P., Fink, D., Auer, T., and Bennett, J. R. (2019). Optimizing the conservation of migratory species over their full annual cycle. *Nature communications*, 10(1):1754.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Servajean, M., Joly, A., Shasha, D., Champ, J., and Pacitti, E. (2017). Crowdsourcing thousands of specialized labels: a bayesian active training approach. *IEEE Transactions on Multimedia*, 19(6):1376–1391.

Shah, M. and Coulson, S. (2019). Artportalen (swedish species observation system). version 92.150. artdatabanken. Occurrence dataset https://doi.org/10.15468/kllkyl accessed via GBIF.org on 2019-07-29.

Shen, G., Yu, M., Hu, X.-S., Mi, X., Ren, H., Sun, I.-F., and Ma, K. (2009). Species–area relationships explained by the joint effects of dispersal limitation and habitat heterogeneity. *Ecology*, 90(11):3033–3041.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Si-Moussi, Sara, G. E. H. M. D. T. T. W. (2019). Species recommendation using environment and biotic associations. *proceedings of CLEF 2019*.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24:467–471.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., and McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of ispot. *ZooKeys*, 480:125–146.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Slade, N. A., Alexander, H. M., and Dean Kettle, W. (2003). Estimation of population size and probabilities of survival and detection in mead's milkweed. *Ecology*, 84(3):791–797.

Soberón, J. and Nakamura, M. (2009). Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 106(Supplement 2):19644–19650.

Soberón, J. and Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):689–698.

Soultan, A. and Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PloS one*, 12(11):e0187906.

Stockwell, D. (1999). The garp modelling system: problems and solutions to automated spatial prediction. *International journal of geographical information science*, 13(2):143–158.

Stohlgren, T. J. and Schnase, J. L. (2006). Risk analysis for biological hazards: what we need to know about invasive species. *Risk Analysis: An International Journal*, 26(1):163–173.

Stolar, J. and Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, 21(5):595–608.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292.

Syfert, M. M., Smith, M. J., and Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of maxent species distribution models. *PloS one*, 8(2):e55158.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Taubert, S., Mauermann, M., Kahl, S., Kowerko, D., and Eibl, M. (2018). Species prediction based on environmental variables using machine learning techniques. In *CLEF working notes 2018*.

Teacher, A. G., Griffiths, D. J., Hodgson, D. J., and Inger, R. (2013). Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecology and Evolution*, 3(16):5268–5278.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F., De Siqueira, M. F., Grainger, A., Hannah, L., et al. (2004). Extinction risk from climate change. *Nature*, 427(6970):145.

Thuiller, W. (2003). Biomod–optimizing predictions of species distributions and projecting potential future shifts under global change. *Global change biology*, 9(10):1353–1362.

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12):2020–2027.

Thuiller, W., Albert, C., Araujo, M. B., Berry, P. M., Cabeza, M., Guisan, A., Hickler, T., Midgley, G. F., Paterson, J., Schurr, F. M., et al. (2008). Predicting global change impacts on plant species' distributions: future challenges. *Perspectives in plant ecology, evolution and systematics*, 9(3-4):137–152.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., and Prentice, I. C. (2005). Climate change threats to plant diversity in europe. *Proceedings of the National Academy of Sciences*, 102(23):8245–8250.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tilman, D. (1982). *Resource competition and community structure*. Princeton university press.

Tscharntke, T., Clough, Y., Wanger, T. C., Jackson, L., Motzke, I., Perfecto, I., Vandermeer, J., and Whitbread, A. (2012). Global food security, biodiversity conservation and the future of agricultural intensification. *Biological conservation*, 151(1):53–59.

Tsybakov, A. (2009). Introduction to nonparametric estimation. In *Springer Series in Statistics, ISBN 978-0-387-79051-0*. Springer-Verlag New York.

Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11):1084–1091.

Violle, C., Choler, P., Borgy, B., Garnier, E., Amiaud, B., Debarros, G., Diquelou, S., Gachet, S., Jolivet, C., Kattge, J., et al. (2015). Vegetation ecology meets ecosystem science: permanent grasslands as a functional biogeography case study. *Science of the Total Environment*, 534:43–51.

Vitousek, P., D'Antonio, C., Loope, L., and Westbrooks, R. (1996). Biological invasions as global environmental change. *American scientist (USA)*.

Waagepetersen, R., Guan, Y., Jalilian, A., and Mateu, J. (2016). Analysis of multispecies point patterns by using multivariate log-gaussian cox processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):77–96.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, 63(1):252–258.

Wäldchen, J., Rzanny, M., Seeland, M., and Mäder, P. (2018). Automated plant species identification—trends and future directions. *PLoS computational biology*, 14(4):e1005993.

Walker, P. and Cocks, K. (1991). Habitat: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*, pages 108–118.

Wallace, A. R. (1860). On the zoological geography of the malay archipelago. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 4(16):172–184.

Ward, G. (2007). *Statistics in ecological modeling; presence-only data and boosted mars*. stanford.edu.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. (2009). Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563.

Warton, D. I., Renner, I. W., and Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168.

Warton, D. I., Shepherd, L. C., et al. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402.

Weber, E. and Gut, D. (2004). Assessing the risk of potentially invasive plant species in central europe. *Journal for Nature Conservation*, 12(3):171–179.

Werbos, P. (1974). Beyond regression:" new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*.

Whittaker, R. H. (1960). Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338.

Whittaker, R. H. (1967). Gradient analysis of vegetation. *Biological reviews*, 42(2):207–264.

Whittaker, R. H., Levin, S. A., and Root, R. B. (1973). Niche, habitat, and ecotope. *The American Naturalist*, 107(955):321–338.

Whittaker, R. H. and Niering, W. A. (1975). Vegetation of the santa catalina mountains, arizona. v. biomass, production, and diversity along the elevation gradient. *Ecology*, 56(4):771–790.

Wilf, P., Zhang, S., Chikkerur, S., Little, S. A., Wing, S. L., and Serre, T. (2016). Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences*, 113(12):3305–3310.

Wintle, B. A., Kavanagh, R. P., McCARTHY, M. A., and Burgman, M. A. (2005). Estimating and dealing with detectability in occupancy surveys for forest owls and arboreal marsupials. *The Journal of Wildlife Management*, 69(3):905–917.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C., Guisan, A., and Group, N. P. S. D. W. (2008). Effects of sample size on the performance of species distribution models. *Diversity and distributions*, 14(5):763–773.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE.

Yanikoglu, B., Aptoula, E., and Tirkaz, C. (2014). Automatic plant identification from photographs. *Machine vision and applications*, 25(6):1369–1383.

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of vegetation science*, 2(5):587–602.

Yoccoz, N. G., Nichols, J. D., and Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in ecology & evolution*, 16(8):446–453.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zarnetske, P. L., Edwards, T. C., and Moisen, G. G. (2007). Habitat classification modeling with incomplete data: pushing the habitat envelope. *Ecological Applications*, 17(6):1714–1726.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., et al. (2013). On rectified linear units for speech processing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3517–3521. IEEE.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

11 Appendices

11.1 Appendix of chapter 2

S1 Appendix. Texts and mathematical proofs of Manuscript: "Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection"

Christophe Botella * [1,2,3,5], Alexis Joly[1], Pascal Monestiez[5], Pierre Bonnet[3,4], and François Munoz[6]

[1]INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.

[2]INRA, UMR AMAP, F-34398 Montpellier, France.

[3]AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France.

[4]CIRAD, UMR AMAP, F-34398 Montpellier, France.

[5]BioSP, INRA, Site Agroparc, 84914 Avignon, France.

[6]Université Grenoble Alpes, Laboratoire d'Ecologie Alpine, CS 40700, 38058 Grenoble cedex 9621, France.

*christophe.botella@cirad.fr

1 Text A: Poisson process induced on the environmental domain and factorization of its intensity.

In this part we show how the Poisson process modeling the distribution of observed points in the geographic space, explained in section 2.3 of the manuscript, induce a Poisson point process into the environmental space $Im(x)$ whose intensity function, namely the expected points count per unit of space for a given environment, factorizes to the product of the species intensity function λ and the observation intensity named $\bar{s}$, both defined over $Im(x)$. We justify the interest of looking at bias in the environmental space. Besides, we justify several important hypothesis made in section **2**, such as the almost everywhere continuity of x over D, the almost everywhere continuity of λ over $\mathbb{R}$ and the assumption that it is bounded on any bounded subset of $\mathbb{R}$.

Poisson process induced in the environmental domain. We show hereafter that Z_r follows a general Poisson process [Chiu et al., 2013, Haenggi, 2013] of intensity measure $\Lambda : \mathscr{L}(\mathbb{R}) \to \mathbb{R}^+, W \to \int_{x^{-1}(W)} s\lambda \circ x d\mu$, i.e. (i) for any $W \in \mathscr{L}(\mathbb{R})$, $|Z_r \cup x^{-1}(W)| \sim \mathscr{P}(\Lambda(W))$, and (ii) $\forall W_1, W_2 \in \mathscr{L}(\mathbb{R})$ such that $W_1 \cap W_2 = \emptyset$, $|Z_r \cup x^{-1}(W_1)|$ and $|Z_r \cup x^{-1}(W_2)|$ are independent random variables.

First, (i) is straightforward. Let $W \in \mathscr{L}(\mathbb{R})$, then by definition $|Z_r \cup x^{-1}(W)| \sim \mathscr{P}(\int_{x^{-1}(W)} s\lambda \circ x d\mu)$ because Z_r follows a Poisson process over D of intensity measure $s\lambda \circ x$, which is indeed a measure over $\mathscr{L}(\mathbb{R})$ because it is positive by definition, and it is finite because λ is bounded on any bounded subset of $\mathbb{R}$ and $s \in [0,1]$ by definition.

Secondly, (ii) is also straightforward. Let $W_1, W_2 \in \mathscr{L}(\mathbb{R})$ such that $W_1 \cap W_2 = \emptyset$, then $x^{-1}(W_1) \cap x^{-1}(W_2) = \emptyset$ (no spatial point has two different values of x), then $\forall n_1, n_2 \in \mathbb{N}^2, p(|Z_r \cup x^{-1}(W_1)| = n_1, |Z_r \cup x^{-1}(W_2)| = n_2) = p(|Z_r \cup x^{-1}(W_1)| = n_1) p(|Z_r \cup x^{-1}(W_2)| = n_2)$ because Z_r follows a Poisson process over D.

Remark: The Poisson process in the environmental space is equivalent to the one in D if and only x achieves a bijection, or a one-to-one correspondance, between D and the environmental space, which is not the case here as $\mathbb{R}$ is only one dimensional.

Intensity in the environmental domain. We now show that the intensity measure Λ can also be written, for any $W \in \mathscr{L}(\mathbb{R}), \Lambda(W) = \int_W \lambda \bar{s} d\mu_x$ where $\lambda \bar{s}$ is the intensity function of the induced Poisson process over the environmental space $\mathbb{R}$ relatively to the measure μ_x (which is null outside of $Im(x)$) and $\bar{s}$ is defined by:

$$\forall w \in \mathbb{R}, \bar{s}(w) = \begin{cases} \lim\limits_{\delta \to 0} \frac{\int_{x^{-1}([w-\frac{\delta}{2}, w+\frac{\delta}{2}])} s d\mu}{\mu_x([w-\frac{\delta}{2}, w+\frac{\delta}{2}]))} & \text{if } w \in Im(x) \\ 0 & \text{otherwise, by convention.} \end{cases} \tag{1}$$

Firstly, we show the case where $w \in \mathbb{R} \backslash Im(x)$. We have that $\Lambda(\mathbb{R}) = \Lambda(Im(x))$. This is because $\mu(x^{-1}(\mathbb{R} \backslash Im(x))) = \mu(\{z \in D, x(z) \notin Im(x)\}) = \mu(\{z \in D, x \text{ not continuous at } z\}) = 0$ because x is continuous almost everywhere on D. It implies that any Lebesgue integral computed relatively to μ (Lebesgue measure on $\mathbb{R}^2$) over $x^{-1}(\mathbb{R} \backslash Im(x))$ also equals 0. Thus we could define any value for $\bar{s}$ outside of $Im(x)$, we set it to 0 by convention (which means no observation intensity outside the geographic domain under study).

It remains to show that the writing of $\bar{s}$ is legitimate on $Im(x)$. $Im(x)$, as any subset of $\mathbb{R}$, is a union of intervals and singletons. However, the singletons of $Im(x)$ have an important particularity, they are all atoms of μ_x. More precisely, any singleton w in the connected components of $Im(x)$ is necessarily an atom for the measure μ_x, i.e. $\mu_x(w) > 0$. Indeed, there exists an open subset of the geographic domain $O \subset D$ where x is continuous and reaches the value w somewhere in O. Then, $x(O)$ is an element of an interval of $Im(x)$ that contains w, but as w is not included in any continuous interval of $Im(x)$, this interval is necessarily the singleton $\{w\}$, which implies that $\forall z \in 0, x(z) = w$. Consequently, $\mu_x(w) = \mu(x^{-1}(w)) \geq \mu(O) > 0$ because O is an open subset of $\mathbb{R}^2$ and by definition of the Lebesgue measure on $\mathbb{R}^2$. We have shown that if $w \in Im(x)$ is a singleton of $Im(x)$, it is an atom for μ_x. We can then write $\Lambda(w) = \int_{x^{-1}(w)} s\lambda \circ x\mu = \lambda(w) \int_{x^{-1}(w)} s\mu = \lambda(w) \frac{\int_{x^{-1}(w)} sd\mu}{\mu_x(w)} \mu_x(w)$ $= \lambda(w) \left[\lim_{\delta \to 0} \frac{\int_{x^{-1}([w-\frac{\delta}{2}, w+\frac{\delta}{2}])} sd\mu}{\mu_x([w-\frac{\delta}{2}, w+\frac{\delta}{2}]))} \right] \mu_x(w)$. Thus, the definition of $\bar{s}$ in equation 1 holds for singletons of $Im(x)$.

