UNIVERSITE DE RENNES

UFR Sciences de la Vie et de l'Environnement

Master de biologie, écologie et évolution

Parcours modélisation en écologie

Internship report

Predicting earthworm diversity and distribution: A comparative approach at national scale using multiple algorithms

---

Abdourahmane DIALLO

Presented on june 19, 2024, in Rennes

Academic year 2023 – 2024

Host institutions:	University supervisor:
UMR 6553 ECOBIO –Université de Rennes UMR 1347 Agroécologie (INRAE de Dijon)	Mrs. Marie-Pierre Etienne
Internship supervisors:	Program coordinators:
Mr. Daniel Cluzeau Mr. Kevin Hoeffner Mr. Walid Horrigue	Mr. Cédrick Wolf Mr. Frederic Hamelin Mrs. Marie-Pierre Etienne

Internship from january 01 to june 30, 2024

Acknowledgements

I would like to thank my three internship supervisors, Mr. Daniel Cluzeau, Mr. Kevin Hoeffner, and Mr. Walid Horrigue, for the time you took to mentor me, and for your invaluable advice and feedback.

I would also like to extend my gratitude to all my colleagues in the LMCU team at UMR ECOBIO and the BIOCOM team at UMR Agroecology for their kindness and good spirits.

I am grateful to all my classmates for their insightful discussions.

Special thanks to my friend Mamadou Saidou Diallo and to my family for their encouragement and attention to my work.

Table of Contents

Acknowledgements ii

1 Introduction 1

2 Materials and methods 3

2.1 Earthworm and land use data collections 3

2.2 Environmental data collection 5

2.3 Modeling strategy 6

2.3.1 Overview Detection Mapping Application Protocol (ODMAP) 6

2.3.2 Variable selection, importance and effects 7

2.3.3 Model fitting and calibration 9

2.3.4 Model evaluation and selection 11

2.3.5 Predictions, interpolation and mapping of earthworm communities 12

1 Results 13

1.1 Model performance 13

1.2 Contribution of the explanatory variables 14

1.3 Effects of changes in environmental variables when predicting earthworm communities 15

1.4 Spatial distribution of earthworms at the scale of France 16

2 Discussion 19

2.1 Improved performance of ensemble methods when predicting earthworm communities 19

2.2 Importance of land use, climate and soil properties 20

2.3 Uncertainty 22

Conclusion 24

References 1

Appendices 1

1 Introduction

Soil fauna provides a wide range of ecosystem services (Bardgett and Van der Putten, 2014; FAO, 2020), and among them, earthworms are referred to as "ecosystem engineers" (Jones et al., 1994) as they act on other soil organisms by modifying soil properties (Blouin et al., 2013): they contribute to soil structure development (Lavelle et al., 1997; Sharma et al., 2017; Edwards and Arancon, 2022), water infiltration and water retention through burrows and casts deposited in the soil (Capowiez et al., 2014; Cunha et al., 2016), organic matter dynamic 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 et al., 2014; Marchán et al., 2015; Gabriac et al., 2022) and regional scales (Marchán et al., 2016, 2021; Marchán and 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 and Vercauteren, 2022; Salako et al., 2023). 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 texture), and latitude. Recently, another study by Phillips et al. (2019) on 6 928 sites distributed across 57 countries demonstrated that at a global scale, climatic factors (annual mean temperature, temperature seasonality, temperature annual range, annual precipitation, and precipitation seasonality) are the most important environmental filters in shaping earthworm communities than soil properties (pH, organic carbon, soil clay content, soil silt content, and CEC) or habitat cover. However, these studies are limited by using a single type of predictive modeling algorithm: generalized linear models (GLM) for Rutgers et al. (2016) and generalized linear mixed models (GLMM) for Phillips et al. (2019).

Commented [KH1]: Idem pas sur que les ref soient adaptées…
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. Sci. Rep. 4.

Dans cette ref on peut voir que la composition des com vdt influence la croissance des plantes…

Commented [AD2R1]: J'ai ajouter les liens de ces ref dans la biblio et ils sont adapté pour montrer les effets de ma perte de la biodiversité...

Indeed, several studies have shown that the results can vary considerably depending on the type of predictive model used (Elith et al., 2006; Elith and Graham, 2009; Oppel et al., 2012; Li and Wang, 2013; Salako et al., 2023). Moreover, other studies have revealed that traditional regression models such as GLMs or GLMMs can be less robust, as they are sensitive to extreme values and handle complex relationships less effectively. Therefore, comparing different predictive models appears to be a crucial step in ensuring the quality of predictions (Salako et al., 2023). It is also important to highlight that a model might predict total abundance well but not necessarily total biomass or earthworm total taxonomic richness. Therefore, it would be relevant to compare multiple algorithms to find the best model for each earthworm parameter (total abundance, total biomass, and total taxonomic richness).

In the case of 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 for 44 earthworm species. Their research shows a decline in functional richness between 1960 and 2012. Additionally, their models predict that this reduction could continue in the future across different temporal periods and climate change scenarios. However, this latter study is based on the presence/absence of earthworm data collected in the 1960s by Bouché (1972), and therefore, they no longer reflect the current earthworm community assemblage in France. Furthermore, the study of earthworm communities should include variables describing species total abundance, total biomass and total taxonomic richness, which are essential parameters for biodiversity assessments. Relying solely on functional diversity is not sufficient to fully evaluate the status of earthworm communities, as, for instance, two communities may be identical in terms of diversity but differ in terms of density or biomass (Groves, 2022).

To understand earthworm diversity, distribution, and the environmental factors influencing this distribution, several tools have been developed in recent years. For example, species distribution models (SDMs) are models based on ecological niches that allow for modeling correlations between species or communities and environmental factors (Elith and 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 and Wang, 2013; Valavi et al., 2021). The comparative approach proposed by Salako et al. (2023) using multiple SDMs to predict the geographical distribution area

and diversity of earthworms in Germany is promising. Therefore, the main aim of this study is to compare different modeling algorithms to predict the earthworm community composition and distribution in France. Specifically, we sought to: (i) quantify and rank the influence of environmental factors (land use, soil properties, location, and climate data) on total abundance (individuals per m^2), total biomass (g per m^2), and total taxonomic richness of earthworms (number of taxa in the plot) in France (excluding Corse), and (ii) to build predicting map using these same earthworm parameters but based on environmental factors. To address these objectives, we selected five SDM algorithms and compared their predictive performance. 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 and Wang, 2013; Rutgers et al., 2016; Valavi et al., 2021; Salako et al., 2023). This comparative approach reduces uncertainty and identifies the best model for each earthworm variable (total abundance, total biomass, and total taxonomic richness).

2 Materials and methods

2.1 Earthworm and land use data collections

We used data from the LandWorm project (2023-2026 FRB-MTE-OFB), which aims to quantify the effects of land use and management on earthworm communities, taking into account the soil and the climate heterogeneity on a national scale in France. This project seeks to understand and predict earthworm community assembly and identify land management practices that are favorable to earthworms. The database created from the LandWorm project contained approximately 8,019 earthworm observations. Thus, we have used this database in this study by applying different filters.

