**Plan** (25 pages numérotées +/- 10 %)

**Remerciements**

**Abréviations**

**Table des matières**

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

# Introduction

La biodiversité des écosystèmes terrestres est majoritairement constituée par la faune vivant dans le sol (FAO, 2020). En effet, la faune du sol fournit une large gamme de services écosystémiques (Bardgett & Van der Putten, 2014; FAO, 2020) dont la plupart résulte du processus conduit par les groupes fonctionnels de différentes communautés biologiques tels que les vers de terre (Gardi & Jeffery, 2009; Turbé et al., 2010). Les vers de terre sont qualifiés comme étant les « ingénieurs des écosystèmes » car ils participent au développement de la structure du sol (Schon et al., 2017; Sharma et al., 2017), à l'infiltration de l'eau à travers les galeries verticales et horizontales creusés dans le sol (Capowiez et al., 2014; Cunha et al., 2016) et à la minéralisation des nutriments grâce à la dégradation de la matière organique (Van Groenigen et al., 2019). Ces dernières années, plusieurs études ont démontré les rôles importants assurés par les vers de terre, (Blouin et al., 2013; Babu Ojha & Devkota, 2014; Bertrand et al., 2015). Ces travaux ont été justifiés par la prise de conscience croissante du rôle de la diversité des vers de terre, mais aussi par le fait qu'on pense que les changements dans leur composition entraîneront une altération des services écosystémiques fournis par les vers de terre (Cardinale et al., 2012; Hooper et al., 2012).

Bien que les vers de terre jouent un rôle important dans les services écosystémiques (Turbé et al., 2010; Blouin et al., 2013; Bertrand et al., 2015), ils sont sensibles à de nombreux facteurs anthropiques et environnementaux (Pelosi et al., 2014; Bai et al., 2018; Gabriac et al., 2022). Les activités humaines notamment la surexploitation des ressources naturelles à cause du développement de l'agriculture intensive et l'industrialisation est la source principale de l’altération des vers de terre (Pelosi et al., 2009). Les effets des facteurs environnementaux sur la communauté des vers de terre ont déjà été examinés à l’échelles locale (Marchán et al., 2015; Gabriac et al., 2022) et régionale (Marchán et al., 2016, 2021; Marchán & Domínguez, 2022).

Cependant, peu d’études se sont concentrées sur les effets des facteurs environnementaux et anthropiques qui influencent la biodiversité et la répartition des vers de terre à des échelles plus larges, c’est-à-dire suprarégionale ou nationale (De Wandeler et al., 2016). Les raisons qui peuvent expliquer ce manque de connaissances sont la faible disponibilité des données à l’échelle des pays, des incohérences taxonomiques et des difficultés pour fusionner les bases de données existantes (Rutgers et al. 2016). Par conséquent, cela pose des défis pour la conservation des vers de terre et donc les fonctions écosystémiques associées. L'un des rares travaux réalisés à une échelle très grande (continentale) a été effectué par Rutgers et al. (2016), qui ont décrit l’abondance totale et la richesse en espèces sur 3838 sites échantillonnés dans 8 pays d’Europe. Ils ont observé un gradient latitudinal inverse de l'abondance des vers de terre. De plus, leur étude montre aussi que l'occupation du sol, la texture du sol, la matière organique et le pH du sol sont les facteurs qui affectent plus les vers de terre. Ces résultats sont en accord avec ceux trouvés par (Jänsch et al., 2013). Récemment, une autre étude réalisée par Phillips et al. (2019) sur 9212 sites répartis dans 57 pays a démontré que les facteurs climatiques et le type d’habitat sont les filtres environnementaux les plus importants dans l’altération des paramètres lombriciens que les propriétés du sol. Les études de Rutgers et al. (2016) et de Phillips et al. (2019) sont les pionniers dans le domaine de la prédiction des vers de terre et ils ont permis d’obtenir des cartes de l’abondance et de la richesse des vers de terre à l'échelle continentale (Europe) et mondiale. Toutefois, leurs études sont limitées par l’utilisation d’un seul type d’algorithme de modélisation prédictive : des modèles linéaires généralisés pour Rutgers et al. (2016) et des modèles linéaires généralisés à effets mixtes pour Phillips et al. (2019). Or, il serait pertinent de comparer plusieurs algorithmes afin de trouver le meilleur modèle pour chaque paramètre des vers de terre (abondance, biomasse et richesse). De plus, les valeurs prédites par les modèles de Rutgers et al. (2016) et de Phillips et al. (2019) sont différents des valeurs observées sur le terrain en France par (Cluzeau et al., 2012). Il existe donc un réel besoin de connaissances approfondies de la distribution des vers de terre en France, et cela peut se faire par le développement des modèles prédictifs performants permettant de fournir des données de référence précises et d’identifier les besoins de conservation spécifiques des vers de terre en France.

