\*\*Slide 1:\*\*

Hello everyone, my name is Abdou and I will present you the my internship progress. The topic of my internship is: "Modeling and predicting earthworm diversity and distribution in France: A comparative approach using multiple algorithms."

During my internship, I am supervised by Mr. Daniel and Kevin from the ECOBIO research laboratory in Rennes, and Walid, Agroecology research laboratory in Dijon.

\*\*Slide 2:\*\*

To begin with, I'd like to give a brief introduction to the prediction algorithms. 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). SDMs can be used to model the correlations between species and their environment as well as between communities and their environments. There are several SDMs, each with its own advantages and disadvantages.

\*\*Slide 3:\*\*

The question, therefore, is which SDM algorithms is the most appropriate to study and predict the earthworm community. For this question, there are several possible approaches, but the most pertinent seems to be the comparative approach as proposed by Salako et al. (2023). Indeed, this comparative approach allows us to identify the best model to predict each earthworm variables and also to reduce the uncertainty related to the predictions.

\*\*Slide 4:\*\*

For our study, we selected 5 prediction algorithms that we can categorized into two groups. For the first groups, we have the traditional regression algorithms, and we chose GLM and GAM. For the next groups, we have machine learning algorithms, and we chose three: RF, GBM, and ANN. We chose these 5 algorithms because they have recently been used in ecology, and some of them have already been independently used to predict earthworms.

\*\*Slide 5:\*\*

In ours study, we have 3 main objectives. The first objective was to compare the 5 selected algorithms to determine which one is the best for predicted the earthworms community. Our second objective was to quantify and rank the effects of environmental factors on earthworms, namely abundance, biomass, and richness. Our last objective was to predict the diversity and distribution of earthworms across metropolitan France.

\*\*Slide 6:\*\*

Now, I will present the materials and methods we used to achieve our objectives. For the modeling strategy, we followed the ODMAP protocol as recommended by Zurell et al. (2020). This figure represents the 8 steps of our modeling strategy, and I will now detail each one.

\*\*Slide 7:\*\*

The first step was collecting earthworm data. We used the LandWorm database, which contained approximately 8,000 observations of earthworms distributed across different land uses. These data varied from 1821 to 2023.

\*\*Slide 8:\*\*

Subsequently, we applied some filters to this database by selecting only sampling protocols of the Hand-sorting type and/or application of a chemical expellant.

\*\*Slide 9:\*\*

So, our database was reduced from 8,000 to about 6,000 observations, ranging from 1990 to 2023.

\*\*Slide 10:\*\*

The second step of our modeling strategy was the collection of environmental data. For the environmental data, we used some external databases to complete our LandWorm database. For the land use and spatial variables, we kept those that were available on LandWorm. We selected 6 land use types corresponding to the level 3 of the nomenclature of CORINE LAND COVER, with some adaptations. For the for forests, we grouped Broad-leaved forest, Coniferous forest, and Mixed forest into one category because we did not have enough data for each type of forest.

For the climatic variables, we extracted the 19 standard variables, such as precipitation and temperature, from the CHELSA database.

For the soil variables, we used this two databases to obtain the variables like the texture, the pH or the Carbon content

\*\*Slide 11:\*\*

To standardize the variables and match them to the same resolution, we referenced all the maps to the same reference, cropped them to keep only the boundaries of France, and resampled or disaggregated all the variables to a resolution of 30 arc seconds, which corresponds to about 800 meters in France. These steps were performed using Python with the GDAL module.

\*\*Slide 12:\*\*

Do you have any questions or suggestion ?

\*\*Slide 13:\*\*

Once the earthworm data and environmental data were merged, we cleaned them by removing NAs and outliers using the Grubbs test. For land use, which was a categorical variable, we transformed the 6 levels into binary variables. It is important to note that some data were temporally repeated. We decided to keep these plots because when we tested the models with and without the repeated data, the models performed better with the repeated data.

\*\*Slide 14:\*\*

Here is the Cleveland plot of the three response variables for earthworms, namely total abundance, total biomass, and total richness.

\*\*Slide 15:\*\*

For the numeric explanatory variables, we standardized them to facilitate model fitting. For abundance and biomass, we applied a square root transformation. We tested several transformations and even without any transformation, the square root transformation was the best.

\*\*Slide 16:\*\*

For example, for biomass, here is the distribution and the QQ plot before and after transformation. We can see that after transformation, the distribution is closer to a Gaussian distribution.

\*\*Slide 17:\*\*

Do you have any questions

\*\*Slide 18:\*\*

To reduce the number of variables and avoid model overfitting, we adjusted random forest models and identified the most important variables for earthworms using a permutation procedure. Thus, the variables identified for prediction were those listed in this table.

\*\*Slide 19:\*\*

The fifth step of the modeling strategy was data partitioning. There are different methods of partitioning to prepare data evaluation, such as K cross-validation. In our case, we chose the simple cross-validation method, which involves dividing the dataset into two parts with 80% of the data for training and 20% for validation. We then tested the homogeneity of the partition with the Kolmogorov test, which indicated that the data were similar between the training and test sets. This method was chosen for its simplicity, ease of understanding and implementation, and fast compilation.

