THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Ecologie Fonctionnelle

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Portée par

Unité de recherche ECO&SOLS

Apports de la fouille de données à la modélisation des communautés écologiques

Présentée par Sara SI-MOUSSI

Le 18 décembre 2020

Sous la direction de Mickaël HEDDE, Esther GALBRUN et Wilfried THUILLER

Devant le jury composé de

Thibaud DECAËNS, Professeur des universités, Université de Montpellier	Président de jury
Olivier FRANÇOIS, Professeur des universités, Université Grenoble Alpes	Rapporteur
Elisa FROMONT, Professeure des universités, Université de Rennes I	Examinatrice
Tamara MÜNKENMÜLLER, Chargée de recherche avec habilitation, CNRS-UGA-USMB	Rapportrice
Mickaël HEDDE, Directeur de recherche, INRAE	Directeur
Esther GALBRUN, Chargée de recherche, University of Eastern Finland	Co-directrice
Wilfried THUILLER, Directeur de recherche, CNRS-UGA-USMB	Co-encadrant

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

Remerciements

Ce manuscrit achève un long travail de trois ans qui a commencé par un coup de chance, une invitation opportune à m'embarquer dans une aventure scientifique "outside the box" de ma formation initiale et du chaleureux cocon familial. Une aventure pendant laquelle j'ai acquis des connaissances et pris conscience de mon ignorance.

Et quoi de mieux pour entamer ce manuscrit que d'exprimer ma gratitude envers toutes les personnes qui ont contribué à son aboutissement de près ou de loin.

Tout d'abord, ce travail n'aurait jamais pu se faire sans mes directeurs de thèse Mickael Hedde, Esther Galbrun et Wilfried Thuiller qui m'ont permis de naviguer entre deux voire trois champs de recherche sans perdre pied. Merci de m'avoir admis dans votre discipline en dépit de mon background différent et de m'avoir permis d'explorer des pistes de recherche de manière autonome et pour m'avoir toujours rappelé l'essentiel et tendu la main lorsque je me perdais dans les bifurcations des questionnements scientifiques. Je vous suis aussi reconnaissante à tous les trois pour avoir été compréhensibles et à l'écoute quand les choses n'allaient pas bien.

Merci à Mike pour m'avoir très vite mise dans le bain lors de ma première année, et de m'avoir initiée à l'écologie fonctionnelle, aux lombriciens, tout en t'assurant que je ne perdais pas ma casquette math-info. Merci à Wilfried pour m'avoir vite fait voir les ponts entre l'écologie et l'informatique, pour m'avoir guidé tout au long de ma thèse, pour m'avoir intégré dans ton équipe et pour m'avoir permis d'élargir mon réseau scientifique. Mille mercis à Esther pour ta disponibilité, pour nos échanges sur Skype, pour tes conseils toujours pertinents et pour m'avoir permis de garder le lien avec la communauté de fouille de données.

Un grand merci au département MathNum de l'INRAE et au comité de la bourse nationale INRAE-INRIA 2017 sans quoi cette thèse n'aurait pu avoir lieu. Merci au projet BISE pour avoir financé ma participation aux journées de la Société Française de Biométrie en 2018, et également au projet GlobNets pour avoir financé mes missions, mes formations et mes participations aux congrès EGC et EEF.

Je remercie les membres du jury pour avoir accepté d'évaluer mon travail. En particulier, mes rapporteurs Tamara Munkenmuller et Olivier François pour avoir accepté de relire ce manuscrit, et Elisa Fromont pour avoir accepté d'examiner mon travail. Je suis particulièrement reconnaissante envers mon référent Thibaud Decaens pour ta disponibilité, pour ton support scientifique lors des comités et moral durant cette année particulièrement éprouvante. Merci d'avoir accepté de prendre part à mon jury de thèse.

Par ailleurs, je remercie les membres de mon comité de pilotage Marc Plantevitn Tanguy Daufresne et Eric Blanchart pour leurs conseils et nos discussions. Merci à Tanguy Daufresne à la fois via l'axe écologie qui m'a permis de me sentir moins dépaysée au début et pour m'avoir invité aux échanges et discussions diverses autour des modèles de niche, de coexistence et d'évolution. Je ne dirai pas que j'ai appris à coder en Mathématica mais au moins j'en sais un peu plus sur les approches numériques.

Je suis chanceuse d'avoir pu conduire mes travaux de thèse au sein de l'UMR Eco&Sols, dirigée par un équipe d'enfer: Laurent Cournac, Isabelle Bertrand et Philippe Thaler. Je remercie ses gestionnaires en particulier Farid pour m'avoir facilité mes démarches administratives et de séjour. Merci à Mélanie, Cassandre, Philippe B. et Eric M. pour le travail exceptionnel qu'ils font. Merci aux chercheurs du Thème 1 pour m'avoir été une source de connaissance parfois au-delà du périmètre de ma thèse et à ses chercheurs exceptionnels:

Laetitia Bernard, Eric Blanchart, Jean Trap, Elisa Taschen, Claude Plassard, ... En particulier, merci à Benoît Jaillard avec qui j'ai beaucoup appris de nos échanges sur les modèles, la combinatoire et l'approche systémique.

Je suis reconnaissante pour les nouvelles amitiés tissées durant ces trois années avec les permenants et non-permanents de l'unité: Camille, Najat, Saloua, Sitor, Damien ainsi qu'à mes cobureaux actuels et passés Esther, Atoumane, Amandine, Marie, Ivi et Rafa.

Je remercie le LECA pour m'avoir accueilli à plusieurs reprises. Merci à Florence pour m'avoir facilité les démarches administratives. Merci aux chercheurs, doctorants et stagiaires pour m'avoir permis de toujours repartir de Grenoble inspirée et revigorée par nos échanges et discussions. Special thanks for Laura Pollock and Tamara Munkenmuller for being always available to answer my questions. Merci aussi à Maya Guéguen et à l'équipe du GRICAD pour m'avoir permis d'accéder à et d'utiliser les ressources de calcul de l'UGA.

Je suis éternellement reconnaissante envers Camille Martinez pour ton crash-course en écologie des réseaux au début de ma thèse qui m'a permis de prendre vite des repères. Mille merci à Giovanni, Marc et Nico pour vos remarques/relectures de mes manuscrits et de manière générale pour m'avoir inspiré par votre enthousiasme et capacité d'écoute exceptionnelles. A special gratitude for my dear friend Dasha, our virtual coffee breaks made my quarantine days but nearly not as much as our in-person exchanges. Merci à Irene, Marie, Bills, Louise et tous les autres pour avoir maintenu le contact malgré la distance et pour l'inspiration que me suscite votre passion pour l'écologie et la biodiversité.

Merci au département d'informatique de l'Université de Montpellier, en particulier Philippe Janssen, Bruno Grenet et Stéphane Bessy pour m'avoir accueilli dans leurs équipes d'enseignement et pour avoir été des supers mentors tout au long de ma mission. Merci à mes stagiaires Maeva et Salah pour leur travail et leur curiosité scientifique.

Je remercie l'équipe InfoSols de l'INRAE Orléans en particulier Ghislain Girot, Sébastien Lehmann et Jean-Philippe Chenu pour leur accueil et pour m'avoir fait découvrir la pédologie et le travail exceptionnel du GISSOL. Merci à toutes les personnes et équipes ayant contribué aux données de vers de terre et à leur analyse. I'm grateful for Daniel Fernandez Marchàn and his team for providing the earthworm phylogenetic data and for guiding me through their analysis.

Merci aux consortiums TEBIS et groupement de recherche Ecologie Statistique (EcoStat) et Ecologie Trophique (GRET) ainsi que les équipes de projet EcoNet et Gambas pour m'avoir permis de présenter mes travaux et pour leurs feedbacks.

D'autre part, merci aux personnes qui m'ont soutenu moralement de près ou de loin parfois sans qu'ils s'en rendent compte. Je pense à Agnès et Nils Ferrand qui par leur gentillesse ont permis de faire passer l'amère période de confinement. Je pense aussi à mon amie Sihem pour son hospitalité à Grenoble, mes meilleurs amis Aymen, Islem, Achwak et Rahma.

Enfin, un énorme merci pour mes parents, mes soeurs, mes beaux-frèrs qui ont été tout près de moi malgré la distance du début jusqu'à la fin ainsi que mes neveux qui m'ont cajolée par leurs messages réguliers et leurs tendres attentions. Je n'aurais sans doute rien pu achever sans votre support inconditionnel à tous.

Synthèse

Contexte

La biodiversité est au cœur du fonctionnement des écosystèmes et des services fournis par la nature à l'homme. Dans le contexte des changements globaux, la biodiversité terrestre se trouve menacée par le réchauffement climatique, la modification des habitats naturels par l'homme et l'augmentation de l'artificialisation des sols. Afin d'anticiper les conséquences des changements globaux sur la biodiversité, les scientifiques s'efforcent de comprendre la relation entre les organismes et leur environnement. En biogéographie, ce problème est abordé en analysant les observations de présence ou d'absence d'espèces dans diverses conditions environnementales.

Cadre théorique

La théorie d'assemblage des communautés explique la composition des communautés à différentes échelles spatiales par un ensemble de processus fonctionnant comme des filtres qui peuvent favoriser ou défavoriser l'apparition d'une espèce. À l'échelle globale, les processus phylogéographiques, notamment l'historique évolutif et les migrations à large échelle, limitent le pool d'espèces potentielles à l'échelle régionale. Aux échelles régionale et locale, trois processus écologiques déterminent conjointement si une espèce du pool restreint peut être présente sur un site donné. Tout d'abord, les conditions environnementales doivent permettre la survie, la croissance et la reproduction de l'espèce. On parle d'adéquation de l'habitat, ou de filtre abiotique (1). Au sein d'une communauté, l'effet net des interactions directes et indirectes constitue le filtre biotique (2). Enfin, le site doit être accessible par voie de dispersion à partir d'une localisation antérieure. On parle de filtre de dispersion ou mouvement (3).

La théorie de la niche écologique associe à une espèce donnée une position dans un écosystème, définie par ses exigences écophysiologiques ou préférences d'habitat et ses capacités d'adaptation (niche de Grinnell), ainsi que son rôle au sein de la communauté, i-e sa place dans le réseau d'interaction (niche d'Elton). Les niches de Grinnell et d'Elton expliquent la réponse de l'espèce aux filtres abiotique et biotique respectivement. L'apprentissage de la première est la cible des modèles de distribution d'espèces (Species Distribution Models, SDM), la seconde celle des modèles d'apprentissage de réseau écologique (Ecological Network Learning, ENL). Les réseaux d'associations interspécifiques font le pont entre la biogéographie et l'écologie communautaire en reliant les interactions écologiques aux distributions des espèces. Les modèles joints de distribution des espèces concilient les deux tâches en permettant d'inférer les préférences abiotiques des espèces et leurs associations et de prédire la composition des communautés en réponse à l'environnement.

Objectifs

Ma thèse vise le développement de méthodes statistiques et d'outils informatiques appliqués à la fouille des données de la biodiversité dans le but de (1) comprendre les facteurs qui déterminent la distribution des organismes vivants; afin de (2) prédire leur réponse aux changements des conditions environnementales.

Tout d'abord, nous formalisons un cadre méthodologique modulable pour la modélisation des données de

communauté (multi-espèces). Par la suite, nous montrons comment les modules de ce cadre peuvent être exploités dans trois applications courantes en écologie des communautés: la prédiction de la réponse des communautés à l'environnement, la recommandation d'espèces, et l'inférence de réseaux d'associations interspécifiques. Enfin, nous terminons par un cas d'étude détaillé de l'utilisation du cadre pour comprendre les déterminants de la distribution des Lombriciens en France.

Recommandation d'espèces par emplacement géographique

La prédiction automatique de la composition en espèces dans des lieux géographiques est d'une grande importance dans le contexte des questionnements actuels sur la biodiversité. Inspirés par les concepts de niche abiotique et biotique, nous étudions deux architectures de réseaux neuronaux qui visent à exploiter les caractéristiques respectives de ces deux types de niches afin d'aborder la tâche de recommandation de plantes. La première proposition utilise des trames environnementales et exploite des techniques avancées d'extraction de caractéristiques, basées sur des représentations distribuées et des réseaux de neurones convolutifs. La seconde proposition s'appuie sur des cooccurrences proches de plantes et d'organismes provenant d'une liste de taxons établie par des experts. Nous constatons que la première solution surpasse la seconde en termes de précision des prédictions, mais la seconde solution fournit des indicateurs intéressants et plus faciles à interpréter. Les deux approches donnent des résultats prometteurs dans le cadre du défi GeoLifeCLEF 2019.

Inférence de réseaux d'associations asymétriques

La question de savoir si les interactions biotiques peuvent être déduites des modèles de cooccurrence a récemment fait l'objet de nombreux débats. En théorie, les interactions biotiques au sein des communautés devraient créer des motifs spatiaux d'association entre les espèces. Ces associations spatiales peuvent découler de tous types d'interactions, qui peuvent être symétriques ou asymétriques, et qui peuvent varier à la fois en intensité et en direction.

Lorsque l'observation directe des interactions est impossible et que seules les associations peuvent être observées, des méthodes statistiques ont été proposées pour tenter de déduire les interactions sous-jacentes à partir des associations. Cependant, ces méthodes soit ignorent la réponse des espèces à l'environnement (par exemple, modèles nuls basés sur les matrices d'incide uniquement), soit se limitent à l'apprentissage de dépendances symétriques (par exemple via des modèles joints de distribution d'espèces ou des modèles graphiques probabilistes non dirigés i-e champs de Markov aléatoires) ou de liens strictement dirigés (par exemple via des réseaux bayésiens). Leur capacité relativement faible à détecter des interactions connues et le fait que la plupart des interactions sont asymétriques nécessitent le développement d'un nouveau cadre d'inférence d'associations asymétriques.

Nous proposons ici un nouveau cadre qui (a) modélise les associations comme des influences dirigées d'une espèce source vers une espèce cible, paramétrées par deux variables latentes spécifiques à l'espèce : la réponse de la cible à la communauté et l'effet de la source sur la communauté ; et (b) intègre ces associations avec un modèle d'adéquation d'habitat (*habitat suitability model*). Nous évaluons d'abord la capacité du cadre à détecter les associations connues à partir de simulations de processus d'assemblage. Ensuite, nous mettons en évidence les propriétés intéressantes des réseaux d'associations appris grâce à une application sur des données de comptages de végétation le long d'un gradient méso-topographique alpin.

Le modèle que nous proposons permet de distinguer différents types d'associations, comme l'ont montré nos expériences de simulation. Appliqué à des données réelles, il permet de récupérer la majorité des principales associations connues des communautés végétales alpines. Dans l'ensemble, les résultats indiquent que les données de coabondance reflètent mieux les dépendances des espèces que les cooccurrences. Plusieurs facteurs influencent l'inférence de l'intensité d'une association notamment la quantité de chevauchement des niches des espèces, la densité réelle du réseau d'interaction et la présence de multiples associations confondantes. Enfin, le regroupement des espèces sur la base de leurs représentations latentes fournit des indications complémentaires sur la similarité écologique de nombreuses espèces.

Modèle joint de distributions d'espèces

Les modèles joints de distribution d'espèces (Joint Species Distribution Models, JSDM) visent à résoudre le problème d'inférence de réseaux d'associations tout en apprenant simultanément les préférences abiotiques des espèces, ce qui permet de prédire la composition des communautés en fonction des conditions environnementales. Contrairement aux approches qui modélisent une seule espèce à la fois qui bénéficient d'une boîte à outils riche, les implémentations actuelles de JSDM reposent principalement sur la famille des modèles linéaires généralisés (Generalized Linear Models, GLM) pour modéliser les réponses abiotiques. Les réseaux de neurones (Artificial Neural Networks, ANN) constituent une classe de méthodes d'apprentissage automatique qui a fait ses preuves dans plusieurs disciplines. Un ANN peut être considéré comme une composition d'un GLM et de fonctions déterministes d'apprentissage de caractéristiques à partir des données brutes.

Sur cette base, nous proposons une extension des JSDM qui exploite le pouvoir de représentation des réseaux neuronaux pour prendre en charge des co-variables complexes et modéliser des réponses non linéaires. Le modèle proposé supporte des distributions de probabilité dans la famille exponentielle par opposition aux autres JSDM basés sur les réseaux neuronaux qui ont été spécifiquement optimisés pour le modèle probit multivarié.

Les données de communautés écologiques présentent plusieurs défis, notamment un nombre d'espèces souvent plus élevé, avec quelques espèces prévalentes et de nombreuses espèces rares. Nous proposons l'utilisation des réseaux de neurones multi-tâches, tels que chaque sortie du modèle (espèce prédite) est associée à une tâche de prédiction. De plus, une partie des couches d'apprentissage des caractéristiques (représentation intermédiaire) est partagée entre les différentes tâches. Cela permet de réduire le nombre de paramètres et répartir la puissance statistique entre les tâches de prédiction des espèces.

La complexité de la modélisation des associations dans les JSDM est proportionnelle à la taille de la matrice de covariance. Celle-ci augmente avec le nombre d'espèces considérées. Nous suivons l'approche par variable latente utilisée dans les implémentations modernes de JSDM. Le processus d'inférence avec des variables latentes est un goulot d'étranglement majeur en matière de calcul, en particulier dans les jeux de données comportant de nombreuses observations. Pour contourner ce problème, nous avons recours à l'inférence variationnelle amortie (amortized variational inference) pour contrôler la complexité du modèle et assurer son passage à l'échelle. Pour ce faire, nous utilisons un type de réseau de neurones appelé "auto-encodeurs variationnels", intégré au modèle multi-tâche précédents.

La flexibilité des modèles proposés permet d'améliorer les performances prédictives. Cependant, elle rend plus ardue l'interprétation des résultats et du processus qui les a produits. Sans modifier le modèle, nous proposons d'utiliser un ensemble de routines décrites dans la littérature pour mettre en lumière l'importance et la forme de la contribution de chaque facteur environnemental à chaque espèce prédite.

Cas d'étude: biogéographie des lombriciens de France

Dans la présente étude, nous modélisons la distribution spatiale des assemblages de vers de terre, de l'échelle locale à l'échelle nationale en France métropolitaine, en réponse au climat régional et aux conditions locales de l'habitat. Ainsi, nous utilisons le JSDM à base de réseau de neurones multitâches introduit précédemment pour modéliser les associations biotiques entre espèces et prédire conjointement les réponses des espèces à l'environnement. Le modèle est entraîné sur les données d'occurrence de Lombriciens relevées par Marcel Bouché (1972).

Premièrement, nous comparons les performances de notre modèle à celles de modèles de distribution d'espèces, sur chaque taxon et en moyenne. Deuxièmement, à l'aide d'outils d'interprétabilité de modèle d'apprentissage, nous analysons le comportement et les prédictions du modèle pour identifier les facteurs abiotiques essentiels qui déterminent la distribution générale des vers de terre. En nous concentrant sur chaque taxon, nous mettons en lumière ses préférences en matière d'habitat et sa réaction à divers gradients environnementaux. Nous explorons les points communs des réponses des espèces pour révéler des groupes de réponse. Troisièmement, nous analysons les associations identifiées par le modèle. Enfin, en utilisant le modèle cali-

bré, en considérant une grille d'environ 1km de résolution à travers la France, nous calculons des projections géographiques d'indicateurs calculés sur les communautés prédites comprenant : les catégories écologiques, les traits moyens pondérés de la communauté et la diversité des assemblages de vers de terre au niveau local (diversité alpha), régional (diversité gamma) et le turnover (diversité bêta), et ce dans ses trois aspects : taxonomique, fonctionnel et phylogénétique.

Contribution of data mining to ecological communities modeling

Sara SI-MOUSSI

A thesis to be presented for the degree of
Doctor of Philosophy





Doctoral school GAIA
Department of Functional Ecology and Agrosciences
L'Institut Agro

December 18th, 2020

"If the path be beautiful, let us not ask where it leads."

—Anatole France

Abstract

To anticipate the potential impact of environmental changes on ecological communities, it is crucial to understand the drivers of the geographic distribution of living organisms and their interactions. In biogeography, this problem is addressed by analyzing observations of species under various environmental conditions. In recent decades, large amounts of data on the occurrence of species have been made available in open databases (e.g. GBIF) along with environmental data describing climate, land use and soil conditions, etc.

All this provides a breeding ground for the development of statistical methods and computer tools applied to biodiversity data mining in order to (1) understand the factors that determine the observed distribution of living organisms; in order to (2) predict their response to changing environmental conditions. This thesis falls within this interdisciplinary context.

By adopting the community assembly and ecological niche theories as conceptual framework, we present a methodological framework for the analysis of multi-species occurrence data. Next, we illustrate its usefulness via three concrete applications.

First, we address the problem of recommending species according to geographical location, applied to vegetation data from the Pl@ntNet mobile app. We evaluate two alternative architectures enabled by our framework for this task: a convolutional neural network applied to environmental rasters only, and a species embedding architecture applied to animal occurrences. Second, we presented a novel framework for ecological association inference from co-abundance that supports asymmetric and bidirectional associations via a response-effect decomposition of dependency networks. We evaluated the proposal in comparison to state of art network inference methods on a simulation experiment. Then, we applied it to the analysis of potential interactions between alpine plants on a meso-topographic gradient of snow cover duration.

Third, we extended joint species distribution models to support rich features and non-linear responses using multi-task neural networks with latent variables to model species associations. We applied the model to the study of French earthworms and their associations, as well as to the prediction of the geographical distribution of earthworm diversity in France. Last, we illustrated on the earthworm application the use of machine learning interpretability tools to unravel the abiotic preferences captured by the model.

Résumé

Pour anticiper l'impact potentiel des changements environnementaux sur les communautés écologiques, il est essentiel de comprendre les facteurs contrôlant la répartition géographique des organismes vivants et leurs interactions. En biogéographie, ce problème est abordé en analysant les observations des espèces dans diverses conditions environnementales. Au cours des dernières décennies, de grandes quantités de données sur la présence d'espèces ont été mises à disposition dans des bases de données ouvertes (par exemple GBIF), ainsi que des données environnementales décrivant le climat, l'occupation du sol, etc.

Tout cela constitue un cadre propice au développement de méthodes de fouille de données de la biodiversité afin de (1) comprendre les facteurs qui déterminent la répartition observée des organismes vivants ; (2) prédire leur réponse aux conditions environnementales changeantes. Cette thèse s'inscrit dans ce contexte interdisciplinaire.

En adoptant les théories de l'assemblage des communautés et des niches écologiques comme cadre conceptuel, nous présentons un cadre méthodologique pour l'analyse des données d'occurrence multi-espèces. Ensuite, nous illustrons son utilité par le biais de trois applications concrètes.

En premier lieu, nous abordons le problème de la recommandation des espèces en fonction de leur localisation géographique, appliqué aux données de végétation de l'application mobile Pl@ntNet. Nous avons évalué deux architectures alternatives rendues possibles par notre cadre méthodologique pour cette tâche : un réseau neuronal convolutif appliqué aux rasters environnementaux uniquement, et une architecture à base d'embeddings appliquée aux occurrences animales.

En second lieu, nous présentons un nouveau cadre pour l'inférence d'associations asymétriques et bidirectionnelles à partir de co-abondances via une décomposition réponse-effet des réseaux de dépendance. Nous avons évalué notre proposition par rapport aux méthodes d'inférence de réseau les plus récentes dans le cadre d'une expérience de simulation, puis nous avons appliqué l'approche à l'analyse des comptages de plantes alpines sur un gradient méso-topographique.

En troisième lieu, nous avons étendu les modèles joints de distribution d'espèces afin de prendre en charge des descripteurs environnementaux riches et des réponses non linéaires à l'aide de réseaux neuronaux multitâches. Nous avons utilisé des variables latentes pour modéliser les associations d'espèces. Le modèle a été appliqué à l'étude des vers de terre de France et leurs associations, puis à la prédiction de la répartition géographique de leur diversité. Enfin, nous avons illustré sur le même cas d'étude l'utilisation d'outils d'interprétabilité pour démêler les préférences abiotiques capturées par le modèle joint.

List of publications

- Si-Moussi, S., Hedde, M., & Thuiller, W. 2019. Plant recommendation using the environment and biotic associations. In Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum.
- Si-Moussi, S. Galbrun, E., Hedde, M., & Thuiller, W. Uncovering bidirectional ecological associations from community and environmental data. [In preparation]
- Si-Moussi, S., (In collaboration with Galbrun, E., Thuiller, W. & Hedde, M.) (In prep). Ecological preferences and associations of French earthworms revealed by a deep multi-task distribution model. [In preparation]
- Bloor, JMG., Si-Moussi, S., Taberlet, P., Carrère P., & Hedde M. Complex soil trophic networks in contrasting agroecosystems show consistent responses to land-use intensification. [In preparation]

Table of Contents

Abstract 3

Résumé 4

List of figures 10

List of tables 11

List of Acronyms 12

Notation 14

1 Introduction 15
- 1.1 General context . . . 15
- 1.2 Theoretical context . . . 16
  - 1.2.1 Community assembly rules . . . 16
  - 1.2.2 Biotic interactions and ecological networks . . . 17
  - 1.2.3 The interplay between abiotic and biotic filters . . . 18
  - 1.2.4 Ecological niche . . . 18
  - 1.2.5 Patterns, processes and models . . . 18
  - 1.2.6 Traits and phylogeny . . . 19
- 1.3 Methodological context . . . 20
  - 1.3.1 Species distribution models . . . 20
  - 1.3.2 Ecological network learning . . . 23
- 1.4 Objectives and methodological contributions . . . 26
- 1.5 Thesis outline . . . 27

2 From GLMs to neural networks: learning and interpretability 29
- 2.1 Introduction . . . 29
  - 2.1.1 Machine learning problems . . . 29
  - 2.1.2 A probabilistic approach . . . 29
  - 2.1.3 Outline . . . 30
- 2.2 Generalized linear models . . . 30
  - 2.2.1 GLMs and the exponential family . . . 30
  - 2.2.2 Limits of linear models . . . 33
- 2.3 Feedforward neural networks . . . 34
  - 2.3.1 Multilayer perceptrons . . . 34
  - 2.3.2 Convolutional neural networks . . . 36
  - 2.3.3 Autoencoders and Variational autoencoders . . . 38
  - 2.3.4 Training feedforward neural networks . . . 39
- 2.4 Multi-task learning . . . 40
  - 2.4.1 Overview . . . 40
  - 2.4.2 Auxiliary tasks . . . 40

2.4.3 Multi-task neural networks . . . 40
2.4.4 Learning what to share . . . 41
2.5 Machine learning interpretability . . . 42
2.5.1 Overview . . . 42
2.5.2 Permutation feature importance . . . 43
2.5.3 Surrogate models . . . 44
2.5.4 Shapley values . . . 45
2.5.5 SHapley Additive exPlanations (SHAP) . . . 47
2.5.6 Response curves . . . 49
2.6 Conclusion . . . 53

3 Multi-task modeling of ecological communities 55
3.1 Introduction . . . 55
3.2 Generic framework . . . 55
3.2.1 Context builder . . . 56
3.2.2 Feature extraction . . . 57
3.2.3 Prediction . . . 58
3.2.4 Opening the black-box . . . 58
3.2.5 Incorporating expert knowledge . . . 59
3.3 Multi-task neural model of the joint environmental response . . . 61
3.3.1 Generative model . . . 61
3.3.2 Inference as optimization . . . 61
3.3.3 Sparse associations estimation . . . 63
3.3.4 Extending Joint Species Distribution Models . . . 64
3.4 Conclusion . . . 64

4 Location-based recommendation 65
4.1 Introduction . . . 65
4.2 Dataset description . . . 65
4.3 Task and proposals . . . 66
4.3.1 GrinnellNet: a CNN with categorical rasters embedding . . . 66
4.3.2 EltonNet: a species embedding network . . . 69
4.4 Training and evaluation . . . 70
4.4.1 Optimization and evaluation metrics . . . 70
4.4.2 Implementation and learning setting . . . 71
4.4.3 Evaluation . . . 71
4.5 Conclusion . . . 72

5 Ecological association inference 75
5.1 Introduction . . . 75
5.1.1 Objectives and outline . . . 76
5.2 The framework . . . 76
5.2.1 Spatial associations and species embeddings . . . 77
5.2.2 A conditional generative model of abundance . . . 78
5.2.3 Inference and model selection . . . 80
5.2.4 Unraveling inter-specific association networks . . . 80
5.3 Simulation experiment: process-based assembly with additive drivers . . . 82
5.3.1 Methods . . . 82
5.3.2 Results . . . 84
5.4 Simulation experiment: predator-prey co-occurrences . . . 87
5.4.1 Methods . . . 87
5.4.2 Results . . . 90
5.5 Empirical case study . . . 90
5.5.1 Data preparation . . . 92

5.5.2 Statistical analysis . . . 92
5.5.3 Results . . . 93
5.6 Discussion . . . 94
5.6.1 Disentangling abiotic and biotic drivers . . . 95
5.6.2 From species representations to biotic associations . . . 96
5.6.3 Perspectives . . . 96
5.7 Conclusion . . . 97
5.8 Supplementary Materials . . . 98
5.8.1 Virtualcomm how-to . . . 98

6 Earthworms biogeography and diversity **101**
6.1 Introduction . . . 101
6.2 Materials and Methods . . . 103
6.2.1 Data acquisition and processing . . . 103
6.2.2 Joint model of earthworms distribution . . . 105
6.2.3 Abiotic drivers of earthworms distribution . . . 108
6.2.4 Projection to current conditions . . . 109
6.3 Results . . . 109
6.3.1 Predictive performances . . . 109
6.3.2 Habitat suitability . . . 110
6.3.3 Earthworm association networks . . . 115
6.3.4 Mapping earthworm communities . . . 118
6.4 Conclusions and future directions . . . 122
6.4.1 Improving the predictive performances . . . 122
6.4.2 Extending the association model . . . 123
6.4.3 Earthworm habitat suitability . . . 123
6.4.4 Characterizing earthworm communities . . . 124
6.4.5 Future directions . . . 125
6.5 Supplementary Materials . . . 126
6.5.1 Dataset . . . 126
6.5.2 Model and preprocessing selection . . . 131
6.5.3 Full predictive performances . . . 132
6.5.4 Drivers of earthworm distribution . . . 134
6.5.5 Measuring the diversity of an ecological community . . . 138

7 Discussion **141**
7.1 Advancing network inference . . . 141
7.1.1 Theoretical advances and outstanding questions . . . 141
7.1.2 Towards bridging network learning paradigms . . . 142
7.2 Advancing joint species distribution models . . . 142
7.2.1 Benefits and limits of neural architectures . . . 142
7.2.2 Contribution of multi-task learning and improvement perspectives . . . 143
7.3 Prospective applications . . . 144

List of Figures

1.1 Patterns, processes and models in community ecology . . . 16
1.2 Classes of long-term interactions (symbioses). For a pair of species (i, j), $a_{i,j}$ and $a_{j,i}$ represent the outcome of the interaction on j and i respectively. [Morales-Castilla et al., 2015] . . . 17
1.3 BAM diagram . . . 19
1.4 Network learning methods classification . . . 24

2.1 GLM parameterization . . . 31
2.2 Examples of feedforward neural architectures. Computation layers are arranged into a directed acyclic graph (i.e. involve no feedback loops). . . . 34
2.3 Multilayer perceptron . . . 35
2.4 Convolutional Neural Network . . . 36
2.5 Convolution padding types . . . 37
2.6 Layer complexity. . . . 38
2.7 Standard autoencoders vs variational autoencoders. . . . 38
2.8 Hard (left) vs soft (right) parameter sharing in multi-task neural networks. . . . 41
2.9 Process of machine learning interpretability. . . . 42
2.10 LIME tabular data. . . . 46
2.11 Marginal expectation vs conditional expectation in PDP and ALE plots. . . . 50
2.12 Accumulated Local Effects plot for a single numerical feature. . . . 51
2.13 Accumulated Local Effects plot for two interacting numerical feature. . . . 52

3.1 Architecture of MTEC. . . . 56
3.2 Interpretability methods for multi-task models. . . . 60
3.3 Learning architecture of the joint multi-task neural network model. . . . 64

4.1 GrinnellNet architecture . . . 67
4.2 Non plants taxa proportion in the retained list. . . . 70
4.3 EltonNet architecture. . . . 71
4.4 Performance of GrinnellNet and EltonNet variants on validation and test set. . . . 72

5.1 Symmetric and asymmetric association types. . . . 77
5.2 A graphical illustration of the interplay between the habitat and biotic associations. . . . 79
5.3 Abiotic and biotic response aggregation. Illustrated examples of potential applications. . . . 81
5.4 Schematic representation of the virtual experiment. . . . 83
5.5 Simulation convergence diagnosis. . . . 85
5.6 Fitted association strengths per association type and simulation configuration. . . . 86
5.7 Association strength and niche differences. . . . 87
5.8 Association type classification performances. . . . 88
5.9 Food web simulation. . . . 89
5.10 Asymmetric association network inference quality. . . . 90
5.11 Habitat Suitability Model variable importance and prediction performances per genus. . . . 93
5.12 Plant associations on an Alpine mesotopographic gradient. . . . 99

5.13 Plant traits and embeddings mutual information. . . . 100
5.14 Virtualcomm simulation procedure. . . . 100

6.1 Geographic coverage of the calibration and evaluation dataset. . . . 104
6.2 Earthworm phylogenetic tree. . . . 104
6.3 MTEC, SSDMs performance ranking tests. . . . 110
6.4 Feature importance for earthworm distribution model. . . . 111
6.5 Limiting temperature factors for common earthworms. . . . 112
6.6 Response to precipitation. . . . 113
6.7 Response to land cover. . . . 115
6.8 Response to edaphic properties. . . . 116
6.9 Organic matter content preferences and typhlosolis types of earthworms. . . . 117
6.10 Earthworm association network. . . . 118
6.11 Ecological categories and land cover distribution map. . . . 119
6.12 Climate conditions during the Last Glacial Maximum. . . . 120
6.13 Earthworm taxonomic, functional and phylogenetic diversity maps. . . . 121
6.14 Geographic and taxonomic coverage of the occurrence dataset. . . . 126
6.15 Geographic and taxonomic (species) coverage of the evaluation dataset. . . . 126
6.16 Earthworm MTEC architecture . . . 131
6.17 Earthworm habitat suitability prediction scores. . . . 132
6.18 Earthworm global response groups. . . . 134
6.19 Mantel test of the phylogenetic and functional structure of earthworms abiotic responses. . . . 135
6.20 Earthworm preferred tree traits . . . 137
6.21 Principal component analysis biplot of tree traits in forest sites sampled by Bouché [1972] . . . 138

List of Tables

2.1 Generalized linear models parameterization for common distributions 32
2.2 Common regularization penalties and formulation in terms of prior over the parameters. . . . 33

5.1 Description of the evaluated frameworks and their respective configuration. 84
5.2 Summary performances of inference on food webs. 91
5.3 Embedding similarity and niche distance. 91

6.1 List of features and their source for Earthworm modeling task. 105
6.2 Average predictive performances on earthworms. 110
6.3 Median of ecological categories assignment probabilities amongst observed taxa in various land cover types. 114
6.4 Earthworm MTEC grid search. 132
6.5 Functional traits of observed trees in forest sites. 136

List of Acronyms

AE Autoencoders.

ALE Accumulated Local Effects.

BAM Biotic, Abiotic, Movement diagram.

BI-SDM Biotic Interactions Species Distribution Models.

BN Bayesian Networks.

BP Back-Propagation algorithm.

BRT Boosted Regression Trees.

CART Classification and Regression Trees.

CG Covariance Graphs.

CNN Deep Neural Networks.

CSA Conspecifics Spatial Aggregations.

DN Dependency Networks.

DNN Convolutional Neural Networks.

DR-JSDM Dimensionality Reduction Joint Species Distribution Models.

ENL Ecological network learning.

ES Evaluation Strip.

GAM Generalized Additive Model.

GGM Gaussian Graphical Models.

GLLVM Generalized Linear Latent Variable Model.

GLM Generalized Linear Model.

GLMM Generalized Linear Mixed effects Model.

HGLMM Hierarchical Generalized Linear Mixed effects Model.

HSM Habitat Suitability Models.

IML Interpretable Machine Learning.

ISA Interspecific Spatial Associations.

JSDM Joint Species Distribution Models.

LA Laplace Approximation.
LIME Locally Interpretable Model-agnostic Explanations.
LV-JSDM Latent Variable Joint Species Distribution Models.
LVM Latent Variable Model.

MAP Maximum A Posteriori.
MCMC Markov-Chain Monte-Carlo.
MLE Maximum Likelihood Estimate.
MLP Multi-Layer Perceptrons.
MRF Markov Random Fields.
MSDM Multiple Species Distribution Models.
MTEC Multi-Task modeling of Ecological Communities.
MTL Multi-Task Learning.

NN Artificial Neural Networks.

PCA Principal Component Analysis.
PDP Partial Dependence Plots.
PGM Probabilistic Graphical Models.
PIMP Permutation feature IMPortance.

RF Random Forest.

SAM Species Archetype Models.
SDM Species Distribution Models.
SGD Stochastic Gradient Descent.
SGVB Stochastic Gradient Variational Bayes.
SHAP SHapley Additive exPlanations.
SSDM Stacked Species Distribution Models.

VAE Variational autoencoders.
VI Variational Inference.

Notation

Probability distributions

$\mathcal{B}$	Bernoulli distribution
$\mathcal{F}$	distribution from the exponential family
$\mathcal{N}$	Normal or Gaussian distribution

Functions and operators

$\cdot$	matrix/dot product
δ	Dirac point-mass
$\perp\!\!\!\perp$	independence
$\odot$	element-wise multiplication
Φ^{-1}	probit function, inverse of the cumulative distribution function of the standard normal distribution
σ	logistic function

Parameters

W, w	abiotic feature extraction weights matrix, resp. vector
B, β	abiotic response weights matrix, resp. vector
C	correlation matrix
Σ	covariance of the multivariate normal distribution
ϕ	dispersion parameter
A, α	factor loadings matrix, resp. vector
Ω	inverse covariance/precision matrix
μ	mean
η	natural parameter
ρ	partial correlation matrix
σ^2	variance of the univariate normal distribution
τ	sufficient statistic

Data matrices and vectors

Y, y	response matrix, resp. vector
E, e	raw environmental covariates matrix, resp. vector
X, x	extracted features matrix, resp. vector
H, h	latent factors matrix, resp. vector
Z, z	latent responses matrix, resp. vector
T, t	trait matrix, resp. vector

Chapter 1

Introduction

1.1 General context

Biodiversity, the variability of living organisms and their complex organizational structures [Gaston and Spicer, 2013], supports ecosystems functioning and provides significant contribution to people [Haines-Young and Potschin, 2010, Díaz et al., 2018]. These services include supplying renewable resources (food, wood, medicines, etc.), regulating environmental threats (pests/diseases, erosion control, etc.), in addition to providing cultural services for the recreational and spiritual enjoyments of human societies [Barrios, 2007].

Currently, we are witnessing a deterioration of natural (and also human-modified) habitats driven by global changes, i.e. over-exploitation of natural resources, climate change, pollution, and the introduction of invasive species (IPBES 2019 global assessment report [Brondizio et al., 2019]). This degradation of ecosystems affects biodiversity at various levels of organization starting from the individual organisms, to groups of individuals (species' population), to the collection of two or more interacting species populations (community) and their surrounding physical environment (ecosystem) [Brodie et al., 2014].

To anticipate the consequences of global change on biodiversity, ecological research strives towards a deeper understanding of the interactions between organisms and with their physical environment [Thuiller et al., 2013]. In biogeography, this problem is addressed by analyzing observations of the presence or absence of species (or abundances) under various environmental conditions [Humboldt et al., 1805] and relating them to the ecology and evolution of organisms.

Over the last couple of decades, large amounts of data on species occurrences have been made available in open databases (e.g. GBIF [Lane and Edwards, 2007]). These data originated from research programs (e.g. soil organisms occurrences: Edaphobase [Burkhardt et al., 2014]), or from citizen science applications (e.g. PlantNet [Goëau et al., 2013], iNaturalist [Nugent, 2018]). In addition, increasingly precise environmental data are becoming available in the form of rasters describing climate (e.g. Worldclim [Fick and Hijmans, 2017], CHELSA [Karger et al., 2016]), land use (e.g. Corine Land Cover [Cover, 2000]), physical and chemical characteristics of soils (e.g. ESDB [Panagos, 2006], ISRIC [ISRIC, 2013]), but also raw high-resolution satellite pictures and multi-sensor imagery (e.g. Landsat [Woodcock et al., 2008], sentinel [Schmitt et al., 2019]).

All these data provides a breeding ground for the development of statistical and computational methods to mine biodiversity data allowing in turn to understand the factors that determine the observed distribution of living organisms, and predict their response to changing environmental conditions. My thesis work falls within this interdisciplinary context.

In the following, section 2 introduces the theoretical foundations of community ecology. Section 3 reviews the current advances in community modeling, specifically species distributions models (3.1) and ecological network learning (3.2). Section 4 summarizes the thesis objectives and main contributions. Finally, section 5 outlines the structure of this document.

1.2 Theoretical context

A community is an assemblage of populations of organisms living together in the same area or habitat, such that spatial boundaries are either natural (islands) or arbitrary (study plots) [Fauth et al., 1996, Krebs, 1972]. Community ecology examines how interactions between species in a community and their environment affect the resulting species diversity and abundances [Johnson and Stinchcombe, 2007].

1.2.1 Community assembly rules

The assembly rules framework which stems from the seminal work of Diamond [1975] and depicted in Fig 1.1 explains the composition of communities at different spatial scales by a set of processes operating as filters that can explain the local abundance of a species [Weiher and Keddy, 2001]. Globally, **phylogeographic processes**, including evolutionary history and large-scale migration, restrict the pool of potential species regionally. At regional and local scales, three **ecological processes** co-determine whether a species from the restricted pool can be present at a given site. First, the abiotic conditions should enable the species population to satisfy its physiological needs in order to survive, grow and reproduce. This is referred to as habitat suitability, it constitutes the **abiotic filter**. Second, the net impact of direct and indirect biotic interactions in which a species is involved should allow it to sustain its populations. This is referred to as the **biotic filter** [Guisan et al., 2017, Ovaskainen and Abrego, 2020]. Third, the **dispersal or movement filter** requires the site to be accessible from a previous location. This depends both on the dispersal capacity of the species but also on the existence of corridors or, on the contrary, barriers to migration. A distinction is made between active (unassisted) and passive dispersion (via animals, wind, rivers or human activity). Sometimes, populations go locally extinct in response to stochasticity. The dispersal capacity of the species controls how fast it can recolonise the empty patch. Finally, some species may attain high abundance in unsuitable "sink" sites due to frequent immigration from neighbouring suitable "source" sites, a phenomenon known as **source-sink dynamics** [Pulliam, 1988].

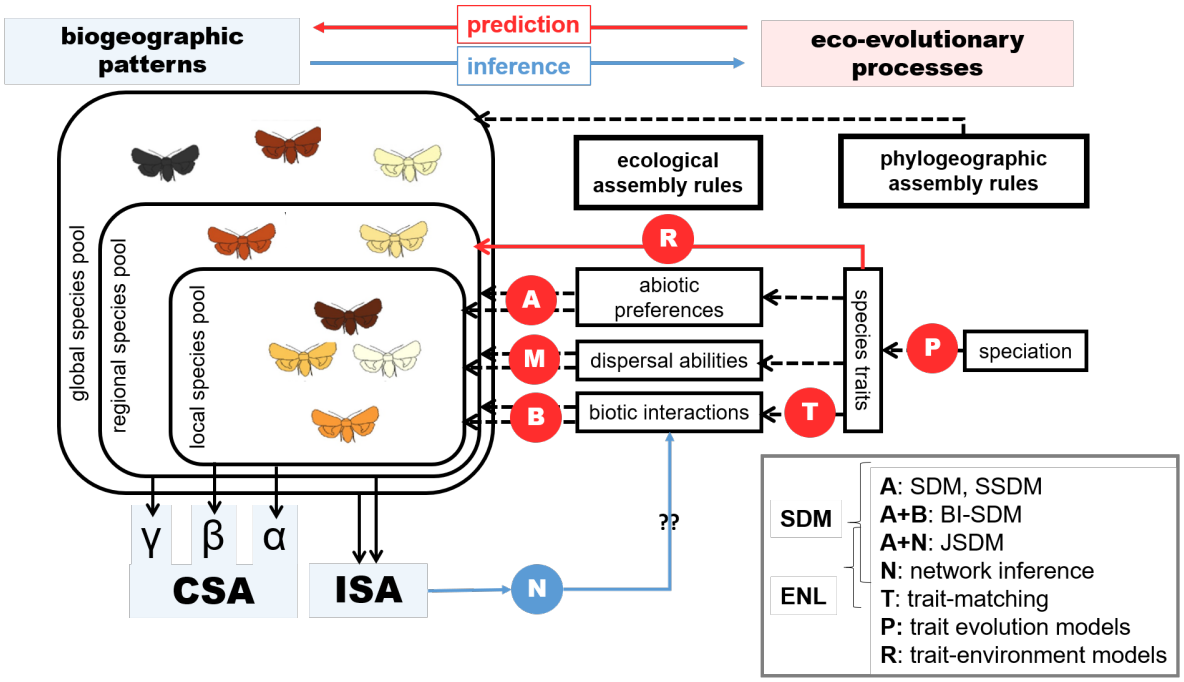


Figure 1.1: The assembly rules framework and models for learning eco-evolutionary assembly processes from biogeographical patterns. Schema adapted from Ovaskainen and Abrego [2020]. CSA: Conspecifics Species Aggregations, ISA: Interspecific Spatial Associations, SDM: Species Distribution Models, ENL: Ecological Network Learning.

1.2.2 Biotic interactions and ecological networks

Organisms co-existing in a community may affect each other through biotic interactions. Interactions are either **intraspecific**, if they occur between individuals of the same species (e.g reproduction, cooperation, and competition), or **interspecific** if they involve different species (e.g the pollination of a flower by a bee). Haskell [1949] distinguishes **short-term** and long-term interactions. The former include trophic or predator-preys and plant-pollinators interaction. The latter, also known as **symbioses**, can be of different types depending on whether their outcome is beneficial for one (commensalism) or both partners (mutualism), harmful for one (ammensalism) or both partners (competition), or beneficial for one and harmful for the other (parasitism), as depicted in Fig 1.2.

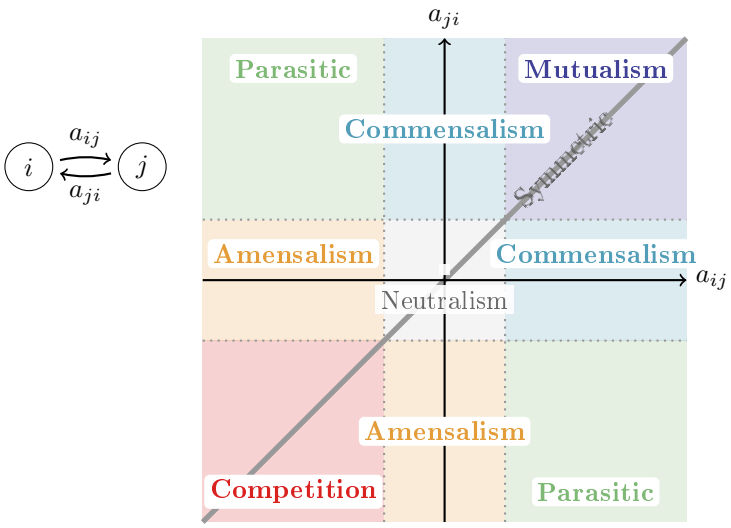


Figure 1.2: Classes of long-term interactions (symbioses). For a pair of species (i, j), $a_{i,j}$ and $a_{j,i}$ represent the outcome of the interaction on j and i respectively. [Morales-Castilla et al., 2015]

It is customary to use networks as a tool to represent the direct interactions taking place in an ecosystem, yielding **ecological networks**. Food webs are a special type of ecological networks restricted to trophic interactions i-e to consumer–resource relationships. Networks are important ecological features to understand the functioning of ecosystems [Thompson et al., 2012] and their resilience to disturbances [Poisot et al., 2016, Domínguez-García et al., 2019]. They are also useful to make projections on future communities' composition under climate or land use change scenarios [Clark et al., 2020, Woodward et al., 2010].

In addition to the direct effects resulting from physical interactions, interaction networks also highlight indirect effects. Indirect interactions emerge in multi-species assemblages when "*the impact of one species on another requires the presence of a third species*" [Wootton, 1994]. They operate through two possible modes: interaction chain or interaction modification. The classic example of interaction chain is the trophic cascade: when a predator feeds on a herbivore, it reduces the consumption pressure of the latter on the vegetation, thus the predator has a positive effect on the vegetation. Interaction modification occurs when the presence or activity of one species alters the probability or outcome of an interaction between two other species. For example, vegetation growth offers shelter from predators, thus reducing the probability of the prey consumption by the predator. Also referred to as higher-order interactions, interaction modification operates via several mechanisms: behavioral, chemical, morphological or response modification [Wootton, 1994]. Both modes of indirect interactions have been shown to promote the stability and resilience of communities [Grilli et al., 2017].

1.2.3 The interplay between abiotic and biotic filters

Species can influence each other indirectly through habitat modification. For example, the presence of cushion plants in dry alpine environments creates a moist micro-habitat, facilitating the establishment of drought-intolerant species. This process is a form of **facilitation**. Similarly, earthworms consume soil and release casts rich in mineral salts and humus and enriched with bacteria that make nutrients quickly available to plants. By their movement in the soil and through geophagy (soil consumption), earthworms create underground galleries (soil bioturbation) that allow water infiltration and vertical root development. They are therefore considered as soil **ecosystem engineers** along with ants and termites. Habitat modification can also be negative. For instance, the high acidity of some trees litter (e.g *Picea*) suppresses non acid-tolerant litter-dwelling earthworms [Schelfhout et al., 2017].

Furthermore, abiotic conditions may potentially change the probability of interaction. For instance, increasing temperatures from global warming alter the metabolism of organisms and consequently their life cycles, which can lead to phenological mismatches between potentially interacting species as observed in plant-pollinator networks [Gérard et al., 2020]. Abiotic conditions can also change the type and/or magnitude of interactions. For example, the stress-gradient hypothesis (SGH) predicts that the frequency of facilitative and competitive interactions will vary inversely along abiotic stress gradients, with facilitation being more common in conditions of high abiotic stress relative to more benign abiotic conditions [Maestre et al., 2005]. The SGH has been supported by numerous studies in many ecosystems, including Alpine plants along elevation gradients [Callaway et al., 2002] and crops associations under increasing soil phosphorus contents [Boudsocq et al., 2016].

1.2.4 Ecological niche

The niche of a species represents its position in an ecosystem, defined by its ecophysiological requirements or habitat preferences and its adaptive capacities (abiotic or Grinnellian niche [Grinnell, 1917]), as well as its role within the community and its place in the interaction network (Eltonian niche [Elton, 2001]). Hutchinson [1957] formalized the abiotic niche of a species as an n-dimensional hypervolume in the multivariate space of environmental and resource characteristics. The fundamental niche comprises all abiotic environmental conditions where a species can survive indefinitely, meaning where it has a positive population growth. The realised niche is what remains of this hypervolume after accounting for biotic effects (e.g. facilitation, competition) [Bruno et al., 2003, Lortie et al., 2004]. The combined influence of all eco-evolutionary processes produces the observed composition of the communities (observed niche).

From the perspective of a given species of interest, Fig 1.3 illustrates the possible scenarios of observation in a prescribed geographic area. The fact that a species can be observed in locations where not all its ecological requirements are met highlights the potential complementary nature of ecological processes: habitat suitability, biotic interactions and dispersal and the non-linearity of their combined effects.

1.2.5 Patterns, processes and models

Individuals of the same species cluster spatially into populations of varying densities, forming *Conspecific Spatial Aggregations* (CSA) [Keil et al., 2019]. The resulting spatial patterns reflect the realised niche, intraspecific interactions and dispersal patterns. CSA motifs are used to quantify the biodiversity on a regional (gamma) or local (alpha) scale and its turnover between sites (beta). Populations of species with similar environmental preferences or dependent on each other through interactions assemble into communities or on the contrary dissociate spatially (in case of competitive exclusion for example) spatially, producing patterns of co-occurrence or co-abundance called *Interspecific Spatial Associations* (ISA) [Keil et al., 2019]. ISA motifs reflect niche overlap and interaction networks. Several metrics exist to quantify ISA based on co-occurrence or co-abundance (cf. review in [Keil et al., 2019]).

The objective of species community modeling is to analyze biogeographic patterns (CSA, ISA) at the observed sites and across scales, to infer the assembly processes governing the distribution of modeled species and then use these to predict community composition on new conditions. The Grinnellian and Eltonian niches explain the species' response to abiotic and biotic filters respectively. Therefore, inferring these two filters

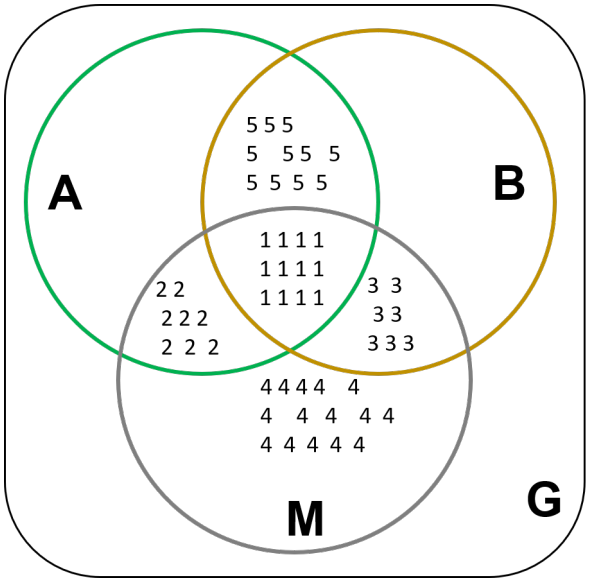


Figure 1.3: BAM diagram [Soberón and Nakamura, 2009] highlighting the interplay between the three drivers of species distribution: abiotic or habitat suitability, biotic interactions, and dispersal/movement. For a given species in a given geographic area of interest, denoted as G, at a given time, we denote by M the subset of sites that are accessible to the species (colonizable by movement/dispersal), by A the subset of sites with suitable habitat or abiotic conditions (fundamental niche), and by B the subset of sites with suitable biotic conditions. The species may be observed on the field in the configurations marked as 1–4 resulting from different intersections of the previous subsets. 1: realized niche (suitable with regard to all factors), 2: suitable habitat but unsuitable biotic conditions (e.g competition). 3: Colonization outside the suitable environment, maybe due to facilitation (sink). 4: Sink in unsuitable biotic and abiotic conditions, maybe due to historical presence. A further configuration (5) corresponds to a potential distributional area that is not accessible within the time-frame of interest, it is an invadable area. Schema adapted from [Guisan et al., 2017]

is equivalent to learning the Grinnellian and Eltonian niches. Learning the former is the target of Habitat Suitability or Species Distribution Models (HSM / SDM), learning the latter is the target of Ecological Network Learning (ENL) models.

1.2.6 Traits and phylogeny

A functional trait is "a morphological, physiological or phenological characteristic of an organism measured at the individual level that affects its individual performance" [Violle et al., 2007]. The functional trait approach considers that the response of species to different filters is explained by their traits. Several groups of traits can be distinguished. **Response traits** [Lavorel and Garnier, 2002, Suding et al., 2008] influence the response of organisms to abiotic factors. **Effect traits** [Lavorel and Garnier, 2002, Suding et al., 2008] modulate the contribution of species to ecosystem functions. **Dispersal traits** measure the ability of the species to actively disperse [Schleuning et al., 2020]. Finally, interaction or **matching traits** [Bartomeus et al., 2016, Gravel et al., 2016] determine the potential partners of a species for a given type of interaction.

A trait may be significantly correlated to the probability of occurrence of a species even if it has nothing to do with it. This occurs because phylogenetically related species share patterns of occurrence and traits, some of which are effectively response traits and others are legacies of ancestral traits. This phenomenon is called niche conservatism and results in **phylogenetic attraction** [Münkemüller et al., 2015]. Paradoxically, the principle of competitive exclusion [Hardin, 1960] in community ecology states that species that consume similar resources and have similar habitat preferences are mutually exclusive, inducing **phylogenetic repulsion**

[Helmus et al., 2007].

The functional traits paradigm allows us to free ourselves from taxonomy and to get closer to the response mechanisms of organisms by linking them to the processes of adaptation and speciation at the origin of the evolution of these traits. Moreover, the models created can be applied to species other than those modeled, including rare species. Nevertheless, existing approaches still lack genericity because the traits measured vary between large taxonomic groups and are often incomparable in multi-trophic assemblages. Databases of traits are made available for different groups: soil invertebrates (BETSI [BETSI, 2012]), plants (TRY [Kattge et al., 2011]), fungi (FunGUILD [Nguyen et al., 2016]), nematodes (NemaGUILD [Nguyen et al., 2016], Nemaplex [Ferris, 1999]), etc. Since it is impossible to measure all relevant traits, phylogeny is sometimes used to approximate unmeasured traits [Münkemüller et al., 2015, Desjardins-Proulx et al., 2017, Laigle et al., 2018]. Phylogenetic analyses are also used to infer missing values of partially observed traits using trait evolution models [Paradis, 2011].

1.3 Methodological context

In this thesis, we refer to the modeled entities as species although the same described approaches can be applied to infra- (subspecies, variety) and supra-species levels (genus, family). In addition, we do not impose any phylogenetic or taxonomic or functional restrictions on the composition of sampled communities.

We refer to observations as sites. In general, we will assume sites are independent, unless stated otherwise. This assumption might be invalidated for instance in cases of nested sampling designs, spatial auto-correlation where distance between sites is accounted for, and temporal auto-correlation where the same site is sampled at multiple points in time.

A typical community dataset comprises the observation of presence/absence or abundance (count of individuals, biomass) or relative abundance of M species, in N sites. The dataset may also include a S-by-T matrix of traits for each species. Traits can be discrete (binary or categorical) or continuous. The dataset can also include a phylogenetic distance matrix S-by-S between all or a subset of the modeled species.

Community data contain the footprint of eco-evolutionary processes. Inferring these processes from the data is one of the most active topics of research at the intersection of biogeography and community ecology. Focusing on habitat suitability and biotic interactions, we hereafter review models used to tackle these processes separately and jointly.

1.3.1 Species distribution models

General assumptions

Habitat Suitability or Species Distribution Models (HSM/SDM) seek to identify statistical relationships between the occurrence or abundance of a species and the abiotic factors describing its habitat/environment. Concretely, SDMs relate species' observations to environmental covariates using a combination of response curves that describes the species' realized niche [Guisan et al., 2017].

As simplified depictions of the real world, SDMs rely on strong assumptions that should be taken in consideration while using these models to answer basic and applied questions [Guisan et al., 2017]. First, species-environment relationships are assumed to be at equilibrium or pseudo-equilibrium, meaning that they will not change across space or time. The species is expected to have colonized most of its suitable habitats in the studied area. There are situations where this assumption is violated, such as when modeling invasive species that are still undergoing expansion, in which case such species' niches are modeled using their native range or the union of the invasive and the native ranges [Gallien et al., 2010]. Second, all important environmental predictors required to describe the species niche are assumed to be available, at a relevant resolution and with no measurement error. Third, a suitable measure of composition (occurrence, abundance, frequency) for the study of interest with an appropriate probability distribution must be employed. Fourth, observations/sites are assumed to be independent. Last, the sampling design of species occurrences is assumed to be objective and unbiased. For instance, in citizen science data occurrences tend

to be concentrated along transportation axes (roads, hiking trails) and public spaces (parks) making this assumption invalid and hence precluding the direct use of these models.

Statistical tools for single species distribution models

Species distribution modeling is possible via a variety of statistical and machine learning methods [Guisan et al., 2017], including: **regression based approaches** (e.g. Generalized Linear Models (GLM) [McCullagh, 81], Generalized Additive Models (GAM) [Hastie and Tibshirani, 1990], Multivariate Adaptive Regression Splines (MARS) [Friedman, 1991]), **machine learning and classification methods** (Support Vector Machines (SVM) [Hearst et al., 1998], Artificial Neural Networks (NN) [Goodfellow et al., 2016], Classification and Regression Trees (CART) [Breiman et al., 1983], **boosting** (e.g Boosted Regression Trees (BRT) [Elith et al., 2008]) and **bagging** approaches (e.g Random Forests (RF) [Breiman, 2001]), **Maximum ENTropy** [Phillips et al., 2006] in addition to **ensemble models** [Thuiller et al., 2009]. The availability of advanced tools and software (e.g `biomod` [Thuiller et al., 2009], `dismo` [Hijmans et al., 2017], `ENMtools` [Warren et al., 2010], etc.) disseminated the use of SDMs in various applications in ecology and conservation. Recently, a standardized protocol for SDMs was proposed [Zurell et al., 2020a].

Multi-species distribution models

At the community or assemblage level, two different approaches are used to predict spatial patterns of species assemblages. The first approach (macroecological modelling, MEM) [Gotelli et al., 2009] focuses directly on realized properties of species assemblages, whereas the second approach (stacked species distribution modelling, S-SDM) starts with constituent species to approximate the properties of assemblages [Schmitt et al., 2017]. The properties of species assemblages include the number of co-occurring species (richness), inter-specifc abundance patterns, and compositional, functional and structural community characteristics. SESAM [Guisan and Rahbek, 2011] was proposed as a unifying framework of both approaches, whereby MEMs are used to constrain S-SDMs. For instance, d'Amen et al. [2015] used predictions of richness and community traits averages to constrain S-SDMs predictions of mountain grassland communities in the Western Swiss Alps.

The advantage of S-SDM is a flexibility in terms of model choice for each species. However, this approach still requires calibrating several distinct models thus multiplying sources of uncertainty and error as well as the computational overhead, although the latter can be solved using parallel computation. It is sometimes reasonable to assume the input-output mapping (i-e the environmental predictors involved and the shape of response curves) is similar across different species models, so better performances can be obtained by fitting all the parameters simultaneously. This setup is known as multi-task learning (MTL) [Caruana, 1993]. Hierarchical Bayesian regression models [Bakker and Heskes, 2003], particularly Hierarchical Generalized Linear Models, are the most common method adopted in the context of multi-species distribution models [Ovaskainen and Abrego, 2020].

When abundance data on natural populations are collected, two patterns are often observed. First, the most common species accounts for a substantial fraction of the individuals sampled. Second, a substantial fraction of the species sampled are very rare. These patterns are omnipresent in community data, leading to long-tailed distributions of species' prevalences. Thus, it might lead to excluding rare species when modelling lots of species. At the same time, it is not desirable to use the same model for all species because some species may have very different response to the environment. As a compromise, Hierarchical Bayesian regression models fit dedicated parameters for each species, but encourage parameters to be similar across species by drawing the parameters from a common distribution. In this way, rare species borrow statistical strength from prevalent species (assuming that rare species respond in a similar way to the environment than common species). In the same setting, functional traits have been used to explain the regression weights of species to each environmental predictor [Pollock et al., 2012]. For instance, Carboni et al. [2018] highlighted using this approach the role of traits on the invasion success of alien herbaceous plants in French permanent grasslands. Additionally, phylogenetic distances have been used to explain the covariance of species responses, as a model-based approach to test for phylogenetic conservatism [Ives and Helmus, 2011].

Dunstan et al. [2011] and [Hui et al., 2015] have proposed using finite mixture of regression (FMR) for

multi-species distribution modeling, whereby species are clustered based on their environmental responses into a small number of "archetypal responses." Archetypes can be interpreted as abiotic response groups. These models are referred to as Species Archetype Models (SAM). When compared to SSDMs [Hui et al., 2013], SAMs improved model accuracy and discriminatory capacity, especially for rare species as they borrow strength from common species for which more data is available.

Biotic interactions in species distribution models

Although species distribution models are fitted to the realized niche, which results of both abiotic and biotic filtering, the models previously described often overlook the effect of biotic interactions. The relevance of this problem depends on the scale of study. Indeed, some studies [Thuiller et al., 2015, Araújo and Rozenfeld, 2014, Mod et al., 2020, Zurell et al., 2018] found that at sufficiently large spatial scales or coarse resolution, the effect of biotic interactions becomes negligible, especially competition. Other studies highlighted the impact of interactions on species ranges at large spatial scales [Gotelli and McCabe, 2002, Wisz et al., 2013], thus advocating for their inclusion in SDMs. BI-SDMs (Biotic Interactions SDM, *sensu* [Kissling et al., 2012], [Dormann et al., 2018]) account for biotitc interactions by including as a predictor in the model the abundance/occurrence of other species [Araújo et al., 2011, Araújo and Luoto, 2007], by restricting the species' predicted distribution with that of interacting species it depends on (e.g. prey, predators, parasites) [Schweiger et al., 2012], or more generally by refining SDMs predictions with biotic interactions [Staniczenko et al., 2017]. However, BI-SDMs require knowledge of the interactions in which the focal species is involved, and a measure of the abundance of the interacting partners a every site, in order to allow predicting its occurrence. Furthermore, these approaches only incorporate unidirectional interactions and do not take into account the reciprocal effect of the focal species on the others [Kissling et al., 2012, Pollock et al., 2014].