Pour déterminer la distribution des communautés des vers de terre, plusieurs outils ont été développés ces dernières années. On distingue les modèles de distribution des espèces « species distribution modèles (SDMs) » qui sont des modèles basés sur les niches écologiques « Niche-based-modèles » et permet de modéliser les corrélations entre les espèces ou les communautés et leur environnement (Elith & Leathwick, 2009; Guisan et al., 2017). Néanmoins, ce type de modèle exige d’une part, des données sur les paramètres qui caractérise les communautés et d’autre part des données environnementales spatialisées. Divers SDMs existent, chacun présentant des avantages et des inconvénients (Li and Wang, 2013 ; Valavi et al., 2021). On peut citer les modèles linéaires généralisés (GLM), appropriés lorsque les relations entre les variables sont linéaires, mais susceptibles d'être influencés par des valeurs aberrantes. Les modèles additifs généralisés (GAM), capables de capturer les relations non-linaires des variables, mais présentent une complexité supérieure par rapport aux GLM (S. N. Wood, 2017). Les forêts aléatoires (RF), reconnues pour leur performance, sont en revanche moins interprétables (Li & Wang, 2013; Mi et al., 2017). Les modèles de régression boostée généralisée (GBM) démontrent une bonne précision, mais leur paramétrage peut être complexe (Friedman, 2001). Les réseaux de neurones artificiels (ANN), tout en ayant la capacité de modéliser des relations complexes, demandent des volumes importants de données pour éviter le surajustement (Li & Wang, 2013). Les choix méthodologiques exigent donc une évaluation approfondie en fonction des caractéristiques spécifiques des données et des objectifs de modélisations. Selon (Maes et al., 2012), les SDMs peuvent être des solutions pour l'aide à la décision pour les programmes de conservation de la biodiversité du sol.

Les SDMs ont déjà été mises en œuvre pour prédire la distribution des vers de terre selon divers facteurs environnementaux. En France, la plus récente étude a été réalisée sur 1366 sites par Fourcade and Vercauteren (2022), qui ont utilisé des arbres de régression boostés pour construire des prévisions spatiale de la diversité fonctionnelle avec 44 espèces des vers de terre. Cependant, l'étude de Fourcade and Vercauteren (2022) repose uniquement sur des données de présence/absence des espèces de vers de terre en France collectées dans les années 1960 par Bouché (1972). Or, ces données ne reflètent plus l'assemblage actuel des vers de terre en France. De plus, l'étude de la communauté des vers de terre doit inclure les variables décrivant la richesse en espèces, l'abondance et la biomasse totale, qui sont des paramètres vitaux pour les évaluations de la biodiversité. Utiliser seulement la diversité fonctionnelle n'est pas suffisant pour évaluer la biodiversité, car, par exemple, deux communautés peuvent être identiques en termes de richesse, mais différer en densité (Groves, 2022).