First, we remove all observations for which the year of sampling and/or GPS coordinates were not recorded, as these details were necessary for collecting environmental data (section 2.2). We also excluded observations located outside of France. Subsequently, we selected the six main land cover types with sufficient observations in our database (see appendix 1), corresponding to the level 3

nomenclature of Corine Land Cover (CLC, https://land.copernicus.eu/pan-european/corine-land-cover). The other land cover types were not included due to the lack of available data. Additionally, we could not distinguish between broad-leaved forests and mixed forests, so we grouped these two types under a single land use category: "Forest" (all types). This decision was made considering the significant impact of land use on earthworm populations, as previously highlighted by Spurgeon et al. (2013).

Then we selected only observations conducted using the hand sorting protocol (ISO 23611-1:2018) and/or the application of a chemical expellant (formaldehyde or allyl isothiocyanate; ISO 23611-1:2006) (see appendix 2). Earthworm sampling was primarily conducted in spring, corresponding to the period of maximum earthworm activity. This choice was made to limit the influence of the sampling protocol (Rutgers et al., 2016) on the results.

Finally, we filtered the database to retain only community-level data: total abundance (ind./m²), total biomass (g/m²), and total taxonomic richness. For each of these three variables, we ecologically and statistically removed observations with outlier values (Grubbs test, p-value < 0.05). Thus, earthworm abundance ranged from 0 to 1 075 ind./m² with a mean of 228 ind./m². Total biomass ranged from 0 to 364 g/m² with a mean of 89 g/m². Total richness (at the species level) ranged from 0 to 16 species per plot with a mean of 5 species. Ultimately, our database included 3 822 observations (Fig. 1), of which 48 % did not have total biomass data. In summary, this step allowed us to obtain the six land cover types, GPS coordinates, and the three earthworm variables.

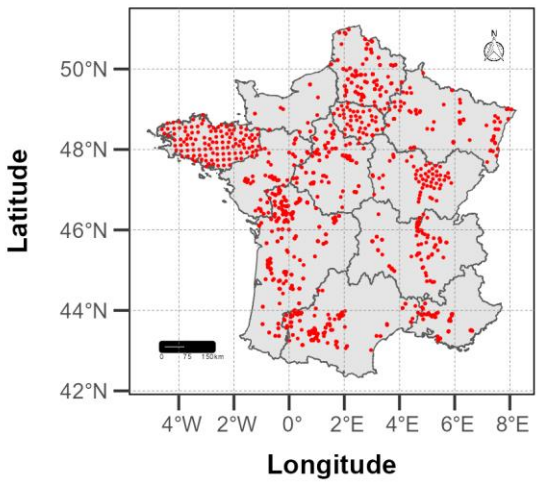


Fig. 1: *Map of the study area (France (excluding Corse)) showing the location of earthworm sampling sites.*

2.2 Environmental data collection

In addition to land use data and GPS coordinates, we also compiled information about abiotic 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 (see https://chelsa-climate.org/wp-admin/download-page/CHELSA_tech_specification_V2.pdf) 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 properties, 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 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 collected data from three soil 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 LUCAS database (Land Use/Cover Area Frame Statistical Survey; Ballabio et al., 2019) to collect information on the following variables: cation exchange capacity (CEC), calcium carbonate ($CaCO_3$), C/N ratio, nitrogen (N), phosphorus (P), potassium (K), and soil pH in H2O. These variables were available at a resolution of 250 m on a European scale and based on 2009/2012 LUCAS data. Initially, we referenced all preselected variables to the World Geodetic System (WGS84) coordinate system, cropped and masked them to match the geographical boundaries of France. To standardize the variables and match them to the same resolution, we resampled or disaggregated them to a resolution of 30 arc-seconds, approximately 800 m in France. These steps were carried out using Python with the GDAL module (https://pypi.org/project/GDAL/). We then removed all lines with NA (227 observations) and/or outliers (19 observations; Grubbs test, p-value < 0.05). Thus, our final database contained 3 576 observations.

2.3 Modeling strategy

2.3.1 Model workflow

Our modeling strategy followed the ODMAP (Overview, Data, Model, Assessment and Prediction) protocol recommended by Zurell et al. (2020), and all steps are detailed in Fig. 2 (see appendix 3). Briefly, we collected data on earthworms and environmental variables. We then merged, cleaned, and transformed them. Environmental variables (quantitative) were centered (on the mean) and scaled (by the standard deviation), the six land cover types were transformed into dummy variables (binary), and the earthworm variables (total abundance and total biomass) were transformed using the square root to approximate a Gaussian distribution. After selecting the most important variables, we partitioned the dataset into training and test sets. We then calibrated the models on the training data, evaluated the models on the test data, compared the models to select the best ones, and finally performed predictions and interpolations. All modeling steps were performed using R software version 4.3.1, 2023 (R Core Team, 2023).

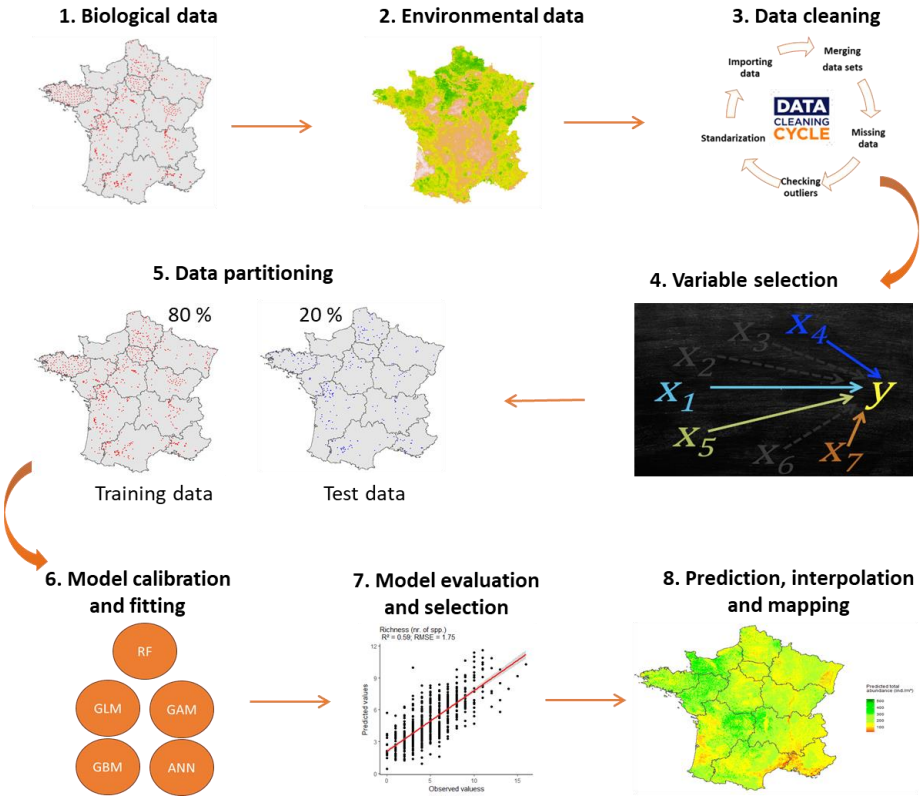


Fig. 2: *Modeling strategy according to the ODMAP protocol: (1) biological data collection, (2) environmental data collection, (3) data cleaning, (4) variable selection, (5) data partitioning, (6) model calibration and fitting, (7) model evaluation and selection, and (8) Prediction, interpolation and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