Joint species distribution models

Joint species distribution models (JSDMs) have emerged as extensions of SDMs that simultaneously model multiple species' distributions, accounting for both environmental relationships and residual associations that might arise from species interactions. Thus, they inform about both abiotic and biotic filters [Pollock et al., 2014] and can potentially improve the predictions of community composition [Harris, 2015].

Promising advances in JSDMs have led to the development of several statistical models and computational implementations. JSDM implementations are essentially based on Generalized Linear Mixed Effect Models (GLMM) [Ovaskainen and Abrego, 2020] combining fixed effects capturing the response to the abiotic predictors, with random effects capturing residual correlations between species. The core statistical differences between different JSDMs can be broadly explained by three methodological choices [Wilkinson et al., 2019]: (1) the inclusion of a hierarchical structure on the regression coefficients, (2) the use of either the full residual covariance or its decomposition into latent factors (LV-JSDM), (3) the use of dimension reduction (DR-JSDM).

Latent factors are interpreted as hypothetical unmeasured environmental variables impacting species co-occurrence patterns. From a computational perspective, latent factors allow the reduction of model complexity as they require substantially fewer parameters than classic GLMM with full-covariance. DR-JSDM allow further complexity reduction by clustering species residual correlations (into clusters akin to biotic response groups) in contrast to SAMs which cluster species on the basis of their abiotic responses (abiotic response groups).

However, current JSDM implementations present several limitations.

On one hand, **JSDMs did not show much improvement in marginal predictive performances compared to univariate GLMs** [Zurell et al., 2020b, Norberg et al., 2019]. This is due to the fact that marginal predictions are based solely on environmental attributes. This explains why different JSDMs tended to have similar predictive performances [Pichler and Hartig, 2020]. In theory, per the likelihood formulation, the estimation of the residual covariance does not affect the estimation of regression coefficients of the environmental response [Poggiato et al., in revision, Wilkinson et al.], i-e the abiotic niche estimation. This statement is backed by several empirical results reporting similar regression coefficients regardless of

the covariance modeling strategy (full-covariance, latent variable, dimension reduction) [Wilkinson et al., 2019, Pichler and Hartig, 2020].

On the other hand, **empirical comparisons of JSDMs to SDMs (aside from GLMs) showed either similar to inferior marginal predictive performances** [Zurell et al., 2020b], or large variation among the models and datasets, especially for communities comprising many species that are rare [Norberg et al., 2019]. With regard to rare species, one of the assumed strengths of multi-species and joint models, both studies concluded that marginal predictions did not benefit from the joint modeling. A recent study [Zhang et al., 2020] found that multivariate machine learning models (multi-output neural networks, and multi-output Random Forest) which include dependencies implicitly were better at rare species prediction than JSDMs which include dependencies explicitly.

The predictive superiority of non-linear machine learning algorithms (RF, BRT, NN) compared to conventional regression models and their multivariate extensions (JSDMs) highlights **the importance of the representation power of the distribution models**. Indeed, JSDMs rely heavily on GLMs to model species' environmental response, which limits their ability to capture complex response shapes. A typical workaround employed by most studies consists in the manual inclusion of pre-transformed features: logarithmic, polynomial, or feature interaction effects. As JSDMs are applied to increasingly large species pools, the ability to discriminate the high number of potential combinations of species using simple linear mappings is limiting. Additionally, complex features such as images, sounds or large sets of features require **consequent feature engineering efforts prior to applying current JSDMs**.

Finally, most JSDMs rely on variants of Markov Chain Monte Carlo (MCMC) algorithms for the posterior inference. While they allow a better quantification of the parameter uncertainty in a Bayesian framework, these **inference algorithms scale poorly with the dataset size** (number of observations). This is due to the estimation of random effects (as many as the number of species or the number of latent factors) for each observation/site. Approximate inference algorithms have been developed to improve the computational efficiency of JSDMs, relying on either Variational Inference (VI) (e.g. GLLVM [Niku et al., 2019]) or Monte Carlo integration (e.g. sJSDM [Pichler and Hartig, 2020, Chen et al., 2018]). However, these optimizations were tailored to the GLMM model, whereas **little effort has been devoted to extending the abiotic response beyond the classical GLM**.

1.3.2 Ecological network learning

Ecological networks are paramount tools in community ecology and food web ecology [Boitani et al., 2007, Gravel et al., 2011, Bohan et al., 2017]. Building an ecological network requires data on the community composition and ecological interactions. In general, community composition is directly observed in the field, but in certain circumstances could also be predicted by (multi-/joint-)species distribution models. Nowadays, the emergence of environmental DNA sampling of soil or water allowed near-exhaustive (depending on the markers used) inventory of species currently or previously populating communities [Taberlet et al., 2018, Derocles et al., 2018, Bohan et al., 2017]. However, interspecific interactions are not always easy to establish. While the proliferation of interaction databases (e.g. GLOBI [Poelen et al., 2014], MANGAL [Vissault et al., 2019]) allows at least the partial description of some networks, most of the available knowledge is limited to well-studied ecosystems and to a few mostly direct interactions: trophic interactions (e.g European tetrapods [Maiorano et al., 2020], Grassland invertebrates [Hines et al., 2019]), plant–pollinators (e.g. [Boreux and Klein, 2019]), plant–frugivores (e.g. [Dehling et al., 2016]) host–parasite (e.g. [Dallas, 2016]), seed dispersal (e.g. [Timóteo et al., 2018]), root–fungi (mycorrhizal complex) (e.g. [Chaudhary et al., 2016, Soudzilovskaia et al., 2020]). Some interactions are impossible to observe for practical reasons (e.g. interactions happening below-ground in the soil). Inference techniques dedicated to specific types of interactions are being developed (e.g. stable isotopes in food webs [Layman et al., 2012]). Ecologists also resort to proxies based on species traits, phylogenetic conservatism of interactions within clades, co-occurrence patterns, and so on [Morales-Castilla et al., 2015]. These proxies take the form of heuristic rules either formulated by experts (e.g [Maiorano et al., 2020, Hines et al., 2019]) or learnt by machine learning models [Majdi et al., 2018, Vacher et al., 2016].

There are two paradigms for machine learning ecological networks as outlined in Fig 1.4. The Interaction

prediction paradigm consists of methods that aim at predicting the probability of an interaction between a pair of species. The association inference paradigm, on the other hand, consists of methods that infer the consequence of interactions on the realized niche of species from their co-distributions.

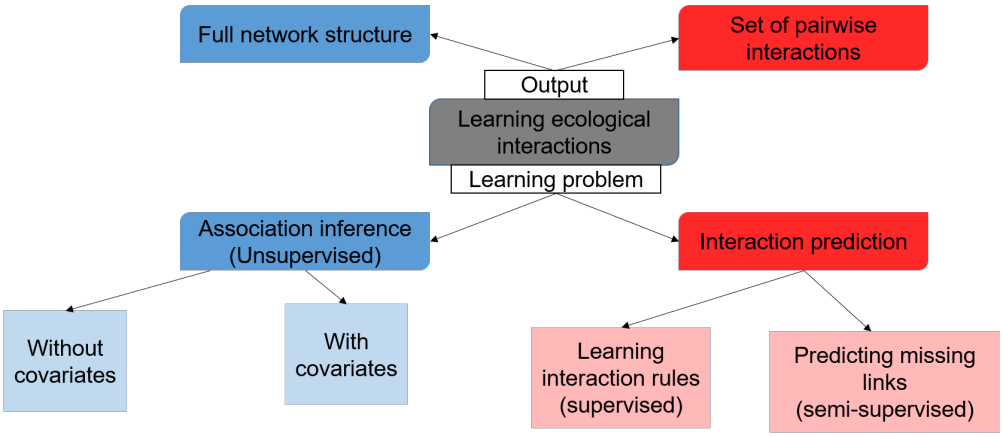


Figure 1.4: Classification of ecological network learning methods.

In the former, the interaction is the target of the prediction, so that the problem can be solved by **supervised or semi-supervised learning** algorithms. These approaches require at least a partial observation of the ecological network. Their output is a list of pairwise interactions that can be expressed as an **interaction network**. This learning problem is sometimes referred to as network reconstruction [Momal et al., 2019] in the literature.

In the latter, the interaction is thought of as a latent process which impacts co-occurrences resulting in correlated or dependent spatial distributions. Such dependencies referred to as "associations" are encoded in a covariance structure represented with a graph known as "association network". Learning the structure of this network is an **unsupervised learning** problem known in computational biology as **network inference**.

Predicting pairwise interactions

Rooted in community and functional ecology, the predictive approach seeks to predict the probability of interaction of a pair of species according to either their attributes such as functional traits (e.g body size, beak shape, physiological, phenological or life history traits), functional/trophic groups (e.g herbivores, decomposers, predators, etc.), or to the interactions of closely related species.

Learning interaction rules Supervised approaches relate known interactions to species attributes to understand the determinant of biotic interactions. In other words, they aim to answer the question "why particular species interact and others not ?" A key hypothesis regarding this question known as "trait-matching" [Schleuning et al., 2015] is that species interact when their traits are compatible [Eklöf et al., 2013]. For example, a long narrow flower is more compatible with a hummingbird species having a very long beak whereas a flat wide flower makes a more convenient landing platform for bees [Garibaldi et al., 2015]. Trait-matching was found relevant in other mutualistic networks (plant-frugivore and seed-dispersal networks [Dehling et al., 2016, Bender et al., 2018]), but also in antagonistic networks (host-parasites [McQuaid and Britton, 2013], food webs [Gravel et al., 2013, Laigle et al., 2018, Brousseau et al., 2018]). Learning trait-matching rules can be achieved using either logic-based [Bohan et al., 2011, 2017] or probabilistic machine learning approaches. In the latter, non-linear models (random forest, deep neural networks) were found to have better performances than conventional regression models [Pichler et al., 2020]. The extent to which trait-matching is useful to learn ecological networks depends on the relative proportion of specialists as opposed to generalists amongst modeled species. So far this approach was applied to direct interactions, in well-studied systems with documented interaction databases, with **little transferability** to other systems. Finally, **the approach loses all its interest in the absence of relevant traits information.**

Predicting missing links Empirical ecological networks are sparse and hence any unobserved interactions could be either "missing links" (interactions that would be observed with complete sampling), or true absences [Terry and Lewis, 2020]. Ecological networks are manifestly non-random [Vázquez et al., 2009, Pascual et al., 2006]. Several semi-supervised learning techniques have been developed to harness the underlying structure of the network to help predict missing links. They include hierarchical structuring models [Clauset et al., 2008], matching-centrality models [Rohr et al., 2016], stochastic block models [Guimerà and Sales-Pardo, 2009], k-nearest neighbor recommenders [Desjardins-Proulx et al., 2017], and collaborative filtering [Fu et al., 2019], etc. Unlike trait-matching methods which are interested in the mechanisms and requirements of interactions, these approaches originating from graph mining are more focused on the network properties and patterns such as the existence of a modular structure: communities/roles, the degree distribution, etc.

Inferring association networks

In many applications, **knowledge of existing interactions is not available or is highly context-dependent, rendering the previous methods inapplicable**. In contrast, occurrence data is easier to obtain. In the study of community response to environmental change, uncovering the exact mechanism underlying the interaction is less important than detecting the existence of the interaction and assessing its effect on species abundances. Inference methods aim to assess the imprint of biotic interactions on species co-occurrences. To achieve this objective, species associations are first quantified from the data using either **descriptive** or **model-based** techniques. Afterwards, associations are linked to possible processes (biotic interactions, unmeasured abiotic factors) by an expert or in light of empirical studies.

Descriptive approaches A battery of indices are used to describe interspecific associations (cf. [Keil et al., 2019] for review). This includes spatially-implicit indices calculated solely on community data (e.g Jaccard index, odds ratios) [Ulrich and Gotelli, 2013] and spatially-explicit indices (e.g community variograms, point patterns analysis)[Wiegand and Moloney, 2013]. Null models [Gotelli and Graves, 1996, Sanderson and Pimm, 2015] are used to assess the statistical significance of associations and standardize their value to enable their comparison between studies. Pairwise associations can eventually be thresholded and visualized as a network. **These approaches are primarily descriptive, rather than predictive.**

Model-based approaches The model-based approaches rely on a formalization of the joint probability of species occurrences/abundances, such that associations are encoded in a probabilistic graphical model. Probabilistic graphical models (PGM) [Koller and Friedman, 2009] use a graph-based representation to encode a joint distribution over a multi-dimensional space, whereby the presence (resp. absence) of an edge between two nodes represents a dependence (resp. independence) conditional to the rest of the nodes. PGMs are particularly useful because they distinguish direct effects between species from indirect effects mediated by other species [Popovic et al., 2019]. The core differences between network inference models revolves around three interlinked axes.

First, the type of graphical model determines the structural constraints of the associations: directed acyclic graph or Bayesian Network (BN), undirected graphical models including Markov Random Fields (MRF) and covariance graphs (CG). A covariance graph [Peña, 2013] is a graph derived from a sparse covariance matrix. Therefore, it represents unconditional dependencies as opposed to a graph derived from a precision matrix which encodes conditional dependencies [Murphy, 2012]. Application of Covariance Graphs in community ecology include exclusively JSDMs, whereas BNs received little uptake from the ecological community [Staniczenko et al., 2017, Faisal et al., 2010] compared to MRFs [Harris, 2016, Clark et al., 2018] and CGs.

Second, models differ in terms of supported probability distributions of the observed data. Amongst MRF types, with the exception of the Ising Model which was applied to binary presence/absence data [Harris, 2016], most frameworks rely on the Gaussian Random Field or Gaussian Graphical Model (GGM). A GGM assumes a multivariate gaussian distribution of the data. Data transformations are typically used to accomodate non-gaussian data. Alternatively, hierarchical models (including JSDMs) use gaussian latent (unobserved) variables to encode dependencies and flexible data generating distributions to support different observed data types (presence/absence, abundance).

Third, models can be separated into conditional/hierarchical approaches, on one hand, and marginal approaches, on the other hand. Conditional models condition the dependencies on latent variables. They include the multivariate Poisson Log-Normal model (PLN) [Aitchison and Ho, 1989] and its extensions [Chiquet et al., 2018, Momal et al., 2019], as well as current implementations of JSDMs (GLMM/GLLVM) [Pollock et al., 2014, Warton et al., 2015, Ovaskainen et al., 2017, Clark et al., 2017, Niku et al., 2019, Pichler and Hartig, 2020]. In contrast, marginal models represent dependencies between species marginal distributions. This category includes copulas [Popovic et al., 2018] and generalized estimation equations [Warton, 2011].

Most network inference frameworks account for environmental conditions when inferring biotic associations (but see [Harris, 2016, Faisal et al., 2010]) by including covariates in the generative model of the data. The underlying incentive for the association learning task is to prevent the inference of spurious associations (false positives). In the community modeling context, the inclusion of abiotic covariates aims to disentangle abiotic preferences and biotic processes driving species distributions. Currently, this goal is far from being achieved [Blanchet et al., 2020, Barner et al., 2018, Holt and Barfield, 2009]. We posit several reasons.

On one hand, the outcome of an ecological interaction for the involved species is rarely symmetric both in polarity (e.g parasitism, predation) and intensity. Consequently, the associations that arise from direct and indirect influences should not be expected to be symmetric either. Yet, so far **inference models used in community ecology rely heavily on symmetric associations** modeled using Gaussian Graphical Models, or Markov Random Fields in general. For their part, Bayesian networks are even more constraining as they impose a direction of influence and do not allow feedback links.

On the other hand, **most inference frameworks employ linear abiotic responses**, which limits the niche shapes that can be expressed by the model. Incorrectly specifying the abiotic model induces errors captured by the residuals, affecting the association estimation for methods that exploit the correlation of residuals to infer associations (e.g JSDM). While the inclusion of covariates reduces spurious associations, the remaining associations may still reflect the effect of hidden factors [Dormann et al., 2018, Blanchet et al., 2020, Harris, 2015] (e.g dispersal limitations, micro-habitat conditions, etc).

Finally, **most network inference frameworks assume that the response of a species is the sum of its abiotic and biotic responses**. This additive link can detect associations that compensate the effect of the environment, either positively (e.g facilitation) or negatively (competition), but is ill-suited in other situations. For instance, if we assume a system where the presence of a predator is conditioned on the habitat suitability of the site and the presence of at least one prey, the response can be expressed as a product of two probabilities describing the satisfaction of both requirements [Gravel et al., 2019]. Using an additive form will likely fail to detect the positive dependence of the predator on its preys. Moreover, one can naively assume that in a predator-prey relationship, the prey would be negatively affected by the predator (at least in terms of abundance). But, in some systems it has been observed that predators induce range expansion of their preys, which can be translated into a positive association [Holt and Barfield, 2009]. This highlights a fundamental conceptual shortcoming for network inference from community data: the same biotic interactions can translate into different associations. In this regard, **a flexible framework that allows testing various hypotheses is much needed.**

1.4 Objectives and methodological contributions

Adopting the assembly theory as our theoretical framework, our objectives in this thesis was to develop new methods to **analyze community data to disentangle the effects of assembly drivers, specifically habitat suitability and biotic interactions**. These tools also aim to predict how community composition changes in response to the environment. We break down the former objectives into the following tasks and highlight the main contributions hereafter:

Joint modeling of species distributions

Our first objective is to develop a joint species distribution model that conserves the benefits of existing solutions while allowing for the automatic discovery of relevant features to reduce feature engineering on the

users end, a task known as feature or **representation learning**. Current state of art of feature learning points towards the use of artificial neural networks (ANN). As natural extensions of Generalized Linear Models, the use of ANN preserves the **flexibility in response types** of GLMs, enabled by the exponential family of distributions. We also propose the use of **multi-task neural networks** to share learnt features between species prediction tasks. We hypothesize that this feature/representation sharing would improve the generalization ability of the model.

The complexity of associations modeling increases with the size of species pools. We follow the Latent Variable approach used in modern JSDM implementations to reduce the number of parameters. Each site is associated to a latent factor capturing unmeasured environmental descriptors affecting species or their co-occurrences. Since latent variables are not observed, their posterior inference from data is a major computational bottleneck, especially in datasets with many observations. To circumvent this issue, we resort to **amortized variational inference** to control model complexity and ensure its **scalability**. This is achieved by using a type of neural network called Variational Autoencoders, integrated within the previous multi-task model.

We describe our proposed solution in Chapter 3, and illustrate it with a concrete application in Chapter 6.

Interpretability of black-box models

Community modeling has two goals: understanding the abiotic and biotic drivers of species distributions and predicting their response to environmental changes. The shift towards black-box machine learning models to achieve more accurate predictions, our second goal, comes at the cost of reduced interpretability, which goes against the first goal. **Interpretable Machine Learning** (IML) tools make it possible to extract hidden knowledge captured by black box models (like our previous proposal). We review relevant IML techniques in Chapter 2 and their application to answer specific questions regarding species responses in Chapters 3 and 6.

Network inference from community data

In this thesis, we present a novel model for bidirectional association inference. We use a type of graphical model known as **dependency network** which allows for more flexibility in defining the network structure. The behavior of each species depends on its role (effect on other species) and its response to the effect of others and to the environment. We compare the ability of our approach to previously described network inference methods in simulation experiments, and illustrate its benefits on an empirical application. This task is addressed in Chapter 5.

1.5 Thesis outline

We dedicate the first part of this document to review methodologies upon which this thesis work is built.

- In chapter 1, we covered the theoretical ecology background needed to highlight our objectives and position this thesis work in the statistical community ecology literature,
- In chapter 2, we present an overview of feedforward neural networks and introduce principles and techniques of machine learning interpretability that are relevant to subsequent applications.

In the second part of the manuscript, we present the methods developed during this thesis and their applications.

- In chapter 3, we present a generic multi-task learning framework for modeling and inference from community data. Specifically, we delineate and detail its various components, and how they relate to existing species distribution models. Subsequent chapters report adaptations of the framework to address different ecological questions.
- In chapter 4, we apply the framework to a location-based recommendation problem aiming to predict the vegetation structure given a spatial location. The model is trained on a big dataset of opportunistic

citizen science observations of plants in France.

- In chapter 5, we present a novel method for learning bidirectional inter-specific associations from co-abundance data. As part of the study, we evaluate the ability of the model to recover associations in a simulation study against existing frameworks. Additionally, we apply the model to an empirical dataset of Alpine plants from a well-studied system.
- In chapter 6, we present a detailed case study highlighting the tools developed during the thesis and applied to the study of the biogeography of earthworms in France. For this purpose, we
- Finally, in chapter 7, we discuss the overall contributions of this thesis work, conceptual and technical limitations as well as outstanding questions and potential ways of improvement.

Chapter 2

From GLMs to neural networks: learning and interpretability

2.1 Introduction

The digitisation of scientific experiments along with advances in sensors and data capture tools, and the decrease in storage costs led to the generation of huge amounts of data. A core task of data-driven research is to make sense of this deluge of data. Data mining provides the tools to process, analyze and gain insights from the data [Aggarwal, 2015]. Machine learning, the science of getting computers to decide without being explicitly programmed, provides the "*methods to automatically detect patterns in data, and then use the uncovered patterns to predict future data or perform other kinds of decision making.*" [Murphy, 2012]

2.1.1 Machine learning problems

Machine learning comprises three main paradigms: supervised, unsupervised and reinforcement learning. Supervised learning aims to predict the state of a modeled entity given some conditions described by a set of variables called covariates. On the other hand, unsupervised learning attempts to discover structural patterns in the data. Reinforcement learning aims to train software agents to learn in an interactive environment by trial and error, to maximize a given reward. This category is not relevant to this work.

Predictive or supervised approaches learn a mapping from inputs x to outputs y, given a set of pairs $\mathcal{D} = (x_i, y_i)_{i=1}^N$ called the *training/calibration set*. In the simplest case, each training input x_i is a vector of scalars, referred to as **features**, **attributes**, or **covariates**. More generally, x can have a more complex structure (image, text sentence, graph, etc.). Similarly, the output y, referred to as **response** or target or label, can be anything. But for many real world problems it can be represented by a scalar. The type of output variable differentiates various supervised learning problems. When y is a categorical or nominal variable, the problem is known as **classification**, and when y is a real-valued scalar, the problem is known as **regression**. An ordinal response y yields an **ordinal regression** problem.

Unsupervised approaches deal with datasets with only inputs $\mathcal{D} = (x_i)_{i=1}^N$, but no targets/labels, hence the name. These methods aim to discover interesting patterns in the data. For instance, the task of discovering groups of similar instances is an unsupervised problem known as **clustering**. Similarly, discovering factors of variation using for example Principal Component analysis or Factor analysis also falls within this category.

2.1.2 A probabilistic approach

A supervised learning algorithm defines a function f to map the input to the output $y = f(x)$. The goal of learning is to estimate this function from the training set, then use it to make predictions of y given new

values of x. Each learning problem can be solved via different algorithms, which vary in the form of the learnt function f (e.g linear, decision tree, neural network etc).

In this thesis, we adopt a probabilistic model-based approach [Murphy, 2012], whereby instead of modeling Y, we model the distribution of y conditioned on x given the data. Concretely, we write $P(y = a \mid x = b, \mathcal{D}) = f(a, b) \in [0, 1]$. From this model f, we can compute the expected prediction, its variance and its mode (most probable value). This allows us to quantify for each possible value a of y, its probability given the value $x = b$, thus providing a measure of confidence, or conversely of uncertainty in the predictions.

In the unsupervised setting, since we do not have access to labels, we formalize the task as one of density estimation. This means that we assume our data have been generated from a given family of distributions with parameter Θ, for instance the family of multivariate normal distributions with mean zero parameterized with a covariance matrix Σ. In this case, the goal is to learn the parameters of the assumed data distribution.

2.1.3 Outline

Most models described further in this thesis build on the generalized linear model (GLM) [McCullagh, 81]. These supervised models apply for both discrete and continuous responses. We start by describing the univariate GLM, used to model scalar responses. Then, we present an overview of feedforward neural networks as extensions of GLMs which constitute a crucial building block of our proposals in subsequent chapters. We limit ourselves to multi-layer perceptrons, convolutional neural networks and autoencoders which are relevant for the subsequent applications. Finally, we present an overview of machine learning interpretability tools applicable to neural networks and relevant to subsequent applications. The mathematical details of methods used in the following chapters are detailed. An exhaustive overview of the topic can be found in the Interpretable Machine Learning book Molnar [2020].

2.2 Generalized linear models

2.2.1 GLMs and the exponential family

The Exponential Family of distributions

The Exponential Family (EF) defines a set of probability distributions whose probability density function can be expressed in the following canonical form:

$$f(y; \theta, \phi) = exp\left[\frac{y\theta - a(\theta)}{\phi} + c(y, \phi)\right]$$

c :normalizing term (constant wrt the parameter)

a :partition function

θ :natural or canonical parameter

ϕ :dispersion parameter

The EF includes most commonly used distributions: Gaussian (normal), Bernoulli, Poisson, Dirichlet, etc. For a given distribution in the EF, the functions a and c can be uniquely derived from its probability density function. Moreover, the density parameter of this distribution μ (usually the mean) can be mapped to the natural parameter θ of the canonical form using a uniquely determined and invertible function ψ. This way, any distribution of this family can be written in the canonical form described previously, simply by plugging the appropriate expression for ψ, that satisfies:

$$\mu = \psi^{-1}(\theta)$$

Probabilistic formulation of the GLM

In the GLM framework, a scalar response y conditioned on a vector of covariates x follows a distribution in the exponential family $\mathcal{F}$ with natural or canonical parameter θ and dispersion ϕ, such that the natural parameter is a function of x.

$$y \sim \mathcal{F}(\theta(x), \phi)$$

In a GLM model, the mean is assumed to be an invertible monotonic function g^{-1} of a linear combination of the covariates. The functions g and its inverse g^{-1} are known as the **link** and inverse mapping or **mean** function respectively.

$$\eta_i = \beta \cdot x_i$$
$$\mu_i = g^{-1}(\eta_i)$$

Commonly written as:

$$g(\mu_i) = \beta \cdot x_i$$

Fig 2.1 illustrates the relationship between the link and mean function.

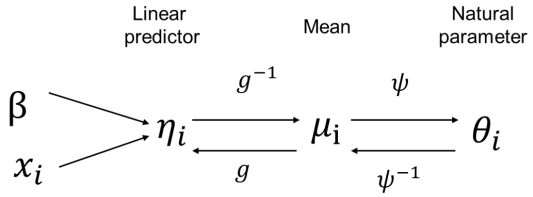


Figure 2.1: Schematic illustration of the mapping between the natural parameter, the linear predictor and the mean response in a generalized linear model.

We provide in Table 2.1 some of the most used link functions for distributions employed in this thesis. A special case is the canonical link function $g = \psi$ which satisfies: $\theta_i = \eta_i$.

Maximum Likelihood and Maximum A Posteriori estimates

The goal of learning in a GLM is to estimate the parameters β that maximize the conditional likelihood w.r.t the training data. Since the logarithm is a monotonically increasing function, we can maximize the conditional log-likelihood equivalently.

The conditional log-likelihood for a univariate GLM, with probability density function f and a dataset $\mathcal{D}$ is given by:

$$\ell(\beta) = \log p(\mathcal{D} \mid \beta) = \sum_{i=1}^{N} \log f(y_i; x_i, \beta, \phi) = \frac{1}{\phi} \sum_{i=1}^{N} \ell_i$$
$$\ell_i \triangleq \theta_i y_i - a(\theta_i) \quad ; \quad \theta_i = \psi \circ g^{-1}(\beta.x_i)$$

The likelihood can be optimized for instance using gradient descent [Bottou, 2010], or second-order methods [Battiti, 1992] for improved efficiency. Both procedures require computing the gradient of the likelihood w.r.t the parameters. By properties of the exponential family and the partition function we have the following

Model	Distribution, mean μ and dispersion ϕ	Link function $\theta = g(\mu)$	Inverse mapping (mean) function $\mu = g^{-1}(\theta)$
Linear regression	Normal $N(\mu, \sigma^2)$ $\phi = \sigma^2$	identity $\theta = \mu$	identity $\mu = \theta$
Binary logistic regression	Bernoulli $\mathcal{B}(p)$ $\mu = p$ $\phi = 1$	logit $\theta = \log \frac{\mu}{1-\mu}$	logistic/sigmoid $\mu = sigmoid(\theta)$ $\mu = \frac{1}{1+exp(-\theta)}$
Multinomial logistic regression	Multinomial of size K $\mathcal{M}(\pi)$ $\mu = \pi = (\pi_1, ..., \pi_K)$ $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$ $\phi = 1$	Inverse softmax (generalized logit) $\theta = (\log \frac{\pi_1}{\pi_K}, \ldots, \log \frac{\pi_{K-1}}{\pi_k}, 0)$	Softmax (generalized logistic) $\mu = (\frac{exp(\theta_1)}{1+\sum_{i=1}^{K-1} exp(\theta_i)}, \ldots, \frac{exp(\theta_{K-1})}{1+\sum_{i=1}^{K-1} exp(\theta_i)}, 1)$
Negative binomial regression	Negative binomial with fixed number of failures $r > 1$ $NB(p, r)$ $\mu = p$ $\phi = \frac{1}{r}$	log $\theta = \log(\mu)$	exponential $\mu = exp(\theta)$
Poisson regression	Poisson $\mathcal{P}(\lambda)$ $\mu = \lambda$ $\phi = 1$	log $\theta = \log(\mu)$	exponential $\mu = exp(\theta)$

Table 2.1: Generalized linear models for common distributions, link functions. The inverse of the link function yields the mean for most distributions. Note that in the case of the negative binomial, the distribution is parameterized with the probability of success instead.

expressions of the expectation and variance:

$$\mathbb{E}[y_i \mid x_i; \beta, \phi] = \mu_i = \frac{da}{d\theta}(\theta_i)$$
$$\mathbb{V}[y_i \mid x_i; \beta, \phi] = \phi_i = \frac{d^2a}{d\theta}(\theta_i)\phi$$

Hence, the partial derivative w.r.t the regression parameters over the jth covariate β_j:

$$\frac{\partial l_i}{\partial \beta_j} = (y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}$$

Furthermore, if the canonical link is used, i-e $\eta_i = \theta_i$, then the partial derivative is:

$$\nabla_\beta l(\beta) = \frac{1}{\phi}\left[\sum_{i=1}^{N}(y_i - \mu_i).x_i\right] = \frac{1}{\phi}\left[\sum_{i=1}^{N}(y_i - g^{-1}(\beta.x_i)).x_i\right]$$

Given a prior distribution $p(\beta)$ over the regression parameters, we can compute the point of maximum posterior probability of the parameters, known as the maximum-a-posteriori (MAP) estimate. The MAP combines background information represented as a priori distribution over the value of the parameter $p(\beta)$ (kind of regularization) and the likelihood $p(D \mid \beta)$.

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \left[\log p(\mathcal{D} \mid \beta) + \log p(\beta)\right]$$

Regularization and priors

Regularization techniques are used to prevent machine learning models from overfitting. With some choices of priors, finding the MAP estimate is equivalent to adding regularization terms to the log-likelihood. However,

not all forms of regularization can be expressed in terms of prior distribution (e.g data-dependent penalties). For instance, using a Gaussian prior is equivalent to adding L2 penalties, known as ridge or Tikhonov regularization [Tikhonov and Arsenin, 1977]. Conversely, using the Laplace prior amounts to minimizing L1 penalties, known as lasso regularization [Tibshirani, 1996] (Table 2.2)

Name	Parameter	Penalty	Prior
Lasso, L1	λ	$\lambda \sum_k \lvert\beta_k\rvert$	$\beta \sim \text{Laplace}(\lambda)$
Tikhonov, ridge, L2	λ	$\lambda \sum_k \beta_k^2$	$\beta \sim \mathcal{N}(0, \frac{1}{\lambda} I)$
Elastic net, L1_L2	α or (λ_1, λ_2)	$(1-\alpha) \sum_k \lvert\beta_k\rvert + \alpha \sum_k \beta_k^2$ $\lambda_1 \sum_k \lvert\beta_k\rvert + \lambda_2 \sum_k \beta_k^2$	/

Table 2.2: Common regularization penalties and formulation in terms of prior over the parameters.

Both L1 and L2 techniques of regularization prevent overfitting by shrinking the coefficients. L2 shrinks all the coefficient by the same proportions but eliminates none. The resulting objective is convex and the estimates stable but the coefficients are less sparse. On the other hand, L1 can shrink some coefficients to zero, performing variable selection. However, the number of selected variables is bounded by the number of observation on datasets with more features than observations. Moreover, it performs poorly in the presence of correlated groups of variables. Elastic net (or L1-L2) regularization combines both types of penalties linearly. The absolute (L1) term generates a sparse model, while the quadratic (L2) term removes the limitation on the number of variables, encourages correlated variables to be selected together (grouping effect) and stabilizes the lasso selection procedure [Zou and Hastie, 2005].

2.2.2 Limits of linear models

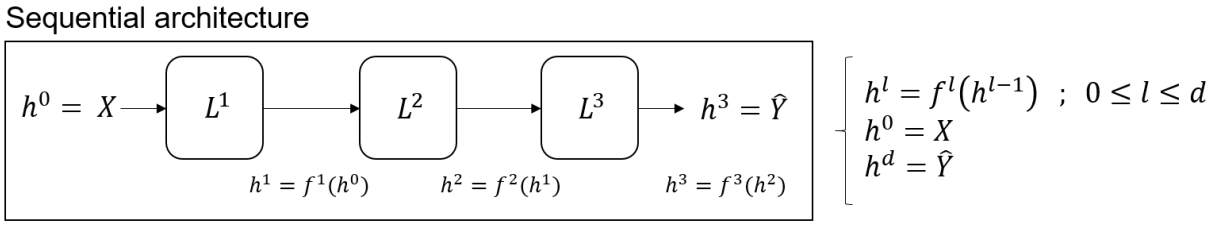
Linear models are appealing because they can be fit reliably, either in closed form or with convex optimization. However, they cannot model non-linear functions of the input covariates, including interactions between inputs. To extend linear models, we can apply them to non-linear transformations of the input $\xi(x)$. It can be interpreted as providing a new derived **representation** of the covariates called **features**. There are various strategies for choosing ξ. One option involves **feature engineering** and human effort, and is specific to the domain of the predictive task. Another possibility is to use **kernel machines** [Scholkopf and Smola, 2018] within GLMs $\xi(x) = (k(x, u_1), ..., k(x, u_N))$ which perform a sort of template matching by measuring the similarity of observed covariates to stored prototypes by means of a kernel function k. Finding a good kernel function is task-specific. Last, one can learn ξ directly from the data. Models which follow this strategy are part of the **Adaptive Basis Function Models** (ABM) family [Murphy, 2012] which includes Classification and Regression Trees (CART) [Breiman et al., 1983], Generalized Additive Models (GAM) [Hastie and Tibshirani, 1990], Random Forest (RF) [Breiman, 2001] and feedforward neural networks (NN) [Goodfellow et al., 2016].

2.3 Feedforward neural networks

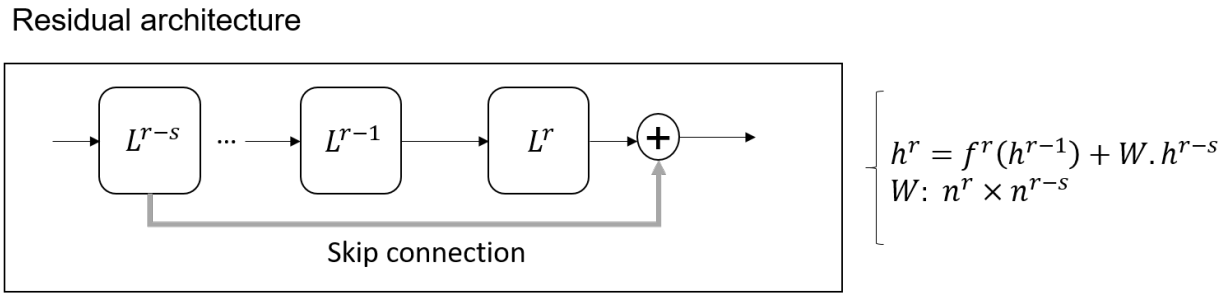
2.3.1 Multilayer perceptrons

A **feedforward neural network** approximates the response variable using a composition of many different functions applied to the input covariates. The information flows from the input through the intermediate computations to the output without feedback, hence the adjective *feedforward*. Neural networks that contain feedback connections belong to the class of *recurrent neural networks* [Rumelhart et al., 1986], which we do not use in this thesis.

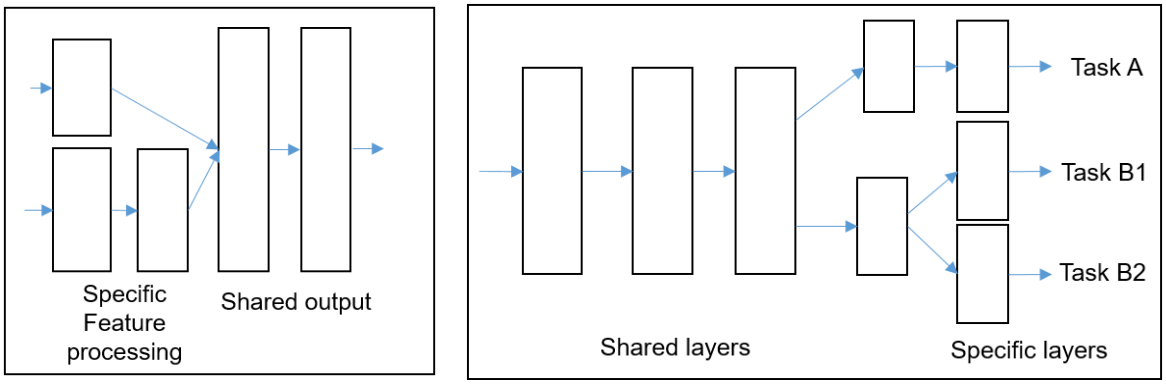
The network architecture, i.e. how the functions are composed together, can be represented as a directed acyclic graph. The most common architecture used is a sequential chain. Other topologies have been explored for various applications, we illustrate some in Fig 2.2.



(a) Sequential models.



(b) Residual or skip-connection networks.



(c) Multi-input and multi-output models with shared layers for the prediction and feature extraction respectively.

Figure 2.2: Examples of feedforward neural architectures. Computation layers are arranged into a directed acyclic graph (i.e. involve no feedback loops).

In general, theses functions are referred to as **layers**. For this reason, feedforward are sometimes referred to as Multi-Layer Perceptrons (MLP). The first or **input layer** simply feeds the input covariates to the

network. The intermediate or **hidden layers** perform feature extraction meaning they learn non-linear transformations to generate new features from the raw inputs.

Each feature is generated by a unit called **neuron**. Each neuron (Fig 2.3) computes a linear combination of its inputs followed by a non-linear transformation called **activation** or **transfer** function. The role of activations in hidden layers is to incorporate non-linearity.

Finally, the last or **output layer** computes the model prediction. In classification and regression problems, the output layer is equivalent to a generalized linear model applied to the result of the penultimate layer with the output activation playing the role of the inverse mapping function (Table 2.1).

In the deep learning lingo, the identity, logistic, and generalized logistic functions are known as linear, sigmoid and softmax activations respectively. Other common activation functions include the hyperbolic tangent (tanh), the rectified linear unit (relu), the exponential linear unit (elu), and leaky rectified linear unit (leaky relu), etc.

Fig 2.3 illustrates the architecture and functioning of a multi-layer perceptron.

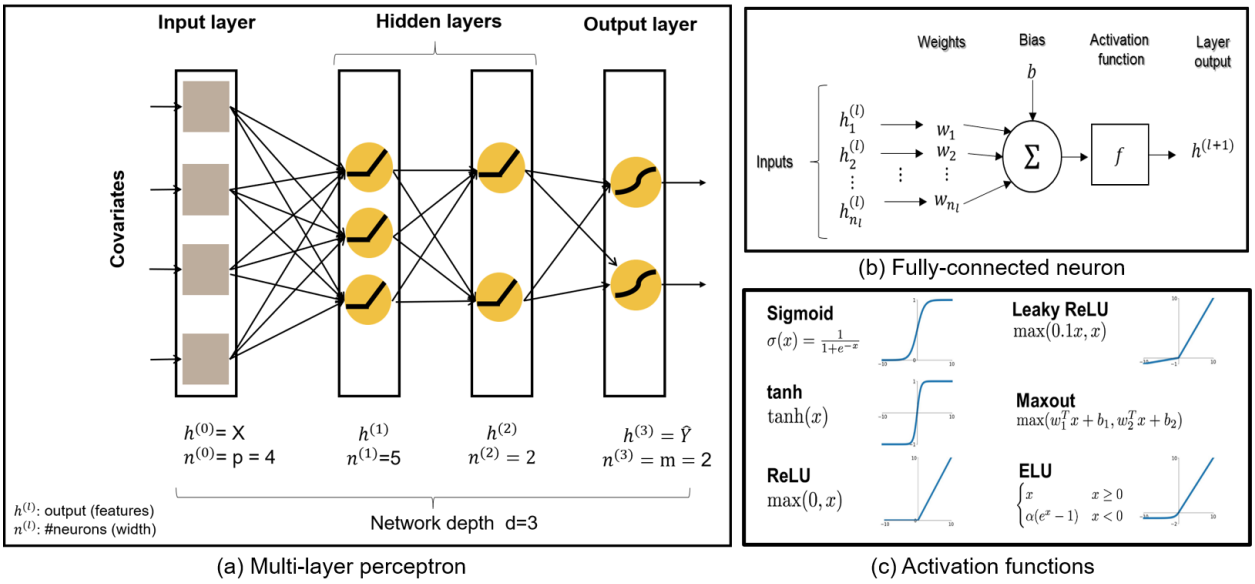


Figure 2.3: Multilayer perceptron with fully connected layers. Square nodes are observed data, circular nodes are functions. (a) MLP architecture, (b) the functioning of a neuron (c) Common activation functions.

The number of layers quantifies the **depth** of the network. Based on that, we distinguish shallow from deep neural networks. The **width** of a layer is the number of neurons it contains or equivalently the number of new features it generates. The MLP is a *universal approximator*, meaning it can model any smooth function given enough hidden units [Hornik, 1991]. With limited data, designing a neural network architecture requires balancing depth and width while maintaining enough complexity to correctly approximate the function $y \approx f(x)$. Deep networks are better than shallow networks at generalization whereas wide networks are better than narrow network at memorization.

Model complexity

In a multi-layer perceptron, each neuron in each layer computes a linear combination of all the layer inputs (outputs of the previous layer) followed by an activation. In practice, the layer output is computed using matrix multiplication followed by the activation function. If we denote by N_{out} the number of hidden units (neurons) of the layer, and by N_{in} the number of inputs, then the number of parameters of the layer is equal to $N_{out} \cdot (N_{in} + 1)$.

2.3.2 Convolutional neural networks

Convolutional Neural Networks (CNN) [LeCun et al., 1989] are specialized networks for processing data with grid-like topologies (e.g. images). The name "convolutional" refers to the fact that these networks use the **convolution operation** in lieu of matrix multiplication in at least one of their layers.

CNNs use three types of layers depicted in Fig 2.4: convolutional (A), pooling (B) and fully connected layers (C).

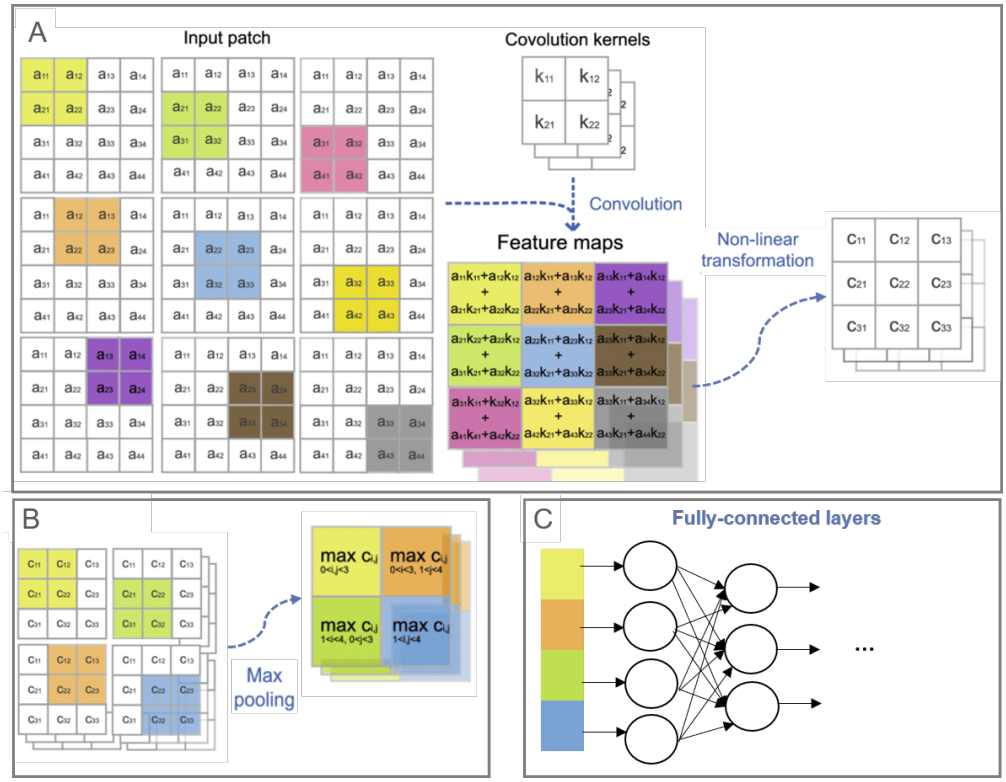


Figure 2.4: Convolutional neural network layer types functioning.

Convolution layer

The convolution layer uses filters that perform convolution operations. Each filter is defined by a kernel. The filter scans the input image from left to right and from top to bottom. At each step, it computes the linear combination of its kernel with the part of the image it covers, followed by a non-linearity (activation function). Each convolution filter learns to detect specific features (e.g. edges, circles). Since they apply to raw data, the initial convolution layers learn low-level features such as edges. Subsequent layers apply to these low-level features to learn mid-level features such as various shapes (circles, triangles, etc.), the deeper the layer position, the more it learns high-level features.

Assuming the input image/patch has a dimension $I \times I' \times C$. For instance, a RGB picture of size 64×64 is represented by a multidimensional matrix (tensor) of dimension $64 \times 64 \times 3$, the third dimension represents the RGB values. Input images are often assumed have a squared dimension (I=I'). In general, the filter dimensions are small odd numbers $F \times F$ (e.g. 3×3, 5×5 or 7×7). The associated kernel is a hypervolume of dimension $F \times F \times C$ ($F <<< I$), as opposed to $I \times I \times C$ if a fully-connected layer is used on the flattened image. Therefore, the use of the convolution operator instead of general matrix multiplication reduces the parameters drastically by sharing the kernel parameters amongst the image regions.

Each filters produces a feature map or activation map of dimension $O \times O \times 1$. Therefore, a convolution layer that uses N filters outputs a tensor of dimension $O \times O \times N$ representing a stack of N feature maps.

In addition to the number and dimension of the filters, a convolution layer includes two hyperparameters which control the dimension O of the feature maps, namely the stride S and the zero-padding P. The stride denotes the number of pixels by which the convolution filter moves after each operation.

Zero-padding denotes the process of adding P zeroes to each side of the boundaries of the input. This value can either be manually specified or automatically set using one of the following strategies:

- **Valid padding**: Does not use any padding, instead it drops the last convolution if dimensions do not match.
- **Same or half padding**: padding chosen such that the dimension of the feature map is $\lceil \frac{I}{S} \rceil$
- **Full or maximum padding**: chosen such that end convolutions are applied on the limits of the input.

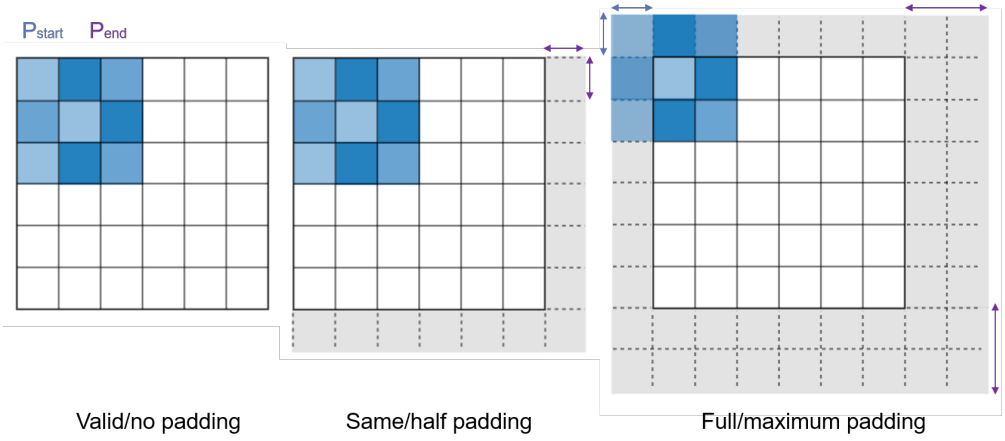


Figure 2.5: Convolution layer padding types. Modified from (CS230 Deep Learning course Stanford University)

The output size O of the feature map is given by:

$$O = \frac{I - F + P_{start} + P_{end}}{S} + 1$$

Pooling layer

The pooling layer is typically applied after a convolution layer to down-sample the data by aggregating neighboring pixels, taking the average (average pooling) or maximum (max-pooling) value. The pooling operation can be viewed as adding an infinitely strong prior that the feature learnt by the layer must be invariant to small translations. This property might be useful for some features (channels) but not for others.

Fully connected layers

Fully connected layers (FC) are the same as those used in Multilayer Perceptrons. They are typically applied on a flattened input where each input is connected to all the neurons of the layer. If present, FC layers are usually found towards the end of CNN architectures to optimize supervised objectives such as class scores.

Model complexity

The model complexity of a convolutional neural network is measured by the number of parameters. For each type of layer, Fig 2.6 summarizes the input, output and parameter sizes.

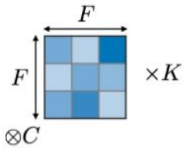
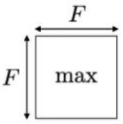
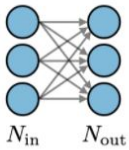
	CONV	POOL	FC
Illustration			
Input size	$I \times I \times C$	$I \times I \times C$	N_{in}
Output size	$O \times O \times K$	$O \times O \times C$	N_{out}
Number of parameters	$(F \times F \times C + 1) \cdot K$	0	$(N_{\text{in}} + 1) \times N_{\text{out}}$
Remarks	• One bias parameter per filter • In most cases, $S < F$ • A common choice for K is $2C$	• Pooling operation done channel-wise • In most cases, $S = F$	• Input is flattened • One bias parameter per neuron • The number of FC neurons is free of structural constraints

Figure 2.6: Complexity of layer types in a convolutional neural network. (Reprinted from CS230 Deep Learning course Stanford University)

2.3.3 Autoencoders and Variational autoencoders

An autoencoder is a feedforward neural network that learns to copy its input to its output. It is composed of two MLPs: an encoder network that maps the input into the code, and a decoder network that maps the code to a reconstruction of the original input. Several variants exist to the basic model, that force the learned representations of the input to have useful properties: *sparse*, *denoising*, *contrastive* [Goodfellow et al., 2016].

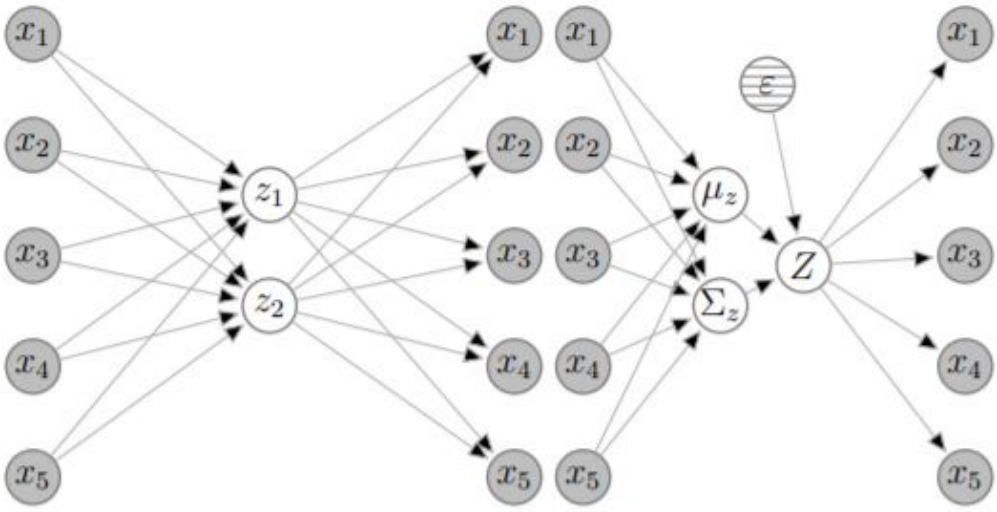


Figure 2.7: Graphical models depicting the architecture of a standard autoencoder (left) and a variational autoencoder (right). The standard autoencoder defines the code/latent layer as a deterministic function of the data. The variational autoencoder defines the latent code as a random vector and incorporates a sampling mechanism to draw samples from its distribution. [Mazoure, 2018]

Variational autoencoders (VAEs) are generative neural networks. Generative models like the VAE try to simulate how the data is generated, in order to understand the underlying factors driving it. Concretely VAEs are directed probabilistic graphical models of the data conditioned on latent variables whose posterior is approximated by a neural network, forming an autoencoder-like architecture. VAEs use a variational approach for latent representation learning, which results in an additional loss component and a specific estimator for the training algorithm called the Stochastic Gradient Variational Bayes (SGVB) estimator [Kingma and Welling, 2013].

Fig 2.7 illustrates the difference between an autoencoder (left) and a variational autoencoder (right). In VAEs, injecting stochastic noise (shaded node) allows for sampling new observations instead of reconstructing the input exactly. While an autoencoder can be seen as a nonlinear version of Principal Component Analysis (PCA), a variational autoencoder can be seen as a nonlinear equivalent to Factor Analysis or probabilistic PCA [Kingma and Welling, 2019].

2.3.4 Training feedforward neural networks

The goal of any supervised learning algorithm is to find a function that best maps a set of inputs to their correct output. Training a feedforward neural network aims to find the parameters (weights, biases) $\Theta = \{\mathcal{W}, b\}$ that minimize the predictive error on the data $\mathcal{D}$ as measured by a loss function $\mathcal{L}(\mathcal{D}, \Theta)$. Common loss function include the Mean Squared Error for regression tasks, and cross-entropy for classification.

Gradient descent

The gradient descent method involves calculating the derivative of the loss function with respect to the weights of the network. The simplest algorithm for parameter estimation in neural networks is known as stochastic gradient descent (SGD) [Bottou, 2010].

Algorithm 1: Stochastic Gradient Descent

Inputs : Observed values $x_1, .., x_n$, responses $y_1, .., y_n$, iterations number T, learning rate $\alpha(t)$, loss function $\mathcal{L}_w$, batch size B;
Outputs Converged parameters $\theta^{(T)}$;
:
Initialize $\theta(0)$ according to some distribution;
foreach $t = 1, 2, ..., T$ **do**
  foreach $b = 1, 2, ..., B$ **do**
    $\theta^{(t)} \leftarrow \theta^{(t-1)} - \alpha(t) \frac{1}{B} \sum_{i \in B} \nabla_w \mathcal{L}_w(y_i, \hat{y}_i)$;
  end
end

Backpropagation

Feedforward neural networks can be expressed as composition of functions. The chain rule dictates how to compute the gradients for composition of differentiable functions. Assuming each layer/neuron computes a differentiable transformation of its input, the chain rule can be used to compute the derivative of the loss function w.r.t any parameter at any layer.

$$
\begin{aligned}
y &= f(u) \\
u &= g(x) \\
y &= f(g(x)) = f \circ g \\
\frac{\delta y}{\delta x} &= \frac{\delta y}{\delta u} \frac{\delta u}{\delta x}
\end{aligned}
$$

The backpropagation (BP) algorithm starts from random initial parameters (weights and biases). Typically, weights are initialized using the glorot uniform function [Glorot and Bengio, 2010], while biases start at

zero. Then, at each iteration (epoch), the algorithm computes for each observation (x, y), the corresponding network prediction by applying the successive transformations of the network layers from the input to the output. This step is known as the forward pass. After that, the algorithm evaluates the loss/error of prediction as a function of the actual prediction $\hat{y}$ and the true outcome $y : \mathcal{L}(\hat{y}, y)$.

Then, it applies gradient descent to update each of the network parameters, which requires computing the gradient of the loss w.r.t each parameter. To achieve that efficiently, the BP algorithm exploits the DAG structure of the feedforward network architecture to compute these gradients starting from the output layer to the input layer using the chain rule. This step is referred to as the backward pass.

In summary, gradient descent is used within the backpropagation loop to evaluate the likelihood of each parameter combination and update their values accordingly.

2.4 Multi-task learning

2.4.1 Overview

Multi-task Learning (MTL) is a technique that aims to improve model generalization by solving several machine learning tasks in parallel while using sharing parts of their models (intermediary features/representations).

Caruana [1993] highlighted mechanisms explaining why MTL works. MTL increases the sample size that we are using for training our model, since some parameters are shared (**implicit data augmentation**). As all tasks are at least somewhat noisy, learning a single task bears the risk of overfitting whereas jointly learning two tasks enables the model to learn more general features/representations. As such MTL has a **regularizing** effect. Moreover, solving for multiple problems allows the model to focus on relevant features (**attention focusing**). Sometimes, useful features are easy to learn for some tasks, but difficult for other tasks. Through MTL, we can allow the model to **eavesdrop**, i.e. learn useful features through other tasks [Ruder, 2016].

2.4.2 Auxiliary tasks

One can attempt learning a group of principal tasks using a group of auxiliary tasks, unrelated to the principal ones. In many applications, joint learning of unrelated tasks which use the same input data can be beneficial. Experiments on synthetic and real data have indicated that incorporating unrelated tasks can result in significant improvements over standard multi-task learning methods. For instance, in Chapter 5, we propose a network inference model that solves two tasks simultaneously: abundance prediction and association inference. The former is a supervised learning problem, and the latter an unsupervised problem. The embeddings learnt in the association model (auxiliary task) are shared with the model predicting species abundances (principal task).

2.4.3 Multi-task neural networks

In the deep learning literature, multi-task learning is traditionally achieved using two types of neural network architectures: hard or soft sharing of hidden layers depicted in Fig 2.8. In hard sharing, a subset of the top hidden layers is shared between all tasks. However, each task has specific output layers. The more tasks are learned simultaneously, the more the model is forced to find a general representation, which reduces overfitting. Conversely, in soft parameter sharing, each task has its own model with its own parameters. The distance between the parameters of the model is then regularized in order to encourage the parameters to be similar. A similar approach is used in hierarchical models of species distributions.

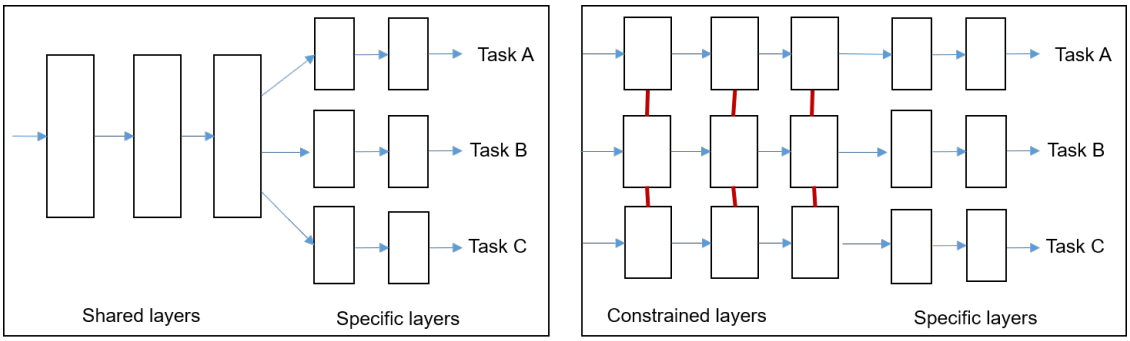


Figure 2.8: Hard (left) vs soft (right) parameter sharing in multi-task neural networks.

2.4.4 Learning what to share

A natural question arises when considering multi-task neural networks: what and with whom to share representations ? This question is an active research topic (see task grouping, [Ruder, 2016] for review). For instance, some studies suggest an adaptive approach to learning the sharing configuration. First, training the networks starting with zero sharing. Then, grouping tasks with similar parameters and repeating the process until the predictive performances deteriorate consequently. For the purpose of this thesis applications, we will address the problem of sharing architecture in the same way as hyper parameter selection, using cross-validation and grid search or random search to find the optimal sharing configuration.

2.5 Machine learning interpretability

2.5.1 Overview

In many scientific disciplines including ecology, there is a shift from qualitative to quantitative methods. Modern scientific problems are solved using big datasets enabled by advances in sensors and data capture, analyzed with black box machine learning models. However, when such models are used in research, scientific findings remain unjustifiable if the models only give predictions without explanations. To facilitate learning and satisfy curiosity as to why certain predictions are made by machines, interpretability and explanations are crucial.

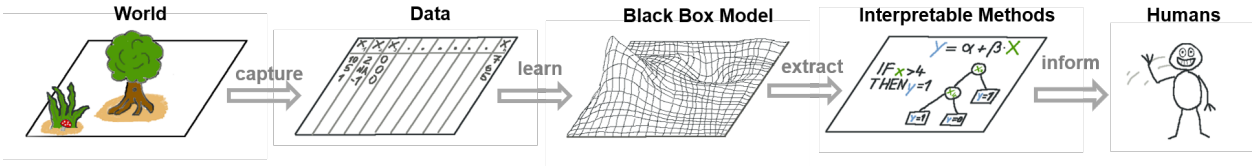


Figure 2.9: The place of interpretability in the machine learning process. The real world (e.g nature, ecological community) processes are captured through measures or descriptions and digitised prior to being fed to the black box machine learning model. The interpretability methods operate on trained machine learning models, yielding useful insights to the human user. (Modified from Molnar [2020])

"Interpretability is the degree to which a human can understand the cause of a decision." [Molnar, 2020] The aim of interpretability in machine learning is to make explicit what patterns lead to the prediction so that the human analyst can track, understand and verify the inference process (decision) of a model. Interpretability and explainability can be use interchangeably. However, the term "explanation" is used when we attempt to explain an individual prediction.

Interpretability should not be confused with algorithm transparency. Algorithm transparency evaluates the learning algorithm in theory, regardless of the data it is applied to, whereas interpretability considers a specific model trained on some data and how individual predictions are made.

Classification of interpretability methods

Machine learning interpretability (*IML*) methods are classified depending on several criteria [Molnar, 2020]:

Intrinsic vs post-hoc : Intrinsic or ad-hoc methods apply to models that are interpretable due to their simple structure, such as sparse linear regression, short decision trees, decision rules, etc. In contrast, extrinsic or post-hoc methods analyze models after training.

Model-specific vs model-agnostic : Model-agnostic methods are applicable to all models. They work by analyzing the (input feature, output prediction) pairs. Model-specific tools are developed for a restricted class of models, they exploit model internals (learnt parameters) to explain predictions.

Local vs global Local methods explain single predictions whereas global methods explain the entire model behavior.

Result of the method : interpretability methods produce different types of outputs:

- **Feature summary statistics**. For instance, feature importance measures the contribution of a feature to the prediction of either a small group (usually one) of instances i-e *local importance*, or to the overall model performance i-e *global importance*.
- **Feature visualization**. For instance, response curves are used visualize the (marginal/conditional/differential) response of the model to variations of the input features,

- **Model internals** visualization including parameters (e.g weights, embeddings) and intermediary activations, specific to the machine learning model.
- **Data point**. This category includes methods that return data points (sampled from the training data or synthetic instances) to make a model interpretable. The data points should be interpretable. This works well for images and texts, but is less useful for tabular data with many features. This category includes counterfactual explanations, class prototypes (representative instance of a class), etc.
- **Local approximation** by an intrinsically interpretable model (local surrogate, LIME).
- **Global approximation** by an intrinsically interpretable model (global surrogate).

2.5.2 Permutation feature importance

Permutation feature importance is a **global** model-agnostic method, producing feature summary statistics. It measures the increase (or decrease) in the prediction error (or quality) of the model after applying a random permutation to the values of a feature of interest. A feature is "important" (resp. "unimportant") if shuffling its values increases (resp. does not change) the model error.

The permutation importance was first introduced as part of Random Forest [Breiman, 2001]. Altmann et al. [2010] proposed a heuristic to estimate p-values for importance values in Random Forest. Later on, Fisher et al. [2018] extended the method to work with any machine learning model (model-agnostic).

