**Title : Modeling and predicting earthworm diversity and distribution in France: A comparative approach using multiple algorithms**

# Introduction

The biodiversity of terrestrial ecosystems is predominantly constituted by soil-dwelling fauna (FAO, 2020; Anthony et al., 2023). This soil fauna provides a wide range of ecosystem services (Bardgett & Van der Putten, 2014; FAO, 2020), most of which result from processes conducted by the functional groups of various biological communities, such as earthworms (Gardi & Jeffery, 2009; Turbé et al., 2010). Earthworms are referred to as "ecosystem engineers" because they contribute to soil structure development (Sharma et al., 2017; Edwards & Arancon, 2022; Lavelle et al., 1997), water infiltration through vertical and horizontal burrows in the soil (Capowiez et al., 2014; Cunha et al., 2016), and nutrient mineralization through the degradation of organic matter (Van Groenigen et al., 2019). These studies have been justified by the growing awareness that changes in earthworm composition will lead to an alteration of the ecosystem services provided by these organisms (Cardinale et al., 2012; Hooper et al., 2012; Van Groenigen et al., 2014).

Numerous studies have examined the effects of anthropogenic and environmental factors on earthworms at local scales (Pelosi, Pey, et al., 2014; Marchán et al., 2015; Gabriac et al., 2022) and regional scales (Marchán et al., 2016, 2021; Marchán & Domínguez, 2022; Diallo et al., 2023). However, few studies have focused on the effects of these factors on earthworm biodiversity and distribution at larger scales, i.e., supranational or national levels (Fourcade & Vercauteren, 2022; Salako et al., 2023; Zeiss et al., 2024). The reasons for this lack of knowledge include the limited availability of data at the country or continent level, taxonomic inconsistencies, and difficulties in merging existing databases (Rutgers et al. 2016). The first study conducted at a continental scale was carried out by Rutgers et al. (2016), who mapped the earthworm community at 3,838 sampled sites across 8 European countries. They observed that earthworm abundance and richness were affected by land use, soil properties (pH, organic matter, and textures), and latitude. Recently, another study by Phillips et al. (2019) on 9,212 sites distributed across 57 countries demonstrated that climatic factors and habitat type are the most important environmental filters affecting earthworm parameters. However, their studies are limited by the use of a single type of predictive modeling algorithm: generalized linear models for Rutgers et al. (2016) and generalized linear mixed models for Phillips et al. (2019). Therefore, it would be relevant to compare multiple algorithms to find the best model for each earthworm parameter (abundance, biomass, and richness).

In France, the most recent study was conducted at 1,366 sites by Fourcade and Vercauteren (2022), who utilized boosted regression trees to build spatial predictions of functional diversity with 44 earthworm species. Their research shows a decline in functional richness between 1960 and 2012. Additionally, their models predict that this reduction in functional richness could continue in the future across different temporal periods and climate change scenarios. However, Fourcade and Vercauteren's (2022) study is based on presence/absence data of earthworm species in France collected in the 1960s by Bouché (1972). These data no longer reflect the current earthworm assemblage in France. Furthermore, the study of earthworm communities should include variables describing species richness, abundance, and total biomass, which are essential parameters for biodiversity assessments. Relying solely on functional diversity is not sufficient to fully evaluate biodiversity, as, for instance, two communities may be identical in terms of diversity but differ in terms of density (Groves, 2022).

To understand biodiversity, its distribution, and the environmental factors influencing this distribution, several tools have been developed in recent years. We distinguish species distribution models (SDMs), which are models based on ecological niches and allow for modeling correlations between species or communities and their environment (Elith & Leathwick, 2009; Guisan et al., 2017). However, this type of model requires, on the one hand, data on parameters characterizing communities and, on the other hand, spatial environmental data. Various SDMs exist, each with its advantages and disadvantages (Li & Wang, 2013; Valavi et al., 2021). The comparative approach of multiple SDMs to predict the geographical distribution area and biodiversity of earthworms in Germany, as proposed by Salako et al. (2023), seems to be a promising approach. The general objective of our study was to develop models to predict the earthworm community in metropolitan France using a comparative approach of different modeling algorithms. Specifically, we sought to address the following objectives: (i) quantify and rank the influence of environmental factors (land use, soil properties, location, and climate data) based on their contribution to predictive models on abundance (individuals per m²), biomass (g per m²), and total taxonomic richness of earthworms (number of taxa in the plot) in metropolitan France (excluding Corsica), and (ii) predict and map these same earthworm parameters based on environmental factors. To address these objectives, we selected five SDM methods or algorithms and compared their predictive performances. This comparative approach reduces the uncertainty associated with predictions and identifies the best model for each earthworm variable (abundance, biomass, and richness). The selected algorithms were generalized linear models (GLM), generalized additive models (GAM), random forests (RF), generalized boosted regression models (GBM), and artificial neural networks (ANN). These algorithms were chosen based on their classification as regression algorithms (GLM and GAM) and machine learning algorithms (RF, GBM, and ANN), as well as their widespread use in recent studies (Li & Wang, 2013; Rutgers et al., 2016; Valavi et al., 2021; Fourcade & Vercauteren, 2022; Salako et al., 2023).

