**Titre : Prédiction de l'abondance, la biomasse et la richesse totale des vers de terre   
en France métropolitaine : un outil pour la surveillance biologique du territoire et la conservation de la biodiversité des sols.**

# Introduction

La biodiversité des écosystèmes terrestres est majoritairement constituée par la faune vivant dans le sol (FAO, 2020; Anthony et al., 2023). Cette faune du sol fournit une large gamme de services écosystémiques (Bardgett & Van der Putten, 2014; FAO, 2020) dont la plupart résulte du processus conduit par les groupes fonctionnels de différentes communautés biologiques tels que les vers de terre (Gardi & Jeffery, 2009; Turbé et al., 2010). Les vers de terre sont qualifiés comme étant les « ingénieurs des écosystèmes » car ils participent au développement de la structure du sol (Sharma et al., 2017; Edwards & Arancon, 2022; Lavelle et al., 1997), à l'infiltration de l'eau à travers les galeries verticales et horizontales creusés dans le sol (Capowiez et al., 2014; Cunha et al., 2016) et à la minéralisation des nutriments grâce à la dégradation de la matière organique (Van Groenigen et al., 2019). Ces travaux ont été justifiés par la prise de conscience croissante que les changements dans la composition des vers de terre entraîneront une altération des services écosystémiques fournis par les lombriciens (Cardinale et al., 2012; Hooper et al., 2012; Van Groenigen et al., 2014).

De nombreuse études ont examiné les effets des facteurs anthropiques et environnementaux sur les vers de terre à des échelles locales (Pelosi, Pey, et al., 2014; Marchán et al., 2015; Gabriac et al., 2022) et régionales (Marchán et al., 2016, 2021; Marchán & Domínguez, 2022; Diallo et al., 2023). Cependant, peu d’études se sont concentrées sur les effets des facteurs environnementaux et anthropiques qui influencent la biodiversité et la répartition des vers de terre à des échelles plus larges, c’est-à-dire suprarégionale ou nationale (Fourcade & Vercauteren, 2022; Salako et al., 2023; Zeiss et al., 2024). Les raisons qui peuvent expliquer ce manque de connaissances sont la faible disponibilité des données à l’échelle des pays ou des continents, les incohérences taxonomiques et les difficultés pour fusionner les bases de données existantes (Rutgers et al. 2016). Le premier travail réalisés à une échelle continentale a été effectué par Rutgers et al. (2016), qui ont cartographie la communauté de vers de terre sur 3838 sites échantillonnés dans 8 pays d’Europe. Ils ont observé que l’abondance et la richesse des vers de terre été affecté par l’occupation du sol, les propriétés du sol (pH, matières organiques et textures) et la latitude. Récemment, une autre étude réalisée par Phillips et al. (2019) sur 9212 sites répartis dans 57 pays a démontré que les facteurs climatiques et le type d’habitat sont les filtres environnementaux les plus importants dans l’altération des paramètres lombriciens. Toutefois, leurs études sont limitées par l’utilisation d’un seul type d’algorithme de modélisation prédictive : des modèles linéaires généralisés pour Rutgers et al. (2016) et des modèles linéaires généralisés à effets mixtes pour Phillips et al. (2019). Or, il serait pertinent de comparer plusieurs algorithmes afin de trouver le meilleur modèle pour chaque paramètre des vers de terre (abondance, biomasse et richesse).

En France, la plus récente étude a été réalisée sur 1366 sites par Fourcade and Vercauteren (2022), qui ont utilisé des arbres de régression boostés pour construire des prévisions spatiales de la diversité fonctionnelle avec 44 espèces de vers de terre. Leurs études montrent une diminution de la richesse fonctionnelle entre 1960 et 2012. De plus, leurs modèles prédisent que cette réduction de la richesse fonctionnelle pourrait se poursuivre à l'avenir selon différentes périodes temporelles et scénarios de changement climatique. Cependant, l'étude de Fourcade and Vercauteren (2022) repose sur des données de présence/absence des espèces de vers de terre en France, collectées dans les années 1960 par Bouché (1972). Ces données ne reflètent plus l'assemblage actuel des vers de terre en France. Par ailleurs, l'étude de la communauté des vers de terre doit inclure des variables décrivant la richesse en espèces, l'abondance et la biomasse totale, qui sont des paramètres essentiels pour les évaluations de la biodiversité. Utiliser seulement la diversité fonctionnelle n'est pas suffisant pour évaluer pleinement la biodiversité, car, par exemple, deux communautés peuvent être identiques en termes de diversité, mais différer en termes de densité (Groves, 2022).

Pour comprendre la biodiversité, sa répartition et les facteurs environnementale influençant cette répartition, plusieurs outils ont été développés ces dernières années. On distingue les modèles de distribution des espèces « species distribution modèles (SDMs) » qui sont des modèles basés sur les niches écologiques « Niche-based-modèles » et permet de modéliser les corrélations entre les espèces ou les communautés et leur environnement (Elith & Leathwick, 2009; Guisan et al., 2017). Néanmoins, ce type de modèle exige d’une part, des données sur les paramètres qui caractérise les communautés et d’autre part des données environnementales spatialisées. Divers SDMs existent, chacun présentant des avantages et des inconvénients (Li & Wang, 2013; Valavi et al., 2021). L’approche comparative de plusieurs SDMs pour prédire l’aire de répartition géographique et la biodiversité des vers de terre en Allemagne, telle que proposée par Salako et al. (2023), semble être une bonne approche.

L'objectif général de notre étude était de développer des modèles pour prédire la communauté des vers de terre en France métropolitaine en utilisant une approche comparative de différents algorithmes de modélisation. Plus précisément, nous avons cherché à répondre aux objectifs suivants : (i) quantifier et hiérarchiser l'influence des facteurs environnementaux (occupation du sol, propriétés du sol, localisation, et données climatiques) en fonction de leur contribution aux modèles prédictifs sur l'abondance (individu par m²), la biomasse (g par m²) et la richesse taxonomique totale des vers de terre (nombre de taxons dans la parcelle) en France métropolitaine (hors Corse) et (ii) prédire et cartographier ces mêmes paramètres lombriciens en fonction des facteurs environnementaux. Pour répondre à ces objectifs, nous avons sélectionné cinq méthodes ou algorithmes de modélisation SDM et comparé leurs performances prédictives. Cette approche comparative permet de réduire l'incertitude associée aux prédictions et de trouver le meilleur modèle pour chaque variable des vers de terre (abondance, biomasse et richesse). Les algorithmes sélectionnés étaient les modèles linéaires généralisés (GLM), les modèles additifs généralisés (GAM), les forêts aléatoires (RF), les modèles de régression boostée généralisée (GBM) et les réseaux de neurones artificiels (ANN). Ces algorithmes ont été choisis sur la base de leur classification en algorithmes de régression (GLM et GAM) et d’apprentissage automatique (RF, GBM et ANN), ainsi que de leur large utilisation dans les études récentes (Fourcade & Vercauteren, 2022; Li & Wang, 2013; Rutgers et al., 2016; Salako et al., 2023; Valavi et al., 2021).

Nos hypothèses reposent sur le fait que les variables climatiques (précipitation et température) et le type d'occupation du sol seraient les facteurs les plus influençant sur l'abondance et la biomasse des vers de terre, et que le type d'occupation du sol et la texture du sol influenceraient davantage la richesse en espèces.

