



SoildiverAgro

Soil biodiversity enhancement in European agroecosystems to promote their stability and resilience by external inputs reduction and crop performance increase

D4.3 - REPORT ON CALIBRATION OF FTIR SPECTROSCOPY AS COMBINED WITH CHEMOMETRICS FOR HIGH-THROUGHPUT PREDICTION OF MICROBIAL AND NEMATODE COMMUNITY CHARACTERISTICS

Universidade de Vigo





This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

D4.3. REPORT ON CALIBRATION OF FTIR SPECTROSCOPY AS COMBINED WITH CHEMOMETRICS FOR HIGHTHROUGHPUT PREDICTION OF MICROBIAL AND NEMATODE COMMUNITY CHARACTERISTICS

Summary

Purpose: to develop prediction models for soil biological indicators (microbial genetic and functional groups, community composition and biodiversity indices and microbial abundance) using Fourier-Transform Infrared spectroscopy (FTIR) and multivariate chemometric models. This has been performed to give solution to the need for development of more time- and cost-efficient methodologies to assess soil biological properties, which are highly expensive and time consuming.

Intended audience: Mostly companies performing soil analyses, but also land planners and the scientific community.

Description of the main activities: Soil samples from different European pedoclimatic regions, cropping systems and management practices have been used to assess whether FTIR spectroscopy could be used to predict soil biological properties. A total dataset of 589 samples were used to obtain FTIR spectra at 7000–400 cm^{-1} . The continuous wavelet transform (CWT) was applied for successful data processing. For predictions, we used processed spectra, CWT coefficients and some environmental and soil physicochemical properties (CP) selected, to improve the accuracy of the models that had just used FTIR spectra. Three different predictions algorithms were implemented: partial least squares regression (PLRS), Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP). To train the algorithms, we adopted a stratified cross-validation scheme based on the pedoclimatic regions of the soil samples used in this study.

Key results: Models using FTIR spectra together with key environmental properties (temperature and precipitation) to predict the following soil biological indicators: Fungal chao1, Fungal guilds, Fungal PathotrophsSaprotrophs, Fungal Pathotrophs-Saprotrophs-Symbiotrophs, Fungal Saprotrophs-Symbiotrophs, Fungal Symbiotrophs, Fungal Pathotrophs, Fungal Pathotrophs-Symbiotrophs, Nematodes Globodera / Heterodera sp, *amoA* gene, Prokaryotes Observed number of OTUs, Prokaryotes Shannon index, Prokaryotes Fisher index, Firmicutes abundance, Actinomycota abundance, Gram positive bacteria abundance, Gram negative bacteria abundance, Arbuscular mycorrhiza fungi abundance, Total bacteria abundance, Zygomycota abundance, Total fungi abundance, and Total microbial biomass (total PLFAs).

Research and practice implications: There is a great demand for rapid and predictive soil data to be used in soil health assessment and monitoring and precision agriculture. Consequently, models based on FTIR spectroscopy improved with key environmental and soil physicochemical properties can be considered as an alternative to complement (or even replace) conventional analytical methods for soil biological indicators owing to their high price and lengthy processing time.

Policy implications: Soil biological indicators for soil health monitoring could be further encouraged to be employed as indirect estimations to reduce costs.

Conclusion: This deliverable confirms the potential of the FTIR spectra to predict soil biological properties, including outputs from DNA metabarcoding such as prokaryotic, fungal, and nematodes diversity indices and functional guilds, nematodes species abundances visually identified, functional genes by qPCR analysis, and microbial abundance by phospholipid fatty acid (PLFA) analysis. However, these models should be further improved with the incorporation of new samples from other pedoclimatic regions, cropping systems or even land uses to generalise their use as a routine basis.

Deliverable Number

D4.3

Work Package

WP4. Best Tools for soil biodiversity evaluation

Lead Beneficiary

Deliverable Author(s)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

UPCT

Fernando Terroso, Mathieu Kessler, Eva Lloret, Raúl Zornoza (UPCT)

Versions (updates)	Date
--------------------	------

V1

09.04.2025

Deliverable Quality Check	Date
---------------------------	------

Kristian Koefoed Brandt
David Fernández Calviño

16.04.2025
29.04.2025

Planned Delivery Date	Final Delivery Date
-----------------------	---------------------

31.05.2025

29.04.2025

Type of deliverable	R	Document, report (excluding periodic and final reports)	X
	DEC	Websites, patents filing, press & media actions, videos	
	E	Ethycs	

Dissemination Level	PU	Public	X
	CO	Confidential, only for members of the consortium	



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

Table of contents

1 INTRODUCTION.....	4
2 MATERIALS AND METHODS.....	5
3 RESULTS.....	14
4 CONCLUSIONS	23
5 CODE AVAILABILITY.....	24
6 REFERENCES.....	25

1 INTRODUCTION

The need for the development of more time- and cost-efficient methodologies for soil analysis is increasing, and mainly for biological properties, which are highly expensive and time consuming. There is a great demand for rapid and predictive soil data that can be used in environmental monitoring, soil health assessment and precision agriculture (Cohen et al., 2005; Viscarra Rossel et al., 2006). For this reason, Fourier-Transform Infrared spectroscopy (FTIR) is considered as an alternative to complement (or even replace) conventional analytical methods.

Over the past few decades, FTIR spectroscopy has rapidly developed to become a fast and robust analytical method for many agricultural, pharmaceutical and food products (Blanco and Villarroya, 2002; Zhang et al., 2022). In particular for soils, this technique permits the evaluation of different properties related to moisture and organic content matter, including carbon and nitrogen content or cation exchange capacity (Chodak et al., 2004). The conventional analytical methods are lengthy, expensive, destructive -consuming the samples during the analytical process-, and often use many chemical reagents. The advantages of using FTIR reflectance spectroscopy include the simplicity of sample pre-treatment (sieving and grinding of soils), its lack of chemical reagents, its non-destructive nature, and the fact that it is rapid, inexpensive and accurate for analysis (Zornoza et al., 2008). There have been few attempts to predict variables related to the soil microbial community composition by infrared spectroscopy. The first study we know that measured relationships between microbial community data and soil reflectance was that expounded by Johnson et al. (2003). The authors did not develop quantitative prediction models, but observed that the grouping of soil samples based on their soil reflectance properties was similar to the grouping based on DNA fingerprinting. From that publication, some attempts have been carry out to calibrate robust models using infrared spectra for soil microbiological properties, although more research and efforts are needed (Ng et al., 2019; Yang et al., 2022; Zhang et al., 2022). In this line, the use of FTIR to estimate microbial genetic and functional diversity provided by next-generation sequencing data has not been properly developed yet.