Our hypotheses are based on the assumption that climatic variables (precipitation and temperature) and land use type would be the most influential factors on earthworm abundance and biomass, and that land use type and soil texture would have a greater influence on species richness.

# Materials & methods

## Study area and earthworm data collection

The study area encompasses the entire mainland territory of metropolitan France (excluding Corse) (Fig. 1). We utilized 3758 earthworm observations from the LandWorm database (2023-2026 FRB-MTE-OFB), which includes earthworm communities across metropolitan France under various land use conditions (see Appendix XX). Earthworm sampling was conducted between 1990 and 2023, primarily during spring, corresponding to the period of maximum earthworm activity. Sampling methods for earthworm communities varied depending on the data sources (see Appendix XX). Sampling methods included manual sorting (ISO 23611-1:2018) and/or the application of a chemical expellant (formaldehyde or allyl isothiocyanate; ISO 23611-1:2006). In each site, earthworms were collected, sorted manually, and stored in ethanol. In the laboratory, earthworms were counted, sometimes weighed, and identified at the species/taxon level. However, 48% of the plots lacked total biomass data. From this database, we also retrieved GPS coordinates and Corine Land Cover information provided by project managers (https://land.copernicus.eu/pan-european/corine-land-cover).

Une image contenant texte, diagramme, carte

Description générée automatiquement

#### **Fig. 1 :** Map of study area ( metropolitan France (excluding Corsica)) showing location of earthworm sampling sites.

## Environmental aata collection

In addition to land use data and GPS coordinates obtained from the LandWorm database (2023-2026 FRB-MTE-OFB), we utilized external variables known to affect earthworms (Rutgers et al., 2016; Phillips et al., 2019; Salako et al., 2023). For climatic variables, we employed the 19 standard bioclimatic variables from the CHELSA project (Karger et al., 2017). These climatic data represented average values between 1981 and 2010 at a resolution of 30 arc-seconds. This time frame corresponded best to the primary period of biological data.