Algorithm 2: Permutation feature IMPortance (PIMP) [Fisher et al., 2018]

Inputs :
- X: input features matrix (p: number of features),
- y: true response vector,
- f : prediction function of the black-box model, $\hat{y} = f(x)$,
- $L(y, \hat{y})$: evaluation function (prediction quality or error).
- R: number of repetitions to assess the statistical strength of importance values.

Outputs $I^{(avg)}, I^{(sd)}$: vectors of (average and standard deviation of) feature importance values.
:

FI : vector of size R;
/* Initialization */
$I^{(avg)} := 0$;
$I^{(sd)} := 0$;
Estimate the original error: $e = L(y, f(x))$;
foreach feature $j = 1 \dots p$ **do**
  $FI := 0$;
  foreach $r = 1 \dots R$ **do**
    /* Generate feature matrix by permuting the feature column j in original data. */
    Generate a random permutation of instances (rows) σ;
    $\tilde{X}_j := \sigma(X_j)$, s.t X_j is the jth column in X;
    $\tilde{X} := \left(X_1, \dots, X_{j-1}, \tilde{X}_j, X_{j+1}, \dots, X_p\right)$;
    /* Compute importance */
    Estimate error based on predictions on permuted feature matrix $\tilde{e} := L(y, f(\tilde{X}))$;
    Calculate feature importance: $FI_r := \tilde{e}/e$ or $FI_r := \tilde{e} - e$;
  end
  /* Compute empirical mean and standard deviation */
  $I_j^{(avg)} = mean(FI)$;
  $I_j^{(sd)} = stddev(FI)$;
end

Multi-output feature importance

Most of the problems addressed in subsequent chapters relate to predicting multiple entities simultaneously using models with multiple responses, a.k.a multi-output models. Assuming that an error measure is available for each response variable L_k, Algorithm 2 can be applied to each output separately or simultaneously. An overall feature importance for all outputs can be estimated in a similar fashion by assuming an aggregated evaluation measure of error (e.g sum of individual errors) or quality (e.g macro-average or micro-average AUC, recall, ...).

Group feature importance

To compute the importance of a feature, the algorithm breaks its association to the output by shuffling its values. When two or more features are correlated, permuting one while keeping the other(s) in the original order can yield unrealistic feature vectors.

Consider two features Weight(kg) and Height(cm), and the following synthetic instances:

T(toddler): (Height=0.86, Weight=12.5)
K(kid): (Height=1.2, Weight=23)
A(adult) (Height=1.8, Weight=70)

Permuting values of the weight column can yield unrealistic instances (e.g [Height=0.86,Weight=70]). This problem is relevant for all permutation-based approaches. To address this issue within the PIMP framework, we group features per degree of correlation. For each group of features, the permutation is applied jointly to the corresponding columns to evaluate a group-wise importance.

Tricks to improve the efficiency

The complexity of the permutation importance algorithm is linear on the cost c of a call to the black-box model prediction. For r repetitions, p features, and n data instances, the overall complexity is $\mathcal{O}(cnpr)$. In practice, we can run most of the computations (Algorithm 2) in parallel since importance calculation of different features as well as repetitions are independent. For multioutput machine learning models, we can estimate the importance of features for all response variables in the same loop to optimize the number of calls to the prediction function of the multioutput model.

2.5.3 Surrogate models

A surrogate model is an interpretable model (linear model, decision tree, decision rules, etc.) that is trained to approximate the predictions of a black box model. The choice of the black box model type and of the surrogate model type is decoupled. The true responses and the performance of the black box model do not play a role in training the surrogate model. The interpretation of the surrogate model makes statements about the model and not about the real world [Molnar, 2020].

Global surrogate

The goal of a global surrogate is to approximate the black box model's prediction function f as closely as possible with the surrogate model prediction function g, under the constraint that g is interpretable.

1. Select a dataset of features X. This can be the training dataset of the black box model, a subset of the training dataset or a new dataset from the same distribution.
2. Obtain the predictions of the black-box model on this dataset $\hat{y} = f(X)$.
3. Train an interpretable surrogate model on the dataset $(X, \hat{y})$.
4. Evaluate the quality of the approximation $L(\hat{y}, g(X))$.
5. Interpret the surrogate model.

Local surrogate

Local surrogate models are interpretable models that are used to explain the predictions of black box machine learning models on individual instances. For each instance, a local surrogate is built as opposed to a single model in the global surrogate approach.

Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016] is a framework that implements local surrogate models. The goal is to understand how the black-box machine learning model arrived to a certain prediction on a particular instance. LIME tests the response of the black-box model to variations of the instance of interest. Concretely, LIME works as follow:

1. Select a instance of interest x to explain.
2. Generate synthetic samples $\tilde{X}$ by applying local variations to the instance x.
3. Compute the prediction of the black-box model on the sampled instances $\tilde{y} = f(\tilde{X})$
4. Assign to each synthetic instance a weight proportional to its proximity to the instance of interest. w_x is the resulting vector of weights.
5. Train a weighted interpretable surrogate model g on the dataset $(\tilde{X}, \tilde{y})$, while keeping the surrogate model complexity $\Omega(g)$ low: $explanation(x) = \arg\min_{g \in G} L(f, g, w_x) + \Omega(g)$
6. Explain the prediction by interpreting the local model.

The surrogate model should be a good approximation of the machine learning model predictions locally, but it does not have to be a good global approximation. This kind of accuracy is also called **local fidelity**.

To generate local variations of the focal instance, LIME authors proposed different approaches per type of data: tabular, text, image. Here, we describe the method used for tabular data with each row representing an instance and each column a feature.

Most machine learning models are smooth, meaning that if the feature values are slightly perturbed, the model prediction does not vary much $f(x+\epsilon) \approx f(x)$. For this reason, LIME samples in tabular data are not generated by adding noise to the feature of interest but rather sampled from the training data distribution of the features. This increases the probability that the result for some of the sample points predictions differ from the data point of interest and that LIME can learn at least some explanation.

2.5.4 Shapley values

For a given instance, highlighting the relative contribution of a feature value to the prediction compared to the average prediction, is of major importance. In linear regression models, the answer is straightforward. The contribution of each feature is the weight of the feature times the feature value. For more complex models, a model-agnostic approach is needed.

$$f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$
$$\phi_j(f) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

In coalitional game theory, Shapley values [Shapley, 1953] provides a way to fairly distribute the payout among multiple cooperating players in a game. Štrumbelj and Kononenko [2014] proposed the use of Shapley values to evaluate the contribution of individual features, assuming the "game" is the prediction, the "payout" is the black-box model's prediction on the focal instance minus the average prediction for all instances, the "players" are the feature values of the instance that are weighted and aggregated to make the prediction. As the method is applied to single instances, the absolute Shapley value is a measure of **local importance**.

Computing Shapley values

The Shapley value of a feature value is its contribution to the payout, weighted and summed over all possible feature value combinations (coalitions):

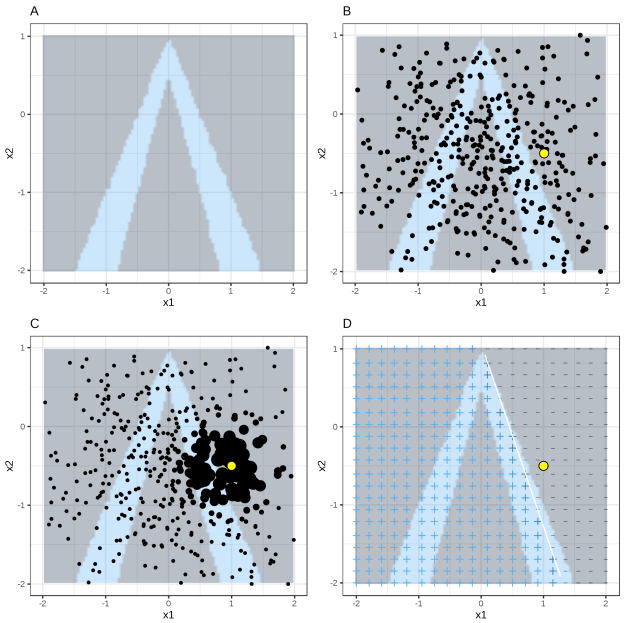


Figure 2.10: LIME algorithm for tabular data. A) Predictions given features x1 and x2. Predicted classes: 1 (dark grey) or 0 (light blue). B) Instance of interest (yellow dot) and data sampled from a normal distribution (small dots). C) Assign higher weight to points near the instance of interest. D) Signs of the grid show the classifications of the locally learned linear model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5). (Reprinted from Molnar [2020])

$$\begin{aligned}\phi_j(val_x) &= \sum_{S\subseteq\{x_1,\ldots,x_p\}\setminus\{x_j\}} \frac{1}{\binom{p}{1}} \frac{1}{\binom{p-1}{|S|}} \left(val_x\left(S\cup\{x_j\}\right) - val_x(S)\right) \\ &= \sum_{S\subseteq\{x_1,\ldots,x_p\}\setminus\{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(val_x\left(S\cup\{x_j\}\right) - val_x(S)\right)\end{aligned}$$

where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained and p the total number of features. val_x is the payout function for a coalition of feature values. It yields the prediction for feature values in set S that are marginalized over features that are not included in set S.

$$val_x(S) = \int f(x_1,\ldots,x_p)d\mathbb{P}_{x\notin S} - E_X(f(X))$$

All possible coalitions (sets) of feature values have to be evaluated with and without the j-th feature to calculate the exact Shapley value. For more than a few features, computing the exact solution to this problem becomes intractable as the number of possible coalitions exponentially increases as more features are added. Štrumbelj and Kononenko [2014] proposed an approach relying on Monte-Carlo sampling in the set of possible coalitions:

$$\hat{\phi}_j = \frac{1}{M}\sum_{m=1}^{M}\left(f(x^m_{+j}) - f(x^m_{-j})\right)$$

At each iteration m, a random instance z is drawn from the dataset. The algorithm then draws a coalition S from the features except the feature of interest. x^m_{-j} is built using values in x for features in S and using values in z for the others. x^m_{+j} is constructed similarly except the value of the jth feature is chosen from x. The difference of prediction on the two constructed instances yields the marginal contribution for the sampled coalition. The Shapley value is the average of marginal contribution across all samples.

Properties of Shapley values

The Shapley value satisfies the following properties:

- **Dummy**: A feature that does not change the predicted value, regardless of which coalition of feature values it is added to, should have a Shapley value of 0.
- **Efficiency**: The feature contributions must add up to the difference of prediction for the focal instance and the average.
- **Symmetry**: The contributions of two feature values should be the same if they contribute equally to all possible coalitions.
- **Additivity**: For a game with multiple combined payouts (e.g ensemble average) the respective Shapley values are obtained by averaging the Shapley values on individual payouts.

Unlike LIME, Shapley values are backed by a solid theory. However, it returns a simple value per feature but no prediction model as opposed to LIME. This means it cannot be used to assess changes in prediction for changes in the input. Moreover, explanations created with the Shapley value method always use all the features, rendering this approach ill-suited for applications where users seek sparse explanations. SHAP, an alternative method to estimate Shapley values addresses some of these issues as it connects Shapley values and LIME.

2.5.5 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations [Lundberg and Lee, 2017] represents the Shapley value explanation as a linear model. This view connects LIME and Shapley Values.

In SHAP, we represent a coalition S by a binary vector $z \in \{0,1\}^P$ of dimension equal to the maximum coalition size P, such that $z_j = 1$ is the jth feature is part of the coalition, $z_j = 1$ otherwise.

$$g(z) = \phi_0 + \sum_{j=1}^{P} \phi_j z_j$$

The end goal is to estimate the individual feature contributions (Shapley values) ϕ_j.

KernelSHAP

KernelSHAP estimates for an instance x the contributions of each feature value to the prediction. It consists of the following steps:

1. Sample coalition $z_k \in \{0,1\}^P$ for $k = 1 \dots M$.
2. Convert each coalition vector to a feature vector: $\hat{x}_k = h_x(z_k)$ s.t $h_x : \{0,1\}^M \mapsto \mathbb{R}^p$. The function h_x maps 1's to the corresponding value from the instance x that we want to explain. For tabular data, it maps 0's to the values of another instance sampled from the data.
3. Get the prediction of the black-box model on the constructed instances $\hat{y}_k = f(\hat{x}_k)$
4. Compute the weight for each coalition vector z_k using the SHAP kernel: $w_k = \pi_x(z_k)$. The SHAP

kernel is computed as follows:

$$\pi_x(z) = \frac{(P-1)}{\binom{P}{|z|}|z|(P-|z|)}$$

5. Fit a weighted linear model on the dataset $(Z, \hat{Y}, w)$:

$$\begin{aligned} L(f, g, \pi_x) &= \sum_{z \in Z} [f(h_x(z)) - g(z)]^2 \pi_x(z) \\ &= \sum_{k=1}^{M} [\hat{y_k} - g(z_k)]^2 w_k \end{aligned}$$

6. Return the coefficients ϕ of the linear model.

Global interpretation of SHAP values

Shapley values can be combined into global explanations. If we run SHAP for every instance, we get a matrix of Shapley values. This matrix has one row per data instance and one column per feature. We can interpret the model globally by analyzing the Shapley values in this matrix.

SHAP feature importance To obtain the global importance for each feature, we average the absolute Shapley values per feature across the data instances.

$$I_j = \sum_{i=1}^{n} |\phi_j^{(i)}|$$

This measure is an alternative to permutation importance. Permutation feature importance is based on the decrease in model performance whereas SHAP is based on the magnitude of feature attributions.

SHAP interaction The interaction effect is defined as the additional combined feature effect after accounting for the individual feature effects. For two features of interest $i \neq j$:

$$\begin{aligned} \phi_{i,j} &= \sum_{S \subseteq \setminus \{i,j\}} \frac{1}{\binom{p}{2}\binom{p-2}{|S|}} \delta_{ij}(S) \\ &= \sum_{S \subseteq \setminus \{i,j\}} \frac{|S|!(p-|S|-2)!}{2(p-1)!} \delta_{ij}(S) \\ \delta_{ij}(S) &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \end{aligned}$$

To compute the interaction effect in each coalition $\delta_{ij}(S)$, we subtract the main effect of the features so that we get the pure interaction effect after accounting for the individual effects. Afterwards, we average the values over all possible feature coalitions S, as in the Shapley value computation.

Clustering SHAP values The goal of clustering is to find groups of similar instances. SHAP clustering works by clustering on Shapley values of each instance. This means that we cluster instances by explanation similarity. Unlike clustering on the original feature space, SHAP values have the same unit (the unit of the prediction space). Any clustering algorithm can be applied at this point.

2.5.6 Response curves

Once important features are selected for each prediction task, response curves are feature visualization techniques which allow to visualize the shape (monotonic, linear, quadratic, complex, etc.) of the prediction along the domain of the features of interest (two at most simultaneously). The following methods work for both continuous and categorical features. They are all classified as global methods, since they describe the global relationship of a feature with the predicted outcome.

SHAP dependence plots

When SHAP valeus are computed for each instance and each features, we can visualize SHAP values as a function of feature values i-e $\{(x_j^{(i)}, \phi_j^{(i)})\}_{i=1}^n$. The resulting graph is known as SHAP dependence plot. We can fit a curve to this plot or make do with a scatter plot.

Evaluation strip

Developed in the context of Species Distribution Models [Elith et al., 2005], the *evaluation strip* (ES) produces a response curve by holding all variables other than the variable of interest constant (at their mean or some other appropriate value), and making predictions at regular intervals over the range of the focal variable.

Given a vector of reference values z (e.g mean feature values), $S = \{j\}$ the feature of interest and $C = S - \{j\}$ the rest of the features.

$$f^{ES}(X_S) = f(X_S, z_C)$$

Partial dependence plot

The *partial dependence plot* (PD) shows the marginal effect of one or two features on the model predictions Friedman [2001]. It works by marginalizing the model output over the distribution of non-focal features C in order to show the relationship between the feature(s) of interest S and the predicted outcome. S contains at most two features.

$$f^{PD}(X_S) = E_{X_C}\left[f(X_S, X_C)\right] = \int f(X_S, X_C) d\mathbb{P}(X_C)$$

A Monte-Carlo approach is typically used to approximate the previous integral:

$$\hat{f}^{PD}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)})$$

While for numerical features, the PD computes the partial dependence at regular intervals to produce the PD plot. With categorical features, the PD computation are conducted for each category of the feature. An assumption of the PD is that the features in C are not correlated with the features in S. If this assumption is violated, the averages calculated for the partial dependence plot will include data points that are very unlikely or even impossible.

Bias in ES and PD plots

When computing the value of the response curve at a given value of the focal feature, the ES and PD aggregate predictions on artificial instances that can be physically impossible or unrealistic if they are generated without accounting for feature correlation Molnar [2020]. The PD shows what the model predicts on average when each data instance has the value v for the focal feature. It ignores whether the value v makes sense for all data instances. On the other hand, the ES gives the model prediction of an average instance at various levels v of the focal feature. If the value v is incompatible with the other feature values, we fall into the same trap as for PD.

Accumulated local effects

Accumulated local effects(ALE) [Apley and Zhu, 2016] describe how features influence the prediction of a machine learning model on average. ALE plots are a faster and unbiased alternative to partial dependence plots (PDPs). Unlike PDP, the ALE computation relies on conditional expectation to account for the correlations of the focal feature with other features (see Fig 2.11 for an illustration).

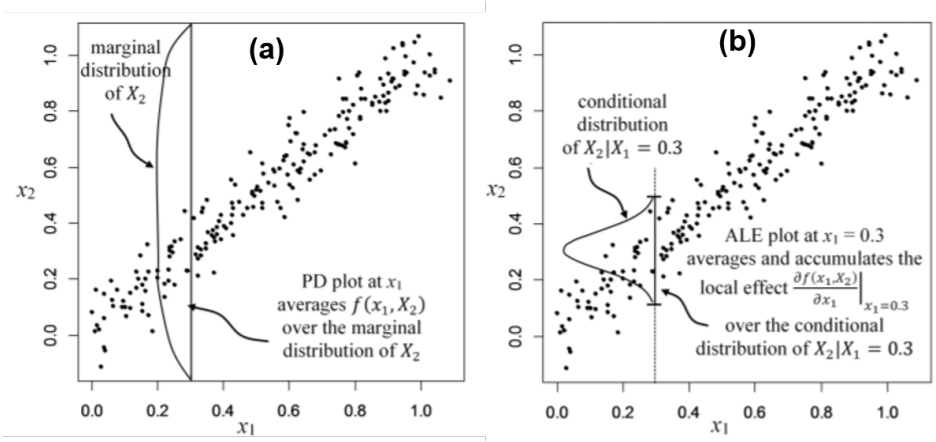


Figure 2.11: Difference between marginal computations of Partial Dependence Plots (PDP), and conditional computations of Accumulated Local Effects (ALE). (Modified from Apley and Zhu [2016])

Concretely, for a given value v of the focal feature, the ALE computes the **change** in model predictions in a **small window** of the focal feature around v for data instances in that window. By averaging changes, the ALE gets rid of confounding effects of other features and estimates the main effect of a feature locally. The main effects are then centered by subtracting the global average, resulting in a plot centered on zero. Finally, local effects are accumulated to facilitate the visualization of the underlying global effect.

Mathematical formulation

$$f^S(x_S, x_C) = \frac{\delta f(x_S, x_C)}{\delta x_S} \tag{2.1}$$

$$f^{ALE}(x_s) = \int_{z_{0,1}}^{x_S} E_{X_C|X_S}\left[f^S(X_s, X_c)|X_S = z_S\right] dz_S - \text{constant} \tag{2.2}$$

$$= \int_{z_{0,1}}^{x_S} \int_{x_C} f^S(z_s, x_c)\mathbb{P}(x_C|z_S)dx_C dz_S - \text{constant} \tag{2.3}$$

ALE for a single numerical feature First, we formulate the computations for a single numerical feature (Fig 2.12). Let j be the feature of interest. We assume that the feature has a compact support: $\mathcal{S}_j = [x_{min,j}, x_{max,j}]$. Let $\mathcal{P}_j^K \equiv \{z_{k,j}^K : k = 0, 1, \ldots, K\}$ be a partition of $\mathcal{S}_j$ into K intervals, with $z_{0,j} = x_{min,j}$ and $z_{K,j} = x_{max,j}$. It follows that: $\lim_{K\mapsto\infty} \max_K(|z_{k,j}^K - z_{k-1,j}^K|) = 0$. For any $x \in \mathcal{S}_j$, $k_j^K(x)$ is the index of the interval in $\mathcal{P}_j^K$ in which the value of x falls i-e $x \in (z_{k-1,j}^K, z_{k,j}^K]$ for $k = k_j^K(x)$. We note $n_j(k)$ the number of training observations that fall into the kth interval $N_j(k) = (z_{k-1,j}^K, z_{k,j}^K]$.

Eq 2.3 can be generalized to non-differentiable prediction functions using limiting finite differences (Eq 2.4), whereas integrals are approximated using summations (Eq 2.5).

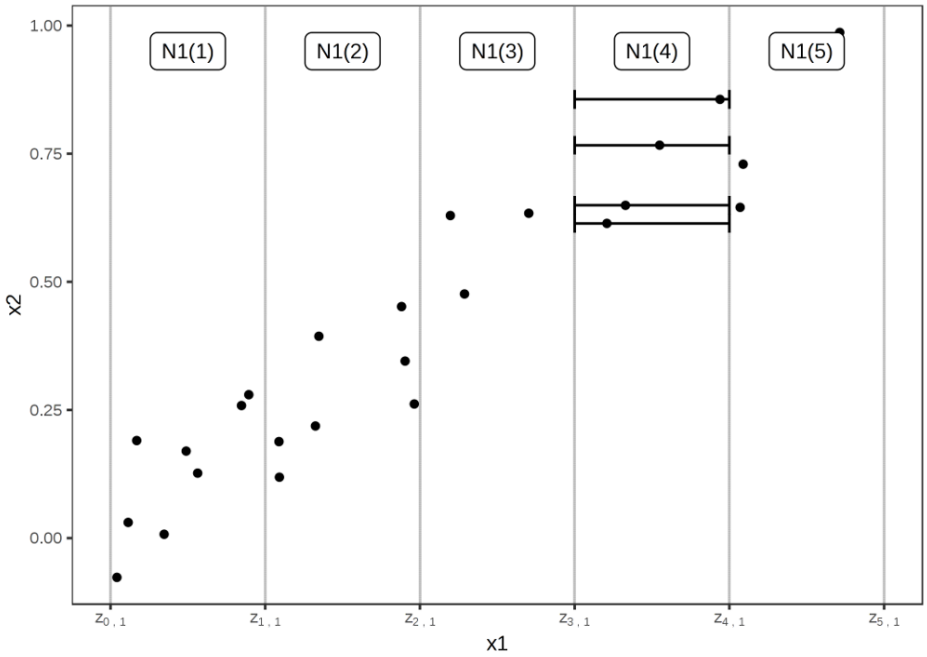


Figure 2.12: Calculation of ALE for feature x1, which is correlated with x2. First, we divide the feature into intervals (vertical lines). For the data instances (points) in an interval, we calculate the difference in the prediction when we replace the feature with the upper and lower limit of the interval (horizontal lines). These differences are later accumulated and centered, resulting in the ALE curve. (Source: Molnar [2020])

$$g_j(x) \equiv \lim_{K \mapsto \infty} \sum_{k=1}^{k_j^K(x)} \mathbb{E}\left[f(z_{k,j}^K, X_{\setminus j}) - f(z_{k-1,j}^K, X_{\setminus j}) \mid X_j \in (z_{k-1,j}^K, z_{k,j}^K]\right] \tag{2.4}$$

$$\hat{g}_j(x) = \sum_{k=1}^{k_j^K(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}^K, x_{\setminus j}^{(i)}) - f(z_{k-1,j}^K, x_{\setminus j}^{(i)})\right] \tag{2.5}$$

The function g computes the uncentered ALE effects. To obtain the centered ALE plot (i-e with mean zero), we simply substract the expected value of g (Eq 2.6). Thus, the value of the ALE is interpreted as the main effect of the feature at a certain value compared to the average prediction of the data.

$$f^{ALE}(x_j) \equiv g_j(x_j) - \mathbb{E}\left[g_j(X_j)\right] \tag{2.6}$$

$$= g_j(x_j) - \int p_j(z_j) g_j(z_j) dz_j \tag{2.7}$$

$$\hat{f}^{ALE}(x_j) = \hat{g}_j(x_j) - \frac{1}{n} \sum_{k=1}^{K} n_j(k) \hat{g}_j(z_{k,j}) \tag{2.8}$$

ALE plot for the interaction of two numerical features 2D ALE plots show the additional interaction effect for two features (second-order effect). The calculation principles are the same as for a single feature, but we work with rectangular cells instead of intervals, because we have to accumulate the effects in two dimensions (Fig 2.13). The estimation is conducted similarly to the 1D configuration. Eq 2.9 estimates the uncentered effect $\hat{h}$ as a summation of second-order finite differences. In Eq 2.10, $\Delta_f^{(j,l)}(K, k, m; x_{i,\setminus\{j,l\}})$ is the second-order finite difference evaluated at the ith observation. Eq 2.11 computes the centered effect.

$$\hat{h}_{(j,l)}(x_j, x_l) = \sum_{k=1}^{k_j(x_j)} \sum_{m=1}^{k_l(x_l)} \frac{1}{n_{(j,l)}(k,m)} \sum_{i:x_{i,(j,l)} \in N_{(j,l)}(k,m)} \Delta_f^{(j,l)}(K,k,m; x_{i,\setminus\{j,l\}}) \tag{2.9}$$

$$\begin{aligned} \Delta_f^{(j,l)}(K,k,m; x_{i,\setminus\{j,l\}}) = &[f(z_{k,j}, z_{m,l}, x_{i,\setminus\{j,l\}}) - f(z_{k-1,j}, z_{m,l}, x_{i,\setminus\{j,l\}})] \\ &- [f(z_{k,j}, z_{m-1,l}, x_{i,\setminus\{j,l\}}) - f(z_{k-1,j}, z_{m-1,l}, x_{i,\setminus\{j,l\}})] \end{aligned} \tag{2.10}$$

for each $(x_j, x_l) \in (z_{0,j}, z_{K,j}] \times (z_{0,l}, z_{K,l}]$.

$$f_{(j,l)}^{ALE}(x_j, x_l) = \hat{h}_{(j,l)}^{ALE}(x_j, x_l) - \frac{1}{n} \sum_{k=1}^{K} \sum_{m=1}^{K} n_{(j,l)}(k,m) \hat{h}_{(j,l)}(z_{k,j}, z_{m,l}) \tag{2.11}$$

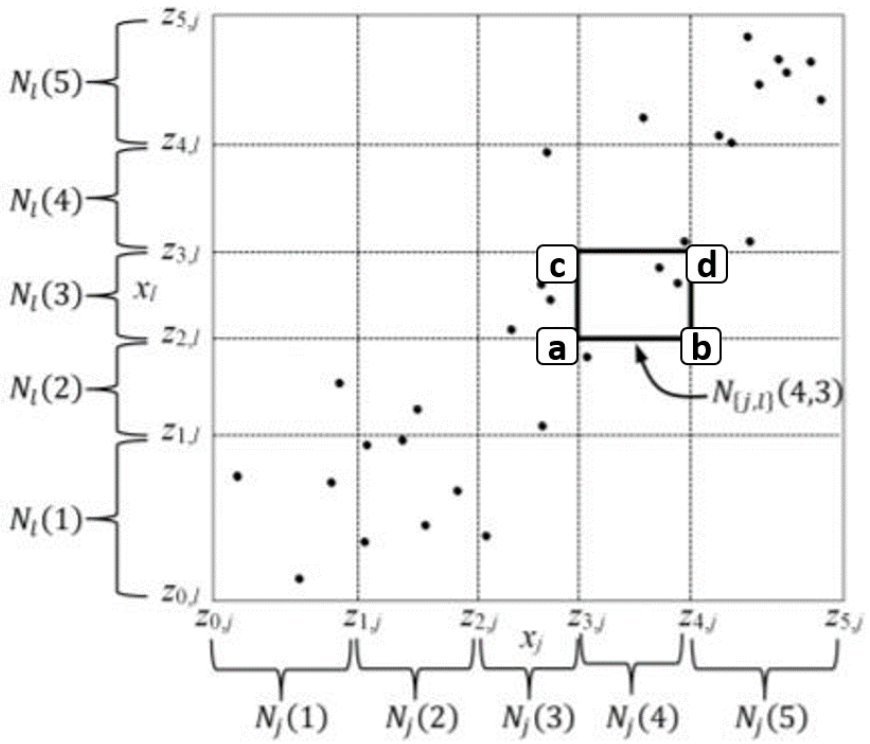


Figure 2.13: Calculation of 2D-ALE. We place a grid over the two features j and l. In each grid cell we calculate the 2nd-order differences for all instance within. We first replace values of x1 and x2 with the values from the cell corners. If a, b, c and d represent the "corner"-predictions of a manipulated instance (as labeled in the graphic), then the 2nd-order difference is (d - c) - (b - a). The mean 2nd-order difference in each cell is accumulated over the grid and centered. (Modified from Apley and Zhu [2016])

ALE plot for a single categorical feature In the case of ordinal features, the previous process is applicable. For categorical features, an order of categories has to be provided. Otherwise, the algorithm computes the distance between instances with respect to the other (non-focal) features, followed by a multi-dimensional scaling to reduce the distance matrix into a one-dimensional distance measure. This process yields a similarity-based order of the categories.

Synthesis: interpreting response curves

- Evaluation Strip (ES): "what the model predicts for the value v if all other features are kept at a constant value (average, median)."
- Partial Dependence Plots (PDP): "what the model predicts on average when each data instance has the value v for that feature, ignores whether the value v makes sense for all data instances."
- Accumulated Local Effects (ALE): "how the model predictions change in a small "window" of the feature around v for data instances in that window."
- SHAP dependence plot: "what is the contribution of the feature to the prediction on v compared to the average prediction"

2.6 Conclusion

In this chapter, we presented the building blocks of the methodological developments proposed in this thesis. First, we introduced Generalized Linear Models (GLMs) which define our response model. Afterwards, we described neural networks as extensions of GLMs with a built-in upstream feature learning mechanism. We described two neural architectures used for supervised learning: multi-layer perceptrons and convolutional neural networks. The latter uses particular operators such as the convolution and the pooling to learn features from images and grid-like data efficiently. We briefly introduced autoencoders and variational autoencoders, two architectures for unsupervised deep learning. The latter will be detailed in the following chapter where it is used as an inference tool. We also introduced an brief overview of multi-task learning in the context of neural networks and deep learning. Finally, we reviewed principles and methods of machine learning interpretability that are relevant for subsequent applications. In the following chapter, we will highlight how the various methods of interpretability can be applied to answer specific questions in community modeling.

Chapter 3

Multi-task modeling of ecological communities

3.1 Introduction

A paramount question in biogeography is to understand and predict the response of ecological communities to environmental conditions. The ecological communities response can be captured by multiple indicators describing various facets of diversity: taxonomic (e.g species abundances, biomass), functional (e.g traits), network (e.g biotic interactions, co-occurrences). Indicators are dependent, due to direct interactions or retroactive effects on the environment. Such dependencies are encoded in networks of associations. Considering each indicator as a response variable in a prediction task, we obtain a multi-task learning problem.

Very often, environmental conditions are described by features encoded in a raster format with varying resolutions. In addition to the previous computed information, rich contextual information can be obtained from aerial/satellite images. In this case, relevant features are typically extracted from raw images using signal processing techniques or using convolutional neural networks [LeCun et al., 1995] as upstream feature extraction components. Botella et al. [2018] applied similar CNN-based architectures to multi-channel images constructed from the environmental rasters (one channel per raster), restricted to a radius around the point of prediction. Unfortunately, these feature learning approaches are incompatible with existing ecological community modeling frameworks.

We propose **M**ulti-**T**ask modeling of **E**cological **C**ommunities (MTEC), a methodological framework to model the response of ecological communities. It extends the traditional setting based on multivariate linear regression models [Niku et al., 2019, Ovaskainen et al., 2017] to allow the automatic extraction of features and non-linear responses via feedforward neural networks.

3.2 Generic framework

We illustrate the generic conceptual framework and the common probabilistic components shared by subsequent applications in Fig 3.1. To avoid excessive abstractions, without loss of genericity, we will describe the framework as used for predicting species abundances. The same mechanisms apply if the targets of the prediction are at other taxonomic levels (e.g family or subspecies), representing groups of taxa (e.g functional group), or a set of traits.

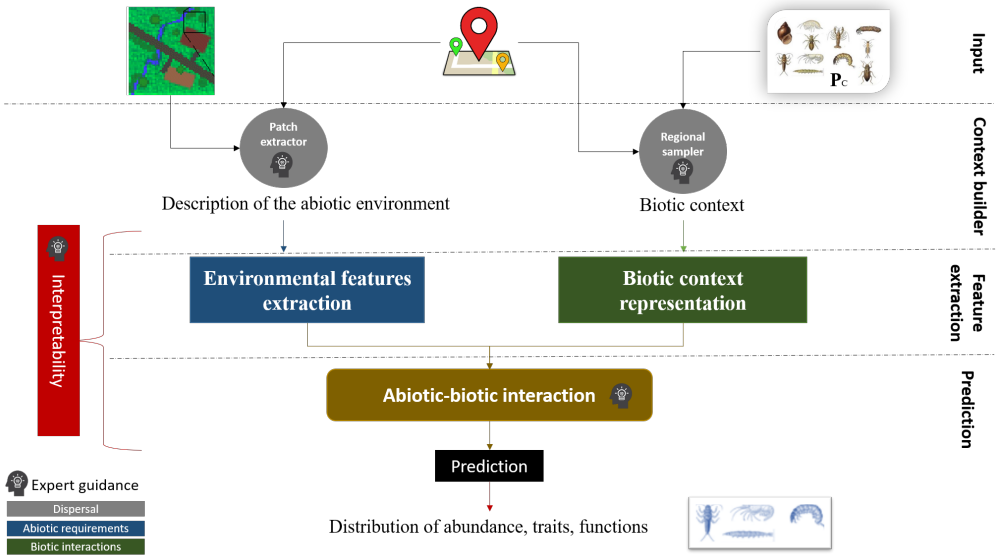


Figure 3.1: Architecture of MTEC. The framework is organized in tiers. The context builder selects the potential covariates: abiotic conditions as well as biotic actors (a.k.a biotic context) relevant to the prediction, and determines their spatial scope. The feature extraction tier transforms the raw covariates into meaningful features exploitable by the down-stream prediction tier. The latter aggregates the effect of features representing the environmental conditions and the biotic context to predict the response of the species (or traits, etc.)

3.2.1 Context builder

The framework takes as input a set of locations, environmental rasters and an occurrence database. The context builder leverages expert knowledge to select the subset of relevant abiotic and biotic factors for the prediction task.

Abiotic descriptors

Environmental data come in various formats and coverage, including metadata accompanying community data in tabular format or raster databases provided by third-party institutions. In the latter case, a preliminary step consists in extracting the portion of raster surrounding the location of interest around a radius in km or pixels on the map. We refer to it using the term *patch*. After normalization and missing values filling, patches can be used directly, or be transformed into single values using summary statistics (average, minimum, maximum, median).

Biotic context

The biotic context encodes our assumptions about the potential biotic effects which could affect the prediction at a given site. In the case of species abundance prediction, it imposes implicit constraints on the structure of species association networks by restricting the set of potential associations a priori. In the simplest case, without any prior knowledge, the biotic context consists of individuals from other species observed at the same site. Formally, the biotic context of species $i \in \mathcal{S}$ at site $s \in \mathcal{K}$, denoted $\mathcal{C}_{si}$, is defined as follows:

$$\mathcal{C}_{si} = \{j \in \mathcal{S}, j \neq i \text{ and } y_{sj} > 0\} \tag{3.1}$$

From this basic definition, we can create straightforward extensions, with corresponding representations (Box).

Biotic context extensions for species abundance prediction

Temporally-explicit extension When longitudinal data are available, we denote the abundance of species i at site s at time-point t as $y_{si}^{(t)}$. Accordingly, the definition of the biotic context for a target species at a given time-point is extended to contain the species, including the target, that were observed in the previous time-point:

$$\mathcal{C}_{si}^{(t)} = \{j \in \mathcal{S}, y_{sj}^{(t-1)} > 0\}$$
$$z_{si}^{(t)} = \frac{1}{\left|\mathcal{C}_{si}^{(t)}\right|} \sum_{j \in \mathcal{C}_{si}^{(t)}} y_{sj}^{(t-1)} \alpha_j$$

Spatially-explicit extension Given a function d that measures the distance between any pair of sites and a radius r, we consider a spatial extension of the base model where the biotic context contains species that were observed at locations within a distance r of the considered site.

$$\mathcal{C}_{si} = \{(j, l) \in \mathcal{S} \times \mathcal{K}, y_{lj} > 0 \text{ and } \mathrm{d}(s, l) \leq r\}$$

One can use multiple radius values, for instance customized to the dispersal abilities of each target group or species. The effect of each contextual element is weighted inversely to its distance to the target location. The hyperparameter τ controls the decrease in weight per unit of distance. Similarly, τ can be customized for each group of species based on expert knowledge.

$$z_{si} = \sum_{(j,l) \in \mathcal{C}_{si}} y_{lj} \cdot \exp(-\tau\, \mathrm{d}(s, l))$$

Graph extension So far, we defined the biotic context using the community composition in terms of species, possibly involving their abundances. At this point, we are able to capture pairwise additive effects. However, we miss the impact of interactions between context species or the whole network structure around the target location on the abundance distribution of the target group, referred to as the contextual network. Fortunately, graph embedding algorithms permit the incorporation of structured data such as knowledge graphs into predictive models. For instance, we can redefine the biotic context as the interaction network at site of interest s after removing the target species i, noted $G_{s/i}$. The context embedding is then obtained by applying a graph embedding function k with parameter θ on the contextual network.

$$z_{si} = \mathrm{k}(G_{s/i}; \theta)\,.$$

3.2.2 Feature extraction

This module maps raw data to meaningful features using a function whose parameters are learnt in an end-to-end learning way. These features represented in vector formats are shared across prediction tasks. The feature extraction function can be as simple as a lookup table returning a vector representation (embedding) given an index, we use this formulation to represent species in a biotic context (Eq 3.2). Conversely, a single layer neural network can be used for feature extraction, it acts as a Principal Component Analysis adjusted for all the prediction tasks (supervised PCA). Finally, when the spatial organization of feature values matters, convolutional neural networks can be used to extract meaningful features.

$$z_{si} = \frac{1}{|\mathcal{C}_{si}|} \sum_{j \in \mathcal{C}_{si}} y_{sj} \alpha_j \tag{3.2}$$

Covariate-dependent biotic representations

In the base model, the estimation of the biotic context effects is oblivious to the abiotic or biotic conditions surrounding it. To account for these neighborhood conditions, we extend the base model by allowing the embeddings used to represent the biotic context to depend on some chosen variables. Each site is associated to p conditioning covariates. These covariates are stored alongside an offset in a $n \times (p+1)$ matrix V, such that each of the first p columns of V contains the values of the corresponding covariates for the different sites while the last column is filled with ones. Then, given an embedding dimension d, the covariates are mapped to each dimension by applying a regression with a weight matrix $W \in \mathbb{R}^{d \times (p+1)}$. The resulting conditioning vectors are such that $\beta_s = W v_s^T$.

The extended biotic context is then represented as follows, where $*$ is the element-wise vector product:

$$z_{si} = \beta_s * \left(\frac{1}{|\mathcal{C}_{si}|} \sum_{j \in \mathcal{C}_{si}} y_{sj} \alpha_j\right) = \frac{1}{|\mathcal{C}_{si}|} \sum_{j \in \mathcal{C}_{si}} y_{sj} \cdot (\beta_s * \alpha_j)$$

3.2.3 Prediction

The prediction module combines the incident abiotic and biotic features using a parametric aggregation function to determine the target of the prediction, whether it is species abundances, trait distribution, etc. This aggregation function can be one of the following:

- *Hierarchical filter*: each factor operates as a selection filter, the result of the first filter is fed to the second filter,
- *Multiplicative filter (intersection)*: the product of both filters outcomes.
- *Additive or linear filter*: the abiotic and biotic effects are added, they can compensate each other,
- *Learnt filter*: e.g. a neural network

We provide some illustrative examples for the specific case of modeling species abundances in Fig 5.3.

We assume within this framework that target variables are independent conditionally to the abiotic and biotic features. Thus, we model each output variable as following a univariate distribution within the Exponential Family. As opposed to the feature extraction components that are shared across tasks, the prediction module has specific prediction parameters for each task. As with GLMs, the type of data considered (presence/absence, abundance, biomass, etc.) determines the choice of distribution and link function.

3.2.4 Opening the black-box

The flexibility of the generative model allows to improve its predictive accuracy, but it comes at the cost of a reduction in the interpretability of the model's decision process. Without altering the model, we propose to use a set of routines from the *machine learning interpretability* literature to shed light on the magnitude and direction of the contribution of each factor to the task prediction function. The mathematical details of the methods are described in Section 2.5.

Fig 3.2 illustrates the functioning of the interpretability module. A set of algorithms interact with the black-box model (here multi-task model) via calls to the prediction function and produce artefacts that support the interpretation of the model's outputs which can take different forms: global importance, surrogate model, response curves, etc. The parameters (model internals) of the black-box model in turn can be examined, as well its intermediary activations (outputs of successive hidden layers). The right-most column contains example of visualizations (a-j) representative of different interpretability tools/artefacts.

Artefact (a) uses shared environmental features for observations at various locations in a grid. These observations are clustered on the basis of the corresponding learnt features. The result is a spatial mapping of **environmental types/clusters** as seen by the modeled species. Artefact (b) uses abiotic response weights to visualize and assess species similarities identifying **response groups**. Artefacts (c,d) explore **feature**

importance globally (for all the observations) for **individual species** (c) or **communities** (d). Artefact (e) visualizes the identity of the most contributing feature per observation. Metadata of these observations' locations is used to map the dominant feature spatially. The resulting map highlights **limiting factors** and their geographic relevance. Artefact (f-h) represent response plots. (f) visualizes the **Accumulated Local Effect** (ALE), or the difference to the mean of the prediction at various values of the environmental feature. (g) showcases the effect (color scale) of two **features' interactions** on a species distribution (here Temperature and Relative humidity). (h) is a **SHAP dependence plot**. It visualizes the contribution of the feature (x-axis) to the model prediction at a given feature value. There is a point for each species,observation pair. Colors allow the differentiation of species. Here, species are colored per trait value. Artefact (i) is a **multispecies summary** of species (y-axis) responses (x-axis) to a particular feature at various values (color scale) of the focal feature. Artefact (j) is a PCA of the data in (i) per species, which aims to identify **feature(s)-specific responses groups**.

3.2.5 Incorporating expert knowledge

Expert guidance can be incorporated at various levels: from selecting relevant covariates and their scope, defining the biotic contexts, choosing an appropriate aggregation function to assessing model interpretability. Additionally, expert knowledge can be harnessed to reduce model complexity by exploiting known ecological similarities via the mechanisms of parameter sharing, at various levels:

- Environmental features sharing (Applied in chapter 4);
- Abiotic response sharing;
- Biotic representations sharing (Applied in chapter 3).

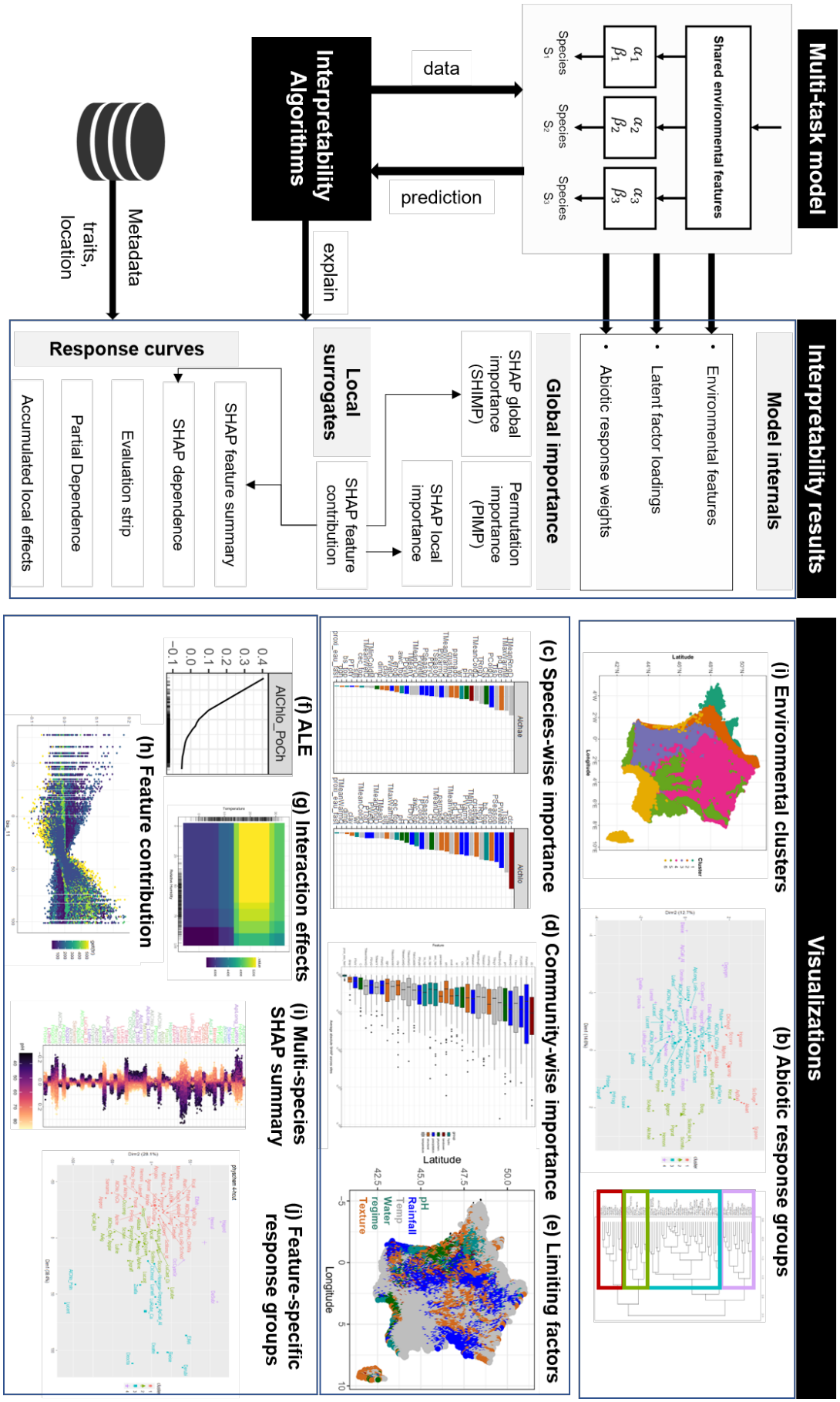


Figure 3.2: Interpretability methods for multi-task models.

3.3 Multi-task neural model of the joint environmental response

The goal is to predict the **joint response** of multiple ecological entities (e.g species or traits) to **common influencing factors** (e.g environment, other taxonomic groups). The method should satisfy the following requirements:

- (R1) Accommodate **various types of covariates**: continuous, ordinal, categorical.
- (R2) Address the potential for multicollinearity induced by the **large number of covariates**.
- (R3) Allow for **non-linear responses** to the environment.
- (R4) Support **heterogeneous responses**: binary, ordinal, categorical, continuous.
- (R5) Incorporate a mechanism to estimate **residual** (unexplained by measured covariates) **associations** between response variables.
- (R6) Enable an **efficient inference** procedure.

3.3.1 Generative model

Let us first formalize the general probabilistic model. We address requirements **R[1-3]** by using a neural network f^{EXT} with weights W dedicated to learning relevant representations of the environment $\mathrm{X} = (x_s)_{s \in S}$ for the prediction tasks from the raw input features $\mathrm{E} = (e_s)_{s \in S}$:

$$x_s = f^{EXT}(e_s). \tag{3.3}$$

We adopt a latent factor approach similar to Warton et al. [2015]. The response depends on the fixed effects induced by measured covariates and random effects arising from a set of K latent factors (Eq 3.5). We encode the dependence of each response variable y_i on the observed and latent factors in respectively the abiotic response $\mathrm{B} = (\beta_i)$ matrix and loadings matrix $\mathrm{A} = (\alpha_i)$. We further assume independence of latent factors (Eq 3.7), and responses conditioned on the latent variables (Eq 3.4). Dependencies between responses are mediated by the latent factors, and are thus encoded in the residual covariance matrix Σ.

$$y_{si} \perp\!\!\!\perp y_{sj} \mid h_s \quad , \forall i \neq j \tag{3.4}$$

To enable flexibility in response types (in accordance with **R4**), we model response variables in the exponential family of distributions (Eq 3.9).

$$g_i(m_{si}) = \beta_i x_s + z_{si} \tag{3.5}$$
$$z_{si} = h_s \alpha_i \tag{3.6}$$
$$h_s \sim \mathcal{N}(0, I_d) \tag{3.7}$$
$$z_{si} \sim \mathcal{N}(0, \mathrm{A}.\mathrm{A}') \tag{3.8}$$
$$y_{si} \mid h_s \sim \mathcal{F}_i(m_{si}, \phi_i) \tag{3.9}$$

In summary, the full generative model parameters are: $\theta = \{\mathrm{W}, \mathrm{B}, \mathrm{A}, \phi\}$.

3.3.2 Inference as optimization

The goal is to maximize the marginal likelihood w.r.t the generative parameters:

$$\log p_\theta(\mathrm{Y}) = \log \int p_\theta(\mathrm{Y}, \mathrm{H}) d\mathrm{H} \tag{3.10}$$

Naively, this requires computing an intractable integral to marginalize the unobserved latent variables.

$$\log p_\theta(\mathrm{Y}) = \log \int p_\theta(\mathrm{Y}, \mathrm{H}) d\mathrm{H} \tag{3.11}$$

To circumvent this issue, we rely upon a **variational approximation** of the true posterior of the hidden variables $p(\mathrm{H} \mid \mathrm{Y})$ within a class $\mathcal{Q}$ of distributions. Here, we choose the family of factorized Gaussian distributions.

$$\mathcal{Q} = \{q : q(\mathrm{H}) = \prod_{s=1}^{N} \mathcal{N}(h_s; \mu_s, diag(\sigma_s^2))\} \tag{3.12}$$

From Eq 3.12, the number of posterior parameters (means and variances) increases linearly with N the number of observations $O(2KN)$. To ensure the scalability of the inference, we follow the $AEVB$ algorithm[1] and use a sufficiently complex parametric function $f^{ENC} : y_s \mapsto (\mu_s, \sigma_s^2)$. In practice, f^{ENC} is a neural network parameterized with weights ξ, referred as *recognition or encoder network*. The combination of the encoder and the generative model (also called *decoder* (Eq 3.9), leads to a *variational autoencoder* [Kingma and Welling, 2013].

We transform the inference problem to the optimization (maximization) of the likelihood w.r.t the global parameters θ minus the information loss induced by the posterior approximation (Eq 3.13). The latter is measured by a divergence measure $\mathcal{D}$ and parameterized with the variational parameters. By setting $\mathcal{D}$ to the *Kullback-Leibler* divergence [Lin, 1991], we obtain the *Variational EM* algorithm which optimizes a lower-bound of the marginal log-likelihood, commonly known as the *evidence lower-bound* or *ELBO* [Yang, 2017].

$$\mathcal{J}(\mathrm{Y}; \theta, \xi) = \log p_\theta(\mathrm{Y}) - \mathcal{D}(q_\xi(\mathrm{H}) \quad || \quad p_\theta(\mathrm{H} \mid \mathrm{Y})) \tag{3.13}$$

$$= \log p_\theta(\mathrm{Y}) - \mathbb{E}_{\mathrm{H} \sim q(\mathrm{H}|\mathrm{Y})} \log \left(\frac{q_\xi(\mathrm{H} \mid \mathrm{Y})}{p_\theta(\mathrm{H} \mid \mathrm{Y})} \right) \tag{3.14}$$

$$= \log p_\theta(\mathrm{Y}) - \mathbb{E}_{\mathrm{H} \sim q(\mathrm{H}|\mathrm{Y})} \log \left(\frac{q_\xi(\mathrm{H} \mid \mathrm{Y}) \cdot p_\theta(\mathrm{Y})}{p_\theta(\mathrm{Y} \mid \mathrm{H}) \cdot p(\mathrm{H})} \right) \tag{3.15}$$

$$\mathcal{J}_{reg}(\mathrm{Y}; \theta, \xi) = \underbrace{\mathbb{E}_{\mathrm{H} \sim q(\mathrm{H}|\mathrm{Y})} \log p_\theta(\mathrm{Y} \mid \mathrm{H})}_{\text{Reconstruction } \mathcal{L}} + \underbrace{\mathcal{D}(p(\mathrm{H}) || q_\xi(\mathrm{H} \mid \mathrm{Y}))}_{\text{KL term } \mathcal{KL}} + \underbrace{\log p(\theta) + \log p(\xi)}_{\text{Parameter regularization } \mathcal{R}} \tag{3.16}$$

$$\tag{3.17}$$

The KL divergence between the prior Eq 3.7 and approximate posterior Eq 3.12 of the latent variables, both normally distributed, can be computed in closed form.

Kullback-Leibler divergence

The KL-divergence or relative entropy is a measure of how one probability distribution is different from a second reference distribution.
The KL-divergence of two multivariate normal distribution $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ is defined as:

$$KL(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = \frac{1}{2} \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{tr}\{\Sigma_2^{-1} \Sigma_1\} + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right].$$

We use the reparameterization trick to transform the reconstruction term into an expectation over an auxiliary noise that does not depend on the parameters and can be estimated using Monte-Carlo sampling, thus enabling gradient-based optimization.

[1] Autoencoding variational bayes

$$\mathcal{L}(y_s;\theta,\xi) = \mathbb{E}_{h_s\sim\mathcal{N}_K(\mu_s,diag(\sigma_s^2))}\left(\sum_{i=1}^{M}\log p_\theta(y_{si}\mid h_s)\right) \tag{3.18}$$

$$= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_K)}\left(\sum_{i=1}^{M}\log p_\theta(y_{si}\mid h_s=\mu_s+\sigma_s^2\odot\epsilon)\right) \tag{3.19}$$

$$\tilde{\mathcal{L}}(y_s;\theta,\xi) \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{M}\log p_\theta(y_{si}\mid h_s^{(l)}=\mu_s+\sigma_s^2\odot\epsilon^{(l)}) \quad \text{s.t} \quad \epsilon^{(l)}\sim\mathcal{N}(0,I_K) \tag{3.20}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{M}\log \mathbf{f}_i(m_{si}^{(l)},\phi_i) \quad / \tag{3.21}$$

s.t

$$\begin{gathered} m_{si}^{(l)} = g_i^{-1}(\beta_i x_s + \alpha_i(\mu_s+\sigma_s^2\odot\epsilon^{(l)})) \\ (\mu_s,\sigma_s^2) = f^{ENC}(y_s;\xi) \quad and \quad x_s = f^{EXT}(e_s;\theta) \end{gathered} \tag{3.22}$$

Given N data points, we construct an estimator of the reconstruction term using mini-batches of size N':

$$\tilde{\mathcal{L}}(\mathrm{Y};\theta,\xi) = \frac{N}{N'}\sum_{s=1}^{N'}\tilde{\mathcal{L}}(y_s;\theta,\xi) \tag{3.23}$$

Experiments from Kingma and Welling [2013] showed empirically that given a sufficiently large batch size, it is possible to use a single posterior sample, i-e $L = 1$. The objective function with the addition of regularization terms on the global parameters (θ,ξ) is optimized using Stochastic Gradient Descent [Bottou, 2010] or its variants [Ruder, 2016]. We illustrate the computation graph of the data generating process and the inference in Fig 3.3

Machine learning frameworks such as *Tensorflow* [Abadi et al., 2016] and *PyTorch* [Paszke et al., 2019] offer automatic differentiation and parallel optimization capabilities by exploiting hardware architectures of both CPUs and GPUs. We used the former library combined with *Keras* [Chollet et al., 2018], a high-level library to specify neural network architectures.

$$\mathcal{J}_{reg}(\mathrm{Y};\theta,\xi) = \tilde{\mathcal{L}}(\mathrm{Y};\theta,\xi) + \mathcal{K}(\mathrm{Y};\xi) + \mathcal{R}_\theta + \mathcal{R}_\xi \tag{3.24}$$

3.3.3 Sparse associations estimation

We define the following posterior statistics over the latent factors H:

$$\begin{array}{rll} \text{Variational mean matrix} & \mathcal{U} = (\mu_s)_{s=1}^N & N\times K \\ \text{Accumulated variance matrix} & \mathcal{S} = \sum_{s=1}^{N} diag(\sigma_s^2) & K\times K \\ \text{Latent covariance matrix estimator} & \hat{\Sigma} = \frac{1}{N}(\mathcal{U}'\mathcal{U}+\mathcal{S}) & K\times K \end{array}$$

From there, the estimated residual covariance matrix (Eq 3.25). We use the *Graphical lasso* algorithm [Friedman et al., 2008][2] to estimate a sparse inverse (Eq 3.26) for the posterior residual covariance matrix $\hat{\Omega}_r = \hat{\Sigma}_r^{-1}$, with a sample size of N. The association network estimated from the data is the Gaussian Graphical Model encoded in the corresponding partial correlation matrix $\mathcal{P}_r$.

[2]R package *glasso* https://CRAN.R-project.org/package=glasso

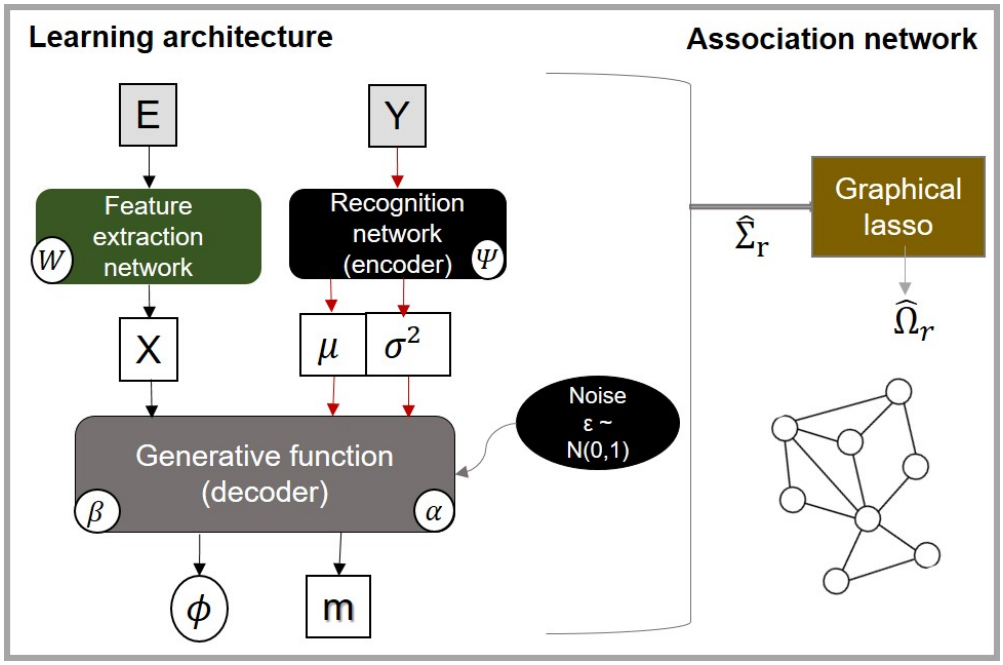


Figure 3.3: Learning architecture of the joint multi-task neural network model. Parameters are shown in white circles. Observed data (input) are highlighted with a grey background. Colored boxes represent the components of the model. Black (resp. red) arrows indicate the data generating workflow (resp. inference-related operations).

$$\hat{\Sigma}_r = \mathrm{A} \cdot \hat{\Sigma} \cdot \mathrm{A}' \qquad M \times M \tag{3.25}$$

$$\hat{\Omega}_r = \mathrm{argmin}_{\Omega \geq 0} \left(\mathrm{tr}(\hat{\Sigma}_r \Omega) - \log\det(\Omega) + \lambda \sum_{j \neq k} |\Omega_{jk}| \right) \qquad \lambda : \text{ penalizing hyperparameter.} \tag{3.26}$$

3.3.4 Extending Joint Species Distribution Models

We can recover the generalized linear latent variable JSDM used in [Warton et al., 2015, Niku et al., 2019, Ovaskainen and Abrego, 2020] by setting $f^{EXT} = Id : x \mapsto x$ and by choosing the link function g as in *Generalized linear models*.

While the latent variable are combined linearly to the environmental features similarly to aforementioned work, our proposal additionally allows non-linear responses to the environment, introduced via the f^{EXT} function. Note that, Eq 3.5 can be traded for a non-linear function $f^{DEC}_{\theta'} : (x_s, h_s) \mapsto m_s$. However, by doing so the derivation of species associations is not straightforward. Here, we assume species dependencies are linear and captured by the pairwise correlations of random effects Z.

3.4 Conclusion

In this chapter, we introduced a generic framework for modeling ecological communities using multi-task neural networks. We introduced a declination of the framework into a joint model of species distributions with support for non-linear responses, with an efficient inference algorithm. We illustrate the proposed solution on an empirical application in Chapter 6. In the following chapters, we describe the application of the framework to two other problems: location-based recommendation (Chapter 4) and ecological association inference (Chapter 5).

Chapter 4

Plant recommendation using environment and biotic associations

4.1 Introduction

Predicting the most likely species occurring in a given location is of great importance in biodiversity studies. This age-old task in biogeography can be formulated as learning a density function of the species over the geographic space from a set of observed geolocalized occurrences. In practice, due to sampling bias, limited examples and local habitat heterogeneity, geographic coordinates are not used directly as predictors. Instead, species abundance is modeled as a function of the environmental conditions at the given locations. Such models are called Species Distribution Models (SDMs) or Habitat Suitability Models (HSMs).

The local environment is usually described by abiotic features such as climate and pedology. Recently, more studies include biotic covariates in the form of other living beings abundance, motivated by the need to account for the dependencies between species that can affect their co-distributions. Indeed, two species distributions may be correlated indirectly through a latent abiotic variable or directly if they interact in some way that creates a dependency between their respective populations as in the case of plant-pollinators, host-parasites, predator-preys, etc.

The set of locations with suitable abiotic conditions for a given species define its *Grinnellian niche* [Grinnell, 1917]. On the other hand, the role it occupies within its community through feeding on other organisms or interacting with them defines its *Eltonian niche* [Elton, 2001] or its biotic requirements. The locations at the intersection between the two niches constitute the species' Fundamental niche. What we observe is an accessible subset of it called the *Realized niche* [Peterson et al., 2011].

In the context of the GeoLifeCLEF 2019 challenge, part of the LifeCLEF evaluation campaign [Joly et al., 2019], we evaluate two models on location-based plant recommendation. The first one relies purely on abiotic features while the second harnesses regional level co-occurrences of the target species with other organisms selected with expert assistance. We describe both solutions and discuss results obtained during training, validation and test phases.

4.2 Dataset description

The task organizers provided a training dataset containing about 280K observations from the Global Biodiversity and Information Facility database. Nearly 2M plant occurrences from automatic species identification of pictures produced in 2017–2018 by the smartphone application Pl@ntNet were added. We used the complete provided plant occurrences involving almost 3.5K plant taxa. Besides, 10M occurrences from other biodiversity kingdoms were also included. Finally, 33 environmental rasters covering the French territory were provided. They describe the climate, topology and pedological landscape. They were constructed from

various open datasets as explained in the protocol note [Botella et al., 2019].

4.3 Task and proposals

The task consists in training a model that predicts at a geographic location the dominant plant species. We formulate this task as a multi-class classification problem where the output class is the dominant plant identity. We evaluate two architectures for solving the task:

- *GrinnellNet*: A convolutional neural network architecture using environmental rasters. It aims to learn features of species Grinnellian niche.
- *EltonNet*: A species embedding network leveraging associations with nearest non-plant taxa occurrences. Its purpose is to identify community composition patterns that are positively associated with a specific plant species, possibly related to the Eltonian niche concept.

Hereafter, we describe both proposals and motivate our architecture, preprocessing and optimization choices.

4.3.1 GrinnellNet: a CNN with categorical rasters embedding

GrinnellNet's architecture (illustrated in Figure 4.1) is organized in a stack of components trained end-to-end: input preprocessing, feature extraction, feature interaction and classification. It takes environmental rasters as inputs and returns the identity of the dominant plant for each location.