FTIR spectroscopy is based on the use of calibrations, coupled with chemometrics techniques, which utilize absorbances at many wavelengths to predict particular properties of a sample (Omer et al., 2020). Normally, FTIR spectra are used to establish models in which the significant information contained in the spectra is concentrated into a few variables, optimized to produce the best correlation with the predicted property. Nevertheless, practically all published authors agree that to assure the reliability of this technique, it is necessary to include a great number of samples from zones with a wide range in the values of soil properties.

Hence, the **objective** of this study was to develop models for the prediction of soil microbial genetic and functional groups, community composition and biodiversity indices, and microbial abundance using FTIR spectroscopy and multivariate chemometric models.

2 MATERIALS AND METHODS

2.1 Soil sampling

Soil samples belonging to WP3 were collected from organically and conventionally arable fields across nine pedoclimatic regions of Europe, namely Atlantic Central, Atlantic North, Boreal, Continental, Lusitanian, Mediterranean North, Mediterranean South, Nemoral, and Pannonian. Soil samples were collected right after harvest of wheat (*Triticum spp.*) in the late summer of 2019, except for the Pannonian region, which was sampled in the late summer of 2020 due to logistical difficulties. At least, 10 organically certified and 10 conventionally farmed fields were sampled per region, resulting in a total of 188 soil samples. When possible, pairwise organically and conventionally farmed fields were included (i.e., farms located in close proximity to each other). From each field, a composite soil sample was collected composed of 60 cores (depth of 25 cm) taken by a core sampler across an area of approximately 1 ha following a zigzag pattern across the field.

Soil samples belonging to WP5 were collected from all case studies. WP5 includes 15 field case studies (Figure 2.1) where different cropping systems and management have been tested to check their effects on soil biodiversity enhancement, crop growth, development and health, and the delivery of ecosystem services. The cropping systems and/or management practices included are related with: the use of soil mycorrhiza and plant growth promoting microorganisms; management of soil organisms (e.g. fungivores); the application of suitable crop rotations, multiple cropping and intercropping; the development of pest alert systems; the use of nutrient catch crops; the use of trap crops for pest control; the use of by-products as soil ameliorants; and the application of adequate tillage systems. As general rule, the soil samplings were carried out according to a selected phenological stage of the crop:

- a) For wheat. Flag leaf stage: flag leaf fully unrolled, ligule just visible.
- b) For potato. Crop cover complete: about 90% of plants meet between rows, regardless of the weather and following the experimental design and applied agricultural practices in every case study.
- c) For grape. Beginning of stem elongation: no internodes ("rosette").
- d) For pea. Visibly extended internodes.
- e) For cabbage. Heads begin to form: the two youngest leaves do not unfold.
- f) For leek. Leaf bases begin to thicken or extend.
- g) For turnip. Leaf rosette has reached 30% of the expected diameter typical for the variety.
- h) For broccoli. Leaf rosette has reached 30% of the expected diameter typical for the variety.
- i) For bean/soybean. Nine or more side shoots visible.