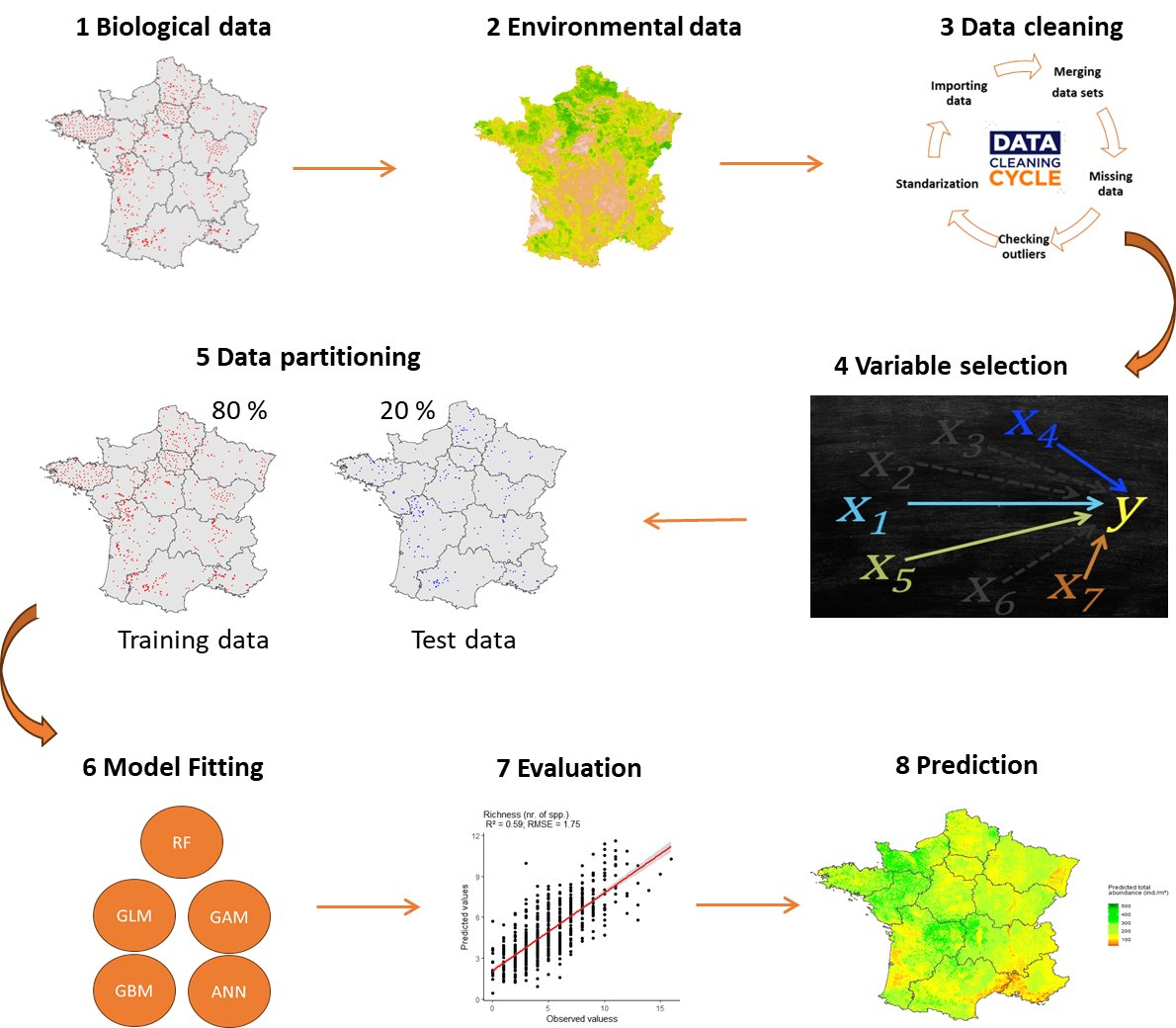
Regarding soil variables, we initially accessed the Research Data Gouv repository (https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/N4E4NE; Roman Dobarco et al., 2022) to retrieve soil sand, clay, and silt contents. These three variables, available at a resolution of 90 meters, originated from 2022 data and were provided at different depths. We chose to utilize the first three layers (0 to 5 cm, 5 to 15 cm, and 15 to 30 cm) and then averaged them to obtain data from 0 to 30 cm. This choice was made to account for earthworm habitat variation.

Subsequently, we used the Joint Research Centre (JRC) database (Ballabio et al., 2019) to supplement soil texture with the following variables: cation exchange capacity (CEC), calcium carbonate (CaCO₃), C/N ratio, nitrogen (N), phosphorus (P), potassium (K), and soil pH in H2O. These variables were available at a resolution of 250 meters on a European scale and based on 2009/2012 LUCAS (Land Use and Cover Area frame Survey) data on arable soils. Initially, we referenced all preselected variables to the World Geodetic System (WGS84) coordinate system, cropped and masked them to match the geographical boundaries of mainland France (excluding Corsica). To standardize the variables and match them to the same resolution, we resampled or disaggregated the variables to a resolution of 30 arc-seconds, approximately 800 meters in France. These steps were carried out using Python with the GDAL module (https://pypi.org/project/GDAL/).

## Modeling strategy

### ODMAP protocol

Our study's modeling strategy followed the "ODMAP" protocol as recommended by Zurell et al. (2020), and all steps are detailed in Appendix XX. The various modeling steps are depicted in Fig. 2. All modeling steps were performed using R software version 4.3.1, 2023 (R Core Team, 2023). Initially, all explanatory variables (numeric) were standardized. Additionally, total abundance and biomass were square-root transformed to approximate a Gaussian distribution.



#### **Fig. 2:**Modeling strategy according to the ODMAP protocol: (1) biological data collection, (2) environmental data collection, (3) data cleaning, (4) variable selection, (5) data partitioning, (6) model fitting, (7) model evaluation, and (8) Prediction.