It remains to show that 1 also holds for any non-singleton interval $W \subset Im(x)$. Let $W \subset Im(x)$ be a non-singleton interval. We define the sequence $(C_j := \{C_j^1, ..., C_j^j\})_{j \in \mathbb{N}^*}$ of finite partitions of $[\inf W, \sup W[$. We define it with $\forall j \geq 1, i \leq j$, $C_j^i = [\inf W + (i-1)(\sup W - \inf W)/j, \inf W + i(\sup W - \inf W)/j[$. Then, we note $I_j(W) := \sum_{i=1}^{j} \int_{x^{-1}(C_j^i)} s\lambda \circ x d\mu$ where we can see that $\forall j, I_j(W) = \int_{x^{-1}(W)} s\lambda \circ x d\mu$. Besides,

$$
\begin{aligned}
\lim_{j \to \infty} I_j(W) &= \lim_{j \to \infty} \sum_{i=1}^{j} \lambda(\inf W + i(\sup W - \inf W)/j) \int_{x^{-1}(C_j^i)} sd\mu && (\lambda \text{continuous almost everywhere}) \\
&= \lim_{j \to \infty} \sum_{i=1}^{j} \lambda(\inf W + i(\sup W - \inf W)/j) \frac{\int_{x^{-1}(C_j^i)} sd\mu}{\mu_x(C_j^i)} \mu_x(C_j^i) && (x \text{ continuous on } x^{-1}(C_j^i) \Rightarrow \mu_x(C_j^i) > 0) \\
&= \int_W \lambda \bar{s} d\mu_x
\end{aligned}
$$

Where $\forall w \in W$:

$$
\bar{s}(w) = \lim_{\delta \to 0} \frac{\int_{x^{-1}([w-\frac{\delta}{2}, w+\frac{\delta}{2}])} sd\mu}{\mu_x([w-\frac{\delta}{2}, w+\frac{\delta}{2}]))}
$$

Finally, we have shown that for $\forall W \in \mathscr{L}(\mathbb{R}), \Lambda(W) = \int_W \lambda \bar{s} d\mu_x$ where $\lambda \bar{s}$ is the intensity function of the induced Poisson process over the environmental space. We see that this intensity, representing the expected the number of points per unit of space corresponding to a given environment

value, factorizes into a species intensity and $\bar{s}$ that we call observation intensity which depends on the sampling effort s and the environmental variable x, as defined in equation 1. Roughly speaking, $\bar{s}(w)$ is the average of the sampling effort function s over the limit subspace $x^{-1}(w) \subset D$.

Why do we analyse bias in the environmental domain. If s is heterogeneous in space, we may encounter a bias when estimating θ_0 from Z_r, but there is no direct link between the spatial form of s and the bias. Indeed, our target f is a function of x values. So even if s is distributed heterogeneously in D, its variations could cancel in $Im(x)$ and entail no difference on the density of species observed points on $Im(x)$ compared to a uniform sampling on D. That is why it is more relevant to look at the distribution of s over $Im(x)$.

Environmental variable continuity assumption The assumption of almost everywhere continuity of x over D, which means that $\mu(\{z \in D/x \text{ is discontinuous at } z\}) = 0$, is necessary to ensure that $\bar{s}$ and s_x are well defined on $Im(x)$. Let's recall that $Im(x) = \{w \in \mathbb{R}/\exists z \in D/x(z) = w \text{ and } x \text{ is continuous at } z\}$ is the set of values for which there exist fibers of x in D at which x is continuous. The almost everywhere continuity allows discontinuities of x over negligible areas of D, basically points and lines, which is useful because it allows x is a rasterized environmental variable, a continuously varying variable, or even a mixture of both.

Species intensity continuity assumption λ which is continuous almost everywhere over $Im(x)$. This hypothesis is useful to allow this function to be not continuous on certain points. For instance, Maxent [Phillips and Dudík, 2008] uses threshold functions in its model. Besides, this hypothesis doesn't seem limiting, because it is hard to imagine a species density function that would have discontinuity points over an infinite and non-countable number of points, even if such function can be theoretically built.

2 Text B: Modeling the species niche with a gaussian density

Here we describe our choice of gaussian density for f in simulation. Of course, we cannot cover the huge variety of niche models, so we chose to illustrate classic ecological types. We assume that the realized niche of a species corresponds to its fundamental niche, in the sense of Hutchinson [1957]. The expected species abundance only depends on the suitability of environment described by x. Even if the spatial variation of the abiotic environment is known to be a strong determinant of species distribution, it is not the only factor affecting it, there is also the spatial dispersal constraints and the interactions with other organisms (Pulliam [2000],Soberón [2007]). Species distribution along environmental gradients are often thought to be unimodal and tappered, and the more precise choice of modeling the species density as a gaussian function along environmental gradient is quite

comon in ecology (Franklin [2010]). The maximum of f is called the **optimum**, and the inverse of its variance, its **specialization**. Indeed, those quantities are of main interest for ecological applications, and it is crucial to study their biases. Chosing the gaussian density for f can be interpreted as setting the constraints that the expected x of a given species individual is μ_0 ($\int_{\mathbb{R}} f(w)wdw = \mu_0$, optimum constraint), the variance of x over many individuals is σ_0^2 ($\int_{\mathbb{R}} f(w)(w-\mu_0)^2 dw = \sigma_0^2$, specialization constraint), and f is of maximum entropy.

3 Text C: Fitting the UB model to data

We here present the details of the UB fitting method, as described in Berman and Turner [1992] and Renner et al. [2015]. The UB method is fitted by maximizing the log-likelihood of the Poisson point process model of intensity λ_θ, defined over the domain D, with observed species occurrences $Z = \{z_1, ..., z_n\}$, with respect to the model parameters $\theta := (\alpha, \beta_1, \beta_2)$:

$$\begin{aligned}
\mathscr{L}(z_1, ..., z_n|\theta) &= \log(p(z_1, ..., z_n|\theta)) \\
&= \log\left(e^{-\int_D \lambda_\theta \circ x d\mu} \prod_{i=1}^{n} \lambda_\theta(x(z_i))\right) \\
&= \sum_{i=1}^{n} \log(\lambda_\theta(x(z))) - \int_D \lambda_\theta \circ x d\mu
\end{aligned}$$

In general, the integral term cannot be computed exactly. We rather use a numerical approximation. The integral is replaced by a weighted sum of λ_θ computed at some background/quadrature points, $Z^q = \{z_1^q, ..., z_Q^q\}$ where Q is the number of background points. In MAXENT literature, quadrature points are often called pseudo-absences. Berman and Turner [1992] re-express the likelihood by including $z_1, ..., z_n$ among background points, and defining samples weights. It gives a classic Poisson regression likelihood:

$$\begin{aligned}
\mathscr{L}(z_1, ..., z_n|\theta) &\approx \sum_{j=1}^{Q} 1_{z \in Z} \log(\lambda_\theta(x(z_j^q))) - w_j \lambda_\theta(x(z_j^q)) \\
&= \sum_{j=1}^{Q} w_j \left(y_j \log(\lambda_\theta(x(z_j^q))) - \lambda_\theta(x(z_j^q))\right)
\end{aligned}$$

Where the y_j correspond to the Poisson regression counts (called pseudo-counts because they can be non integers), and w_j the samples weights. We define **the background points** $Z^q \backslash Z$ and their weights so that $\sum_{i=1}^{n} w_i \lambda_\theta(x(z_i^q)) \approx \int_D \lambda_\theta \circ x d\mu$. A unbiased and popular manner to approximate the integral is the Monte Carlo method, which uses the average over uniformly sampled points on D to approximate the integral. However, we must prevent $z_1, ..., z_n$ from biasing our approximation, because they are not uniformly distributed in D. We give them a total weight in the sum that is negligible compared to the background points drawn uniformly :

$$\forall j \in [|1, Q|],\ w_j \begin{cases} = \frac{\mu(D)}{100n} & \text{if } z_j^q \in Z \\ = \frac{99\mu(D)}{100(Q-n)} & \text{otherwise} \end{cases}$$

With this setting, all weights sum to $\mu(D)$, while weights of species reported points alone represent only 1% of this value. This way, the approximation of $\int_D \lambda_\theta \circ x d\mu$ with background

points is not affected by reported points. Now, the standard formulation of the Poisson parameter in the Poisson Generalized Linear Model is slighlty different from our model. It doesn't use our parametrization of λ_θ with the gaussian distribution parameters $\theta = (K, \mu, \sigma^2)$, but another equivalent parametrization. We note this equivalent function λ'_γ, called the log-linear predictor, for any z like this : $\lambda_\theta(x(z)) = \lambda'_\gamma(x(z)) = \exp\left(\alpha + \beta_1 x(z) + \beta_2 x(z)^2\right)$ where $\gamma = (\alpha, \beta_1, \beta_2)$ are the parameters of the log-linear predictor that are returned by standard Generalized Linear Model softwares. We can now easily recover our parameters of interest μ and σ by identification:

$$\begin{aligned} \forall z \in D,\ \lambda_\theta(x(z)) &= \exp\left(K - \tfrac{(x(z)-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\alpha + \beta_1 x(z) + \beta_2 x(z)^2\right) \quad \text{with} \begin{cases} \beta_1 = \frac{\mu_0}{\sigma_0^2} \\ \beta_2 = \frac{-1}{2\sigma_0^2} \end{cases} \Leftrightarrow \begin{cases} \mu_0 = \frac{\beta_1}{2\beta_2} \\ \sigma_0 = \frac{1}{\sqrt{-2\beta_2}} \end{cases} \end{aligned}$$

Where β_2 is strictly negative. We can now compute the Generalized Linear Model (with R package `glm`) to estimate parameter values β_1, β_2 that maximize the likelihood, for given y_js, $x(z_j^q)$s and w_js.

4 Text D: Proof of asymptotic UB estimate (Equation 2)

This part proves equation **2** (section 4.1 in manuscript) which expresses the expected UB estimate as the minimzer of a divergence to the observed species density fs_x. We are interested in the asymptotical estimate of the environmental density of the UB method given that the observed points follow the Poisson process: $IPP(s\lambda \circ x)$. Our target is the intensity function $\lambda(x(.)))$ but we can only estimate it, at best, up to a constant factor as it is multiplied by s, of unknown global scale, in the generating process as already shown in Fithian and Hastie [2013] and Hastie and Fithian [2013]. We may still estimate the relative intensity function by maximizing the joint likelihood of points position, conditional to the number of points generated by the process. For a finite sample $z_1, ..., z_n \in D$ of point realizations of the process, it is written:

$$p(z_1, ..., z_n | n, \theta) = \prod_{i=1}^{n} \frac{\lambda_\theta(x(z_i))}{\int_D \lambda_\theta \circ x d\mu}$$

Thus, the maximum likelihood parameter estimate of the intensity function is

$$\hat{\theta}_{UB} = \underset{\theta}{\operatorname{argmax}}\, P(z_1, ..., z_n | n, \theta) = \underset{\theta}{\operatorname{argmin}}\ -\frac{1}{n}\log(P(z_1, ..., z_n | n, \theta))$$