Input preprocessing

Observing a species in a given location does not necessarily imply the suitability of the abiotic environment. Indeed, some species survive in locations with unfavorable conditions (known as sink locations) as long as new individuals continuously join the population from a nearby suitable habitat (source locations), through seed dispersal for instance via wind currents. These so called source-sink dynamics [Holt, 1985] ensure an indefinite sustainability of the sink populations despite unfavorable abiotic conditions. Consequently, failing to account for these spatial processes may result in overestimation of the species' environmental niche's breadth. In addition to the effects of such stochastic assembly processes, the geographic coordinates provided by smartphone devices come with measurement errors, which represent a further source of uncertainties that need to be accounted for.

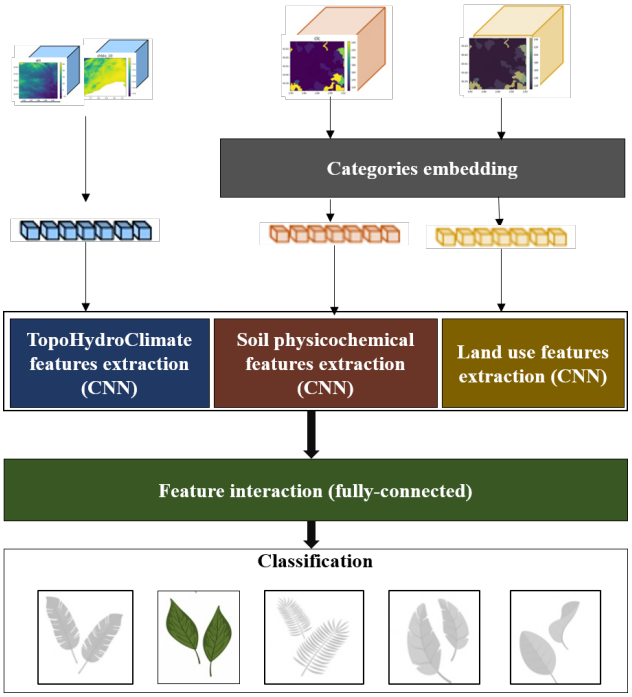
For these reasons, it is necessary to consider a landscapewise rather than a pointwise description of the environment around the observation geolocation by using environmental patches instead of local values.

We divide the environmental rasters into three groups based on their semantic, the resolution at which they vary and their data type (quantitative, ordinal and categorical).

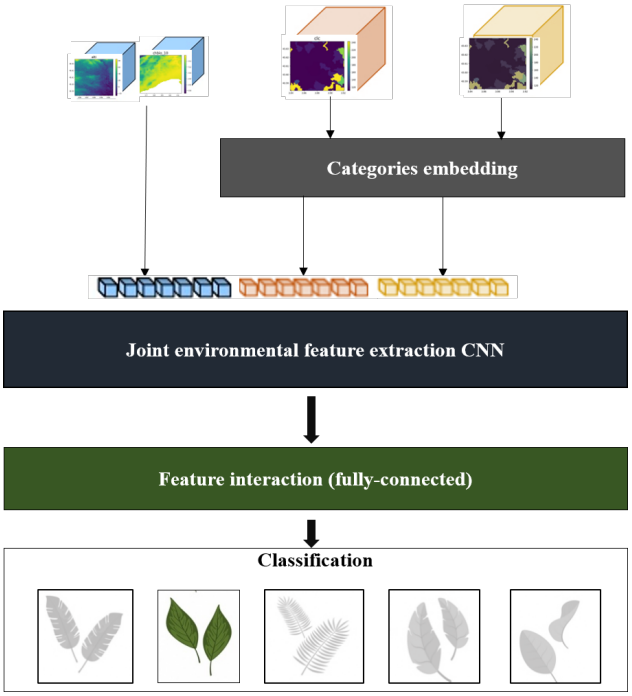
- **TopoHydroClimate** group: quantitative variables describing global bioclimate (CHBIO, ETP), hydrology (water proximity) and topology (altitude).
- **Pedology** group: ordinal and categorical variables describing the physico-chemical structure of the soil.
- **Land use** group: includes *Corine Land Cover* class.

Embedding categorical features

Part of the soil's physico-chemical properties are described by categorical or pseudo-ordinal features such as texture, land cover, erodibility and crusting class. Therefore, for each of the f_c categorical features, before feeding it to the CNN layers, we replace each of its n_c categories by a real vector representation of size k_c, where k_c is a tunable hyperparameter, typically chosen in the interval $[2, \frac{n_c}{2}]$. In practice, this is implemented by a feature-specific embedding lookup layer parameterized by a (n_c, f_c) matrix E_c, such that $E_c[i,]$ is the k_c sized embedding of the ith value of f_c. We apply this transformation batchwise in parallel to all patch cells (light grey module in Figure 4.1). For an input of dimension $(batchSize, patchRadius, patchRadius, 1)$,



(a) SepGrinnellNet: separate feature extraction modules



(b) JointGrinnellNet: joint feature extraction

Figure 4.1: GrinnellNet architecture

the embedding layer of f_c returns a $(batchSize, patchRadius, patchRadius, k_c)$ tensor. These vector representations are trained along with other network parameters.

Categorical embeddings capture richer relationships than raw categories. They are also considered as a dimensionality reduction technique, more practical than one-hot-encoding when dealing with high-cardinality yet sparse features (a typical example in our case is land cover). Note that each embedding dimension could have multiple meanings that do not necessarily line up with ordinal dimensions. In the end, categories with similar representations translate a similar effect on the target variable.

Resulting preprocessed inputs

Pedological features embeddings resulting from the previous step are concatenated into a single tensor of dimension: $(batchSize, patchRadius, patchRadius, d_c)$ s.t $d_c = \sum_{c \in C} k_c$ (brown module in Figure 4.1). C refers to the set of categorical features. In our case, $d_c = 18$. Land cover is embedded separately (khaki box in Fig 4.1).

Climate, topological and hydrological features are input to the neural network as batches of $(batchSize, patchRadius, patchRad$ multi-channel images, $d_{num} = 22$ being the number of features.

Feature extraction

We investigated two modes of feature extraction on the preprocessed inputs:

- **JointGrinnellNet**: A joint mode where all input tensors are concatenated into a single multi-channel image. The result with dimension $(batchSize, patchRadius, patchRadius, \sum_{c \in C} k_c + n_{num}))$ is fed into a feature extraction module, as presented in Fig 4.1.
- **SepGrinnellNet**: A separate mode where the three groups of features are treated separately by dedicated feature extraction modules then merged afterwards, as shown in Fig 4.1.

Handling spatial data requires to use appropriate feature extraction techniques that are able to harness the spatial structure of such inputs. Convolutional neural networks[LeCun et al., 1995], a class of artificial neural networks inspired from the virtual cortex of animals, constitute an ideal choice as they allow to extract features from spatially-structured inputs within an end-to-end learning process. They have been previously shown to provide substantial improvements in predicting species abundance [Botella et al., 2018, Deneu et al., 2018].

Each of our CNN-based feature extraction modules comprises a two-block architecture similar to VGG [Simonyan and Zisserman, 2014]. Each block contains two 3×3 convolution layers set to extract 256 features, followed by MaxPooling then a Leaky Rectified Linear Unit activation. The latter function choice prevents the model from falling into a dying ReLu problem (experienced during first tests with ReLu activation) [Maas et al., 2013]. Retained embedding sizes, patch radius and resulting number of channels are shown in the detailed architecture are provided in the source code.

Feature interaction and classification components

Extracted features from the different components are flattened and eventually concatenated into a single large vector. This vector is then fed into a fully connected neural network dedicated to learning the separation of the plant classes in the learnt feature space. This feature interaction component (green box in Figure 4.1) comprises 3 dense layers of respectively 8192, 4096 and 3353 neurons. We applied a 0.75 dropout rate on the intermediary layers to prevent overfitting. The classification layer consists on a softmax activation applied on the output to determine the probabilities of each class. Class probabilities sum to one by definition. Naturally, the class with the highest probability is attributed to the instance.

4.3.2 EltonNet: a species embedding network

Here we propose to rely purely on associations between plants and other taxonomic groups. The goal is to predict the dominant plant from knowledge of the occurrences of other taxa around it, up to a certain radius. In order to reduce the number of co-occurring organisms, address the rarity of some of them and capture stronger associations with plants, the following processing steps were applied to the records of non-plant occurrences.

Taxonomic grouping and biogeographical filtering

We aggregated taxa according to ecological knowledge on the taxonomic level where biogeographical correlations to plants are meaningful. This level differs from one group to another, thus different preprocessing schemes were applied to different taxonomic groups. Then, for some groups we used domain-knowledge heuristics to filter irrelevant groups. Proportions of retained groups in the taxa list are illustrated in Figure 4.2. Finally, we assigned internal codes, i.e. unique identifiers, to each group.

Fungi selection . We grouped fungal species by genus. Then, we used the FunGUILD database [Nguyen et al., 2016] to select fungi from guilds (groups with similar diets and functions in the ecosystem) that are dependent on plants for feeding. We kept the following guilds: Pathotrophs (parasites of plants) and Symbiotrophs (involved in positive associations with plants such as mycchorizea). We deleted Saprotrophs (organic matter decomposers) as they do not depend on plants whatsoever. In the end, we retained 195 out of 531 genuses.

Insects selection . We aggregated insects to the order level except for Coleoptera and Orthoptera which were grouped in families as they exhibit significant intraorder variability in terms of habitat preferences and diet. We chose among insect orders those with known co-evolution history with plants (such as Hymenoptera) and/or established potential for direct interaction with plants (such as pollinators and herbivores) [Sauvion et al., 2013]. The intuition is that some insects have strong affinities or preferences (such as specialist pollinators) towards specific plants which leads to a greater chance of co-existence. This process led to the selection of 464 families of Coleoptera and Orthoptera in addition to 9 other orders.

Aves selection . Most birds breed in their preferential habitat during the period spanning from March to July. The rest of the year, during their migration phase, they travel through other areas where they can occasionally be observed. We considered these observations as spurious and removed them to avoid establishing false associations to plants. We then aggregated birds to the genus level. Afterwards, we used www.oiseaux.net to identify and remove some introduced/invasive genuses. We ended up with 240 bird taxa.

Amphibians, mammals and reptiles aggregation . Lacking expert knowledge on these phyla, we simply grouped them to the genus level, yielding 21 amphibians, 93 mammals, and 33 reptiles.

Biotic context calculation

To accelerate the training phase, we precomputed for each training example i, given its coordinates, the set V_i of non plant observations that occur within a radius of at most 8 Km. Starting with 500m, we iteratively doubled the radius until we identified a non-empty set of neighbor species, up until at most 8km. Afterwards, we randomly draw with repetition w observations from V_i with a uniform probability. That way, more abundant taxa (present multiple times in V_i) have a higher probability of being included. At the end of this process, we had associated each training example to its biotic context made of w observations of organisms from other kingdoms.

The species embedding network architecture

The learning model used is an adaptation of the Continuous Bag of Words model first introduced in Mikolov et al. [2013]. The architecture, illustrated in Figure 4.3, is based on a neural network composed of 3 layers:

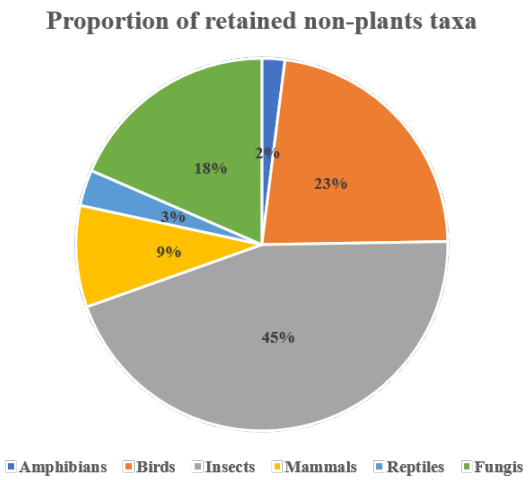


Figure 4.2: Non plants taxa proportion in the retained list.

- The input layer of size W receives the identifiers of the biotic context components.
- An embedding layer that associates a real-valued vector of size k_{np} for each taxa (non plants). This embedding vector captures the effect of observing this organism on the odds of each plant class.
- A lambda averaging layer that aggregates the biotic context embeddings.
- A dense layer that computes for each target plant species the dot products of its weight vector to the aggregated context embedding. This layer uses a softmax activation to return the probabilities of each target plant to occur given the observations of the surrounding non plants.

4.4 Training and evaluation

4.4.1 Optimization and evaluation metrics

Given:
- P: the set of plant classes (here species-level identifications).
- c: the expected or true class.
- s_c: the neural network output probability for the true class.
- w_c: weight of the true class.

In both proposals, we optimize the class-weighted sparse categorical cross-entropy loss given by equation Eq 4.1

$$CE = -w_c log(\frac{\exp^{s_c}}{\sum_j^P \exp^{s_j}}) \tag{4.1}$$

Some species were observed more often than others leading to a class-imbalance problem within the training set. To address this issue, we weighted each training example by the weight of its expected class (see Eq 4.2). This strategy allowed us to give more importance to the misclassification of rare classes observations (correcting for false negatives). Each class c with a frequency of occurrence p_c on the training set was attributed a weight computed as the ratio of its points of absence to its points of presence.

$$w_c = \frac{1 - p_c}{p_c} \tag{4.2}$$

This process is particularly useful for endemic species of undersampled locations.

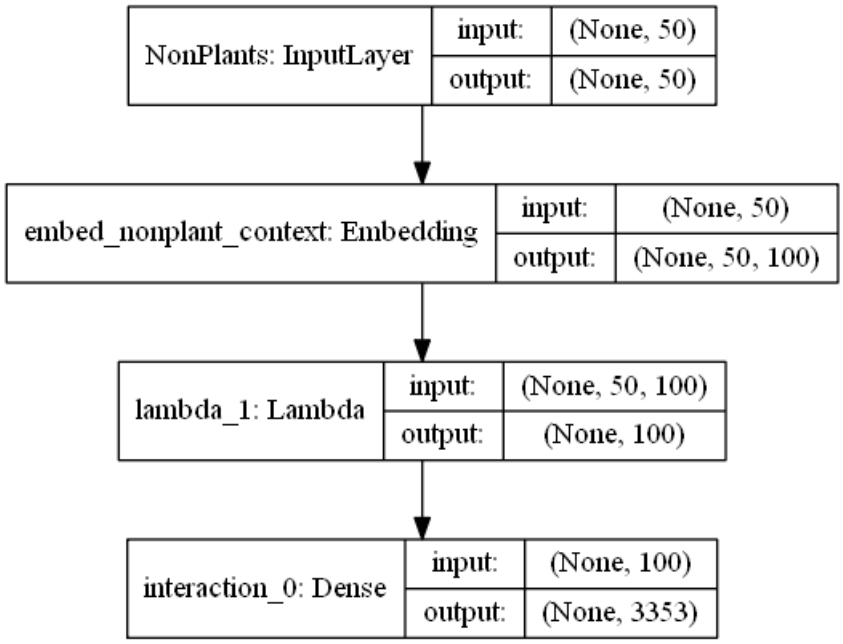


Figure 4.3: EltonNet architecture. Each box represents a layer described by its name and type as well as its input and output dimensions. None refers to the undefined batch size. The context size and embedding dimension shown here are respectively $w = 50$ and $k_{np} = 100$. The last layer (bottom level) gives the probabilities of each plant class, here 3353 classes.

4.4.2 Implementation and learning setting

We implemented GrinnellNet and EltonNet in Python[1] using Keras deep learning framework with Tensorflow[2] backend. We trained the models using multigpu data parallelism on a single computing node equiped with 4 GPUS V100 with NVlink[3]. We used the adam optimization algorithm [Kingma and Ba, 2014] with a decaying learning rate starting from 0.001 and reduced by a 10 factor whenever the validation loss stops improving after 5 epochs.

4.4.3 Evaluation

We sampled 80% of the dataset for training and kept the remaining 20% for validation. We used a stratified cross-validation split procedure to ensure coverage of all classes in the training set. At the end of every epoch, we evaluated the prediction accuracy of the models on the validation set. Figure 4.4 summarizes the performances obtained during training (left axis) along with results on test set (right axis). Note that different evaluation metrics are used. As a result, we can only compare the models ranking.

`ELT100` and `ELT300` correspond to EltonNet applied to occurrences of test species (evaluated in the test set) with the embedding size parameter k_{np} set respectively to 100 and 300. `GRIN_SEP` and `GRIN_SEP+` (trained longer) apply GrinnellNet on occurrences of all plant species whereas in `GRIN_SEP_TEST` the model is trained only on test species. `GRIN_SEP` uses separate feature extraction components for each feature group while `GRIN_JOINT` uses the joint feature extraction mode.

Unsurprisingly, GrinnellNet performs much better than EltonNet. Indeed, we would expect such results as covariates used in the former are richer and unbiased. Besides, biogeographical theory recognizes the

[1]Source code: https://github.com/SoccoCMOS/GeoLifeCLEF2019-GrinnEltonNet
[2]https://www.tensorflow.org/guide/keras
[3]Ciment cluster, UMS GRICAD, Grenoble Alpes University

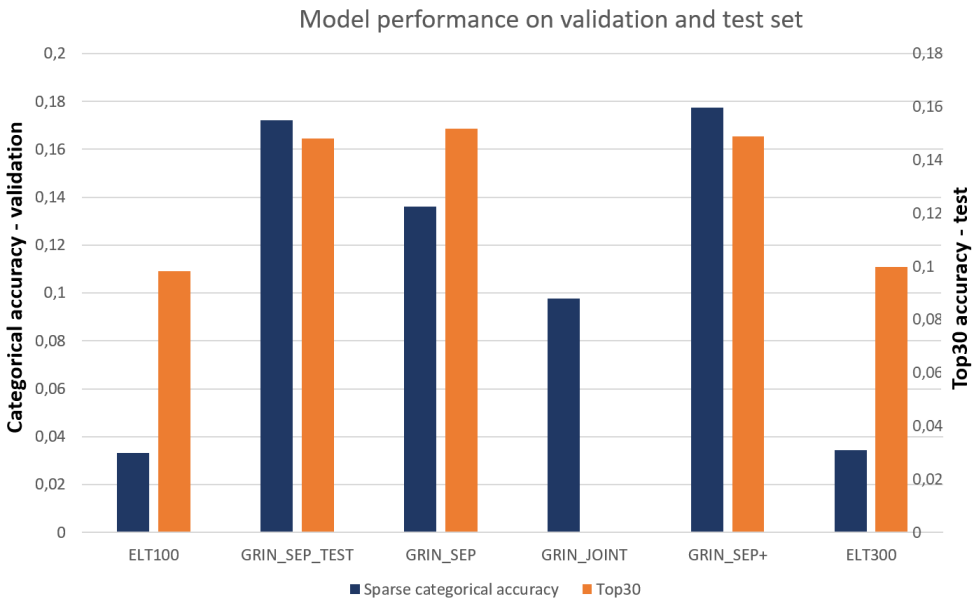


Figure 4.4: Performance of GrinnellNet and EltonNet variants on validation and test set.

superiority of the abiotic filter in selecting species [Boulangeat et al., 2012], as it is directly related to their physiological traits. On the other hand, EltonNet resulted from a series of arbitrary domain heuristics. Nevertheless, it still performs better than random with relatively strong associations learnt between plants and other taxa, a non-negligible insight for community ecologists.

In the case of GrinnellNet, the choice of the feature extraction mode clearly affects its predictive performances. Results show that treating the feature groups separately leads to better performances. This can be explained by the nature of the data encoded in the rasters that were created/interpolated from different data collection protocols. Indeed, pedological characteristics for instance are mainly determined by subjective field observations whereas climate data are calculated with advanced mathematical models. Another possible reason to separate the feature extraction processes is the scale at which the rasters were constructed. While bioclimatic variables are interpolated to the kilometer in regular grids, soil data are aggregated using anthropo-topological polygons to the landscape level, which translates in several kilometers. Consequently, we only submitted `GRIN_SEP` runs for the clef challenge.

We also observed that that the weighting strategy yielded significant performance improvements over the unweighted variant (not shown here). Furthermore, the runs ranking on the validation set is roughly the same as in the test set results except for `GRIN_SEP` and `GRIN_SEP_TEST`. During validation, we found that GrinnellNet performs better when it is trained solely on test species than when it uses occurrences of all plant species. At test time, the order was reversed which might be a sign of overfitting in `GRIN_SEP`. But also, because `GRIN_SEP_TEST` is exposed to more observations it probably learns more robust features.

4.5 Conclusion

We presented two proposals for the location-based species recommendation problem. The first solution leverages the concept of Grinnellian niche by building its predictions on only abiotic features, automatically extracted from environmental rasters using convolutional neural networks. This approach can be extended to any taxonomic group beyond plants. Moreover, we investigated the use of distributed representations as a means to reduce feature dimensionality as well as to capture rich semantic associations.

In the second proposal, we attempted to learn the Eltonian niche of the plants by embedding the biotic contexts where they are observed. We relied heavily on domain knowledge with expert assistance to filter co-occurrences in order to learn strong associations. Although the assumptions and rules used to select non plants species were specific to plant modeling, the learning architecture itself can be used for any taxonomic group. Additionally, this approach suffered from the heterogeneity of the sampling effort. Ideally, one could use projection maps predicted by species distribution models when available as input to a convolutional neural network.

Overall, our proposed CNN solution outperformed the species embedding approach. But the latter allowed us to identify associations between plants and other taxa which can be used to develop bioindicators. In the end, one could train both models jointly with shared layers that can capture the interactions and possible feedbacks between biotic and abiotic variables.

Chapter 5

Uncovering bidirectional ecological associations from community data

5.1 Introduction

Species co-abundances across space, recorded as *community data*, result from the interplay of abiotic filtering, biotic interactions and dispersal limitations [Weiher and Keddy, 2001, Peterson et al., 2011]. As a result of species *Grinnellian* and *Eltonian* niches, together with species dispersal abilities, species co-abundances vary in space. These data, measured as *community data*, are usually the corner-stone of analyses that aim to tease apart the relative importance of these processes [Weiher and Keddy, 2001, Peterson et al., 2011].

A natural way to address this objective is to jointly model multiple species distributions against environmental covariates, and then, analyse the pairwise co-dependencies between species knowing the environmental effects. These pairwise co-dependencies (i.e. associations) represent *the net effect* of one species on another, resulting from direct interactions or indirect interactions mediated by non modeled species or unmeasured environmental factors [Wootton, 1994].

Several statistical frameworks allow to infer these associations, either as their main objective or as a byproduct of the modeling process. These approaches differ in the type of dependencies they can model, in how they accommodate abundance data, and in the way they incorporate environmental covariates.

Joint Species Distribution Models (JSDM) address the problem by jointly predicting co-distributions of multiple species. Basically, once abiotic factors are accounted for, the residual correlation matrix captures species associations that are unexplained by the modeled covariates Pollock et al. [2014]. Early JSDMs used (Hierarchical) Generalized Mixed Effect Models to estimate full covariance matrices. Most recent implementations incorporate latent factors as a way to account for missing environmental variables as well as to reduce the parameter space size Ovaskainen et al. [2017]. Support for abundance data, particularly counts, was achieved through either data transformation techniques or appropriate link functions Clark et al. [2017], Niku et al. [2019], Ovaskainen and Abrego [2020].

Beside multivariate extensions of GLMs, the Poisson-Log-Normal distribution Aitchison and Ho [1989] has been used to model observed counts that arise from a Poisson distribution controlled by fixed environmental effects and latent variables describing species dependencies. The latter are modeled by means of a multivariate normal distribution whose inverse covariance (precision) encodes conditional dependencies or direct associations, which defines a Gaussian Graphical Model (GGM). Chiquet et al. [2018] proposed `PLNetwork` an efficient variational procedure to fit the PLN model while estimating a sparse GGM with the *graphical lasso* algorithm [Friedman et al., 2008]. On the other hand, Momal et al. [2019] developed `EMtree` to model the association network as a mixture of tree-shaped GGMs, additionaly providing edge probabilities.

To support different count distributions as well as mixed response types, Popovic et al. [2018] developed an algorithm that can handle discrete data (presence/absence and count) in Gaussian Copula Graphical Models,

implemented in the *EcoCopula* software. Their approach fits species-wise Generalized Linear Models as a function of the environment. Then, residuals are coupled together in a Gaussian Copula [Nelsen, 2007] to capture species associations. The underlying dependencies are encoded in a GGM. GGMs have also been used directly to model continuous multivariate species abundances as well as multi-trophic communities turnover [Ohlmann et al., 2018]. Gaussian Graphical Models are a special case of undirected graphical models or Markov Random Fields (MRF). Clark et al. [2018] used L1-penalized regression to learn a MRF from species occurrences. The method called *MRFcov* accomodates count data through a non-paranormal transformation prior to model fitting.

In addition to MRFs, Bayesian networks (BN) have been used to model directed associations involving plants [Aderhold et al., 2012], host-parasites [Pilosof et al., 2014] and predator-preys [Trifonova et al., 2015, Majdi et al., 2018].

JSDMs and MRFs are limited to estimating symmetric associations where the involved parties influence one another with the same polarity and strength whereas BNs impose an acyclic structure that does not support bidirectional influences. Lany et al. [2018] proposed a JSDM that allows to capture asymmetric associations but requires longitudinal data (see also Hollingdale et al. [2018], Barraquand et al. [2019]).

Alternatively, dependency networks (DN) [Heckerman et al., 2000] are a type of graphical model that tolerates cycles. DNs are conditionally-specified, they are built by independently performing a probabilistic regression or classification for each variable/node. The resulting node-wise conditional responses are then combined. Given enough data, DNs can learn consistent joint probabilities. As opposed to GGMs, they do not assume linear relationships and can be learnt using various methods ranging from simple linear regression to non-linear machine learning algorithms such as Random Forest or Neural Networks. However, given their representation power DNs require more parameters than BNs and MRFs.

5.1.1 Objectives and outline

Here, we aim to efficiently learn dependency networks encoding bidirectional associations from community data. We present a response-effect framework to infer both species habitat suitability and biotic associations.

Through two experiments on synthetic datasets and an empirical case study, we illustrate three possible modes of interplay of biotic associations with environmental filters. First, we simulate abundance data to compare the ability of our method as well as existing frameworks to recover the known associations. Second, we evaluate the ability of our model to recover simulated predator-prey associations under different food web structures. Finally, in an empirical case study, we apply the framework to a well-studied Alpine vegetation gradient [Choler, 2005, Warton et al., 2015], we use it to demonstrate how to analyze the structure of learnt biotic association networks.

5.2 The framework

Notation We consider a dataset consisting of the abundances of a collection $\mathcal{S}$ of m species observed at a collection $\mathcal{K}$ of n sites, as well as abiotic variables measured at these same sites or nearby. The abundance of species i at site s is denoted y_{si}, while the vector x_s represents the abiotic variables at site s.

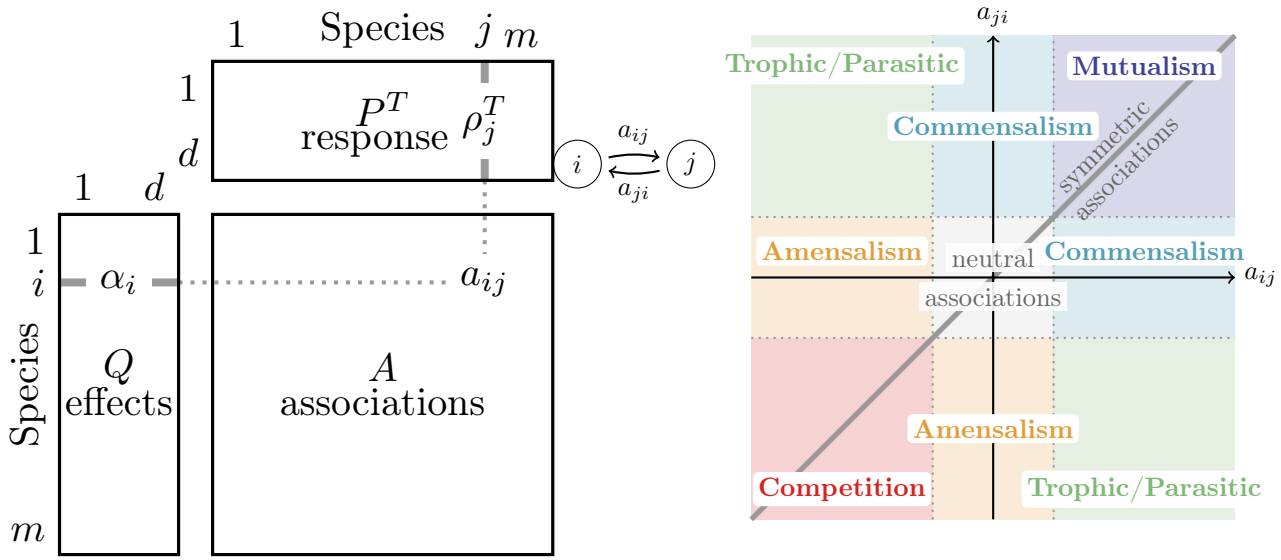
Three main conditions should be satisfied for a species to be present at a given site. First, the site must be accessible. This relates to the species' intrinsic *dispersal* capacity and the presence of migration opportunities or, instead, of barriers. Second, the abiotic conditions should allow the species' population to maintain a positive growth rate. Third, the impact of other species through direct or indirect biotic interactions should allow it to sustain its populations Guisan et al. [2017]. Although we recognize the importance of spatial dispersal processes, in this study we focus on the latter two factors, namely *habitat suitability* and *biotic associations*.

5.2.1 Spatial associations and species embeddings

Representing species responses and effects using embeddings

For a given pair of species, a *spatial association* describes the influence that one species has on the other species' abundance. The influences between two species along opposite directions can be of different polarity (positive, negative or neutral) and have different intensities (Fig 5.1b). Several mechanisms can lead to such association: a direct interaction between these two species (e.g. pollination, predation), an indirect interaction through the environment (e.g. resource competition) or a shared correlation to an unmeasured environmental variable or an unobserved group of organisms.

Here, we represent the influence of species j on species i and, vice-versa, of species i on species j as scalars a_{ij} and a_{ji}, respectively. More specifically, a_{ij} represents the change (excess if positive, deficit if negative) in *target* species i's abundance induced by the *source* species j. These values across all pairs of species, and in both directions, can be collected into an $m \times m$ asymmetric association matrix A.



(a) Association matrix factorization. (b) Pairwise associations classification

Figure 5.1: Association strengths are computed from species response and effects (a). Pairwise association strengths are mapped to potential interaction classes (b). The different quarters of the bi-plot represent the various types of associations between species. The first bissector represents the association domain covered by correlation-based approaches (JSDM) and undirected graphical models (MRF).

The association strength depends on two parameters: the *effect* applied by the source on the target species, and the *response* of the target species. We assume these parameters are controlled by intrinsic traits or properties of the species, which we encode as two separate d-dimensional real-valued vectors referred to as embeddings.

The *effect embedding* of species i, α_i, captures the type of organisms the species allows or deters when it is present. The *response embedding* of species i, ρ_i, measures the type of biotic context the species would strive in. For instance, trees with spreading canopy create shade (effect) that selects only shade-tolerant (response) species and exclude others. The response and effect embeddings of the different species can be collected into two $m \times d$ matrices, respectively denoted as P and Q.

The association matrix is then written as $A = PQ^T$ (cf. Fig 5.1a).

Biotic context

As described in chapter 3, the biotic context encodes our assumptions about the potential biotic effects a target species is exposed to at a given site. We obtain the aggregated effect of the biotic context by averaging the effect embeddings of its elements weighted by their respective abundances. Recall:

$$z_{si} = \frac{1}{|\mathcal{C}_{si}|} \sum_{j \in \mathcal{C}_{si}} y_{sj} \alpha_j \,.$$

This formulation allows the presence of opposing effects to balance one another. By weighting with abundance, we implicitly consider that all individuals from the same species are similar and contribute equally to the community structure. Rare species have a noticeable impact only if their per capita effect is stronger than the aggregated effect of dominant groups.

The biotic context implicitely imposes constraints on the structure of species association networks by restricting the set of potential associations a priori. For instance, the biotic context can be customized for each species according to known interactions. Moreover, it can include species from neighboring locations (**spatially-explicit**) up to a chosen radius within which their influence is relevant (e.g. species with low/high mobility). Similarly, we can construct a **temporally-explicit** biotic context from previous observations to account for time-lag. We provide the mathematical details for the formulation variants corresponding to these alternative definitions along with the associated data requirements and relevant effect aggregation functions in 3.2.1.

5.2.2 A conditional generative model of abundance

Inferring species habitat preferences and biotic associations consists in learning the structure and parameters of the graphical model depicted in Fig 5.2a. Here, we assume that the habitat-species dependency is strictly directed from the environment to the species, but that species associations can be bidirectional.

Each species incoming edges weights (responses) are learnt using multiple regression over all relevant influences including the abiotic covariates and the species in its biotic context Fig 5.2b. By including all factors in a single formula, we aim to uncover *conditional dependencies* and disentangle direct associations from indirect ones. To avoid variance inflation due to multicollinearity [Dormann et al., 2013], we add regularization penalties to select a few meaningful associations for each species.

As opposed to Morueta-Holme et al. [2016], Clark et al. [2018] who used a similar node-wise approach to learn a Gaussian Markov Random Field, we do not symmetrize the pairwise associations in order to detect bidirectional influences. Moreover, we factorize outgoing and incoming edges using species-specific effect and response embeddings respectively to discover structure in the association networks. Another advantage of this approach is the flexibility in terms of data types and functional forms at the level of species joint response to abiotic factors and biotic associations.

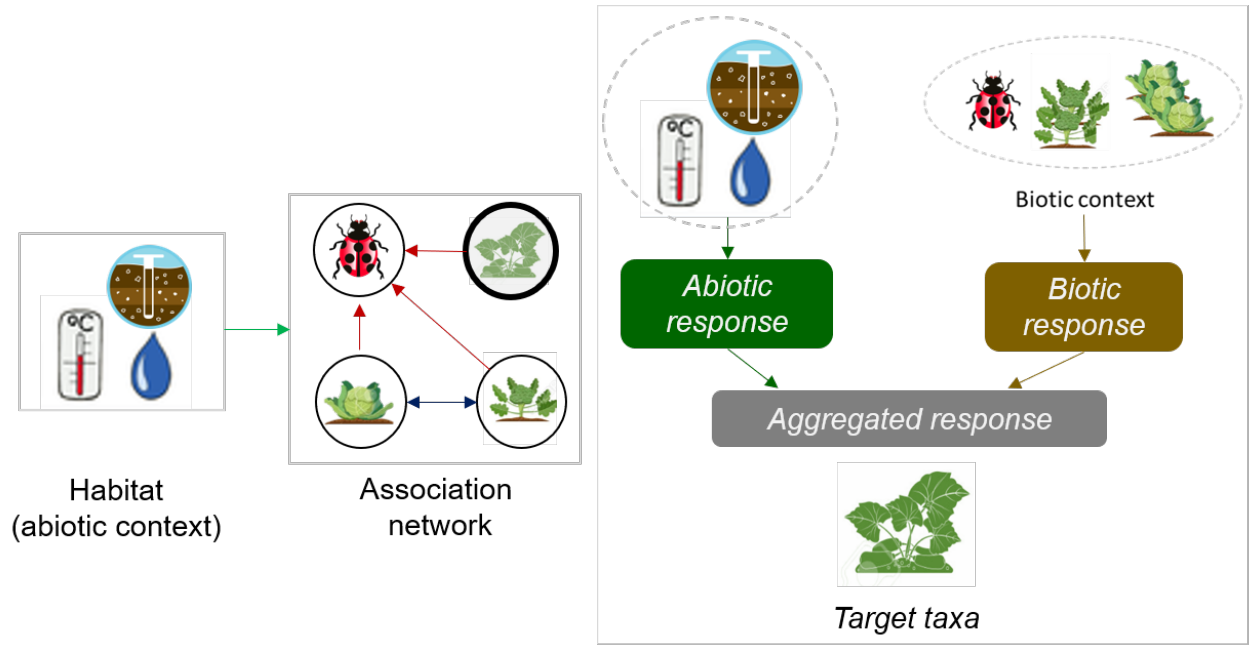
We write the target species abundance as an aggregation f_{agg} of abiotic η^A_{si} and biotic η^B_{si} responses at site s (Eq. (5.1a)).

The abiotic response is given by a habitat suitability model h_i (see Eq. (5.1b)). The biotic response η^B_{si} depends on the response embedding ρ_i of the target species and on the biotic context effect z_{si} resulting in a weighted sum of pairwise association strengths (see Eq. (5.1c)). An offset o_i is used to represent the outcome in the event of an empty biotic context.

$$g(y_{si}) = f_{agg}(\eta^A_{si}, \eta^B_{si}) \tag{5.1a}$$

$$\eta^A_{si} = h_i(x_s) \tag{5.1b}$$

$$\eta^B_{si} = o_i + \rho_i z_{si} = o_i + \sum_{j \in \mathcal{C}_{si}} (y_{sj} * a_{ij}) \tag{5.1c}$$



(a) Overall graphical model.

(b) Node-wise neighborhood inference.

Figure 5.2: A graphical illustration of the interplay between the habitat and biotic associations. (a) Species in a community form a network of associations of different signs: positive (red), negative (blue). Each edge represents an association from the source to the target species. When both species influence each other, the association is bidirectional. All species, along with their associations, respond to the habitat conditions (green). We ignore the reciprocal effect of species on habitat. (b)The abundance of a given target species results of its aggregated response to the abiotic (habitat) and biotic contexts (other species).

Different choices of habitat suitability model and aggregation function yield different modes of interplay between abiotic and biotic factors. We illustrate some examples with potential applications in Fig 5.3. Additionally, the type of data considered (presence/absence, abundance, biomass, etc.) might lead to different choices of probability distributions. As with GLMs, this only requires changing the link function g to accomodate the abundance distribution.

5.2.3 Inference and model selection

We minimize the negative loglikelihood of the observed abundances or occurrences, with lasso penalties on the associations to promote their sparsity. We use the Stochastic Gradient Descent [1] algorithm Bottou [2010] to learn the parameters (esp. response and effect embeddings matrices and HSM parameters). The proposed model includes a set of hyperparameters that need to be selected carefully: the hyper-parameters for the abiotic suitability model, embedding dimension, the vector of species offsets and the regularization coefficient. The best model can be selected using one of various information criterias Konishi and Kitagawa [2008] to penalize complex models: Akaike Information Criteria (AIC), Bayesian Information criteria (BIC) or its extended version (eBIC). Alternatively, the best model can be selected through cross-validation, based on predictive performances, considering for instance the AUC for presence/absence data or poisson deviance for counts.

To assess the robustness of the association estimates, the model can be fitted on multiple bootstrap samples of the original dataset. A confidence interval for the mean of each pairwise association can then be obtained. Due to the invariance of the associations to rotations of the embeddings, we advise against averaging embedding estimates obtained by fitting the model on different subsamples of the data as opposed to averaging associations. Moreover, when using bootstrap, lasso regularization should be avoided [Hastie et al., 2015].

5.2.4 Unraveling inter-specific association networks

In principle, regularization allows to reduce the number of associations for each species. When using bootstrap estimates, we perform an additional step which consists in setting to zero all associations with a confidence interval containing zero and keeping the mean value for the rest. In general, the user can choose a given threshold to obtain a discrete version of the association matrix, defined as follows:

$$I_{ij} = \begin{cases} \text{positive} & \text{if } a_{ij} > \epsilon^+, \\ \text{negative} & \text{if } a_{ij} < -\epsilon^-, \\ \text{neutral} & \text{otherwise.} \end{cases}$$

such that ϵ^+ and ϵ^- represent user-defined thresholds on the strength of the positive and negative associations, respectively. The resulting matrix can be seen as a network, where each species is represented by a vertex and a directed edge labelled as positive (resp. negative) from vertex i to vertex j represents a positive (resp. negative) influence of species i on species j.

By design (Fig. 5.1a), species with similar response embeddings constitute clusters of rows in the adjacency matrix, called *response groups*. Conversely, species with similar effect embeddings constitute clusters of columns called *effect groups*. These groups can be computed simultaneously using a *bi-clustering* algorithm [Govaert and Nadif, 2013]. The product of both types of groups results in the emergence of clusters of exchangeable or redundant species in the resulting network, called *structural roles* [Gauzens et al., 2015].

To validate the ability of the model to recover true associations under different scenarios of associations, we setup two virtual experiments that differ in the data process used to simulate communities.

[1] The framework is implemented atop the Tensorflow 2.0 and Keras stack, and uses automatic differentiation.

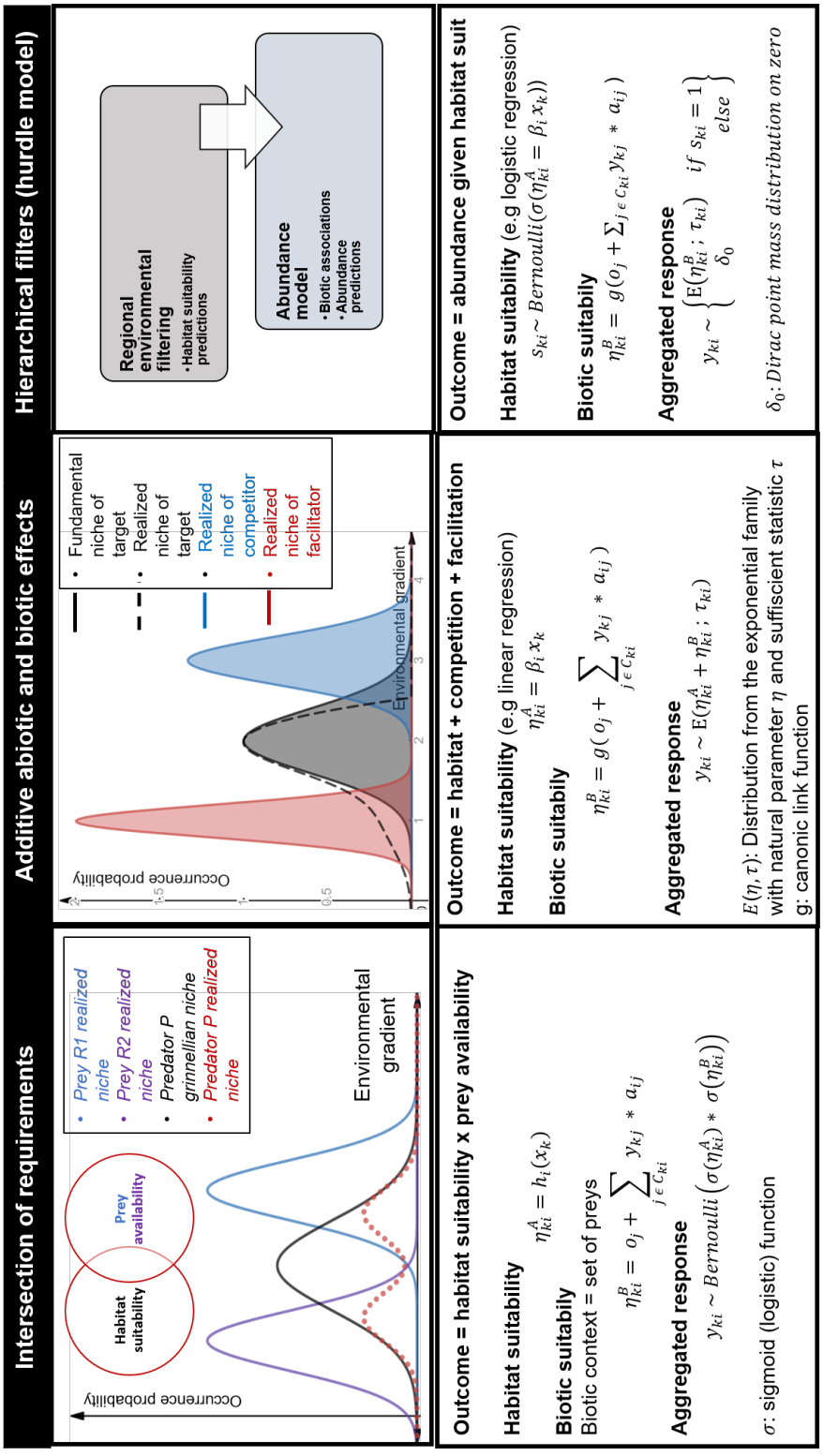


Figure 5.3: Abiotic and biotic response aggregation. Illustrated examples of potential applications.

5.3 Simulation experiment: process-based assembly with additive drivers

5.3.1 Methods

Data generation

In this experiment, we used a process-based stochastic model adapted from `Virtualcom` Münkemüller and Gallien [2015] to simulate the assembly of individuals from a regional species pool into communities, on different locations sampled along an environmental gradient. The assembly process is controlled by three filtering mechanisms: the response to the abiotic environment, the outcome of biotic interactions, and reproduction.

The simulation starts with a given or random initial composition for each community independently. In each iteration, individuals are replaced through time until an equilibrium state is reached or a user-defined number of iterations is completed. In the end, the final composition of the communities is returned as the result of the simulation.

We set up an experiment where multiple simulations were run on random points on a single abiotic gradient ranging from 0 to 100 with different randomly selected configurations of the prior association matrix: absence of association, positive associations only, negative associations only and a mix of positive and negative associations. In each configuration mode, we varied the pool size i-e the number of species (10, 20 or 50), the *density* in terms of *number of associated pairs* as a function of the pool size (sparse 1/3 or dense 2/3) and whether the association matrix included asymmetric effects: semi-attraction (as in commensalism) or semi-repulsion (as in amensalism). Positive (resp. negative) effects were all set to $+1$ (resp. -1) as we are interested in the polarity of the associations rather than their intensity. The factorial design of this experiment produced 33 simulation datasets (Fig 5.4).

Simulation diagnosis

To assess whether the virtual communities reflect the simulated species dependencies, we defined the relative abundance index (RAI), an asymmetric pairwise that measures the change in abundance of the target species when the source species is present as compared to its mean abundance irrespective of whether the source species is present.

$$\bar{y}_t = \text{avg}\left(\{\, y_{st},\ \text{for all } s \in \mathcal{K} \text{ such that } y_{st} > 0 \,\}\right), \quad \text{and}$$
$$\Delta_{st} = \{\, y_{st} - \bar{y}_t,\ \text{for all } s \in \mathcal{K} \text{ such that } y_{st} > 0 \text{ and } y_{ss} > 0 \,\}.$$

Then $\text{RAI}_{st} = \text{avg}(\Delta_{st})$. The larger the standard deviation $\text{std}(\Delta_{st})$, the more ambiguous the strength of the effect of species s on species t. If the confidence interval $\text{avg}(\Delta_{st}) \pm 1.96\,\text{std}(\Delta_{st})$ does not contain zero, then the simulated dependencies unambiguously translate a polarized effect of species s on species t. Otherwise, the polarity of the effect is ambiguous, due to either confounding effects of other species or a neutral association if the mean is close to zero. We also compute the Jaccard coefficient between the binary presence/absence vectors of species s and t, a.k.a. Jaccard co-occurrence index, denoted as J_{st}.

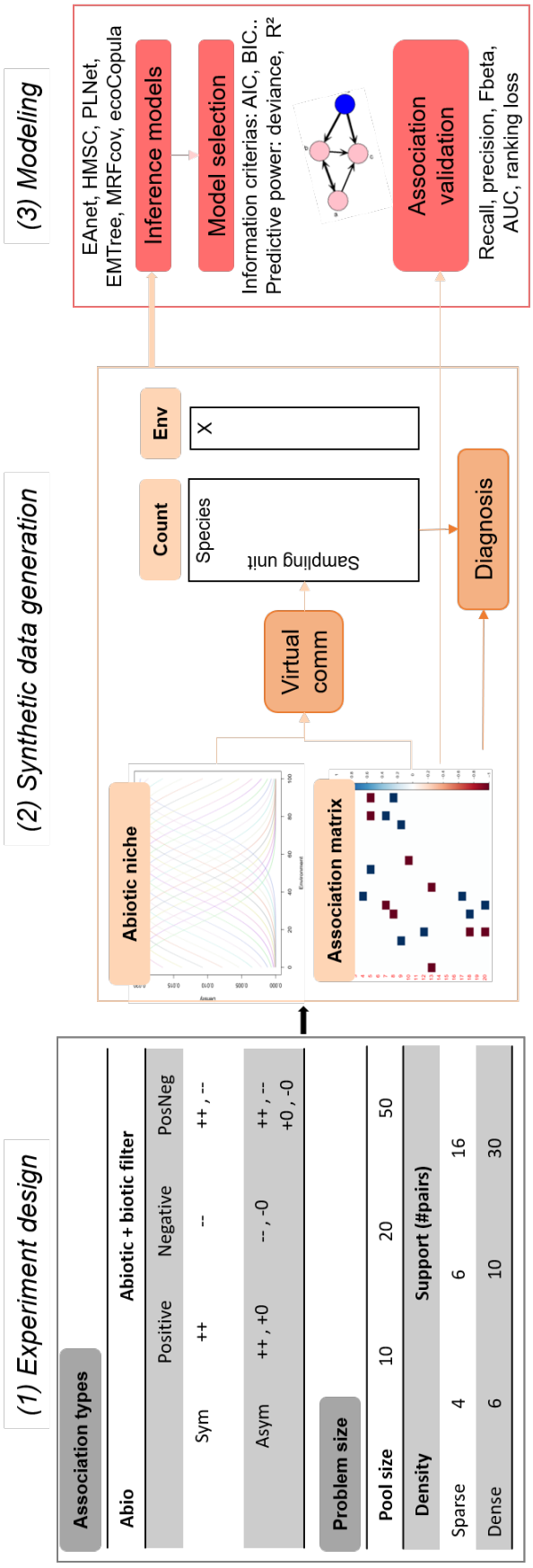


Figure 5.4: Schematic representation of the virtual experiment.

Framework	Count distribution	Association structure	Graph selection procedure	Learning configuration
Ecological Association Network (EA)	-Negative binomial -Unknown dispersion -Link: log	Dependency network	Cross-validation	Optimizer: adam Maximum number of epochs: 200 Batch size: 16 Early stopping (convergence by monitoring validation loss): - Patience: 5 epochs - Tolerance: 1E-3
EcoCopula [Popovic et al., 2018, 2019]	-Negative binomial -Unknown dispersion -Link: log	Copula Gaussian Graphical Model	Graphical lasso	Importance sampling: 1000
EMtree [Momal et al., 2019]	Poisson log-normal	Mixture of tree-shaped Gaussian Graphical Models	Edge probabilty Support: 2/pool_size	Covariance mode: full Number of iterations: 50 Convergence tolerance: 1E-8 Resampling: 5
Hierarchical Modeling of Species Communities (HMSC) [Ovaskainen et al., 2017]	Poisson	Residual correlation (Covariance graph)	MAP residual covariance Support level 95%	MCMC: Hamiltonian Monte-Carlo thinning = 10 nChains = 2 Burn-in = 500 nSamples = 5000-7500 Cross-validation: 2 folds
MRFcov [Clark et al., 2018]	Gaussian with non-paranormal transformation of count data	Conditional Markov Random Field	Bootstrap (95% CI) Sample proportion: 70% Symetrization function: mean	Bootstrap samples: 500
Poisson Log-Normal network (PLNetwork) [Chiquet et al., 2018]	Poisson log-normal	Gaussian Graphical Model	Graphical lasso + stability selection	Covariance mode: full Offset: None

Table 5.1: Description of the evaluated frameworks and their respective configuration.

Comparative evaluation

We fit the proposed model on species counts using a negative binomial distribution, with exponential link, and an additive aggregation function to combine abiotic and biotic drivers. We selected the best hyperparameters using 10 fold cross-validation and the poisson deviance as a performance metric.

For comparison, we also fitted 5 other well-established or emergent methods commonly used in community ecology to infer associations from count data. In Table 5.1, we indicate, for each method, the probabilistic model it relies on, along with the data requirements, the training/inference configuration used to fit the model, as well as any additional post-processing.

For all methods, we discretized the associations learnt by each method to obtain the corresponding association classes (positive, negative, neutral), which we compared against the ground-truth of the simulated association classes using standard multi-class performance metrics (recall, precision, F1-score). The recall measures the percentage of associations of a specific class correctly recovered by the model, whereas the precision quantifies the percentage of true associations amongst those classified as the specified class. The F1-score is computed as the harmonic mean of recall and precision. A higher recall indicates lower false negatives whilst higher precision indicates lower false positives.

We compare the performances between inference methods and across simulation scenarios.

5.3.2 Results

Amongst the six statistical frameworks tested including our proposal, some were straightforward to fit, with little control over the model selection procedure including *EMtree, EcoCopula, MRFcov* except the number of iterations and the size and number of samples for the estimation of confidence intervals. In the case of *PLNetwork* and *EA*, the frameworks also required the specification of a network selection criteria amongst either *information criteria* (AIC, BIC, eBIC) or *predictive performance* (R^2, Deviance) on out-of-fold observations of species abundance conditioned on the abiotic environment. For the latter two methods, both classes of metrics selected different network structures. The execution time varied greatly between all methods and was inflated by the model selection procedures used (bootstrapping, lasso path execution, cross-validation). The Bayesian posterior inference step in HMSC made this method significantly slower to fit than others.

Simulation diagnosis

Before fitting the inference models on the simulated data, we checked using an empirical measure of pairwise association RAI_{ij} whether species dependencies reflected the simulated patterns. Fig 5.5 depicts the value of the statistic for each directed association $a_i j$. The distribution of RAI_{ij} values showed a good discrimination of positive and negative associations, albeit with different strengths. Neutral associations translated into small RAI_{ij} values, median-centered on zero for simulations with negative associations only. Whereas they spanned a large spectrum of values in simulations with only positive or a mix of positive and negative effects. The proposed indicator is itself a good proxy for association inference, however since it does not account for environmental covariates we use it as a diagnosis tool.

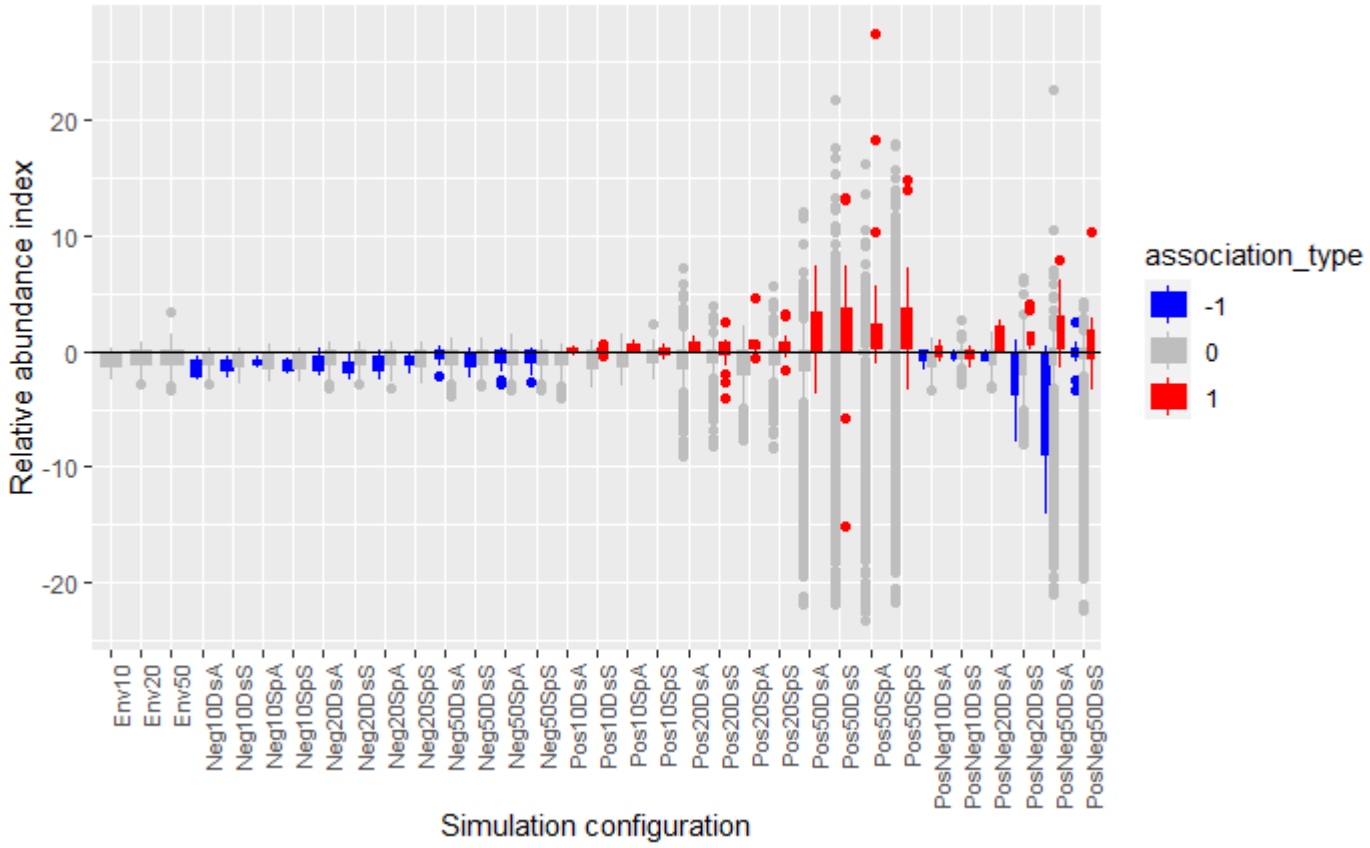


Figure 5.5: **Simulation diagnosis**. Distribution of relative abundance indices RAI_{ij} per simulation. Each data point represents a directed association (positive in red, negative in blue and neutral in gray) involving two species from the corresponding simulation. Labels on the x-axis correspond to simulation configurations.

Association strengths distribution

For each inference method, we represented the distribution of association intensity inferred, grouped by their simulated polarity (see Fig 5.6). Recall that we simulated fixed intensities. All methods except HMSC produced sparse association networks (most association values close to zero), with low strengths values and were good at discriminating positive and negative associations, while maintaining neutral associations median-centered at zero . Therefore, there exists an interval delimiting neutral associations with extremes specific to each inference method and dataset. Most spurious associations, i.e. neutral pairs with inferred value significantly different from zero, were negative especially in simulations involving only positive associations reflecting the implicit constraint induced by the fixed carrying capacity on the total species count. On the other hand, HMSC produced very dense association matrices despite a large support level for association

selection suggesting that some of the inferred associations are indirect associations. However, it was closer to the true strength of positive associations than other approaches. There was no difference in inferred strengths neither between symmetric and asymmetric simulations (for all methods).

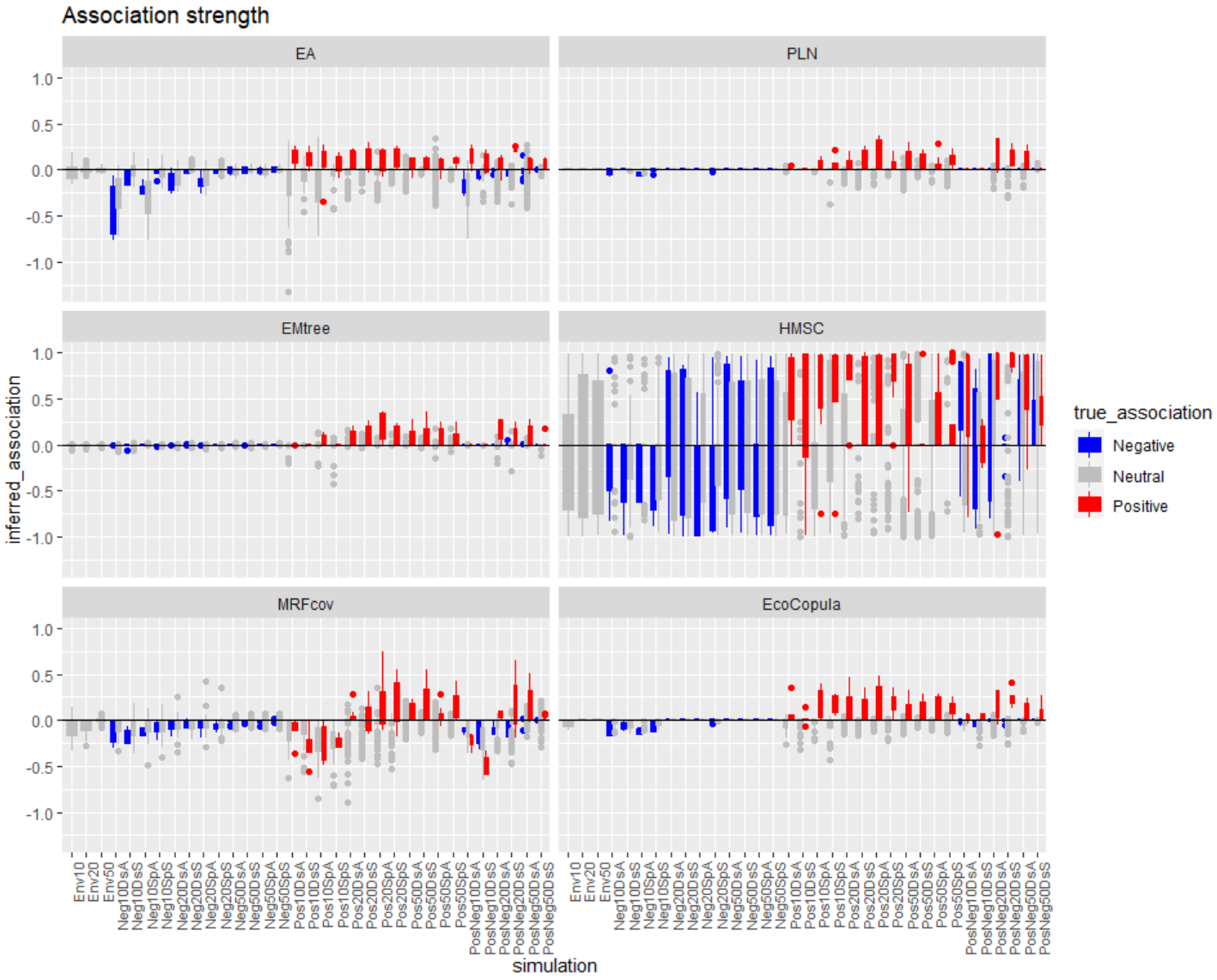


Figure 5.6: Distribution of fitted association strengths per simulated class (positive, negative, neutral) for each simulation configuration. A data point represents a directed association from a species to another, its color encodes the true type of the association, its coordinates on the y axis represents the fitted association strength by the corresponding model for the simulation configuration x-axis.

On larger pool sizes, the distance between niche optima decreases as the probability of niche overlap increases. Consequently, the estimates of species associations varied with the number of modeled species. Although the strengths appeared to be drawn from a small fixed interval, their value was sensitive to the species niche differences(Fig 5.7). For all tested frameworks, the strength inferred for true positive associations decreased with the amount of niche overlap. For large niche differences (no overlap), MRFcov, EMtree and PLN even reported opposite signs. Conversely, inferred strengths of true negative associations were either invariant to niche difference (MRFcov), very small and close to neutral (EMtree, PLN, ecocopula and EA) or increasing in absolute strength (HMSC). At medium to high niche distance, EA reported an increase in the absolute

strength of negative associations. HMSC showed the opposite pattern, suggesting that the negative effects were rather explained by the abiotic covariates.

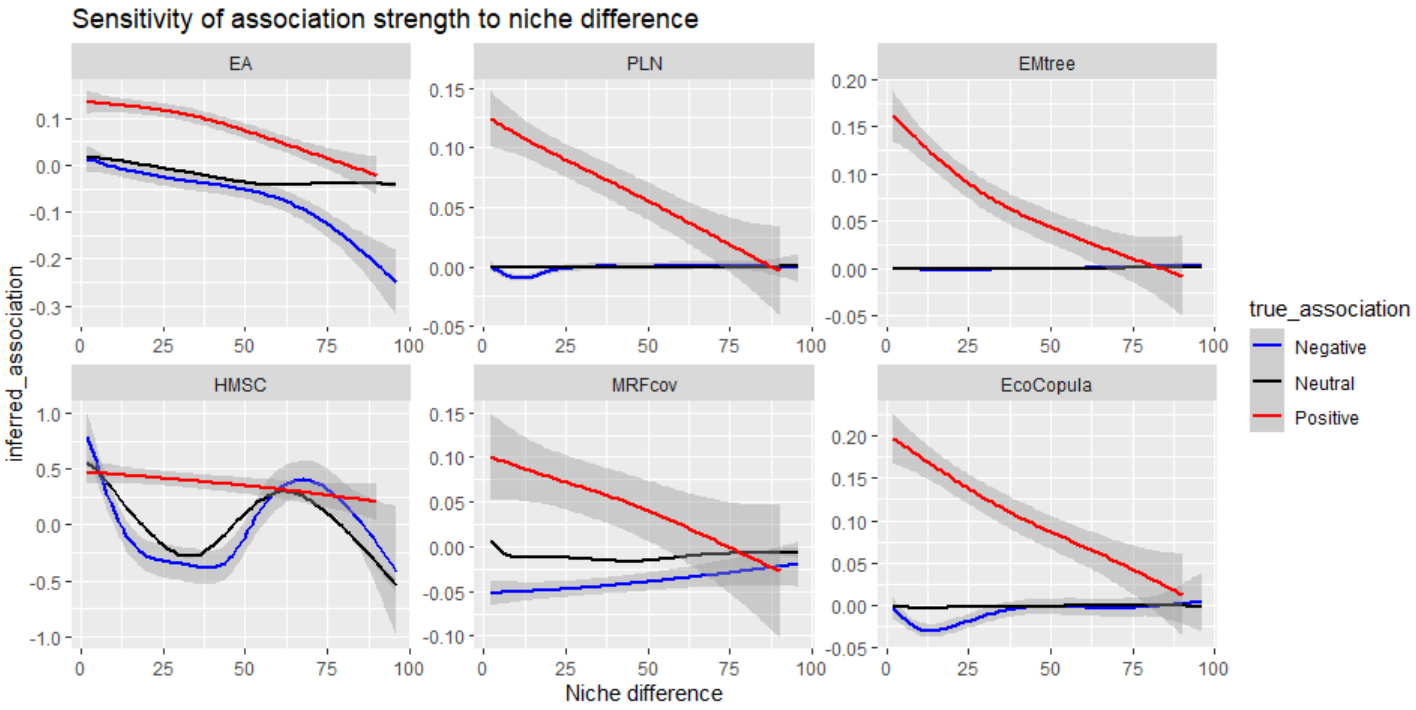


Figure 5.7: Sensitivity of the inferred association strength a_{ij} per association type and inference model to the abiotic niche distance measured by the absolute difference between their niche optima $\mu_i - \mu_j$.