L'objectif généralde notre étude était de développer des modèles de prédictions des vers de terre grâce à différents algorithmes de modélisation, de produire des données de référence et d’identifier les besoins de conservation spécifiques des vers de terre en France. Plus précisément, nous avons cherché à répondre aux objectifs suivants : (i) quantifier et hiérarchiser l'influence des facteurs environnementaux (occupation du sol, propriétés du sol, localisation, et données climatiques) en fonction de leur contribution aux modèles prédictifs sur l'abondance (individu par m²), la biomasse (g par m²) et la richesse taxonomique totale des vers de terre (nombre de taxons dans la parcelle) en France métropolitaine (hors Corse) et (ii) prédire et cartographier ces mêmes paramètres lombriciens en fonction des facteurs environnementaux. Nos hypothèses reposent sur le fait que les variables climatiques (précipitation et température) et le type d'occupation du sol seraient les facteurs les plus influençant sur l'abondance et la biomasse des vers de terre, et que le type d'occupation du sol et la texture du sol influenceraient davantage la richesse en espèces. Pour répondre aux objectifs de l'étude, nous avons appliqué et comparé plusieurs algorithmes de modélisation prédictive afin d'identifier le meilleur modèle pour chacun des trois paramètres des vers de terre.

# Matériels & méthodes

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

La zone de notre étude couvre l’ensemble du territoire de France métropolitaine (hors Corse) (Fig. X). Nous avons utilisé 3576 observations de vers de terre de la base de données LandWorm (2023-2026 FRB-MTE-OFB), qui comprend les communautés de vers de terre à travers la France métropolitaine dans des conditions variées d'utilisation des sols. Les échantillonnages des vers de terre varient entre 1990 et 2023 et ont été réalisés principalement au printemps qui correspondent à la période d’activité maximale des vers de terre. Les protocoles utilisés était le test Bêche (70 % des données ; ISO 23611-1 :2018), le formol (25 % des données ; ISO 23611-1 :2006) et AITCTM (isothiocyanate d’allyle, 5 % des données ; ISO 23611-1 :2018). Dans chaque site, les vers de terre ont été collectés, triés à la main et stocké dans de l’éthanol. Au laboratoire, les vers de terre ont été comptés, pesés, affectés à un stade de développement (juvénile, subadulte et adulte) et identifiés au niveau de l'espèce/taxon. Pour nos analyses, nous avons utilisé uniquement l’abondance totale (ind/m²), la biomasse totale (g/m²) et richesse taxonomique totale (nombre de taxons trouvés dans le site). Par la suite, pour chacun de ces variables de réponse, nous avons enlevé tous les sites ayant de NA ou des valeurs aberrantes en utilisant le test de Grubbs avec la fonction *grubbs.test* du package *outliers* (Komsta, 2022). Cependant, 48 % des parcelles n’avait pas de biomasse totale. Nous avons aussi récupéré toutes les métadonnées associées au vers de terre comme les coordonnées GPS et l’occupation du sol qui correspondait à la classification des données sur la couverture terrestre du projet CORINE (https://land.copernicus.eu/pan-european/corine-land-cover).

Une image contenant texte, diagramme, carte

Description générée automatiquement

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

## Collecte des données environnementales

En plus des données sur l’occupation du sol et des coordonnées GPS obtenues avec le base de données LandWorm (2023-2026 FRB-MTE-OFB), nous avons utilisé des variables externes connues pour avoir des effets sur les vers de terre (Rutgers et al., 2016; Phillips et al., 2019; Salako et al., 2023). Pour les variables climatiques, nous avons utilisé 19 variables bioclimatiques standards du projet CHELSA (Karger et al., 2017). Pour les variables pédologiques, nous avons d’abord utilisé la base de données BDAT pour récupérer les teneurs en sables, argiles et limons (Roman Dobarco et al., 2022). Ces trois variables étaient disponibles à différentes profondeurs et nous avons choisi de télécharger les trois premières couches (0 à 5 cm, 5 à 15 cm et 15 à 30 cm) puis de faire la moyenne pour obtenir des données de 0 à 30 cm. Ce choix a été fait pour tenir compte de la variation de l’habitat des vers de terre. Par la suite, nous avons complété la texture du sol par les variables : la capacité d'échange de cations (CEC), le carbonate de calcium (CaCO3), le rapport C/N, l’azote (N), le phosphore (P), le potassium (K), et le pH du sol dans H2O obtenue grâce à la JRC (Ballabio et al., 2019). Nous avons initialement référencé toutes les variables présélectionnées au système de coordonnées géographiques mondial (WGS84), les avons recadrées et masquées pour correspondre aux limites géographiques de France métropolitaine (hors Corse). Nous les avons aussi rééchantillonnées ou désagrégées à une résolution commune de 30 secondes d'arc, ce qui correspond à environ 800 m en France. Ces étapes ont été réalisées sous Python avec le module GDAL (https://pypi.org/project/GDAL/).