# Matériels & méthodes

## Zone d'étude et collecte des données des vers de terre

La zone de notre étude couvre l’ensemble du territoire de France métropolitaine (hors Corse) (Fig. 1). Nous avons utilisé 3758 observations de vers de terre de la base de données LandWorm (2023-2026 FRB-MTE-OFB), qui comprend les communautés de vers de terre à travers la France métropolitaine dans des conditions variées d'utilisation des sols (Détails ? annexe ?). Les échantillonnages des vers de terre ont été effectué entre 1990 et 2023 et ont été réalisés principalement au printemps qui correspondent à la période d’activité maximale des vers de terre. Les méthodes d'échantillonnage des communautés de vers de terre différaient selon les sources de provenance des données (Détails ? annexe ?). Les méthodes d'échantillonnage étaient un tri manuel (ISO 23611-1 :2018) et/ou l'application d'un expulseur chimique (formaldéhyde ou isothiocyanate d’allyle ; ISO 23611-1 :2006). Dans chaque site, les vers de terre ont été collectés, triés à la main et stocké dans de l’éthanol. Au laboratoire, les vers de terre ont été comptés, parfois pesés, et identifiés au niveau de l'espèce/taxon. Cependant, 48 % des parcelles n’avait pas de biomasse totale. De cette base de données nous avons aussi récupéré les coordonnées GPS et le Corine Land Cover (couverture du sol) renseignés par les gestionnaires « des projets » (https://land.copernicus.eu/pan-european/corine-land-cover).

Une image contenant texte, diagramme, carte

Description générée automatiquement

#### **Fig. 1 :** Carte de la zone d'étude (France métropolitaine (hors Corse)) montrant la localisation des sites d'échantillonnage de vers de terre.

## Collecte des données environnementales

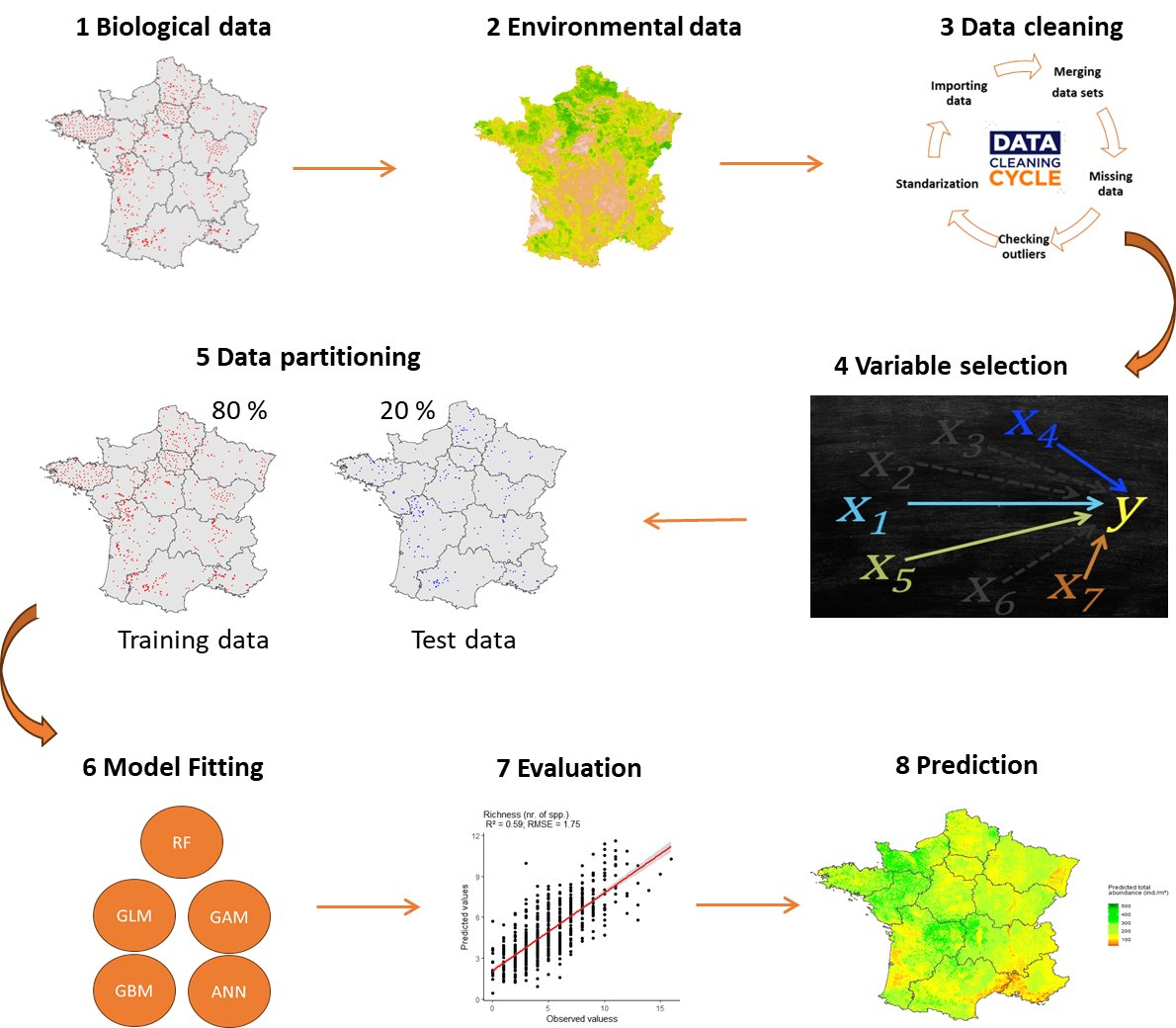
En plus des données sur l’occupation du sol et des coordonnées GPS obtenues avec le base de données LandWorm (2023-2026 FRB-MTE-OFB), nous avons utilisé des variables externes connues pour avoir des effets sur les vers de terre (Rutgers et al., 2016; Phillips et al., 2019; Salako et al., 2023). Pour les variables climatiques, nous avons utilisé les 19 variables bioclimatiques standards du projet CHELSA (Karger et al., 2017). Ces données climatiques représentaient des valeurs moyennes entre 1981 et 2010 à une résolution de 30 secondes d'arc. Cette plage temporelle correspondait le mieux à la période principale des données biologiques.

Pour les variables pédologiques, nous avons d’abord utilisé l’entrepôt de Recherche Data Gouv (<https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/N4E4NE>; Roman Dobarco et al., 2022) pour récupérer les teneurs en sables, argiles et limons. Ces trois variables, disponibles à une résolution de 90 m, provenaient de données de 2022 et étaient fournies à différentes profondeurs. Nous avons choisi d’utiliser les trois premières couches (0 à 5 cm, 5 à 15 cm et 15 à 30 cm) puis de faire la moyenne pour obtenir des données de 0 à 30 cm. Ce choix a été fait pour tenir compte de la variation de l’habitat des vers de terre. Par la suite, nous avons utilisé la base de donnée Joint Research Centre (JRC; Ballabio et al., 2019) pour complété la texture du sol par les variables suivantes : la capacité d'échange de cations (CEC), le carbonate de calcium (CaCO₃), le rapport C/N, l’azote (N), le phosphore (P), le potassium (K), et le pH du sol dans H2O. Ces variables étaient disponibles à une résolution de 250 m à l'échelle européenne et basées sur les données LUCAS 2009/2012 (Land Use and Cover Area frame Survey) sur les sols arables. Nous avons initialement référencé toutes les variables présélectionnées au système de coordonnées géographiques mondial (WGS84), les avons recadrées et masquées pour correspondre aux limites géographiques de France métropolitaine (hors Corse). Pour homogénéiser les variables et les faire correspondre à la même résolution, nous avons rééchantillonnées ou désagrégées les variables à une résolution de 30 secondes d'arc, ce qui correspond à environ 800 m en France. Ces étapes ont été réalisées sous Python avec le module GDAL (https://pypi.org/project/GDAL/).

## Stratégie de modélisation

### Protocole ODMAP

La stratégie de modélisation de notre étude a suivi le protocole « ODMAP » comme recommandé par Zurell et al. (2020) et l’ensemble des étapes sont détaillée dans l’annexe XX. Les différentes étapes sont représentées à la Fig. 2. Toutes les étapes de modélisation ont été effectuées avec le logiciel R version 4.3.1, 2023 (R Core Team, 2023). Initialement, toutes les variables explicatives (numérique) ont été centrées réduites. De plus, l’abondance et la biomasse totale ont été transformées par une racine carrée pour se rapprocher d’une distribution gaussienne.