Comparative performances on association type inference

We reported the Area Under the Precision-Recall Curve (PR-AUC), the recall and f1-score in Fig 5.8 for each association type separately. We found no significant difference between performances on positive vs dense or symmetric/asymetric datasets. However, the quality of inferred associations varied with the pool size.

On positive associations, EA and Ecocopula outperformed the other methods in all pool sizes. EMtree, PLN and MRFcov reported good performances for 20 and 50 species datasets, but they failed to detect positive associations in 10-species datasets. On negative associations, EA, MRFcov and HMSC outperformed other methods. But generally, the detection of negative effects decreased with the pool size.

5.4 Simulation experiment: predator-prey co-occurrences

5.4.1 Methods

Data generation

We generate 6 food webs involving the same number of trophic groups $G = 5$ using different trophic niche models, using the `trophic` R package. To each trophic group (Eltonian niche), we assign $m_G = 5$ species with different abiotic (Grinnellian) niche optima sampled uniformly along an environmental gradient ranging from 0 to 100. We select 500 sites uniformly in the same gradient.

Fig 5.9 illustrates the experimental setup and the procedure used to generate occurrences. Briefly, we assume a bottom-up control so that to be present, a consumer requires the availability of *at least* one resource, in

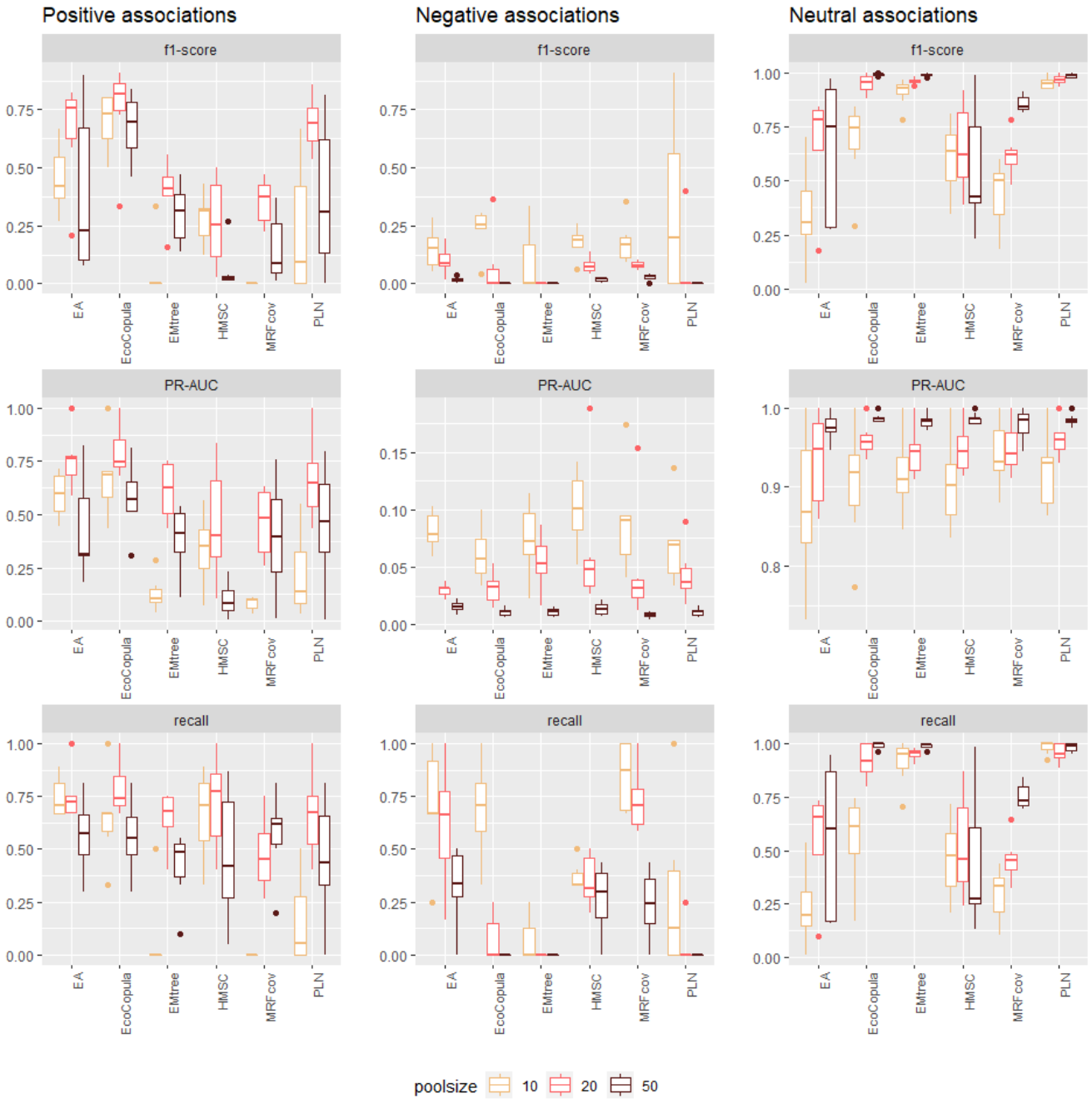


Figure 5.8: Inference of true association class per type of association for each model measured by the recall, f1-score and AUC-PR metrics. The higher the value the better performing is the model. AUC-PR is computed on the raw associations while the recall and f1-score are computed on the discretized associations using as a threshold $\epsilon^+ = \epsilon^- = 5E-2$

addition to habitat suitability. Basal trophic groups do not depend on any resource, only on the environment. The process resulted in 6 datasets, each containing 25 species and 500 sites.

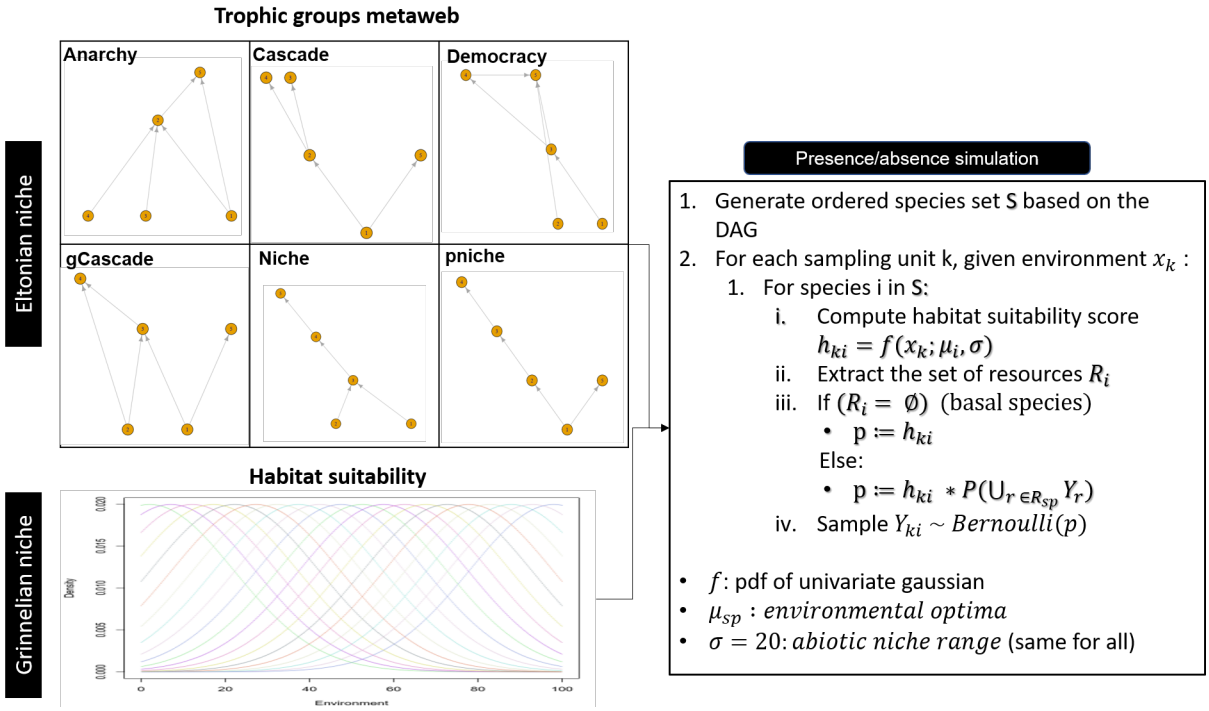


Figure 5.9: Simulation of consumer-resource co-occurrences along an environmental gradient given food web topology and true abiotic niches.

Evaluation

In this experiment, we only evaluate our proposed model. We combines abiotic and biotic responses using a product of probabilities (illustrated in Fig 5.3 as *intersection of requirements*) to fit the model on presence-absence data, combined to a linear logistic regression with a quadratic term to fit the Gaussian abiotic niche.

The simulation model assumes an unrequited positive dependency of the predators on their preys. Thus, we imposed a non-negative constraint to the embedding vectors to prevent the inference of negative associations and promote sparsity of the association matrix Hoyer [2004]. Consequently, we only inferred two types of associations: positive and neutral. We used 10 fold cross-validation to select the combination of embedding dimension and lasso regularization that maximized the accuracy of predicted occurrences.

We first evaluated the quality (accuracy, AUC, sensitivity and specificity) of recovered consumer-resource associations using as ground truth:

- The potential foodweb (metaweb): all potential interactions.
- The realized foodweb: obtained from the potential foodweb by filtering edges involving non co-occurring species.

Afterwards, we evaluated whether sharing embeddings between species of the same trophic group improved the ability of the model to retrieve true potential and realized associations.

Structural regularities conserved in the embeddings

For each topology, we investigated whether learnt embeddings reflect the underlying clustering of species into trophic groups. Concretely, we performed a Mann-Whitney ranking test Mann and Whitney [1947]

checking whether species from the same prior trophic group had more similar embeddings than species from different groups. Finally, we asked to what extent response and effect embeddings captured species abiotic preferences and their biotic requirements. We did that by testing the correlation between embedding similarity and niche overlap measured with environmental optima differences and the proportion of shared preys in both the potential and realized food webs.

5.4.2 Results

Network inference performances

The inferred associations were more faithful to the realized than the potential network (Fig 5.10). All metrics varied amongst topologies. Specificity was notably lower for the niche topology. In all cases, incorporating a parameter-sharing constraint within trophic groups allowed to improve the sensitivity with respect to both ground truth networks.

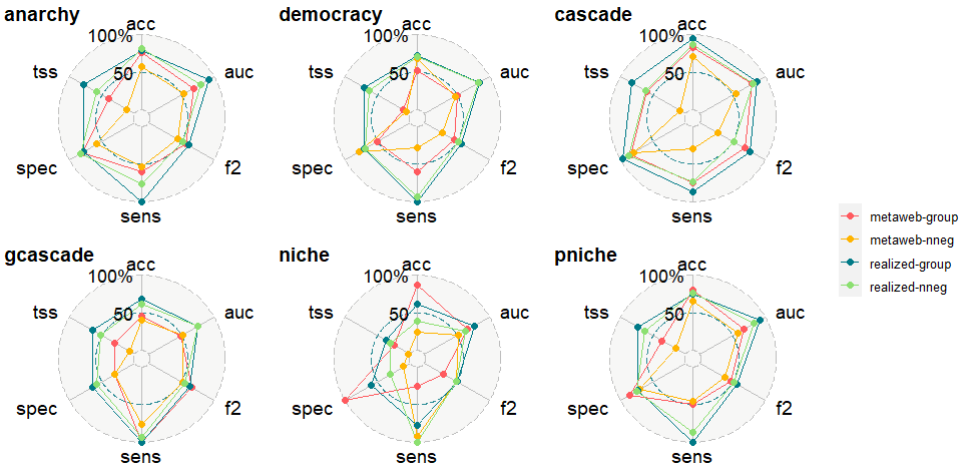


Figure 5.10: Network structure inference quality with respect to the potential (metaweb) and the realized networks under two different constraints: non-negative associations and within-group embedding sharing. Performances are reported separately for each food web topology.

The visualization of inferred networks highlighted that the main source of error was due to the confusion of directed associations with symmetric reciprocal associations, introducing spurious associations. This error was particularly frequent between species of inner groups in trophic chains as in the niche topology. However, the model managed to break this symmetry for species from groups with a higher diversity (number of distinct groups) of preys.

Structural regularities captured by the embeddings

For all topologies, the ranking test showed that both response and effect representations were significantly more similar when species belonged to the same trophic group. Moreover, species embedding similarity correlated positively with both abiotic and biotic niche similarity, with varying significance levels across topologies. Response and effect representations similarity were strongly correlated for two reasons: (1) species with similar responses were likely to have similar effects and vice-versa, (2) both representations were very similar, which explains the symmetry in some of the inferred associations.

5.5 Empirical case study

In our last experimental study, we applied our model to a dataset of Alpine plant abundances (previously analyzed by Warton et al. [2015]) and show how it can uncover meaningful associations from observational data.

	Additive abiotic and biotic responses				Intersection of abiotic and biotic requirements			
	Group-level		Species-level		Group-level		Species-level	
	MW	**RN**	**MW**	**RN**	**MW**	**RN**	**MW**	**RN**
Accuracy (ACC)								
min	0.45	0.80	0.72	0.82	0.45	0.62	0.24	0.38
max	0.83	0.94	0.83	0.94	0.88	0.96	0.72	0.87
median	**0.790**	0.890	0.765	**0.895**	0.780	0.735	0.615	0.735
mean (sd)	0.72 ± 0.15	0.87 ± 0.06	0.77 ± 0.04	0.88 ± 0.05	0.71 ± 0.18	0.76 ± 0.12	0.55 ± 0.19	0.69 ± 0.18
ROC-AUC (AUC)								
min	0.24	0.07	0.49	0.50	0.50	0.75	0.47	0.62
max	0.80	0.85	0.53	0.74	0.80	0.92	0.58	0.83
median	0.470	0.415	0.505	0.565	**0.670**	**0.855**	0.535	0.800
mean (sd)	0.52 ± 0.22	0.43 ± 0.27	0.51 ± 0.01	0.58 ± 0.09	0.64 ± 0.12	0.84 ± 0.07	0.53 ± 0.04	0.77 ± 0.08
F2-score (F2)								
min	0.00	0.00	0.00	0.00	0.29	0.50	0.27	0.49
max	0.56	0.65	0.19	0.44	0.68	0.77	0.52	0.55
median	0.120	0.220	0.105	0.220	**0.520**	**0.590**	0.415	0.520
mean (sd)	0.19 ± 0.23	0.25 ± 0.25	0.10 ± 0.08	0.19 ± 0.16	0.52 ± 0.14	0.61 ± 0.09	0.40 ± 0.11	0.52 ± 0.02
True Skill Statistic (TSS)								
min	-0.06	0.00	-0.01	0.00	0.11	0.37	0.02	0.30
max	0.36	0.77	0.06	0.46	0.60	0.83	0.16	0.63
median	0.000	0.150	0.005	0.120	**0.335**	**0.715**	0.085	0.600
mean (sd)	0.08 ± 0.16	0.24 ± 0.30	0.02 ± 0.03	0.15 ± 0.17	0.34 ± 0.16	0.68 ± 0.16	0.09 ± 0.05	0.54 ± 0.13

Table 5.2: Summary of trophic network structure inference performances using two architectures: Additive abiotic and biotic effects, intersection of requirements (reported in the main text), with associations at the species level or the group level (by sharing embeddings within prior trophic groups). Metrics are computed taking in turn the potential **MW** (metaweb) and the realized **RN** network as ground truth. We highlight the best median score for each metric and reference network.

	Anarchy	Democracy	Cascade	gCascade	Niche	pNiche
Internal clustering validation						
response	38278(<**0.001**)	36408(0.0021)	40024(<**0.001**)	35983(0.0044)	37188(<**0.001**)	38100(<**0.001**)
effect	36683(0.0013)	37300(<**0.001**)	39164(<**0.001**)	36708(0.0013)	36448(0.002)	38188(<**0.001**)
Abiotic niche similarity						
response	0.21(<**0.001**)	0.13(0.0014)	0.13(<**0.001**)	0.16(<**0.001**)	0.23(<**0.001**)	0.16(<**0.001**)
effect	0.21(<**0.001**)	0.12(0.0033)	0.18(<**0.001**)	0.24(<**0.001**)	0.07(0.065)	0.15(<**0.001**)
Eltonian niche similarity						
response	0.04(0.29)	0.1(0.015)	0.15(<**0.001**)	0.09(0.03)	0.12(0.0039)	0.13(0.0011)
effect	0.04(0.31)	0.11(0.0052)	0.2(<**0.001**)	0.08(0.06)	0.08(0.041)	0.15(<**0.001**)
Filtered prey similarity						
response	0.14(<**0.001**)	-0.03(0.39)	0.34(<**0.001**)	-0.11(0.0052)	0.13(0.0013)	0.03(0.48)
effect	0.13(<**0.001**)	0.02(0.59)	0.31(<**0.001**)	-0.09(0.033)	0.05(0.25)	0.08(0.035)

Table 5.3: Structural regularities captured by the response and effect embeddings, for each food web topology.

5.5.1 Data preparation

The dataset consists of records of abundance for 82 distinct plant species collected around July 2000 on 75 vegetation plots of size 5×5m each Choler [2005]. In addition, a set of environmental and topographic variables was recorded on each plot:

slope : the slope inclination in degrees,

snow : the average snowmelt date in Julian days between 1997 and 1999,

physd : the percentage of non vegetated soil due to physical processes,

zoogd : the percentage of non vegetated soil due to marmot activity,

aspect : the relative south aspect, and

form : the microtopographic landform index.

We applied a one-hot encoding scheme to the two categorical features (aspect and form) and scaled the numerical features. We split the observations into a training and a test dataset using a multi-label stratification scheme[2] to ensure that all species were covered and their proportions were preserved in both sets.

5.5.2 Statistical analysis

Framework adaptation

We assumed that habitat suitability drove the species occurrences whereas local within-community biotic associations determined their abundance, and could in turn induce local absence. Thus, we used the hierarchical filtering mode (Fig 5.3). We defined the biotic context for a target species as the set of plants observed on the location of interest, and we used a negative binomial distribution to fit the plant counts.

Training

For each plant species, we pre-trained a generalized linear model (GLM) with a logit link to relate species occurrences to the environmental variables. We used the learnt weights as initial parameter values in the habitat suitability component of our framework. On the other hand, the embedding vectors were initialized using random samples from a uniform distribution on the $[-0.01, 0.01]$ interval, and subjected to lasso penalties. Finally, the offset value for each species was set to its average count on occurrence points.

We trained the full model using stochastic gradient descent (with a learning rate of 0.01, a momentum of 0.8 and a batch size of 1) on the training dataset. We monitored the negative log-likelihood of positive examples (presences) on the validation set after each full pass of the training set to assess the convergence of the training. We stopped when the validation loss started increasing or when 200 epochs have elapsed.

The first step in this evaluation was to find appropriate values for the hyperparameters of our model. For a species pool of size m, the embedding dimension d is selected among powers of 2 up to $m/2$, to improve hyperparameter search speed. In our case, with $m = 82$, the embedding dimension is chosen from the set $\{2, 4, 8, 16, 32\}$. Note that this is a maximum dimension, that could shrink to a smaller dimension with the lasso regularization. For each value of d, we apply the training procedure described previously with increasing values of $\lambda \in \{0.01, 0.015, 0.02, 0.025\}$. We evaluate the resulting models on the test set by computing the deviance of the predicted counts on the test set.

Network analysis

We performed a hierarchical bi-clustering on the inferred association matrix, to obtain effect and response groups. In parallel, we applied the modularity maximization algorithm Newman [2006] on the association network to identify densely connected modules, referred to as communities (*sensu* graph theory) Gauzens et al. [2015]. After that, we mapped the structural roles within the modules to create the group-level network. Finally, we analyzed the resulting patterns in light of existing literature on Alpine plants interactions Choler et al. [2001].

[2]Python library `scikit-multilearn`: `http://scikit.ml/`

Relationship between functional traits and embeddings

We investigated the functional determinants of the associations diversity. To do so, we computed the mutual information between the learnt embeddings and the following plant traits (reported in Choler [2005]):

Height : Vegetation height in centimeters,

Spread : Maximum lateral spread of clonal plants in centimeters,

Angle : Leaf elevation angle, estimated at the laminal level,

Area : Area of a single leaf,

Thick : Maximum thickness of a leaf,

$\mathbf{N}_{mass}$: Mass-based leaf nitrogen content,

SLA : Specific Leaf Area,

Seed : Seed mass,

The Mutual Information Shannon and Weaver [1949] is an unbounded symmetric and positive score that measures the amount of information contained in one random variable about another. It quantifies the reduction in uncertainty about one random variable given knowledge of another. Zero mutual information indicates independence.

In general, we expected traits related to dispersal capabilities (seed mass, spread) to impact the prevalence of the species, consequently increasing or decreasing the opportunity to affect other species (interaction probability). As a result, we expected such traits to have a higher mutual information with effect embeddings than with response embeddings. Conversely, traits related to nutrient uptake and biomass accumulation potential capture competitive or cooperative abilities of the plant species. Hence, we would expect a high mutual information between these traits and both responses and effects embeddings.

5.5.3 Results

Plant habitat suitability

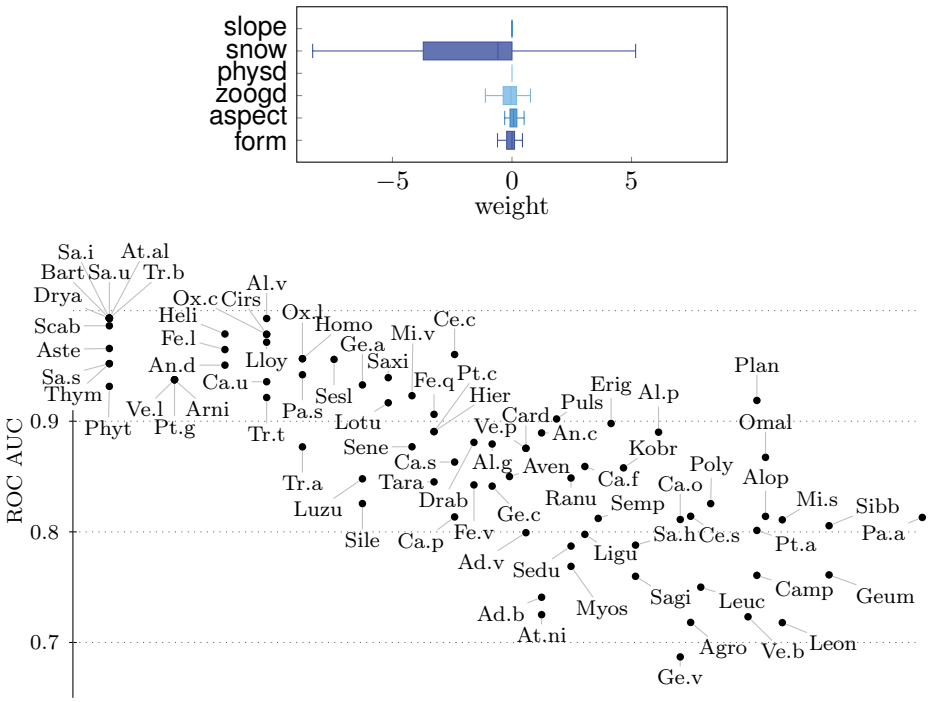


Figure 5.11: Habitat Suitability Model variable importance and prediction performances per genus.

The model predicts habitat suitability with a $87.7 \pm 0.17\%$ AUC score for all genera (Fig. 5.11). The analysis of environmental variable importance showed the dominance of snow duration followed by zoogenic disturbances, the site form and aspect. Physical disturbance and slope weights were negligible, probably due to their correlation with snow.

Plant associations along the mesotopographic gradient

The application of the proposed approach to the Alpine dataset identified four densely connected modules of different sizes, within which species occupied distinct structural roles in the plant association network. Modules were structured along the snow melting date gradient (Fig. 5.12a).

Species from early-melting sites were classified into the same module. We found a prominence of positive associations, specifically an unselective mostly asymmetric attraction of forbs and grasses to tall dominant graminoids (*Carex, Kobresia*). Forbs and grasses formed two distinct groups linked by negative associations. Besides, some of them acted as hubs connecting the high elevation sites to the adjacent sites where they also occurred. The second module encompassed two groups of grasses: (i) Tall herbs occuring in favorables conditions, mostly structured by negative associations (ammensalism and competition); (ii) Short herb meadows, exposed to zoogenic disturbances. They presented higher abundances when co-occuring with tall herbs. The third module consisted of chinopholous (cold-resistant) vegetation appearing on late-melting sites. The last module included north-facing isolated communities dominated by *Salix herbacea* positively associated with high-altitude communities, but disconnected from the other modules (Fig. 5.12).

In general, positive associations were prominent under stressful conditions. For instance, on early melting sites, species are exposed to wind and erosion due to snow melting Choler [2005]. The positive associations could be explained by the facilitative effect of graminoids through multiple hypothetical mechanisms. Graminoids have the ability to maintain the soil stability Callaway [2007], Heilbronn and Walton [1984], they can also prevent dessication and frost heaving on stones in favor of seedling survival Choler et al. [2001]. They also sustain a suitable microclimate for small forbs and grasses, while offering protection from the wind Wardle et al. [1998]. On the other hand, negative associations were found on the richest sites, hypothetically reflecting a competition for resources: water and Nitrogen Choler et al. [2001].

As reported in the literature, the abiotic conditions strongly structured the predicted response Choler [2005] of the plant species and the dominant interaction types Callaway et al. [2002]. Specifically, network modules were distributed along the gradient following their composition's response to the average snow duration. Negative associations inflicted by competitive tall grasses on mid slope communities connected early-melting communities to the chinopholous vegetation from late-melting sites in the resulting association network (Fig. 5.12b). Response and effect groups were included in one or at most two close (in terms of position in the gradient) modules (Fig. 5.12a). Non neutral associations had consistent types (either negative or positive) within effect groups regardless of the responding species, suggesting that species roles (effects) in their community might be predictable from their own characteristics and the surrounding abiotic conditions. At last, learnt associations were symmetric within groups but asymmetric (mostly semi-attraction or semi-repulsion) between them.

Analyzing the functional meaning of plant embeddings

There was a relatively significant contribution of the leaf nitrogen mass and spread to the plants response, whereas leaf angle was found to be independent (Fig. 5.13). The Specific Leaf Area contributed significantly to the effect, in addition to the Nitrogen mass and on a lesser extent Spread. Height wad related to both parameters.

5.6 Discussion

In this chapter, we tackled the challenge of inferring interspecific associations from multiple species co-abundances in hetereogeneous environments. To do so, we formalized pairwise associations as a function of two sets of latent variables representing the response and the effect of each species with respect to the others.

We incorporated these associations into a conditional probabilistic model of abundance that accounts for environmental covariates.

5.6.1 Disentangling abiotic and biotic drivers

Uncovering positive, negative and neutral associations

Comparatively to other tested frameworks, our method (EA) performed well on both positive and negative associations detection, despite the constraint induced by the embedding-based factorization. On average, discriminating positive and negative associations was within reach of most methods, provided an appropriate pair of thresholds was used to delimit the range of neutral associations. An exception concerned HMSC, which recovered multiple spurious associations that were potentially indirect effects despite the high support level. A more appropriate approach would have been to analyze the inverse covariance matrix. However, inverting the posterior estimate of the covariance matrix suffered from various numerical instabilities.

Due to the upper-limit constraint of the fixed carrying capacity on the total count, all models inferred spurious negative associations between non-interacting species, especially in simulation with positive effects only, as a compensation mechanism. Association strengths were sensitive to niche overlap. For all methods, positive associations were easier to detect between species with overlapping niches. The fact that this pattern was observed for all methods as well as on the pairwise relative abundance indices suggested that the abiotic filter outruled these associations during the community assembly simulation.

Amongst the tested methods, HMSC, PLN/EMtree and EcoCopula first fit the abiotic response then species dependencies are estimated either as random effects (HMSC, PLN) or from the marginal residuals (EcoCopula). The implicit importance given by these inference procedures to the abiotic drivers over the species associations explains the low detection rate of negative associations. In constrast, EA and MRFcov which both rely on an explicit regression over species abundances, do not suffer from the same bias, explaining their superiority in detecting repulsive effects.

Uncovering prey-predator associations

Several studies discuss the difficulties of recovering biotic interactions from co-occurrences Sander et al. [2017], Barner et al. [2018], Blanchet et al. [2020]. In our experiment, we assumed that trophic interactions induce a dependence of predators on their preys but not vice-versa (directed positive association).

We showed that using an appropriate coupling with the abiotic drivers allows to detect such associations providing that the species pair co-occur. However, the model detected symmetric dependencies when the abiotic niches of the pair overlapped strongly and especially when the predator did not have alternative preys, as in trophic chains. Moreover, the varying performances in recovery the true network structure amongst food web topologies questions the power of the response-effect factorization to represent arbitrary DAG structures and suggests that a symmetric approach might be more effective, if coupled with knowledge of species trophic levels.

Importance of the abiotic-biotic aggregation function

Species joint responses to abiotic (environment) and biotic (associations) drivers take on different forms, modeled by an aggregation function. Most existing frameworks are limited to linear or additive forms. Linear responses are particularly useful when associations are mediated by the environment (e.g in competition) or can alter it (as in habitat facilitation). In this case, associations compensate the suitability of the environment by either improving microhabitat conditions or exerting a negative force that counterbalances it.

On the other hand, when associations arise from direct interferences, their detection requires conditioning on co-occurrence, hence on habitat suitability Gravel et al. [2019]. When we fitted an additive architecture to the predator-prey occurrences, the model had a very low detection rate confirming that linear combinations of habitat suitability and biotic effects are not sensitive to such direct associations. These results may be specific to presence/absences and not hold true for abundances.

In general, the choice of an aggregation function depends on the type of interactions expected in the studied

system. To guide this choice, several frameworks [Kissling et al., 2012, Boulangeat et al., 2012, Thuiller et al., 2013] conceptualize the incorporation of eco-evolutionary processes into species distribution models (a.k.a *BI-SDMs* [Dormann et al., 2018]). Besides, theoretical developments extended the theory of island biogeography [MacArthur and Wilson, 2001] to account for trophic interactions [Gravel et al., 2011] and more general interaction networks under environmental constraints [Cazelles et al., 2016a,b].

5.6.2 From species representations to biotic associations

The meaning of species embeddings

In theory, the effect embedding of a species A is equivalent to a factor analysis of all other species abundances (residual abiotic responses if coupled with environmental data) when A is present. The effect embedding is a proxy of the species' influence on the community composition. Combining the effect embeddings of occurring species produces an ordination of the community composition in the embedding space of dimension d: R^d. The species response embedding can be mapped into the same space, we can measure through the dot product the compatibility of the species to the observed community.

Since the community ordination is obtained as a linear combination of present species' effects, species response to the community can be rewritten as a sum of one-to-one responses to each observed species. When the response and effect embeddings are forced to be similar, we recover the same structure used by Latent Variable JSDMs.

Analogously to the species embeddings, Kissling et al. [2012] proposed the concept of *interaction currencies* as surrogates for biotic interactions in distribution models in a similar response-effect framework. Hypothetically, these currencies include resources, bionomic variables [Hutchinson, 1957], traits, and other non-consumable environmental conditions that mediate interactions. Our analysis of embeddings learnt from data in food web simulations showed that they captured both abiotic and biotic species preferences. In the case study on Alpine plants, we found that embeddings were mildly related to functional traits.

Constraining embeddings with prior knowledge

In practice, the embedding dimension is typically significantly smaller than the number of species. While species can have distinct habitat preferences, the biotic role expressed in their interactions and the spatial associations they produce is drawn from a limited number (significantly smaller than the pool size) of behaviors represented by functional groups [Walker, 1992]. A species can belong to one or several functional groups with different proportions. Such information can be mined from online databases or provided by experts [BETSI, 2012, Nguyen et al., 2016, Kattge et al., 2020]. While learning graphical models with large species pools requires large datasets, replacing species with fixed groups Ohlmann et al. [2018] has two advantages: (1) to reduce the parameter space size by sharing embeddings within groups, (2) allowing extrapolation to new settings where different taxa are observed yet from the same modeled groups. Besides, as evidenced by our simulated experiment, using group constraints can improve the ability of inference models to recover potential associations even when species did not co-occur.

5.6.3 Perspectives

Beyond group constraints, some frameworks [Lo and Marculescu, 2017, Chiquet et al., 2018, Scutari et al., 2019] support white-lists and black-lists, containing authorized and forbidden associations respectively, by penalizing graphs that do not satisfy those constraints. When interaction networks can be described at least partially, the same approach can be used to complete missing edges by harnessing similarities of species interactions. This semi-supervised problem is referred to as *collaborative fitlering* [Fu et al., 2019] and is one of the main applications of dependency networks. Incorporating this link prediction task within a multispecies distribution model would allow to quantify the effect of known and predicted interactions on species distributions.

We motivated throughout our simulation experiments the use of different joint responses for abiotic and biotic drivers depending on the underlying biotic interactions. The fact that interactions require and affect co-occurrences simultaneously are not mutually exclusive [Gravel et al., 2019]. The availability of multi-trophic

communities datasets [Derocles et al., 2018] where complex interactions are entangled calls for applications coupling different modes of aggregating abiotic drivers with biotic associations.

5.7 Conclusion

Biological interactions and other processes induce spatial patterns of co-occurrence and co-abundance. We presented and validated a model of species co-abundances as a function of the habitat and biotic associations. We proposed an asymmetric scheme for modeling associations that is based on learning latent representations of species' responses and effects. Future efforts should be directed towards an incorporation of prior knowledge of the complete or partial topology of the association networks to guide the inference process. Along with that, a strong theory of how known ecological interactions influence the co-distribution of species is needed to support all these models.

5.8 Supplementary Materials

5.8.1 Virtualcomm how-to

We used a process-based stochastic model adapted from Virtualcomm (Gallien and Münkemüller 2015) to simulate the assembly of individuals from a regional species pool into communities, on different locations sampled along an environmental gradient. The assembly process is controlled by three filtering mechanisms: the response to the abiotic environment, the outcome of biotic interactions and reproduction. For simplicity, the spatial structure of communities and thus dispersal processes are ignored. In other words, there is no exchange of individuals between neighboring communities. The simulation starts with a given or random initial composition for each community independently. Individuals are replaced through time until an equilibrium state is reached or a user-defined number of iterations is completed. The final communities' composition is returned at the end Fig. 5.14.

Notation

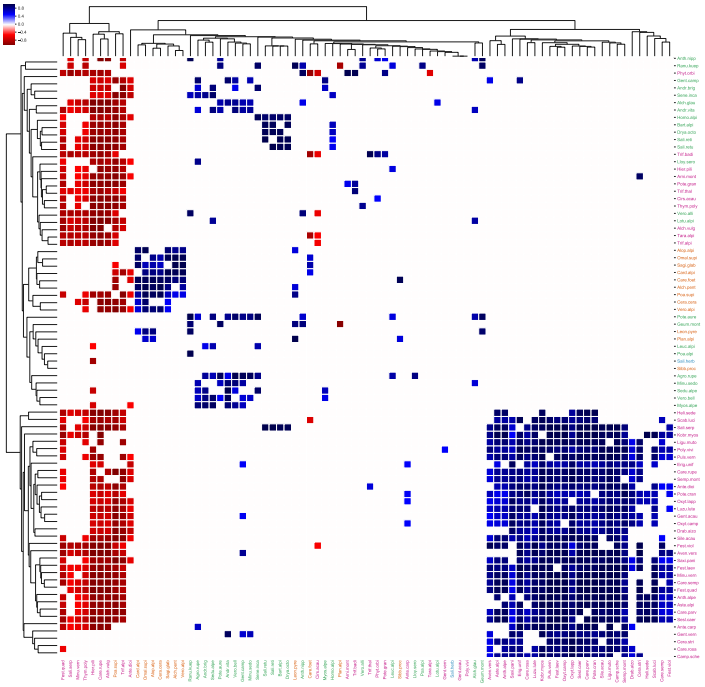
- We start by sampling n locations uniformly on a single environmental gradient E.
- All locations have the same carrying capacity of K individuals from a common pool of m species $S = \{S_j / j \in [1, m]\}$.
- Each species has its own optimal environmental value $\mu_j \in E$ as well as a niche breadth $\delta_j \in E$.
- Biotic interactions are described by a full interaction matrix $I = (I_{jk})/j, k \in [1, m]^2$; $-1 \leq I_{jk} \leq 1$ where I_{jk} represents the effect of the interaction between the pair (S_j, S_k) on species S_k. We also write: $I = I^+ - I^-$ such that:
  - $I^+ = (I^+_{jk})/j, k \in [1, m]^2; 0 \leq I^+_{jk} \leq 1$ represents the matrix of positive effects (facilitation matrix)
  - $I^- = (I^+_{jk})/j, k \in [1, m]^2; -1 \leq I^+_{jk} \leq 0$ represents the matrix of negative effects (competition matrix)

Assembly rules

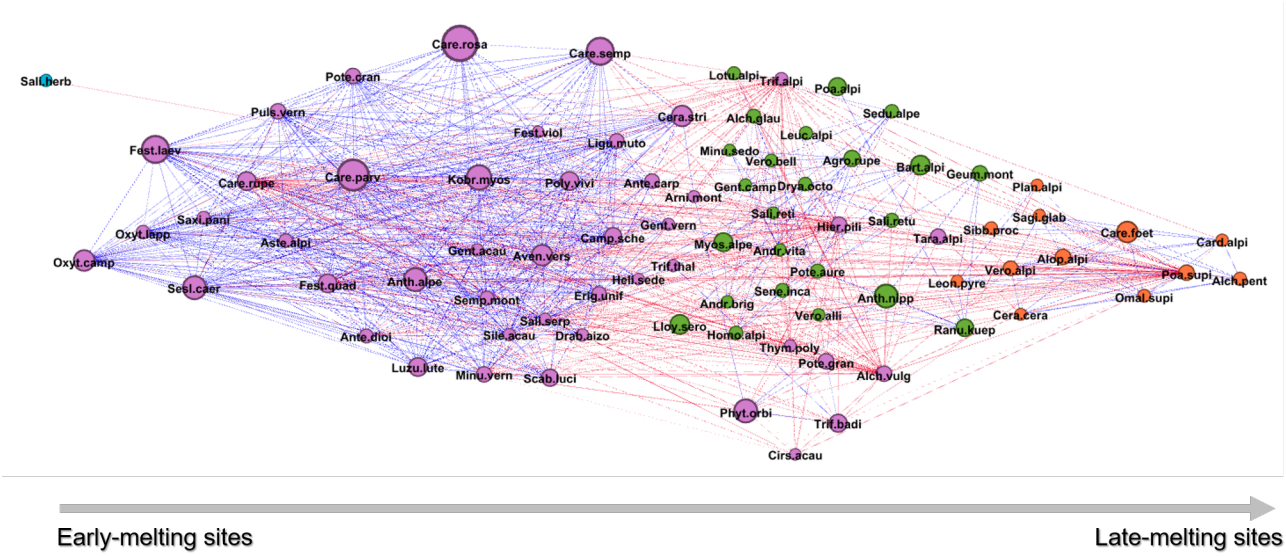
At each timestep (epoch), given an actual composition c, the probability that an individual from a given species i to replace any other individual of c is given by the following equation ; such that:

- B_{env}: weights of the abiotic filter.
- B_{comp}: weight of the competition.
- B_{fac}: weight of the facilitation.
- B_{abun}: weight of the reproduction filter, can be interpreted in terms of growth rate.
- $P_{env,i,c}$: the probability of species i to occur under the environmental value E_c is given by the normalized density on E_c of a Gaussian distribution parameterized by its optimum and niche breadth. The closer to its optima, the higher the probability of the species' occurrence.
- $P_{comp,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its competitors in c.
- $P_{fac,i,c}$: the probability for an individual of species i to join the community given the aggregated effect of its facilitators in c.
- $P_{abund,i,c}$: probability of an individual of species i to join the community as a result of the reproduction of some of the $N_{i,c}$ conspecifics in c.

The unnormalized weights $W_{i,c}$ for each species are then normalized by dividing each one of them by their sum. The result is a vector of probabilities W that sums to 1. Finally, we sample from a multinomial distribution, parameterized with W, K individuals to compose the new community.



(a) Inferred plant association matrix. Species in the association matrix are grouped based on a hierarchical bi-clustering performed row-wise (yielding response groups) and column-wise (yielding effect groups).



(b) Network of plant associations. Blue (resp. red) edges indicate negative (resp. positive) edge weights. Node colors on the graph represent communities identified by the modularity maximization algorithm Newman [2006] whilst node sizes are scaled according to the plant height. Nodes (except *Salix herbacea*, which represents the vegetation on the northern face of the gradient) are placed from left to right following an ascending order of their response to Snow duration (regression coefficient from the Generalized Linear Model used as a Habitat Suitability Model).

Figure 5.12: Plant associations on an Alpine mesotopographic gradient. We highlight the communities (node colors) in figure (b) using colored labels on the matrix (a).

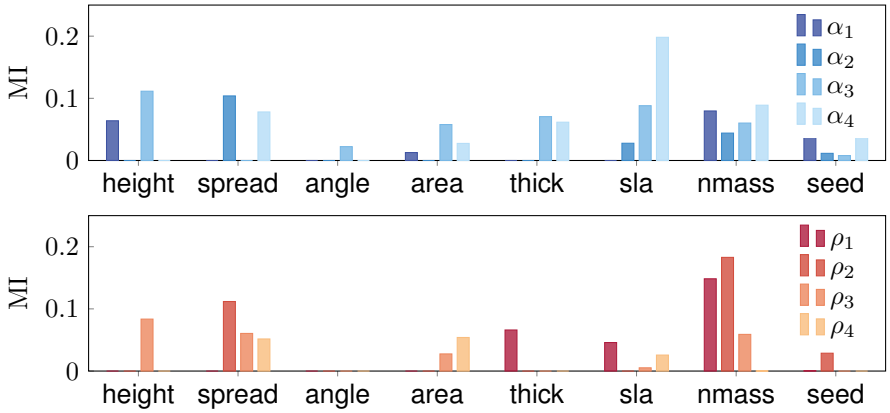


Figure 5.13: Mutual information between plant traits and their latent representations. Each bar concerns a specific trait, it represents the mutual information scores from the first to the last (fourth) embedding dimension. The lower (resp. upper) figure shows the results for the response (resp. effect) embeddings.

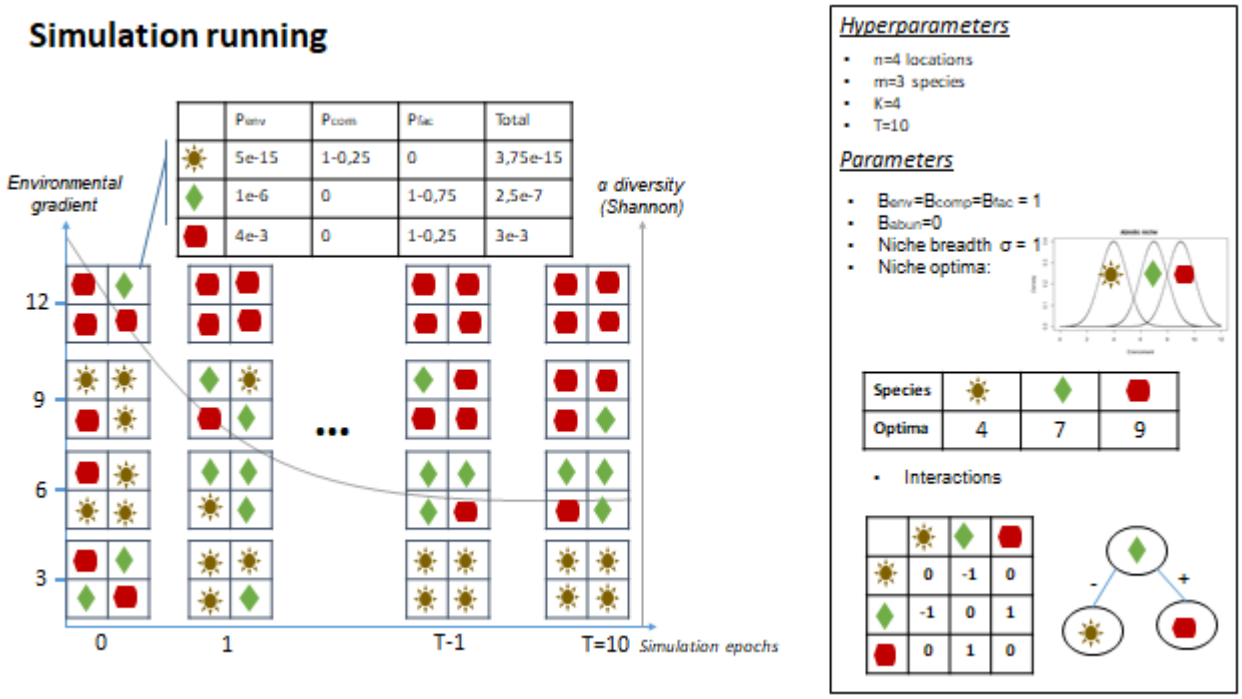


Figure 5.14: Virtualcomm simulation procedure.

Chapter 6

Biogeography and diversity of earthworms in mainland France

> *It may be doubted whether there are many other animals which have played so important a part in the history of the world, as have these lowly organized creatures.*
>
> Darwin [1892]

6.1 Introduction

Earthworms are terrestrial invertebrates from the clitellate annelids subclass [Cameron et al., 2016]. Distributed worldwide Phillips et al. [2019] except in drylands, deserts and polar lands, they account for a major proportion of the soil biomass [Amat et al., 2008]. They are divided into 23 families, more than 700 genera, and more than 7,000 species. However, a high amount of cryptic species [Marchán et al., 2020b] is supposed and many species are not yet described, even in well-studied countries [Marchán et al., 2020a].

Earthworms contribute to many physical and chemical processes crucial to soil-based ecosystem services [Lavelle and Spain, 2001]. As detritivores, they accelerate **nutrient cycling** through physical grinding and chemical digestion of plant debris on the surface (leaf litter) and in the soil (root litter), and the mixing of organic matter with minerals from the soil, which enhances **nutrient availability** and **soil fertility** Jiménez and Filser [2020]. As the engine of soil **bioturbation** Meysman et al. [2006], earthworms enable organic matter, water and air to circulate within the soil, through networks of burrows [Capowiez et al., 2003] that also help the penetration of the root system in depth, and the fast **drainage** of soil [Beven and Germann, 2013, van Schaik et al., 2014]. Through the production of casts and galleries, earthworms contribute to **soil structure maintenance** Blanchart et al. [1999], Pulleman et al. [2005] and **restoration** [Blanchart, 1992] as well as the creation of habitats for the microbiota [Brown, 1995]. For this, they are known as **ecosystem engineers** [Lavelle, 1988, Lavelle et al., 1997].

For its intrinsic and extrinsic values, edaphic biodiversity is receiving increasing attention [Phillips et al., 2020] in light of the threats posed by global and local environmental change. Earthworms are particularly affected by drought resulting from warming [Eisenhauer et al., 2014, Phillips et al., 2019], habitat disturbance, and passive dispersal of invasive species via human trade, in addition to intense farming strategies such as tillage and the excessive use of fertilizers [Kibblewhite et al., 2008, Bünemann et al., 2018]. In the context of global changes, the description of earthworms' niches is a prerequisite for understanding their response to environmental pressures and, in turn, the spatial emergence of their effects on ecosystem functioning. Since the 70s and Marcel Bouché's seminal works on French species, earthworm are regularly grouped into three

guilds a.k.a ecological categories [Bouché, 1972] based on microhabitat preferences, diet and demography. Recently, Bottinelli et al. [2020] proposed a fuzzy classification of earthworms into the latter categories based on morpho-anatomical traits, allowing to compute ecological distances and represent taxa with intermediate strategies.

Recently, functional trait-based approaches have received increasing attention in soil fauna ecology [Hedde et al., 2012]. In particular, anatomical and morphological traits have been linked to earthworm's habitat preferences [Marchán et al., 2016a, Briones and Álvarez-Otero, 2018]. Since ecological categories are built from traits, we assume that species responses would be more similar within ecological categories than between them. Furthermore, because traits are phenotypic expressions of a genetic makeup, it is not far fetched to assume that phylogenetically close taxa would have similar abiotic preferences.

Earthworms are known to be constrained by a combination of abiotic and biotic factors. The main abiotic factors limiting earthworms occurrences and abundances are temperature and soil moisture [Eggleton et al., 2009, Eisenhauer et al., 2014], resource availability especially for species with low mobility [Barot et al., 2007], as well as the physical and chemical soil properties. The latter include soil texture [Nuutinen et al., 1998], pH[Lavelle et al., 1995, Decaëns et al., 2008], and organic carbon content [Palm et al., 2013]. On the other hand, biotic factors include competition [Abbott, 1980, Jiménez et al., 2006, Decaëns, 2010] or facilitation [Jiménez et al., 2012] with other species as well as small-scale dispersal processes [Decaëns et al., 2008, Mather et al., 1988]. The factors affecting earthworms vary with scale. At the very local scale, earthworms are affected by the direct and indirect effects of vegetation cover, land use [Decaëns et al., 2008, Falco et al., 2015]. Management strategies, like liming [Potthoff et al., 2008], tillage [Pelosi et al., 2014] and crop residue management [Frazão et al., 2019] also highly influence them. At the regional scale, Mathieu and Jonathan Davies [2014] investigated post-glacial dispersal processes as an important biogeographical filter in France. However, this process was not identified as relevant at the West-European scale by [Rutgers et al., 2016]. At the continental scale, temperature and precipitation gradients along with land use are the main driving forces of West-european earthworms [Rutgers et al., 2016]. At the global scale, climate along with phylogeographic history rule [Brussaard et al., 2012, Phillips et al., 2019].

Several attempts have been made to identify the driving environmental factors that explain the spatial patterns of species and the habitat preferences of earthworms mainly using species distribution models (SDMs). SDMs have been applied to model single species occurrences or abundances from presence-absence or presence-only data using either mechanistic (process-based) or correlative methods. Johnston et al. [2014] used a process-based model that combines individual energy budgets, movement and soil profiles to predict the abundance and distribution of *Aporrectodea caliginosa* (Savigny, 1826) . Taking a correlative instead of a mechanistic approach, Marchán et al. [2015] applied the MaxEnt algorithm [Elith et al., 2011, Warren et al., 2010] to presence-only data to predict the relative probabilities of occurrence of *Hormogaster elisae* (Alvarez, 1977) in Spain as a function of vegetation, land use and bioclimatic conditions. Marchán et al. [2016b] applied the same approach to compare and disentangle the ecological preferences of different clades of the Hormogastridae, an endemic mediterranean family. Last, Shartell et al. [2013] applied a Generalized Linear Model to predict the percentage cover of the exotic *Lumbricus terrestris* (Linnaeus, 1758) in the Upper Peninsula of Michigan as a function of soil properties (pH, soil drainage) and landscape characteristics (topology).

In a multispecies setting, Schneider et al. [2016] used Boosted Regression Trees (BRT) [Elith et al., 2008] to predict the abundance of common species at the catchment scale in an agricultural system. A BRT model was calibrated for each species independently. The authors used plot-scale edaphic properties and topography-derived landscape covariates as a proxy for local climate and hydrology. BRT were also used in a similar context to model the abundance of epigeic, endogeic and anecic worms using for each ecological group the abundance of the others as covariates to estimate inter-group interactions [Palm et al., 2013]. All studies targeting multiple species assemblages fitted dedicated SDMs for each taxa. Moreover, these studies concerned common taxa with wide ranges (e.g *Lumbricus terrestris* (Linnaeus, 1758) , *Aporrectodea caliginosa* (Savigny, 1826) , etc.) and mainly in agricultural systems and forests, with a limited spatial scale.

Large-scale applications of distributions models at the community level mostly involved modeling the total biomass or taxonomic diversity. The latter were investigated at the country-wide scale in France [Mathieu

and Jonathan Davies, 2014, Rutgers et al., 2016], at the continental scale in Europe [Rutgers et al., 2016] using Generalized Linear Models and at the global scale [Phillips et al., 2019] using Generalized Linear Mixed Effect Models. As predictors, these studies used climate data (temperature, precipitation), soil physico-chemical properties, water retention, habitat cover and elevation. Unlike taxonomic diversity, the distribution of functional and phylogenetic diversity has not yet been explored for earthworm communities.

Many studies argue that biotic interactions can alter species ranges at larger scales than the mechanism's [Wisz et al., 2013] leaving a footprint in species distributions. Whether these findings hold true for soil invertebrates in general and earthworms in particular has never been explored [White et al., 2020]. Most SDMs on earthworms used abiotic conditions, or abundance of vegetation, or other soil fauna as predictors, overlooking potential inter-specific associations between earthworms arising from biotic interactions or other indirect mechanisms of dependency.

In the present study, we model the spatial distribution patterns of earthworm assemblages from the local to the country-wide scales in mainland France, in response to regional climate and local habitat conditions. Indeed, we explore an alternative way to overcome the current limitations in earthworm distribution models. For this purpose, we use a multi-task neural network to model biotic associations and predict species responses jointly with latent variables. The neural network part of the model enables it to capture interaction effects and non-linear responses to the environment. The multi-task part allows the modeling of multiple species simultaneously, avoiding the computational overhead of training many dedicated SDMs, and permitting rare taxa with few observations to gain statistical strength by being modeled alongside common taxa.

We compare (1) the predictive performances of our model on individual taxa and in average to SDMs. Also, using machine learning interpretability tools, we analyze the model's behavior and predictions to (2) identify crucial abiotic factors driving earthworms' distributions overall, (3) focusing on each taxa, we shed light on its habitat preferences, and response to various environmental gradients, and (4) explore commonalities in species' responses to reveal response groups. Afterwards, we analyze the association network uncovered by the model (5). Finally, using the calibrated model, we compute projections, on a grid of 1km across France, of metrics computed on predicted communities (6) including: ecological categories, community weighted mean traits and the diversity of earthworm assemblages at the local (alpha), regional (gamma) and its turnover (beta) in its three aspects: taxonomic, functional and phylogenetic.

6.2 Materials and Methods

6.2.1 Data acquisition and processing

Earthworms occurrences

The earthworm occurrence data, obtained from Bouché [1972], includes 5708 presence/absence points covering 1360 sites in mainland France and Corsica (Fig 6.1). We mapped the taxonomic information from the original dataset (131 taxa) to a more recent taxonomy (121 taxa) using the *drilobase* database [Cameron et al., 2016]. We used the latter in subsequent data processing and reporting.

We selected taxa that occurred in at least 5 sites, and removed sites where no local habitat characteristics were reported. The final dataset comprised 1346 sites and 77 taxa with heterogeneous taxonomic assignment levels among : Species, Subspecies, and Variety.

Phylogeny

The earthworm phylogeny was provided by (Marchàn et al, *in prep*). We transformed the phylogenetic tree into a distance matrix using the `ape` package [Paradis and Schliep, 2019]. For taxa with no phylogenetic information (25/77), we averaged information of siblings within the same subspecies (6/25), species (11/25) or genus (8/25).

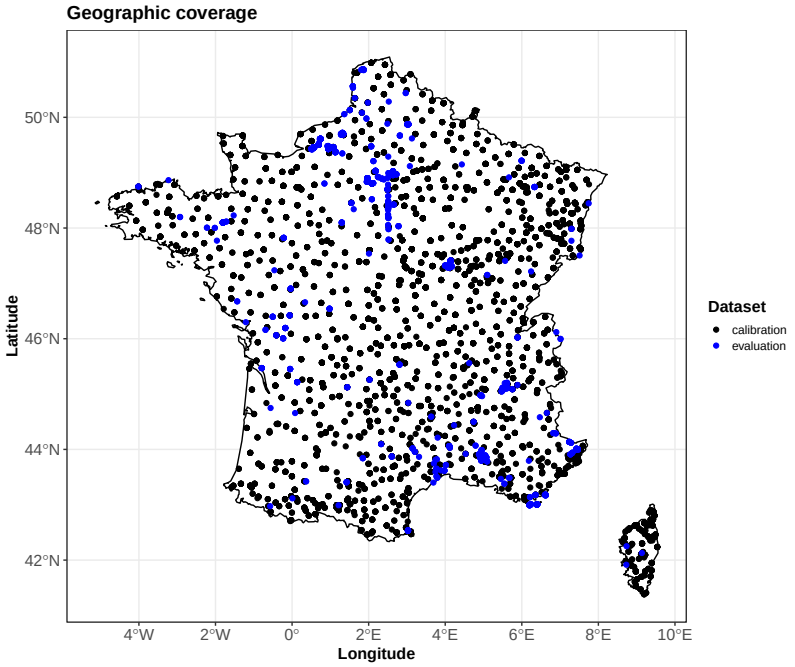


Figure 6.1: Geographic coverage of the calibration and evaluation dataset.

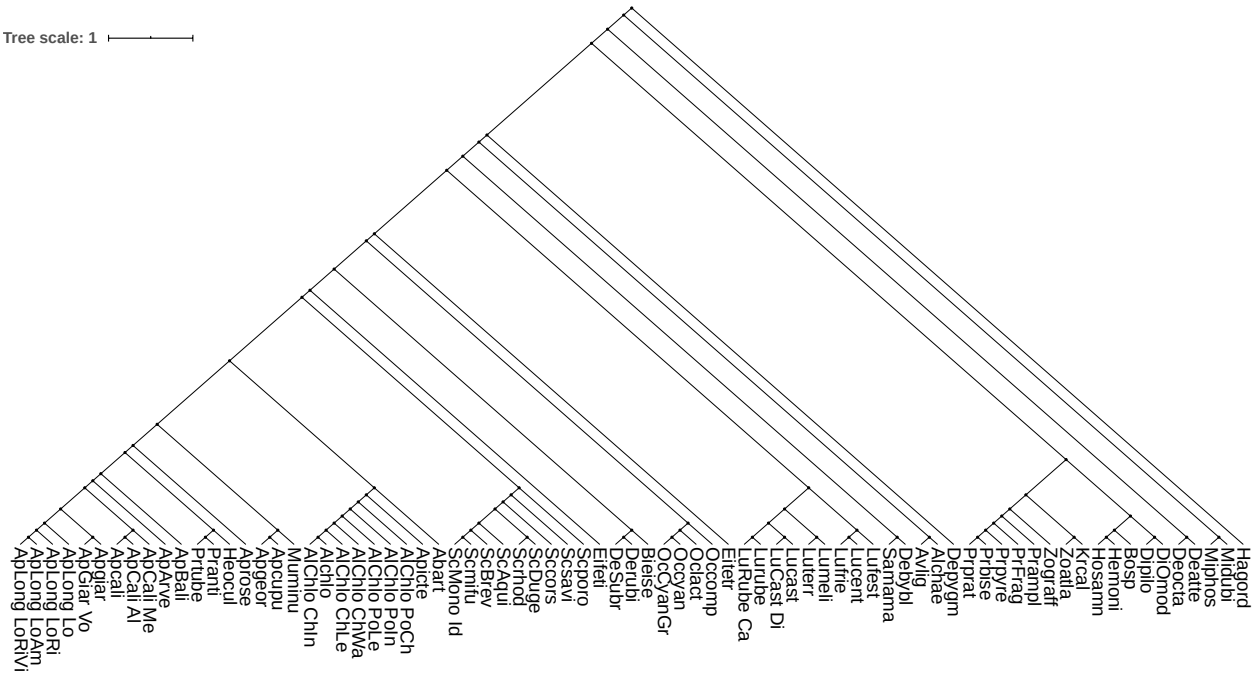


Figure 6.2: Earthworm phylogenetic tree. (provided by Marchan et al, in prep)

Traits and ecological categories

We extracted 25 morpho-anatomical traits from the database of soil invertebrate functional traits [BETSI, 2012]. We filled missing information by aggregating information from closely related taxa. We completed the dataset with a fuzzy classification of earthworms into ecological categories triangle with the three poles:*epigeic*, *endogeic* and *anecic* [Bottinelli et al., 2020] .

Environmental covariates

To model the distribution of earthworms, we collected environmental descriptors describing the climate (temperature and precipitation), soil physico-chemical, structural and hydrological properties, in addition to the land cover type. We built two datasets of environmental conditions with different spatial coverage and temporal relevance. **A calibration dataset**: limited to the occurrence data sites at the period of observation (1960-1980). **A projection dataset**: covering a grid of $1km$ resolution for the current period (2000-2020). We summarize in Table 6.1 the covariates used and the corresponding sources for each dataset.

Group	Variable	Unit, Type	Source (calibration)	Source (projection)
Temperature	- Annual mean temperature (TMeanY) - Mean diurnal range (TMeanRngD) - Isothermality (Tiso) - Temperature seasonality (Tseason) - Maximum of Warmest Month (TMaxWarmM) - Minimum of Coldest Month (TMinColdM) - Annual range (TRngY) Mean temperature of: - Wettest Quarter (TMeanWetQ) - Dryest Quarter (TMeanDryQ) - Warmest Quarter (TMeanWarmQ) - Coldest Quarter (TMeanColdQ)	Celsius	Worldclim [1960-1990] [Fick and Hijmans, 2017]	CHELSA [1979-2013] [Karger et al., 2017]
Precipitation	- Annual precipitation (PTotM) - Precipitation seasonality (Pseason) - Precipitation of - Wettest Month (PWetM) - Dryest Month (PDryM) - Wettest Quarter (PWetQ) - Dryest Quarter (PDryQ) - Warmest Quarter (PWarmQ) - Coldest Quarter (PColdQ)	mm		
Land cover	Land cover class	Categorical (16 levels)	[Bouché, 1972]	Corine land cover 2012 [Soukup et al., 2016]
Soil physico-chemical properties	Water acidity level (pH)	None	[Bouché, 1972]	ISRIC soilgrids [Batjes et al., 2020]
	Organic carbon (C)	mg/Kg soil		
	Carbon to Nitrogen ratio (C/N)	None		
Hydrological properties	Proximity to watercourses <50m (proxieaufast)	Binary	[Christophe et al., 2019]	Invariant
	Available water capacity of topsoil (awctop)	mm	European soil database v2 raster library (ESDB)	
	Base saturation of top soil (bstop)			
	Cation exchange of top soil (cectop)	cmol/Kg soil	[Panagos et al., 2012]	
	Water regime (wr)	Categorical (5 levels)		
Soil structural properties	Texture	clay silt sand	[Bouché, 1972] Donesol [Infosol, 2015]	ISRIC Soilgrids rasters [Batjes et al., 2020]
	Soil crusting class (crusting)	Categorical (6 levels)	ESDB [Panagos et al., 2012]	Invariant
	Erodibility class (erodi)	Categorical (6 levels)		
	Depth to gleyed horizon (dgh)	cm		
	Depth to impermeable layer (dimp)	cm		
	Dominant parent material (parmado)	Categorical (9 levels)		
	Packing density of top soil (pdtop)	g/cm3		

Table 6.1: **List of environmental descriptors**. We used the same sources for structural properties that are not likely to vary between the two periods (1972 todate). These are marked invariant.

Evaluation dataset

We collected a dataset of recently sampled geotagged earthworm observations containing 1126 occurrences collected in the period $[1998 - 2019]$ and covering $37/77$ of the modeled taxa. Due to the multiplicity of sources and the heterogeneous sampling effort, we treated this collection as presence-only data. We extracted environmental data for the corresponding locations from the projection dataset. Some observations contained metadata describing local land cover, when available we used this information instead.