2.3.2 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, and include correlated variables, and variable interactions, all of which improve model performance (Breiman, 2001). To reduce the number of variables and avoid overfitting (Vaughan and Ormerod, 2005), we identified the ten most important variables on earthworms using a permutation procedure (Fourcade and Vercauteren, 2022; Zeiss et al., 2024; Table 1): land use, longitude, latitude, calcium carbonate, nitrogen, phosphorus, clay and

silt content, isothermality, 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 effects of variables (Casalicchio et al., 2024). To study the effects of each variable, we used the accumulated local effects (ALE), which describe how the model's predictions change within a small "window" of the considered variable. ALE effects are a faster and more unbiased alternative to partial dependence plots (PDP; Apley and Zhu, 2019).

Table 1: *Abbreviations and description of variables used in predicting earthworm communities. For the land use (boolean-type) the total sum is provided summarizing the total data set with 3 576 observations. For the continuous variables, three descriptive parameters (minimum, mean and maximum) of the final data set are provided.*

Category	Abbr	Description	Units	Min	Mean	Max	Time scale	Reference
Land use	For	Forest (116)	Boolean				1990 - 2023	LandWorm database
	Gua	Green urban areas (535)	Boolean					
	Nag	Natural grasslands (111)	Boolean					
	Nial	Non-irrigated arable land (1683)	Boolean					
	Pmo	Pastures, meadows and other permanent grasslands under agricultural use (413)	Boolean					
	Viny	Vineyards (718)	Boolean					
Location	Long	Longitude	WGS84	-4.612	1.772	8.056		
	Lat	Latitude	WGS84	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		
	P	Phosphorus	mg·kg^{-1}	7.18	40.85	68.39	2009 - 2012	Ballabio et al. (2019)
	N	Nitrogen	g·kg^{-1}	0.87	1.99	3.84		

2.3.3 Model fitting and 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:

$$Y = glm(y \sim x1 + x2 + \cdots + xn, family =' gaussian', data = data)$$

Where *y* is the response variable (total abundance, total biomass, or total taxonomic richness) and x represents the n explanatory variables.

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

$$Y = gam(y \sim s(x1) + s(x2) + \cdots + s(xn), family =' gaussian',$$
$$method = 'REML', data = data)$$

Where *y* is the response variable (total abundance, total biomass, or total taxonomic richness) and x represents the n explanatory variables.

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

$$Y = randomForest\big(data[-rep.var], data\big[[rep.var]\big], mtry = 3, ntree = 500, maxnodes = NULL, importance = TRUE\big)$$

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 (NULL and 2 to 15 in increments of 1). We selected the model with the highest R^2 and lowest RMSE and MAE from all models.

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

$$Y = gbm(y\sim., data = data, distribution =' gaussian', n.trees = 1000, shrinkage = 0.01, interaction.depth = 5, n.minobsinnode = 10)$$

Where *n.trees* is the number of trees, *shrinkage* is the learning rate, *interaction.depth* is an integer specifying the maximum depth of each tree, 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 choose 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), maximum depth of trees (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 R^2 and lowest RMSE and MAE from all models.

We used the *Keras* package (Kalinowski et al., 2024a) with a sequential architecture for ANN. 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. The last layer consisted of a single neuron corresponding to the predicted variable (total abundance, total biomass, and total taxonomic 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 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 patience of 10 to monitor loss on the validation set and restore weights from the best model. We used the *tuning_run* (Kalinowski et al., 2024b) function to hyperparameterize the set of parameters explained above, and selected the model with the highest R^2 and lowest RMSE and MAE from all models.

2.3.4 Model evaluation and selection

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 and 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 for the large French dataset where 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 (R^2) between observations and predictions, mean absolute error (MAE), and root mean square error (RMSE) as performance evaluation metrics.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

where: n is the number of observations, y_i represents the observed value of observation (i) and $\widehat{y_i}$ represents the predicted value of observation (i).

The objective was to maximize R^2 and minimize MAE and RMSE.

2.3.5 Predictions, interpolation 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 total taxonomic richness). Initially, we sampled at a resolution of approximately 800 m across the entire French territory (excluding Corse). 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 displayed 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 and Graeler, 2023).

$$w(x) = \frac{A}{B}$$

$$A = \sum_{i=1}^{n}\frac{u_i}{\left(d(x,x_i)\right)^p}$$

$$B = \sum_{i=1}^{n}\frac{1}{\left(d(x,x_i)\right)^p}$$

Where: w is the predicted value, d is the distance, x is the unknown point, xi is the n-th know point, ui is the value of the known 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.

Interpolation was primarily conducted in areas located within parcels where land use is not an input variable for the models. This includes heavily urbanized areas (industrial or commercial zones, airports), wetlands, or agroforestry territories. Interpolation was also performed in areas where we could not extract the variables and/or soil properties.

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

3 Results

3.1 Model performance

The performances of the five models varied with an average R^2 of 0.33 (± 0.10 SD) for total 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 total 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 R^2 and the lowest RMSE, indicating that this algorithm provides the best prediction of the earthworm community structure. The evaluation of predictive model performances showed that GLM was the worst-performing model in predicting the total abundance, total biomass, and total taxonomic 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 R^2 of 0.43. Similarly, RF was also the best model for total biomass with an R^2 of 0.35. For total taxonomic richness, the best models were RF and GBM, both with identical R^2 values of 0.59.

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

Algorithms	Response variables	R^2	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

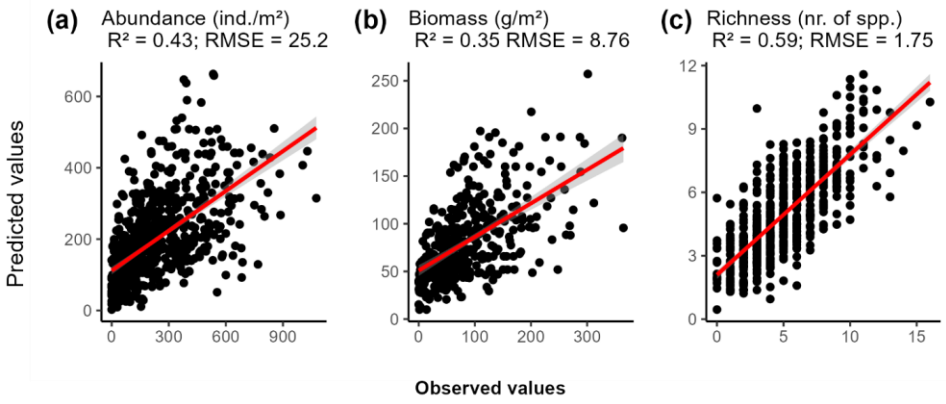


*Fig. **3:** Prediction on the validation dataset with the best algorithms for (a) total abundance (RF), (b) total biomass (RF), and (c) total taxonomic richness (GBM). The X-axis indicates the observed values, and the Y-axis indicates the predicted values. The red line represents the linear regression (trend) between the observed and predicted values, and the gray band indicates the confidence interval around the regression line.*

3.2 Contribution of the explanatory variables

During the adjustment of prediction models, the most important variable was land use (CLC; Fig. 4). When land use was permuted, the RMSE of the total abundance model increased on average by 1.68 ind./m² for total abundance, 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 ones for predicting

earthworm variables, 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 ($CaCO_3$) and nitrogen (N), respectively. $CaCO_3$ was also found to have a great influence on total taxonomic richness and total abundance compared to biomass.