#### **Fig. 2:**Stratégie de modélisation selon le protocole ODMAP : (1) collecte des données biologiques, (2) collecte des données environnementales, (3) Nettoyage des données, (4) sélection de variables, (5) partitionnement des données, (6) ajustement des modèles, (7) évaluation des modèles et (8) prédiction et.

### Sélection, importance et effets des variables

Pour chacune des trois variable de réponse des communautés des vers de terre, nous avons ajusté des modèles en forêts aléatoires pour identifier l’importance de chaque variable explicative (Breiman, 2001). Nous avons choisi le modèle en forêt aléatoire car il peut traiter des données non linéaires, accepter les variables corrélées tout en intégrant les interactions des variables qui peuvent améliore les performances du modèle (Breiman, 2001). Pour réduire le nombre de variable et éviter le surajustement (Vaughan et Ormerod, 2005), nous avons identifié les dix variables les plus importantes sur les vers de terre à l’aide d’une procédure de permutation (Fourcade & Vercauteren, 2022; Zeiss et al., 2024). Ainsi, les variables identifiés (Tableau 1) pour la prédiction étaient : l’occupation du sol, la longitude, la latitude, le carbonate de calcium (CaCO₃), l’azote (N), le phosphore (P), la teneur en argile et en limon, l'isothermité (bio3) et les précipitations annuelles moyennes. Par la suite, tous les modèles ajustés ont utilisé ces dix variables sélectionnées. Par ailleurs, nous avons utilisé le package *iml* pour améliorer l’interprétabilité des modèles notamment en explorant l’importance, les effets et les interactions des variables (Casalicchio et al., 2024).

#### **Tableau 1 :** Abréviations et noms des variables utilisés dans la prédiction des communautés de vers de terre. Pour les prédicteurs catégoriels (de type booléen), la somme totale est fournie, résumant l'ensemble total de données avec 3758 observations. Pour les prédicteurs continus, trois percentiles (0,05, 0,5 et 0,95) de l'ensemble de données final sont fournis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Category** | **Abbr** | **Description** | **Units** | **Time scale** | **Reference** |
| **LandCover** | For | Forest (all types) | Boolean | 1990 -2023 | LandWorm database |
| Gua | Green urban areas | Boolean |
| Nag | Natural grasslands | Boolean |
| Nial | Non-irrigated arable land | Boolean |
| Pmo | Pastures, meadows and other permanent grasslands under agricultural use | Boolean |
| Viny | Vineyards | Boolean |
| **Location** | Long | Longitude | WGS84 decimal system |  |
| Lat | Latitude | WGS84 decimal system |
| **Climate** | bio12 | Annual precipitation amount | mm year-1 | 1981 - 2010 | Karger et al. (2017) |
| bio3 | Isothermality | °C |
| **Soil** | Clay | Clay particles | g·kg -1 | 2022 | Roman Dobarco et al. (2022) |
| Silt | Silt particles | g·kg -1 |
| CaCO3 | Calcium carbonates | g·kg−1 | 2009 - 2012 | Ballabio et al. (2019) |
| P | Phosphorus | mg·kg−1 |
| N | Nitrogen | g·kg−1 |

### Calibrages des modèles

Nous avons comparé cinq algorithmes SDM pour prédire les paramètres lombricien (abondance totale, biomasse totale et la richesse taxonomique totale) en utilisant 10 variables explicatives. Les cinq algorithmes étaient : les modèles linéaires généralisés (GLM), les modèles additifs généralisés (GAM), les modèles de forêts aléatoires (RF), les modèles boostés généralisés (GBM) et les réseaux de neurones artificiels (ANN).

Nous avons ajusté les GLM à l’aide de la fonction *glm* du package *stats* avec la formulation suivante :

Où y est la variable de réponse (abondance, biomasse ou la richesse) et les x sont les variables explicatives.

Pour les GAM, nous avons utilisé la fonction *gam* du package *mgcv* (Wood, 2023) avec la formulation suivante :

Où y est la variable de réponse (abondance, biomasse ou la richesse) et les x sont les variables explicatives.

Les modèles de forêts aléatoires ont été ajustés avec la fonction *randomForest* du package *randomForest* (Breiman, 2001) avec la formulation suivante :

Où *var.rep* = position de la colonne de la variable de réponse, mtry = nombre de variables échantillonnées au hasard, ntree = nombre d’arbres et maxnodes = nombre maximum de nœuds terminaux.

Nous avons effectué l’hyperparametisation des modèles en forêt aléatoire par la méthode de recherche par grille avec toutes les combinaisons possibles des paramètres suivants : nombre de variables échantillonnées au hasard = de 2 à 8 par incrément de 1, nombre d’arbres = de 100 à 2000 par incrément de 200 et nombre maximum de nœuds terminaux = de 2 à 15 par incrément de 1. Nous avons retenu parmi l’ensemble des modèles celui qui donnait le R² le plus élevé et la RMSE la plus faible.

Nous avons ajusté les GBM en utilisant la fonction *gbm* du package gbm (Greg et al., 2024) avec la formulation suivante :

Où n.trees = nombre d’arbres, shrinkage = taux d’apprentissage, interaction.depth = profondeur d'interaction et n.minobsinnode = nombre minimum d'observations dans les nœuds terminaux des arbres.

Plusieurs paramètres doivent être choisis dans les GBM pour contrôler la complexité du modèle. Pour sélectionner les paramètres les plus appropriés, nous avons ajusté les modèles par la méthode de recherche par grille avec toutes les combinaisons possibles des paramètres suivants : nombre d'arbres = de 500 à 2000 par incrément de 100, profondeur d'interaction = 1, 3, 5, 6, 8, taux d’apprentissage = 0.01, 0.02, 0.05, 0.001, 0.002, 0.005 et le nombre minimum d'observations dans les nœuds terminaux des arbres = 2 , 5, 10, 20, 30, 50. Nous avons retenu parmi l’ensemble des modèles celui qui donnait le R² le plus élevé et la RMSE la plus faible.

Pour les ANN, nous avons utilisé le package *keras* (Kalinowski et al., 2024) avec une architecture séquentielle. Le modèle était composé d’une couche d’entrée avec un *input\_shape* de 15 correspondants aux 9 variables explicatives plus les 6 levels de l’occupation du sol que nous avons transformés en variables binaires indépendantes. Nous avons introduit trois couches cachées avec respectivement 32, 16 et 8 neurones denses. La dernière couche était composée d’un seul neurone qui correspondait à la variable prédite (abondance, biomasse et richesse). Toutes les couches étaient accompagnées d’une fonction d’activation *ReLU* sauf la dernière couche qui avait une activation linéaire. Nous avons utilisé la méthode de régression avec la fonction de perte quadratique moyenne (MSE) et l’optimiseur *RMSprop* alors que l’erreur absolue moyenne (MAE) a été utilisée pour évaluer les performances du modèle. Pour la compilation, nous avons défini une *epochs* de 100, un *batch\_size* de 64 et une *validation\_split* de 0.2. Pour limiter l’overfitting, nous avons ajouté quatre *dropouts* et nous avons introduit un *callback* du type *EarlyStopping* avec une *patience* 10 pour surveiller la perte sur l’ensemble de validation et restaurer les poids du meilleur modèle.

### Evaluation des modèles

Nous avons évalué les modèles en utilisant la méthode de la validation croisée en attribuant aléatoirement 80 % des données à l’entraînement et 20 % des données à la validation des modèles (Horrigue et al., 2016; Hijmans & Elith, 2019; Salako et al., 2023). Cette méthode a été choisie pour sa simplicité de compréhension et de mise en œuvre, ainsi que pour sa rapidité de compilation. De plus, elle s'est avérée efficace dans notre cas, étant donné que l'ensemble de données était volumineux et que la distribution entre données de l’entraînement et données de validation était similaire (test de Kolmogorov-Smirnov non significatif , p-valeur > 0,05) (Guisan et al., 2017). Les données d’entraînement ont été utilisées pour ajuster les modèles et les données de validation ont été utilisées pour évaluer les performances prédictives des modèles. Etant donnée que tous nos variables de réponse étaient quantitatives, nous avons choisi d’utilisé comme métriques d’évaluation des performances, le coefficient de détermination d'une régression (R²) entre les observations (donnée de validation) et les prédictions ainsi que l’erreur absolue moyenne (MAE) et la racine carrée de l'erreur quadratique moyenne (RMSE « Root Mean Square Error »). L’objectif était donc de maximiser le R² et de minimiser la MAE et la RMSE.