\*\*Slide 20:\*\*

For example, in these figures, we can see the comparison between the training and test sets for the three earthworm variables.

\*\*Slide 21:\*\*

Do you have any questions?

\*\*Slide 22:\*\*

The sixth step of the modeling strategy was model fitting and calibration. We first adjusted the GLM and GAM according to this formulation.

\*\*Slide 23:\*\*

Next, we adjusted and calibrated the Random Forests (RF). We first defined a default model. Then we performed hyperparameter tuning for the random forest models using grid search with all possible combinations of these three parameters. We selected the model that gave the highest R² and the lowest RMSE. The final model formulation was this:

\*\*Slide 24:\*\*

For the GBM, we used the same approach as for the RF, and the final model formulation was this:

\*\*Slide 25:\*\*

For the ANN, we also developed a default model with a sequential architecture as shown in this figure.

\*\*Slide 26:\*\*

We also performed tuning to improve the default model, and the final model was this one shown in the image.

\*\*Slide 27:\*\*

Do you have any questions?

\*\*Slide 28:\*\*

To evaluate the models and perform selection, we used several metrics. The training data were used to fit the models, and the validation data were used to evaluate the predictive performance of the models. Since all our response variables were quantitative, we chose to use the coefficient of determination (R²) between the observations (validation data) and the predictions, as well as the mean absolute error (MAE) and the root mean square error (RMSE). The formulas used were these. The goal was to maximize R² and minimize MAE and RMSE.

\*\*Slide 29:\*\*

Do you have any questions?

\*\*Slide 30:\*\*

The final step of our modeling strategy was prediction and mapping. First, we performed sampling at a resolution of about 800 meters across the entire territory of France. For each sampling point, we extracted, when it possible, the values of the environmental final variables included in the models. We then make the prediction by using the best model of each earthworms variables. For areas where we could not predict the earthworm community, we performed interpolation. There are several types of interpolation. We could simply shade areas with NA values in gray, indicating they were not predicted, or use the proximity interpolation method, which involves assigning the value of the closest observation. We could also use trend surface or kriging methods, which are neighborhood-based interpolation methods with regression using spatial coordinates.

\*\*Slide 31:\*\*

In our case, we chose to use the IDW method because it was the most appropriate for earthworms. It involves interpolating the unknown point using the closest observations but weighted by their distance. Here is the IDW formula, and concretely, this means, for example, on this figure, we want to know the abundance at the center point symbolized by a star here, we will use the 10 closest observations, and each of these observations will have a weight according to its distance from the unknown point. The final value will be a linear combination of all these points with their weights.

\*\*Slide 32:\*\*

Do you have any questions?

\*\*Slide 33:\*\*

I will now present the preliminary results. First, here is the comparison table of the predictive performance of the 5 models for each of the three earthworm variables. We observe that the RF and GBM have the highest R² and the lowest RMSE for abundance, biomass, and richness. This superiority of the RF and GBM is explained by the fact that they use an ensemble of methods and decisions to make predictions. They capture very well the nonlinear relationships between environmental variables and earthworm variables. They handle the effects of extreme values well and also integrate interactions between environmental variables. However, RF and GBM require a lot of data and are not easily interpretable.

\*\*Slide 34:\*\*

Here are the figures comparing observed values on the X-axis and predicted values on the Y-axis. The highest R² for abundance is 0.43, 0.35 for biomass, and 0.59 for total richness. However, we can see that the models do not predict high values very well, which is also one of the problems with machine learning models.

\*\*Slide 35:\*\*

These three figures show the importance of environmental factors on earthworms. On the Y-axis, we have the explanatory variables, and on the X-axis, we have the RMSE. We can see that land use is the most influential variable on earthworms (for abundance, biomass, and richness). This result is explained by the fact that earthworms have certain habitat preferences. We observed that grasslands had the highest earthworm communities, while cultivated lands and vineyards had the lowest. After land use, spatial variables, particularly longitude, were the most important. Climatic variables were the least influential on earthworms. However, precipitation was very important for biomass.

\*\*Slide 36:\*\*

Do you have any questions or suggestions?

\*\*Slide 37:\*\*

Here is the prediction map of the total abundance distribution of earthworms in France. The abundance ranges from 0 to 530 individuals per square meter. We observe that earthworms were more numerous in the west and center of the country.

\*\*Slide 38:\*\*

Here is the prediction map of the total biomass distribution of earthworms in France. Biomass ranges from 0 to 210 grams per square meter. We observe that biomass was very low in the southeast of the country, which corresponded to cultivated areas with vineyards.

\*\*Slide 39:\*\*

Here is the prediction map of the total taxonomic richness distribution of earthworms in France. Richness ranges from 0 to 10 species per plot. This is consistent with recent study results. We observe that earthworms were more numerous in the west and center of the country.

\*\*Slide 40:\*\*

Do you have any questions or suggestions?

\*\*Slide 41:\*\*

Thank you for your attention.