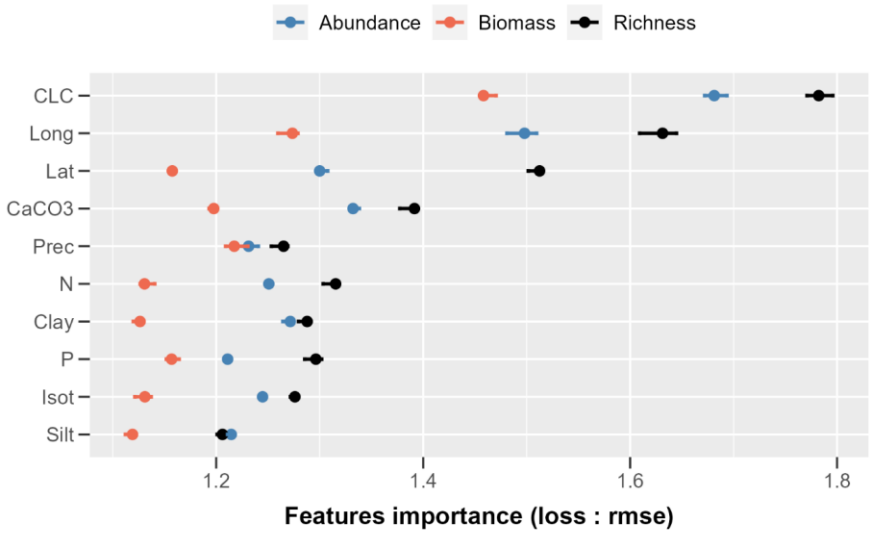


Fig. 4: *Importance of environmental explanatory variables for predicting total abundance, total biomass, and taxonomic richness of earthworms. Values represent the median increase in RMSE (plus or minus the 5% and 95% quantiles with n = 5) of model prediction after the permutation of variable values.*

3.3 Effects of environmental variables on earthworm communities

Analyses of Accumulated Local Effects (ALE) showed that land use had mixed effects on the earthworm community. For example, grassland plots (Nag and Pmo) had the highest earthworm total abundances and showed the highest predicted increase in total abundance by 80 ind./m². Similarly, plots located in urban green spaces (Gua) predicted total abundance increased by 50 ind./m². Plots in Nag, Pmo, and Gua had about one or two more species than any other land use types. ALE effects also showed that plots in vineyards (Viny), forests (For), and annual crops (Nial)

were associated with the lowest earthworm total abundance. Regarding biomass, only grasslands showed an increase in the predicted average weight of earthworms.

Spatial variables also exerted a strong influence on earthworms, with longitude having the most significant impact. Moving from west to east of France, earthworm total abundance and total taxonomic richness decreased, with a decrease of about 50 ind./m² and one species, respectively, whereas biomass showed very little variations. Increasing latitude did not influence total abundance between 43°N and 46°N but decreased between 46°N and 47.5°N before increasing beyond the latter latitude.

Regarding the effect of climatic variables, an increase in annual precipitation up to 700 $kg \cdot m^{-2} \cdot year^{-1}$ resulted in an increase in predicted total abundance by approximately 15 ind./m². Beyond this threshold, abundance remained stable. Similarly, an increase in annual precipitation up to 700 $kg \cdot m^{-2} \cdot year^{-1}$ led to an increase in total biomass, but biomass gradually decreased beyond this threshold. In contrast, total richness decreased when annual precipitation was below 900 kg·m-2·year-1 and increased above this threshold. Additionally, earthworms were little affected by the ratio between diurnal temperature variation and annual variation (Isot).

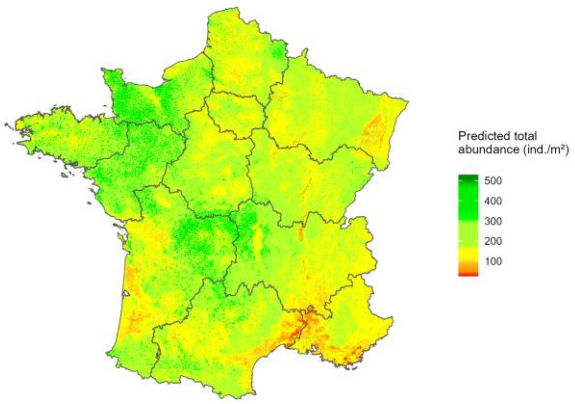
Regarding soil properties, increasing $CaCO_3$ up to approximately 10 $g \cdot kg^{-1}$ led to an increase in predicted total abundance by 25 ind./m² and predicted total biomass by 15 g/m². However, values above 10 $g \cdot kg^{-1}$ $CaCO_3$ led to a decrease in total abundance and total biomass. Increasing $CaCO_3$ from 0 to 100 $g \cdot kg^{-1}$ resulted in a decrease of one species, but total taxonomic richness remained stable beyond 100 g·kg−1. Soil nitrogen was positively correlated with total abundance, total biomass and total taxonomic richness. In addition, increasing nitrogen from 1 to 3 $g\text{-}kg^{-1}$ increased total abundance by 25 ind./m², total biomass by 10 g/m² and total taxonomic richness by one species. Soil texture (clay and silt content) did not influence predicted total abundance up to 40 $g \cdot kg^{-1}$, but higher values resulted in increases in total abundance, reaching an average predicted increase of 45 ind./m². The texture had little influence on total biomass. Phosphorus had a minor negative effect on earthworm communities.

3.4 Spatial distribution of earthworms at the scale of France

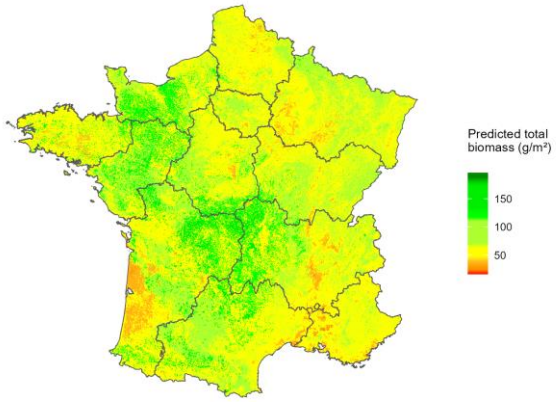
Figure 5 shows the predicted spatial distribution of the earthworm communities in France (excluding Corse). Total predicted total abundance varied from 0 to 530 ind./m² with a mean of

192 ind./m² per plot (Fig. 5a). The average predicted biomass was 72 g/m² (minimum = 0 and maximum = 196; Fig. 5b), while predicted 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 total abundance map and the total taxonomic richness map. It can be observed that the soils in the southeastern part of France have very low (total abundance < 100 ind./m² and total taxonomic richness < 2 species) to low diversity (total abundance < 200 ind./m² and total taxonomic richness < 4 species), while the southwestern part has a medium (total abundance < 300 ind./m² and total taxonomic richness < 6 species) diversity.