## Prédictions et cartographie des communautés des vers de terre

La prédiction de la communauté des vers de terre a été réalisée en utilisant le meilleur algorithme pour chacun des trois variables des vers de terre (abondance totale, biomasse totale et richesse taxonomique). Dans un premier temps, nous avons effectué un échantillonnage d’une résolution d'environ 800 m sur tout le territoire de la France métropolitaine (hors Corse). Ensuite, pour chacun des points d’échantillonnage, nous avons extrait lorsque cela est possible les valeurs des différentes variables finales incluses dans les modèles à partir des bases de données détaillées dans la section 2.1. Par la suite, nous avons utilisé la fonction *predict* en fournissant le modèle final du meilleur algorithme et les variables explicatives extraites pour prédire les vers de terre. Enfin, nous avons présenté les valeurs prédites sous forme de cartes de la communauté des vers de terre. Pour les zones auxquelles, on ne pouvait pas prédire la communauté des vers de terre, nous avons effectué une interpolation en utilisant la méthode IDW ( Inverse Distance Weighting; Pebesma & Graeler, 2023).

Where : w is the predicted value, d is the distance, x is the unknown point, xi is the n-th know point, ui is th value of the know point, p is the power coefficient (p = 10) and n is the number of sampling points used for interpolation (n = 10). The parameter 𝑝 is the weighting parameter that is applied as an exponent to the distance. A large 𝑝 indicates that nearby points exert a much greater influence on the unsampled location than a distant point.

Pour estimer la biodiversité approximative de la communauté des vers de terre, nous avons superposé la carte de l’abondance totale et la carte de la richesse taxonomique totale (Salako et al., 2023).

# Résultats

## Performance des modèles

Les performances des cinq modèles variaient avec une R² moyenne de 0.33 (± 0.10 sd) pour l’abondance, de 0.28 (± 0.05 sd) pour la biomasse et de 0.48 (± 0.11 sd) pour la richesse taxonomique totale. En termes de RMSE, la moyenne était de 29 ind./m² (± 4.32 sd) pour l’abondance, 9.95 g/m² (± 3.69 sd) pour la biomasse et 1.92 espèce par parcelle (± 0.21 sd) pour la richesse taxonomique totale (Tableau 2). Pour les trois paramètres des vers de terre, c’est le RF et le GBM qui présentaient les R² le plus élevées et la RMSE la plus faible traduisant ainsi une bonne prédiction de la communauté des vers de terre (Tableau 2). L'évaluation des performances de prédiction de l’ensemble des modèles montre que les GLM étaient les modèles les moins performants pour prédire l’abondance, la biomasse et la richesse des vers de terre. La fig. 3 illustre la comparaison entre les valeurs observées (jeu de données de validation) et les valeurs prédites par les meilleurs modèles. Pour l’abondance totale, le meilleur modèle était le RF avec un R² de 0,43. Le meilleur modèle pour la biomasse totale était également le RF avec un R² de 0,35. Pour la richesse totale, les meilleurs modèles étaient le RF et le GBM, avec un R² identique de 0,59.

#### **Tableau 2 :** Mesures de performance de la prédiction sur le jeu de donnée validation pour les différents algorithmes de modèles testés sur les trois variables de réponse de la communauté des vers de terre. Les valeurs en gras indiquent le meilleur algorithme pour chaque variable des vers de terre.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Response variables** | **R²** | **RMSE** |
| GLM | Total  abundance | 0.22 | 34.57 |
| GAM | 0.26 | 33.06 |
| **RF** | **0.43** | **25.20** |
| GBM | 0.43 | 25.30 |
| ANN | 0.35 | 28.94 |
|  |  |  |  |
| GLM | Total  biomass | 0.23 | 10.69 |
| GAM | 0.24 | 10.50 |
| **RF** | **0.35** | **8.76** |
| GBM | 0.32 | 9.30 |
| ANN | 0.27 | 10.50 |
|  |  |  |  |
| GLM | Total  taxonomic  richness | 0.36 | 2.18 |
| GAM | 0.44 | 2.04 |
| **RF** | **0.59** | **1.75** |
| GBM | **0.59** | **1.75** |
| ANN | 0.40 | 2.16 |

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

#### Fig**. 3 :** Prédiction sur le jeu de donnée validation avec les meilleurs algorithmes pour, (a) abondance totale (RF ; ind/m²), (b) biomasse totale (RF ; g/m²), et (c) richesse taxonomique totale (GBM ; nombre de taxons dans la parcelle). L'axe X indique les valeurs observées et l'axe Y les valeurs prédites. La ligne rouge représente la régression linéaire (tendance) entre les valeurs observées et prédites et la bande grise indique l’intervalle de confiance autour de la ligne de régression.

## Contribution des variables aux modèles de prédictions

La variable la plus importante lors de l’ajustement des modèles de prédiction était l’occupation du sol (CLC ; Fig. 4). Lorsque l’occupation du sol était permutée, la RMSE du modèle de l’abondance augmentait en moyenne de 1.68 ind./m², de 1.45 g/m² pour la biomasse et de 1.78 espèce par parcelle pour la richesse des vers de terre. Après l’occupation du sol, les variables spatiales étaient les plus importantes pour la prédiction des vers de terre notamment avec la longitude qui entraînait une augmentation moyenne de la RMSE de 1.63 espèce par parcelle. Pour les variables climatiques, c’est la précipitation annuelle qui avait plus importance sur les vers de terre. Les deux variables du sol les plus influents sur les vers de terre était respectivement le CaCO₃ et l’azote (N). De plus, le CaCO₃ s’est révélé être plus influent sur la richesse et l’abondance totale comparée à la biomasse.

Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement

#### **Fig. 4 :** Importance des variables explicatives environnementales aux prédictions de l’abondance totale, de la biomasse totale et de la richesse taxonomique totale des vers de terre.

## Effets des variables sur les vers de terre

Les analyses des effets locales accumulés (« Accumulated Local Effects » (ALE)) montrent que l’occupation du sol avait des effets mixtes sur la communauté des vers de terre. Les effets ALE décrivent comment les variables explicatives influencent en moyenne la prédiction d'un modèle d'apprentissage automatique. Autrement dit, les effets ALE montrent comment la prédiction change localement, lorsque les variables varient. En effet, les parcelles en prairie (Nag et Pmo) avaient des abondances plus élevées et permettait d’augmenter la prédiction moyenne de l’abondance de 80 ind./m². De même, les parcelles en espaces verts urbains (Gua) augmentaient en moyenne les prédictions de l’abondance de 50 ind./m². Les parcelles en Nag, Pmo et Gua avaient environ une ou deux espèces de plus que les autres occupations de sol. Les effets ALE montrent aussi que les parcelles en vigne (Viny), en forêt (For) et en culture annuelle (Nial) étaient les occupations de sol qui entraînaient les plus faibles abondances. En ce qui concerne la biomasse, seules les prairies permettaient d’augmenter le poids moyen des vers de terre.

Les variables spatiales exerçaient une forte influence sur les vers de terre. Cette influence était mieux représentée par la longitude. De l'Ouest vers l'Est, l'abondance et la richesse ont diminué, avec une baisse d'environ 50 individus/m² et une espèce de moins respectivement. Cependant, la biomasse variait faiblement entre l’Ouest et l’Est de la France. L'augmentation de la latitude n'a pas influencé l'abondance totale entre 43°N et 46°N, puis celle-ci a diminué entre 46°N et 47,5°N avant d'augmenter au-delà de cette dernière latitude.