## Stratégie de modélisation

### Protocole ODMAP

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

#### **Fig. X :** Stratégie de modélisation selon le protocole ODMAP : (1) collecte de données biologiques et des données environnementales, (2) Normalization des données, (3) sélection de variables, (4) organisation des données, (5) partitionnement des données en données d'entraînement et de test, (6) ajustement, (7) prédiction et (8) évaluation du modèle.

### Sélection, importance et effets des variables

Pour chacune des trois mesures de la communauté des vers de terre, nous avons ajusté des modèles en forêts aléatoires pour identifier l’importance de chaque variable explicative à l’aide de la fonction *importance* du package *randomForest* (Breiman, 2001). Nous avons choisi le modèle en forêt aléatoire car il peut traiter des données non linéaires tout en intégrant les interactions des variables qui peuvent améliore les performances du modèle (Breiman, 2001). Pour réduire le nombre de variable, nous enlevons à chaque fois la variable la moins importante et nous regardons la variation du R² du modèle. Nous répétons cela jusqu’à l’obtention d’un nombre de variables réduites et qui entraine un R² élevé (Fourcade & Vercauteren, 2022). Ainsi, la sélection de variables nous a permis d’identifier 10 variables environnementales (Tableau XX) influençant les vers de terre : l’occupation du sol, la longitude, la latitude, le carbonate de calcium (CaCO3), l’azote (N), le phosphore (P), la teneur en argile et en limon, l'isothermité (bio3) et les précipitations annuelles moyennes. Par ailleurs, nous avons utilisé le package *iml* pour améliorer l’interprétabilité des modèles notamment en explorant l’importance, les effets et les interactions des variables (Casalicchio et al., 2024).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Abbr** | **Variables names** | **Units** | **Sum** | **Percentiles** | | |
| **0.05** | **0.5** | **0.95** |
| **For** | Forest (all types) | Boolean | 116 |  |  |  |
| **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 |  |  |  |
| **CaCO3** | Calcium carbonates | g·kg−1 |  | 2.59 | 54.43 | 206.03 |
| **Long** | Longitude | WGS84 decimal system |  | -2.84 | 2.24 | 5.43 |
| **Lat** | Latitude | WGS84 decimal system |  | 43.48 | 47.99 | 49.87 |
| **Clay** | Clay particles | g·kg−1 |  | 13 | 23 | 41 |
| **Silt** | Silt particles | g·kg−1 |  | 26 | 48 | 68 |
| **bio12** | Annual precipitation amount | kg m−2 year−1 |  | 666 | 816 | 1040 |
| **bio3** | Isothermality | °C |  | 0.31 | 0.34 | 0.37 |
| **P** | Phosphorus | mg· kg·kg−1 |  | 21.11 | 42.05 | 57.77 |
| **N** | Nitrogen | g·kg−1 |  | 1.39 | 1.85 | 2.94 |

#### **Tableau X :** Abréviations et noms des variables utilisés dans la prédiction des communautés de vers de terre. Pour les prédicteurs catégoriels (de type booléen), la somme totale est fournie, résumant l'ensemble total de données avec 3576 observations. Pour les prédicteurs continus, trois percentiles (0,05, 0,5 et 0,95) de l'ensemble de données final sont fournis.

### Calibrages des modèles

Nous avons comparé cinq algorithmes SDM pour prédire les paramètres lombricien (abondance totale, biomasse totale et la richesse taxonomique totale) en utilisant 10 variables explicatives. Les cinq algorithmes sont : les modèles linéaires généralisés (GLM), les modèles additifs généralisés (GAM), les modèles de forêts aléatoires (RF), les modèles boostés généralisés (GBM) et les réseaux de neurones artificiels (ANN). Ces algorithmes ont été choisis sur la base de la classification des modèles en algorithmes de régression et d’apprentissage automatique mais aussi par leur large utilisation dans les études récentes (Li & Wang, 2013; Phillips et al., 2019; Valavi et al., 2021; Salako et al., 2023). Plusieurs études ont décrit en détail ces algorithmes (Li & Wang, 2013; Guisan et al., 2017; Mi et al., 2017; Yiu, 2021). Ainsi, nous avons ajusté ces 5 algorithmes avec les 10 variables explicatives pour prédire l’abondance totale (ind/m²), la biomasse totale (g/m²) et richesse taxonomique totale (nombre de taxons trouvés dans le site).