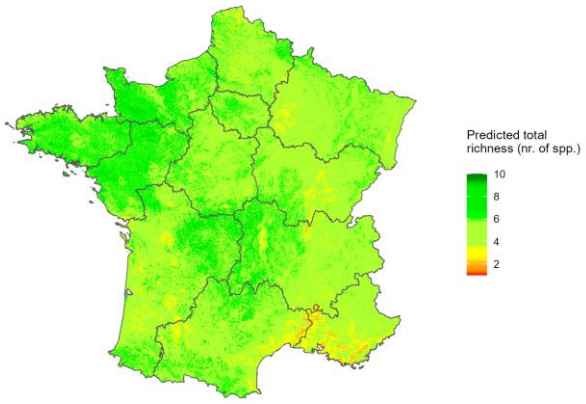
(a)



(b)



(c)



(d)

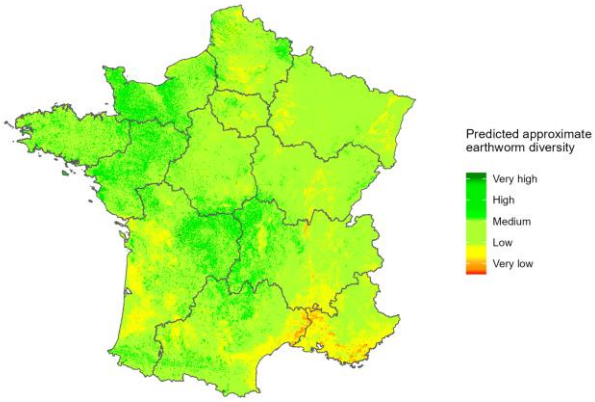


Fig. 5: *Predicted spatial distribution of (a) total abundance (ind./m²), (b) total biomass (g/m²), (c) total taxonomic richness (number of taxa per plot), and (d) approximate diversity of earthworms. Legend: Very low (total abundance < 100 ind./m² and total taxonomic richness < 2 species); Low (total abundance < 200 ind./m² and total taxonomic richness < 4 species); Medium (total abundance < 300 ind./m² and total taxonomic richness < 6 species); High (total abundance < 400 ind./m² and total taxonomic richness < 8 species); Very high (total abundance > 500 ind./m² and total taxonomic richness > 10 species).*

4 Discussion

The adjustment of five different model algorithms to predict the distribution of earthworms using land use, spatial, climatic, and soil variables allowed us to determine the accuracy of the models tested. Our comparative approach showed that earthworm total abundance was predicted with an R^2 of 0.43 (RMSE = 25 ind./m²), biomass with an R^2 of 0.35 (RMSE = 8.76 g/m²), and total taxonomic richness with an R^2 of 0.59 (RMSE = 1.7 species per plot). We observed that the spatial distribution of earthworms in France is primarily driven by land use and spatial variables, particularly longitude. All stages of the modeling strategy were reported following the ODMAP protocol (see appendix 3), as recommended by Zurell et al. (2020).

4.1 Improved performance of ensemble methods when predicting earthworm communities

Our results demonstrate that random forest (RF) models and generalized boosted models (GBM) provided better results of predicted earthworm total abundance, total biomass, and total taxonomic richness 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 also observed that RF had very high predictive performances. Similarly, Salako et al. (2023) using different prediction algorithms concluded that RF was the most effective algorithm for predicting earthworm communities in Germany. 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 and Wang, 2013). Their effectiveness also stems from their ability to capture better nonlinear relationships between explanatory and response variables, robustness against outliers, and better handling of variable interactions (Breiman, 2001). Considering the high predictive potential of RF and GBM, it would be advisable to use them in ecological studies instead of or in addition 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). From this, it is clear that model performances would be improved if several databases can be merged and standardized (e.g., derived from different research units and across countries). Another limitation of RF and GBM is their low degree of interpretability, but this is becoming less true as numerous

tools now exist to better understand and interpret machine learning models. For example, the "iml" package provides very useful 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).

4.2 Importance and effects of environmental variables

Our results demonstrate that land use was the most important variable on earthworms. This is consistent with the findings of Rutgers et al. (2016); Fourcade and Vercauteren (2022) and Salako et al. (2023), which showed that land use strongly affects the distribution of earthworm communities. All predictive models used in this study predicted the positive effects of grasslands (Pmo and Nag) on earthworms. This result is in agreement with earthworms exhibiting habitat preferences (Rutgers et al., 2016; Hoeffner et al., 2021). For example, Rutgers et al. (2016) demonstrated that earthworm total abundance and diversity were better explained by the presence/absence of certain types of land use such as grasslands, cultivated lands, forests, moorlands, and vineyards. Our models predicted a high number of earthworms (80 ind./m²), average biomass (40 g/m²), and approximately one to two species in grasslands compared to the other land uses. These results can be explained by the fact that grassland plot are conducive to the development and growth of earthworms, providing more food resources, refuge from predators, and protection against extreme climatic events (Iordache, 2010; Zhu and Zhu, 2015; Niswati et al., 2022). However, many parameters can mitigate this effect of grasslands: for example, the use of inputs, frequency of grazing by cattle, stocking density, mowing management and seasonality(Postma-Blaauw et al., 2006; van der Wal et al., 2009; Cluzeau et al., 2012). 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 negative impact on earthworm communities (Smith et al., 2008; Spurgeon et al., 2013). These negative effects of intensive agriculture can be mainly attributed to the significant impact of soil tillage, fertilization, and pesticides (Pelosi et al., 2013, 2014; Maggi and Tang, 2021; Niswati et al., 2022). For example, in a global meta-analysis, Briones and Schmidt (2017) showed that disturbing the soil less (no-tillage and conservation agriculture) significantly increased earthworm abundance (mean increase of 137% and 127%, respectively) and biomass (196% and 101%, respectively) compared to conventional ploughing. The low earthworm community in forests could be explained by the fact that food resources may not be easily

assimilable by earthworms due to the presence of lignin, which makes the food resources harder to degrade.

After land use, the variables with the greatest influence on earthworms were spatial variables. We observed that the earthworm community was more abundant in the northwest and center of France compared to the east of the country. This result is consistent with the findings of Zeiss et al. (2024), who also observed that earthworm total taxonomic richness was high in the west-central part of Europe and low in the northeast. Furthermore, Rutgers et al. (2016) concluded that the large-scale distribution of earthworm densities is positively correlated with latitude, longitude, and climatic factors at the European scale. These discrepancies with our study could be explained by differences in land use: more grasslands plots were sampled in the west, while the east forests and vineyard plots were more abundant.