### Variable selection, importance and effects

For each of the three response variables of earthworm communities, we fitted random forest models to identify the importance of each explanatory variable (Breiman, 2001). We chose the random forest model because it can handle non-linear data, accept correlated variables, and incorporate variable interactions that can improve model performance (Breiman, 2001). To reduce the number of variables and avoid overfitting (Vaughan & Ormerod, 2005), we identified the ten most important variables on earthworms using a permutation procedure (Fourcade & Vercauteren, 2022; Zeiss et al., 2024). Thus, the variables identified (Table 1) for prediction were: land use, longitude, latitude, calcium carbonate (CaCO₃), nitrogen (N), phosphorus (P), clay and silt content, isothermality (bio3), and average annual precipitation. Subsequently, all fitted models used these ten selected variables. Additionally, we used the iml package to improve the interpretability of the models, particularly by exploring the importance, effects, and interactions of variables (Casalicchio et al., 2024).

#### **Table 1 :** Abbreviations and description of variables used in predicting earthworm communities.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Category** | **Abbr** | **Description** | **Units** | **Sum** | **Min** | **Mean** | **Max** | **Time scale** | **Reference** |
| **Land use** | For | Forest | Boolean | 116 |  |  |  | 1990 - 2023 | LandWorm database |
| Gua | Green urban areas | Boolean | 535 |  |  |  |
| Nag | Natural grasslands | Boolean | 111 |  |  |  |
| Nial | Non-irrigated arable land | Boolean | 1683 |  |  |  |
| Pmo | Pastures, meadows and other permanent grasslands under agricultural use | Boolean | 413 |  |  |  |
| Viny | Vineyards | Boolean | 718 |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Location** | Long | Longitude | WGS84 decimal system |  | -4.612 | 1.772 | 8.056 |  |
| Lat | Latitude | WGS84 decimal system |  | 43.014 | 47.504 | 50.982 |
|  |  |  |  |  |  |  |  |  |  |
| **Climate** | bio12 | Annual precipitation amount | kg m-2 year-1 |  | 599 | 820 | 1283 | 1981 - 2010 | Karger et al. (2017) |
| bio3 | Isothermality | °C |  | 0.247 | 0.338 | 0.394 |
|  |  |  |  |  |  |  |  |  |  |
| **Soil** | Clay | Clay particles | g·kg -1 |  | 0.4 | 24.4 | 53.0 | 2022 | Roman Dobarco et al. (2022) |
| Silt | Silt particles | g·kg -1 |  | 2.3 | 47.0 | 81.6 |
| CaCO3 | Calcium carbonates | g·kg−1 |  | 0 | 75 | 332 | 2009 - 2012 | Ballabio et al. (2019) |
| P | Phosphorus | mg·kg−1 |  | 7.18 | 40.85 | 68.39 |
| N | Nitrogen | g·kg−1 |  | 0.87 | 1.99 | 3.84 |

### Model calibration

We compared five SDM algorithms to predict earthworm parameters (total abundance, total biomass, and total taxonomic richness) using 10 explanatory variables. The five algorithms were: Generalized Linear Models (GLM), Generalized Additive Models (GAM), Random Forest models (RF), Generalized Boosted Models (GBM), and Artificial Neural Networks (ANN).

We fitted GLMs using the *glm* function from the *stats* package with the following formulation:

Where *Y* is the response variable (abundance, biomass, or richness) and x represents the explanatory variables.

For GAMs, we utilized the *gam* function from the *mgcv* package (Wood, 2023)with the following formulation:

Where *Y* is the response variable (abundance, biomass, or richness) and x represents the explanatory variables.

Random Forest models were fitted using the *randomForest* function from the *randomForest* package (Breiman, 2001) with the following formulation:

Where rep.var represents the position of the response variable column, mtry is the number of variables randomly sampled, ntree is the number of trees, and maxnodes is the maximum number of terminal nodes.

We performed hyperparameter tuning for Random Forest models using a grid search method with all possible combinations of the following parameters: number of variables randomly sampled (2 to 10 in increments of 1), number of trees (100 to 2000 in increments of 200), and maximum number of terminal nodes (2 to 15 in increments of 1). We selected the model with the highest R² and lowest RMSE from all models.

GBMs were fitted using the *gbm* function from the *gbm* package (Greg et al., 2024) with the following formulation:

Where n.trees is the number of trees, shrinkage is the learning rate, interaction.depth is the interaction depth, and n.minobsinnode is the minimum number of observations in terminal nodes.