En ce qui concerne l’effet des variables climatiques, les prédictions de l’abondance et de la biomasse totale ont augmenté en dessous de 700 mm/ande précipitation annuelle, mais reste stable au-dessus de ce seuil. La biomasse des vers de terre a été peu affectée par le rapport entre la variation de la température diurne et la variation annuelle (Isot).

Pour les paramètres du sol, l’augmentation du CaCO₃ jusqu’à environ 10 g·kg−1 entrainait une augmentation des prédictions de 25 ind./m² pour l’abondance et de 15 g/m² pour la biomasse. Cependant, le CaCO₃ au-dessus de 10 g·kg−1 conduisait à une diminution des vers de terre. Le phosphore diminuait faiblement les vers de terre. L’augmentation de l’azote de 2 g·kg−1 entrainait une augmentions de l’abondance de 25 ind.m² et de la biomasse de 5 g/m². La texture du sol (teneur en argile et en limon) n’a pas influencé les prédictions de l’abondance totale jusqu’à 40 g·kg−1, niveau auquel l’abondance a commencé à augmenter pour atteindre une hausse moyenne de 45 ind./m². La texture avait peu d’influence sur la biomasse totale.

## Répartition spatiale des vers de terre à l’échelle de la France

La fig. 5 montre la répartition spatiale prédite de la communauté des vers de terre en France métropolitaine (hors Corse). L’abondance totale variait de 0 à 530 ind./m² avec une moyenne de 192 ind./m² par parcelle (Fig. 5a). La biomasse moyenne était de 72 g/m² (minimum = 0 et maximum = 196 ; Fig. 5b) alors que la richesse taxonomique totale allait de 0 à 10 avec une moyenne de 5 espèces par parcelle (Fig. 5c). La fig. 5d représente la diversité approximative des vers de terre issus de la superposition de la carte de l’abondance et la carte de la richesse taxonomique. On observe que le sud-est du pays présentait une diversité très faible à faible alors que la partie sud-ouest avait une diversité moyenne.

Une image contenant texte, carte, diagramme

Description générée automatiquement

#### Fig. 5 : Distribution spatiale prédite de, (a) abondance totale (ind./m²), (b) biomasse totale (g/m²), (c) richesse taxonomique totale (nombre de taxons dans la parcelle) et (d) la diversité approximative de vers de terre.

# Discussion

Les premiers résultats notables que nous avons obtenus concernent la précision estimée de nos modèles qui permettent de prédire la biomasse totale avec un R² de 0.35 (RMSE = 8.76 g/m²), l’abondance totale avec un R² de 0.43 (RMSE = 25 ind./m²) et la richesse taxonomique totale avec un R² de 0.59 (RMSE = 1.7 espèce par parcelle). Nous avons observé que la répartition spatiale des vers de terre en France est principalement due à l’occupation du sol et aux variables spatiales, notamment avec la longitude. Notre étude a permis de créer des cartes de prédiction de la communauté des vers de terre, fournissant ainsi des données de référence précises et sensibilisant à la protection de la biodiversité du sol et donc des services écosystémiques associés. Toutes les étapes de la stratégie de modélisation ont été rapportées sous forme de protocole ODMAP (ANNEX XX ), comme recommandé par Zurell et al. (2020).

Dans cette étude, nous avons ajusté cinq algorithmes de modèle différents pour modéliser les prédictions de distribution des vers de terre en utilisant des variables de l’occupation du sol, des variables spatiales, du climat et des variables pédologiques.

## Supériorité des méthodes d'ensemble dans la prédiction des communautés de vers de terre par rapport aux modèles de régression traditionnels

Nos résultats montrent que les modèles de forêts aléatoires (RF) et les modèles de boosting généralisés (GBM) prédisaient mieux la communauté des vers de terre comparée aux modèles de régression traditionnels (GLM et GAM) et aux réseaux de neurones artificiels (ANN). Ce résultat est conforme avec ceux trouvés par Li & Wang (2013); Mi et al. (2017); Valavi et al. (2021), qui ont observé que le RF présentait des performances prédictives très élevées. De plus, une étude réalisée à l’échelle de l’Allemagne par Salako et al. (2023) utilisant différents algorithmes de prédiction a conclu que le RF était l’algorithme le plus performant pour prédire la communauté des vers de terre. Les performances élevées de ces algorithmes s’expliquent par le fait que les RF et les GBM sont comme des classificateurs d'ensemble et qui consiste à utiliser plusieurs arbres alternatifs dans la prise de décision lors de la construction de prédictions de modèles (Breiman, 2001; Li & Wang, 2013). Leurs performances s’expliquent aussi par le fait qu’il capture mieux les relations non linéaires entre les variables explicatives et les variables de réponses, ils sont plus robustes aux valeurs aberrantes et ils gèrent aussi mieux les interactions entre les variables (Breiman, 2001). Compte tenue donc du potentiel de prédiction élevée des RF et des GBM, il serait intéressant de les utiliser dans les études écologique et ne pas se limiter à utiliser uniquement les algorithme de régression traditionnelle (Rutgers et al. (2016) Phillips et al. (2019). Cependant, les RF et les GBM demandent des quantités de données importante pour avoir des bonnes performances prédictives (Yiu, 2021). Cela montre donc que les performances des modèles pourront être améliorer à l’avenir si nous arrivons à fusionnée et homogénéisé les diffèrent base de données qui existent entre les unités de recherche mais aussi entre les pays. L’autre limites des RF et des GBM est leurs faibles interprétabilités mais cela semble être de moins en moins vrai car des nombreux outils existent maintenant pour mieux comprendre et interprété les modèles de machine learning. Par exemple, le package « iml » fournit des outils très pertinentes pour analyser n'importe quel modèle d'apprentissage automatique en boîte noire. Le package permet d’explorer l’importance, les effets et les interactions des variables tout en proposant des modèles de substituions (Casalicchio et al., 2024).

Il est important de souligner que les variables explicatives liées au climat et au sol proviennent de bases de données externes, car toutes les parcelles de notre base de données ne contenaient pas ces informations. Cependant, cette dépendance à l’égard de base de données externes constitue une limite, comme déjà mentionné par Rutgers et al. (2016) ainsi que Salako et al. (2023). En effet, la qualité des données utilisées lors de l’entraînement des modèles influence grandement leur performance. Néanmoins, nous avons choisi de ne pas utiliser de bases de données externes pour l’occupation du sol, car ces informations étaient bien renseignées dans notre base de données. Ce choix évite toute rupture temporelle entre les observations des vers de terre et les occupations du sol, qui peuvent évoluer au fil des années. Une autre limite de notre étude réside dans la sélection restreinte des types d’occupation du sol. En effet, nous avons utilisé uniquement six types d’occupation du sol de niveau 3 définies par Corine Land Cover. Les autres types d’occupation du sol n’ont pas été inclus en raison du manque de données disponibles. De plus, nous n’avons pas pu distinguer les forêts de feuillus, les forêts de conifères et les forêts mixtes. Par conséquent, nous avons regroupé ces trois types de forêts sous une seule types d’occupation du sol : “Forêt” (tous types). Cette décision a été prise en tenant compte du fait que l’utilisation des terres a un impact significatif sur les populations de vers de terre, comme l’ont déjà souligné Spurgeon et al. (2013). Nous sommes aussi conscients que notre base de données est déséquilibrée en termes d’échantillonnage avec plus d’observation au nord qu’au sud et cela pourrai expliquer la faible R² pour la biomasse totale d’autant plus certaine observation des vers de terre n’avait de biomasse.