Regarding climate, our study demonstrated the positive effect of precipitation on earthworm communities, confirming the results of Rutgers et al. (2016), Salako et al. (2023) and Zeiss et al. (2024). It has been shown that various climatic factors, such as temperature, precipitation, soil moisture, and extreme weather events like droughts and floods, alter the composition and functioning of soil communities (Singh et al., 2019). Moreover, at global scales and according to Phillips et al. (2019), the most influential variables on earthworms (abundance and richness) are precipitation and annual temperature. This is because climatic parameters play a crucial role at large spatial scales (Rutgers et al., 2016; Phillips et al., 2019), while soil-related factors become more important at local spatial scales (Palm et al., 2013; Marchán et al., 2015). We found that increasing precipitation up to approximately 700 $kg \cdot m^{-2} \cdot year^{-1}$ increased total abundance and total biomass. This result could be explained by the fact that up to this threshold, moisture conditions were suitable for earthworms (Edwards and Arancon, 2022). Beyond 700 $kg \cdot m^{-2} \cdot year^{-1}$, total abundance remained constant, while total biomass gradually decreased. This observation could be due to the fact that the increased frequency and intensity of extreme precipitation events can lead to mortality by altering the life cycle and nutrition of soil animals (Bates et al., 2008; Thakur et al., 2018), as well as making soils more vulnerable to erosion (Nearing et al., 2004) and impairing their habitat function for soil fauna (Singh et al., 2019). However, we believe that the effect of precipitation beyond 900 $kg \cdot m^{-2} \cdot year^{-1}$ is not very reliable because we did not have enough observations with high precipitation levels. This would have influenced the accumulated local

effects (ALE), which are sensitive to the number of observations and the number of intervals chosen for each environmental variable.

Although our study confirmed the combined role of land use, spatial variables, and precipitation, it also identified soil variables such as $CaCO_3$, soil texture, and nitrogen as important factors. Indeed, $CaCO_3$ had positive effects on earthworms below 10 $g \cdot kg^{-1}$ and negative effects beyond this threshold. This result could be explained by the fact that low amounts of $CaCO_3$ would favor the alkalinization of food and facilitate the passage of food at the level of Morren's glands, while high amounts of $CaCO_3$ would be toxic to earthworms. We found that soil texture (clay and silt content) above 40 $g \cdot kg^{-1}$ increased the total abundance of earthworms. This result is consistent with the conclusions of Edwards and Arancon (2022), who state that silts generally favor earthworm populations. Coarse elements such as sands are easier to burrow through but harder to ingest and are abrasive, whereas finer and denser textures slow down movement but are more easily ingested by geophages (Perreault and Whalen, 2006). However, other factors such as the presence of organic matter, vegetation, management practices, and soil pH, as well as their interactions, can greatly influence earthworm abundance and activity (Hoeffner et al., 2021; Edwards and Arancon, 2022). In this study, increasing nitrogen was positively correlated with all three earthworm variables. This increase in the earthworm community as nitrogen levels rise could be explained by the fact that increased nitrogen boosts primary production, which, for earthworms, increases food supply and provides refuge from predators and extreme climatic events (Iordache, 2010; Zhu and Zhu, 2015; Niswati et al., 2022)

4.3 Uncertainty

We are aware that there are other variables not included in this study that can strongly influence earthworms, such as tillage (Ernst and Emmerling, 2009; Crittenden et al., 2014; Pelosi et al., 2014; Briones and Schmidt, 2017), pesticides (Pelosi et al., 2013; Maggi and Tang, 2021) and fertilization (Leroy et al., 2008; Niswati et al., 2022). However, these variables are not available in high-resolution maps at national scale. Despite this shortcoming, our study includes a large number of the most important environmental variables known to affect earthworms (Edwards and Arancon, 2022).

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.

It is important to note that the explanatory variables related to climate and soil properties 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 noted by Rutgers et al. (2016) and Salako et al. (2023). Indeed, the quality of the data used during model training greatly influences their performance. For this reason, we chose not to use external databases for land use because this information was well documented in our database, and thus, preventing the temporal gap between collecting earthworm observations and land use changes over time. Another limitation of our study lies in the restricted selection of land use types (six types of level 3 land use defined by Corine Land Cover) due to the lack of available data. Additionally, we could not distinguish between broad-leaved forests and mixed forests and 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 R^2 for total biomass, especially since some earthworm observations had no biomass.

Conclusion

In this study, we developed 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 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 R^2 of 0.35 (RMSE = 8.76 g/m²), total abundance with an R^2 of 0.43 (RMSE = 25 ind./m²), and total taxonomic richness with an R^2 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. 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 total abundance or presence-absence, as well as the total abundance and total biomass of specific ecological categories of earthworms.

References (under verification)

Apley, D.W., Zhu, J., 2019. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. https://doi.org/10.48550/arXiv.1612.08468

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, 505–511. https://doi.org/10.1038/nature13855

Bates, B., Kundzewicz, Z., Wu, S., 2008. Climate Change and Water. Intergovernmental Panel on Climate Change Secretariat.

Blouin, M., Hodson, M.E., Delgado, E.A., Baker, G., Brussaard, L., Butt, K.R., Dai, J., Dendooven, L., Peres, G., Tondoh, J.E., Cluzeau, D., Brun, J.-J., 2013. A review of earthworm impact on soil function and ecosystem services: Earthworm impact on ecosystem services. Eur. J. Soil Sci. 64, 161–182. https://doi.org/10.1111/ejss.12025

Bouché, M.B., 1972. Lombriciens de France. Ecologie et systématique. INRA Editions.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Briones, M.J.I., Schmidt, O., 2017. Conventional tillage decreases the abundance and biomass of earthworms and alters their community structure in a global meta-analysis. Glob. Change Biol. 23, 4396–4419. https://doi.org/10.1111/gcb.13744

Capowiez, Y., Sammartino, S., Michel, E., 2014. Burrow systems of endogeic earthworms: Effects of earthworm abundance and consequences for soil water infiltration. Pedobiologia 57, 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, 59–67. https://doi.org/10.1038/nature11148

Casalicchio, G., Molnar, C., Schratz, P., 2024. iml: Interpretable Machine Learning.

Cluzeau, D., Guernion, M., Chaussod, R., Martin-Laurent, F., Villenave, C., Cortet, J., Ruiz-Camacho, N., Pernin, C., Mateille, T., Philippot, L., Bellido, A., Rougé, L., Arrouays, D., Bispo, A., Pérès, G., 2012. Integration of biodiversity in soil quality monitoring: Baselines for microbial and soil fauna parameters for different land-use types. Eur. J. Soil Biol., Bioindication in Soil Ecosystems 49, 63–72. https://doi.org/10.1016/j.ejsobi.2011.11.003

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. Appl. Soil Ecol., XVI International Colloquium on Soil Zoology & XIII International Colloquium on Apterygota, Coimbra, 2012 – Selected papers 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 Sci. 181, 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. Appl. Soil Ecol. 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, New York, NY. https://doi.org/10.1007/978-0-387-74943-3

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x

Elith, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M., E. Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. 40, 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. Eur. J. Soil Biol. 45, 247–251. https://doi.org/10.1016/j.ejsobi.2009.02.002

FAO, 2020. FAOSTAT [WWW Document]. URL https://www.fao.org/faostat/fr/#data/QCL (accessed 5.10.23).