We recall that $\bar{s}$, λ and λ_θ are continuous μ_x-almost everywhere. Then, the limit of the above averaged negative Log likelihood when $n \to +\infty$ can be rewritten as follows:

$$
\begin{aligned}
& \lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\lambda_\theta(x(z_i))}{\int_D \lambda_\theta\circ x d\mu}\right) \\
= \;& \mathbb{E}\left(-\log\left(\frac{\lambda_\theta(x(z_1))}{\int_D \lambda_\theta\circ x d\mu}\right)\right) \\
= \;& -\int_D \frac{s\lambda\circ x}{\int_D s\lambda\circ x d\mu}\log\left(\frac{\lambda_\theta\circ x}{\int_D \lambda_\theta\circ x d\mu}\right)d\mu \\
= \;& -\lim_{\delta\to 0^+}\sum_{k=0}^{N(\delta)}\int_D 1_{z\in x^{-1}([a_k,a_k+\delta])}\frac{s(z)\lambda(x(z))}{\int_D s\lambda\circ x d\mu}\log\left(\frac{\lambda_\theta(x(z))}{\int_D \lambda_\theta\circ x d\mu}\right)dz \\
= \;& -\lim_{\delta\to 0^+}\sum_{k=0}^{N(\delta)}\frac{\lambda(a_k)}{\int_D s\lambda\circ x d\mu}\log\left(\frac{\lambda_\theta(a_k)}{\int_D \lambda_\theta\circ x d\mu}\right)\int\limits_{x^{-1}([a_k,a_k+\delta])} s d\mu \qquad (x(z)\to a_k \text{ and} \\
& \qquad \lambda,\lambda_\theta \text{ continuous} \\
& \qquad \mu_x\text{-almost everywhere}) \\
= \;& -\lim_{\delta\to 0^+}\sum_{k=0}^{N(\delta)}\left[\frac{\lambda(a_k)}{\int_D s\lambda\circ x d\mu}\log\left(\frac{\lambda_\theta(a_k)}{\int_D \lambda_\theta\circ x d\mu}\right)\right. \\
& \left.\frac{\int_{x^{-1}([a_k,a_k+\delta])} s d\mu}{\mu_x([a_k,a_k+\delta])}\mu_x([a_k,a_k+\delta])\right] \\
= \;& -\int_{\mathbb{R}}\frac{\bar{s}\lambda}{\int_D s\lambda\circ x d\mu}\log\left(\frac{\lambda_\theta}{\int_D \lambda_\theta\circ x d\mu}\right)d\mu_x \\
\alpha \;& -\int_{\mathbb{R}}\frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\log\left(\frac{\lambda_\theta}{\int_D \lambda_\theta\circ x d\mu}\right)d\mu_x \qquad (\text{factor } >0 \text{ and} \\
& \qquad \text{independent of } \theta) \\
= \;& -\int_{\mathbb{R}}\frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\log\left(\frac{\lambda_\theta}{\int_{Im(x)} \lambda_\theta d\mu_x}\right)d\mu_x \qquad (\text{same method}) \\
= \;& \int_{\mathbb{R}}\frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\left[\log\left(\frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\right)-\log(f_\theta)\right]d\mu_x \qquad (\text{adding term} \\
& \qquad \text{independent of } \theta) \\
= \;& \mathscr{D}_{KL}^{\mu_x}(f s_x||f_\theta)
\end{aligned}
$$

Where $a_k = inf(Im(x)) + k\delta$ and $N(\delta)$ is the quotient of the euclidean division of $|Im(x)|$ by δ. This way, we have $[a_0, a_{N(\delta)}] \subset Im(x) \subset [a_0, a_{N(\delta)}+\delta]$.

5 Text E: μ_x weighted KL-Divergence

The divergence is weighted by μ_x, the measure of the spatial area associated with any x value. It means that on parts of $Im(x)$ where $\mu_x = 0$ (*i.e.* environment not in D or of negligible area), $f_{\hat{\theta}_{UB}}$ is unconstrained, so it is allowed to take any shape, and it will depend on estimated parameters. As a consequence, the prediction of species intensity outside the enviromental range covered in D will be highly influenced by a misspecification of distribution model. This remark is also true for the following methods. The μ_x weighting will also lead approximation error compromises when $f_{\hat{\theta}_{UB}}$

cannot fit exactly to $f \circ s_x$. For example, if the parametrization of f_θ doesn't allow it to fit well to $s_x f$ over both subset $W_1, W_2 \subset Im(x)$ with $W_1 \cap W_2 = \emptyset, |W_1| = |W_2|$, and $\mu_x(W_1) > \mu_x(W_2)$, then the estimate should fit better on W_1 than on W_2.

6 Text F: A sample from the sampling effort proportional density as background.

In this part we demonstrate the optimality of the theoritical method consisting of using an large sample of background points directly drawn independantly from the sampling effort proportional density over D, which is the method **ApproxFactorBiasOut** introduced in Dudík et al. [2006].

We now assume that we have a sample $z_1^s, ..., z_{n_0}^s$ from $s/\int_D s(z)dz$, *i.e.* points distributed according to the proportional density of the sampling effort. We use these points as equally weighted background points in the Poisson process likelihood. We re-express the asymptotic estimator associated with this procedure. Like previously, we write the limit of the averaged negative log-likelihood, with now both n and n_0 tend to infinity:

$$
\begin{aligned}
& \lim_{\substack{n\to\infty \\ n_0\to\infty}} -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\lambda_\theta(x(z_i))}{\sum_{j=1}^{n_0}\frac{\mu(D)}{n_0}\lambda_\theta(x(z_j^s))}\right) \\
= \; & \lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\lambda_\theta(x(z_i))}{\int_D \frac{s}{\int_D s d\mu}\lambda_\theta \circ x \, d\mu}\right) \\
= \; & -\int_D \frac{s\lambda\circ x}{\int_D s\,\lambda\circ x\, d\mu} \log\left(\frac{\lambda_\theta\circ x}{\int_D \frac{s}{\int_D s d\mu}\lambda_\theta\circ x\, d\mu}\right) d\mu \\
= \; & -\int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x} \log\left(\frac{\lambda_\theta(w)}{\int_D \frac{s}{\int_D s d\mu}\lambda_\theta\circ x\, d\mu}\right) d\mu_x \qquad \text{(previous method, factor independent of } \theta) \\
= \; & -\int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x} \log\left(\frac{\lambda_\theta}{\int_{\mathbb{R}} f_\theta \frac{\bar{s}}{\int_D s d\mu} d\mu_x}\right) d\mu_x \qquad \text{(previous method)} \\
= \; & \int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x} \log\left(\frac{f s_x}{f_\theta s_x}\right) d\mu_x \qquad \text{(adding constant term and neg-entropy of } f s_x) \\
= \; & \mathscr{D}_{KL}^{\mu_x}(f s_x || f_\theta s_x)
\end{aligned}
$$

$$\mathbb{E}(\hat{\theta}_{AFBO}) = \underset{\theta}{\text{Argmin}}\; \mathscr{D}_{KL}^{\mu_x}(f s_x || f_\theta s_x)$$

Thus, $f_{\hat{\theta}_{AFBO}}$ will converge to f, except on parts where $s_x \mu_x = 0$ and the method gives an unbiased estimate of the species niche.

7 Text G: Proof of asymptotic TGOB estimate (Equation 5)

This part proves equation **5** (section 4.7 in manuscript) which expresses the expected TGOB estimate as the minimzer of a divergence from $f_\theta a$ to f, which means it fits to f/a. For this part we assume, on top of previous conditions (Riemann integrability of λ and $\bar{s}$), that a is Riemann inte-

grable on $\mathbb{R}$. We recall that the ensemble of observed points of species from the Target-Group is noted Z^{tg}. On the same principle than previously, we re-express the limit of the averaged negative log likelihood when the background points are drawn according to the TG species density:

$$
\begin{aligned}
& \lim_{\substack{n\to\infty \\ |Z^{tg}|\to\infty}} -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\lambda_\theta(x(z_i))}{\sum_{z\in Z^{tg}}\frac{\mu(D)}{|Z^{tg}|}\lambda_\theta(x(z))}\right) \\
= \; & \lim_{n\to\infty} -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log\left(\frac{\lambda_\theta(x(z_i))}{\int_D s\, a\circ x\, \lambda_\theta\circ x\, d\mu}\right) && \text{(TG points drawn from density } s\, a\circ x) \\
= \; & -\textstyle\int_D s\, \lambda\circ x\, \log\left(\frac{\lambda_\theta\circ x}{\int_D s\, a\circ x\, \lambda_\theta\circ x\, d\mu}\right) d\mu && \text{(species points drawn from density } s\, \lambda\circ x) \\
\alpha \; & -\textstyle\int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\log\left(\frac{\lambda_\theta}{\int_D s\, a\circ x\, \lambda_\theta\circ x\, d\mu}\right) d\mu_x && \text{(previous method, factor independent of } \theta) \\
= \; & -\textstyle\int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\log\left(\frac{\lambda_\theta}{\int_{\mathbb{R}} \bar{s}a\lambda_\theta d\mu_x}\right) d\mu_x && \text{(previous method)} \\
= \; & -\textstyle\int_{\mathbb{R}} \frac{s_x f}{\int_{\mathbb{R}} s_x f d\mu_x}\log\left(f_\theta s_x a\right) d\mu_x && \text{(adding term independent of } \theta) \\
= \; & D_{KL}^{\mu_x}\left(f s_x || f_\theta s_x a\right) && \text{(substrating entropy of } \lambda\bar{s}p_x)
\end{aligned}
$$

References

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Dudík, M., Phillips, S. J., and Schapire, R. E. (2006). Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917.

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Haenggi, M. (2013). Stochastic geometry for wireless networks, cambridge uni.

Hastie, T. and Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867.

Hutchinson, G. E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding remarks*, 22:415–427.

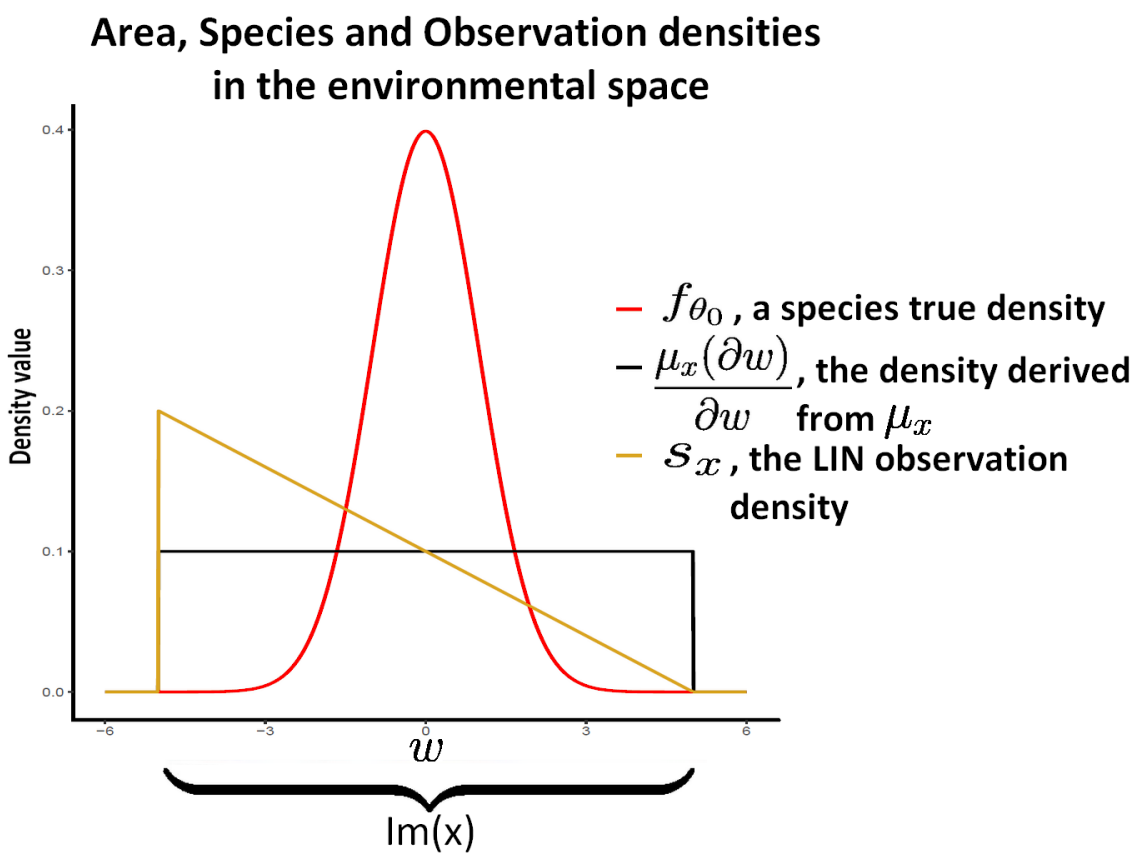
Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology letters*, 3(4):349–361.