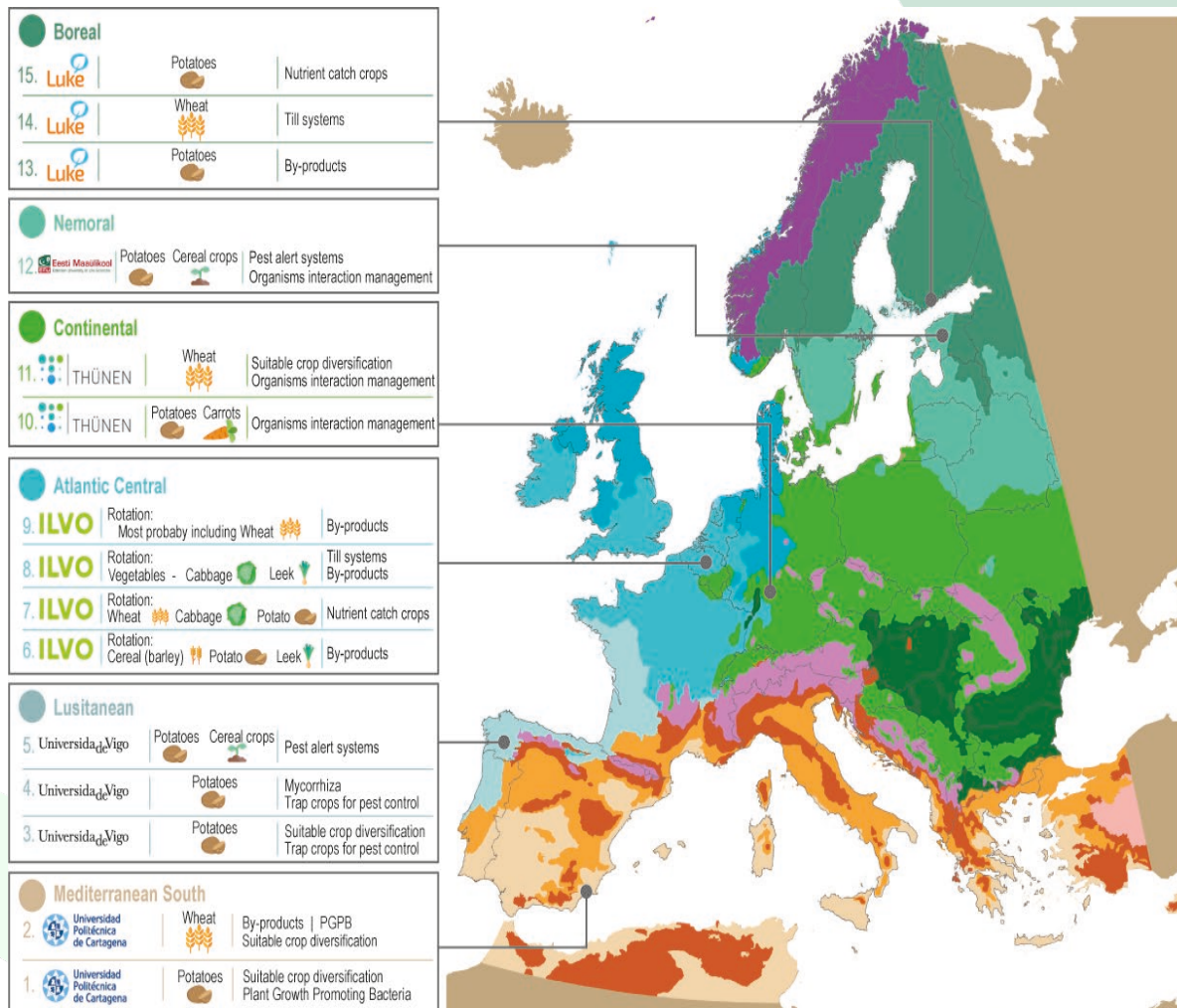


Figure 2.1. Location of the 15 case studies in the different European pedoclimatic areas together with the cropping systems and/or management practices tested on each case study.

In each case study, two soil samplings were carried out to get the FTIR spectra, once a year during the first two crop cycles. In each case study, there were between 3-4 treatments, with four field replications (four plots per treatment). One treatment was in all case studies the traditional management historically performed by the farmer in the farm. The other treatments were selected with the objective of increasing soil biodiversity. One composite soil sample was collected from each plot each year. This brings the total number of soil samples of 401 in WP5. Thus, the total number of soil samples used to obtain FTIR spectra to calibrate and validate chemometric models was 589.

Immediately after sampling, in all sampling campaigns from WP3 and WP5, the composite samples were thoroughly mixed and separated into two aliquots. One aliquot was kept at ambient temperature and, once in the laboratory, air-dried for 10 days, sieved at < 2 mm and stored at room temperature for physicochemical analyses. The other aliquot was stored cold in a cooler (~5 °C) until freezing (-80 °C) in the laboratory facilities for biological analyses. All datasets, together with proper metadata, can be found in the SoildiverAgro community of the



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

public repository of Zenodo (<https://zenodo.org/communities/soildiveragro-h2020/records?q=&l=list&p=1&s=10&sort=newest>).

2.2 Soil analyses and FTIR spectroscopy

Analyses for soil physicochemical properties and biological properties are explained in detail in deliverables “D3.2. Report and database of chemical and physical properties of different agricultural fields across the major European pedoclimatic regions”, “D3.3. Report on biodiversity status of soil microorganisms and soil fauna across the major European pedoclimatic regions”, and “D5.2. Report on biodiversity of soil micro- and macroorganisms in the different case studies in terms of cropping system, management practices and pedoclimatic region”.

A fraction of air-dried samples was ground for Fourier-Transform Infrared Spectroscopy (FTIR) analysis (100 mg). Attenuated total reflectance – FTIR spectra were obtained by means of a Nicolet 5700 spectrometer (Thermo Inc., Waltham, MA, USA) using a single-bounce diamond attenuated total reflection accessory (Smart Orbit, Thermo Inc.) at room temperature (25°C). Spectral data were collected in the range of 7000–400 cm^{-1} with OMNIC8.0 software (Thermo Fisher Scientific Inc.), to avoid the low signal to noise ratio at the start and end of the spectral range. Each spectrum was recorded by accumulating 32 scans with a resolution of 4 cm^{-1} . At least two spectra replicates for each soil sample were measured, and the average spectra was computed.

For each spectra, 6845 wavelengths were recorded, but contiguous wavelength measurements were highly collinear, which may cause numerical difficulties when implementing the regression and prediction algorithms. Thus, the spectra were therefore thinned, to a 10 nm resolution. As a result, each spectra consisted of 385 wavelengths. Following the methodology proposed by Yang et al. (2022), the reflectance (R) was transformed to apparent absorbance, $A = \log_{10}(1 / R)$, and the Savitzky Golay filter was applied, using a window of size 7, a polynomial order of 2, and first derivative method. The SVN transformation was applied to each spectra for a better comparison between samples. As an example, the resulting 188 spectra from soils collected in WP3 are represented in Figure 2.2.