Fourcade, Y., Vercauteren, M., 2022. Predicted changes in the functional structure of earthworm assemblages in France driven by climate change. Divers. Distrib. 28, 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. https://doi.org/10.1101/2022.10.13.512055

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.

Groves, C.P., 2022. Biogeographic region | Definition, Features, Locations, & Facts | Britannica [WWW Document]. URL https://www.britannica.com/science/biogeographic-region (accessed 4.17.24).

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

Hijmans, R.J., Elith, J., 2019. Spatial Distribution Models.

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. Appl. Soil Ecol. 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, 105–108. 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. Ecol. Indic. 64, 203–211. https://doi.org/10.1016/j.ecolind.2015.12.004
Iordache, M., 2010. Abundance of earthworms under fertilization with organo-mineral fertilizers in a chernozem from west of Romania. J. Food Agric. Environ. 10, 1103–1105.
Jones, C.G., Lawton, J.H., Shachak, M., 1994. Organisms as Ecosystem Engineers, in: Samson, F.B., Knopf, F.L. (Eds.), Ecosystem Management: Selected Readings. Springer, New York, NY, pp. 130–147. https://doi.org/10.1007/978-1-4612-4018-1_14
Kalinowski, T., Falbel, D., Allaire, J.J., Chollet, F., RStudio, Google, Tang [ctb, Y., cph, Bijl, W.V.D., Studer, M., Keydana, S., 2024a. keras: R Interface to "Keras."
Kalinowski, T., Falbel, D., Allaire, J.J., RStudio, https://d3js.org/), M.B. (D3 library-, http://c3js.org/), M.T. (C3 library-, library), jQuery F. (jQuery, inst/views/components/jquery-AUTHORS.txt), jQuery contributors (jQuery library; authors:, plugin), S.B. (jQuery visibilityChanged, https://materializecss.com/), M. (Materizlize library-, https://vuejs.org/), Y.Y. (Vue js library-, https://github.com/kpdecker/jsdiff/), K.D. (jsdiff library-, https://diff2html.xyz/), R.F. (diff2html library-, https://highlightjs.org/), I.S. (highlight js library-, library), Y.P. (highlightjs-line-numbers, 2024b. tfruns: Training Run Tools for "TensorFlow."
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. Sci. Data 4, 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. Eur. J. Soil Biol.
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, 139–150. https://doi.org/10.1016/j.pedobi.2008.07.001
Li, X., Wang, Y., 2013. Applying various algorithms for species distribution modelling. Integr. Zool. 8, 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 Secur. 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). J. Zool. Syst. Evol. Res. 59, 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, 337. 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). Eur. J. Soil Biol. 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 Biol. Biochem. 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
Nearing, M.A., Pruski, F.F., O'Neal, M.R., 2004. Expected climate change impacts on soil erosion rates: A review. J. Soil Water Conserv. 59, 43–50.
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 Conf. Ser. Earth Environ. Sci. 1018, 012012. https://doi.org/10.1088/1755-1315/1018/1/012012
Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A.F., Miller, P.I., Louzao, M., 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biol. Conserv., Seabirds and Marine Protected Areas planning 156, 94–104. https://doi.org/10.1016/j.biocon.2011.11.013
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, 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.
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. Appl. Soil Ecol., XVI International Colloquium on Soil Zoology & XIII International Colloquium on Apterygota, Coimbra, 2012 – Selected papers 83, 79–87. https://doi.org/10.1016/j.apsoil.2013.10.005
Pelosi, Toutous, L., Chiron, F., Dubs, F., Hedde, M., Muratet, A., Ponge, J.-F., Salmon, S., Makowski, D., 2013. Reduction of pesticide use can increase earthworm populations in wheat crops in a European temperate region. Agric. Ecosyst. Environ. 181, 223–230. https://doi.org/10.1016/j.agee.2013.10.003
Perreault, J.M., Whalen, J.K., 2006. Earthworm burrowing in laboratory microcosms as influenced by soil temperature and moisture. Pedobiologia 50, 397–403. https://doi.org/10.1016/j.pedobi.2006.07.003
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., Mathieu, J., Mulder, C., van der Putten, W.H., Ramirez, K.S., Rillig, M.C., Russell, D., Rutgers, M., Thakur, M.P., de Vries, F.T., Wall, D.H., Wardle, D.A., Arai, M., Ayuke, F.O., Baker, G.H., Beauséjour, R., Bedano, J.C., Birkhofer, K., Blanchart, E., Blossey, B., Bolger, T., Bradley, R.L., Callaham, M.A., Capowiez, Y., Caulfield, M.E., Choi, A., Crotty, F.V., Crumsey, J.M., Dávalos, A., Diaz Cosin, D.J., Dominguez, A., Duhour, A.E., van Eekeren, N., Emmerling, C., Falco, L.B., Fernández, R., Fonte, S.J., Fragoso, C., Franco, A.L.C., Fugère, M., Fusilero, A.T., Gholami, S., Gundale, M.J., López, M.G., Hackenberger, D.K., Hernández, L.M., Hishi, T., Holdsworth, A.R., Holmstrup, M., Hopfensperger, K.N., Lwanga, E.H., Huhta, V., Hurisso, T.T., Iannone, B.V., Iordache, M., Joschko, M., Kaneko, N., Kanianska, R., Keith, A.M., Kelly, C.A., Kernecker, M.L., Klaminder, J., Koné, A.W., Kooch, Y., Kukkonen, S.T., Lalthanzara, H., Lammel, D.R., Lebedev, I.M., Li, Y., Jesus

Lidon, J.B., Lincoln, N.K., Loss, S.R., Marichal, R., Matula, R., Moos, J.H., Moreno, G., Morón-Ríos, A., Muys, B., Neirynck, J., Norgrove, L., Novo, M., Nuutinen, V., Nuzzo, V., Mujeeb Rahman P, Pansu, J., Paudel, S., Pérès, G., Pérez-Camacho, L., Piñeiro, R., Ponge, J.-F., Rashid, M.I., Rebollo, S., Rodeiro-Iglesias, J., Rodríguez, M.Á., Roth, A.M., Rousseau, G.X., Rozen, A., Sayad, E., van Schaik, L., Scharenbroch, B.C., Schirrmann, M., Schmidt, O., Schröder, B., Seeber, J., Shashkov, M.P., Singh, J., Smith, S.M., Steinwandter, M., Talavera, J.A., Trigo, D., Tsukamoto, J., de Valença, A.W., Vanek, S.J., Virto, I., Wackett, A.A., Warren, M.W., Wehr, N.H., Whalen, J.K., Wironen, M.B., Wolters, V., Zenkova, I.V., Zhang, W., Cameron, E.K., Eisenhauer, N., 2019. Global distribution of earthworm diversity. Science 366, 480–485. https://doi.org/10.1126/science.aax4851

Postma-Blaauw, M.B., Bloem, J., Faber, J.H., van Groenigen, J.W., de Goede, R.G.M., Brussaard, L., 2006. Earthworm species composition affects the soil bacterial community and net nitrogen mineralization. Pedobiologia 50, 243–256. https://doi.org/10.1016/j.pedobi.2006.02.001

R Core Team, 2023. A language and environment for statistical computing.