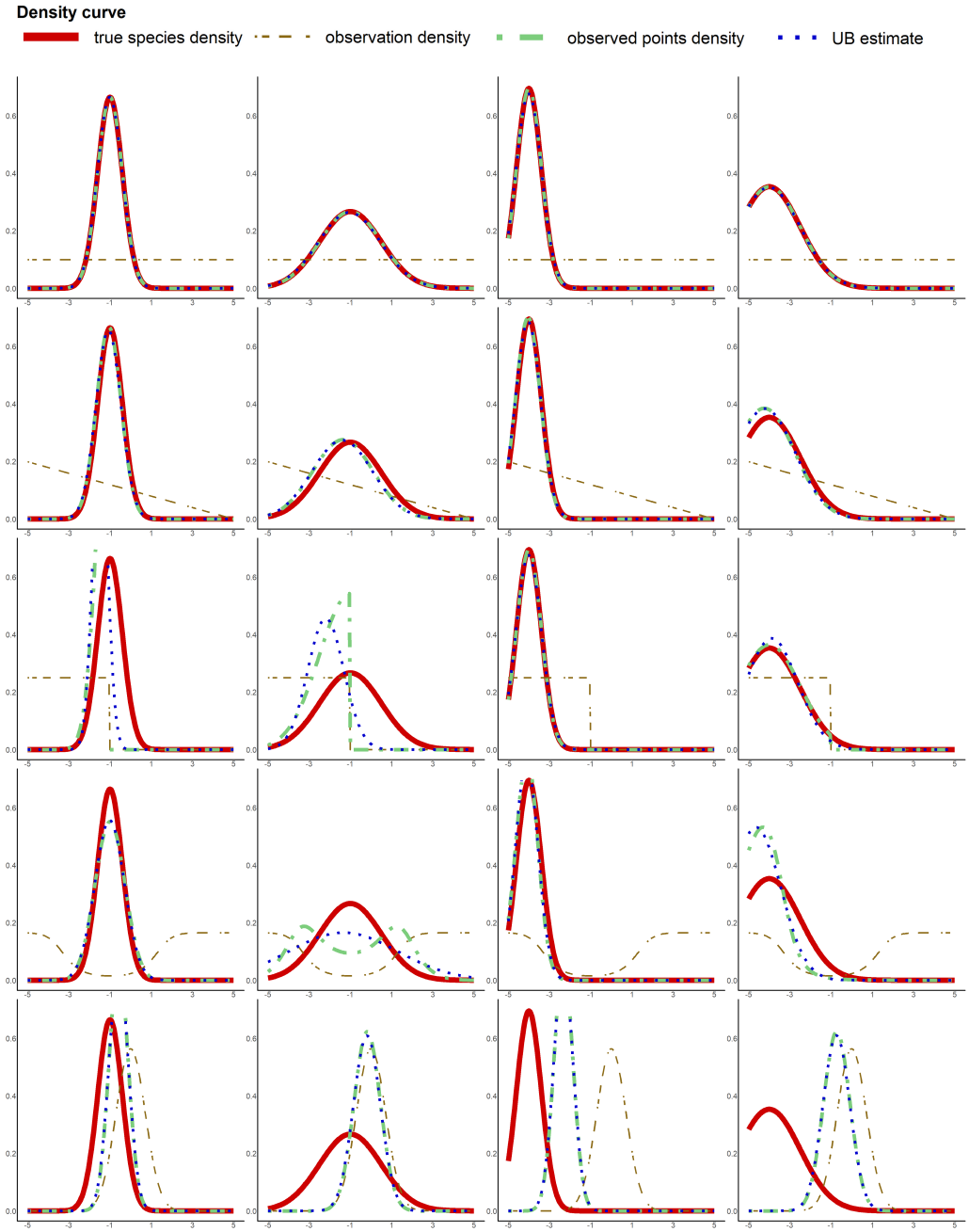
Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and

Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

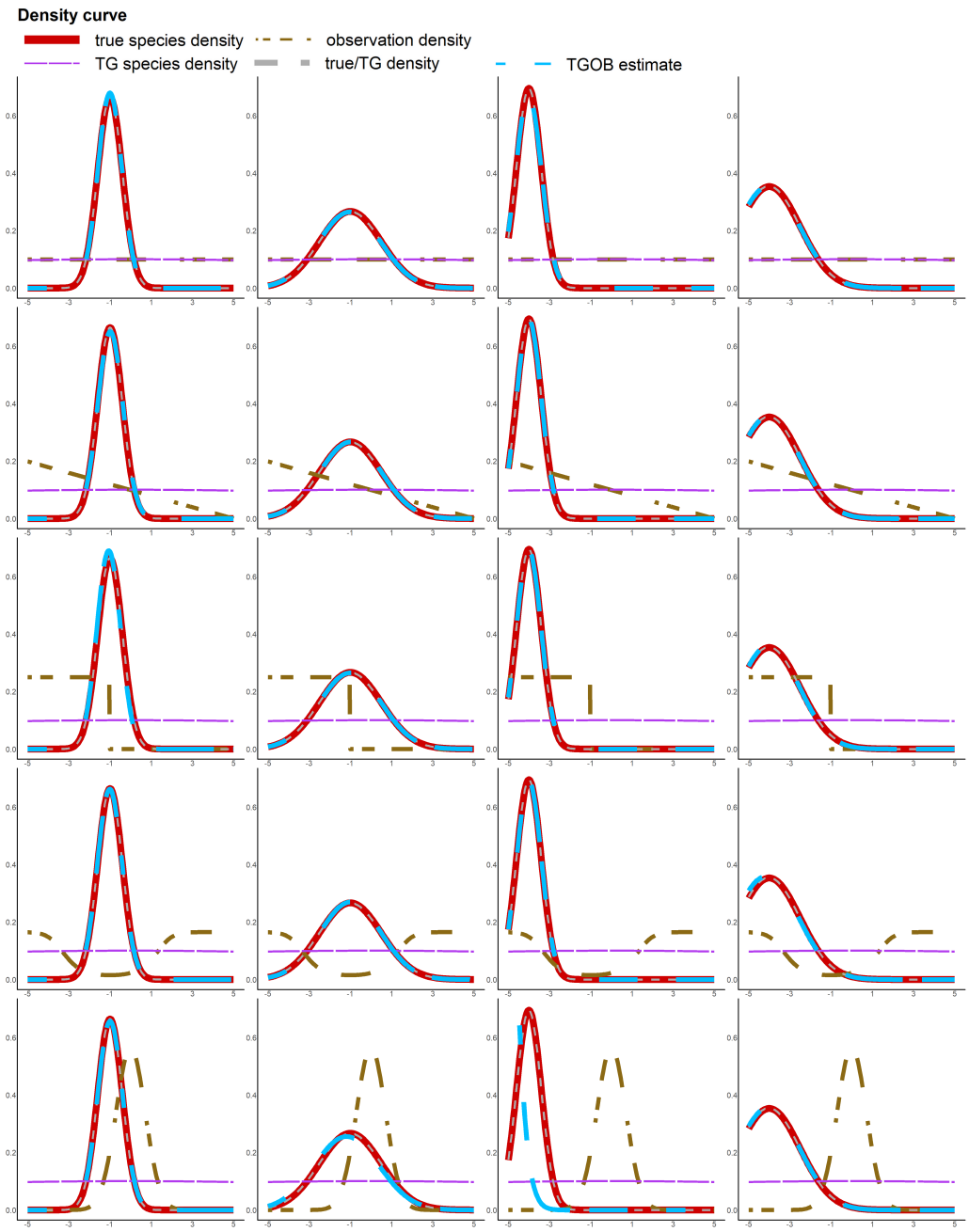
Soberón, J. (2007). Grinnellian and eltonian niches and geographic distributions of species. *Ecology letters*, 10(12):1115–1123.



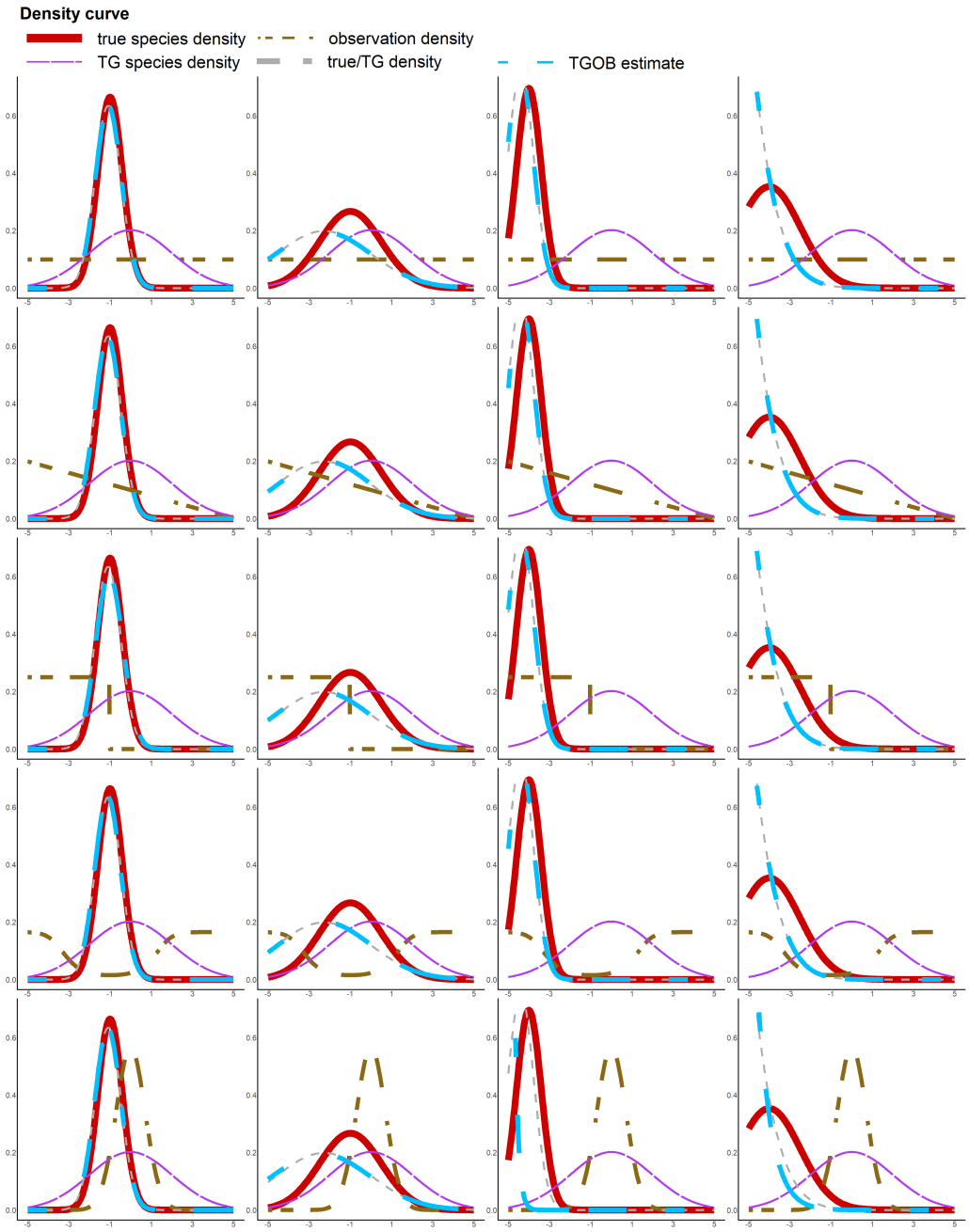
S1 Fig. Illustrations of μ_x, f_{θ_0} and s_x along x values. An example species density with the standard normal distribution (red curve), the density derived from μ_x chosen uniform over $[-5, 5]$ for the simulation study (black curve), and the observation density s_x of type LIN (gold curve).



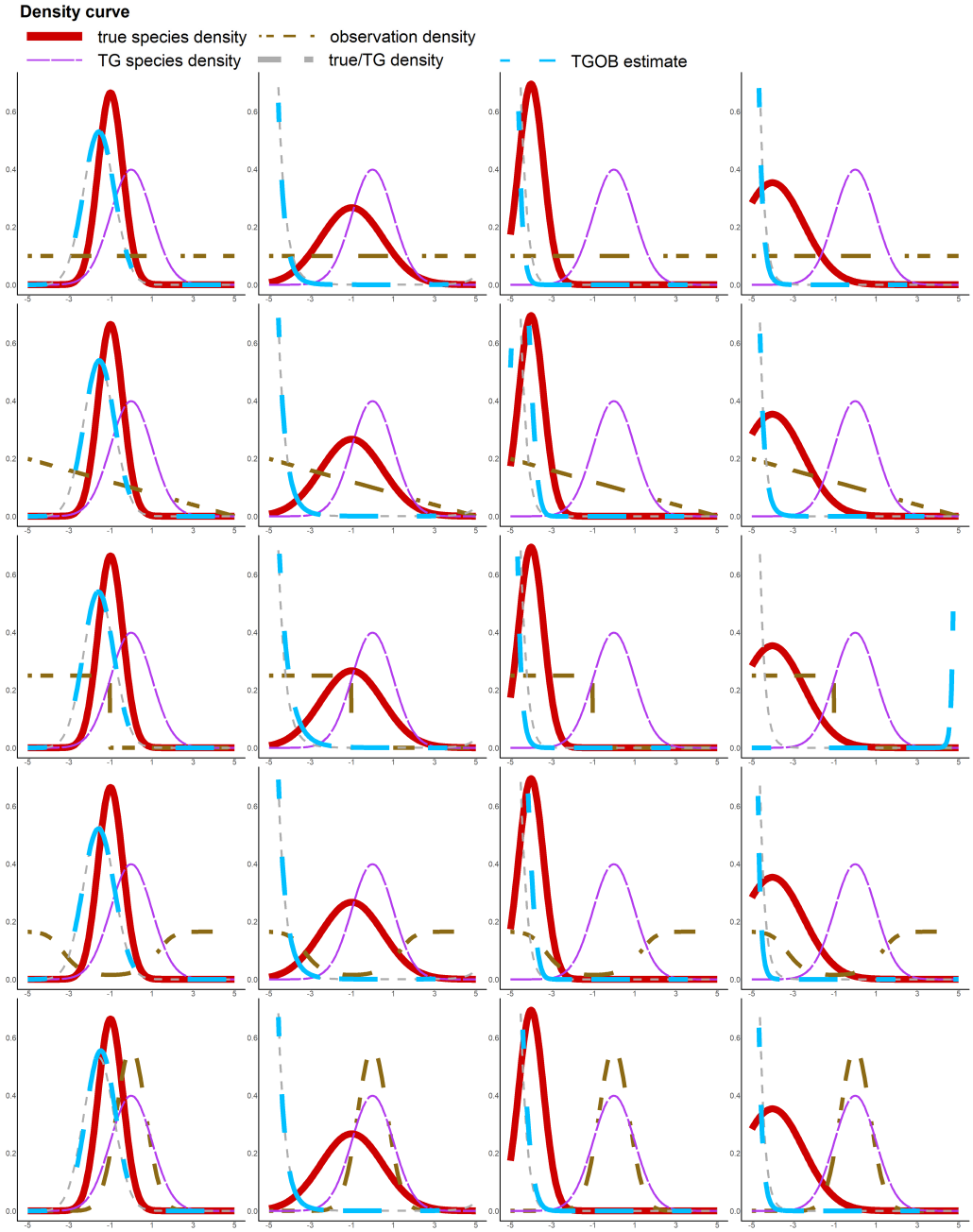
S2 Fig. Illustrations of all simulation results for UB. Plotted true species density (f_{θ_0}), observation density (s_x), observed points density ($f_{\theta_0} s_x$) and UB estimate of species density in the environmental space. Each situation of the simulation study is represented.