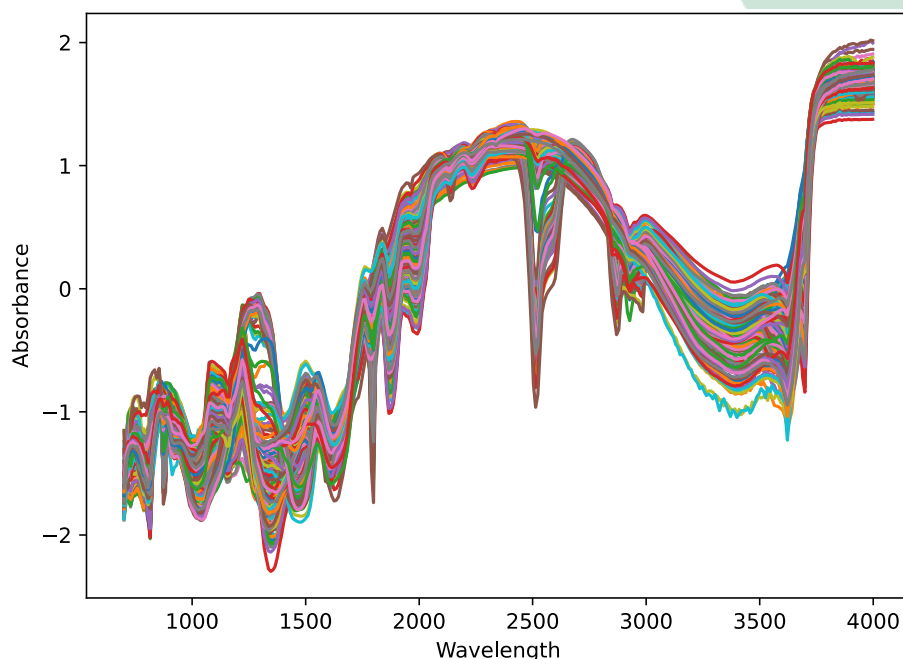


Figure 2.2. Absorbance processed spectra of all 188 samples collected in WP3.

2.3 The Continuous Wavelet Transform

The continuous wavelet transform (CWT) is a powerful mathematical tool used in signal processing and data analysis to analyze and represent data in both the time and frequency domains simultaneously. Unlike the discrete Fourier-transform, which offers a fixed resolution in time and frequency, the CWT adapts to the local characteristics of a signal. It does this by convolving the signal with a family of wavelet functions, which are essentially small, localized oscillatory patterns of varying frequencies and scales. The CWT provides a time-frequency representation of a signal, revealing details about its transient features and how they evolve over time. This makes it particularly valuable in applications such as image processing, audio analysis, and pattern recognition. For an introduction, see for example Chapter 1 of Antoine et al. (2004).

The CWT has been previously used with success for spectra processing (Chen et al., 2015). Once a wavelet basis is selected, the signal is decomposed at different scales and a matrix of CWT coefficients is obtained for each spectra. In the CWT matrix, each row corresponds to a scale and each column to a wavelength localization. In Figure 2.3, for a single spectra, the whole matrix of CWT coefficients is represented for integer scales ranging from 1 to 32, and the same wavelet basis.

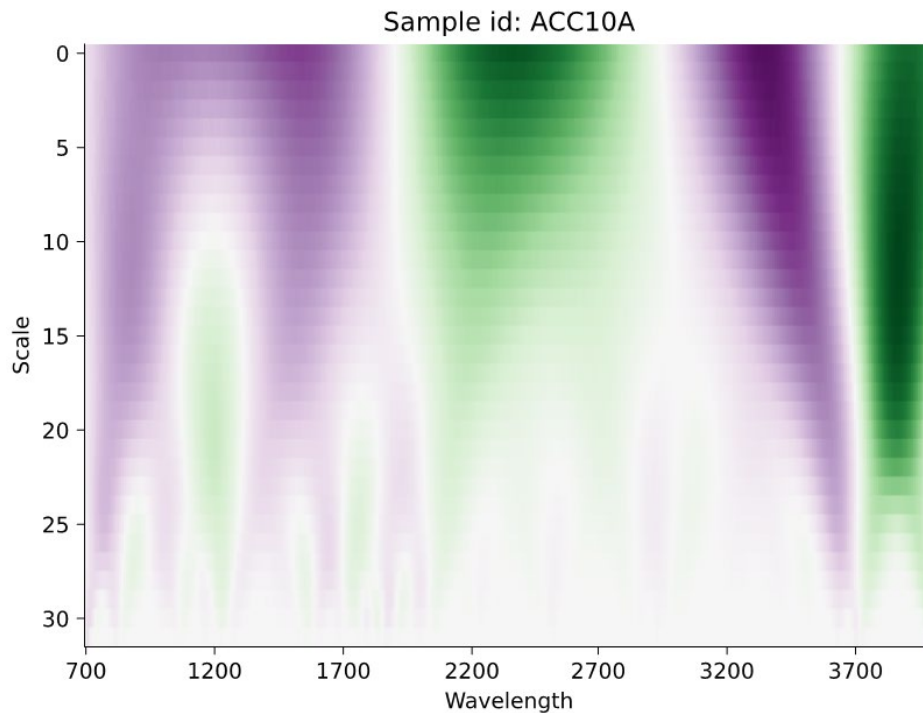


Figure 2.3. Example of CWT coefficients at scales 1 to 32, computed from on sample spectra.

2.4 Predictions

The soil biological indicators used as dependent variables to be predicted are the following:

- Fungal diversity and functionality: Fungal chao1, Fungal Shannon index, Fungal Inverse Simpson index, Fungal guilds, Fungal Pathotrophs-Saprotrophs, Fungal Pathotrophs-Saprotrophs-Symbiotrophs, Fungal Saprotrophs-Symbiotrophs, Fungal Symbiotrophs, Fungal Pathotrophs, Fungal Pathotrophs-Symbiotrophs.
- Nematodes diversity and abundance: Nematodes observed number of OTUs, Nematodes Shannon index, Nematodes absolute number of individuals, Nematodes *Pratylenchus sp*, Nematodes *Meloidogyne sp*, Nematodes *Globodera / Heterodera sp*, Nematodes *Trichodoridae*, Nematodes *Xiphinema / Longidorus sp*, Nematodes *Ditylenchus sp*, Nematodes *Paratylenchus sp*, Nematodes *Rotylenchus sp*.
- Functional genes: *amoA* gene, *nirK* gene, *cbbL* gene, GH7 gene.
- Prokaryotes diversity: Prokaryotes Observed number of OTUs, Prokaryotes Shannon index, Prokaryotes Simpson index, Prokaryotes InvSimpson index, Prokaryotes Fisher index
- Microbial biomass: Firmicutes abundance, Actinomycota abundance, Gram positive bacteria, Gram negative bacteria, Arbuscular mycorrhiza fungi, Total bacteria, Zygomycota, Total fungi, Total microbial biomass (total PLFAs).