## Importance de l’occupation du sol, du climat et des propriétés du sol

Nos résultats montrent que l’occupation du sol était la variable la plus importante sur les vers de terre. Cela est cohérent avec les résultats de Fourcade & Vercauteren (2022) et de Salako et al. (2023) qui ont montré que l’occupation du sol affecte la distribution de la communauté des vers de terre. De plus, l’étude réalisé par Rutgers et al. (2016) à travers l’Europe a également mis en évidence l’importance de l’occupation du sol. Toutes les modèles de prédiction utilisée dans cette étude prédisaient des effets positifs des prairies (Pmo et Nag) sur les vers de terre. Ce résultat est conforme aux préférences connues en matière d'habitat de vers de terre (Rutgers et al., 2016; Hoeffner et al., 2021). De plus, Rutgers et al. (2016) ont démontré que l’abondance et la diversité des vers de terre étaient étroitement liées à la présence ou à l’absence de certains types de l’occupation du sol tels que les prairies, les terres cultivées, les forêts, les landes et les vignobles. Cela s’explique par le fait que les vers de terre ont une certaine préférence en matière d’habitat. Dans notre cas, nos modèles prédisent un nombre élevé de vers de terre (80 ind./m²), une biomasse moyenne (40 g/m²) et environ une à deux espèces dans les prairies par rapport aux autres occupations du sol. Nos modèles prédisaient aussi un effet négatif des cultures et des vignes. Cela est cohérant car, il est connu que l'intensité de l'utilisation des terres peu avoir un impact important sur les communautés de vers de terre (Smith et al., 2008; Spurgeon et al., 2013).

Notre étude as permis aussi de mettre en évidence que, après l’occupation du sol, les variables spatiales étaient les plus importantes, suivies des variables climatiques, puis des variables pédologiques.

En effet, le CaCO₃ avait des effets positifs sur les vers de terre en dessous de 10 g·kg⁻¹ et des effets négatifs au-delà de ce seuil. Ce résultat pourrait s’expliquer par le fait que de faibles quantités de CaCO₃ favoriserait l’alcalinisation des aliments et faciliterait le passage des aliments au niveau des glandes de Morren, tandis que des quantités élevées de CaCO₃ seraient toxiques pour les vers de terre.

## Influences de l'échelle sur la biodiversité des vers de terre : Facteurs climatiques et pédologiques

Cependant, selon Phillips et al. (2019), les variables les plus influentes sur les vers de terre sont les précipitations et la températures annuelle. En effet, il semblerait que les facteurs déterminants de la biodiversité des vers de terre dépendent de l’échelle (Salako et al., 2023). Les paramètres climatiques joueraient un rôle crucial à l’échelle mondiale et continentale (Rutgers et al., 2016; Phillips et al., 2019), tandis que les facteurs liés au sol deviennent plus importants à des échelles spatiales plus restreintes (Palm et al., 2013; Marchán et al., 2015). Dans notre étude, l’importance du climat était principalement portée par la précipitation annuelle, tandis que le carbonate de calcium (CaCO₃) était la variable du sol la plus significative pour les vers de terre.

Nous sommes conscients qu’il existe d’autres variables non incluses dans cette étude et qui peuvent influencer fortement les vers de terre. C’est le cas par exemple des pratiques agricoles comme le travail du sol (Crittenden et al., 2014; Ernst & Emmerling, 2009; Pelosi, Pey, et al., 2014), les pesticides (Pelosi, Barot, et al., 2014; Maggi & Tang, 2021) et la fertilisation (Leroy et al., 2008; Niswati et al., 2022) qui sont bien connues pour leur impact sur les vers de terre. Cependant, ces variables relatives aux pratiques agricoles ne sont pas disponibles sous forme de cartes à haute résolution à l’échelle de la France métropolitaine. Par conséquent, nous ne les avons pas incluses dans nos modèles et cela empêcherait la production des cartes basées sur les prédictions des modèles. En résumé, bien que nous ne soyons pas en mesure de prendre en compte tous les facteurs potentiels de répartition des vers de terre, notre étude incluent un grand nombre de variables environnementaux les plus importants connus pour affecter les vers de terre (Edwards & Arancon, 2022).

Nos modèles prédisaient une richesse taxonomique moyenne de 5 espèces par parcelle, avec un maximum de 10 espèces. Ces prédictions se situent dans la gamme de valeurs intermédiaires, en accord avec les études antérieures. En effet, Phillips et al. (2019) et Fourcade & Vercauteren (2022) ont rapporté une richesse maximale de 6 espèces par parcelle. Rutgers et al. (2016) ont observé jusqu’à 7 espèces par parcelle, tandis que Zeiss et al. (2024) ont signalé une richesse allant jusqu’à 18 espèces. Nous avons observé que la communauté des vers de terre était plus importante au nord-ouest et dans le centre du pays comparé à l’est du pays. Ce résultat concorde avec les résultats obtenus par Zeiss et al. (2024), qui ont observé que la richesse en espèces de vers de terre était élevée dans le centre-ouest de l’Europe et faible dans le nord-est de l’Europe. De plus, Rutgers et al. (2016) affirme que la distribution à grande échelle des densités de vers de terre apparaît grâce à des corrélations positives avec la latitude, la longitude et les facteurs climatiques à l’échelle Européen. Dans notre étude, ce résultat pourrait s’explique par les différences d’occupation des sols : davantage de prairies à l’ouest, tandis que l’est est caractérisé par des forêts et des parcelles de vignes.

Nous avons prédit la répartition des vers de terre dans les zones pour lesquelles nous avons pu extraire les prédicteurs des modèles grâce aux cartes de hautes résolutions, ainsi que dans certaines zones supplémentaires pour produire des cartes continues. Ces zones supplémentaires étaient principalement localisées dans des parcelles avec des occupation du sol qui ne sont pas des variables d’entrés des modèles. Ce sont par exemple les zones à forte artificialisation (zones industrielle ou commerciale, aéroports), les zones humides, les surfaces en eau ou les territoires agroforestières. Les répartitions attendues des vers de terre dans ces zones ont été indirectement dérivées par interpolation et doivent donc être interprétées avec prudence. Les cartes ont été créées principalement pour sensibiliser et pour protéger la biodiversité des sols et donc des services écosystémiques associées.

# Conclusions

Dans cette étude, nous avons développé des modèles prédictive grâce à une approche comparative entre les modèles de régression traditionnels (GLM, GAM) et les algorithmes d'apprentissage automatique (RF, GBM et ANN) afin d'identifier le meilleur modèle pour prédire la communauté des vers de terre en France métropolitaine. Les modèles boostés généralisés et les forêts aléatoires avait les meilleures performances prédictives. Les premiers résultats notables que nous avons obtenus concernent la précision estimée de nos modèles qui permettent de prédire la biomasse totale avec un R² de 0.35 (RMSE = 8.76 g/m²), l’abondance totale avec un R² de 0.43 (RMSE = 25 ind./m²) et la richesse taxonomique totale avec un R² de 0.59 (RMSE = 1.7 espèce par parcelle). Notre étude a permis de mettre en évidence que l’occupation du sol était la variable la plus importante sur les vers de terre suivie des variables spatiales, climatiques et pédologiques. Notre étude a permis aussi de créer des cartes de prédiction de la communauté des vers de terre en France, fournissant ainsi des données de référence et sensibilisant à la protection de la biodiversité du sol et donc des services écosystémiques associés. Afin de compléter cette étude et obtenir un aperçu global de la communauté des verts de terre en France, il serait pertinent de développer d’autres modèles permettant de prédire l’abondance ou la présence absence des espèces mais aussi l’abondance et la biomasse des catégories écologiques des vers de terre.