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

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

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

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

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

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

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

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

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

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

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

## Evaluation des modèles

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

## Prédictions et cartographie des communautés des vers de terre

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

# Résultats

## Performance des modèles

##### Tableau ou fig (nuage de point pred vs obs) des metriques R², RMSE,…

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

##### Production des cartes : 3

## Contribution relative des variables/catégories aux modèles de prédictions

##### Fig des contributions

## Effets des variables/catégories sur les vers de terre

##### Comment chaque variables finale affecte les vers de terre

# Discussion

# Conclusions et perspectives (page 25)

# Bibliographie

# Annexes

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

**Objectifs détaillés**

1. Quantifier et hiérarchiser l’influence des facteurs environnementaux (occupation et gestion du sol, propriétés de sol, localisation et climatique) sur l’abondance, biomasse et la richesse totales lombricienne sur le territoire métropolitain et sur certaines régions françaises (Bretagne, Dijon métro/Bourgogne ? Aquitaine ? /Poitou-Charentes ?)

*Pour y parvenir nous devrons identifier le meilleur algorithme (RF) pour quantifier et hiérarchiser l’influence des facteurs environnementaux sur les communautés lombriciennes.*

1. Prédire l’abondance, la biomasse et la richesse totales lombriciennes en fonction de différents facteurs environnementaux (occupation et gestion du sol, propriétés de sol, localisation et climatique) sur le territoire métropolitain et sur certaines régions françaises (Bretagne, Dijon métro/Bourgogne ? Aquitaine ? /Poitou-Charentes ?)

*Pour y parvenir nous devrons identifier le meilleur algorithme pour prédire chaque variable des communautés de vers de terre identifiée : abondance totale, biomasse totale et richesse totale*

1. Identifier les meilleures (résolutions des) représentations cartographiques de l’abondance totale, la biomasse totale et la richesse totale des communautés de vers de terre sur le territoire métropolitain et sur certaines régions françaises (Bretagne, Dijon métro/Bourgogne ? Aquitaine ? /Poitou-Charentes ?)

*Pour y parvenir nous utiliserons les différents algorithmes susmentionnés et nous devrons identifier/trouver les données géoréférencées correspondantes aux variables d’entrée des algorithmes (exemple BFC & NA). Nous chercherons à identifier les meilleures représentations cartographique (nbre de classe ?).*

*Cas sup : Modèle hors Dijon : cas 1*

**Données**

Les données viennent de la base de données LANDWORM (2023-2026 FRB-MTE-OFB).

* Variables de réponses = abondance, biomasse et richesse des vers de terre par parcelle.
* Variables explicatives envisager = occupations du sol, paramètres physico-chimiques du sol (pH, matière organique, texture, carbone, …) et facteurs climatique (précipitation et température, …).
* Autres variables = coordonnées GPS, année, site, …
* Échelle = la France

**Méthodologie / Planning prévisionnel**

**Janvier :** Rédaction du contexte, problématique et objectifs.

**Du 1ER au 15 février :** Analyse exploratoire des données de LANDWORM

-> taux de complétion des données ;

-> exploration des variables quantitative (dispersion, graphe, … ) ;

-> exploration des variables qualitatives (nombre, box plot, … ) ;

-> nettoyage ;

\* Utilisation de base de données externe pour compléter la BDD LANDWORM

-> sélection d’une série de variable explicatives « pertinentes » ;

\* ACP, analyse de variance, … ;

\* à dire d’experts ;

\* selon la bibliographie ;

**Du 16 au 28 février :** Modélisation avec le logiciel R, validation croisée et analyse de sensibilité / qualitative.

Validation du modèle avec des données indépendantes. Rédactions de matériels et méthodes.

Productions des graphiques et rédaction des résultats.

**Etape XX :** Cartographie avec QGIS et R ?.