We have used as inputs of the prediction algorithms either:

- i. The processed spectra (343 wavelengths for each spectra). These inputs correspond to the data represented in Figure 2.2.
- ii. The CWT coefficients at a given scale (343 coefficients for each spectra). These inputs correspond to the data represented in Figure 2.3.
- iii. Some environmental and soil physicochemical properties (CP) selected to improve the accuracy of the models just using the FTIR spectra. These properties were stratified into six different prioritized groups to be included in the models determined by facility to be obtained by analysts and their importance at contributing to explain variation in biological properties, such as soil organic matter or pH (Table 2.1).

Table 2.1. Prioritized blocks of selected environmental and soil physicochemical properties for models input to estimate soil biological properties together with CWT coefficients derived from FTIR spectra.

PRIORITY	PROPERTIES				
1	Mean annual temperature	Total annual precipitation	pH (measured in H ₂ O)	Soil organic matter	Total N
2	pH (measured in KCl)	Electrical conductivity	Total organic C	CaCO ₃	Particulate organic C
3	Cation exchange capacity	Exchangeable Ca	Exchangeable Mg	Exchangeable K	Exchangeable Na
4	Field moisture	NH ₄ ⁺	NO ₃ ⁻	Olsen P	
5	Available Fe	Available Cu	Available Zn	Available Mn	
6	Aggregate fraction > 2000 μm	Aggregate fraction between 250-2000 μm	Aggregate fraction between 53-250 μm	Aggregate fraction < 53 μm	

From the datasets described above, we can now define the prediction model as follows: given a soil sample s , its CWT matrix c_s and its physicochemical properties p_s , find a mapping function F :

$$F(c_s, p_s) \rightarrow b_s^i$$

where b_s^i is the i -th biological property of the sample.

In order to implement the aforementioned mapping function F , we have tested three prediction algorithms. The first one comes from the Machine Learning (ML) field while the other two comes from the Deep Learning field:

- Partial Least Squares Regression (PLSR) as a linear chemometric regression model which is a powerful tool for handling complex data structures often encountered in chemical analysis, such as spectroscopic data or chromatographic data. It combines a dimension reduction and a regression using the derived components to predict the response. PLSR can be applied on the processed spectra (i. input scenario) or the CWT coefficients at a

given scale (ii. input scenario). PLS has been extensively used for soil prediction properties from spectroscopic data and is used as reference method to compare the other prediction algorithms.

- A Convolutional Neural Network (CNN) is a specialized type of artificial neural network designed for processing and analyzing visual data, such as images and videos. CNNs are inspired by the human visual system and use a hierarchical structure of interconnected layers to automatically learn and extract features from raw pixel data. They employ convolutional layers to detect patterns like edges and textures and pooling layers to reduce spatial dimensions. CNNs have revolutionized computer vision tasks by excelling in tasks like image classification, object detection, and image segmentation, making them a fundamental technology in fields ranging from image recognition to medical imaging and autonomous vehicles. Recently, Padarian et al. (2019) and Ng et al. (2019) used successfully CNN for prediction of soil properties. Moreover, Yang et al. (2022) compared different machine learning prediction algorithms to estimate the soil fungal abundance using samples from a soil spectra database and concluded that CNN outperforms other widely used algorithms like XGBoost or SVM.
- Multilayer Perceptron (MLP) as a feedforward artificial neural network is a powerful tool for modeling complex relationships in various domains, such as pattern recognition and function approximation. It consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer, enabling it to learn nonlinear mappings from input to output. MLP combines feature extraction and regression through backpropagation-based weight optimization to minimize the prediction error. MLP can be applied using raw input features (i. input scenario) or transformed representations obtained through feature engineering (ii. input scenario). MLP has been extensively used for classification and regression tasks and serves as a benchmark method to compare the performance of other machine learning models.

In that sense, we have evaluated different versions of these algorithms, each one taking as input different variations of the datasets shown in Table 2.1. Table 2.2 summarizes such variations based on their input data. It is important to remark that one of the variations of the CNN, CNNcwt,pca(CP) used Principal Component Analysis (PCA) (Pearson, 1901) to reduce the dimensionality of the CP dataset from 26 to 5 dimensions. Moreover, Figure 2.4 shows the layer configuration composing the inner architecture of the CNN and MLP variations.

Table 2.2. Input and layers configuration of the prediction algorithms.

ID.	MODEL	INPUT DATA
CNNcwt	CNN	CWT
CNNcwtCP		CWT, CP
CNNcwt,pca(CP)		CWT, CP _{pca}
PLS	PLS	SP
MLP	MPL	CWT, CP

In this way, we trained and evaluated the five algorithms listed in Table 2.2 using an incremental approach with environmental and soil physicochemical properties (Table 2.1). First, we provided them with the first set of properties (mean annual temperature, total annual precipitation, pH measured in water, soil organic matter and total nitrogen). Next, we trained and evaluated them again by incorporating the properties included in groups 1 and 2. Then, we repeated the process using the properties from groups 1 to 3, and so on. Finally, we trained the models using all the properties included in the six groups. This approach allows us to assess the trade-offs between model accuracy and the difficulty of extracting the CP properties used as inputs for the predictors.