6.2.2 Joint model of earthworms distribution

We aimed to predict the probability of presence of all selected taxa simultaneously under variable environmental conditions. We applied the joint model described in 3.3 with a *Bernoulli* distribution and a probit link $g = \Phi$ to accommodate the binary presence/absences. We fix the number of latent factors to at most 3 to facilitate the visualization. Moreover, we use L1-L2 regularization of the generative θ and variational ξ parameters.

Model training

In the following, we detail the steps followed to preprocess the data, train the model and select the optimal hyperparameters.

Feature preprocessing We applied a one-hot-encoding to categorical features and standardize numerical and ordinal features. Due to the large number of abiotic covariates and responding taxa, we investigated different feature selection/transformation approaches: (1) End-to-end (learnt transformation); (2) preprocessing via correlation-based selection using the Variation Inflation Factor; (3) preprocessing via principal component analysis.

Learning architecture To find the optimal configuration of network complexity and transfer amongst prediction tasks, we varied the shared feature learning network depth, and the width (number of neurons) of each layer.

Dealing with imbalance Due to the imbalance of presences and absences, random sampling of observations may result in training sets that contain only absences for rare taxa. To circumvent this issue and to enable a robust estimation of the abiotic niche for all taxa, we used Algorithm 3 that ensures at least 5 occurrences for each taxa in the sampled set i-e $min_occur = 5$.

Algorithm 3: Train-test split with imbalance

Input : Y: Occurrence matrix of dimensions $(N \times M)$, min_occur: integer, tsize: training set size
Output: $\mathcal{T}, \mathcal{V}$: set of training and validation observations

```
/* Initialization */
```
$\mathcal{T} := \varnothing$;
$\mathcal{V} := \{1, 2, \ldots, N\}$;
Req := REP(min_occur, M) ; `// Required occurrences`
S := $\{1, 2, \ldots, M\}$; `// Unsatisfied taxa`
Prev := REP$(0, M)$;
while S$\neq \varnothing$ **do**
  `/* Satisfy rarest taxa first */`
  potential := available presences for each unsatisfied taxa ;
  candidates := taxa with the smallest potential ;
  choice := select randomly from candidates ;
  $nidx$:= Req[choice] ;
  `/* Draw occurrence points */`
  selected := draw $nidx$ indices from potential[choice] ;
  $\mathcal{T} := \mathcal{T} \cap$ selected ;
  $\mathcal{V} := \mathcal{V} \setminus$ selected ;
  `/* Update requirements */`
  Prev := ROWSUM$(Y[\mathcal{T},])$;
  Req := Req - Prev ;
  S := $\{1 \leq t \leq M \mid Req[t] > 0\}$;
end
`/* Fill rest of training set randomly */`
ncomp := SIZE$(\mathcal{T}) - tsize$;
if ncomp < 0 **then**
  selected := draw uniformly $ncomp$ sites from $\mathcal{V}$;
  $\mathcal{T} := \mathcal{T} \cap$ selected ;
  $\mathcal{V} := \mathcal{V} \setminus$ selected ;
end

In addition to ensuring all taxa occur in each training set, we also weighted the individual likelihood presences (positive loss) with their odds relatively to absences (negative loss) (Eq 6.3). Last, we initialize the intercept of the mean predictions to the logit-prevalence (Eq 6.5) of each taxa, to prevent the model from fitting the bias in the data. The rest of the parameters were initialized using the `glorot-uniform` initializer [Glorot and Bengio, 2010].

$$p_i = \frac{1}{N}\sum_{s=1}^{N} y_{si} : \quad \text{Prevalence of } i^{th} \text{ taxa.} \tag{6.1}$$

$$m_{si} : \quad \text{Probability of taxa occurrence in observation site} \tag{6.2}$$

$$\mathcal{L}_{si} = w_i \underbrace{y_{si}\log(m_{si})}_{\text{Positive loss}} + \underbrace{(1-y_{si})\log(1-m_{si})}_{\text{Negative loss}} : \quad \text{Weighted binary cross-entropy} \tag{6.3}$$

$$w_i = \frac{1-p_i}{p_i} : \quad \text{Positive weight of taxa prediction task} \tag{6.4}$$

$$\beta_{0i} = \text{logit}(p_i) = \log(w_i) : \quad \text{Initial intercept value} \tag{6.5}$$

Training configuration We fit the model using the `adam` optimization algorithm for at most 400 iterations (epochs), using a batch size of 32. To prevent overfitting, we used *early-stopping* to stop the training whenever the objective function of the validation set stopped improving after more than 10 epochs.

Model selection and evaluation Since ROC-AUC is threshold-independent, we used the macro average ROC-AUC across taxa as a metric to select the best configuration. Particularly, we performed a 5×2 cross-validation coupled with a paired t-test [Dietterich, 1998] to compare models in pairs. We shortlisted two learning architectures that did not show significant differences.

Combining learning architectures with the preprocessing modes yielded in 6 configurations. We trained all 6 models on the full calibration dataset while varying the regularization strength in the range $\{1, 2, 5, 10\} \cdot 10^{-4}$. For each trained model, we quantified the recall of its predictions on the presence-only evaluation dataset. We retained the model with the best performance on the evaluation dataset. For this model, we evaluated for each taxa the ROC-AUC and True Skill Statistic (TSS) of the selected model predictions.

Comparison to species distribution models

For each taxa, we fit three single species distribution models: a Generalized Linear Model (GLM) and a Gradient Booosted Machine (GBM) with Bernoulli distribution in addition to a Random Forest (RF), using the BIOMOD package [Thuiller et al., 2016]. We evaluate the ROC-AUC and TSS as well as the recall on the evaluation dataset for each taxa and each SDM. We compare the per taxa scores on the training dataset to those reported for the proposed multi-task model (MTEC) using a Wilcoxon ranking test [Wilcoxon, 1992].

Unraveling the association network

We applied the graphical lasso algorithm to the posterior covariance matrix estimated using the full dataset size (n=1346). We used the extended Bayesian Information Criteria [Konishi and Kitagawa, 2008] (eBIC) with tuning parameter (gamma=0.5) to select the best partial correlation graph. The resulting graph encodes the direct dependencies between modeled taxa and is referred to as the association network. We analyze network topological properties and the distribution of associations with respect to the identity, prevalence and ecological categories of the associated taxa.

6.2.3 Abiotic drivers of earthworms distribution

Measuring feature contributions with SHAPley values

To identify the main driving factors of habitat suitability, we analyze the model's predictions across environmental conditions using machine learning interpretability tools. In this study, we use the framework of Shapley values (cf. Section 2.5). This approach is chosen for two reasons. First, it is model-agnostic therefore compatible with the current model and any ulterior updates (e.g extensions to new taxa). Second, it measures for each prediction task (here a given taxa) and a given observation the contribution of each feature to the predicted output (here probability of occurrence). Specifically, it computes the difference of the prediction to the average prediction across observations as a payout. Then, it attempts to partition the payout into individual features contributions fairly. Therefore, the shapley value can be interpreted as a partial response to a feature. In practice, we use `SHAP` an efficient implementation, via the R package `iml` [Molnar, 2020].

Concretely, assuming N observation sites, a pool of M taxa and P features, the method takes in a matrix N-by-P of environmental conditions and a matrix N-by-M containing the corresponding model's predictions. The result is the mean and standard deviation of the SHAP estimate for each feature (P), taxa (M), and observation (N), collected in a matrix of size [N.M.P,3+2].

Positive (resp. negative) SHAP values indicate that the current value of the feature increases (resp. decreases) habitat suitability, negative values indicate that it decreases habitat suitability relatively to the average prediction. A feature can have a positive effect in an interval of its domain and a negative or neutral effect in another interval. Similarly, the SHAP can be different for the same level of the feature value in presence of interaction effects. To assess the response shape of a given earthworm taxon into a feature, we visualize the dependence plots representing the SHAP values of the target (feature, taxa) pair along the feature gradient.

Finally, the absolute value of the SHAP statistic measures the local importance. Local importance values across geotagged observation sites can be mapped to assess the spatial variability of feature relevance and identify limiting factors in different locations. Finally, the average of local importance across observations yields the global feature importance for a given taxa.

Comparing earthworm responses to environmental gradients

First, we adopt a visual approach that consists in displaying the SHAP dependence plots side by side to identify the preferences of each taxon for the given variable (e.g. preferential pH range) and compare them between taxa. Then, we evaluate the variability of their responses by calculating distances between their partial responses (SHAP values). The previous analyses are conducted for each environmental variable and by thematic group of variables. Finally, we identify response groups specific to environmental factors by applying a hierarchical clustering algorithm from the distances computed previously. We selected the smallest number of clusters maximizing the GAP statistic [Tibshirani et al., 2001] locally (up to 6 clusters). We limited our discussion to the resulting groups for edaphic properties, land use, and precipitation.

6.2.4 Projection to current conditions

Using the environmental conditions collected in the projection dataset, we compute the community composition across France with a 1 km grid. We filtered out sites where land cover was not habitable by earthworms (e.g water bodies, roads and urban infrastructures), since they can result in empty communities.

The predicted occurrence probabilities for each taxon are typically used to produce habitat suitability maps for the current conditions. Here, we focus our discussion on community-level indicators computed from the vectors of taxa probabilities of occurrences. Particularly, we compute and project community weighted means of functional traits and relative proportions of ecological categories in France. Finally, we compute earthworm taxonomic, functional and phylogenetic richness at the local (alpha) and regional (gamma) scales as well as the turnover (beta) using a radius of 100 km. We use the framework developed by Leinster and Cobbold [2012] for diversity partitioning (Details in 6.5.5).

6.3 Results

6.3.1 Predictive performances

The comparison with SDMs (Fig 6.3) showed that the use of a multispecies model improved the performances on the calibration dataset in average significantly compared to single species models, regardless of their complexity. Moreover, SDM models did not converge on some taxa with low prevalence (less than 10). For all models, predictive performances were lower for widespread taxa and were variable for less prevalent taxa (Table 6.5.3). In the case of the selected MTEC architecture, AUC values were particularly low (around 0.69) for the widespread *Lumbricus castaneus disjonctus* (Savigny, 1826) and *Octolasium cyaneum* (Savigny, 1826) .

Furthermore, we reported comparable to better performances on the most widespread peregrine species (*Aporrectodea caliginosa* (Savigny, 1826) , *Lumbricus terrestris* (Linnaeus, 1758)) for SDMs (especially RF and GBM) than MTEC. Indeed, dominant species did not benefit from the parameter sharing as much as rare and less prevalent taxa did. However, the ability of the MTEC model to predict recent occurrences of the evaluation dataset was higher than SDMs both in average and for most taxa, including those for which performances were higher for SDMs. This result suggests that some of the SDMs might be prone to overfitting.

MTEC is a Joint Species Distribution Model (JSDM) with latent variables. The underlying probabilistic model is similar to generalized linear latent variable models (GLLVM), except that the proposed method relies on amortized (fixed parameters) inference using a type of neural networks called variational autoencoders. Recent reviews on JSDMs showed that predictive performances are similar or lower in JSDMs than with stacks of SDMs [Wilkinson et al., 2019, Norberg et al., 2019, Pichler and Hartig, 2020]. The results we obtained on earthworms did not follow this trend. A possible explanation is that the mean response of

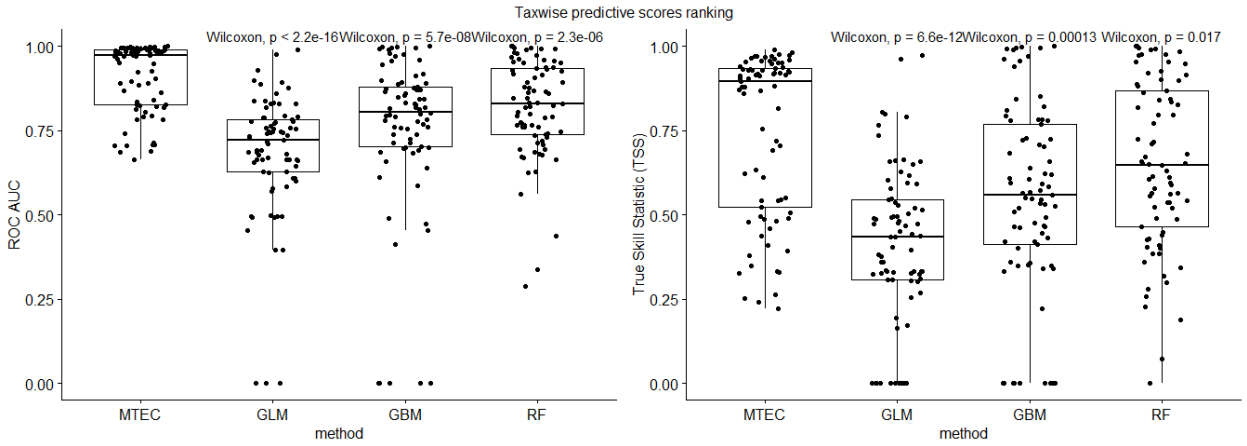


Figure 6.3: Comparison of joint and separate models prediction scores. Wilcoxon ranking test of the hypothesis that the score of the MTEC model is greater than the score of the SDM i-e **"Score(MTEC)> Score(SDM)"** for each of the three single species distribution models : GLM, GBM, RF.

Model	TSS	ROC-AUC	Eval Recall
MTEC	0.746 ± 0.243	0.912 ± 0.102	0.818 ± 0.317
GLM	0.411 ± 0.234	0.681 ± 0.186	0.519 ± 0.389
GBM	0.552 ± 0.281	0.753 ± 0.238	0.302 ± 0.267
RF	0.646 ± 0.247	0.816 ± 0.145	0.250 ± 0.314

Table 6.2: Predictive performances distribution over taxa for the proposed multi-task architecture (MTEC) and single species distribution models (GBM, GLM, RF).

MTEC is a neural network which has the ability to capture complex non-linear response shapes. However, the superiority of MTEC might not hold in comparison to other machine learning models since we could not test all possible methods.

Modeling rare taxa relates to the more general problem of imbalanced classification, which is challenging regardless of the machine learning model at hand. In our case, the better performances observed on rare taxa may be due do two factors: (1) benefits arising from sharing representations with common taxa with higher statistical power, (2) loss weighting strategies that force the model to pay more attention to the presences of these species than their absences.

The better generalization power of MTEC can be explained by the enforced regularization through sharing environmental feature extraction amongst prediction tasks. In this particular setting, the relevance of this technical choice can be justified by the mono-guild nature of the taxonomic pool. Indeed, earthworms broadly share the same habitat, morphological constraints, physiological needs and resource types despite having differential adaptations. Therefore, their perception of the environmental conditions in terms of relevant features does not vary much. But, we cannot conclude that feature sharing is beneficial beyond the scope of this application as it might not extend, for instance, to above-ground multi-trophic communities.

6.3.2 Habitat suitability

Abiotic drivers of earthworm habitat suitability

Land cover, precipitation extremes and temperature variability were the most important abiotic parameters overall (Fig 6.4). These results are consistent with global patterns reported by Phillips et al. [2019] who identified precipitation and land use followed by temperature, soil and altitude as the main drivers of earthworms distribution. Although our results are on a mesoscopic scale (France) in comparison to Phillips et al. [2019], the geographic extent of our work encompasses climates and land uses that are sufficiently contrasted

for the following hierarchy to appear: at fine scales (landscape, catchment) soil properties are predominant but at larger scales where climate and land use change strongly, climate becomes the main driver of species distribution.

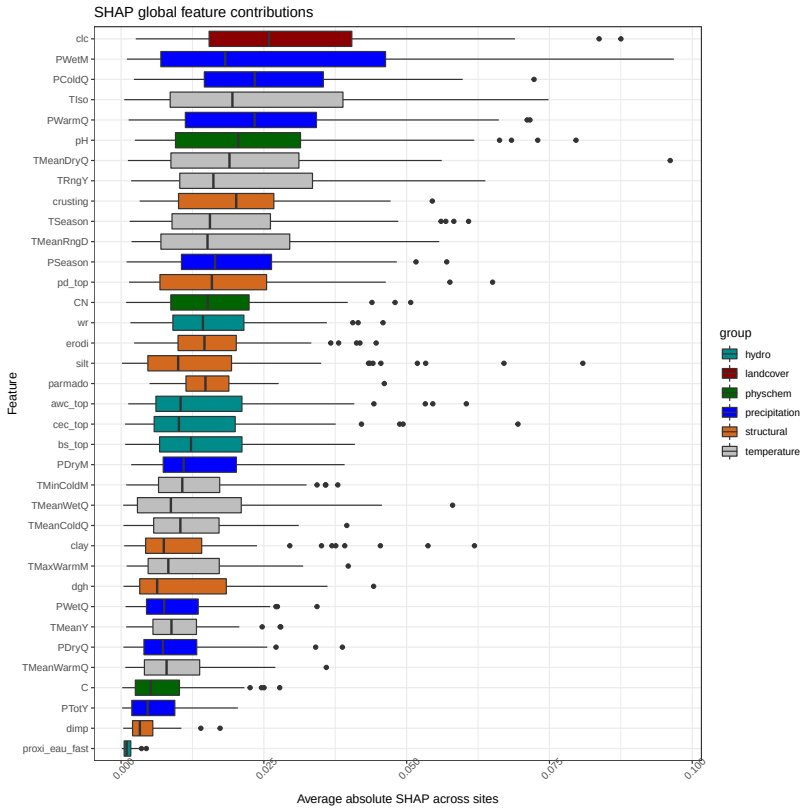


Figure 6.4: Overall feature importance. For a given taxa, feature importance is measured as the average across sites of the feature's contribution to the habitat suitability probability. The distribution of importance amongst taxa for the features used by the model is shown in descending order.

Precipitation regimes affect soil moisture, which conditions earthworms' activity and development. Indeed, Eisenhauer et al. [2014] showed through experimental warming that earthworms development is limited by temperature only to the extent that it induces drought.

In fact, total precipitation had little influence on overall habitat suitability predictions. Instead, the model relied on precipitation of the wettest month (autumn in France), coldest and warmest quarters (winter and summer) and variability over the year (seasonality). While precipitation extremes are known to affect earthworms, the importance of seasonality means that some species have developed strategies to deal with high climate variability both during day and between month (e.g. Scherotheca). On the other hand, several species clearly prefer more uniform climatic conditions (e.g. N-W France). These species are more likely to be species belonging to the 'epigeic' pole of Bouché's triangle.

The influence of temperature in ecoregions was carried by different features for different taxa, but we found similar patterns within ecological categories (Fig 6.5). The epigeics *Lumbricus castaneus* (Savigny, 1826) and *Lumbricus friendi* (Cognetti, 1904) are affected by short-term temperature fluctuations reflected by the mean diurnal range on coastal regions and isothermality elsewhere. On the other hand, endogeics *Aporrectodea caliginosa* (Savigny, 1826) and *Aporrectodea rosea* (Savigny, 1826) were driven by temperature extremes and long-term trends (yearly average, seasonality). Epi-anecics *Aporrectodea giardi* (Ribaucourt, 1901) and *Aporrectodea longa longa* (Mrsic and Sapkarev, 1988) responded to both short-term climate events, long-term trends and extremes. This pattern shows that the soil acts as a protection layer against above-ground pressures in this context heat strokes, and buffers their effects.

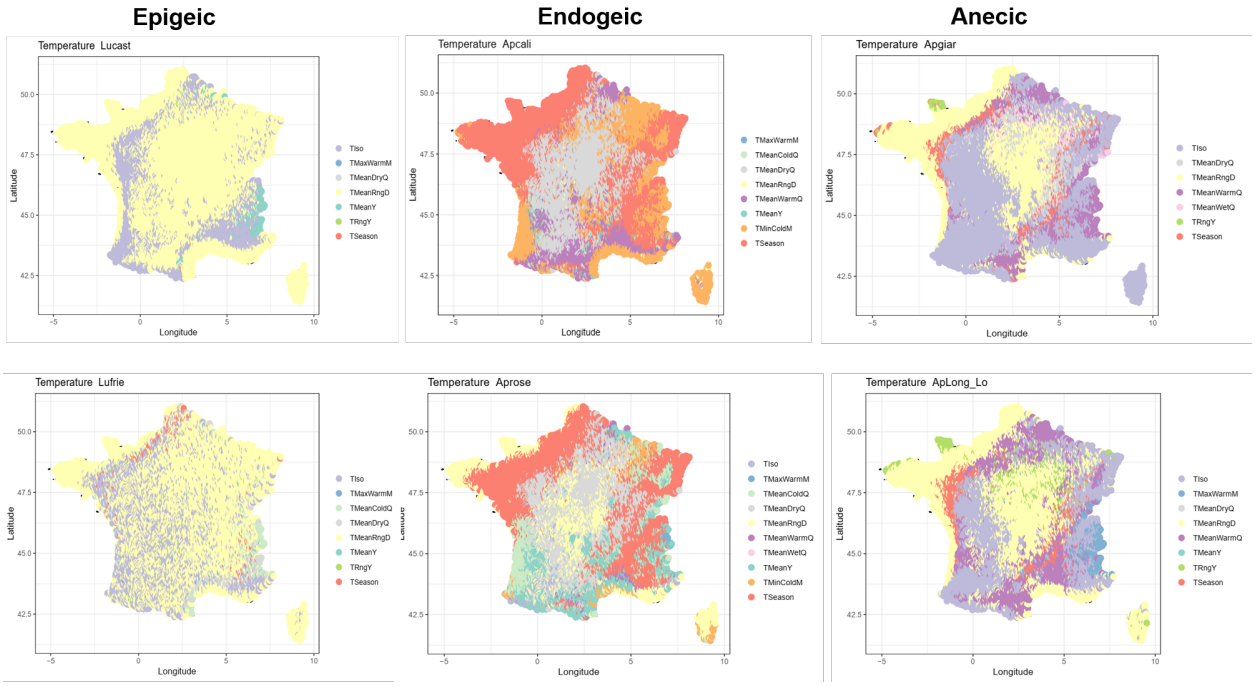


Figure 6.5: Limiting (locally important) temperature factors for 6 common taxa (2 epigeic, 2 endogeic, 2 anecic earthworms).

In terms of edaphic properties, soil acidity was in the most important variables. The pH is regularly recognized as one of the most discriminating edaphic factors [Bouché, 1972, Lavelle et al., 1995]. Soil pH influences plant growth, plant litter on the soil, but also the availability of metals (Al, Zn, Cu, Cd) that can reach toxic levels [Spurgeon and Hopkin, 1996]. On the other hand, the quantity of soil organic carbon content had a lower importance than its quality (seen through C/N).

The water regime as a proxy for soil moisture reflecting the amount and duration of the drainage had a great importance for most taxa whereas proximity to water courses had negligible influence. Land cover already included an "inland wetland" class which must have been exploited by the model instead.

In terms of structural properties, soil packing density had a greater impact at the assemblage level than texture. This result is unsurprising since packing density (pd_top) was estimated from soil texture and structure classes using pedo-transfer rules [Hiederer, 2013]. During the preprocessing phase, we converted descriptions of the texture into classes of the GEPPA texture system [Moeys, 2018], then we used the class centroids in the texture triangles defined by (%clay,%silt,%sand) as features to ensure compatibility with texture rasters. This transformation might have induced an information loss and reduced the discriminating power of textural features. Regardless, texture was the most discriminating factor for other species like *Dendrodrilus subrubicundus* (Eisen, 1874) and *Eiseniella tetraedra* (Savigny, 1826) that are subservient to sandy soils near water bodies.

In conclusion, the ranking shown in Fig 6.4 reflects the role of features for determining community composition overall. However, the hierarchy of features tends to vary amongst taxa as are the shapes of their responses.

Response of earthworms to environmental gradients.

Due to the high number of model output, we only presented a select set of results to illustrate the model's potential. We explored the response shape of taxa to three groups of environmental features of high overall importance and with direct link to global change factors: precipitations, land cover and soil characteristics (C content and pH). We discuss these results in light of the ecological categories *sensu* Bottinelli et al.

[2020].

Precipitation response groups We projected the response (SHAP) values of all taxa using the first two components of a Principal Component Analysis (PCA), explaining 60% of variation (Fig 6.6).

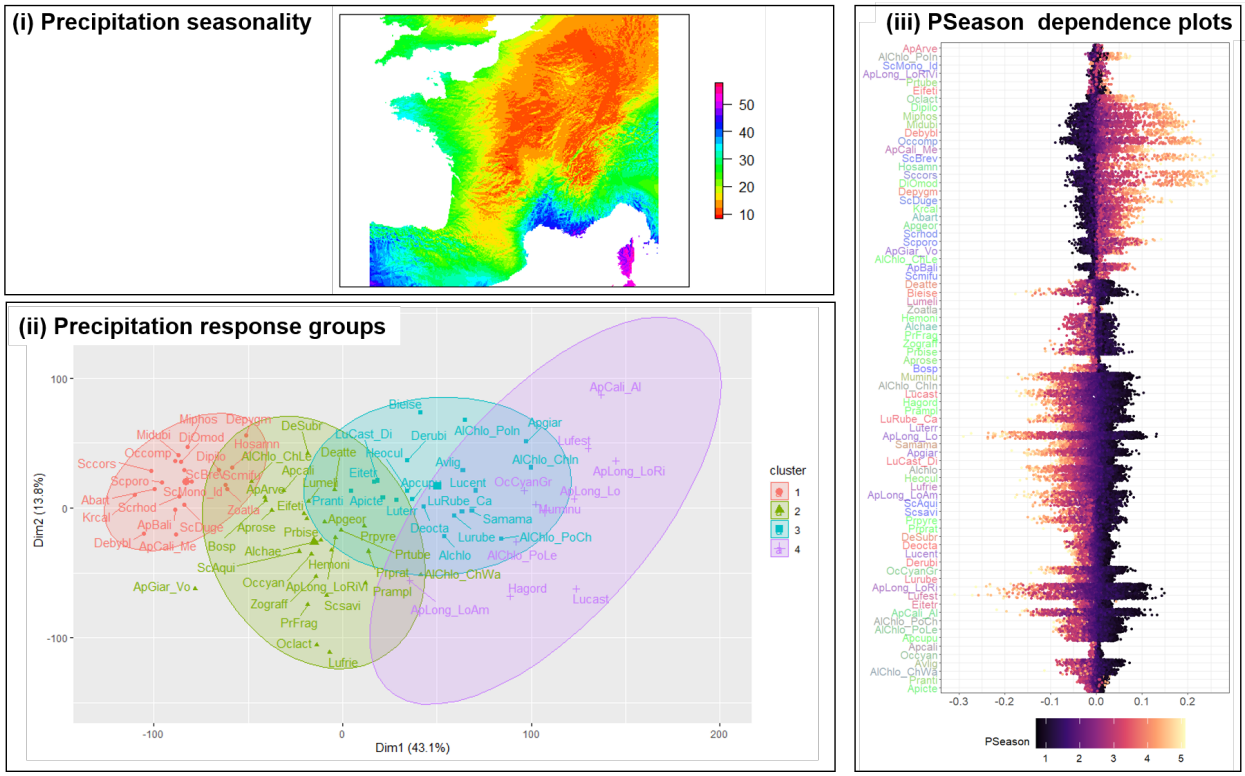


Figure 6.6: Response of earthworms to precipitation and its variability. (i) Precipitation seasonality in current climate (from CHELSA database). (ii) Response groups to precipitation conditions overall: we project the partial response (SHAP) to the precipitation across all observations for each taxa, using Principal Components Analysis. (iii) SHAP dependence plot to Precipitation Seasonality.

We interpreted PCA1 as a gradient of precipitation seasonality, with high seasonality correlated with negative coordinates . PCA2 interpretation was not obvious, and we proposed that it may reflect total precipitation. Four clusters were selected by the GAP criterion in a hierarchical clustering procedure on all PCA axes. The cluster 1 was first separated from the rest. It mostly gathered Mediterranean species with strong ability to live in xeric, warm and dry, environment (e.g. *Scherotheca corsica*, *Hormogaster samnitica*, *Kritodrilus calarensis*). The species in this cluster presented a positive response to precipitation seasonality. Anecic is the dominant ecological group, probably because those species avoid the adverse climate conditions by remaining deeper in the soil. The second cluster contained many species belonging to the genus *Zophoscolex*, *Prosellodrilus*, *Scherotheca* and *Hemigastrodrilus* that are endemic of the South-Western France. It also included some eurytopic, peregrine species like *Al. chlorotica* or *Ap. rosea*. There are many endogeic and anecic species, but also some epigeic. We interpreted this cluster as a mesophilous with warm-tolerant species group. The third cluster is composed by many species located near the epigeic pole of Bouché's triangle with 'full' epigeic species (*E. tetraedra*, *S. mammalis*, *D. octaedra*), epi-anecic species (*A. giardi*, *L. centralis*). Living at soil surface, these species require wet conditions to maintain themselves. The last cluster gathers few species, some of them known as riparious. We propose that the species of this group share a high affinity for a precipitation regime uniformly distributed during the year.

Land cover preferences The multi-species plots of dependency to land cover (Fig 6.7) highlighted striking differences in earthworm preferences. Particularly, inland wetlands and forests had the most contrasting

effects among species. Most species had a positive response to humid soils, but a few riparian epigeics such as *Eiseniella tetraedra* (Savigny, 1826) and endogeics including *Helodrilus oculatus* (Hoffmeister, 1845), *Haplotaxis gordioides* (Hartmann, 1821), and *Octolasium cyaneum gracile* (Savigny, 1826), had a strong preference for marshes and/or peat bogs. On the other hand, most taxa have a neutral or negative response to forest soils, except a very limited and distinct pool made exclusively of litter-dwelling epigeics including *Dendrobaena octaedra* (Savigny, 1826), *Dendrobaena attemsi* (Michaelsen, 1902)and *Eisenia foetida* (Savigny, 1826) , etc. These taxa had a preference (positive SHAP) for closed canopy environments and aversion (negative SHAP) to open-spaces like pastures, grasslands and scrublands. Other epigeics were positively associated with pastures.

Pastures hosted the richest set of species from all ecological categories, including open-space epigeics and epianecics, slightly dominated by endogeics (Table 6.3). This pattern could also be an artefact of the sampling effort. In-between forests and pastures, shrubs and meadows were co-dominated by endogeic and anecic worms. The latter, mainly composed of the genus *Scherotheca*, were preponderant in Mediterranean garrigues of *Golfe du Lion* and *Corsica*.

Forests, pastures and shrubs/herbaceous land cover classes comprised $\sim 80\%$ of the calibration dataset. In contrast, arable and permanent crop lands were underrepresented in the training dataset. Their composition included widespread peregrine species and a few endemic species. Their median ecological categories assignement included mostly anecics and endogeics (Table 6.3). The lack of epigeics could be related to farming practices which affect the surface fauna. These patterns were reflected in the shap responses (Fig 6.7) which showed no species with a strong affinity to agricultural sites, but a few with a strong aversion (negative shap values) towards them.

Land cover class	**Proportion (%) of calibration set**	**%Epigeic**	**%Endogeic**	**%Anecic**
Forests	24	85	11	4
Scrublands & Open-space herbaceous vegetations	20	16	44	40
Pastures	42	30	41	29
Agricultural lands (arable, permanent crops)	5.3	8	43	49
Inland wetlands	5.3	38	38	24

Table 6.3: Median of ecological categories assignment probabilities amongst observed taxa in various land cover types.

Land cover response groups In the PCA projection of the responses, the first two components explained about $\sim 80\%$ of the variation. The first axis (62% of variance) positioned modeled taxa in the gradient from pastures to forests. The clustering algorithm applied to all dimensions identified six response groups (Fig 6.7), three of which were associated with distinct land cover classes: forests, pastures and inland wetlands. Taxa with affinities to shrub and herbaceous vegetation were divided into two groups on the forest-to-pasture gradient: scrub woodlands and meadows. Another intermediate group was identified for species from wet grasslands.

Edaphic preferences and response groups. We analyzed the response shapes of earthworm taxa to soil acidity and organic matter quality (measured by the C:N ratio) and quantity (organic carbone), the top edaphic properties exploited by the model. The dependence plots are reported in Fig 6.8.

The first two components of the PCA of the responses to pH and C:N ratio explained $\sim 70\%$ of the variation and ordinated species according to tolerance to soil acidity and organic matter quality. At one end, we find the typical forest assemblages. They include epigeics and epi-endogeics that survive in spite of acidic soils by taking advantage of the rich organic matter in the litter. In this group, we also found endogeics of the *Zophoscolex* genus, described by Qiu and Bouché [1998] as acid-tolerant. At the other end are taxa that prefer neutral, slightly alkaline environments despite poorer organic matter. An intermediate group

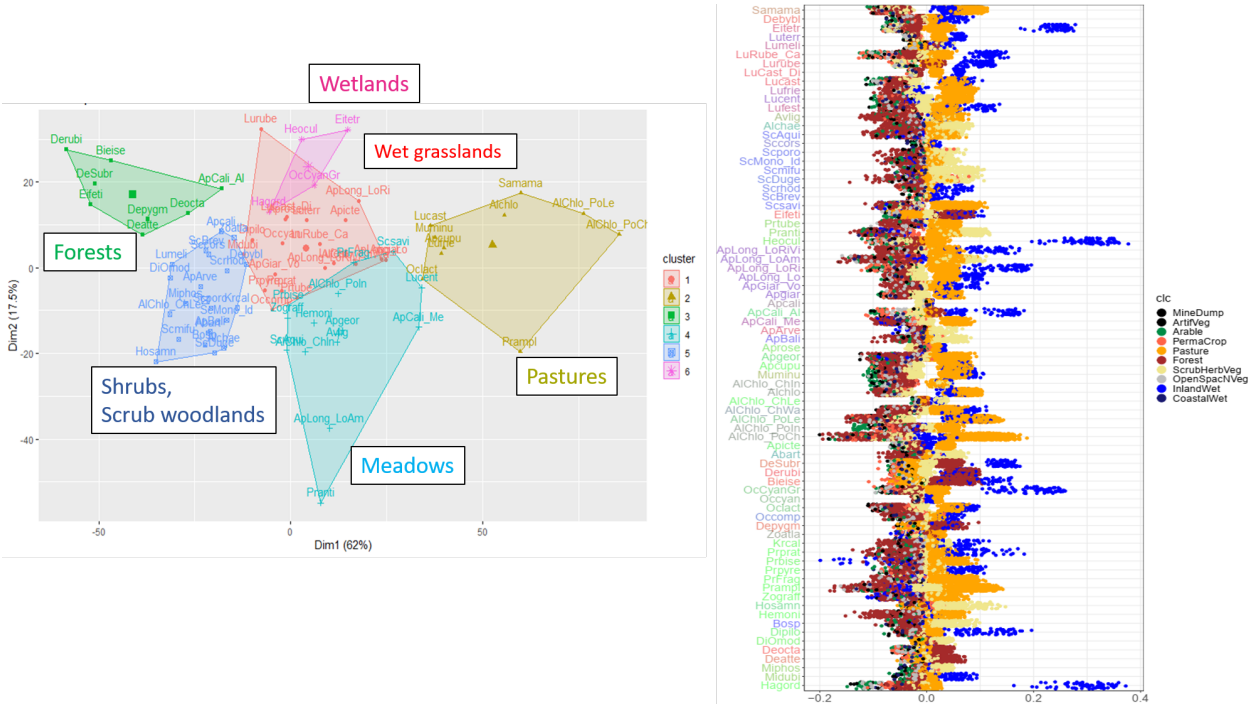


Figure 6.7: Response to land cover type. On the right, SHAP dependence plot of modeled taxa to land cover. Labels are colored according to the ecological categories, and ordered following their phylogenetic relatedness. On the left, we project the partial response (SHAP) to the land cover across all observations for each taxa, using Principal Components Analysis. The first two components explain $\sim 80\%$ of the variation in earthworm's responses.

connects the two extremes of the gradient. It is made of species which showed little sensitivity to pH values (SHAP values closes to zero), or with preference for the average pH value (slightly acidic 6). Riparian taxa stand out in a distinct cluster as they live in neutral wetland soils characterized with a sandy texture and an average quality organic matter.

Earthworms have different preferences in terms of organic matter quality [Curry and Schmidt, 2007]. In tropic ecosystems, endogeics are further classified depending on the quality of the consumed organic matter, from the richest to the poorest: polyhumic, mesohumic and oligohumic [Lavelle, 1988]. In our study, the gradient from "acid soils with rich SOM rich" to "alkaline soil with poor SOM" highlights a trade-off between tolerance to acidity and ability to degrade the carbon.

To assess the ability of the model's outputs to identify relationships between earthworm responses and earthworm traits, Fig 6.9 shows the results of linear models of the earthworms responses (measured by SHAP values) to the amount of soil organic carbon, with different slopes and intercepts per typhlosolis type. The typhlosolis is a dorsal flap that runs along the interior of the intestine, effectively increasing the absorption area by that of its inner surface. Its function is to increase intestine surface area for more efficient absorption of digested nutrients [James and Cunningham, 1989]. The slopes highlighted two opposing preferences towards the amount of organic matter: species with low exchange surface (no lamellae) preferred rich soils with abundant organic matter as opposed to species with wide exchange surface (lamelated typhlosoles) whose habitat suitability decreased in carbon-rich soils. This corroborates the hypothesis linking the greater convolution of intestinal surface to lower quality diets [Marchán et al., 2016a].

6.3.3 Earthworm association networks

The proposed method recovered a network of 31 significant pairwise associations, resulting in a 0.015 density. Only 31 out of the 77 modeled taxa were involved in these associations. We restricted the visualization of

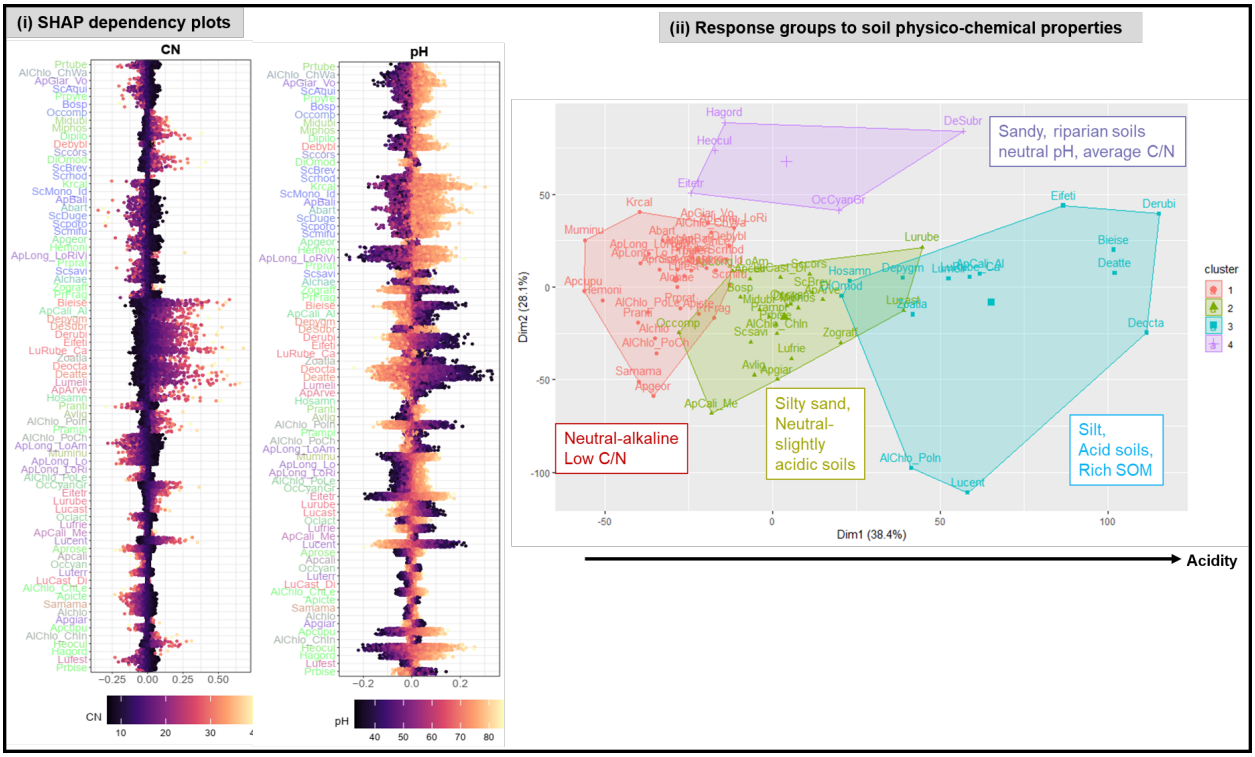


Figure 6.8: Response groups to soil physico-chemical properties. (i) visualizes the response of each taxa to soil acidity level (pH) and Carbone to Nitrogen ratio (CN). (ii) shows the clustering of taxa based on their responses to pH and CN (i) and Organic Carbone content. Clusters represent groups of similar response (response groups) to physical and chemical properties of the soil. (iii) highlights the cluster assignations in a PCA projection of the response of earthworm taxa. Annotations are done manually based on optimal conditions for each taxa, extracted from (i).

nodes to these taxa (Fig 6.10). It is worth noting that the partial correlations (associations strengths, edge attributes in the graph) had low and comparable strengths ranging between 0.14 and 0.22 in absolute value. This shows that earthworm distribution is rather driven by abiotic conditions after all. Nevertheless, some very interesting patterns arose from the association network.

The network was composed of three connected components, one of which included the majority of nodes (80%) and edges (87%). This component involved a hub of 4 of the most prevalent taxa: *L. castaneus*, *S. mammalis* and two subspecies of *A.caliginosa*: *A.caliginosa.caliginosa* which is present throughout the territory but rarely alongside the southern-distributed *A.caliginosa.meridionalis* as reflected in their negative associations. A positive ternary association was detected between *L.rubellus*, *A.longa.ripicola* and *D.rubidus.rubidus* three epigeics-epianecics with a predominantly northern distribution.

Few studies explored earthworm associations, most of them involved laboratory experiments [Edwards and Lofty, 1982, Butt, 1998]. Many authors stressed a greater strength of intraspecific than interspecific interactions [Uvarov et al., 2019]. Competition was regularly proposed as the main driver, e.g. by the trophic competition for the litter-dwelling, whereas the soil dwellers were presumably determined by competition for the space [Uvarov et al., 2019]. Earthworm densities are often high in lab experiments to exacerbate processes. However, in cases of high population density in the field, niche separation or migration away from competitive pressure may help to regulate a multi-species population to a given level. This may not be possible to observe in laboratory experiments, probably leading to an increase in competitive interactions. Since we worked on field-collected data, the detection of many neutral associations and of twice as more positive than negative associations is a clue that interspecific competition is low at the studied scale. Indeed, we did not exclude the possibility of competitive exclusion at a fine scale (plot), but our data set and metadata

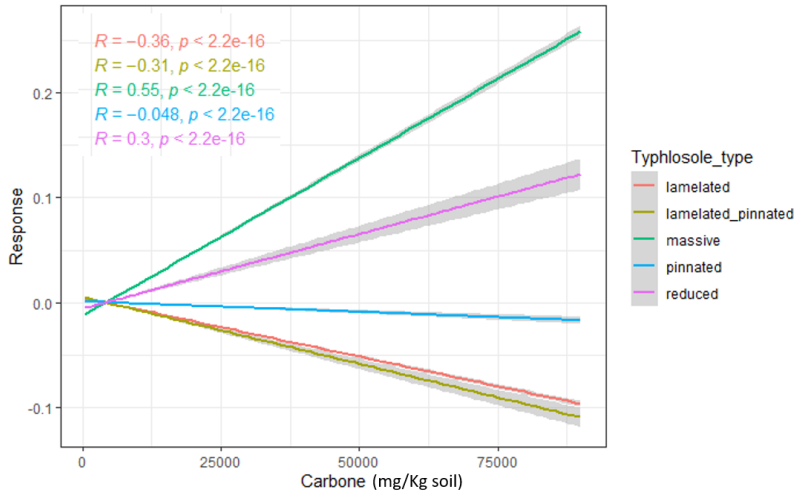


Figure 6.9: Response of different typhlosolis types to soil organic carbon content. Types of typhlosolis by decreasing exchange surface: lamellated, lamellated pinnated, pinnated, massive, reduced.

set do not allow the detection of these dependencies at fine scale [Decaëns, 2010]. This result suggests that associations are scale-dependent.

Richard et al. [2012] explored spatial associations at the same sampling scale (plot) as our data, within meadows in Normandy. Our findings did not agree with their results for several reasons. First, the network estimated by our model is country-wide which further supports the scale-dependency hypothesis. Second, it was inferred across different land uses (not restricted to meadows). The third and most plausible reason is that the nature of associations in the two studies is different. The previous study analyzed the co-occurrence or conversely spatial segregation. In theory, the co-occurrence of two species may result from the fact that they share the same abiotic requirements at the study scale without necessarily having any form of interaction. Similarly, spatial segregation of two species may result from opposite abiotic requirements rather than competition. By taking abiotic factors into account, the associations reflected by our model capture the dependencies that are not due to the environment as measured by abiotic features.

Nevertheless, our proposed approach can still detect associations arising from a shared dependence on an unmeasured environmental factor instead of an interaction. For instance, the majority of species in the network were epigeics, epi-endogeics or epi-anecics. This pattern might be due to an unmodeled effect of an above-ground factor such as the type of vegetation cover which controls the litter quality (leaf C:N) and seasonality (deciduous vs evergreen trees). We also observed that species placed at the "epigeic" pole of the Bouché's triangle (epigeics, epi-endogeics or epi-anecics) were mainly involved in positive associations with endogeics and anecics. This suggests a possible facilitation by the use of burrows and galleries built by soil-dwelling species as a refuge from above-ground disturbances such as cold, wind or drought episodes or to avoid above ground predators (birds, mammals or lizards for instance).

Incidentally, we reported the quasi-absence of endemic/narrow range species in the association network composed of a majority of peregrine species. We assume that the ability of the eurytopic species to spread and colonize previously inaccessible areas might be due not only to their competitive and dispersal abilities [Mathieu and Jonathan Davies, 2014] but also to their positive interactions with other species that are reflected in the association patterns.

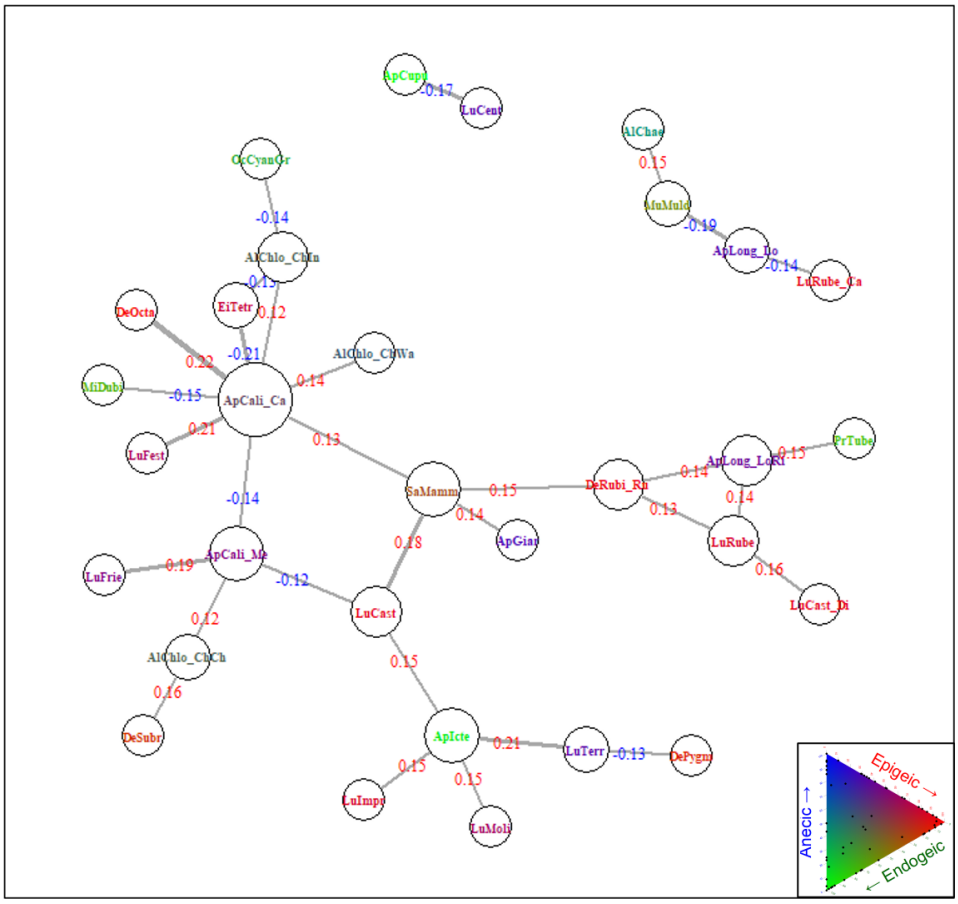


Figure 6.10: Earthworm association network. The graph represents the residual partial correlations (edges) between modeled taxa (nodes). Representation is restricted to taxa that are involved in at lease one significant association. Nodes are colored according to the ecological category of the taxa (colormap provided in the legend), and sized proportionally to their degree (number of associations). Positive (resp. negative) associations are annotated in red (resp. blue).

6.3.4 Mapping earthworm communities

The advantage of modeling community composition instead of modeling diversity directly [Phillips et al., 2019, Rutgers et al., 2016] is that from the predicted communities one can compute and represent geographically not only the distribution of species but also the average trait values and the diversity of the assemblages in all its aspects: taxonomic, functional and phylogenetic.

Ecological categories distribution

The ecological categories distribution map (Fig 6.11, right) shows that different geographical areas are dominated by different ecological categories. Areas of bare rock in alpine heights, watercourses, highways and artificial territories (large metropolises such as Paris, Lyon, Toulouse) are shown in white as they are non habitable land uses for earthworms.

On average, the communities presented a balanced composition with a few exceptions here and there. In the South, the genus *Scherotheca*, whose species are true anecics, dominates communities with low taxonomic richness in the dry environments of the Mediterranean scrublands. High dominance of anecics is limited to the "Golfe du Lion", bounded to the North by the Massif Central, to the East by the Alps and to the West by the Garonne river and the Pyrenées moutains. Epigeics dominate areas in Aquitaine, in the western tip of Brittany in the Armorican Massif, in the marshes and grasslands of the Cotentin penisula as well as in the

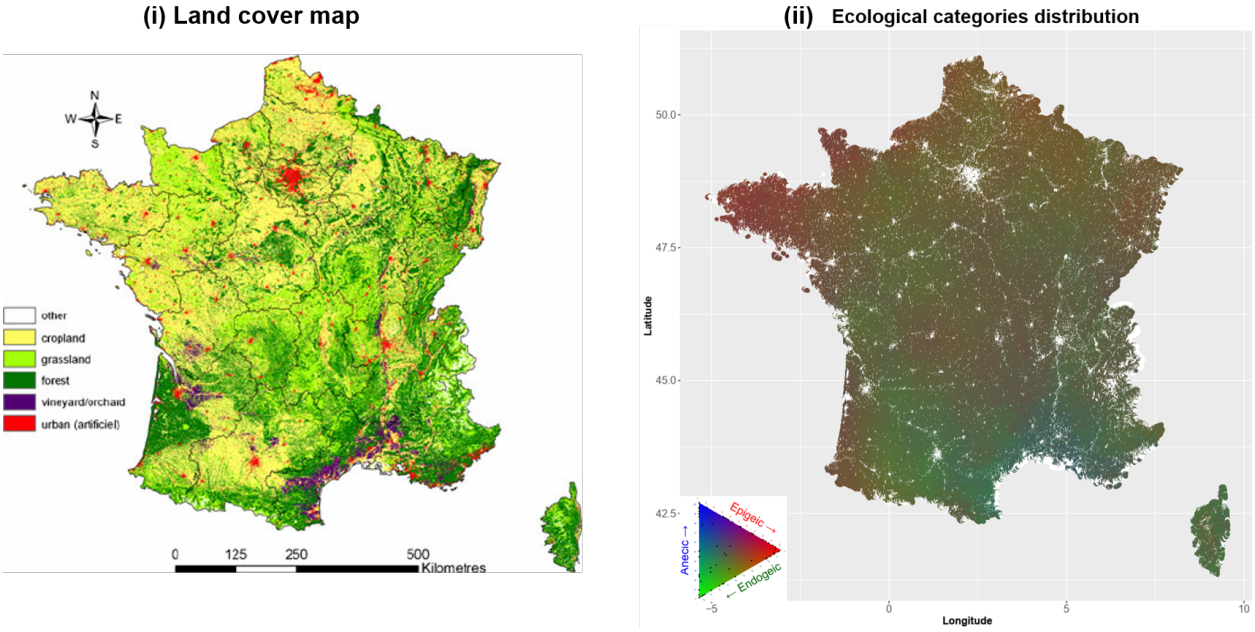


Figure 6.11: (i) The simplified land cover map of France [Meersmans et al., 2012] for ease of interpretation. (ii) Geographic distribution of ecological categories relative proportions, on the right. In each pixel, the local community composition is predicted using a joint distribution model. The ecological categories assignment of taxa predicted as present are aggregated to obtain the community-level relative proportions of the epigeic (reddish), endogeic (greenish) and anecic (bluish) classes respectively. The white pixels correspond to unsuitable land uses for the presence of any earthworm assemblages such as urban surfaces and water bodies.

meadows of northern France. Also, to a lower extent, they represent a high proportion of communities in low mountain such as the Massif Central, the Vosges and the Ardennes, mostly covered by forests. Elsewhere in the meadows and agricultural landscapes, we observed either an endogeic dominance or even a balanced tendency between the three groups.

Distribution of earthworm diversity

Our results showed that the overall variability in taxonomic diversity is much higher than functional and phylogenetic diversity, with highest values $>$ 3 fold vs 20-1.1 fold the lowest values. It reflects the low variability in traits of earthworms as morphology and anatomy is highly constrained by living in soil-related habitats.

Taxonomic diversity Taxonomic diversity maps (Fig 6.13) showed a decrease in the size of regional pools (gamma diversity, over 100 km radius) from South to North accompanied by an increase in community richness (alpha diversity, predicted over 1km pixels) from south to north with the exception of the Ile de France region. This result is consistent with the results described by Mathieu and Jonathan Davies [2014] on the same dataset.

Gamma diversity was maximum in the southeast, in the region bounded by the Massif Central in the North and the "Golfe du Lion" in the south. This area, that globally intersect a former glacial refuge (Fig 6.12), constitutes a corridor devoid of topographic barriers linking Ligurian biodiversity to the East and Catalonia to the West. At the level of this hotspot of earthworm diversity, the communities were restricted in terms of alpha richness with an average of 2 to 3 taxa, inducing a high turnover between sites.

To the West of this hotspot, the second former glacial refuge located in Aquitaine region depicted a different pattern. We did not observe the high regional richness as in the "Golfe de Lion". It can be assumed that

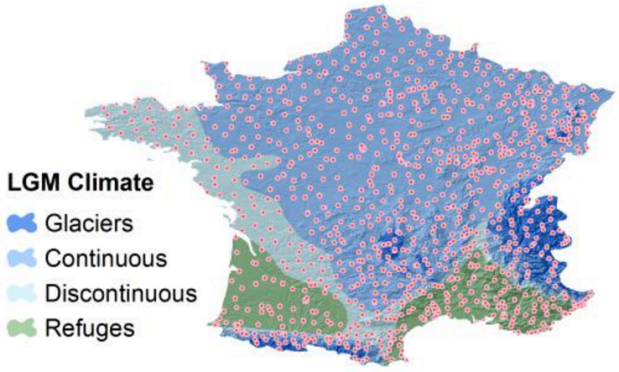


Figure 6.12: Sample sites of earthworms in France and permafrost limits during the Last Glacial Maximum (LGM). Ice, areas continuously frozen and under ice during the LGM; continuous, areas where the soil was continuously frozen during the LGM; discontinuous, areas where the soil was periodically frozen during the LGM; and refugia, areas that were not glaciated during the LGM. Obtained from [Mathieu and Jonathan Davies, 2014]

the barrier of the Atlantic Ocean to the West, the Pyrenees mountains to the South and the Garonne River to the East played against dispersal from the western refuge. This hypothesis was supported by the distinct regional pools with endemic species in Aquitaine, including endogenous species of the genus *Prosellodrilus*. However, high alpha richness in Aquitaine can be attributed to the more favourable climatic conditions in this region than in the southeast. On the other hand, northern communities were more homogeneous in species composition (low turnover, high richness). These locations were recolonized at the end of the last glaciations by peregrine species with high dispersal capacities. Finally, in Corsica, we reported a relatively high total pool size (gamma) but a rather homogeneous richness per community.

Functional and phylogenetic diversity In the southeastern hotspot identified previously, functional richness matched the taxonomic patterns as communities were made of a small number of similar taxa despite a relatively diverse regional pool. This indicates that climate acts as a strong environmental filter, harsher conditions (dryer and warmer climate) selected a restricted combination of traits. The high functional turnover reflected a potential habitat heterogeneity (that can be confirmed from the land cover maps Fig 6.11) and microclimate conditions (North vs South hill slope for instance). It may also relate to the presence of exotic*Microscolex* species in Batz Island that increased the diversity of traits. The high taxonomic turnover observed previously in light of the high functional turnover rules out competition in these communities.

Unlike its uniform taxonomic diversity, the patterns indicated that Corsica hosts a functionally diverse pool of earthworms especially in the southern part of the island. At the Atlantic edge of Aquitaine and in Brittany, low functional diversity level was associated with high functional turnover, probably in relation with small-scale heterogeneity in habitat or in local abiotic filtering (e.g land cover or topological features).

As for taxonomic and functional diversity, alpha phylogenetic diversity is low in South-East France. We assume that beta diversity patterns relies on to the distribution of two genus: *Microscolex* and *Haplotaxis* that are phylogenetically distant from others. The local phylogenetic diversity reached its maximal values around the "Côte d'Azur" and at the end of Brittany, two areas colonized by the exotic *Microscolex* species. This genus belongs to the *Acanthodrilidae* which is an ancient family, a lineage far away from French species that are mostly dominated by *Lumbricidae* and *Hormogastridae*. Additionnaly, it is now recognized that several clades exist within peregrine *Microscolex* species that should modify a bit the observed pattern [Rota et al., 2018]. In addition, high phylogenetic values in Massif Central, Western Pyrénées and North meet the distribution of the riparian *Haplotaxis gordioides* (Hartmann, 1821). The species is the furthest clade (cf. phylogenetic tree Fig 6.2) from the rest of the earthworm families in France.

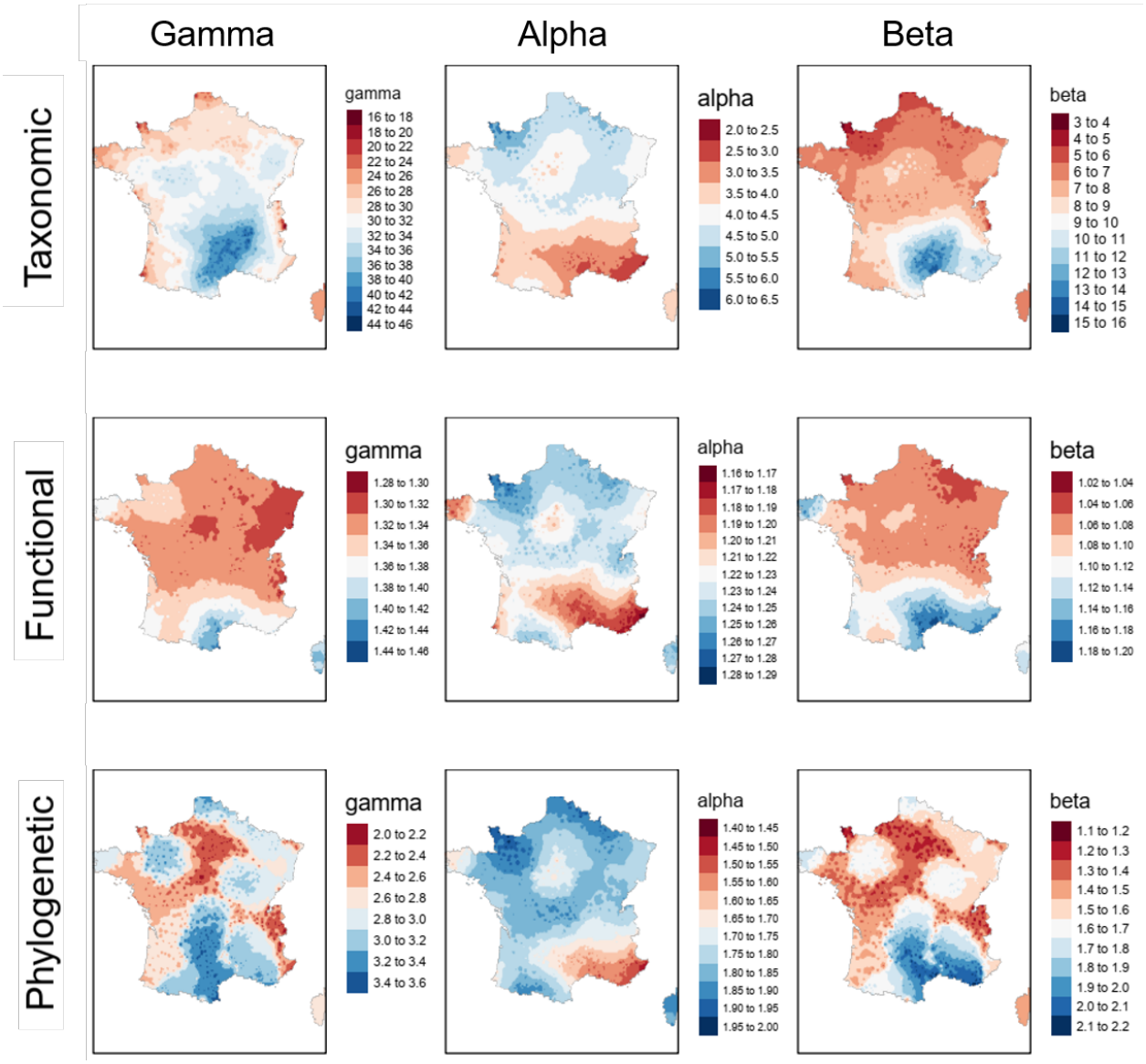


Figure 6.13: Mapping of taxonomic, functional and phylogenetic diversity. The indices are computed on presence/absence predictions on the calibration dataset, using the Leinster and Cobbold [2012] framework with q=0. Alpha diversity is computed on local communities (1km grid), gamma diversity is computed using a radius of 100km.

6.4 Conclusions and future directions

Understanding the drivers of earthworms distribution is a prerequisite to predict their trajectory in response to the ongoing changes in climate and land use, as well as to estimate the spatial emergence of their functional benefits as ecosystem engineers. To achieve that, scientists use species distribution models to establish statistical relationships between observed occurrences and the corresponding environmental conditions. So far, most studies used single species distribution models in experiments involving a limited number of widespread taxa or conversely aggregated indicators such as biomass or richness, irrespective of the species identities.

France hosts a diverse pool of lumbricids, itself part of a wide taxonomic group spanning several families. Studying each species separately using dedicated SDMs can be time-consuming and might be unfeasible for species with few available observations due to their intrinsic rarity or to insufficient sampling effort. Therefore, we aimed to build a multispecies model that identifies the environmental preferences of all species simultaneously, while being able to predict whole community composition. To satisfy this objective, we used a multi-task model such that each taxa's habitat suitability is ensured by a probit regression task applied to a set of features or environment representation shared amongst all tasks. Although we have raised a number of methodological obstacles, we discuss in this section the main conclusions, existing limitations and how our solution can be further improved in the future.

6.4.1 Improving the predictive performances

A few widespread taxa had relatively low performances for all fitted methods. This indicates that forecasting habitat suitability is harder for species with high adaptation capacities. Modeling relative abundances might help understand these species niches better than occurrence. In some cases, this might be an indicator of cryptic species [Marchán et al., 2020b] (e.g *L.terrestris* [James et al., 2010]), and calls for in-depth genetic studies [King et al., 2008, Shekhovtsov et al., 2019].

In our modeling experiment, we can also assume that discriminating environmental conditions were missing. As these species have a wide geographical extent, the climate gradient is not sufficient to discriminate their presence/absence zones. Indeed, they are also highly invasive in other parts of the world such as Northern America [Addison, 2009, Eisenhauer et al., 2014]. Thus, to improve their occurrence prediction performances, we propose the future inclusion of more fine-scale variables describing the micro-habitat. For instance, proximal density data (e.g soil electrical conductivity (EC), near infrared absorbance (NIR) [Schirrmann et al., 2016]), altitude as a proxy for soil depth and other topological features (e.g slope, topographic wetness index, solar radiation [Schneider et al., 2016]).

Many edaphic properties are themselves influenced by earthworm activity. Even if their inclusion in the model highlights correlations with the distribution of earthworms and improves predictive performances, it is not possible to identify the direction of causality without experimentation. Further experimental studies are needed to shed light on this question. On the other hand, process-based models are better equipped to address this problem [Schneider and Schröder, 2012] as they can model the dynamics of earthworm populations and soil properties simultaneously. Unfortunately, these models are difficult to adapt to a large number of species and require temporal data, which limits their applicability.