S3 Fig. Illustrations of all simulation results for TGOB with FLAT TG species density. Plotted true species density (f_{θ_0}), observation density (s_x), **flat** Target Group species density (a), ratio density of species over target group (f_{θ_0}/a) and TGOB estimate of species density in the environmental space. Each situation of the simulation is represented.



S4 Fig. Illustrations of all simulation results for TGOB with THICK TG species density. Plotted true species density (f_{θ_0}), observation density (s_x), **thick** Target Group species density (a), ratio density of species over target group (f_{θ_0}/a) and TGOB estimate of species density in the environmental space. Each situation of the simulation is represented.



S5 Fig. Illustrations of all simulation results for TGOB with THIN TG species density. Plotted true species density (λ_0), observation density (s_x), **thin** Target Group species density (a), ratio density of species over target group (λ_0/a) and TGOB estimate of species density in the environmental space. Each situation of the simulation is represented.

11.2 Appendix of chapter 3

Supporting information for :

Jointly estimating spatial sampling effort and habitat suitability for many species from opportunistic occurrences

Christophe Botella[1,2,3,4], Alexis Joly[1], Pierre Bonnet[3,5], François Munoz[6], and Pascal Monestiez[4]

[1]INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.

[2]INRAE, UMR AMAP, F-34398 Montpellier, France.

[3]Univ Montpellier, UMR AMAP, Montpellier, France.

[4]INRAE, BioSP, Site Agroparc, 84914 Avignon, France.

[5]CIRAD, UMR AMAP, F-34398 Montpellier, France.

[6]Université Grenoble Alpes, 621 avenue Centrale, 38400 Saint-Martin-d'Hères, France.

1 Appendix A: Expected estimators and Information matrix

Expected estimators. From the negative log-likelihood of the model expressed in equation **(3)** of the article manuscript, we derive an expression of the asymptotic density and intercept estimators in the system of equations 1. It shows that the density estimators minimize a weighted sum of Kullback-Leibler divergences from the true to estimated occurrences densities. We note in the following $n_i := |Z_i|$ and $\theta_i = (\alpha_i, \beta_i)$.

$$\begin{aligned} \mathbb{E}((\hat{\gamma}, \hat{\beta}_1, ..., \hat{\beta}_N)) &= \underset{\gamma, \beta_1, ..., \beta_N}{\text{argmin}} \sum_{i=1}^{N} (\int_D s\lambda^i d\mu) D_{KL}^{D}(s\lambda^i || s_\gamma \lambda^i_{(0,\beta_i)}) \\ \forall i \in [|1, N|], \qquad \mathbb{E}(\hat{\alpha_i}) &= \log(\int_D s\lambda^i d\mu / \int_D s_{\mathbb{E}(\hat{\gamma})} \exp(\mathbb{E}(\hat{\beta}_i)^T x) d\mu) \end{aligned} \tag{1}$$

Proof:

$$
\begin{aligned}
&\mathbb{E}(\hat{\theta}) \\
&= \lim_{n_1,\dots,n_N\to\infty} \underset{\theta}{\operatorname{argmin}} -\log(p(Z_1,\dots,Z_n|\theta)) \\
&= \underset{\theta}{\operatorname{argmin}} \lim_{n_1,\dots,n_N\to\infty} \sum_{i=1}^{N} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \frac{\sum_{k=1}^{n_i} \log(s_\gamma(z_i^k)\lambda^i_{\theta_i}(z_i^k))}{n_i} \right) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \lim_{n_i\to\infty} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \frac{\sum_{k=1}^{n_i} \log(s_\gamma(z_i^k)\lambda^i_{\theta_i}(z_i^k))}{n_i} \right) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \lim_{n_i\to\infty} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \int_D \frac{s(z)\lambda^i(z)}{\int_D s\lambda^i d\mu} \log(s_\gamma(z)\lambda^i_{\theta_i}(z))\mu(dz) \right) && \text{Large number law} \\
& && \text{and transfer theorem} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\int_D s\lambda^i d\mu\right) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log(s\lambda^i) d\mu - \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log(s_\gamma \lambda^i_{\theta_i}) d\mu \right) && \text{Large number law} \\
& && + \text{ independent term} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\int_D s\lambda^i d\mu\right) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i}{s_\gamma \lambda^i_{\theta_i}}\right) d\mu \right) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\int_D s\lambda^i d\mu\right) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} - \log\left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu}\right) + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \lambda^i_{\theta_i} d\mu}{s_\gamma \lambda^i_{\theta_i} \int_D s\lambda^i d\mu}\right) d\mu \right) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\int_D s\lambda^i d\mu\right) \left(\mathrm{nlogL}(\alpha_i) + D^D_{KL}(s\lambda^i || s_\gamma \lambda^i_{\theta_i})\right)
\end{aligned}
$$

Where $\mathrm{nlogL}(\alpha_i) := \frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} - \log\left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu}\right) = -\log\left(\frac{\left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu}\right)^1}{1!} \exp\left(-\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu}\right)\right)$ is the negative log-likelihood of a Poisson regression with a single count of value one. The likelihood is maximized when the Poisson parameter $\int_D s_\gamma \lambda^i_{\theta_i} d\mu / \int_D s\lambda^i d\mu = 1$, which then minimizes $\mathrm{nlogL}(\alpha_i)$ with $\mathrm{nlogL}(\alpha_i) = 0$, and translates into $\alpha_i = \log(\int_D s\lambda^i d\mu / \int_D s_\gamma \exp(\beta_i^T x) d\mu)$. In other words, we can chose α_i to minimize $\mathrm{nlogL}(\alpha_i)$ whatever the values of $\gamma, \beta_1, \dots, \beta_N, s, \lambda^1, \dots, \lambda^N$. This means that the minimization of the whole sum with respect to $\gamma, \beta_1, \dots, \beta_N$ is unaffected by the terms $(\int_D s\lambda^i d\mu)\mathrm{nlogL}(\alpha_i)$ which can be removed in the expression of $\mathbb{E}(\hat{\gamma}, \hat{\beta}_1, \dots, \hat{\beta_N})$, and brings us the first equation of system 1. The second equation of 1 is shown by remarking that, conversely, the term $D^D_{KL}(s\lambda^i || s_\gamma \lambda^i_{\theta_i})$ is totally independent of α_i. Indeed, when replacing α_i by $\alpha_i + \delta$ we have :

$$
\begin{aligned}
D^D_{KL}(s\lambda^i || s_\gamma \exp(\alpha_i + \delta + \beta_i^T x)) &= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \exp(\alpha_i + \delta + \beta_i^T x) d\mu}{s_\gamma \exp(\alpha_i + \delta + \beta_i^T x) \int_D s\lambda^i d\mu}\right) d\mu \\
&= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{e^\delta s\lambda^i \int_D s_\gamma \exp(\alpha_i + \beta_i^T x) d\mu}{e^\delta s_\gamma \exp(\alpha_i + \beta_i^T x) \int_D s\lambda^i d\mu}\right) d\mu \\
&= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \exp(\alpha_i + \beta_i^T x) d\mu}{s_\gamma \exp(\alpha_i + \beta_i^T x) \int_D s\lambda^i d\mu}\right) d\mu \\
&= D^D_{KL}(s\lambda^i || s_\gamma \exp(\alpha_i + \beta_i^T x))
\end{aligned}
$$

Finally, the computation of the expected estimators can be separated as follows: First, the densities parameters estimates $\gamma, \beta_1, \dots, \beta_N$ are given by resolving the first equation of the system 1, and then the intercept parameters estimates $\alpha_1, \dots, \alpha_N$ are given by resolving the others equations.

Fisher Information matrix of the model. We write $I(\theta)$, the global Fisher information matrix of our model parameters, and show its particular structure. Note that the Fisher information matrix is also the Hessian, or curvature, matrix of the negative log-likelihood. Indeed, $I(\theta)$ gathers the second and cross derivatives of the negative log-likelihood written previously in **equation (3)** of section **2.2 - Inference** of the article (see also Bickel and Doksum [2015], section 6.2.2 , p.386, for more details on the Fisher information matrix).

Because of our model structure, $I(\theta)$ has many 0. We compute its non-null submatrices in the following. We consider here, for simplifying notations, that all species densities are functions of the same vector of environmental features called x, such that $\forall z \in D, x(z) \in \mathbb{R}^p$.

$\beta_i \in \mathbb{R}^p$ is the vector of parameters that model species i density in the environmental space for any $i \in [|1, N|]$. Its Fisher information matrix for this parameter is derived from the second and cross derivatives of the negative log-likelihood, written in equation **equation (3)** of the article, with respect to the components of β_i. That is:

$I(\beta_i) = \int_D xx^T s\lambda^i_{\theta_i} d\mu$

$\alpha_i \in \mathbb{R}$ is the intercept parameter of species i that is directly linked to the global abundance and detection/reporting probability of the species. Its information equals the total expected occurrences count of species i:

$I(\alpha_i) = \int_D s\lambda^i_{\theta_i} d\mu = \mathbb{E}(n_i)$

$\gamma_j \in \mathbb{R}$ is the parameter of the sampling effort in cell j. The cross information between cell j and j' is null when $j \neq j'$ cells form a partition of D and don't intersect. Its information equals the total expected occurrences count of cell j:

$$\begin{aligned} I(\gamma_j) &= \sum_{i=1}^{N} \int_D s\lambda^i_{\theta_i} d\mu \\ &= e^{\gamma_j} \sum_{i=1}^{N} \int_{c_j} \lambda^i_{\theta_i} d\mu \\ &= \mathbb{E}(n^j) \end{aligned} \tag{2}$$

The cross information of γ_j and β_i is written:

$I(\gamma_j, \beta_i) = \int_{c_j} x e^{\gamma_j} \lambda^i_{\theta_i} d\mu$

The cross information of γ_j and α_i equals the expected occurrences count of species i in cell j:

$I(\gamma_j, \alpha_i) = \int_{c_j} e^{\gamma_j} \lambda^i_{\theta_i} d\mu = \mathbb{E}(n^j_i)$

The cross information of β_i and α_i is written:

$I(\beta_i, \alpha_i) = \int_D x s \lambda^i_{\theta_i} d\mu$

The remaining of the Information matrix is null. In particular we have:

$$I(\gamma) = \begin{pmatrix} I(\gamma_2) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I(\gamma_Q) \end{pmatrix}$$

Thus, we exhibit the structure of $I(\theta)$ as follows:

$$I(\theta) = \begin{pmatrix} I(\gamma) & I(\gamma,\alpha_1)^T & I(\gamma,\beta_1)^T & \dots & I(\gamma,\alpha_N)^T & I(\gamma,\beta_N)^T \\ I(\gamma,\alpha_1) & I(\alpha_1) & I(\beta_1,\alpha_1)^T & 0 & 0 & 0 \\ I(\gamma,\beta_1) & I(\beta_1,\alpha_1) & I(\beta_1) & 0 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 & 0 \\ I(\gamma,\alpha_N) & 0 & 0 & 0 & I(\alpha_N) & I(\beta_N,\alpha_N)^T \\ I(\gamma,\beta_N) & 0 & 0 & 0 & I(\beta_N,\alpha_N) & I(\beta_N) \end{pmatrix} \tag{3}$$

2 Appendix B: Model identifiability and robustness

Necessary and sufficient condition for structural identifiability. Our model is structurally identifiable (for all set of parameters) in the multi-species case if it is in the single species case. The single species case is a Poisson process whose log-linear intensity function may be noted $z \to \theta^T v(z)$ where $\forall z \in D,\ v(z) = (1, 1_{z\in c_2}, ..., 1_{z\in c_Q}, x_1(z), ..., x_p(z))$, containing the intercept, the indicator functions of the cells c_j, and the environmental features vector. Then, according to the CNS identifiability condition shown for log-linear Poisson processes in Rathbun and Cressie [1994], the model is identifiable if and only if the matrix $\int_D v(z)v(z)^T dz$ is of full rank, i.e. of rank $1 + p + Q - 1$.

This condition means that there must exist no linear condition of the non constant functions of v that

is constant. This condition is fulfilled if there is no linear combination of the environmental features that is constant across all sampling cells. For a single environmental feature, it would mean that this feature must vary at least inside one sampling cell. In the multivariate case, a simply interpretable identifiability condition is hard to provide. Fulfilling the condition above is sufficient to insure unicity and convergence of the estimator for any dataset. However, on finite number of occurrences, being close to non-identifiability is often synonym of facing numerical approximation problems in the likelihood optimization, or getting high correlations between distinct parameters estimators. We need stronger conditions to insure a good estimability ([Jacquez and Greif, 1985]) of the model parameters. We thus advise the user, after having fit the model, to check the condition number of the inverse observed Fisher Information Matrix. This matrix may be computed by replacing parameters of the Information matrix in equation 3 by their estimates. The closer the condition number is to 1, the fewest is the global covariance between pairs of distinct parameters estimators.