# Bibliographie

Anthony, M. A., Bender, S. F., & van der Heijden, M. G. (2023). Enumerating soil biodiversity. *Proceedings of the National Academy of Sciences*, *120*(33), e2304663120. https://doi.org/10.1073/pnas.2304663120

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., & Panagos, P. (2019). Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. *Geoderma*, *355*, 113912. https://doi.org/10.1016/j.geoderma.2019.113912

Bardgett, R. D., & Van der Putten, W. H. (2014). Belowground biodiversity and ecosystem functioning. *Nature*, *515*(7528), Article 7528. https://doi.org/10.1038/nature13855

Bouché, M. B. (1972). *Lombriciens de France. Ecologie et systématique* (Vol. 72‑2, Numéro HS, p. 671 p.). INRA Editions. https://hal.inrae.fr/hal-02859798

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5‑32. https://doi.org/10.1023/A:1010933404324

Capowiez, Y., Sammartino, S., & Michel, E. (2014). Burrow systems of endogeic earthworms : Effects of earthworm abundance and consequences for soil water infiltration. *Pedobiologia*, *57*(4), 303‑309. https://doi.org/10.1016/j.pedobi.2014.04.001

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), Article 7401. https://doi.org/10.1038/nature11148

Casalicchio, G., Molnar, C., & Schratz, P. (2024). *iml : Interpretable Machine Learning* (0.11.3) [Logiciel]. https://cran.r-project.org/web/packages/iml/

Crittenden, S. J., Eswaramurthy, T., de Goede, R. G. M., Brussaard, L., & Pulleman, M. M. (2014). Effect of tillage on earthworms over short- and medium-term in conventional and organic farming. *Applied Soil Ecology*, *83*, 140‑148. https://doi.org/10.1016/j.apsoil.2014.03.001

Cunha, L., Brown, G. G., Stanton, D. W. G., Da Silva, E., Hansel, F. A., Jorge, G., McKey, D., Vidal-Torrado, P., Macedo, R. S., Velasquez, E., James, S. W., Lavelle, P., Kille, P., & Network, the T. P. de I. (2016). Soil Animals and Pedogenesis : The Role of Earthworms in Anthropogenic Soils. *Soil Science*, *181*(3/4), 110‑125. https://doi.org/10.1097/SS.0000000000000144

Diallo, A., Hoeffner, K., Guillocheau, S., Sorgniard, P., & Cluzeau, D. (2023). Combined effects of annual crop agricultural practices on earthworm communities. *Applied Soil Ecology*, *192*, 105073. https://doi.org/10.1016/j.apsoil.2023.105073

Edwards, C. A., & Arancon, N. Q. (2022). *Biology and Ecology of Earthworms*. Springer US. https://doi.org/10.1007/978-0-387-74943-3

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models : Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677‑697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ernst, G., & Emmerling, C. (2009). Impact of five different tillage systems on soil organic carbon content and the density, biomass, and community composition of earthworms after a ten year period. *European Journal of Soil Biology*, *45*(3), 247‑251. https://doi.org/10.1016/j.ejsobi.2009.02.002

FAO, I. (2020). *State of knowledge of soil biodiversity – Status, challenges and potentialities. Summary for policy makers*. https://policycommons.net/artifacts/1526136/state-of-knowledge-of-soil-biodiversity-status-challenges-and-potentialities/2214245/

Fourcade, Y., & Vercauteren, M. (2022). Predicted changes in the functional structure of earthworm assemblages in France driven by climate change. *Diversity and Distributions*, *28*(5), 1050‑1066. https://doi.org/10.1111/ddi.13505

Gabriac, Q., Ganault, P., Barois, I., Aranda-Delgado, E., Cimetière, E., Cortet, J., Gautier, M., Hedde, M., Marchán, D. F., Reyes, J. C. P., Stokes, A., & Decaëns, T. (2022). *Environmental drivers of earthworm communities along an altitudinal gradient in the French Alps* (p. 2022.10.13.512055). bioRxiv. https://doi.org/10.1101/2022.10.13.512055

Gardi, C., & Jeffery, S. (2009). *Soil biodiversity*. https://doi.org/10.2788/7831

Greg, R., Edwards, D., Kriegler, B., Schroedl, S., Southworth, H., Greenwell, B., Boehmke, B., Cunningham, J., & Developers (https://github.com/gbm-developers), G. B. M. (2024). *gbm : Generalized Boosted Regression Models* (2.1.9) [Logiciel]. https://cran.r-project.org/web/packages/gbm/index.html

Groves, C. P. (2022). *Biogeographic region | Definition, Features, Locations, & Facts | Britannica*. https://www.britannica.com/science/biogeographic-region

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models : With Applications in R*. Cambridge University Press. https://doi.org/10.1017/9781139028271

Hijmans, R. J., & Elith, J. (2019). *Spatial Distribution Models*. https://rspatial.org/sdm/SDM.pdf

Hoeffner, K., Hotte, H., Cluzeau, D., Charrier, X., Gastal, F., & Pérès, G. (2021). Effects of temporary grassland introduction into annual crop rotations and nitrogen fertilisation on earthworm communities and forage production. *Applied Soil Ecology*, *162*, 103893. https://doi.org/10.1016/j.apsoil.2021.103893

Hooper, D. U., Adair, E. C., Cardinale, B. J., Byrnes, J. E. K., Hungate, B. A., Matulich, K. L., Gonzalez, A., Duffy, J. E., Gamfeldt, L., & O’Connor, M. I. (2012). A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*, *486*(7401), Article 7401. https://doi.org/10.1038/nature11118

Horrigue, W., Dequiedt, S., Chemidlin Prévost-Bouré, N., Jolivet, C., Saby, N. P. A., Arrouays, D., Bispo, A., Maron, P.-A., & Ranjard, L. (2016). Predictive model of soil molecular microbial biomass. *Ecological Indicators*, *64*, 203‑211. https://doi.org/10.1016/j.ecolind.2015.12.004

Kalinowski, T., Falbel, D., Allaire, J. J., Chollet, F., RStudio, Google, Tang [ctb, Y., cph, Bijl, W. V. D., Studer, M., & Keydana, S. (2024). *keras : R Interface to « Keras »* (2.15.0) [Logiciel]. https://cran.r-project.org/web/packages/keras/index.html

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth’s land surface areas. *Scientific Data*, *4*(1), 170122. https://doi.org/10.1038/sdata.2017.122

Lavelle, P., Bignell, D., Lepage, M., Wolters, V., Roger, P., Ineson, P., Heal, O. W., & Dhillion, S. (1997). Soil function in a changing world : The role of invertebrate ecosystem engineers. *European Journal of Soil Biology*. https://www.semanticscholar.org/paper/Soil-function-in-a-changing-world%3A-the-role-of-Lavelle-Bignell/363b3c76b7cef2aa2637d1b3a33034ffa41b1dce

Leroy, B. L. M., Schmidt, O., Van den Bossche, A., Reheul, D., & Moens, M. (2008). Earthworm population dynamics as influenced by the quality of exogenous organic matter. *Pedobiologia*, *52*(2), 139‑150. https://doi.org/10.1016/j.pedobi.2008.07.001

Li, X., & Wang, Y. (2013). Applying various algorithms for species distribution modelling. *Integrative Zoology*, *8*(2), 124‑135. https://doi.org/10.1111/1749-4877.12000

Maggi, F., & Tang, F. H. M. (2021). Estimated decline in global earthworm population size caused by pesticide residue in soil. *Soil Security*, *5*, 100014. https://doi.org/10.1016/j.soisec.2021.100014

Marchán, D. F., Csuzdi, C., Decaëns, T., Szederjesi, T., Pizl, V., & Domínguez, J. (2021). The disjunct distribution of relict earthworm genera clarifies the early historical biogeography of the Lumbricidae (Crassiclitellata, Annelida). *Journal of Zoological Systematics and Evolutionary Research*, *59*(8), 1703‑1717. https://doi.org/10.1111/jzs.12514

Marchán, D. F., & Domínguez, J. (2022). Evaluating the Conservation Status of a North-Western Iberian Earthworm (Compostelandrilus cyaneus) with Insight into Its Genetic Diversity and Ecological Preferences. *Genes*, *13*(2), Article 2. https://doi.org/10.3390/genes13020337

Marchán, D. F., Refoyo, P., Fernández, R., Novo, M., de Sosa, I., & Díaz Cosín, D. J. (2016). Macroecological inferences on soil fauna through comparative niche modeling : The case of Hormogastridae (Annelida, Oligochaeta). *European Journal of Soil Biology*, *75*, 115‑122. https://doi.org/10.1016/j.ejsobi.2016.05.003