Several parameters needed to be selected in GBMs to control the model complexity. To select the most appropriate parameters, we fitted the models using a grid search method with all possible combinations of the following parameters: number of trees (500 to 2000 in increments of 100), interaction depth (1, 3, 5, 6, 8,10), learning rate (0.01, 0.02, 0.05, 0.001, 0.002, 0.005), and minimum number of observations in terminal nodes (2, 5, 10, 20, 30, 50). We selected the model with the highest R² and lowest RMSE from all models.

For ANN, we used the keras package (Kalinowski et al., 2024) with a sequential architecture. The model consisted of an input layer with an input\_shape of 15 corresponding to the 9 explanatory variables plus the 6 levels of land use that we transformed into independent binary variables. We introduced three hidden layers with 32, 16, and 8 dense neurons, respectively. The last layer consisted of a single neuron corresponding to the predicted variable (abundance, biomass, and richness). All layers were accompanied by a ReLU activation function except the last layer, which had a linear activation. We used the mean squared error (MSE) loss function and the RMSprop optimizer, while the mean absolute error (MAE) was used to evaluate model performance. For compilation, we defined an epochs of 100, a batch\_size of 64, and a validation\_split of 0.2. To mitigate overfitting, we added four dropout layers and introduced an EarlyStopping callback with a patience of 10 to monitor loss on the validation set and restore weights from the best model.

### Model evaluation

We evaluated the models using the cross-validation method by randomly assigning 80% of the data for model training and 20% for model validation (Horrigue et al., 2016; Hijmans & Elith, 2019; Salako et al., 2023). This method was chosen for its simplicity in understanding and implementation, as well as its quick compilation. Moreover, it proved to be effective in our case, as the dataset was large and the distribution between training and validation data was similar (non-significant Kolmogorov-Smirnov test, p-value > 0.05) (Guisan et al., 2017). The training data were used to fit the models, while the validation data were used to assess the predictive performances of the models. Since all our response variables were quantitative, we chose the coefficient of determination (R²) between observations and predictions, mean absolute error (MAE), and root mean square error (RMSE) as performance evaluation metrics.

where: n is the number of observations, represents the actual value of observation (i), represents the predicted value of observation (i).

The objective was to maximize R² and minimize MAE and RMSE.

## Predictions and mapping of earthworm communities

The prediction of earthworm communities was conducted using the best algorithm for each of the three earthworm variables (total abundance, total biomass, and taxonomic richness). Initially, we sampled at a resolution of approximately 800 m across the entire territory of metropolitan France (excluding Corsica). Subsequently, for each sampling point, we extracted the values of the different final variables included in the models from the databases detailed in section 2.1, whenever possible. Then, we used the `predict` function, providing the final model of the best algorithm and the extracted explanatory variables to predict earthworms. Finally, we presented the predicted values as maps of earthworm communities. For areas where earthworm communities could not be predicted, we performed interpolation using the Inverse Distance Weighting (IDW) method (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.

To estimate the approximate biodiversity of the earthworm community, we overlaid the map of total abundance with the map of total taxonomic richness (Salako et al., 2023).

# Results

## Model performance

The performances of the five models varied with an average R² of 0.33 (± 0.10 sd) for abundance, 0.28 (± 0.05 sd) for biomass, and 0.48 (± 0.11 sd) for total taxonomic richness. In terms of RMSE, the average was 29 ind./m² (± 4.32 sd) for abundance, 9.95 g/m² (± 3.69 sd) for biomass, and 1.92 species per plot (± 0.21 sd) for total taxonomic richness (Table 2). For all three earthworm parameters, RF and GBM exhibited the highest R² and the lowest RMSE, indicating good prediction of the earthworm community (Table 2). The evaluation of predictive model performances showed that GLM was the least performing model in predicting abundance, biomass, and richness of earthworms. Fig. 3 illustrates the comparison between observed values (validation dataset) and values predicted by the best models. For total abundance, the best model was RF with an R² of 0.43. Similarly, the best model for total biomass was RF with an R² of 0.35. For total richness, the best models were RF and GBM, both with identical R² values of 0.59.

#### **Table 2 :** Performance measures of prediction on the validation dataset for different model algorithms tested on the three response variables of the earthworm community. Bold values indicate the best algorithm for each earthworm variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 :** Prediction on the validation dataset with the best algorithms for (a) total abundance (RF; ind/m²), (b) total biomass (RF; g/m²), and (c) total taxonomic richness (GBM; number of taxa in the plot). The X-axis indicates the observed values, and the Y-axis indicates the predicted values. The red line represents the linear regression (trend) between the observed and predicted values, and the gray band indicates the confidence interval around the regression line.