Still, an option for the user, in the first place before fitting the model, is to compute numerically the condition number of the matrix $\int_D v(z)v(z)^T dz$ when designing the sampling mesh. Then, the user may chose among the possible sampling meshes one that has a condition number inferior to 10^6 (from our experience) while keeping in mind the other conditions provided in the article. This may directly eliminate some designs and is much faster than fitting the model and computing the condition number over for whole information matrix, even though the latter is a more accurate indice of estimability as it accounts for the data points distribution.

Remarks on model robustness. Profile (2) and (3) of the simulation experiment illustrate a limit of the method robustness: The sampling model does not allow the estimate to converge exactly towards the true sampling model. Indeed, the latter varies sharply in the middle of the sampling cells defined for the model. So, our estimation is necessarily an imperfect approximation of the truth. More generally, the following property formalizes a sufficient condition of estimation bias due to a lack of model robustness:

Property: If the model fulfils the structural identifiability conditions (see first paragraph of **Appendix B**) and there exists a non-constant function $g \in (\mathbb{R}^+ \backslash \{0\})^D$ such that $s/g \in \{s_\gamma, \gamma \in \mathbb{R}^{Q-1}\}$ and $\lambda^i g \in \{\lambda^i_{\theta_i}, \theta_i \in \mathbb{R}^{p_i+1}\}$, then the expected density estimates $s_{\hat{\gamma}}, \lambda^1_{\hat{\theta}_1}, ..., \lambda^1_{\hat{\theta}_1}$ are biased with $s_{\hat{\gamma}} = s/g$ and $\forall i \in [|1, N|]$, $\lambda^i_{\hat{\theta}_1} = \lambda^i g$.

<u>Proof:</u> If the model is identifiable, the parameter estimators is consistant whatever are $s, \lambda^1, ..., \lambda^N$. Then,

there is only one set of solutions $(s_{\hat{\gamma}}, \lambda^1_{\hat{\theta}_1}, ..., \lambda^1_{\hat{\theta}_1})$ to equations of system 1. We can easily show that the $(s/g, \lambda^1 g, ..., \lambda^N g)$, where by assumption each component belongs to the associated parametric function class of the model, is the solution because it cancels all the KL Divergences terms in the sum of first equation in 1, and thus minimizes it globally (Any KL Divergence is ≥ 0 by definition). Indeed, $\forall i, D^D_{KL}(s\lambda^i || (s/g)(\lambda^i g)) = D^D_{KL}(s\lambda^i || s\lambda^i) = 0$, by definition of the KL divergence. As g is non-constant, $s/g \neq s$ and $\forall i, \lambda^i g \neq \lambda^i$. In conclusion, we have shown that the estimators converge asymptotically towards biased sampling effort and species intensities.

In real applications, there is no function available in our parametric function classes that exactly fit the true occurrences densities, even with infinite numbers of occurrences. The key question is then: Will our estimation converge asymptotically towards the best approximation of the true density available in our function class? Not necessarily. Indeed, it may exist sets of couples $(s_{\hat{\gamma}}, \lambda^i_{\hat{\theta}_i})$ such that asymptotically each product density $s_{\hat{\gamma}}, \lambda^i_{\hat{\theta}_i}$ best approximates the product density $s\lambda_i$, but neither $s_{\hat{\gamma}}$ is the best approximation of s nor $\lambda^i_{\hat{\theta}_i}$ is the best approximation of λ_i in their respective model function class. This non-optimal approximation can appear when the model is incorrectly specified. It depends on an interplay of the true densities and their model. This is not a particular caveat of our method though, it is intrinsic to any method that would try to separate more than one density from one set of points. It is thus also the case for the single species approach (Warton et al. [2013]).

3 Appendix C: Estimation variance analysis

Our model is in the canonical exponential family, and thus the vector or parameter estimators $\hat{\theta} := (\hat{\gamma}, \hat{\alpha_1}, \hat{\beta_1}, ..., \hat{\alpha_N}, \hat{\beta_N})$ asymptotically follows a multivariate gaussian distribution (see Bickel and Doksum [2015], section 5.3.3, p.322-323). In the present case of one realization from a Poisson process, the variance-covariance matrix is simply the inverse of the Fisher Information matrix, introduced in equation 3 of Appendix A.

$\Sigma(\hat{\theta}) \; = \; I(\theta)^{-1}$.

Effect of occurrence rate. We use this formula and equation 3 in the R script `Variance_Script.R` (downloadable from the article Github repository: `https://github.com/ChrisBotella/SamplingEffort`) to effi-

ciently compute the model parameters variance-covariance matrix for a given scenario: A spatial domain D, sampling effort s, species number N and intensities $\lambda_1, ...\lambda_N$ (defined from their densities and expected number of occurrences $n_1, ..., n_N$) and the model sampling cells. We compute the variance for the profile 2 of the complementary simulation setting (see **Appendix F**). We set the number of occurrences for species 1 to 100 while varying the number of occurrences for the other species, conversely. Figure 1 shows, in the upper panel (resp. lower panel), how species 1 (resp. 2) parameters variance decreases when increasing the number of occurrence of species 1 (resp. 2) through the curve in blue (resp. curve in red). The upper panel (resp. lower panel) also shows through the curve in red (resp. in blue) that the variance of the focal species 1 (resp. 2) parameter decreases when increasing the occurrence rate of the other species 2 (resp. 1) while occurrence rate of the focal one is kept constant. Indeed, increasing the number of occurrence of any species enables the model to better estimate the sampling effort which makes easier the estimation of every other species parameters. Indeed, in equation 2, we see that the information gained on the sampling effort in cell j is the expectation of the total number of occurrences in this cell $\mathbb{E}(n^j)$ of all species so that each species contributes proportionally to its number of occurrences in the cell to improve the estimation of γ_j. Still, note, as shown by Figure 2, the indirect variance reduction mechanism from one species to another is slower than increasing the occurrence rate of the focal species itself.

Effect of removing the parameter. As proposed in the **model design guidelines** paragraph of section 2.1 of the article, we can drastically reduce the estimation variance of all species parameters by excluding an environmental variable from the model of one species (say species i) while keeping its in the model training data. This is a special case of conditional estimation (see next paragraph) where we condition on $\beta_i = 0$. It means that we assume a priori the species i to be indifferent to the variation of the environment variable across the study domain D. In this case, the model knows that the species intensity is constant along this environmental variables (all others kept constant) and can then use the variation of occurrences concentration along this gradient to better estimate the variation of sampling effort. We show this in the same theoretical context as last paragraph, which corresponds to the sampling effort profile 2 of the simulation experiment. We now compute the asymptotic parameters variance of species 1 (β_1) given that we known the exact niche parameters of species 2 (β_2) along the environmental variable x. This variance is simply obtained by removing the columns and lines of the information matrix $I(\theta)$ (see equation 3 in Appendix A) that are associated

with β_2, obtaining $I(\theta_{-\beta_2})$, and numerically inverting $I(\theta_{-\beta_2})$ to get the new estimators variance-covariance matrix $\Sigma(\hat{\theta}_{-\beta_2})$. In the upper panel of Figure 1 we represented the estimation variance on density parameters of species 1 extracted from $\Sigma(\hat{\theta}_{-\beta_2})$ with a growing occurrence rate for species 1 (purple curve) or species 2 (green curve). We can see that (i) the variance is always lower or equal compared to the cases where β_2 is estimated (green *le* red, purple *le* blue), (ii) it is especially lower for small sample size (for 100 occurrences, green is well below red, and purple is well below blue), (iii) it enhances the indirect variance reduction effect from increasing the occurrences rate on another species (green is well below red for all occurrences rates). To lighten the graph, we have not added to the lower panel the effect of removing parameters β_1 on estimation of β_2 but it works the same way.

Variance reduction with conditional estimation, the general case. Last paragraph showed that the estimation variance was reduced, when setting the parameters β_i of some species i to 0, on all other species parameters. We have shown it for a specific simulation scenario and it is only a particular case of conditional estimation, i.e. estimating some parameters when the values of others is given, which can be used more broadly with our method. We show here mathematically that (i) the variance reduction is not specifically due to the chosen simulation scenario but appears in any case, and (ii) it appears whatever are the parameters θ_i over which we condition. We first recall that when we have many occurrences for all species, we have that (see Bickel and Doksum [2015], section 5.3.3, p.322-323):

$$\lim_{n_1,\dots,n_N\to\infty} \mathcal{L}(\hat{\theta}) = \mathcal{N}(\theta, \Sigma(\theta))$$

Here we re-order the parameter estimation vector $\hat{\theta} = (\hat{\gamma}, \hat{\theta}_1, ..., \hat{\theta_{i-1}}, \hat{\theta}_{i+1}, ..., \hat{\theta}_N, \hat{\theta}_i)$ and decompose its variance-covariance matrix as follows:

$$\Sigma(\theta) = \begin{pmatrix} \Sigma_{-\theta_i} & \Sigma_c^T \\ \Sigma_c & \Sigma_{\theta_i} \end{pmatrix}$$

We also note $\hat{\theta}_{-i} := (\hat{\gamma}, \hat{\theta}_1, ..., \hat{\theta_{i-1}}, \hat{\theta}_{i+1}, ..., \hat{\theta}_N)$. The Gaussian conditionning theorem states that the conditional law $\hat{\theta}_{-i}|\hat{\theta}_i$ is a multivariate gaussian distribution with variance covariance matrix $\Sigma(\theta_{-i}) = \Sigma_{-\theta_i} - \Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$. The individual variances of all parameter are the diagonal elements of the latter matrix. We can now easily show that they are all smaller than the original variances, i.e. the diagonal elements of $\Sigma_{-\theta_i}$, because the diagonal elements in the matrix $\Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$ are all strictly positive. Indeed, $\Sigma_{\theta_i}^{-1}$ is positive definite as the inverse of Σ_{θ_i}, which is positive definite as a variance-covariance matrix. Then, the jth diagonal element

of $\Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$ is of the form $a_j^T \Sigma_{\theta_i}^{-1} a_j > 0$ (where a_j is jth column of Σ_c) by definition of positive definite matrices. In summary, the variance reduction of the estimator conditionally to the parameters of species i is strict whatever the value of θ_i.

Effect of the number of sampling cell. With the same setting, we evaluate the effect of the number of modelled sampling cells, evenly spaced along the longitude of the square domain. In Figure 2, we plot the asymptotic estimation variance on species parameters, computed numerically through the inversion of the information matrix, as a function of the number of cells. All estimators variance increase with the number of cells, but not all types of parameters at an equal speed. More precisely, we see that the variances on $\beta_{1,1}$ and $\beta_{2,1}$, which both control the optimum of the species gaussian density along the environmental gradient x, explode very quickly, whereas the parameters controlling the niche breadth remains reasonable even for 20 cells. Above 20 cells, the model shows a weak numerical identifiability, checked through the high condition number of the information matrix. When including too many cells, we decrease the ability of the model to separate the effect of the environmental variable, varying less within each cell, from the cell effect. However, the identifiability may not concern all parameters simultaneously: The species niche breadth parameters seem not very sensitive to the increased number of cells. However, the sampling effort approximation error increases as we decrease the number of cells, and this effect is not taken into account in the estimation variance. Thus, determining the best size of cells should be based on cross-validation using a density evaluation metric (Tsybakov [2009]). For a K-fold cross-validation, we recommend to build the folds so that each one contains approximately a proportion $1/K$ of the occurrences of every individual cell, because no sampling cell should be empty or scarce for training.

4 Appendix D: Inference and implementation details

For a given mesh across which a cellwise constant sampling effort is defined, we fit log-linear Poisson processes for multiple species with a shared term in their linear predictor, i.e. the log-sampling effort. We here introduce a maximum-likelihood fitting procedure. We use an approximation of the Poisson process likelihood by a Poisson regression likelihood using background points, as described in Berman and Turner [1992] and Warton et al. [2010], which we extend to the joint likelihood of a marked Poisson process.