Marchán, D. F., Refoyo, P., Novo, M., Fernández, R., Trigo, D., & Díaz Cosín, D. J. (2015). 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. https://doi.org/10.1016/j.soilbio.2014.10.023

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, *5*, e2849. https://doi.org/10.7717/peerj.2849

Niswati, A., Liyana, Prasetyo, D., & Lumbanraja, J. (2022). Abundance and biomass of earthworm as affected by long-term different types of soil tillage and fertilization on mung bean plantation at Ultisols. *IOP Conference Series: Earth and Environmental Science*, *1018*(1), 012012. https://doi.org/10.1088/1755-1315/1018/1/012012

Palm, J., van Schaik, N. L. M. B., & Schröder, B. (2013). Modelling distribution patterns of anecic, epigeic and endogeic earthworms at catchment-scale in agro-ecosystems. *Pedobiologia*, *56*(1), 23‑31. https://doi.org/10.1016/j.pedobi.2012.08.007

Pebesma, E., & Graeler, B. (2023). *gstat : Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation* (2.1-1) [Logiciel]. https://cran.r-project.org/web/packages/gstat/

Pelosi, C., Barot, S., Capowiez, Y., Hedde, M., & Vandenbulcke, F. (2014). Pesticides and earthworms. A review. *Agronomy for Sustainable Development*, *34*(1), 199‑228. https://doi.org/10.1007/s13593-013-0151-z

Pelosi, C., Pey, B., Hedde, M., Caro, G., Capowiez, Y., Guernion, M., Peigné, J., Piron, D., Bertrand, M., & Cluzeau, D. (2014). Reducing tillage in cultivated fields increases earthworm functional diversity. *Applied Soil Ecology*, *83*, 79‑87. https://doi.org/10.1016/j.apsoil.2013.10.005

Phillips, H. R. P., Guerra, C. A., Bartz, M. L. C., Briones, M. J. I., Brown, G., Crowther, T. W., Ferlian, O., Gongalsky, K. B., van den Hoogen, J., Krebs, J., Orgiazzi, A., Routh, D., Schwarz, B., Bach, E. M., Bennett, J. M., Brose, U., Decaëns, T., König-Ries, B., Loreau, M., … Eisenhauer, N. (2019). Global distribution of earthworm diversity. *Science*, *366*(6464), 480‑485. https://doi.org/10.1126/science.aax4851

R Core Team. (2023). *A language and environment for statistical computing* (R.4.3.0) [Logiciel]. R Foundation for Statistical Computing. URL https://www.R-project.org/.

Roman Dobarco, M., Bourennane, H., Arrouays, D., Saby, N., Cousin, I., & Manuel, M. P. (2022). *Propriétés de granulométrie (argile, limons, sables) et d’éléments grossiers pour la France métropolitaine au pas de 90 m* [jeu de données]. Portail Data INRAE. https://doi.org/10.57745/N4E4NE

Rutgers, M., Orgiazzi, A., Gardi, C., Römbke, J., Jänsch, S., Keith, A. M., Neilson, R., Boag, B., Schmidt, O., Murchie, A. K., Blackshaw, R. P., Pérès, G., Cluzeau, D., Guernion, M., Briones, M. J. I., Rodeiro, J., Piñeiro, R., Cosín, D. J. D., Sousa, J. P., … Zwart, D. de. (2016). Mapping earthworm communities in Europe. *Applied Soil Ecology*, *97*, 98‑111. https://doi.org/10.1016/j.apsoil.2015.08.015

Salako, G., Russell, D. J., Stucke, A., & Eberhardt, E. (2023). Assessment of multiple model algorithms to predict earthworm geographic distribution range and biodiversity in Germany : Implications for soil-monitoring and species-conservation needs. *Biodiversity and Conservation*, *32*(7), 2365‑2394. https://doi.org/10.1007/s10531-023-02608-9

Sharma, D. K., Tomar, S., & Chakraborty, D. (2017). Role of earthworm in improving soil structure and functioning. *Current Science*, *113*(6), 1064‑1071.

Smith, R. G., McSwiney, C. P., Grandy, A. S., Suwanwaree, P., Snider, R. M., & Robertson, G. P. (2008). Diversity and abundance of earthworms across an agricultural land-use intensity gradient. *Soil and Tillage Research*, *100*(1), 83‑88. https://doi.org/10.1016/j.still.2008.04.009

Spurgeon, D. J., Keith, A. M., Schmidt, O., Lammertsma, D. R., & Faber, J. H. (2013). Land-use and land-management change : Relationships with earthworm and fungi communities and soil structural properties. *BMC Ecology*, *13*(1), 46. https://doi.org/10.1186/1472-6785-13-46

Turbé, A., Toni, A., Benito, P., Lavelle, P., Lavelle, P., Camacho, N., Putten, W., Labouze, E., & Mudgal, S. (2010). Soil biodiversity : Functions, threats and tools for policy makers. *Report for European Commission*.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731‑1742. https://doi.org/10.1111/ecog.05615

Van Groenigen, J. W., Lubbers, I. M., Vos, H. M. J., Brown, G. G., De Deyn, G. B., & van Groenigen, K. J. (2014). Earthworms increase plant production : A meta-analysis. *Scientific Reports*, *4*(1), Article 1. https://doi.org/10.1038/srep06365

Van Groenigen, J. W., Van Groenigen, K. J., Koopmans, G. F., Stokkermans, L., Vos, H. M. J., & Lubbers, I. M. (2019). How fertile are earthworm casts? A meta-analysis. *Geoderma*, *338*, 525‑535. https://doi.org/10.1016/j.geoderma.2018.11.001

Wood, S. (2023). *mgcv : Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (1.9-1) [Logiciel]. https://cran.r-project.org/web/packages/mgcv/index.html

Yiu, T. (2021, septembre 29). *Understanding Random Forest*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zeiss, R., Briones, M. J. I., Mathieu, J., Lomba, A., Dahlke, J., Heptner, L.-F., Salako, G., Eisenhauer, N., & Guerra, C. A. (2024). Effects of climate on the distribution and conservation of commonly observed European earthworms. *Conservation Biology*, *38*(2), e14187. https://doi.org/10.1111/cobi.14187

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261‑1277. https://doi.org/10.1111/ecog.04960

# Annexes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithms** | **Response variables** | **LandWorm** | |  | **EcoBioSoil** | |
| **R²** | **RMSE** |  | **R²** | **RMSE** |
| GLM | Abundance | 0.22 | 34.57 |  | 0.25 | 33.99 |
| GAM | 0.26 | 33.06 |  | 0.27 | 32.72 |
| **RF** | **0.43** | **25.20** |  | **0.41** | **28.83** |
| GBM | 0.43 | 25.30 |  | 0.4 | 27.46 |
| ANN | 0.35 | 28.94 |  | 0.25 | 34.22 |
|  |  |  |  |  |  |  |
| GLM | Biomass | 0.23 | 10.69 |  | 0.22 | 10.76 |
| GAM | 0.24 | 10.50 |  | 0.25 | 10.30 |
| **RF** | **0.35** | **8.76** |  | **0.34** | **9.24** |
| GBM | 0.32 | 9.30 |  | 0.32 | 9.36 |
| ANN | 0.27 | 10.50 |  | 0.32 | 9.64 |
|  |  |  |  |  |  |  |
| GLM | Richness | 0.36 | 2.18 |  | 0.33 | 2.16 |
| GAM | 0.44 | 2.04 |  | 0.39 | 2.08 |
| **RF** | **0.59** | **1.75** |  | **0.57** | **1.77** |
| GBM | 0.59 | 1.75 |  | 0.55 | 1.80 |
| ANN | 0.40 | 2.16 |  | 0.38 | 2.16 |

**Quatrième de couverture (**résumé en Anglais en Français, mots clés,…)