## Contribution of variables

The most important variable during the adjustment of prediction models was land use (CLC; Fig. 4). When land use was permuted, the RMSE of the abundance model increased on average by 1.68 ind./m², by 1.45 g/m² for biomass, and by 1.78 species per plot for earthworm richness. After land use, spatial variables were the most important for predicting earthworms, particularly longitude, which led to an average increase in RMSE of 1.63 species per plot. Regarding climatic variables, annual precipitation was the most important for earthworm prediction. The two most influential soil variables on earthworms were calcium carbonate (CaCO₃) and nitrogen (N), respectively. Additionally, CaCO₃ was found to have a greater influence on richness and total abundance compared to biomass.

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

Description générée automatiquement

#### **Fig. 4 :** Importance of environmental explanatory variables for predicting total abundance, total biomass, and taxonomic richness of earthworms.

## Effects of variables on earthworms

Analyses of Accumulated Local Effects (ALE) showed that land use had mixed effects on the earthworm community. ALE effects describe how explanatory variables on average influence the prediction of a machine learning model. In other words, ALE effects show how prediction changes locally as variables vary. Indeed, plots in grasslands (Nag and Pmo) had higher abundances and increased the average abundance prediction by 80 ind./m². Similarly, plots in urban green spaces (Gua) increased the average abundance predictions by 50 ind./m². Plots in Nag, Pmo, and Gua had about one or two more species than other land use types. ALE effects also showed that plots in vineyards (Viny), forests (For), and annual crops (Nial) were the land uses associated with the lowest abundances. Regarding biomass, only grasslands increased the average weight of earthworms.

Spatial variables exerted a strong influence on earthworms, with longitude having the most significant impact. From west to east, abundance and richness decreased, with a decrease of about 50 individuals/m² and one species, respectively. However, biomass varied only slightly from west to east in France. Increasing latitude did not influence total abundance between 43°N and 46°N, then decreased between 46°N and 47.5°N before increasing beyond the latter latitude.

Regarding the effect of climatic variables, predictions of total abundance and biomass increased below 700 mm/year of annual precipitation but remained stable above this threshold. Earthworm biomass was minimally affected by the ratio between diurnal temperature variation and annual variation (Isot).

For soil parameters, increasing CaCO₃ up to approximately 10 g·kg−1 led to an increase in predictions of 25 ind./m² for abundance and 15 g/m² for biomass. However, CaCO₃ above 10 g·kg−1 led to a decrease in earthworms. Phosphorus had a minor negative effect on earthworms. Increasing nitrogen by 2 g·kg−1 led to an increase in abundance of 25 ind./m² and biomass of 5 g/m². Soil texture (clay and silt content) did not influence predictions of total abundance up to 40 g·kg−1, after which abundance began to increase, reaching an average increase of 45 ind./m². Texture had little influence on total biomass.

## Spatial distribution of earthworms at the scale of France

Figure 5 shows the predicted spatial distribution of the earthworm community in metropolitan France (excluding Corsica). Total abundance varied from 0 to 530 ind./m² with a mean of 192 ind./m² per plot (Fig. 5a). The average biomass was 72 g/m² (minimum = 0 and maximum = 196; Fig. 5b), while total taxonomic richness ranged from 0 to 10 with a mean of 5 species per plot (Fig. 5c). Figure 5d represents the approximate diversity of earthworms resulting from the overlay of the abundance map and the taxonomic richness map. It can be observed that the southeastern part of the country exhibited very low to low diversity, while the southwestern part had moderate diversity.

A map of france with different colored areas

Description automatically generated

#### Fig. 5 : Predicted spatial distribution of (a) total abundance (ind./m²), (b) total biomass (g/m²), (c) total taxonomic richness (number of taxa per plot), and (d) approximate diversity of earthworms.