In addition to the abiotic variables, a description of the surrounding biotic context would allow the study of associations between earthworms and other groups of soil fauna (e.g microarthropods [Gutiérrez-López et al., 2010]) and microflora [Zirbes et al., 2012, Gong et al., 2019], which could further explain the micro-habitat preferences of these worms. Beyond the land use class, a more detailed description of the type of vegetation or plant traits from databases TRY [Kattge et al., 2020] would make it possible to clarify the more favourable litter conditions for epigeics, epi-endogeics and epi-anecics. We reported preliminary results on this relationship in Appendix 6.5.4. In any case, the vegetation landscape is a typical information in field data, while modern techniques based on satellite imagery [Rogan and Chen, 2004] can be used to obtain this information for non-sampled areas.

6.4.2 Extending the association model

The organization of the soil layers allows the niche separation and the cohabitation of several species, creating opportunities for a mutual influence via interactions mediated by the availability of resources and the feedback on the soil structure. These interactions can attract or repel species, thus affecting the observed composition of communities and producing positive or negative spatial associations.

A peculiarity of our approach compared to previous studies on earthworms is the estimation of associations between taxa that are not attributable to the environment. Our results showed that these associations, even if weak in comparison to abiotic effects, are still significant. Moreover, the resulting network is structured around a hub of peregrine species with epigeic tendency associated with peripheral nodes from different ecological categories.

Existing studies of spatial associations have been carried out in various land uses: forest, pasture, agricultural land. The variability of the reported results is indicative of the context-dependence of the associations. Thus, an interesting question to investigate further would be to estimate a different network per land use.

As biotic interactions between earthworms are mainly indirect, an interesting extension of the method would be to estimate varying associations as a function of resources or soil characteristics. In contrary to the current method which produced a single association network, we would try to address questions typically asked in community ecology like: Does resource scarcity (e.g organic carbon content) or harsh abiotic conditions (e.g acid soil) lead to facilitation between earthworms? Does resource abundance favor competition between earthworms? (e.g field experiment [Winsome et al., 2006]) The topic can be extended to other soil fauna groups with common resources, such as micro-arthropods [Gutiérrez-López et al., 2010]. The ultimate goal would be to identify determinants of earthworms spatial associations.

6.4.3 Earthworm habitat suitability

What are the main drivers of earthworm habitat suitability?

The analysis of feature importance identified climate and land use as the main abiotic drivers of earthworm distribution. Seasonality was the most discriminating precipitation characteristic, whereas the effect of temperature was carried by short-term events (diurnal range, extreme events, isothermality) for above-ground litter-dwelling species, and long-term trends (annual range, seasonal average) for soil-dwelling species.

The activity of earthworms is strongly conditioned by soil moisture and sensitivity to variations in precipitation and temperature [Edwards and Bohlen, 1996]. Some species developed adaptations to drought. For instance, *Ap.caliginosa.caliginosa* survives drought by forming estivation chambers in the topsoil [McDaniel et al., 2013]. Anecics of the *Scherotheca* genus manage to survive in warm and dry environments by slowing down water losses thanks to their thick cuticle epidermis. Their gigantic size was posited as a hypothesis of adaptation to sclerophyllous vegetation typical of Mediterranean garrigues. The grinding power and traction to the galleries of these animals is considerable and well adapted to the crushing of tough litter. Their ability to burrow deep in the soil allows them to seek refuge in wet soil depths [Bouché, 1984].

Land cover was an important feature in our model as it allowed the distinction of characteristic assemblages per habitat type: riparian soils, forest, pasture, scrubland and grassland, etc. However, the land cover classes were not represented equivalently in the training dataset (e.g few agricultural sites) which could have resulted in biased results, particularly in the assignment of species to response groups.

Can phylogeny and ecological categories predict earthworms responses ?

The ranking of environmental factors varied amongst species and differed from the global importance ranking (Appendix **??**). The analysis of the modeled taxa's responses to various feature gradients also highlighted variability in abiotic preferences even within the same species (e.g *Ap.caliginosa caliginosa* and *meridionalis*) and within the same subspecies (e.g *Ap.chlorotica.chlorotica chlorotype* and *leucotype*). This highlights the need to model infra-specific groups (subspecies, varieties) and calls for in-depth phylogenetic analysis beyond the species level. In some cases, the infra-specific variability in response can be attributed to different ecological categories resulting from trait differences. For instance the subspecies of *Ap.caliginosa alternisetosus*,

caliginosa and *meridionalis* ; which are respectively endogeic, epi-endogeic and epi-anecic ; have different affinities to land cover, organic matter quality and soil acidity. However, **the ecological category alone was not sufficient to infer the abiotic preferences of species**.

The question of whether the response of earthworms to environmental conditions is similar within closely-related taxa cannot be addressed in this context due to the high variability of responses within species (deepest level of genetic analysis). But, by crossing the dependency plots with the phylogenetic tree we found some examples of phylogenetic conservatism within certain genera. For instance, in the genus *Dendrobaena*, species *D.octaedra* and *D.attemsi* had similar responses to all abiotic factors but differed from *D.byblica* and *D.pygmea*, although all of them are epigeics. It turned out that the former are very close phylogenetically and distant from the latter two. On the other hand, closely related *Lumbricus* species: *L.festivus*, *L.centralis* and *L.friendi* all preferred pastures over other habitats. However, they had different tolerances to soil acidity and preferences to organic matter quality. Moreover, their different tolerances to precipitation seasonality reflected in their geographic ranges: northern, central-meridional and southern respectively.

In summary, setting apart infra-specific variability, phylogenetic conservatism existed in inferred earthworms responses but it was only relevant in the land cover and to some extent edaphic properties responses, less so with regard to precipitation for instance.

Is there a congruence of response groups ?

Based on the partial responses of earthworms to the modeled abiotic factors, we identified several clusterings of earthworm taxa into response groups. Some groups were consistent across clusterings. For instance, forest epigeics (Fig 6.7) had roughly the same preferences in terms of organic matter quality (rich SOM, high C:N ratio) and pH (acid soils) (Fig 6.8). Similarly, riparian species were grouped together in response to land cover (inland wetlands) and edaphic properties (sandy soil, neutral pH). This may result from the fact that land use type limits the range of values for soil properties. In contrast, the precipitation response groups provided a complementary picture of species preferences that rather reflected the gradient of precipitation seasonality.

While the groups we identified can be combined to produce an overall classification of earthworm abiotic niches, we deemed it necessary to present these various facets as they have different implications for conservation, bio-indication and climate change studies.

6.4.4 Characterizing earthworm communities

Implications of ecological categories distribution

The relative densities of the ecological categories (epigeic, endogeic and anecic) are often used as bio-indicators of soil health and environmental disturbance [Kibblewhite et al., 2008, Pulleman et al., 2012] since they could reveal the availability of different niches. For instance, the absence of mulch or leaf litter at soil surface is deleterous for epigeic worms while compacted or ploughed soils preclude the presence of soil-dwelling endogeic and anecics. In addition, even if the ecological categories are not functional groups *senu stricto* [Bottinelli and Capowiez, 2020], their complementary effects on the physical and chemical structure of soil have been used as proxy for soil functioning [Schneider et al., 2016]. Any bio-indicators need a referential with a known range of values to be useful. Our results concerning the response of different categories to land use (Table 6.3, Fig 6.7) and the variability of the geographically dominant categories (Fig 6.11) showed that several reference frames need to be constructed per region (i.e. per topoclimatic zone), but also per land use, to fully exploit the potential of ecological categories as bio-indicators.

A more complete picture of earthworm diversity patterns

Several studies addressed the question of diversity distribution [Rutgers et al., 2016, Phillips et al., 2019] using similar abiotic factors as our study (climate, land use, soil parameters) resulting in various maps of total biomass/abundance and richness/diversity in the literature. However, the previous works used data from heterogeneous sources to model complex parameters. In our case, we used data sampled using a

uniform protocol and identified by a single taxonomist, thus preventing biases due to conflicting taxonomic identifications which can occur when combining multiple sources.

Interpreting geographic patterns of diversity can be difficult and arbitrary when modeled directly. Linking the geographic distribution of abiotic factors to diversity is more natural by knowing the underlying species preferences. Our analysis focused on three related aspects of diversity: taxonomic, functional and phylogenetic and was not limited to the first aspect, contrary to existing studies on diversity patterns in France.

Our study highlighted areas of congruence between the different diversity facets, in link with latitudinal and elevation gradients and with last glaciation refuges particularly the Mediterranean basin. The latter is known to be an exceptional ecoregion in which the paleogeography and historical land-use have created a complex mosaic of communities, which is reflected in the diversity patterns we identified.

Our results highlighted hotspots of earthworm diversity. Areas like the Meditteranean basin and Aquitaine region, locations of two previous glaciation refuges, but also Brittany and Corsica are home to phylogenetically and functionally unique groups. Our study paves the way for further exploration of the functional rarity [Violle et al., 2017] of the assemblages in these regions.

6.4.5 Future directions

A natural continuation of our study would be to predict the response of earthworms to climate change scenarios. Climate models project an increase in the frequency of extreme precipitation events, whereas the total amount of rainfall is predicted to remain comparable to its present state but with increasing variability [Singh et al., 2019]. Given the sensitivity of earthworms to precipitation seasonality, these changes will most likely affect earthworms distributions. Increasing temperature, drought, and winter rainfall will affect soil moisture and temperature regimes, which have varying impacts on earthworms. Moreover, the combination of high temperatures, that increase metabolic demands, and the higher frequency of extreme precipitation events (drought, flood) may change the life cycles of earthworms, leading to a rise in mortality. Extreme rainfall and land use change can make soils more vulnerable to erosion, thus destroying the habitat of earthworms [Singh et al., 2019]. Climate change via temperature and precipitation may also impact earthworms indirectly by altering soil properties and the vegetation landscape [Eisenhauer et al., 2009]. In addition to climate conditions, we identified land use as a crucial predictor in our model. Thus, scenarios of land use changes are needed in order to predict future trajectories of earthworm distributions. Finally, the ability to account for species dispersal abilities through spatially-explicit modeling of dispersal corridors and barriers, will be crucial for a more accurate forecasting of earthworm distributions.

6.5 Supplementary Materials

6.5.1 Dataset

Taxonomic and geographic coverage

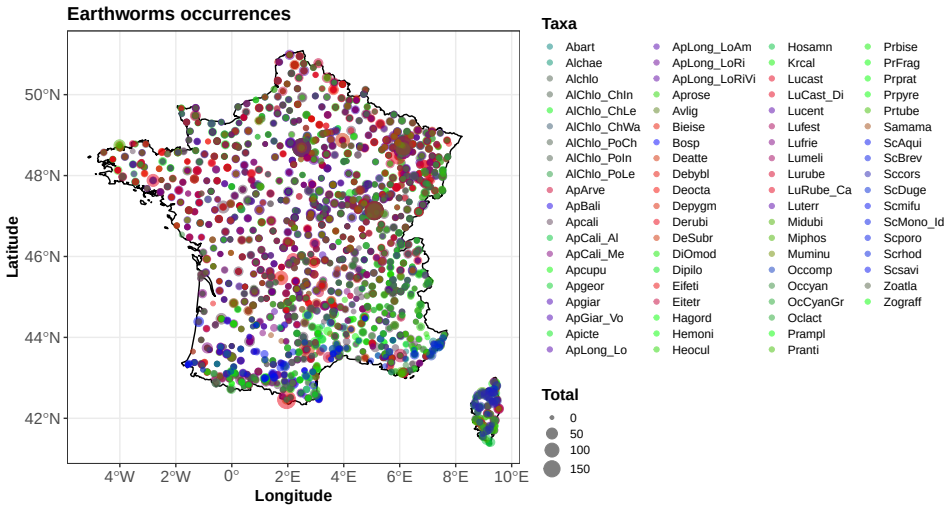


Figure 6.14: Geographic and taxonomic coverage of the occurrence dataset. Colors are derived from the functional groups assignment of each species: epigeic (red), endogeic (green), anecic (blue).

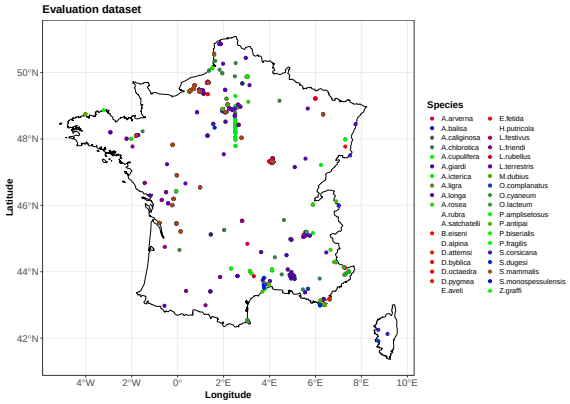


Figure 6.15: Geographic and taxonomic (species) coverage of the evaluation dataset. Colors are derived from the functional groups assignment of each species: epigeic (red), endogeic (green), anecic (blue).

Full description of environmental features

Environmental features distribution and interpretation guide

Feature group	Feature		Dataset	
	Feature (unit)	Interpretation	calibration, N = 1,351	evaluation, N = 1,703
Temperature	**Annual Mean Temperature (x10°C)**	The annual mean temperature approximates the total energy inputs for an ecosystem.	108 ± 21	113 ± 24
	Mean Diurnal Range (*10°C)	This index can help provide information pertaining to the relevance of temperature fluctuation for different species.	89 ± 11	171 ± 21
	Isothermality (%)	Isothermality is generally useful for tropical, insular, and maritime environments (Nix,1986). Isothermality quantifies how large the day-tonight temperatures oscillate relative to the summerto-winter (annual) oscillations. • An isothermal value of 100 indicates the diurnal temperature range is equivalent to the annual temperature range, • Anything less than 100 indicates a smaller level of temperature variability within an average month relative to the year. A species distribution may be influenced by larger or smaller temperature fluctuations within a month relative to the year and this predictor is useful for ascertaining such information.	35.9 ± 2.9	50.5 ± 2.1
	Temperature Seasonality (*10°C)	Temperature seasonality is a measure of temperature change over the course of the year.	5,611 ± 515	5,349 ± 617
	Max Temperature of Warmest Month (*10°C)	This information is useful when examining whether species distributions are affected by warm temperature anomalies throughout the year.	243 ± 23	289 ± 26
	Min Temperature of Coldest Month (*10°C)	This information is useful when examining whether species distributions are affected by cold temperature anomalies throughout the year.	-2 ± 24	-49 ± 31
	Temperature Annual Range (*10°C)	This information is useful when examining whether species distributions are affected by ranges of extreme temperature conditions.	244 ± 23	337 ± 32
	Mean Temperature of Wettest Quarter (*10°C)	This index provides mean temperatures during the (wettest, driest, warmest, coldest) three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions.	105 ± 50	81 ± 39
	Mean Temperature of Driest Quarter (*10°C)		114 ± 69	117 ± 74
	Mean Temperature of Warmest Quarter (*10°C)		180 ± 20	190 ± 26
	Mean Temperature of Coldest Quarter (*10°C)		36 ± 24	44 ± 25

Feature group	Feature (unit)	Feature Interpretation	Dataset calibration, N = 1,351	Dataset evaluation, N = 1,703
Precipitation	**Annual Precipitation (mm)**	Annual total precipitation approximates the total water inputs and is therefore useful when ascertaining the importance of water availability to a species distribution.	808 ± 152	821 ± 148
	Precipitation of Wettest Month (mm)	The wettest and driest month precipitation are useful if extreme precipitation conditions during the year influence a species potential range.	89 ± 18	97 ± 19
	Precipitation of Driest Month (mm)		46 ± 16	48 ± 15
	Precipitation Seasonality (%)	Since species distributions can be strongly influenced by variability in precipitation, this index provides a percentage of precipitation variability where larger percentages represent greater variability of precipitation.	18 ± 9	22 ± 10
	Precipitation of Wettest Quarter (mm)	This index provides total precipitation during the (wettest, driest, warmest, coldest) three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions.	243 ± 49	274 ± 53
	Precipitation of Driest Quarter (mm)		161 ± 44	154 ± 43
	Precipitation of Warmest Quarter (mm)		186 ± 50	167 ± 48
	Precipitation of Coldest Quarter (mm)		209 ± 52	199 ± 37
Physical and chemical properties	**pH_H2O**	Acidity level (x10) <7: acidic, 7: neutral, >7: basic	63 ± 11	65 ± 8
	Carbon (mg/Kg soil)	Quantity of organic carbon per kg of soil	4,266 ± 4,509	3,223 ± 1,964
	Carbon-to-nitrogen ratio	Ratio of organic carbon quantity (mg/Kg) to nitrogen quantity (mg/Kg)	12.15 ± 3.44	11.67 ± 2.13
Hydrological properties	**Proximity to water courses**	Whether there is a water course at <1km	0.09 ± 0.29	0.05 ± 0.22
	Available water capacity topsoil	Available water capacity is the maximum amount of plant available water a soil can provide. It is an indicator of a soil's ability to retain water and make it available for plant use.	165 ± 35	150 ± 60
	Base saturation topsoil	Base saturation is calculated as the percentage of CEC occupied by base cations.	69 ± 22	68 ± 22
	Cation exchange capacity topsoil	A soil's total quantity of negative surface charges. It is the potential of available nutrient supply	11.3 ± 6.8	12.5 ± 7.2
	Water regime (categorical)	Dominant annual average soil water regime class of the soil profile. https://esdac.jrc.ec.europa.eu/ESDB_Archive/sgdbe/wra3.pdf		
	NA	Not applicable (non-soil)	1.5%	11%
	D3D1	Not wet within 80 cm for over 3 months, not wet within 40 m for over a month	81%	82%
	W3-6D1	Wet within 80 cm for 3 to 6 months, but not wet within 40 cm for over 1 month	7.8%	2.9%
	W6D11	Wet within 80 cm for over 6 months, but not wet within 40 cm for over 11 months	9.0%	4.0%
	W11	Wet within 40 cm depth for over 11 months	0.2%	0%

Feature group	Feature		Dataset	
	Feature (unit)	Interpretation	calibration, N = 1,351	evaluation, N = 1,703
Structural properties	**Depth to gleyed horizon (cm)**	The gleyed horizon is characterized by its asphyxiating and reducing conditions where iron in its divalent (ferrous) state gives it a greenish color. https://esdac.jrc.ec.europa.eu/ESDB_Archive/ptrdb/dgha3.pdf Shallow: <40cm ; Moderate:40-80cm ; Deep: 80-120cm ; Very deep:>120cm	129 ± 29	130 ± 29
	Clay (%)	Soil texture is described as the % of clay, %silt, %sand	22 ± 27	19 ± 4
	Silt (%)	%clay+%silt+%sand=100	34 ± 32	43 ± 10
	Packing density of topsoil (g/cm^3)	Refers to the density of solid particles. It is computed using pedo-transfer rules based on texture and soil structure class. https://esdac.jrc.ec.europa.eu/Projects/SINFO/pdf/annex4_7.pdf	1.40 ± 0.49	1.42 ± 0.49
	Crusting (categorical)	Soil crusting refers to the forming processes and the consequences of a thin layer at the soil surface with reduced porosity and high penetration resistance. This phenomenon causes water runoff, causing erosion and inhibiting seedling emergence. https://esdac.jrc.ec.europa.eu/ESDB_Archive/ptrdb/crustinga3.pdf		
	NS	Non-significant	0.1%	0%
	VW	Very weak	1.0%	1.4%
	W	Weak	21%	17%
	M	Medium	40%	53%
	S	Strong	22%	10%
	VS	Very strong	17%	19%
	Unknown	Default class if crusting is not informed on this pixel	/	/
	Erodibility (categorical)	Susceptibility of the soil to be eroded. https://esdac.jrc.ec.europa.eu/ESDB_Archive/ptrdb/erodibilitya3.pdf		
	NS	Non-significant	1.6%	11%
	VW	Very weak	1.0%	1.4%
	W	Weak	25%	15%
	M	Medium	27%	26%
	S	Strong	32%	35%
	VS	Very strong	13%	12%
	Unknown	Default class if crusting is not informed on this pixel	/	/
	Dominant parent material (categorical)	Parent material at the origin of soil formation https://esdac.jrc.ec.europa.eu/ESDB_Archive/sgdbe/par-mat-dom1a3.pdf		
	alluvWeathSlope	Unconsolidated deposits (alluvium, weathering residuum and slope deposits)	13%	14%
	anthropo	Anthropogenic deposits	0.1%	0%
	clastic-sedim	Consolidated clastic sedimentary rocks	1.3%	11%

Feature group	Feature (unit)	Feature Interpretation	Dataset calibration, N = 1,351	Dataset evaluation, N = 1,703
	eolian	Eolian deposits	0.4%	<0.1%
	glacial	Unconsolidated glacial deposits/glacial drift	20%	29%
	igneous	Igneous rocks	29%	20%
	metamorpho	Metamorphic rocks	16%	7.2%
	organic	Organic materials	7.6%	15%
	sediment	Sedimentary rocks (chemically precipitated, evaporated, or organogenic or biogenic)	13%	4.5%
Land use	**Land cover (Corine Land Cover nomenclature)**	Land cover is the physical material at the surface of the earth. Land covers include grass, asphalt, trees, bare ground, water, etc.		
	1.1 Urban	Urban fabric	0%	14%
	1.2 Transport	Industrial, commercial and transport units: roads, ports, airports, etc.	0%	1.2%
	1.3 MineDump	Mine, dump, and construction sites	0.1%	0.4%
	1.4 ArtifVeg	Artificial, non-agricultural vegetated areas: leisure facilities, green urban sites	1.0%	1.6%
	2.1 Arable	Arable lands: irrigated, non-irrigated and rice-fields	2.2%	26%
	2.2 PermaCrop	Permanent crops: vineyards, fruit trees, olive groves, etc.	3.1%	4.2%
	2.3 Pasture	Pastures: land covered with grass and other low plants suitable for grazing animals.	42%	6.3%
	2.4 HeterAgri	Heterogeneous agricultural areas: associations of permanent/annual crops, agroforestry, etc.	0%	13%
	3.1 Forest	Forests: including broad-leaved, coniferous, or mixed	24%	25%
	3.2 ScrubHerbVeg	Shrub and/or herbaceous vegetation associations: grassland, moors, scrubland, woodland.	20%	5.9%
	3.3 OpenSpacNVeg	Open spaces with little to no vegetation: bare rock, beach, burnt areas, glaciers	1.6%	1.0%
	4.1 InlandWet	Inland wetlands: marshes, peat bogs	5.3%	<0.1%
	4.2 CoastalWet	Costal wetlands: salt marshes, salines, intertidal flats	0.3%	0.6%
	5.1 ContinWater	Continental water: water courses, water bodies	0%	0.3%

[1]Statistics presented: Mean ± SD ; %

6.5.2 Model and preprocessing selection

The best architecture selected in terms of cross-validated performances was E2E-Shared16 which contained a single shared layer of depth 16 and a specific output layer per taxa, with a rectified linear unit (ReLU) hidden activation and a probit link (output activation). The best recall scores on the evaluation dataset was obtained for a value $\lambda = (10^{-4}, 10^{-4})$ of the elastic net regularization parameter. Amongst feature preprocessing strategies, the end-to-end approach with a shared feature extraction layer outperformed the others in terms of recall on the evaluation dataset, while scores on the training dataset were comparable (Table 6.4). During training, we found that loss weighting strategies benefited rare taxa especially in terms of TSS (the sensitivity component of the TSS reflected the ability of the model to detect the presence of rare taxa instead of systematically predicting their absence, as is the case when weighting is not used, resulting in a TSS of zero or less).

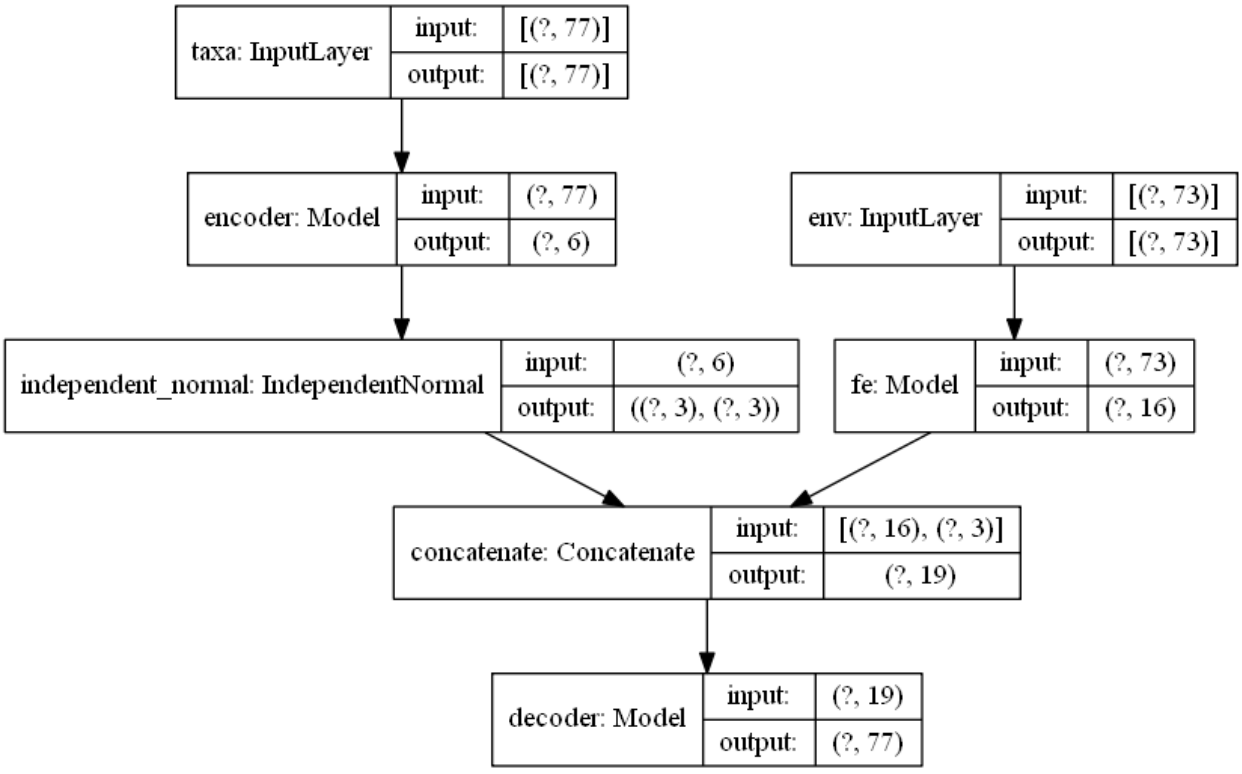


Figure 6.16: Selected MTEC architecture for the prediction of the joint distribution of earthworms.

Configuration			Performances of best model (evaluation)		Retained optimal hyperparameters	
Feature preprocessing	**Feature learning architecture**	**Model name**	**Calibration ROC AUC**	**Evaluation Recall**	**Loss weighting**	**Regularization** ($\lambda \cdot 10^{-4}$)
None	[16]	**E2E-Shared16**	**0.89** ± 0.106	**0.765**	WBCE Beta=3	Lambda=(1,1)
	[32-16]	E2E-Shared32-16	0.861 ± 0.02	0.604	WBCE Beta=3	Lambda=(5,1)
PCA	[16]	**PCA-Shared16**	0.886 ± 0.02	**0.679**	WBCE Beta=2	Lambda=(2,1)
	[32-16]	PCA-Shared32-16	0.864 ± 0.021	0.574	WBCE Beta=3	Lambda=(10,1)
VIF	[16]	VIF-Shared16	0.864 ± 0.022	0.519	WBCE Beta=1	Lambda=(10,1)
	[32-16]	**VIF-Shared32-16**	0.843 ± 0.02	**0.65**	WBCE Beta=3	Lambda=(10,1)

Table 6.4: Grid search cross-validation scores and model selection. Feature learning architecture is represented by an array whose size gives the depth of the network and elements represent the width of its layers (from the input to the output). We highlight the retained model (colored row) and the best performances for each feature preprocessing strategy.

6.5.3 Full predictive performances

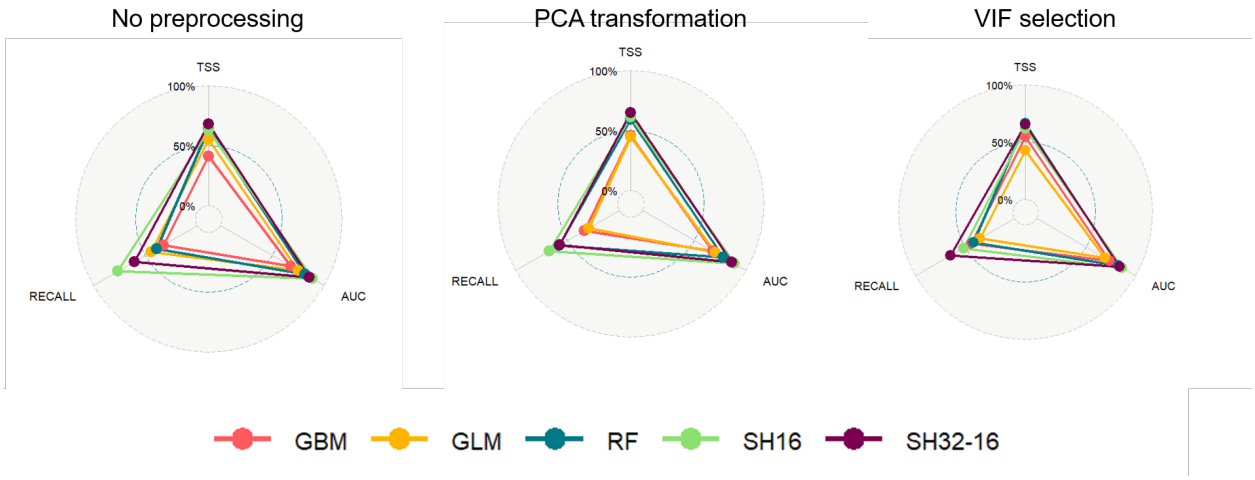


Figure 6.17: Average performances over taxa (AUC, TSS on training set, Recall on evaluation set) per preprocessing strategy and learning architecture.

Target	Prevalence	threshold	True Skill Statistic				ROC-AUC				Evaluation recall			
			MTEC	GLM	GBM	RF	MTEC	GLM	GBM	RF	MTEC	GLM	GBM	RF
Average	0.05243048	0.65012987	**0.746**	0.411	0.552	0.646	**0.912**	0.681	0.753	0.816	**0.818**	0.519	0.302	0.250
Predictive performances for each taxa														
Abart	0.010	0.88	**0.947**	0.326	0.853	0.869	**0.991**	0.663	0.913	0.914	**1**			
Alchae	0.007	0.88	0.95	0.331	**0.958**	0.953	**0.996**	0.663	0.969	0.968	**1**	**1.000**		
Alchlo	0.150	0.46	0.437	**0.475**	0.430	0.405	**0.782**	0.753	0.769	0.761				
AlChlo_ChIn	0.055	0.58	0.333	0.193	0.222	0.227	**0.707**	0.570	0.587	0.562	**1**	0.165		
AlChlo_ChLe	0.202	0.48	0.263	0.334	**0.352**	0.300	0.687	0.692	**0.695**	0.676	**1**			
AlChlo_ChWa	0.013	0.53	**0.859**	0.309	0.558	0.562	**0.95**	0.599	0.761	0.739				
AlChlo_PoCh	0.005	0.69	**0.871**	0.000	0.000	0.629	**0.97**	0.395	0.000	0.718				
AlChlo_PoIn	0.006	0.74	**0.868**	0.330	0.493	0.537	**0.969**	0.626	0.702	0.708				
AlChlo_PoLe	0.029	0.46	**0.612**	0.536	0.508	0.447	**0.883**	0.762	0.814	0.759				
ApArve	0.005	0.86	0.973	0.663	**0.990**	**0.990**	**0.999**	0.830	0.990	0.990	**1**	0.333		
ApBali	0.004	0.89	**0.975**	0.332	0.000	0.947	**0.998**	0.663	0.000	0.952	**1**			
Apcali	0.613	0.6	0.223	0.325	**0.401**	0.383	0.712	0.687	**0.737**	0.736	0.684	0.248	**0.774**	0.607
ApCali_Al	0.016	0.48	0.815	0.736	0.727	**0.882**	0.949	0.837	0.894	**0.954**				
ApCali_Me	0.155	0.41	0.489	0.472	**0.567**	0.556	0.818	0.776	**0.833**	0.817	**1**			
Apcupu	0.045	0.47	0.41	0.438	**0.532**	0.522	0.791	0.754	**0.805**	0.799	**1**	0.067		
Apgeor	0.006	0.84	**0.935**	0.164	0.784	0.652	**0.994**	0.579	0.879	0.804				
Apgiar	0.121	0.41	0.505	0.529	0.534	**0.565**	0.828	**0.831**	0.818	0.825	0.946	**0.973**		
ApGiar_Vo	0.008	0.81	**0.93**	0.592	0.703	0.606	**0.985**	0.783	0.842	0.765				
Apicte	0.136	0.51	0.329	0.375	**0.564**	**0.564**	0.707	0.691	**0.847**	0.846	**0.239**	0.216	0.114	
ApLong_Lo	0.130	0.53	**0.46**	0.434	0.444	0.425	**0.804**	0.759	0.777	0.768				
ApLong_LoAm	0.004	0.89	**0.93**	0.000	0.000	0.189	**0.994**	0.492	0.000	0.438				
ApLong_LoRi	0.036	0.47	0.543	**0.545**	0.521	0.519	**0.834**	0.787	0.760	0.740				
ApLong_LoRiVi	0.008	0.64	**0.915**	0.790	0.359	0.648	**0.982**	0.892	0.657	0.833				
Aprose	0.375	0.5	0.326	0.305	**0.340**	0.279	**0.704**	0.679	0.701	0.669	0.000		**0.014**	**0.129**
Avlig	0.010	0.77	**0.913**	0.617	0.683	0.827	**0.984**	0.793	0.866	0.937	**1**			
Bieise	0.017	0.7	**0.928**	0.766	0.639	0.680	**0.983**	0.897	0.859	0.893				
Bosp	0.008	0.85	0.962	0.494	0.941	**0.985**	**0.996**	0.747	0.977	0.992				
Deatte	0.020	0.65	0.878	0.361	0.822	**0.902**	**0.979**	0.678	0.960	0.974	**1**	**1.000**	0.375	
Debybl	0.004	0.84	**0.917**	0.000	0.000	0.536	**0.986**	0.000	0.472	0.695	**1**			
Deocta	0.051	0.46	**0.72**	0.514	0.591	0.594	**0.926**	0.791	0.870	0.865	**1**	0.750	0.200	
Depygm	0.005	0.89	**0.957**	0.659	0.000	0.915	**0.994**	0.830	0.000	0.925				
Derubi	0.048	0.42	**0.633**	0.443	0.464	0.466	**0.897**	0.736	0.795	0.767	**0.476**			
DeSubr	0.035	0.51	**0.522**	0.269	0.420	0.409	**0.827**	0.644	0.686	0.687	**1**			
DiOmod	0.015	0.86	**0.971**	0.492	0.792	0.834	**0.991**	0.740	0.897	0.931				
Dipilo	0.005	0.88	0.939	0.323	0.657	**0.976**	**0.987**	0.609	0.819	0.978	**1**			
Eifeti	0.007	0.68	**0.913**	0.000	0.000	0.073	**0.986**	0.496	0.412	0.339	**1**			
Eitetr	0.066	0.49	0.486	0.468	0.466	**0.489**	**0.823**	0.746	0.798	0.765	**0.758**			
Hagord	0.011	0.52	**0.755**	0.396	0.348	0.384	**0.922**	0.584	0.638	0.625				
Hemoni	0.007	0.84	**0.93**	0.327	0.844	0.899	**0.995**	0.663	0.958	0.955				
Heocul	0.010	0.51	**0.878**	0.548	0.477	0.401	**0.961**	0.715	0.725	0.627				
Hosamn	0.004	0.88	0.97	0.661	**1.000**	**1.000**	0.997	0.829	**1.000**	**1.000**	**1**			
Krcal	0.008	0.65	**0.919**	0.000	0.549	0.717	**0.977**	0.000	0.758	0.831	**1**			
Lucast	0.064	0.51	0.378	0.451	**0.571**	0.541	0.742	0.775	0.827	**0.829**	0.019	**0.848**		
LuCast_Di	0.263	0.51	0.242	0.255	**0.341**	0.318	0.663	0.627	**0.690**	0.681				
Lucent	0.051	0.45	**0.542**	0.433	0.526	0.487	**0.841**	0.743	0.802	0.789	**0.625**			
Lufest	0.049	0.44	**0.706**	0.627	0.623	0.651	**0.903**	0.861	0.871	0.880				
Lufrie	0.127	0.45	0.55	0.650	**0.725**	0.711	0.861	0.877	0.889	**0.893**	**0.957**	0.130		
Lumeli	0.004	0.87	**0.965**	0.000	0.609	0.659	**0.992**	0.494	0.779	0.793	**0.8**			
Lurube	0.113	0.48	0.543	0.578	**0.606**	0.614	0.82	0.818	0.854	**0.863**				
LuRube_Ca	0.009	0.59	**0.691**	0.172	0.000	0.000	**0.891**	0.396	0.491	0.288	**1**			
Luterr	0.216	0.49	0.35	**0.361**	0.348	0.360	0.705	0.687	**0.713**	0.693	**0.302**	0.054		
Midubi	0.014	0.59	**0.899**	0.381	0.594	0.772	**0.97**	0.656	0.792	0.876	**0.154**	0.077		
Miphos	0.016	0.52	**0.918**	0.594	0.779	0.796	**0.972**	0.743	0.874	0.867				
Muminu	0.040	0.56	**0.481**	0.301	0.411	0.344	**0.794**	0.607	0.682	0.662	**0.9**			
Occomp	0.014	0.58	**0.904**	0.659	0.781	0.887	**0.971**	0.837	0.916	0.930	**0.5**			
Occyan	0.223	0.56	0.252	0.308	**0.357**	0.258	0.685	0.670	**0.699**	0.671	**0.925**	0.388		0.015
OcCyanGr	0.025	0.44	**0.621**	0.488	0.421	0.439	**0.891**	0.777	0.712	0.712				
Oclact	0.045	0.54	0.494	0.521	0.468	**0.578**	0.814	0.758	0.814	**0.822**				
Prampl	0.008	0.61	**0.897**	0.308	0.544	0.671	**0.978**	0.628	0.738	0.775				
Pranti	0.031	0.4	0.48	0.603	0.722	**0.723**	0.867	0.789	0.886	**0.918**	**1**			
Prbise	0.004	0.84	0.937	0.000	**0.996**	0.797	0.987	0.494	**0.996**	0.859	**1**			
PrFrag	0.009	0.74	0.926	**0.962**	0.707	0.653	**0.986**	0.975	0.830	0.745	**1**	**1.000**		
Prprat	0.004	0.89	**0.97**	0.333	0.333	0.926	**0.994**	0.662	0.610	0.935				
Prpyre	0.010	0.69	**0.914**	0.482	0.584	0.844	**0.982**	0.772	0.844	0.937				
Prtube	0.006	0.76	**0.925**	0.490	0.549	0.611	**0.988**	0.729	0.756	0.788				
Samama	0.107	0.37	0.392	0.404	**0.461**	0.429	0.783	0.712	**0.791**	0.760	**0.909**			
ScAqui	0.010	0.76	**0.905**	0.504	0.784	0.840	**0.982**	0.722	0.875	0.907	**1**			
ScBrev	0.004	0.87	0.98	0.331	0.997	**1.000**	**1**	0.662	0.997	**1.000**				
Sccors	0.027	0.67	0.968	0.973	0.963	**0.976**	0.991	0.988	**0.992**	**0.992**	0.000	**0.333**	**0.333**	
ScDuge	0.005	0.85	**0.989**	0.000	0.000	0.953	**0.999**	0.495	0.000	0.960				
Scmifu	0.004	0.87	0.953	0.000	0.994	**0.995**	**0.995**	0.498	0.994	**0.995**	**1**	**1.000**		
ScMono_Id	0.006	0.87	0.958	0.799	0.969	**0.973**	**0.993**	0.888	0.976	0.980	**1**			
Scporo	0.004	0.88	**0.951**	0.000	0.621	0.589	**0.981**	0.000	0.782	0.729				
Scrhod	0.007	0.84	**0.932**	0.473	0.770	0.920	**0.981**	0.733	0.871	0.951	**1**	0.286		
Scsavi	0.031	0.45	**0.883**	0.803	0.809	0.863	**0.974**	0.929	0.945	0.963				
Zoatla	0.008	0.83	0.956	0.495	**0.993**	**0.993**	0.996	0.747	**0.997**	0.995				
Zograff	0.004	0.88	**0.922**	0.000	0.000	0.817	**0.987**	0.453	0.454	0.845	**1**	**1.000**		

6.5.4 Drivers of earthworm distribution

Taxa clustering

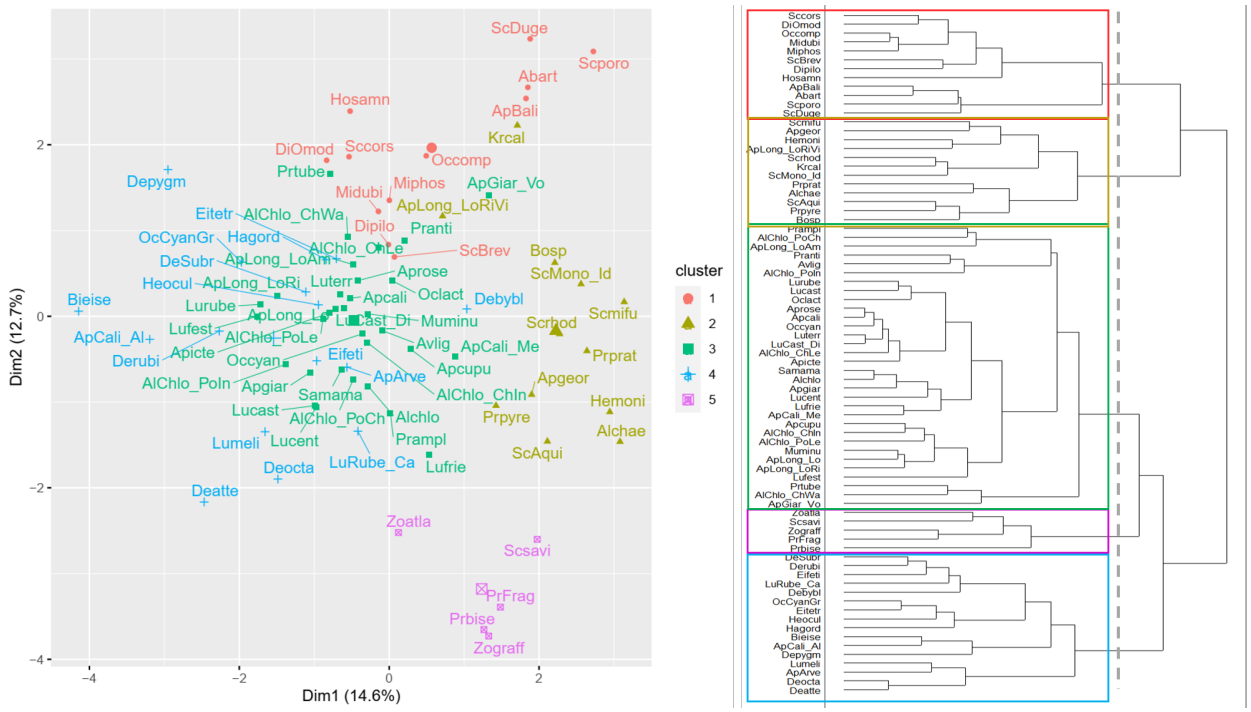


Figure 6.18: Abiotic response groups. On the left, a principal component analysis projection of the response embeddings (output regression parameters). On the right, the agglomerative hierarchical clustering tree of taxa's response embeddings. The optimal clusters (5) are annotated in colors.

Phylogenetic and functional structure of earthworm responses

We evaluate the variability of earthworm responses by calculating distances between their partial responses (SHAP values) to each environmental variable and by thematic group of variables. Then, we try to explain the differences in responses between taxa by their phylogenetic, functional (between their vectors of morpho-anatomical traits) or ecological (between their percentages of assignment to the 3 ecological categories) distances using a partial Mantel test (ecodist::mantel function on R). The test quantifies the partial correlation (direct effects) between the distances, so it takes into account the existing multicollinearity including the fact that phylogeny and functional traits, as well as ecological traits and categories are correlated. We report the correlation values and their p-values (with a significance threshold of 0.01).

We reported the correlations as well as their significance level (*:<0.01, **:<0.001) in Fig 6.19



Figure 6.19: Mantel test phylogenetic and functional structure of earthworms abiotic responses.

.

Response to tree traits

For a subset of forest sites (316 observations) in the training dataset, Bouché [1972] also described the surrounding vegetation in terms of trees, shrubs, mosses and ferns. We selected trees amongst the described plants. We queried the plant traits database TRY (Version 5, 2019-03-26). [Kattge et al., 2020] for a set of litter and leaf traits namely: the C:N ratio of tree leaves (1), of its litter (2), litter decomposition rate (3), and the litter nitrogen (4) and tannin content per dry mass (5). Their traits are reported in Table 6.5.

	Leaf C:N ratio	Litter C:N ratio	Litter decomposition rate	Litter Nitrogen content	Litter Tannin content
Abies sp.	12	1.75	4.2	1.4	n.a
Acer campestre	7	1	13	3	n.a
Acer pseudoplanatus	12	1	20	3	n.a
Alnus sp.	7	0.5	11	3	n.a
Betula pubescens	7	1	10	5	3
Buxus sempervirens	3	n.a	5	n.a	n.a
Carpinus betulus	5	2	16	3	n.a
Castanea sativa	23	2	10	1	n.a
Corylus avellana	2	1	12	n.a	n.a
Fagus sylvatica	70	5	20	6	n.a
Fraxinus excelsior	17	2	13	n.a	n.a
Ilex aquifolium	8	n.a	13	1	n.a
Larix decidua	9	8	11	1	n.a
Picea abies	73	2	20	7	2
Pinus nigra	263	n.a	6	1	n.a
Pinus silvestris	446	4	34	9	3
pinus.silvestris	446	4	34	9	3
Populus sp.	8	n.a	3	4	n.a
Prunus avium	13	2	12	1	n.a
Prunus spinosa	3	n.a	12	1	n.a
Quercus coccifera	3	1	2	1	n.a
Quercus ilex	398	1	13	2	n.a
Quercus pubescens	4	n.a	2	1	n.a
Quercus robur	392	1	24	3	n.a
Quercus sessilifolia	1	n.a	6	1	n.a
Quercus sp.	151	1	9	2	n.a
Quercus suber	108	1	4	3	n.a
Robinia pseudoacacia	7	1	10	2	n.a
Salix alba	3	n.a	10	n.a	n.a
Ulmus minor	9	n.a	3	1	n.a

Table 6.5: Functional traits of observed trees in forest sites.

We restricted our analysis to forest epigeic earthworms. For each of these taxa, we computed the mean traits of the trees with which it co-occurs. Then, we performed a Principal Component Anlysis on the traits (Fig 6.20).

The first component carried solely by the litter decomposition rate explained 88% of the variation in forest epigeics preferences. The remaining variation was attributed to leaf and litter C:N ratios.

We ordinated the tree species in a similar way in Fig 6.21.

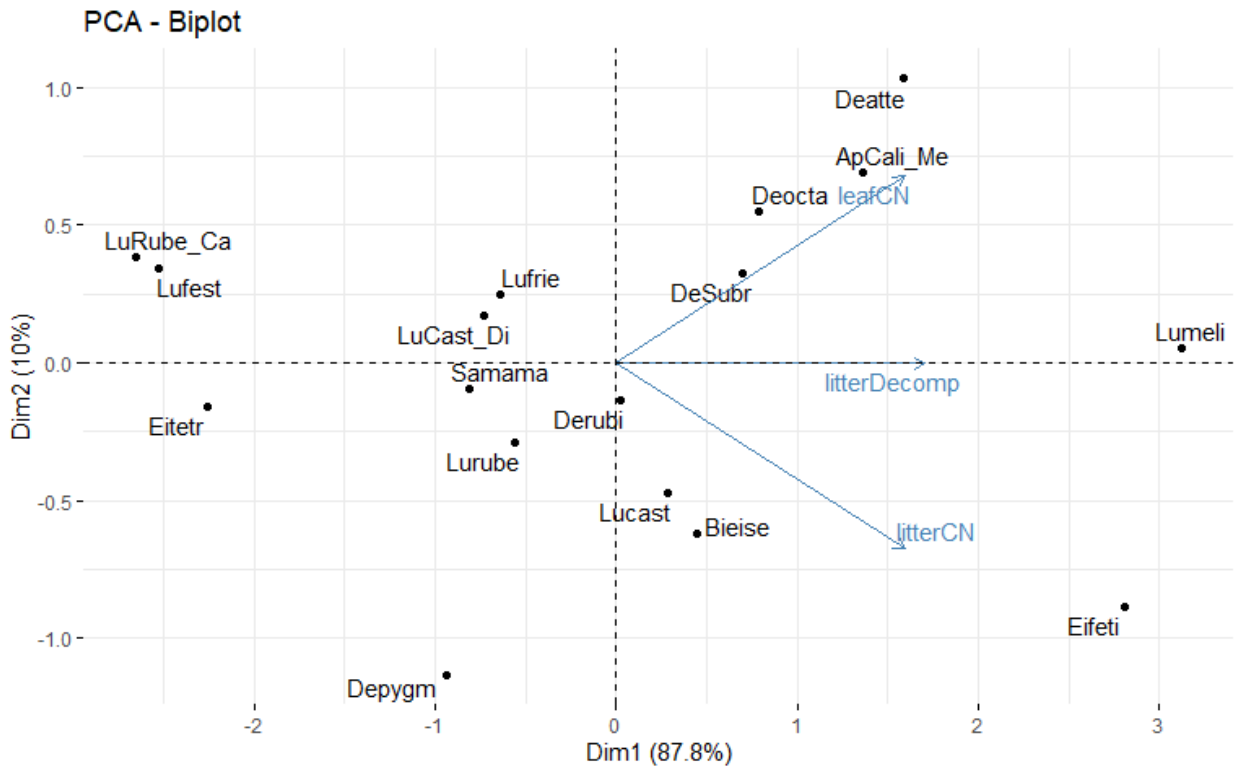


Figure 6.20: Principal component analysis biplot of mean tree traits associated to epigeics occurrences in forest sites sampled by Bouché [1972]

.

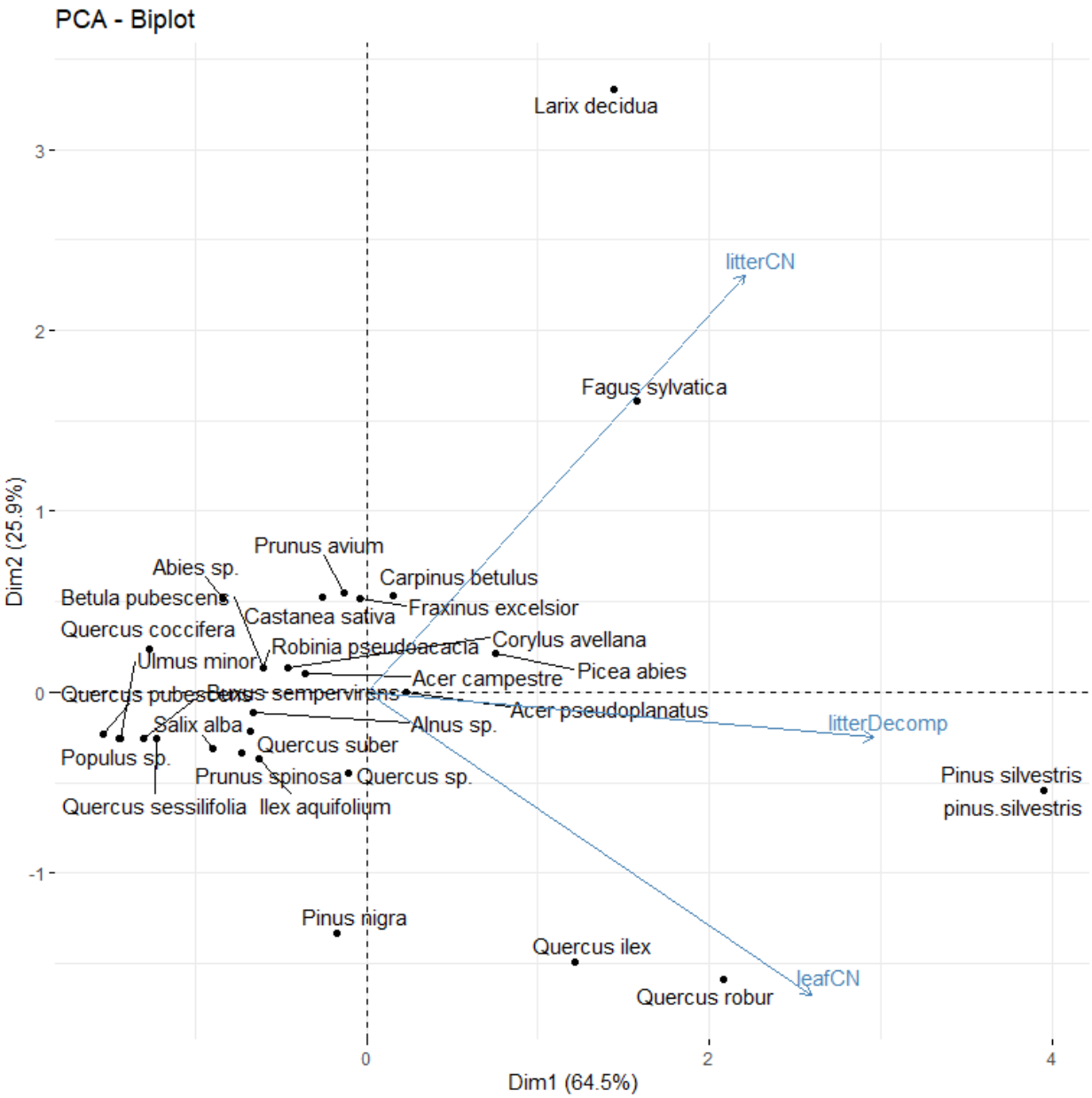


Figure 6.21: Principal component analysis biplot of tree traits in forest sites sampled by Bouché [1972]

.

6.5.5 Measuring the diversity of an ecological community

To quantify the diversity of a censused ecological community of $\mathcal{S}$ species, we follow the distance-based diversity framework by Leinster and Cobbold [2012]. It takes as input:

- A relative abundance vector p describing the proportions of each species. The term "species" refers to the biological unit chosen to describe the community composition.
- A species by species similarity matrix $Z = (Z_{ij})$. It is assumed that $0 \leq Z_{ij} \leq 1$. $Z_{ij} = 1$ indicates total dissimilarity, whereas $Z_{ij} = 1$ for identical species. If no information is available, the naive setting considers all species to be completely unique, i-e $Z = Id$. Distances (values in $[0, +\infty[$) are transformed into similarity using a negative exponential transformation: $Z_{ij} = e^{-d_{ij}}$.

- The sensitivity parameter, $q = 0...\infty$, which controls the relative importance attached to species abundance. For $q = 0$, prevalent and rare species are equally important. At the other extreme i-e $q = \infty$, rare species are completely ignored.

The nature of the similarity described by Z determines the type of diversity measured: taxonomic, phylogenetic, functional, trophic, etc. The diversity of order q is computed as follows:

$$(O)_i = \sum_{j=1}^{S} Z_{ij} p_i$$ *Ordinariness*: relative abundance of species similar to the ith

$$C^q = \left(\sum_{i=1}^{S} p_i O_i^{q-1} \right)^{\frac{1}{q-1}}$$ *Generalized mean ordinariness* (concentration)

$$D^q = \frac{1}{C^q} = \left(\sum_{i=1}^{S} p_i O_i^{q-1} \right)^{\frac{1}{1-q}}$$ *Diversity of order* $q \neq 1, \infty$

$$D^1 = \lim_{q \to 1} D^q = \frac{1}{(O_1)^{p_1} (O_2)^{p_2} ... (O_S)^{p_S}} = \prod_{i=1}^{S} (O_i)^{-p_i}$$

$$D^\infty = \lim_{q \to \infty} D^q = \frac{1}{\max(O_i)}$$

Diversity across spatial scales

Whittaker [1972] defined three terms for measuring biodiversity over spatial scales: alpha, beta, and gamma diversity. The γ-diversity is a measure of the total diversity within a region. The α-diversity is a measure of the local diversity in the study unit. The β-diversity is a pairwise measure of the turnover of species between communities at different study units. It is computed as the total number of species that are not common to both units, i-e: $\beta_{ss'} = |i \in S \mid Y_{si} > 0, Y_{s'i} > 0|$. The average beta diversity in a given study unit s can be deduced as $\beta_s = \gamma / \alpha_s$.

Chapter 7

Discussion

7.1 Advancing network inference

7.1.1 Theoretical advances and outstanding questions

The complex interplay of different assembly processes operating at various scales generates spatial co-abundance patterns, quantified by multivariate abundance data a.k.a community data [Weiher and Keddy, 2001]. Community data contain the footprint of eco-evolutionary filters and reflect species ecological niches. While Species Distribution Models (SDM) in biogeography attempt to learn the abiotic preferences of species, i.e their Grinnellian niches, Ecological Network Learning in community ecology aims to infer species interactions, i.e. their Eltonian niches. Interspecific association networks bridge biogeography and community ecology by connecting ecological interactions to species distributions.

Biotic interactions are physical processes that occur between individuals on a one-off or long-term basis, whereas associations are quantified at the level of the species (meta-)population on the scale of data sampling, which seldom matches the scale of interactions. Thus, associations are highly scale-dependent [Zurell et al., 2018, Araújo and Rozenfeld, 2014, Thuiller et al., 2015]. Whether two species on a spatial unit attract (positive association) or repulse (negative association) each other results from several factors including: the total effect of direct interactions between individuals of their populations, but also indirect effects mediated by the shared habitat. Even when models include environmental descriptors to rule out the effects mediated by the habitat, dependencies may still arise from indirect interactions mediated by non-sampled species, model mis-specifications or missing variables. All these arguments highlight the fact that pinpointing the exact type of ecological interactions is conceptually impossible [Blanchet et al., 2020, Barner et al., 2018, Peterson et al., 2020, Holt and Barfield, 2009]. Nevertheless, association networks can be powerful hypothesis generators for empirical studies.

Much has been written on the question of inferring interactions from co-occurrences, the current consensus being that co-occurrence is not evidence for interaction [Blanchet et al., 2020, Peterson et al., 2020]. Some studies argued that taking into account the effect of the environment would allow the effect of interactions to be isolated [Momal et al., 2019], so long as all relevant predictors are measured. Simulation studies using process-based models [Münkemüller and Gallien, 2015] or generative models showed promising results with idealistic error-free data. However, the unsupervised nature of the association inference problem made it difficult to evaluate the existing frameworks on empirical data [Barner et al., 2018, Pichler and Hartig, 2020]. The evaluation was further complicated by the lack of theory on how various types of biotic interactions translate into associations. For instance, does a host-parasite interaction lead to a negative or a positive association between the host and parasite species? In this example, one may argue that despite the co-occurrence of the two antagonists, abundances would be more suitable to capture the negative association [Blanchet et al., 2020]. Our virtual experiments showed that regardless of the inference framework, association strength depended on niche overlap. In our empirical application, it depended on species prevalence and mean abundance (also reported by [Zurell et al., 2018]).

Ecological assembly processes can interact in an additive (compensatory effects), multiplicative (intersection of requirements), interactive (e.g. an abiotic factor may alter the type or magnitude of an interaction) or even hierarchical or nested (e.g. trophic relationship requires co-occurrence which is conditioned by the suitability of the habitat for prey and predator) manner. Thus, untangling their respective effects is based on a good choice of the response model. Most inference models that integrate covariates do so in a linear fashion (additive model). We showed in Chapter 5 (5.9) via a simulation of predator-prey co-occurrences that an additive model fails to detect the dependency between species when their abiotic niche overlap strongly. Furthermore, the positive dependency of predators on their preys is sometimes confused by a negative association with non prey species that don't overlap with its preys. This result supports the claim that several interaction networks under different niches can generate similar co-occurrence patterns, which makes the problem non-identifiable. Expert knowledge of potential interactions can inform the choice of the response function.

Association inference from community data is well-suited for the detection of interactions that are inherently related to habitat and space such as indirect effects mediated by micro-habitat conditions, resource competition, facilitation, etc. This is often the case when modeling species of the same guild (e.g plant-plant). On the other hand, with novel community data obtained via eDNA sampling, more diverse interactions can be expected, thus the ability to detect a wide range of association types (symmetric, asymmetric) becomes more crucial.

7.1.2 Towards bridging network learning paradigms

The decomposition of associations into a dot product of embeddings is an old idea in computer science [Heckerman et al., 2000]. In recommendation systems, this factorization is akin to a problem known as "collaborative filtering". It is typically applied to matrix completion problems. The method developed can thus be applied to scenarios where the association matrix is only partially observed, bridging network inference with missing link prediction approaches.

There is also a promising avenue for further research in combining trait-based ecological network reconstruction that mostly targets direct interactions (trophic, pollination, parasitism) with network inference methods that identify residual associations to species environmental responses. Interactions can be used to constrain association networks, or conversely associations can be used to select the most plausible links. Ultimately, associations should be considered as a standalone facet in the multi-layer network of species dependencies, along with biotic interactions.

Since biotic interactions are dynamic processes, a more ideal setting for network inference would be to have time-series of community data. Aside from Bayesian Networks which have a natural extension to time-series (Dynamic Bayesian Neworks [Auclair et al., 2016]), existing frameworks cannot straightforwardly be applied to dynamic data. Our proposed model can easily be extended to time-series by simply changing the biotic context definition (3.2.1).

7.2 Advancing joint species distribution models

7.2.1 Benefits and limits of neural architectures

Joint species distribution models address the network inference problem while simultaneously learning about species abiotic preferences, thus allowing the prediction of community composition in response to the environmental conditions. Current JSDM implementations rely on the Generalised Linear Model to estimate the abiotic responses (but see Generalized Additive Mixed models [Knape, 2016] and stochastic neural networks [Harris, 2015]). This limitation is due to the difficulty to integrate random effects in other machine learning models (e.g Random Forest, Boosted Regression Trees). Neural networks can be viewed as deterministic parametric feature learning functions composed with generalized linear models or other probabilistic response distribution. Based on this, we proposed an extension to existing JSDMs that combines the function representation power of neural networks with the inferential capacity of hierarchical models to support complex features and model non-linear responses. The proposed model supports response distribution in

the exponential family as opposed to other neural network based JSDMs that were specifically optimized for the multivariate probit model [Chen et al., 2016, Pichler and Hartig, 2020, Chen et al., 2018, Harris, 2015].

The use of a variational autoencoder (VAE) to fit the latent variable part of the model can be interpreted as a non-linear factor analysis on the residuals of the abiotic response model. The main advantage of a VAE is the fixed complexity (number of parameters) with respect to dataset size. VAEs are optimized using gradient descent which makes them composable with other feedforward neural networks. One important limitation of VAEs is the prior assumption that latent sample representations are independent and identically distributed. Extensions of VAEs that account for spatial/temporal structure of observations (e.g spatial autocorrelation) have emerged recently, for instance using Gaussian Process priors (GPPVAE [Casale et al., 2018]) on the latent factors instead of isotropic multivariate Gaussian priors [Kingma and Welling, 2013]. Aside from the gaussian graphical model, VAEs have been applied outside ecological research to learn the structure of Bayesian Networks [Zhang et al., 2019] and Gaussian Copulas [Suh and Choi, 2016, Wang and Wang, 2019, Mazoure, 2018].

The ill-conditioned and highly multimodal likelihood functions associated with most neural networks make it difficult to quantify the uncertainty associated with parameter estimates. Thus, only point estimates of local optima are provided. To quantify parameter uncertainty, on moderate-size networks, Bayesian neural networks [Tran et al., 2019] enable the specification of a prior over the parameters. Then, stochastic variational inference (SVI) [Blei et al., 2017] is used to perform posterior inference on those parameters. An alternative solution called deep ensembles consist on optimizing the neural network parameters starting from different initial parameter values [Lakshminarayanan et al., 2017]. Other approaches rely on the post-hoc analysis of stochastic gradient descent trajectories to weight various regions of the parameter space [Izmailov et al., 2020].

7.2.2 Contribution of multi-task learning and improvement perspectives

Ecological community data presents several challenges, including often a higher number of species than observations, a few abundant species and many rare species. By sharing feature (representation) learning layers, multi-task neural networks reduce the number of parameters and distribute statistical power amongst species prediction tasks. Moreover, parameter sharing operates as a regularization force. In our case study, it made sense to share representations between all species, particularly as we model similar organisms that are generally affected by the same abiotic factors (earthworms). Multi-task models are susceptible to task interference, which occurs when the representation sharing does more harm than good. In general, determining the proportion of shared versus specific layers is application-specific. Future efforts should be directed towards an adaptive selection of the feature sharing architecture [Ruder et al., 2019, Rosenbaum et al., 2017, Misra et al., 2016].