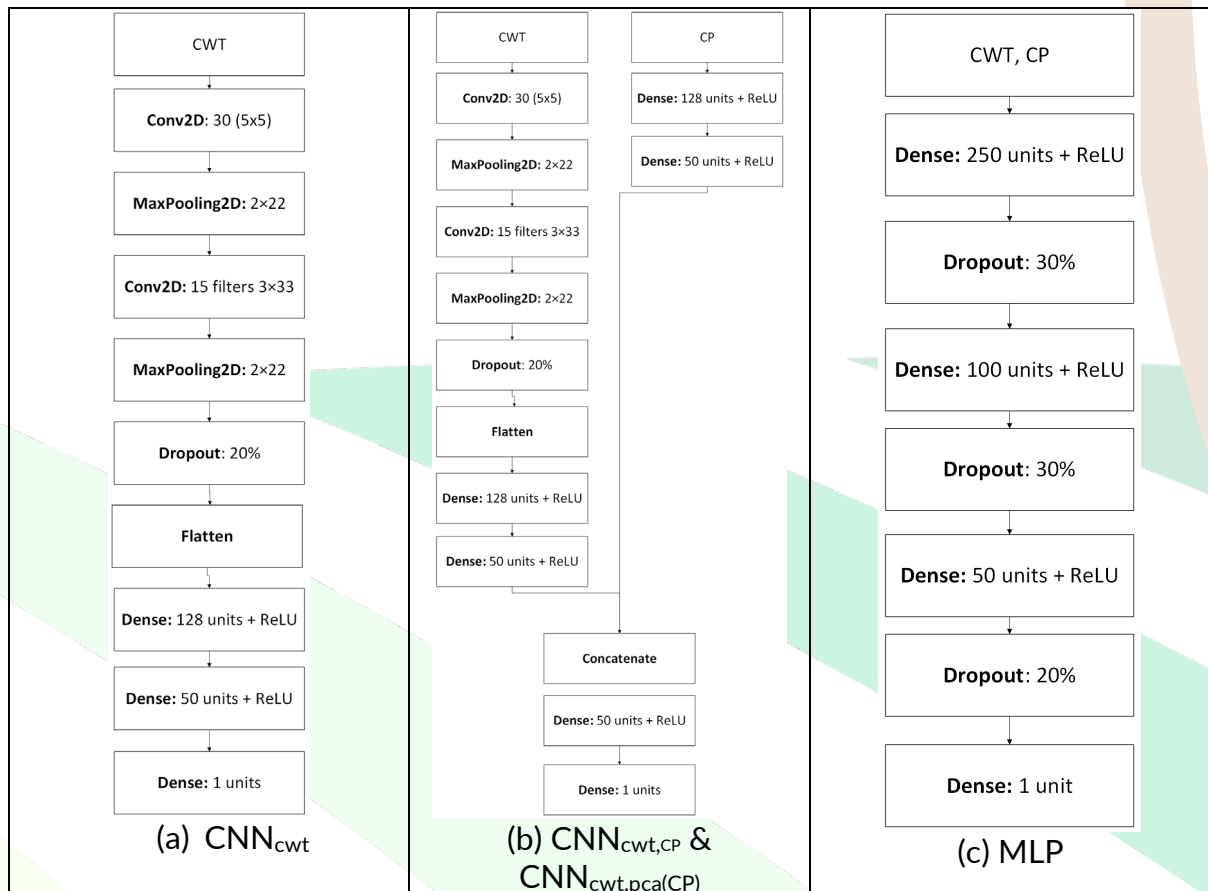


Figure 2.4. Layer architecture of the CNNs and MLP algorithms.

2.5 Training policy

To train the algorithms described in Section 2.4 (Table 2.2), we adopted a stratified cross-validation scheme based on the pedoclimatic areas of the soil samples used in this study. Moreover, Table 2.3 presents the most important hyperparameters used to train the three CNN variations and the MLP.

Table 2.3. Hyperparameters to train the DL models.

PARAMETER	DESCRIPTION	VALUE
Batch size	Size of batch used for training/forecasting	16
Epochs	Number of epochs used in training	1000 + Early Stopping(patience=30)
Optimizer	Function that optimizes the learning of the model	Adam
Loss function	Function used for evaluating the error each epoch	Mean Squared Error (MSE)
Learning rate	Percentage of weight adjustment each iteration	0.001 + ReduceLROnPlateau (patience=30, factor=0.4)

2.6 Evaluation metrics

In the evaluation of predictive models, three commonly used metrics are R^2 (coefficient of determination), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). Each of these metrics provides a different perspective on the accuracy and performance of a model. The formulas for these metrics are presented below:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{Equation 1}$$

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad \text{Equation 3}$$

where y_i is the actual value of the i -th data sample, \hat{y}_i is the value predicted by the model, \bar{y} is the mean of the actual values and n the total number of observations.

3 RESULTS

Regarding the most important results obtained from the model evaluation, Figure 3.1 shows the average values of the three evaluation metrics for the five models when all the SP parameters are used as input. As we can see, there is high variability in the results; however, it is important to note that the CNN combining CWT and the SP properties (CNNcwtCP) achieved the best R^2 and the second-best RMSE and MAE.