We consider the set of observed occurrences for any species $i \in [1, N]$ $Z_i = \{(z_1^i, i, 1), ..., (z_{n_i}^i, i, 1)\}$, i.e. a set

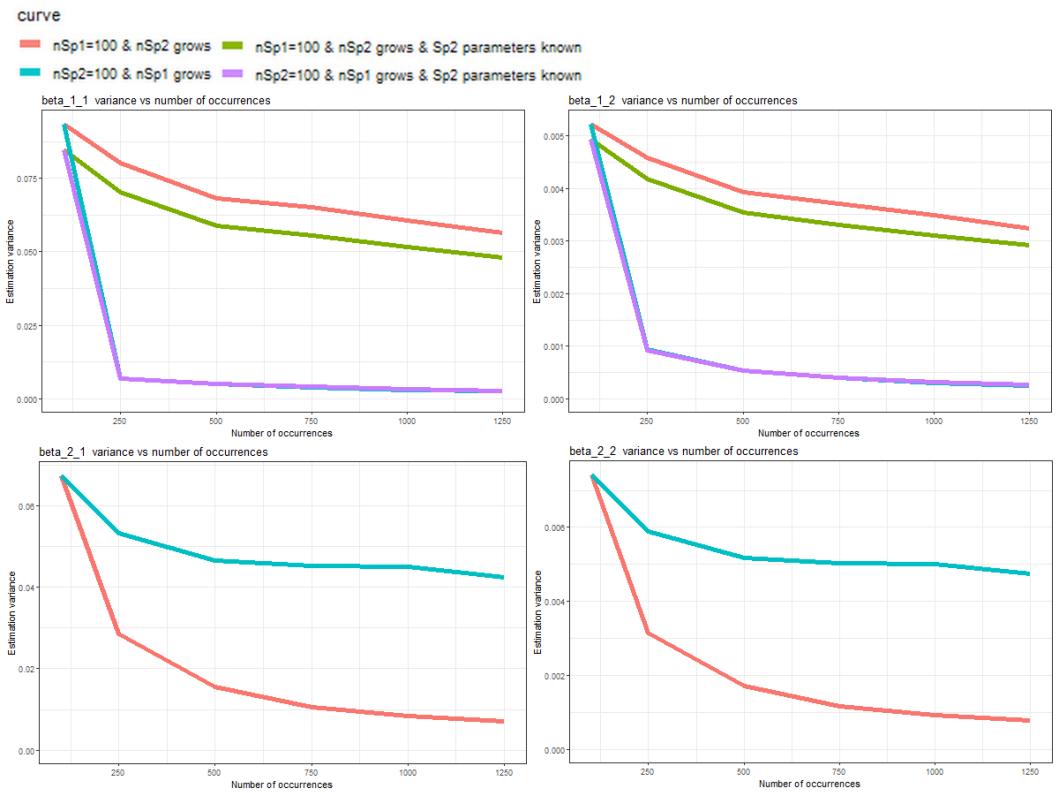


Figure 1: Asymptotic species densities parameters estimation variance as a function of the number of each species occurrence for the simulation setting of profile 2 described in section 2.4 of the article manuscript. $\beta_{1,1}$ and $\beta_{1,2}$ (resp. $\beta_{2,1}$ and $\beta_{2,2}$) are respectively the first and second parameters modeling the gaussian density of species 1 (resp. species 2) along the environmental gradient x.

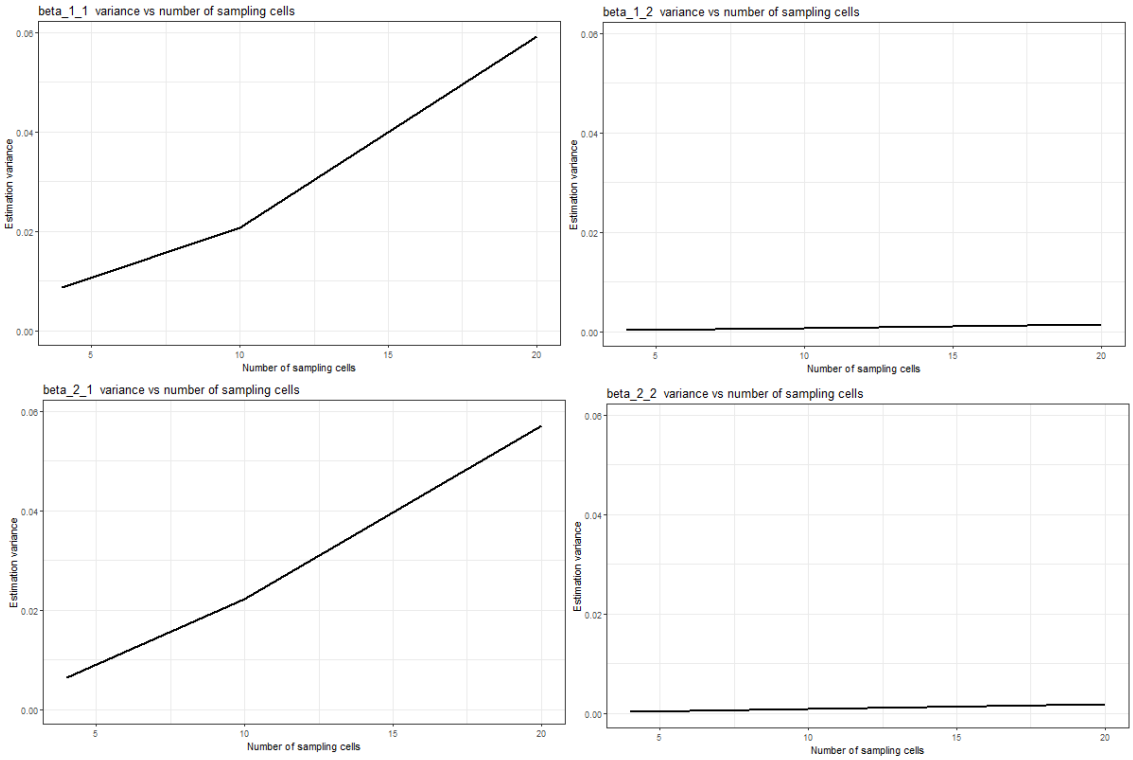


Figure 2: Asymptotic species densities parameters estimation variance as a function of the number modelled sampling cells (regularly spaced along the longitude of the domain) in the simulation setting of profile 2 described in section 2.4 of the article manuscript. $\beta_{1,1}$ and $\beta_{1,2}$ (resp. $\beta_{2,1}$ and $\beta_{2,2}$) are respectively the first and second parameters modeling the gaussian density of species 1 (resp. species 2) along the environmental gradient x. Above 20 cells, we began to diagnose weak numerical identifiability (through the condition number of $I(\theta)$) of the model making the variance-covariance matrix unreliable.

of points marked with the species label i and the state 1. We have to maximize the joint likelihood of $Z_1, ..., Z_N$ with respect to all model parameters introduced in the previous section $\theta := (\alpha_1, ..., \alpha_N, \beta^1, ..., \beta^N, \gamma_1, ..., \gamma_C)$:

$$\begin{aligned} p(Z_1, ..., Z_N|\theta) &= \prod_{i=1}^N \left[\frac{(\int_D s(z)\lambda_i(z)dz)^{n_i}}{!n_i} \exp\left(-\int_D s(z)\lambda_i(z)dz\right) \prod_{k=1}^{n_i} \frac{s(z_k^i)\lambda_i(z_k^i)}{\int_D s(z)\lambda_i(z)dz} \right] \\ \Leftrightarrow \quad p(Z_1, ..., Z_N|\theta) &\propto \prod_{i=1}^N \left[\exp\left(-\int_D s(z)\lambda_i(z)dz\right) \prod_{k=1}^{n_i} s(z_k^i)\lambda_i(z_k^i) \right] \\ \Leftrightarrow \quad log(p(Z_1, ..., Z_N|\theta)) &= \sum_{i=1}^N \left[\sum_{k=1}^{n_i} log(s(z_k^i)\lambda_i(z_k^i)) - \int_D s(z)\lambda_i(z)dz \right] \end{aligned} \tag{4}$$

The likelihood is factorized over species as we assume that their processes are independent given the environment.

The integral terms are often very costly to compute exactly when we deal with multiple high resolution raster of environmental variables. We rather use a numerical approximation. Each integral is replaced by a weighted sum of $s\lambda_i$ computed at some quadrature points $Z_i^q = \{(z_1^q, i, 0), ..., (z_Q^q, i, 0)\}$ marked with their species label i and state 0 indicating it is a background point, associated with some weights $w_1^i, ..., w_Q^i$, selected such that $\int_D s(z)\lambda_i(z)dz \approx \sum_{k=1}^Q w_k s(z_k^q)\lambda_i(z_k^q)$. Background points are also called quadrature points, or pseudo-absences in the Poisson process SDM literature (Warton et al. [2010]).

Numerical quadrature strategy and background points. We chose to draw uniformly background points to achieve the approximation of the integral through the unbiased Monte Carlo estimator. More precisely, Berman and Turner [1992] re-expressed the likelihood by including the points of Z_i among the quadrature points Z^q, and defining adapted weights. We note $w(z, i, e)$ the weight associated with the marked point (z, i, e).

$$\begin{aligned} \log(p(Z_1, ..., Z_N|\theta)) &\approx \sum_{i=1}^N \sum_{(z,i,e)\in Z_i \cup Z_i^q} 1_{e=1} \log(s(z)\lambda_i(z)) - w(z, i, e)s(z)\lambda_i(z) \\ &= \sum_{(z,k,e)\in \cup_i (Z_i \cup Z_i^q)} w(z, k, e) \left[y(z, k, e) \log(s(z)\lambda_i(z)) - s(z)\lambda_i(z) \right] \end{aligned} \tag{5}$$

Where the $y(z, k, e) := 1_{e=1}/w(z, k, e)$ are the Poisson regression pseudo-counts (non-integers), and we recall that by construction of our model $s(z)\lambda_i(z) = \exp(\sum_{j=1}^C \gamma_j 1_{z\in c_j} + \alpha_i + \beta^{iT} x_i(z))$. We end up with a Poisson regression log-likelihood that approximates well our initial log-likelihood when there are enough properly selected quadrature points. We use the same quadrature points and associated weights for all species. Now, we need to explain how are selected those points, and how are computed their weights $w(z, i, e)$. An

unbiased manner to approximate the integral is the Monte Carlo method: We use the average of $s\lambda_i$ over uniformly sampled background points on D to approximate the integral $\int_D s(z)\lambda_i(z)dz$. However, occurrences in Z_i's are not uniformly distributed over D and we need to ensure that they will not bias our approximation. For this purpose, the sum of weights of occurrences is negligible compared to the sum of weights of quadrature points and altogether:

$$\forall(z,i,e) \in \cup_i(Z_i \cup Z_i^q) w(z,i,e) = \begin{cases} \frac{|D|}{100 n_i} & \text{if } e = 1 \\ \frac{99|D|}{100Q} & \text{if } e = 0 \end{cases}$$

This yields the following expression for the approximation of integral term $\int_D s(z)\lambda_i(z)dz$:

$$\begin{aligned} \int_D s(z)\lambda_i(z)dz &\approx \sum_{z \in Z_i \cup Z_i^q} w(z)s(z)\lambda_i(z) \\ &= \frac{1}{100} \sum_{z \in Z_i} \frac{|D|}{n_i} s(z)\lambda_i(z) + \frac{99}{100} \sum_{z \in Z_i^q} \frac{|D|}{Q} s(z)\lambda_i(z) \end{aligned}$$

With this setting, all weights sum to $|D|$ (area of D), while weights of species occurrences alone represent only 1%, which we have noticed to be enough to not bias the approximation in our experience.

Particularity of the application to real dataset. For the real dataset of occurrences, we use an alternative strategy to insure that all the sampling cells have background points and that they capture the environmental variability of each cell. We uniformly draw a fixed number (6) of background points uniformly in each sampling cell. As each sampling cell has the same size in the present case, we can keep the same weighting scheme as previously, and the procedures weighted sum will also converge to the target integral. We can show this by decomposing the integral into a sum of integrals over each sampling cell multiplied by the inverse of the total number of cells and then using the Monte Carlo (because points are uniformly drawn inside cells).

Implementation details. The inference is performed using a software for Generalized Linear Model penalized with L1 (with R package `glmnet`) to estimate parameter values that maximize the penalized version of the likelihood, for given y_j, $Z_1, ..., Z_N$ and w.

The `R` code used for fitting the model can be found on the following Github repository: `https://github.com/ChrisBotella/SamplingEffort`. Equation 6 gives the R formula building the model design matrix passed to `glmnet`.

$$
\begin{aligned}
\texttt{y} \quad \sim \quad & 1 + \texttt{SamplingCell} + \texttt{species1} : (x_1^1 + ... + x_{p_1}^1) + \texttt{species2} : (1 + x_1^2 + ... + x_{p_2}^2) \\
& ... + \texttt{speciesN} : (1 + x_1^N + ... + x_{p_N}^N)
\end{aligned} \tag{6}
$$

The categorical effect of a point `SamplingCell` is the effect of its cell. There are $C - 1$ parameters for the sampling effort because it is impossible to identify the global intercept and the parameters of all sampling cells. Thus, we need to chose a way to constrain the effects of the C cells with $C - 1$ parameters, or in other words, to define contrasts. We chose the `SamplingCell` contrasts as `contr.sum`, $\sum_{j=1}^{C} \gamma_j = 0$. This way the L1 penalty induces a shrinkage of all sampling cells parameters toward zero, rather than a shrinkage toward a reference cell as would have done the `contr.treat` contrasts. Concerning the species niche parameters, there are $p_i + 1$ parameters for species i and different species can depend on different environmental predictors. Note that the intercept of species 1 is grouped with the global intercept, again for identifiability reason. It explains that we can only estimate the species intensities and the sampling effort up to a constant factor. Using `glmnet` allows handling sparse matrices and performing our model with large number sampling cells, environmental features, background points, occurrences as explained in the real data illustration section.

5 Appendix E: Environmental variables tables

Name	Description	Values	Resolution (m)
CHBIO_1	Annual Mean Temperature	[-10.6,18.4]	1000
CHBIO_5	Max Temperature of Warmest Month	[36.4,6.2]	1000
CHBIO_12	Annual Precipitation	[318,2543]	1000
etp	Potential Evapo Transpiration	[133,1176]	1000
alti	Elevation	[-188,4672]	90
slope	Absolute elevation gradient	[0,13457]	90
awc_top	Topsoil available water capacity	$\{0, 120, 165, 210\}$	1000
bs_top	Base saturation of the topsoil	$\{35, 62, 85\}$	1000
spht	Aggregated land cover	{culti.,for.,past.,urb.,other}	100

Table 1: Table of environmental variables used in this study.