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. 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., Suhadolc, M., Kos, I., Krogh, P.-H., Faber, J.H., Mulder, C., Bogte, J.J., Wijnen, H.J. van, Schouten, A.J., Zwart, D. de, 2016. Mapping earthworm communities in Europe. Appl. Soil Ecol., Soil biodiversity and ecosystem functions across Europe: A transect covering variations in bio-geographical zones, land use and soil properties 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. Biodivers. Conserv. 32, 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. Curr. Sci. 113, 1064–1071.

Singh, J., Schädler, M., Demetrio, W., Brown, G.G., Eisenhauer, N., 2019. Climate change effects on earthworms - a review. Soil Org. 91, 114–138. https://doi.org/10.25674/so91iss3pp114

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 Tillage Res. 100, 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 Ecol. 13, 46. https://doi.org/10.1186/1472-6785-13-46

Thakur, M.P., Reich, P.B., Hobbie, S.E., Stefanski, A., Rich, R., Rice, K.E., Eddy, W.C., Eisenhauer, N., 2018. Reduced feeding activity of soil detritivores under warmer and drier conditions. Nat. Clim. Change 8, 75–78. https://doi.org/10.1038/s41558-017-0032-6

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

van der Wal, A., Geerts, R.H.E.M., Korevaar, H., Schouten, A.J., op Akkerhuis, G.A.J.M.J., Rutgers, M., Mulder, C., 2009. Dissimilar response of plant and soil biota communities to long-term nutrient addition in grasslands. Biol. Fertil. Soils 45, 663–667. https://doi.org/10.1007/s00374-009-0371-1

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. Sci. Rep. 4, 6365. 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. J. Appl. Ecol. 42, 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.

Yiu, T., 2021. Understanding Random Forest [WWW Document]. Medium. URL https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed 1.18.24).

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. Conserv. Biol. 38, e14187. https://doi.org/10.1111/cobi.14187

Zhu, X., Zhu, B., 2015. Diversity and abundance of soil fauna as influenced by long-term fertilization in cropland of purple soil, China. Soil Tillage Res., Soil Structure and its Functions in Ecosystems: Phase matter & Scale matter 146, 39–46. https://doi.org/10.1016/j.still.2014.07.004

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., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography 43, 1261–1277. https://doi.org/10.1111/ecog.04960

Appendix 1 (in progress)

Subsequently, we selected the six main land cover types with sufficient observations (more than 110 observations) in our database (see appendix 1), corresponding to the level 3 nomenclature of Corine Land Cover (CLC, https://land.copernicus.eu/pan-european/corine-land-cover).

Appendix 2 (in progress)

Then we selected only observations conducted using the hand sorting protocol (ISO 23611-1:2018) and/or the application of a chemical expellant (formaldehyde or allyl isothiocyanate; ISO 23611-1:2006) (see appendix 2).

Appendix 3 (in progress)

Our modeling strategy followed the "ODMAP" protocol recommended by Zurell et al. (2020), and all steps are detailed in Fig. 2 (see appendix 3).

Master de modélisation en écologie de l'Université de Rennes (2023 – 2024)

Abdourahmane DIALLO

Abstract

Earthworms, as crucial ecosystem engineers, contribute significantly to various soil functions and ecosystem services such as water regulation, nutrient dynamics, and biomass production. Recognized as sensitive indicators of soil health, their conservation is pivotal for maintaining biodiversity and ecosystem stability. However, the factors influencing their biogeography and community composition are still not well understood. This study addressed this knowledge gap by comparing different algorithms for predicting the spatial distribution of earthworms across France. The aim was to identify the best algorithm and to quantify and prioritize the effects of environmental factors on earthworms. By using the comprehensive LandWorm database (3 576 observations), we compared five modeling algorithms: Generalized Linear Models, Generalized Additive Models, Random Forest, Generalized Boosted Models, and Artificial Neural Networks. These models were employed to predict key community parameters (total abundance, total biomass, and total taxonomic richness) using ten environmental variables related to climate, soil and land use. Our results highlight that the Random Forest model performed "best" for predicting earthworm abundance, achieving the highest R^2 of 0.43 with an RMSE of 25 individuals/m². Similarly, for taxonomic richness, the Random Forest model and the Generalized Boosted Models yielded the best R^2 of 0.59 with an RMSE of 1.7 taxa. Our study highlighted that land use was the most important variable for earthworms, followed by spatial, climatic, and soil variables. This research not only advances our understanding of earthworm community distribution but also supports the design of targeted conservation strategies, ensuring the protection and sustainability of vital soil functions.

Keywords: earthworm community, predictions, SDMs, land use, environmental factors

Résumé

Les vers de terre sont les ingénieurs des écosystémiques et participant à des nombreux services écosystémiques tels que la régulation de l'eau, la dynamique des nutriments et la production de biomasse. Ils sont reconnus comme des bio-indicateurs sensibles de la santé des sols et leur conservation est essentielle pour maintenir la biodiversité et la stabilité des écosystèmes. Cependant, les facteurs influençant leur biogéographie et la composition de leurs communautés sont encore mal compris. Dans cette étude, nous comblons cette lacune en comparant différents algorithmes de prédiction de la distribution spatiale des vers de terre à travers la France. L'objectif était d'identifier le meilleur algorithme et de quantifier et hiérarchiser les effets des facteurs environnementaux sur les vers de terre. En utilisant la base de données LandWorm (3 576 observations), nous avons comparé cinq algorithmes de modélisation: modèles linéaires généralisés, modèles additifs généralisés, forêts aléatoires, modèles boostés généralisés et réseaux de neurones artificiels. Ces modèles ont été utilisés pour prédire l'abondance totale, la biomasse totale et la richesse taxonomique des vers de terre en utilisant dix variables environnementales liées au climat, au sol et à l'utilisation des terres. Nos résultats montre que le modèle de forêt aléatoire a obtenu les meilleures performances pour prédire l'abondance des vers de terre ($R^2 = 0.43$, RMSE = 25 individus/m²). De même, pour la richesse taxonomique, le modèle de forêts aléatoires et le modèle boosté généralisé ont donné le meilleur ($R^2 = 0,59$, RMSE = 1,7). De plus, notre étude montre que l'utilisation des terres était la variable la plus importante pour les vers de terre, suivie par les variables spatiales, climatiques et pédologiques. Cette étude non seulement améliore notre compréhension de la distribution des vers de terre, mais elle soutient également la conception de stratégies de conservation ciblées, garantissant la protection et la durabilité des fonctions vitales du sol.

Mots-clés: communauté de vers de terre, prédictions, SDMs, utilisation des terres, facteurs environnementaux