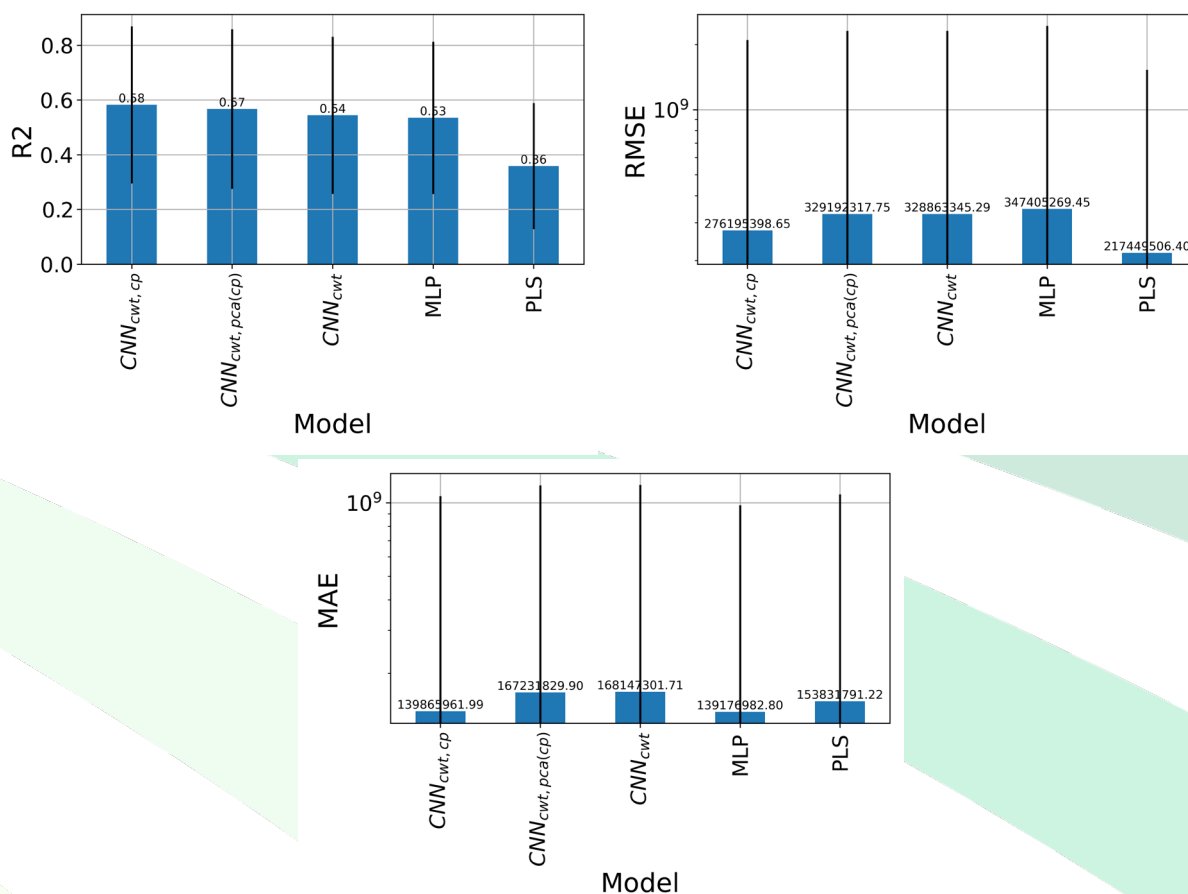


Figure 3.1. Mean value of the evaluation metrics (R^2 , RMSE and MAE) when all the PS parameters are used as input. The vertical lines depict the standard deviation of each model. Note that the y axis of RMSE and MAE are in logarithmic scale.

Due to the heterogeneity of the results, we further investigated the best combination of model and input data for each target biological property. In this regard, Tables 3.1, 3.2, and 3.3 present this combination, considering the R^2 , RMSE, and MAE metrics. Regarding Table 3.1, we obtained an average R^2 of 0.738 for all the soil biological properties when the best model and SP were used. In this context, it is noteworthy that we achieved an R^2 above 0.80 in 19 out of 36 properties.

Table 3.1. Best model and SP input for each biological property considering the R^2 metric. The best CP input column indicates the groups of CP properties taken as input according to the stratification in Table 2.1.

TARGET PROPERTY	R^2	BEST MODEL	BEST CP INPUT
Fungal chao1	0.916	CNN _{CWT,CP}	1-3
Fungal Shannon index	0.637	CNN _{CWT,CP}	1
Fungal Inverse Simpson index	0.669	CNN _{CWT,CP}	1-5
Fungal yields	0.876	CNN _{CWT}	-
Fungal Pathotrophs-Saprotrophs	0.894	CNN _{CWT,CP}	1
Fungal Pathotrophs-Saprotrophs-Symbiotrophs	0.860	CNN _{CWT,PCA(CP)}	1-2
Fungal Saprotrophs-Symbiotrophs	0.876	CNN _{CWT,PCA(CP)}	1-6
Fungal Symbiotrophs	0.808	CNN _{CWT,PCA(CP)}	1
Fungal Pathotrophs	0.845	CNN _{CWT,PCA(CP)}	1-6
Fungal Pathotrophs-Symbiotrophs	0.921	CNN _{CWT,PCA(CP)}	1-2
Nematodes observed number of OTUs	0.678	CNN _{CWT,CP}	1-5
Nematodes Shannon index	0.611	CNN _{CWT,CP}	1-4
Nematodes absolute number of individuals	0.579	MLP	1-5
Nematodes <i>Pratylenchus sp</i>	0.715	CNN _{CWT}	1
Nematodes <i>Meloidogyne sp</i>	0.121	CNN _{CWT,PCA(CP)}	1-5
Nematodes <i>Globodera / Heterodera sp</i>	0.911	CNN _{CWT,PCA(CP)}	1
Nematodes <i>Trichodoridae</i>	0.682	CNN _{CWT,CP}	1-2
Nematodes <i>Xiphinema / Longidorus sp</i>	0.657	CNN _{CWT,CP}	1-4
Nematodes <i>Ditylenchus sp</i>	0.604	CNN _{CWT,PCA(CP)}	1-6
Nematodes <i>Paratylenchus sp</i>	0.692	CNN _{CWT,CP}	1- 6
Nematodes <i>Rotylenchus sp</i>	0.579	CNN _{CWT,CP}	1
<i>amoA</i> gene	0.837	MLP	1-6
<i>nirK</i> gene	0.542	PLS	1
<i>cbbL</i> gene	0.878	MLP	1-3
GH7 gene	0.588	MLP	1-6
Prokayotes Observed number of OTUs	0.989	CNN _{CWT,CP}	1-3
Prokayotes Shannon index	0.882	CNN _{CWT}	1-3
Prokayotes Simpson index	0.061	CNN _{CWT,PCA(CP)}	1-6
Prokayotes InvSimpson index	0.638	MLP	1-6
Prokayotes Fisher index	0.991	CNN _{CWT,CP}	1-3
Firmicutes abundance	0.857	CNN _{CWT,CP}	1-6
Actinomycota abundance	0.893	CNN _{CWT,CP}	1-6
Gram positive bacteria abundance	0.899	CNN _{CWT,CP}	1-5
Gram negative bacteria abundance	0.811	CNN _{CWT,CP}	1-5
Arbuscular mycorrhiza fungi abundance	0.864	CNN _{CWT,CP}	1
Total bacteria abundance	0.882	CNN _{CWT,CP}	1-5
Zygomycota abundance	0.890	CNN _{CWT,CP}	1-3