# Discussion

In this study, we utilized the LandWorm database (2023-2026 FRB-MTE-OFB), containing earthworm communities across metropolitan France under varied land use conditions, to predict total abundance, total biomass, and total taxonomic richness of earthworms. We employed multiple SDM algorithms to compare their performances and select the most effective model for predicting the spatial distribution of earthworms. The notable initial results we obtained pertain to the estimated accuracy of our models, allowing for the prediction of total biomass with an R² of 0.35 (RMSE = 8.76 g/m²), total abundance with an R² of 0.43 (RMSE = 25 ind./m²), and total taxonomic richness with an R² of 0.59 (RMSE = 1.7 species per plot). We observed that the spatial distribution of earthworms in France is primarily influenced by land use and spatial variables, particularly longitude. Our study enabled the creation of prediction maps for earthworm communities, providing precise reference data and raising awareness about the protection of soil biodiversity and associated ecosystem services. All stages of the modeling strategy were reported following the ODMAP protocol (ANNEX XX), as recommended by Zurell et al. (2020). In this study, we adjusted five different model algorithms to model the distribution predictions of earthworms using land use, spatial, climatic, and soil variables.

## Superiority of ensemble methods in predicting earthworm communities compared to traditional regression models

Our results demonstrate that random forest (RF) models and generalized boosted models (GBM) better predicted earthworm communities compared to traditional regression models (GLM and GAM) and artificial neural networks (ANN). This finding aligns with those of Li & Wang (2013); Mi et al. (2017); Valavi et al. (2021), who observed that RF exhibited very high predictive performances. Additionally, a study conducted at the scale of Germany by Salako et al. (2023) using different prediction algorithms concluded that RF was the most effective algorithm for predicting earthworm communities. The high performances of these algorithms can be attributed to the fact that RF and GBM act as ensemble classifiers, utilizing multiple alternative trees in decision-making during model predictions (Breiman, 2001; Li & Wang, 2013). Their effectiveness also stems from their ability to better capture nonlinear relationships between explanatory and response variables, robustness to outliers, and superior handling of variable interactions (Breiman, 2001). Considering the high predictive potential of RF and GBM, it would be beneficial to use them in ecological studies and not restrict oneself to traditional regression algorithms alone(Rutgers et al., 2016; Phillips et al., 2019). However, RF and GBM require substantial amounts of data to achieve good predictive performances (Yiu, 2021).This indicates that model performances could be further improved in the future if we manage to merge and standardize the various existing databases between research units and countries. Another limitation of RF and GBM is their low interpretability, but this seems to be becoming less true as numerous tools now exist to better understand and interpret machine learning models. For example, the "iml" package provides very relevant tools for analyzing any black-box machine learning model. The package allows for exploring the importance, effects, and interactions of variables while also proposing surrogate models (Casalicchio et al., 2024).

It is important to note that the explanatory variables related to climate and soil come from external databases, as not all plots in our database contained this information. However, this reliance on external databases constitutes a limitation, as already mentioned Rutgers et al. (2016) and Salako et al. (2023). Indeed, the quality of the data used during model training greatly influences their performance. Nevertheless, we chose not to use external databases for land use because this information was well documented in our database. This choice avoids any temporal gap between earthworm observations and land uses, which may change over the years. Another limitation of our study lies in the restricted selection of land use types. We only used six types of Level 3 land use defined by Corine Land Cover. Other land use types were not included due to the lack of available data. Additionally, we could not distinguish between deciduous forests, coniferous forests, and mixed forests. Therefore, we grouped these three types of forests under a single land use type: "Forest" (all types). This decision was made considering that land use has a significant impact on earthworm populations, as highlighted by Spurgeon et al. (2013). We are also aware that our database is unbalanced in terms of sampling, with more observations in the north than in the south, which could explain the low R² for total biomass, especially since some earthworm observations had no biomass.

## Importance of land use, climate and soil properties

Our results demonstrate that land use was the most important variable on earthworms. This is consistent with the findings of Fourcade & Vercauteren (2022) et de Salako et al. (2023), which showed that land use affects the distribution of the earthworm community. Additionally, the study conducted by Rutgers et al. (2016) across Europe also highlighted the importance of land use. All predictive models used in this study predicted positive effects of grasslands (Pmo and Nag) on earthworms. This result is in line with known habitat preferences of earthworms (Rutgers et al., 2016; Hoeffner et al., 2021). Furthermore, Rutgers et al. (2016) demonstrated that earthworm abundance and diversity were closely linked to the presence or absence of certain types of land use such as grasslands, cultivated lands, forests, moorlands, and vineyards. This is because earthworms have certain habitat preferences. In our case, our models predict a high number of earthworms (80 ind./m²), average biomass (40 g/m²), and approximately one to two species in grasslands compared to other land uses. Our models also predicted a negative effect of crops and vineyards. This is coherent because it is known that land use intensity can have a significant impact on earthworm communities (Smith et al., 2008; Spurgeon et al., 2013).