In our case study on earthworms, the main benefit of the proposed architecture was a regularizing effect of representation sharing that led to a better generalization of the model compared to single species models, as measured by the predictive performances on the evaluation dataset (6.2). Just as these shared representations have benefited the modeled pool, they can be extended to new taxa by plugging new regression tasks on the shared features, and fitting only their specific parameters. Furthermore, the feature learning layers can be re-purposed and fine-tuned via the techniques of transfer learning [Torrey and Shavlik, 2010] for another related application subject to the same environmental conditions such as modeling earthworm traits responses, or modeling the habitat suitability of other edaphic groups such as collembola, nematodes, ants, etc.

The multi-task neural network architecture consists of shared layers representing a transformation function of the raw features, and specific layers for each taxa. In our case study, specific layers were set to be GLMs. The taxa-wise GLM regression weights allow to embed species in a shared response space, analogous to the Hutchinson's niche hypervolume. But, in general, specific layers can have different depths per taxa. In this case, species response weights cannot be represented in the same space. However, it is possible to use partial responses derived for instance using the SHAP framework (or local surrogate models) to cluster species. The drawback of this approach is that the resulting species clusters will depend on the dataset used for computing

SHAP values. Species clustering can also be performed in an ad-hoc fashion similarly to Species Archetype Models [Hui et al., 2013]. This can be achieved by using a discrete (multinomial) latent variable layer with a Dirichlet prior (e.g Dirichlet Autoencoder [Joo et al., 2020]).

7.3 Prospective applications

The flexibility in the response types of the proposed multi-task model paves the way for applications on mixed data such as the joint prediction of functional traits responses to the environment. In practice, it's been shown that multi-task learning can benefit from combining mildly related tasks [Liebel and Körner, 2018]. For instance, one can jointly predict species and functional traits. The proposed framework can also be applied outside community modeling, to predict functions (e.g decomposition, respiration, enzymatic activity, etc.) Finally, while we focused on ecological assembly processes, our case study on earthworms highlighted the importance of phylogeographic history, which calls for further investigations into modeling genetic diversity distribution.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.

Ian Abbott. Do earthworms compete for food? Soil Biology and Biochemistry, 12(6):523–530, 1980.

JA Addison. Distribution and impacts of invasive earthworms in canadian forest ecosystems. Biological Invasions, 11(1):59–79, 2009.

Andrej Aderhold, Dirk Husmeier, Jack J Lennon, Colin M Beale, and V Anne Smith. Hierarchical bayesian models in ecology: reconstructing species interaction networks from non-homogeneous species abundance data. Ecological Informatics, 11:55–64, 2012.

Charu C Aggarwal. Data mining: the textbook. Springer, 2015.

John Aitchison and CH Ho. The multivariate poisson-log normal distribution. Biometrika, 76(4):643–653, 1989.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10):1340–1347, 2010.

Jean-Paul Amat, Lucien Dorize, and Emmanuèle Gautier. Éléments de géographie physique. Editions Bréal, 2008.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468, 2016.

Miguel B Araújo and Miska Luoto. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography, 16(6):743–753, 2007.

Miguel B Araújo and Alejandro Rozenfeld. The geographic scaling of biotic interactions. Ecography, 37(5): 406–415, 2014.

Miguel B Araújo, Alejandro Rozenfeld, Carsten Rahbek, and Pablo A Marquet. Using species co-occurrence networks to assess the impacts of climate change. Ecography, 34(6):897–908, 2011.

E Auclair, N Peyrard, and R Sabbadin. Learning network structure using parameterized dynamic bayesian networks. Journées Française des Réseaux Bayésiens, Clermond-Ferrand, 2016.

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. Journal of Machine Learning Research, 4(May):83–99, 2003.

Allison K Barner, Kyle E Coblentz, Sally D Hacker, and Bruce A Menge. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. Ecology, 99(3):557–566, 2018.

Sébastien Barot, J-P Rossi, and Patrick Lavelle. Self-organization in a simple consumer–resource system, the example of earthworms. Soil Biology and Biochemistry, 39(9):2230–2240, 2007.

Frederic Barraquand, Coralie Picoche, Matteo Detto, and Florian Hartig. Inferring species interactions using granger causality and convergent cross mapping. arXiv preprint arXiv:1909.00731, 2019.

Edmundo Barrios. Soil biota, ecosystem services and land productivity. Ecological economics, 64(2):269–285, 2007.

Ignasi Bartomeus, Dominique Gravel, Jason M Tylianakis, Marcelo A Aizen, Ian A Dickie, and Maud Bernard-Verdier. A common framework for identifying linkage rules across different types of interactions. Functional Ecology, 30(12):1894–1903, 2016.

Niels H Batjes, Eloi Ribeiro, and Ad Van Oostrum. Standardised soil profile data to support global mapping and modelling (wosis snapshot 2019). Earth System Science Data, 12(1):299–320, 2020.

Roberto Battiti. First-and second-order methods for learning: between steepest descent and newton's method. Neural computation, 4(2):141–166, 1992.

Irene MA Bender, W Daniel Kissling, Pedro G Blendinger, Katrin Böhning-Gaese, Isabell Hensen, Ingolf Kühn, Marcia C Muñoz, Eike Lena Neuschulz, Larissa Nowak, Marta Quitián, et al. Morphological trait matching shapes plant–frugivore networks across the andes. Ecography, 41(11):1910–1919, 2018.

A BETSI. Database for functional traits of soil invertebrates. French Foundation for Biodiversity Research, 2:4, 2012.

Keith Beven and Peter Germann. Macropores and water flow in soils revisited. Water resources research, 49(6):3071–3092, 2013.

Eric Blanchart. Restoration by earthworms (megascolecidae) of the macroaggregate structure of a destructured savanna soil under field conditions. Soil Biology and Biochemistry, 24(12):1587–1594, 1992.

Eric Blanchart, Alain Albrecht, Julio Alegre, Arnaud Duboisset, Cécile Gilot, Beto Pashanasi, Patrick Lavelle, and Lijbert Brussaard. Effects of earthworms on soil structure and physical properties. Earthworm management in tropical agroecosystems, 5:149–171, 1999.

F Guillaume Blanchet, Kevin Cazelles, and Dominique Gravel. Co-occurrence is not evidence of ecological interactions. Ecology Letters, 2020.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518):859–877, 2017.

David A Bohan, Geoffrey Caron-Lormier, Stephen Muggleton, Alan Raybould, and Alireza Tamaddoni-Nezhad. Automated discovery of food webs from ecological data using logic-based machine learning. PLoS One, 6(12):e29028, 2011.

David A Bohan, Corinne Vacher, Alireza Tamaddoni-Nezhad, Alan Raybould, Alex J Dumbrell, and Guy Woodward. Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. Trends in Ecology & Evolution, 32(7):477–487, 2017.

Luigi Boitani, Alessandra Falcucci, Luigi Maiorano, and Carlo Rondinini. Ecological networks as conceptual frameworks or operational tools in conservation. Conservation biology, 21(6):1414–1422, 2007.

Virginie Boreux and Alexandra-Maria Klein. Global pollinator database. 10 2019. doi: 10.6084/m9.figshare.9980471.v1. URL https://figshare.com/articles/dataset/Global_pollinator_database/9980471.

Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A deep learning approach to species distribution modelling. In Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pages 169–199. Springer, 2018.

Christophe Botella, Maximilien Servajean, Pierre Bonnet, and Alexis Joly. Overview of geolifeclef 2019: plant species prediction using environment and animal occurrences. In CLEF working notes 2019, 2019.

Nicolas Bottinelli and Yvan Capowiez. Earthworm ecological categories are not functional groups. Biology and Fertility of Soils, pages 1–3, 2020.

Nicolas Bottinelli, Mickaël Hedde, Pascal Jouquet, and Yvan Capowiez. An explicit definition of earthworm ecological categories–marcel bouché's triangle revisited. Geoderma, 372:114361, 2020.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, pages 177–186. Springer, 2010.

Marcel B Bouché. Lombriciens de France: écologie et systématique, volume 72. Institut national de la recherche agronomique Paris, 1972.

Marcel B Bouché. Les modalités d'adaptation des lombriciens à la sécheresse. Bulletin de la Société Botanique de France. Actualités Botaniques, 131(2-4):319–327, 1984.

Simon Boudsocq, Hans Lambers, and Philippe Hinsinger. Shifts in belowground interactions in a legume-cereal intercrop: a test of the stress gradient hypothesis. In Sfécologie-2016, International Conference of Ecological Sciences, 2016.

Isabelle Boulangeat, Dominique Gravel, and Wilfried Thuiller. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecology letters, 15(6): 584–593, 2012.

L Breiman, J Friedman, R Olshen, C Stone, D Steinberg, and P Colla. Cart: Classification and regression trees (p. 156). Belmont, CA: Wadsworth, 1983.

Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

MJI Briones and R Álvarez-Otero. Body wall thickness as a potential functional trait for assigning earthworm species to ecological categories. Pedobiologia, 67:26–34, 2018.

Jedediah F Brodie, Clare E Aslan, Haldre S Rogers, Kent H Redford, John L Maron, Judith L Bronstein, and Craig R Groves. Secondary extinctions of biodiversity. Trends in Ecology & Evolution, 29(12):664–672, 2014.

Eduardo S Brondizio, J Settele, S Díaz, and HT Ngo. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. IPBES Secretariat: Bonn, Germany, 2019.

Pierre-Marc Brousseau, Dominique Gravel, and I Tanya Handa. Trait matching and phylogeny as predictors of predator–prey interactions involving ground beetles. Functional Ecology, 32(1):192–202, 2018.

George G Brown. How do earthworms affect microfloral and faunal community diversity? Plant and soil, 170(1):209–231, 1995.

John F Bruno, John J Stachowicz, and Mark D Bertness. Inclusion of facilitation into ecological theory. Trends in ecology & evolution, 18(3):119–125, 2003.

L Brussaard, DK Aanen, MJI Briones, T Decaëns, GB De Deyn, TM Fayle, SW James, and T Nobre. Biogeography and phylogenetic community structure of soil invertebrate ecosystem engineers, global to local patterns, implications for ecosystem, functioning and services and global, environmental change impacts. In Soil Ecology Ecosystems Services, pages 201–232. Oxford University Press, 2012.

Else K Bünemann, Giulia Bongiorno, Zhanguo Bai, Rachel E Creamer, Gerlinde De Deyn, Ron de Goede, Luuk Fleskens, Violette Geissen, Thom W Kuyper, Paul Mäder, et al. Soil quality–a critical review. Soil Biology and Biochemistry, 120:105–125, 2018.

U Burkhardt, DJ Russell, P Decker, M Döhler, H Höfer, S Lesch, S Rick, J Römbke, C Trog, J Vorwald, et al. The edaphobase project of gbif-germany—a new online soil-zoological data warehouse. Applied Soil Ecology, 83:3–12, 2014.

Kevin R Butt. Interactions between selected earthworm species: a preliminary, laboratory-based study. Applied Soil Ecology, 9(1-3):75–79, 1998.

Ragan M Callaway. Positive interactions and interdependence in plant communities. Springer, 2007.

Ragan M Callaway, RW Brooker, Philippe Choler, Zaal Kikvidze, Christopher J Lortie, Richard Michalet, Leonardo Paolini, Francisco I Pugnaire, Beth Newingham, Erik T Aschehoug, et al. Positive interactions among alpine plants increase with stress. Nature, 417(6891):844, 2002.

Erin K Cameron, Thibaud Decaëns, Emmanuel Lapied, David Porco, and Nico Eisenhauer. Earthworm databases and ecological theory: Synthesis of current initiatives and main research directions. Applied Soil Ecology, 104:85–90, 2016.

Yvan Capowiez, Alain Pierret, and Christopher J Moran. Characterisation of the three-dimensional structure of earthworm burrow systems using image analysis and mathematical morphology. Biology and Fertility of Soils, 38(5):301–310, 2003.

Marta Carboni, Irene Calderon-Sanou, Laura Pollock, Cyrille Violle, DivGrass Consortium, and Wilfried Thuiller. Functional traits modulate the response of alien plants along abiotic and biotic gradients. Global Ecology and Biogeography, 27(10):1173–1185, 2018.

Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In Proceedings of the Tenth International Conference on Machine Learning, pages 41–48. Morgan Kaufmann, 1993.

Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. In Advances in Neural Information Processing Systems, pages 10369–10380, 2018.

Kévin Cazelles, Miguel B Araújo, Nicolas Mouquet, and Dominique Gravel. A theory for species co-occurrence in interaction networks. Theoretical Ecology, 9(1):39–48, 2016a.

Kévin Cazelles, Nicolas Mouquet, David Mouillot, and Dominique Gravel. On the integration of biotic interaction and environmental constraints at the biogeographical scale. Ecography, 39(10):921–931, 2016b.

V Bala Chaudhary, Megan A Rúa, Anita Antoninka, James D Bever, Jeffery Cannon, Ashley Craig, Jessica Duchicela, Alicia Frame, Monique Gardes, Catherine Gehring, et al. Mycodb, a global database of plant response to mycorrhizal fungi. Scientific data, 3(1):1–10, 2016.

Di Chen, Yexiang Xue, Shuo Chen, Daniel Fink, and Carla Gomes. Deep multi-species embedding. arXiv preprint arXiv:1609.09353, 2016.

Di Chen, Yexiang Xue, and Carla P Gomes. End-to-end learning for the deep multivariate probit model. arXiv preprint arXiv:1803.08591, 2018.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for sparse network reconstruction from count data. arXiv preprint arXiv:1806.03120, 2018.

Philippe Choler. Consistent shifts in alpine plant traits along a mesotopographical gradient. Arctic, Antarctic, and Alpine Research, 37(4):444–453, 2005.

Philippe Choler, Richard Michalet, and Ragan M Callaway. Facilitation and competition on gradients in alpine plant communities. Ecology, 82(12):3295–3308, 2001.

François Chollet et al. Keras: The python deep learning library. ascl, pages ascl–1806, 2018.

Botella Christophe, Servajean Maximilien, Bonnet Pierre, and Joly Alexis. Overview of geolifeclef 2019: plant species prediction using environment and animal occurrences. 2019.

James S Clark, Diana Nemergut, Bijan Seyednasrollah, Phillip J Turner, and Stacy Zhang. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. Ecological Monographs, 87(1):34–56, 2017.

James S Clark, C Lane Scher, and Margaret Swift. The emergent interactions that govern biodiversity change. Proceedings of the National Academy of Sciences, 117(29):17074–17083, 2020.

Nicholas J Clark, Konstans Wells, and Oscar Lindberg. Unravelling changing interspecific interactions across environmental gradients using markov random fields. Ecology, 99(6):1277–1283, 2018.

Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. Nature, 453(7191):98–101, 2008.

Corine Land Cover. Corine land cover. European Environment Agency, Copenhagen, 2000.

James P Curry and Olaf Schmidt. The feeding ecology of earthworms–a review. Pedobiologia, 50(6):463–477, 2007.

Tad Dallas. helminthr: an r interface to the london natural history museum's host–parasite database. Ecography, 39(4):391–393, 2016.

Manuela d'Amen, Anne Dubuis, Rui F Fernandes, Julien Pottier, Loïc Pellissier, and Antoine Guisan. Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. Journal of biogeography, 42(7):1255–1266, 2015.

Charles Darwin. The formation of vegetable mould through the action of worms: with observations on their habits, volume 37. Appleton, 1892.

Thibaud Decaëns. Macroecological patterns in soil communities. Global Ecology and Biogeography, 19(3): 287–302, 2010.

Thibaud Decaëns, Pierre Margerie, Michael Aubert, Mickaël Hedde, and Fabrice Bureau. Assembly rules within earthworm communities in north-western france—a regional analysis. Applied Soil Ecology, 39(3): 321–335, 2008.

D Matthias Dehling, Pedro Jordano, H Martin Schaefer, Katrin Böhning-Gaese, and Matthias Schleuning. Morphology predicts species' functional roles and their degree of specialization in plant–frugivore interactions. Proceedings of the Royal Society B: Biological Sciences, 283(1823):20152444, 2016.

Benjamin Deneu, Maximilien Servajean, Christophe Botella, and Alexis Joly. Location-based species recommendation using co-occurrences and environment-geolifeclef 2018 challenge. In CLEF 2018, 2018.

Stephane AP Derocles, David A Bohan, Alex J Dumbrell, James JN Kitson, Francois Massol, Charlie Pauvert, Manuel Plantegenest, Corinne Vacher, and Darren M Evans. Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis. In Advances in ecological research, volume 58, pages 1–62. Elsevier, 2018.

Philippe Desjardins-Proulx, Idaline Laigle, Timothée Poisot, and Dominique Gravel. Ecological interactions and the netflix problem. PeerJ, 5:e3644, 2017.

Jared M Diamond. The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. Biological conservation, 7(2):129–146, 1975.

Sandra Díaz, Unai Pascual, Marie Stenseke, Berta Martín-López, Robert T Watson, Zsolt Molnár, Rosemary Hill, Kai MA Chan, Ivar A Baste, Kate A Brauman, et al. Assessing nature's contributions to people. Science, 359(6373):270–272, 2018.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7):1895–1923, 1998.

Virginia Domínguez-García, Vasilis Dakos, and Sonia Kéfi. Unveiling dimensions of stability in complex ecological networks. Proceedings of the National Academy of Sciences, 116(51):25714–25720, 2019.

Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitao, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1):27–46, 2013.

Carsten F Dormann, Maria Bobrowski, D Matthias Dehling, David J Harris, Florian Hartig, Heike Lischke, Marco D Moretti, Jörn Pagel, Stefan Pinkert, Matthias Schleuning, et al. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. Global ecology and biogeography, 27(9):1004–1016, 2018.

Piers K Dunstan, Scott D Foster, and Ross Darnell. Model based grouping of species across environmental gradients. Ecological Modelling, 222(4):955–963, 2011.

CA Edwards and JR Lofty. Nitrogenous fertilizers and earthworm populations in agricultural soils. Soil Biology and Biochemistry, 14(5):515–521, 1982.

Clive A Edwards and Patrick J Bohlen. Biology and ecology of earthworms, volume 3. Springer Science & Business Media, 1996.

Paul Eggleton, Kelly Inward, Joanne Smith, David T Jones, and Emma Sherlock. A six year study of earthworm (lumbricidae) populations in pasture woodland in southern england shows their responses to soil temperature and soil moisture. Soil Biology and Biochemistry, 41(9):1857–1865, 2009.

Nico Eisenhauer, Alexandru Milcu, Alexander CW Sabais, Holger Bessler, Alexandra Weigelt, Christof Engels, and Stefan Scheu. Plant community impacts on the structure of earthworm communities depend on season and change with time. Soil Biology and Biochemistry, 41(12):2430–2443, 2009.

Nico Eisenhauer, Artur Stefanski, Nicholas A Fisichelli, Karen Rice, Roy Rich, and Peter B Reich. Warming shifts 'worming': effects of experimental warming on invasive earthworms in northern north america. Scientific Reports, 4(1):1–7, 2014.

Anna Eklöf, Ute Jacob, Jason Kopp, Jordi Bosch, Rocío Castro-Urgal, Natacha P Chacoff, Bo Dalsgaard, Claudio de Sassi, Mauro Galetti, Paulo R Guimarães, et al. The dimensionality of ecological networks. Ecology letters, 16(5):577–583, 2013.

Jane Elith, Simon Ferrier, Falk Huettmann, and John Leathwick. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. Ecological modelling, 186(3): 280–289, 2005.

Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4):802–813, 2008.

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

Charles S Elton. Animal ecology. University of Chicago Press, 2001.

Ali Faisal, Frank Dondelinger, Dirk Husmeier, and Colin M Beale. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. Ecological Informatics, 5(6):451–464, 2010.

Liliana B Falco, R Sandler, F Momo, C Di Ciocco, L Saravia, and C Coviella. Earthworm assemblages in different intensity of agricultural uses and their relation to edaphic variables. PeerJ, 3:e979, 2015.

JE Fauth, J Bernardo, M Camara, WJ Resetarits Jr, Jr Van Buskirk, and SA McCollum. Simplifying the jargon of community ecology: a conceptual approach. The American Naturalist, 147(2):282–286, 1996.

H Ferris. Nematode plant expert information system (nemaplex). University of California. Recuperado de: http://plpnemweb. ucdavis. edu/nemaplex/index. htm, 1999.

Stephen E Fick and Robert J Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology, 37(12):4302–4315, 2017.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective. arXiv preprint arXiv:1801.01489, 68, 2018.

Joana Frazão, Ron GM de Goede, Tamás E Salánki, Lijbert Brussaard, Jack H Faber, Mickaël Hedde, and Mirjam M Pulleman. Responses of earthworm communities to crop residue management after inoculation of the earthworm lumbricus terrestris (linnaeus, 1758). Applied Soil Ecology, 142:177–188, 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441, 2008.

Jerome H Friedman. Multivariate adaptive regression splines. The annals of statistics, pages 1–67, 1991.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

Xiao Fu, Eugene Seo, Justin Clarke, and Rebecca A Hutchinson. Link prediction under imperfect detection: Collaborative filtering for ecological networks. IEEE Transactions on Knowledge and Data Engineering, 2019.

Laure Gallien, Tamara Münkemüller, Cécile H Albert, Isabelle Boulangeat, and Wilfried Thuiller. Predicting potential distributions of invasive species: where to go from here? Diversity and Distributions, 16(3):331–342, 2010.

Lucas A Garibaldi, Ignasi Bartomeus, Riccardo Bommarco, Alexandra M Klein, Saul A Cunningham, Marcelo A Aizen, Virginie Boreux, Michael PD Garratt, Luísa G Carvalheiro, Claire Kremen, et al. Editor's choice: Review: Trait matching of flower visitors and crops predicts fruit set better than trait diversity. Journal of Applied Ecology, 52(6):1436–1444, 2015.

Kevin J Gaston and John I Spicer. Biodiversity: an introduction. John Wiley & Sons, 2013.

Benoit Gauzens, Elisa Thébault, Gérard Lacroix, and Stéphane Legendre. Trophic groups and modules: two levels of group detection in food webs. Journal of The Royal Society Interface, 12(106):20141176, 2015.

Maxence Gérard, Maryse Vanderplanck, Thomas Wood, and Denis Michez. Global warming and plant–pollinator mismatches. Emerging Topics in Life Sciences, page ETLS20190139, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256, 2010.

Hervé Goëau, Pierre Bonnet, Alexis Joly, Vera Bakić, Julien Barbe, Itheri Yahiaoui, Souheil Selmi, Jennifer Carré, Daniel Barthélémy, Nozha Boujemaa, et al. Pl@ ntnet mobile app. In Proceedings of the 21st ACM international conference on Multimedia, pages 423–424, 2013.

Xin Gong, Shuai Wang, Zhenwei Wang, Yuji Jiang, Zhengkun Hu, Yong Zheng, Xiaoyun Chen, Huixin Li, Feng Hu, Manqiang Liu, et al. Earthworms modify soil bacterial and fungal communities through enhancing aggregation and buffering ph. Geoderma, 347:59–69, 2019.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1. MIT press Cambridge, 2016.

Nicholas J Gotelli and Gary R Graves. Null models in ecology. 1996.

Nicholas J Gotelli and Declan J McCabe. Species co-occurrence: a meta-analysis of jm diamond's assembly rules model. Ecology, 83(8):2091–2096, 2002.

Nicholas J Gotelli, Marti J Anderson, Hector T Arita, Anne Chao, Robert K Colwell, Sean R Connolly, David J Currie, Robert R Dunn, Gary R Graves, Jessica L Green, et al. Patterns and causes of species richness: a general simulation model for macroecology. Ecology letters, 12(9):873–886, 2009.

Gérard Govaert and Mohamed Nadif. Co-clustering: models, algorithms and applications. John Wiley & Sons, 2013.

Dominique Gravel, François Massol, Elsa Canard, David Mouillot, and Nicolas Mouquet. Trophic theory of island biogeography. Ecology letters, 14(10):1010–1016, 2011.

Dominique Gravel, Timothée Poisot, Camille Albouy, Laure Velez, and David Mouillot. Inferring food web structure from predator–prey body size relationships. Methods in Ecology and Evolution, 4(11):1083–1090, 2013.

Dominique Gravel, Camille Albouy, and Wilfried Thuiller. The meaning of functional trait composition of food webs for ecosystem functioning. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1694):20150268, 2016.

Dominique Gravel, Benjamin Baiser, Jennifer A Dunne, Jens-Peter Kopelke, Neo D Martinez, Tommi Nyman, Timothée Poisot, Daniel B Stouffer, Jason M Tylianakis, Spencer A Wood, et al. Bringing elton and grinnell together: a quantitative framework to represent the biogeography of ecological interaction networks. Ecography, 42(3):401–415, 2019.

Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. Nature, 548(7666):210–213, 2017.

Joseph Grinnell. The niche-relationships of the california thrasher. Auk, 34(4):427–433, 1917.

Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 106(52):22073–22078, 2009.

Antoine Guisan and Carsten Rahbek. Sesam–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography, 38(8):1433–1444, 2011.

Antoine Guisan, Wilfried Thuiller, and Niklaus E Zimmermann. Habitat suitability and distribution models: with applications in R. Cambridge University Press, 2017.

Mónica Gutiérrez-López, Juan B Jesús, Dolores Trigo, Rosa Fernández, Marta Novo, and Darío J Díaz-Cosín. Relationships among spatial distribution of soil microarthropods, earthworm species and soil properties. Pedobiologia, 53(6):381–389, 2010.

Roy Haines-Young and Marion Potschin. The links between biodiversity, ecosystem services and human well-being. Ecosystem Ecology: a new synthesis, 1:110–139, 2010.

Garrett Hardin. The competitive exclusion principle. science, 131(3409):1292–1297, 1960.

David J. Harris. Generating realistic assemblages with a joint species distribution model. Methods in Ecology and Evolution, 6(4):465–473, 2015. doi: 10.1111/2041-210X.12332. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12332.

David J Harris. Inferring species interactions from co-occurrence data with markov networks. Ecology, 97 (12):3308–3314, 2016.

Edward F Haskell. A clarification of social science. Main Currents in Modern Thought, 7(2):45–51, 1949.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015.

Trevor J Hastie and Robert J Tibshirani. Generalized additive models, volume 43. CRC press, 1990.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. IEEE Intelligent Systems and their applications, 13(4):18–28, 1998.

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research, 1(Oct):49–75, 2000.

Mickaël Hedde, Benjamin Pey, Appoline Auclerc, Yvan Capowiez, Daniel Cluzeau, Jérôme Cortet, Thibaud Decaëns, Louis Deharveng, Florence Dubs, Fabien Grumiaux, et al. Betsi, a complete framework for studying soil invertebrate functional traits. In XVI ICSZ-International Colloquium on Soil Zoology, page np, 2012.

TD Heilbronn and David WH Walton. Plant colonization of actively sorted stone stripes in the subantarctic. Arctic and Alpine Research, 16(2):161–172, 1984.

Matthew R Helmus, Kristina Savage, Matthew W Diebel, Jeffrey T Maxted, and Anthony R Ives. Separating the determinants of phylogenetic community structure. Ecology letters, 10(10):917–925, 2007.

Roland Hiederer. Mapping soil properties for europe—spatial representation of soil database attributes. Luxembourg: Publications Office of the European Union, 2013.

Robert J Hijmans, Steven Phillips, John Leathwick, and Jane Elith. dismo: Species distribution modeling. R package version, 1(4):1–1, 2017.

Jes Hines, Darren P Giling, Michael Rzanny, Winfried Voigt, Sebastian T Meyer, Wolfgang W Weisser, Nico Eisenhauer, and Anne Ebeling. A meta food web for invertebrate species collected in a european grassland. Ecology, 100(6):e02679, 2019.

Edward Hollingdale, Francisco Javier Pérez-Barbería, and David McPetrie Walker. Inferring symmetric and asymmetric interactions between animals and groups from positional data. PloS one, 13(12), 2018.

Robert D Holt. Population dynamics in two-patch environments: some anomalous consequences of an optimal habitat distribution. Theoretical population biology, 28(2):181–208, 1985.

Robert D Holt and Michael Barfield. Trophic interactions and range limits: the diverse roles of predation. Proceedings of the Royal Society B: Biological Sciences, 276(1661):1435–1442, 2009.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2):251–257, 1991.

Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of machine learning research, 5(Nov):1457–1469, 2004.

Francis KC Hui, David I Warton, Scott D Foster, and Piers K Dunstan. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology, 94(9): 1913–1919, 2013.

Francis KC Hui, David I Warton, Scott D Foster, et al. Multi-species distribution modeling using penalized mixture of regressions. The Annals of Applied Statistics, 9(2):866–882, 2015.

Alexander von Humboldt, Aimé Bonpland, et al. Essai sur la géographie des plantes. Chez Levrault, Schoell et compagnie, libraires, 1805.

GE Hutchinson. The multivariate niche. In Cold Spring Harbor Symposia on Quantitative Biology, volume 22, pages 415–421, 1957.

Us1106 Infosol. Donesol: la base de données nationale des sols de france. In Conférence Année Internationale des Sols (AIS), 2015.

ISRIC. Soilgrids: an automated system for global soil mapping. 2013.

Anthony R Ives and Matthew R Helmus. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs, 81(3):511–525, 2011.

Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In Uncertainty in Artificial Intelligence, pages 1169–1179. PMLR, 2020.

Samuel W James and Mary R Cunningham. Feeding ecology of some earthworms in kansas tallgrass prairie. American Midland Naturalist, pages 78–83, 1989.

Samuel W James, David Porco, Thibaud Decaëns, Benoit Richard, Rodolphe Rougerie, and Christer Erséus. Dna barcoding reveals cryptic diversity in lumbricus terrestris l., 1758 (clitellata): resurrection of l. herculeus (savigny, 1826). PloS one, 5(12):e15629, 2010.

Juan J Jiménez and Juliane Filser. Soil fauna: key to soil organic matter dynamics and modelling. 2020.

Juan-José Jiménez, Thibaud Decaëns, and Jean-Pierre Rossi. Stability of the spatio-temporal distribution and niche overlap in neotropical earthworm assemblages. Acta Oecologica, 30(3):299–311, 2006.

Juan-José Jiménez, Thibaud Decaëns, and Jean-Pierre Rossi. Soil environmental heterogeneity allows spatial co-occurrence of competitor earthworm species in a gallery forest of the colombian 'llanos'. Oikos, 121(6): 915–926, 2012.

Marc TJ Johnson and John R Stinchcombe. An emerging synthesis between community ecology and evolutionary biology. Trends in ecology & evolution, 22(5):250–257, 2007.

Alice SA Johnston, Martin Holmstrup, Mark Edward Hodson, P Thorbek, T Alvarez, and RM Sibly. Earthworm distribution and abundance predicted by a process-based model. Applied Soil Ecology, 84:112–123, 2014.

Alexis Joly, Hervé Goëau, Christophe Botella, Stefan Kahl, Maximillien Servajean, Hervé Glotin, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Fabian-Robert Stöter, and Henning Müller. Overview of lifeclef 2019: Identification of amazonian plants, south & north american birds, and niche prediction. In Proceedings of CLEF 2019, 2019.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. Pattern Recognition, 107:107514, 2020.

Dirk Nikolaus Karger, Olaf Conrad, Jürgen Böhner, Tobias Kawohl, Holger Kreft, Rodrigo Wilber Soria-Auza, Niklaus E Zimmermann, H Peter Linder, and Michael Kessler. Climatologies at high resolution for the earth's land surface areas. Scientific data, 4:170122, 2017.

DN Karger, O Conrad, J Böhner, T Kawohl, H Kreft, RW Soria-Auza, NE Zimmermann, HP Linder, and M Kessler. Chelsa climatologies at high resolution for the earth's land surface areas (version 1.1). World Data Cent. Clim. doi: https://doi. org/10.1594/WDCC/CHELSA_v1_1, 2016.

Jens Kattge, Sandra Diaz, Sandra Lavorel, I Colin Prentice, Paul Leadley, Gerhard Bönisch, Eric Garnier, Mark Westoby, Peter B Reich, Ian J Wright, et al. Try–a global database of plant traits. Global change biology, 17(9):2905–2935, 2011.

Jens Kattge, Gerhard Bönisch, Sandra Díaz, Sandra Lavorel, Iain Colin Prentice, Paul Leadley, Susanne Tautenhahn, Gijsbert DA Werner, Tuomas Aakala, Mehdi Abedi, et al. Try plant trait database–enhanced coverage and open access. Global change biology, 26(1):119–188, 2020.

Petr Keil, Thorsten Wiegand, Anikó B Tóth, Daniel J McGlinn, and Jonathan M Chase. Interspecific spatial associations as a facet of biodiversity, and a review of approaches to measure them. 2019.

MG Kibblewhite, K Ritz, and MJ Swift. Soil health in agricultural systems. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1492):685–701, 2008.

R Andrew King, Amy L Tibble, and William OC Symondson. Opening a can of worms: unprecedented sympatric cryptic diversity within british lumbricid earthworms. Molecular ecology, 17(21):4684–4698, 2008.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691, 2019.

W Daniel Kissling, Carsten F Dormann, Jürgen Groeneveld, Thomas Hickler, Ingolf Kühn, Greg J McInerny, José M Montoya, Christine Römermann, Katja Schiffers, Frank M Schurr, et al. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography, 39(12):2163–2178, 2012.

Jonas Knape. Decomposing trends in swedish bird populations using generalized additive mixed models. Journal of Applied Ecology, 53(6):1852–1861, 2016.

Daphne Koller and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Sadanori Konishi and Genshiro Kitagawa. Information criteria and statistical modeling. Springer Science & Business Media, 2008.

Charles J Krebs. The experimental analysis of distribution and abundance. Ecology. New York: Harper and Row, pages 1–14, 1972.

Idaline Laigle, Isabelle Aubin, Christoph Digel, Ulrich Brose, Isabelle Boulangeat, and Dominique Gravel. Species traits as drivers of food web structure. Oikos, 127(2):316–326, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in neural information processing systems, pages 6402–6413, 2017.

Meredith A Lane and James L Edwards. The global biodiversity information facility (gbif). Biodiversity databases: Techniques, politics, and applications, pages 1–4, 2007.

Nina K Lany, Phoebe L Zarnetske, Erin M Schliep, Robert N Schaeffer, Colin M Orians, David A Orwig, and Evan L Preisser. Asymmetric biotic interactions and abiotic niche differences revealed by a dynamic joint species distribution model. Ecology, 99(5):1018–1023, 2018.

P Lavelle. Earthworm activities and the soil system. Biology and fertility of soils, 6(3):237–251, 1988.

P Lavelle, D Bignell, M Lepage, V Wolters, Pierre-Armand Roger, POWH Ineson, OW Heal, and S Dhillion. Soil function in a changing world: the role of invertebrate ecosystem engineers. European Journal of soil biology, 33(4):159–193, 1997.

Patrick Lavelle and AV Spain. Soil ecology. Springer Science & Business Media, 2001.

Patrick Lavelle, Armand Chauvel, and C Fragoso. Faunal activity in acid soils. In Plant-Soil Interactions at Low pH: Principles and Management, pages 201–211. Springer, 1995.

Sandra Lavorel and Eric Garnier. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the holy grail. Functional ecology, 16(5):545–556, 2002.

Craig A Layman, Marcio S Araujo, Ross Boucek, Caroline M Hammerschlag-Peyer, Elizabeth Harrison, Zachary R Jud, Philip Matich, Adam E Rosenblatt, Jeremy J Vaudo, Lauren A Yeager, et al. Applying stable isotopes to examine food-web structure: an overview of analytical tools. Biological Reviews, 87(3): 545–562, 2012.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995, 1995.

Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. Ecology, 93(3):477–489, 2012.

Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334, 2018.

Jianhua Lin. Divergence measures based on the shannon entropy. IEEE Transactions on Information theory, 37(1):145–151, 1991.

Chieh Lo and Radu Marculescu. Mplasso: Inferring microbial association networks using prior microbial knowledge. PLoS computational biology, 13(12):e1005915, 2017.

Christopher J Lortie, Rob W Brooker, Phillipe Choler, Zaal Kikvidze, Richard Michalet, Francisco I Pugnaire, and Ragan M Callaway. Rethinking plant community theory. Oikos, 107(2):433–438, 2004.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in neural information processing systems, pages 4765–4774, 2017.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In Proc. icml, volume 30, page 3, 2013.

Robert H MacArthur and Edward O Wilson. The theory of island biogeography, volume 1. Princeton university press, 2001.

Fernando T Maestre, Fernando Valladares, and James F Reynolds. Is the change of plant–plant interactions with abiotic stress predictable? a meta-analysis of field results in arid environments. Journal of Ecology, 93(4):748–757, 2005.

Luigi Maiorano, Alessandro Montemaggiori, Gentile Francesco Ficetola, Louise O'connor, and Wilfried Thuiller. Tetra-eu 1.0: A species-level trophic metaweb of european tetrapods. Global Ecology and Biogeography, 29(9):1452–1457, 2020.

Nabil Majdi, Nicolas Hette-Tronquart, Etienne Auclair, Alexandre Bec, Tiphaine Chouvelon, Bruno Cognie, Michaël Danger, Priscilla Decottignies, Aurélie Dessier, Christian Desvilettes, et al. There's no harm in having too much: A comprehensive toolbox of methods in trophic ecology. Food webs, 17:e00100, 2018.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, pages 50–60, 1947.

Daniel F Marchán, Pablo Refoyo, Marta Novo, Rosa Fernández, Dolores Trigo, and Darío J Díaz Cosín. Predicting soil micro-variables and the distribution of an endogeic earthworm species through a model based on large-scale variables. Soil Biology and Biochemistry, 81:124–127, 2015.

Daniel F Marchán, Marta Novo, Rosa Fernández, Irene De Sosa, Dolores Trigo, and Darío J Díaz Cosín. Evaluating evolutionary pressures and phylogenetic signal in earthworms: a case study-the number of typhlosole lamellae in hormogastridae (annelida, oligochaeta). Zoological Journal of the Linnean Society, 178(1):4–14, 2016a.

Daniel F Marchán, Pablo Refoyo, Rosa Fernández, Marta Novo, Irene de Sosa, and Darío J Díaz Cosín. Macroecological inferences on soil fauna through comparative niche modeling: the case of hormogastridae (annelida, oligochaeta). European Journal of Soil Biology, 75:115–122, 2016b.

Daniel F Marchán, Mickaël Hedde, Emmanuel Lapied, Marie-Eugénie Maggia, Marta Novo, Jorge Domínguez, and Thibaud Decaëns. Contrasting phylogeographic patterns of earthworms (crassiclitellata, lumbricidae) on near-shore mediterranean islands. European Journal of Soil Biology, 101:103242, 2020a.

Daniel Fernández Marchán, Marta Novo, Nuria Sánchez, Jorge Domínguez, Darío J Díaz Cosín, and Rosa Fernández. Local adaptation fuels cryptic speciation in terrestrial annelids. Molecular phylogenetics and evolution, 146:106767, 2020b.

JG Mather, O Christensen, et al. Surface movements of earthworms in agricultural land. Pedobiologia, 32 (5/6):399–405, 1988.

Jerome Mathieu and T Jonathan Davies. Glaciation as an historical filter of below-ground biodiversity. Journal of Biogeography, 41(6):1204–1214, 2014.

Bogdan Mazoure. Estimating the structure of probabilistic graphical models through a Gaussian copula with discrete marginals. PhD thesis, McGill University Montreal, Quebec, 2018.

P McCullagh. Nelder. ja (1989), generalized linear models. CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York, 81.

Jacob P McDaniel, Kenneth A Barbarick, Mary E Stromberger, and Whitney Cranshaw. Survivability of aporrectodea caliginosa in response to drought stress in a colorado soil. Soil Science Society of America Journal, 77(5):1667–1672, 2013.

CF McQuaid and NF Britton. Host–parasite nestedness: a result of co-evolving trait-values. Ecological Complexity, 13:53–59, 2013.

Jeroen Meersmans, Manuel Pascal Martin, Eva Lacarce, Sarah De Baets, Claudy Jolivet, Line Boulonne, Sébastien Lehmann, Nicolas Philippe Anthony Saby, Antonio Bispo, and Dominique Arrouays. A high resolution map of french soil organic carbon. Agronomy for Sustainable Development, 32(4):841–851, 2012.

Filip JR Meysman, Jack J Middelburg, and Carlo HR Heip. Bioturbation: a fresh look at darwin's last idea. Trends in Ecology & Evolution, 21(12):688–695, 2006.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3994–4003, 2016.

Heidi K Mod, Mathieu Chevalier, Miska Luoto, and Antoine Guisan. Scale dependence of ecological assembly rules: insights from empirical datasets and joint species distribution modelling. Journal of Ecology, 108 (5):1967–1977, 2020.

Julien Moeys. The soil texture wizard: R functions for plotting, classifying, transforming and exploring soil texture data. CRAN. R-Project, 2018.

Christoph Molnar. Interpretable Machine Learning. Lulu. com, 2020.

Raphaëlle Momal, Stéphane Robin, and Christophe Ambroise. Tree-based inference of species interaction network from abundance data. arXiv preprint arXiv:1905.02452, 2019.

Ignacio Morales-Castilla, Miguel G Matias, Dominique Gravel, and Miguel B Araújo. Inferring biotic interactions from proxies. Trends in ecology & evolution, 30(6):347–356, 2015.

Naia Morueta-Holme, Benjamin Blonder, Brody Sandel, Brian J McGill, Robert K Peet, Jeffrey E Ott, Cyrille Violle, Brian J Enquist, Peter M Jørgensen, and Jens-Christian Svenning. A network approach for inferring species associations from co-occurrence data. Ecography, 39(12):1139–1150, 2016.

Tamara Münkemüller and Laure Gallien. Virtualcom: a simulation model for eco-evolutionary community assembly and invasion. Methods in Ecology and Evolution, 6(6):735–743, 2015.

Tamara Münkemüller, Florian C Boucher, Wilfried Thuiller, and Sébastien Lavergne. Phylogenetic niche conservatism–common pitfalls and ways forward. Functional ecology, 29(5):627–639, 2015.

Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

Roger B Nelsen. An introduction to copulas. Springer Science & Business Media, 2007.

Mark EJ Newman. Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23):8577–8582, 2006.

Nhu H Nguyen, Zewei Song, Scott T Bates, Sara Branco, Leho Tedersoo, Jon Menke, Jonathan S Schilling, and Peter G Kennedy. Funguild: an open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecology, 20:241–248, 2016.

Jenni Niku, Francis KC Hui, Sara Taskinen, and David I Warton. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. Methods in Ecology and Evolution, 10(12):2173–2182, 2019.

Anna Norberg, Nerea Abrego, F Guillaume Blanchet, Frederick R Adler, Barbara J Anderson, Jani Anttila, Miguel B Araújo, Tad Dallas, David Dunson, Jane Elith, et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs, 89(3):e01370, 2019.

Jill Nugent. inaturalist: citizen science for 21st-century naturalists. Science Scope, 41(7):12, 2018.

Visa Nuutinen, Jyrki Pitkänen, Eeva Kuusela, Timo Widbom, and Hanne Lohilahti. Spatial variation of an earthworm community related to soil properties and yield in a grass–clover field. Applied Soil Ecology, 8 (1-3):85–94, 1998.

Marc Ohlmann, Florent Mazel, Loïc Chalmandrier, Stéphane Bec, Eric Coissac, Ludovic Gielly, Johan Pansu, Vincent Schilling, Pierre Taberlet, Lucie Zinger, et al. Mapping the imprint of biotic interactions on β-diversity. Ecology letters, 21(11):1660–1669, 2018.

Otso Ovaskainen and Nerea Abrego. Joint Species Distribution Modelling: With Applications in R. Cambridge University Press, 2020.

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. Ecology Letters, 20(5):561–576, 2017.

Juliane Palm, N Loes MB van Schaik, and Boris Schröder. Modelling distribution patterns of anecic, epigeic and endogeic earthworms at catchment-scale in agro-ecosystems. Pedobiologia, 56(1):23–31, 2013.

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

Panos Panagos, Marc Van Liedekerke, Arwyn Jones, and Luca Montanarella. European soil data centre: Response to european policy support and public data requirements. Land use policy, 29(2):329–338, 2012.

Emmanuel Paradis. Analysis of Phylogenetics and Evolution with R. Springer Science & Business Media, 2011.

Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. Bioinformatics, 35(3):526–528, 2019.

Mercedes Pascual, Jennifer A Dunne, et al. Ecological networks: linking structure to dynamics in food webs. Oxford University Press, 2006.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems, pages 8026–8037, 2019.

Céline Pelosi, Benjamin Pey, Mickaël Hedde, Gaël Caro, Yvan Capowiez, Muriel Guernion, J Peigné, Denis Piron, M Bertrand, and Daniel Cluzeau. Reducing tillage in cultivated fields increases earthworm functional diversity. Applied Soil Ecology, 83:79–87, 2014.

Jose M Peña. Reading dependencies from covariance graphs. International journal of approximate reasoning, 54(1):216–227, 2013.

A Townsend Peterson, Jorge Soberón, Richard G Pearson, Robert P Anderson, Enrique Martínez-Meyer, Miguel Nakamura, and Miguel B Araújo. Ecological niches and geographic distributions (MPB-49), volume 56. Princeton University Press, 2011.

A Townsend Peterson, Jorge Soberón, Janine Ramsey, and Luis Osorio-Olvera. Co-occurrence networks do not support identification of biotic interactions. Biodiversity Informatics, 15(1):1–10, 2020.

Helen RP Phillips, Carlos A Guerra, Marie LC Bartz, Maria JI Briones, George Brown, Thomas W Crowther, Olga Ferlian, Konstantin B Gongalsky, Johan Van Den Hoogen, Julia Krebs, et al. Global distribution of earthworm diversity. Science, 366(6464):480–485, 2019.

Helen RP Phillips, Léa Beaumelle, Nico Eisenhauer, Jes Hines, and Linnea C Smith. Lessons from the wbf2020: extrinsic and intrinsic value of soil organisms. Soil organisms, 92(2):121, 2020.

Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. Ecological modelling, 190(3-4):231–259, 2006.

Maximilian Pichler and Florian Hartig. A new method for faster and more accurate inference of species associations from novel community data. arXiv preprint arXiv:2003.05331, 2020.

Maximilian Pichler, Virginie Boreux, Alexandra-Maria Klein, Matthias Schleuning, and Florian Hartig. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. Methods in Ecology and Evolution, 11(2):281–293, 2020.

Shai Pilosof, Miguel A Fortuna, Jean-Francois Cosson, Maxime Galan, Chaisiri Kittipong, Alexis Ribas, Eran Segal, Boris R Krasnov, Serge Morand, and Jordi Bascompte. Host–parasite network structure is associated with community-level immunogenetic diversity. Nature communications, 5:5172, 2014.

Jorrit H Poelen, James D Simons, and Chris J Mungall. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics, 24:148–159, 2014.

Giovanni Poggiato, Tamara Münkemüller, Daria Bystrova, Julyan Arbel, James S Clark, and Wilfried Thuiller. On the interpretations of joint modelling in community ecology. Trends in Ecology and Evolution, in revision.

Timothée Poisot, Daniel B Stouffer, and Sonia Kéfi. Describe, understand and predict: why do we need networks in ecology? Functional Ecology, 30(12):1878–1882, 2016.

Laura J Pollock, William K Morris, and Peter A Vesk. The role of functional traits in species distributions revealed through a hierarchical model. Ecography, 35(8):716–725, 2012.

Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O'Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). Methods in Ecology and Evolution, 5(5):397–406, 2014.

Gordana C Popovic, Francis KC Hui, and David I Warton. A general algorithm for covariance modeling of discrete data. Journal of Multivariate Analysis, 165:86–100, 2018.

Gordana C Popovic, David I Warton, Fiona J Thomson, Francis KC Hui, and Angela T Moles. Untangling direct species associations from indirect mediator species effects with graphical models. Methods in Ecology and Evolution, 10(9):1571–1583, 2019.

Martin Potthoff, Norbert Asche, Benjamin Stein, Annette Muhs, and Friedrich Beese. Earthworm communities in temperate beech wood forest soils affected by liming. European Journal of Soil Biology, 44(2): 247–254, 2008.

Mirjam Pulleman, Rachel Creamer, Ute Hamer, Johannes Helder, Céline Pelosi, Guenola Peres, and Michiel Rutgers. Soil biodiversity, biological indicators and soil ecosystem services—an overview of european approaches. Current Opinion in Environmental Sustainability, 4(5):529–538, 2012.

MM Pulleman, J Six, A Uyl, JCY Marinissen, and AG Jongmans. Earthworms and management affect organic matter incorporation and microaggregate formation in agricultural soils. Applied Soil Ecology, 29 (1):1–15, 2005.

H Ronald Pulliam. Sources, sinks, and population regulation. The American Naturalist, 132(5):652–661, 1988.

JP Qiu and MB Bouché. Zophoscolex, un nouveau genre de lumbricidae (annelida: Oligochaeta) d'iberie et de france. Documents Pédozoologiques et Intégrologiques, 4:20–36, 1998.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

Benoit Richard, Marc Legras, Pierre Margerie, Jerôme Mathieu, Sébastien Barot, Gaël Caro, Thierry Desjardins, Florence Dubs, Lise Dupont, and Thibaud Decaëns. Spatial organization of earthworm assemblages in pastures of northwestern france. European Journal of Soil Biology, 53:62–69, 2012.

John Rogan and DongMei Chen. Remote sensing technology for mapping and monitoring land-cover and land-use change. Progress in planning, 61(4):301–325, 2004.

Rudolf P Rohr, Russell E Naisbit, Christian Mazza, and Louis-Félix Bersier. Matching–centrality decomposition and the forecasting of new links in networks. Proceedings of the Royal Society B: Biological Sciences, 283(1824):20152702, 2016.

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. arXiv preprint arXiv:1711.01239, 2017.

EMILIA Rota, SVANTE Martinsson, CHRISTER ErsÉus, VALENTIN N Petushkov, NATALJA S Rodionova, and PIETRO Omodeo. Green light to an integrative view of microscolex phosphoreus (dugès, 1837)(annelida: Clitellata: Acanthodrilidae). Zootaxa, 4496(1):175–189, 2018.

Sebastian Ruder. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 4822–4829, 2019.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. nature, 323(6088):533–536, 1986.

Michiel Rutgers, Alberto Orgiazzi, Ciro Gardi, Jörg Römbke, Stephan Jänsch, Aidan M Keith, Roy Neilson, Brian Boag, Olaf Schmidt, Archie K Murchie, et al. Mapping earthworm communities in europe. Applied Soil Ecology, 97:98–111, 2016.

Elizabeth L Sander, J Timothy Wootton, and Stefano Allesina. Ecological network inference from long-term presence-absence data. Scientific Reports, 7(1):1–12, 2017.

James G Sanderson and Stuart L Pimm. Patterns in Nature: The analysis of species co-occurrences. University of Chicago Press, 2015.

Nicolas Sauvion, Paul-André Calatayud, Denis Thiéry, and Frédéric Marion-Poll. Interactions insectes-plantes. Editions Quae, 2013.

Stephanie Schelfhout, Jan Mertens, Kris Verheyen, Lars Vesterdal, Lander Baeten, Bart Muys, and An De Schrijver. Tree species identity shapes earthworm communities. Forests, 8(3):85, 2017.

Michael Schirrmann, Monika Joschko, Robin Gebbers, Eckart Kramer, Mirjam Zörner, Dietmar Barkusky, and Jens Timmer. Proximal soil sensing–a contribution for species habitat distribution modelling of earthworms in agricultural soils? PloS one, 11(6):e0158271, 2016.

Matthias Schleuning, Jochen Fründ, and Daniel García. Predicting ecosystem functions from biodiversity and mutualistic networks: an extension of trait-based concepts to plant–animal interactions. Ecography, 38(4):380–392, 2015.

Matthias Schleuning, Eike Lena Neuschulz, Jörg Albrecht, Irene MA Bender, Diana E Bowler, D Matthias Dehling, Susanne A Fritz, Christian Hof, Thomas Mueller, Larissa Nowak, et al. Trait-based assessments of climate-change impacts on interacting species. Trends in Ecology & Evolution, 35(4):319–328, 2020.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. arXiv preprint arXiv:1906.07789, 2019.

Sylvain Schmitt, Robin Pouteau, Dimitri Justeau, Florian de Boissieu, and Philippe Birnbaum. ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models. Methods in Ecology and Evolution, 8(12):1795–1803, 2017.

A-K Schneider, L van Schaik, A Zangerlé, JA Eccard, and B Schröder. Which abiotic filters shape earthworm distribution patterns at the catchment scale? European Journal of Soil Science, 67(4):431–442, 2016.

Anne-Kathrin Schneider and Boris Schröder. Perspectives in modelling earthworm dynamics and their feedbacks with abiotic soil properties. Applied soil ecology, 58:29–36, 2012.

Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive Computation and Machine Learning series, 2018.

Oliver Schweiger, Risto K Heikkinen, Alexander Harpke, Thomas Hickler, Stefan Klotz, Otakar Kudrna, Ingolf Kühn, Juha Pöyry, and Josef Settele. Increasing range mismatching of interacting species under global change is related to their ecological characteristics. Global Ecology and Biogeography, 21(1):88–99, 2012.

Marco Scutari, Maintainer Marco Scutari, and Hiton-PC MMPC. Package 'bnlearn'. Bayesian network structure learning, parameter learning and inference, R package version 4.4, 1, 2019.

Claude E Shannon and Warren Weaver. A mathematical model of communication. Urbana, IL: University of Illinois Press, 11, 1949.

Lloyd S Shapley. A value for n-person games. Contributions to the Theory of Games, 2(28):307–317, 1953.

Lindsey M Shartell, Erik A Lilleskov, and Andrew J Storer. Predicting exotic earthworm distribution in the northern great lakes region. Biological invasions, 15(8):1665–1675, 2013.

Sergei V Shekhovtsov, Nikita I Ershov, Gennady V Vasiliev, and Sergey E Peltek. Transcriptomic analysis confirms differences among nuclear genomes of cryptic earthworm lineages living in sympatry. BMC evolutionary biology, 19(1):50, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

Jaswinder Singh, Martin Schädler, Wilian Demetrio, George G Brown, and Nico Eisenhauer. Climate change effects on earthworms-a review. Soil organisms, 91(3):114, 2019.

Jorge Soberón and Miguel Nakamura. Niches and distributional areas: concepts, methods, and assumptions. Proceedings of the National Academy of Sciences, 106(Supplement 2):19644–19650, 2009.

Nadejda A Soudzilovskaia, Stijn Vaessen, Milagros Barcelo, Jinhong He, Saleh Rahimlou, Kessy Abarenkov, Mark C Brundrett, Sofia IF Gomes, Vincent Merckx, and Leho Tedersoo. Fungalroot: global online database of plant mycorrhizal associations. New Phytologist, 2020.

Tomas Soukup, György Büttner, Jan Feranec, Gerard Hazeu, Gabriel Jaffrain, Marketa Jindrova, Miroslav Kopecky, and Erika Orlitova. Corine land cover 2012 (clc2012): Analysis and assessment. European landscape dynamics. CORINE Land Cover data. CRC/Taylor & Francis, Boca Raton, Florida. Pp, pages 93–98, 2016.

DJ Spurgeon and SP Hopkin. Effects of variations of the organic matter content and ph of soils on the availability and toxicity of zinc to the earthworm eisenia fetida. Pedobiologia, 40(1):80–96, 1996.

Phillip PA Staniczenko, Prabu Sivasubramaniam, K Blake Suttle, and Richard G Pearson. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. Ecology letters, 20(6):693–707, 2017.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41(3):647–665, 2014.

Katharine N Suding, Sandra Lavorel, FS Chapin Iii, Johannes HC Cornelissen, Sandra DIAz, Eric Garnier, Deborah Goldberg, David U Hooper, Stephen T Jackson, and Marie-Laure Navas. Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. Global Change Biology, 14(5):1125–1140, 2008.

Suwon Suh and Seungjin Choi. Gaussian copula variational autoencoders for mixed data. arXiv preprint arXiv:1604.04960, 2016.

Pierre Taberlet, Aurélie Bonin, Eric Coissac, and Lucie Zinger. Environmental DNA: For biodiversity research and monitoring. Oxford University Press, 2018.

J Christopher D Terry and Owen T Lewis. Finding missing links in interaction networks. Ecology, 101(7): e03047, 2020.

Ross M Thompson, Ulrich Brose, Jennifer A Dunne, Robert O Hall Jr, Sally Hladyz, Roger L Kitching, Neo D Martinez, Heidi Rantala, Tamara N Romanuk, Daniel B Stouffer, et al. Food webs: reconciling the structure and function of biodiversity. Trends in ecology & evolution, 27(12):689–697, 2012.

Wilfried Thuiller, Bruno Lafourcade, Robin Engler, and Miguel B Araújo. Biomod–a platform for ensemble forecasting of species distributions. Ecography, 32(3):369–373, 2009.

Wilfried Thuiller, Tamara Münkemüller, Sébastien Lavergne, David Mouillot, Nicolas Mouquet, Katja Schiffers, and Dominique Gravel. A road map for integrating eco-evolutionary processes into biodiversity models. Ecology letters, 16:94–105, 2013.

Wilfried Thuiller, Laura J Pollock, Maya Gueguen, and Tamara Münkemüller. From species distributions to meta-communities. Ecology letters, 18(12):1321–1328, 2015.

Wilfried Thuiller, Damien Georges, Robin Engler, Frank Breiner, Maintainer Damien Georges, and Contact Wilfried Thuiller. Package 'biomod2'. Species distribution modeling within an ensemble forecasting framework. software, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2): 411–423, 2001.

Andrey N Tikhonov and Vasiliy Y Arsenin. Solutions of ill-posed problems. New York, pages 1–30, 1977.

Sérgio Timóteo, Marta Correia, Susana Rodríguez-Echeverría, Helena Freitas, and Ruben Heleno. Multilayer networks reveal the spatial structure of seed-dispersal interactions across the great rift landscapes. Nature communications, 9(1):140, 2018.

Lisa Torrey and Jude Shavlik. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pages 242–264. IGI global, 2010.

Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner. Bayesian layers: A module for neural network uncertainty. In Advances in Neural Information Processing Systems, pages 14660–14672, 2019.

Neda Trifonova, Andrew Kenny, David Maxwell, Daniel Duplisea, Jose Fernandes, and Allan Tucker. Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. Ecological informatics, 30:142–158, 2015.

Werner Ulrich and Nicholas J Gotelli. Pattern detection in null model analysis. Oikos, 122(1):2–18, 2013.

AV Uvarov, K Ilieva-Makulec, K Karaban, NS Yakovenko, and J Uchmański. Effects of intra-and interspecific interactions in earthworm assemblages: A comparative study. Biology Bulletin, 46(5):475–482, 2019.

Corinne Vacher, Alireza Tamaddoni-Nezhad, Stefaniya Kamenova, Nathalie Peyrard, Yann Moalic, Regis Sabbadin, Loïc Schwaller, Julien Chiquet, M Alex Smith, Jessica Vallance, et al. Learning ecological networks from next-generation sequencing data. In Advances in Ecological Research, volume 54, pages 1–39. Elsevier, 2016.

Loes van Schaik, Juliane Palm, Julian Klaus, Erwin Zehe, and Boris Schröder. Linking spatial earthworm distribution to macropore numbers and hydrological effectiveness. Ecohydrology, 7(2):401–408, 2014.

Diego P Vázquez, Nico Blüthgen, Luciano Cagnolo, and Natacha P Chacoff. Uniting pattern and process in plant–animal mutualistic networks: a review. Annals of botany, 103(9):1445–1457, 2009.

Cyrille Violle, Marie-Laure Navas, Denis Vile, Elena Kazakou, Claire Fortunel, Irène Hummel, and Eric Garnier. Let the concept of trait be functional! Oikos, 116(5):882–892, 2007.

Cyrille Violle, Wilfried Thuiller, Nicolas Mouquet, François Munoz, Nathan JB Kraft, Marc W Cadotte, Stuart W Livingstone, and David Mouillot. Functional rarity: the ecology of outliers. Trends in Ecology & Evolution, 32(5):356–367, 2017.

Steve Vissault, Dominique Gravel, and Timothée Poisot. Mangal: An open infrastructure for ecological interactions. Biodiversity Information Science and Standards, 2019.

Brian H Walker. Biodiversity and ecological redundancy. Conservation biology, 6(1):18–23, 1992.

Prince Zizhuang Wang and William Yang Wang. Neural gaussian copula for variational autoencoder. arXiv preprint arXiv:1909.03569, 2019.

DA Wardle, GM Barker, KI Bonner, and KS Nicholson. Can comparative approaches based on plant ecophysiological traits predict the nature of biotic interactions and individual plant species effects in ecosystems? Journal of ecology, 86(3):405–420, 1998.

Dan L Warren, Richard E Glor, and Michael Turelli. Enmtools: a toolbox for comparative studies of environmental niche models. Ecography, 33(3):607–611, 2010.

David I Warton. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. Biometrics, 67(1):116–123, 2011.

David I Warton, F Guillaume Blanchet, Robert B O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C Walker, and Francis KC Hui. So many variables: joint modeling in community ecology. Trends in Ecology & Evolution, 30(12):766–779, 2015.

Evan Weiher and Paul Keddy. Ecological assembly rules: perspectives, advances, retreats. Cambridge University Press, 2001.

Hannah J White, Lupe León-Sánchez, Victoria J Burton, Erin K Cameron, Tancredi Caruso, Luís Cunha, Tara Dirilgen, Stephanie D Jurburg, Ruth Kelly, Deepak Kumaresan, et al. Methods and approaches to advance soil macroecology. Global Ecology and Biogeography, 29(10):1674–1690, 2020.

Robert H Whittaker. Evolution and measurement of species diversity. Taxon, 21(2-3):213–251, 1972.

Thorsten Wiegand and Kirk A Moloney. Handbook of spatial point-pattern analysis in ecology. CRC press, 2013.

Frank Wilcoxon. Individual comparisons by ranking methods. In Breakthroughs in statistics, pages 196–202. Springer, 1992.

David P Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A McCarthy. Defining and evaluating predictions of joint species distribution models. Methods in Ecology and Evolution.

David P Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A McCarthy. A comparison of joint species distribution models for presence–absence data. Methods in Ecology and Evolution, 10(2):198–211, 2019.

Thaïs Winsome, Lynn Epstein, Paul F Hendrix, and William R Horwath. Competitive interactions between native and exotic earthworm species as influenced by habitat quality in a california grassland. Applied Soil Ecology, 32(1):38–53, 2006.

Mary Susanne Wisz, Julien Pottier, W Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F Damgaard, Carsten F Dormann, Mads C Forchhammer, John-Arvid Grytnes, Antoine Guisan, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological reviews, 88(1):15–30, 2013.

Curtis E Woodcock, Richard Allen, Martha Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, Samuel N Goward, Dennis Helder, Eileen Helmer, et al. Free access to landsat imagery. SCIENCE VOL 320: 1011, 2008.

Guy Woodward, Jonathan P Benstead, Oliver S Beveridge, Julia Blanchard, Thomas Brey, Lee E Brown, Wyatt F Cross, Nikolai Friberg, Thomas C Ings, Ute Jacob, et al. Ecological networks in a changing climate. In Advances in ecological research, volume 42, pages 71–138. Elsevier, 2010.

J Timothy Wootton. The nature and consequences of indirect effects in ecological communities. Annual review of ecology and systematics, pages 443–466, 1994.

Xitong Yang. Understanding the variational lower bound, 2017.

Chongliang Zhang, Yong Chen, Binduo Xu, Ying Xue, and Yiping Ren. Improving prediction of rare species' distribution from community data. Scientific reports, 10(1):1–9, 2020.

Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. D-vae: A variational autoencoder for directed acyclic graphs. In Advances in Neural Information Processing Systems, pages 1588–1600, 2019.

Lara Zirbes, Philippe Thonart, and Eric Haubruge. Microsale interactions between earthworms and microorganisms, a review. Biotechnologie, Agronomie, Société et Environnement, 16(1):125–131, 2012.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2):301–320, 2005.

Damaris Zurell, Laura J Pollock, and Wilfried Thuiller. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? Ecography, 41(11):1812–1819, 2018.

Damaris Zurell, Janet Franklin, Christian König, Phil J Bouchet, Carsten F Dormann, Jane Elith, Guillermo Fandos, Xiao Feng, Gurutzeta Guillera-Arroita, Antoine Guisan, et al. A standard protocol for reporting species distribution models. Ecography, 2020a.

Damaris Zurell, Niklaus E Zimmermann, Helge Gross, Andri Baltensweiler, Thomas Sattler, and Rafael O Wüest. Testing species assemblage predictions from stacked and joint species distribution models. Journal of Biogeography, 47(1):101–113, 2020b.