CLC category description	spht category name	Raster code
Non-irrigated arable land	cultivated	12
Permanently irrigated land	cultivated	13
Vineyards	cultivated	15
Fruit trees and berry plantations	cultivated	16
Complex cultivation patterns	cultivated	20
Land principally occupied by agriculture, with significant areas of natural vegetation	cultivated	21
Agro-forestry areas	cultivated	22
Pastures	grasslands	18
Natural grasslands	grasslands	26
Moors and heathland	grasslands	27
Sclerophyllous vegetation	grasslands	28
Broad-leaved forest	forest	23
Coniferous forest	forest	24
Mixed forest	forest	25
Transitional woodland-shrub	forest	29
Continuous urban fabric	urban	1
Discontinuous urban fabric	urban	2
Industrial or commercial units	urban	3
Road and rail networks and associated land	urban	4
Airports	urban	6
Green urban areas	urban	10
Sport and leisure facilities	urban	11
Port areas	other	5
Mineral extraction sites	other	7
Dump sites	other	8
Construction sites	other	9
Rice fields	other	14
Olive groves	other	17
Annual crops associated with permanent crops	other	19
Beaches, dunes, sands	other	30
Bare rocks	other	31
Sparsely vegetated areas	other	32
Burnt areas	other	33
Glaciers and perpetual snow	other	34
Inland marshes	other	35
Peat bogs	other	36
Salt marshes	other	37
Salines	other	38
Intertidal flats	other	39
Water courses	other	40
Water bodies	other	41
Coastal lagoons	other	42
Estuaries	other	43
Sea and ocean	other	44
No data	other	48
Unclassified land surface	other	49
Unclassified water bodies	other	50

Table 2: spht (Aggregated land cover) categories correspondance with Corine Land Cover 2012.

6 Appendix F: Complementary simulation study, a closer look on the density estimates

6.1 Methodology

We designed the following simulation study to examine more closely whether our approach allows reliably inferring sampling effort density and species densities from observed occurrences of 2 virtual species with heterogeneous sampling effort. Note that we do not use intercepts in the simulation because, as explained in section **2.1**, we can't estimate the absolute intensities over space but rather the relative intensities. We evaluate the estimation quality as the ability to recover the density over the environmental gradient, because it is the space over which both the species intensity, and the sampling effort are defined by our construction. This space is one-dimensional, enabling visualization. To reproduce this experiment, one must run the script called `Simu_and_graphs.R` on the article Github repository: `https://github.com/ChrisBotella/SamplingEffort`.

Spatial domain and species variable. We consider a square spatial domain $D = [0, 10]^2$ where the only environmental variable x is a linear gradient from west to east, such that $x(z) = z - 5$.

Virtual species. The environmental intensity of virtual species is modeled as a Gaussian function over the gradient x, i.e. $\forall z \in D,\ \lambda_i(z) \propto \exp((x(z) - \mu_i)^2/(2\sigma_i^2))$. It means that the expected x of a given species individual is μ_i (optimum constraint), and the variance of x over many individuals is σ_i^2 (niche breadth constraint), and λ_i is of maximum entropy. We use the following re-parameterization of species density:

$$\begin{aligned} \forall z \in D,\ \lambda_i(z) &\propto \exp\left(-\frac{(x(z)-\mu_i)^2}{2\sigma_i^2}\right) \\ &\propto \exp\left(\beta_1^i x(z) + \beta_2^i x(z)^2\right) \end{aligned}$$

With $\begin{cases} \beta_1^i = \frac{\mu_i}{\sigma_i^2} \\ \beta_2^i = -\frac{1}{2\sigma_i^2} \end{cases} \Leftrightarrow \begin{cases} \mu_i = -\frac{\beta_1^i}{2\beta_2^i} \\ \sigma_i = \frac{1}{\sqrt{-2\beta_2^i}} \end{cases}$

β_2^i being strictly negative. This re-expression will be useful as the method implementation gives us estimates of β_1^i, β_2^i for each i (see Inference section). In our simulation study we have two virtual species $i \in \{1, 2\}$ and we chose the optima to be $\mu_1 = -2.5$, $\mu_2 = 2.5$. Besides, the standard deviation of their intensities are

$\sigma_1 = \sigma_2 = 1.6$.

Types of sampling effort. We designed a case where the relative sampling effort strongly depended on the environment x, which makes harder separating sampling effort from species intensities: The relative sampling effort is a step function over D depending of the longitude only (like the feature x), and not of the latitude. We designed three profiles of relative sampling effort :

1. $s(z) = 1_{x(z)<0}$. This profile has a constant non-null effort on the western half of the domain, and no sampling on the eastern half.

2. $s(z) = 1 + 5\ 1_{x(z)\in[-4.5,-2.5[\cup[-0.5,1.5[\cup[2.5,4.5[}$. This profile has sharp variation inside the sampling cells of the model design.

3. $s(z) = 9 * \dfrac{\exp(-5x(z))}{1 + exp(-5x(z))} + 1$. This profile is a decreasing sigmoïdal function. It has also sharp variations inside sampling cells, plus they are continuous and monotonic all across the domain.

The fitted sampling model is well specified for type (1). Indeed, the point of discontinuity of the simulated sampling effort is a limit between the sampling cells. Thus, we expect to get exact estimates of species niches and sampling effort density. We test how the method recovers the species niches with only a partial sampling of the environmental range. However, for type (2), the simulated sampling effort varies in the middle of some modeled sampling cells so that it's impossible to get a perfect estimation. If the method is robust, we expect sampling effort estimate to approximate the average of the target in every sampling cell. Finally, the estimation can't be perfect for type (3) either. Here, the sampling effort co varies strongly and monotonically with the environmental variable, and it is expected to be the most problematic profile for the method.

Simulating species observed points. We drew $200,000$ occurrences for both species in each of the 3 sampling effort scenarios. For a defined relative sampling effort s and species intensity λ, we drew points according to a conditional Poisson process of intensity function $s\lambda$ over D. It is done using the following acceptance-rejection algorithm:

- Initialization: Determine an upper bound B of $s\lambda$ on D.
- Repeat:

1. Draw a point $z \sim U(D)$.

2. Draw a variable $y \sim U([0, B])$

3. We accept z if $y <= s(z)\lambda(z)$.

4. If $200,000$ points are accepted, finish the procedure, otherwise go back to 1).

We chose $200,000$ points as it is well enough for a satisfying convergence of the sampling effort and species intensities estimates, as shown by the standard deviation bounding curves of Fig. 3.

Background points. For each experiment, $50,000$ background points were uniformly drawn over D, which is enough for likelihood convergence in this simple setting.

6.2 Results

We analyse here the reliability of our joint estimation method for 2 simulated species with 3 scenarios of sampling effort. Fig. 3 which shows the mean and standard deviations of estimated relative sampling effort.

Unbiased niches and sampling effort estimates under good model specifications. Our simulation results first show that estimation of the relative sampling effort and of relative species intensities are unbiased under obervation scenario (1), *i.e.* when the species and sampling model is well designed. In scenario (1), there is no sampling on the eastern part of the domain, and constant sampling on the western part. Left graph of box A on Fig. 3 shows that the model perfectly captures the non sampled area, and the estimate on the western part is almost exact. Center and right graphs of box A show that the species intensities are also well recovered. The model uses the variation of species points occurrences in the western part to fit the whole species intensity model and is then able to make good prediction on the eastern part. Blue curves in Fig. 3 represent the observed standard deviation, which approximately delimit the 95% confidence interval (mean +/- 2 times the standard deviation) of the estimate over the 20 repetitions of the simulation. We remark a small bias likely due and/or to numerical approximation in the fitting algorithm. It is not due to the regularization path as we had a bias of similar order with a implementation `glmn`.

Approximation bias under bad sampling model design. Secondly, graphs of box B illustrate the results of scenario (2). It shows that even though the sampling effort model neglects actual variation inside sampling cells, the method provides a reasonably good approximation as the estimate is often close to the average of the

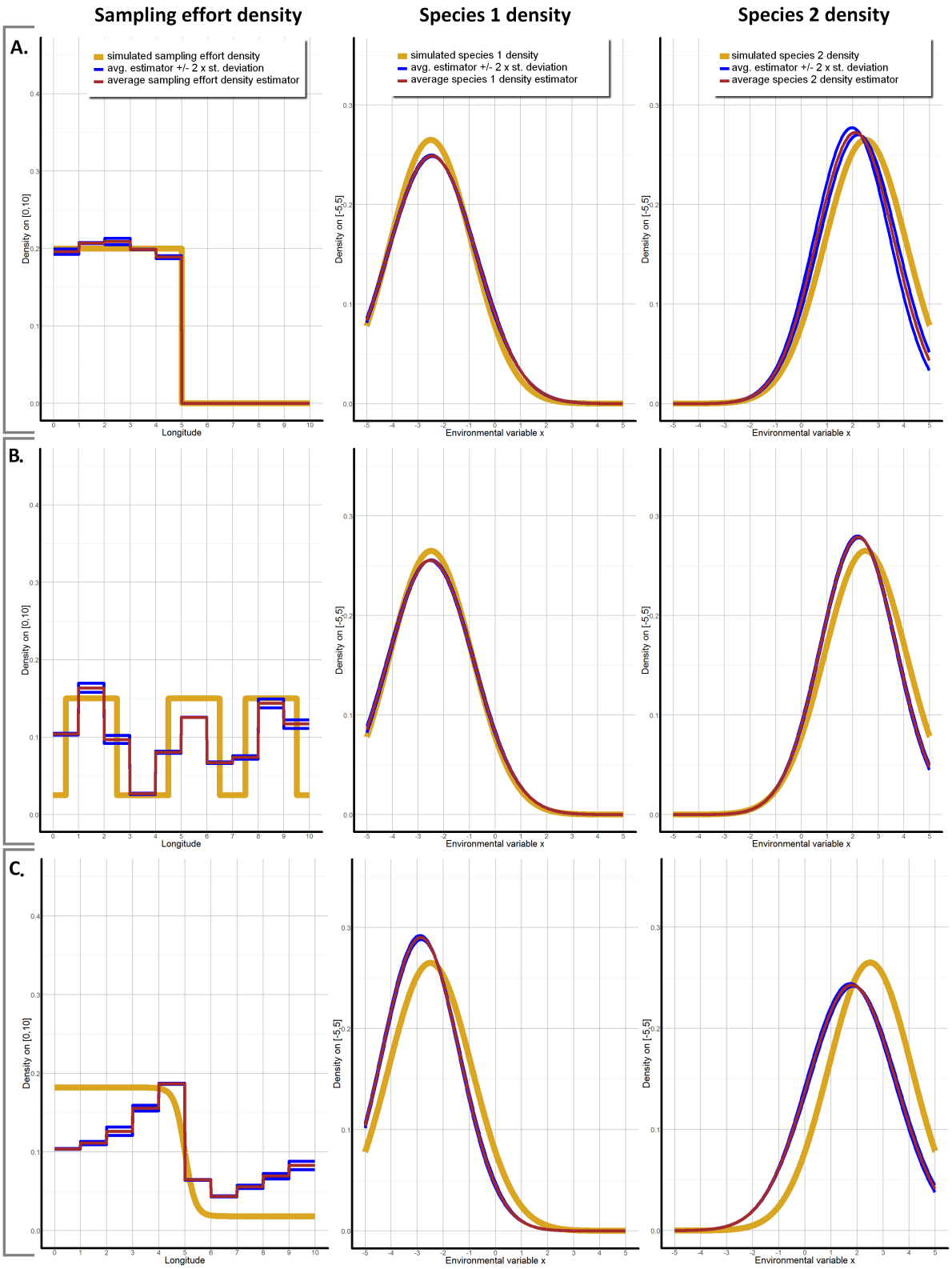


Figure 3: Sampling effort and the two species estimated densities for the three profiles of simulated sampling effort in the simulation experiment. A. type (1); B. type (2); C. type (3); see the paragraph "Types of sampling effort". Red curves are the mean estimates over 20 repetitions the simulation scenario, with the blue curves delimiting the approximate 95% confidence interval. Yellow curves are the targets. Sampling density (graphs on the left) is plotted against longitude, while species densities (graphs on the center and right) are plotted against x values (which are in bijection). The vertical grey lines on the graphs represent the longitudinal limits of sampling effort square cells.

true sampling effort in each cell. Besides, the species intensities estimates, on center and right graphs of box B, are slightly more biased than in case (1). For the scenario (3) illustrated by the densities of box C, we see a joint bias in the estimation of the species densities and the sampling effort. The species densities are deviated on the left, associated with an underestimation of the sampling effort for low x values and an over-estimation for high x values.

References

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38.

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227.

Rathbun, S. L. and Cressie, N. (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous poisson point processes. *Advances in Applied Probability*, 26(1):122–154.

Tsybakov, A. (2009). Introduction to nonparametric estimation. In *Springer Series in Statistics, ISBN 978-0-387-79051-0*. Springer-Verlag New York.

Warton, D. I., Renner, I. W., and Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168.

Warton, D. I., Shepherd, L. C., et al. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402.