Our study also revealed that, after land use, spatial variables were the most important, followed by climatic variables, and then soil variables.

Indeed, CaCO₃ had positive effects on earthworms below 10 g·kg⁻¹ and negative effects beyond this threshold. This result could be explained by the fact that low amounts of CaCO₃ would favor the alkalinization of food and facilitate the passage of food at the level of Morren's glands, while high amounts of CaCO₃ would be toxic to earthworms.

## Influences of scale on earthworm biodiversity: climatic and soil factors

However, according to Phillips et al. (2019), the most influential variables on earthworms are precipitation and annual temperature. Indeed, it seems that the determining factors of earthworm biodiversity depend on the scale (Salako et al., 2023). Climatic parameters would play a crucial role at the global and continental scales (Rutgers et al., 2016; Phillips et al., 2019), while soil-related factors become more important at smaller spatial scales (Palm et al., 2013; Marchán et al., 2015). In our study, the importance of climate was mainly represented by annual precipitation, while calcium carbonate (CaCO₃) was the most significant soil variable for earthworms.

We are aware that there are other variables not included in this study that can strongly influence earthworms. This is the case, for example, of agricultural practices such as tillage (Crittenden et al., 2014; Ernst & Emmerling, 2009; Pelosi, Pey, et al., 2014), pesticides (Pelosi, Barot, et al., 2014; Maggi & Tang, 2021) and fertilization (Leroy et al., 2008; Niswati et al., 2022), which are well known for their impact on earthworms. However, these variables related to agricultural practices are not available in the form of high-resolution maps at the scale of metropolitan France. Therefore, we did not include them in our models, which would hinder the production of maps based on model predictions. In summary, although we are not able to consider all potential factors influencing earthworm distribution, our study includes a large number of the most important environmental variables known to affect earthworms (Edwards & Arancon, 2022).

Our models predicted an average taxonomic richness of 5 species per plot, with a maximum of 10 species. These predictions fall within the range of intermediate values, in accordance with previous studies. Indeed, Phillips et al. (2019) et Fourcade & Vercauteren (2022) reported a maximum richness of 6 species per plot. Rutgers et al. (2016) observed up to 7 species per plot, while Zeiss et al. (2024) reported richness of up to 18 species. We observed that the earthworm community was more abundant in the northwest and center of the country compared to the east. This result is consistent with the findings of Zeiss et al. (2024), who observed that earthworm species richness was high in the west-central part of Europe and low in the northeast. Furthermore, Rutgers et al. (2016) asserts that the large-scale distribution of earthworm densities appears through positive correlations with latitude, longitude, and climatic factors at the European scale. In our study, this result could be explained by differences in land use: more grasslands in the west, while the east is characterized by forests and vineyard plots.

We predicted the distribution of earthworms in areas for which we were able to extract model predictors through high-resolution maps, as well as in some additional areas to produce continuous maps. These additional areas were mainly located in plots with land uses that are not input variables of the models. These include highly artificialized areas (industrial or commercial zones, airports), wetlands, water surfaces, or agroforestry territories. The expected distributions of earthworms in these areas were indirectly derived by interpolation and should therefore be interpreted with caution. The maps were created primarily to raise awareness and protect soil biodiversity and associated ecosystem services.

# Conclusions

In this study, we developed predictive models using a comparative approach between traditional regression models (GLM, GAM) and machine learning algorithms (RF, GBM, and ANN) to identify the best model for predicting the earthworm community in metropolitan France. Generalized boosted models and random forests showed the best predictive performances. The notable initial results we obtained pertain to the estimated precision of our models, which enable prediction of total biomass with an R² of 0.35 (RMSE = 8.76 g/m²), total abundance with an R² of 0.43 (RMSE = 25 ind./m²), and total taxonomic richness with an R² of 0.59 (RMSE = 1.7 species per plot). Our study highlighted that land use was the most important variable for earthworms, followed by spatial, climatic, and soil variables. Additionally, our study created prediction maps of the earthworm community in France, providing reference data and raising awareness about soil biodiversity protection and associated ecosystem services. To complement this study and obtain a comprehensive overview of the earthworm community in France, it would be relevant to develop additional models to predict species abundance or presence-absence, as well as the abundance and biomass of specific ecological categories of earthworms.

# Bibliography

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

Vaughan, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, *42*(4), 720‑730. https://doi.org/10.1111/j.1365-2664.2005.01052.x

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 |