TARGET PROPERTY	R ²	BEST MODEL	BEST CP INPUT
Total fungi abundance	0.885	CNN _{CWT,CP}	1-6
Total microbial biomass (total PLFAs)	0.906	CNN _{CWT,CP}	1-6

Concerning Tables 3.1 and 3.2, we compared the RMSE and MAE results, focusing on the frequency with which the same model is identified as the best for a given property. We found that in 18 out of 36 properties (50%), both the best model architecture and the best SP input were exactly the same across RMSE and MAE. This high level of agreement indicates that these models are not only accurate in terms of minimizing average errors (MAE), but also effective in reducing larger deviations (as captured by RMSE), suggesting stable and robust performance regardless of the specific error metric used. However, we identified several cases where discrepancies occur. For instance, for the property *Fungal yields*, the best model for RMSE is CNN_{CWT} with SP input 2, while for MAE, the best model remains CNN_{CWT}, but the best SP input changes to 4. This indicates that while the same model architecture is optimal across both metrics, the specific SP input that leads to best performance may differ depending on the metric considered.

Table 3.2. Best model and SP input for each biological property considering the RMSE metric. The best CP input column indicates the groups of CP properties taken as input according to the stratification in Table 2.1.

TARGET PROPERTY	RMSE	BEST MODEL	BEST CP INPUT
Fungal chao1	167.834	CNN _{CWT,CP}	1-3
Fungal Shannon index	0.316	CNN _{CWT,CP}	1
Fungal Inverse Simpson index	6.155	CNN _{CWT,CP}	1-5
Fungal yields	5.201	CNN _{CWT}	-
Fungal Pathotrophs-Saprotrophs	4.466	CNN _{CWT,CP}	1
Fungal Pathotrophs-Saprotrophs-Symbiotrophs	4.930	CNN _{CWT,PCA(CP)}	1-2
Fungal Saprotrophs-Symbiotrophs	2.877	CNN _{CWT,PCA(CP)}	1-6
Fungal Symbiotrophs	0.362	CNN _{CWT,PCA(CP)}	1
Fungal Pathotrophs	1.764	CNN _{CWT,PCA(CP)}	1-6
Fungal Pathotrophs-Symbiotrophs	0.916	CNN _{CWT,PCA(CP)}	1-2
Nematodes observed number of OTUs	13.867	CNN _{CWT,CP}	1-5
Nematodes Shannon index	0.343	CNN _{CWT,CP}	1-4
Nematodes absolute number of individuals	1.783.399	MLP	1-4
Nematodes <i>Pratylenchus sp</i>	271.093	CNN _{CWT}	-
Nematodes <i>Meloidogyne sp</i>	775.992	CNN _{CWT,PCA(CP)}	1-5
Nematodes <i>Globodera / Heterodera sp</i>	13.792	CNN _{CWT,PCA(CP)}	1
Nematodes <i>Trichodoridae</i>	4.531	CNN _{CWT,CP}	1-2
Nematodes <i>Xiphinema / Longidorus sp</i>	0.790	CNN _{CWT,CP}	1-4
Nematodes <i>Ditylenchus sp</i>	2.163	CNN _{CWT,PCA(CP)}	1-6
Nematodes <i>Paratylenchus sp</i>	65.456	CNN _{CWT,CP}	1-6

TARGET PROPERTY	RMSE	BEST MODEL	BEST CP INPUT
Nematodes <i>Rotylenchus sp</i>	32.388	CNN _{CWT,CP}	1
<i>amoA</i> gene	4,15x 10 ¹³	CNN _{CWT,PCA(CP)}	1-3
<i>nirK</i> gene	8,17 x 10 ¹⁵	PLS	1
<i>cbbL</i> gene	1,14x 10 ¹³	CNN _{CWT,CP}	1
GH7 gene	881.878.969	CNN _{CWT,PCA(CP)}	1-6
Prokaryotes Observed number of OTUs	391.447	CNN _{CWT,CP}	1-3
Prokaryotes Shannon index	0.197	CNN _{CWT}	-
Prokaryotes Simpson index	0.002	PLS	1
Prokaryotes InvSimpson index	74.976	MLP	1-6
Prokaryotes Fisher index	83.291	CNN _{CWT,CP}	1-3
Firmicutes abundance	1.791	CNN _{CWT,CP}	1-6
Actinomycota abundance	0.918	CNN _{CWT,CP}	1-6
Gram positive bacteria abundance	2.354	CNN _{CWT,CP}	1-5
Gram negative bacteria abundance	1.998	CNN _{CWT,CP}	1-6
Arbuscular mycorrhiza fungi abundance	0.392	CNN _{CWT,PCA(CP)}	1
Total bacteria abundance	5.105	CNN _{CWT,CP}	1-5
Zygomycota abundance	0.694	CNN _{CWT,CP}	1-3
Total fungi abundance	1.239	CNN _{CWT,CP}	1-6
Total microbial biomass (total PLFAs)	7.026	CNN _{CWT,CP}	1-6

In other cases, such as Fungal Pathotrophs, the best model for RMSE is CNN_{CWT,PCA(CP)} with SP input 6, whereas for MAE, the best model is CNN_{CWT} with SP input 5. This illustrates that different model architectures can perform differently depending on whether the focus is on minimizing average error (MAE) or penalizing large errors (RMSE).

Table 3.3. Best model and CP input for each biological property considering the MAE metric. The best CP input column indicates the groups of CP properties taken as input according to the stratification in Table 2.

TARGET PROPERTY	MAE	BEST MODEL	BEST CP INPUT
Fungal chao1	119.182	CNN _{CWT,CP}	1-3
Fungal Shannon index	0.222	CNN _{CWT,CP}	1
Fungal Inverse Simpson index	3.977	CNN _{CWT,CP}	1-5
Fungal yields	2.833	CNN _{CWT}	-
Fungal Pathotrophs-Saprotrophs	2.459	CNN _{CWT,CP}	1-6
Fungal Pathotrophs-Saprotrophs-Symbiotrophs	2.869	CNN _{CWT,PCA(CP)}	1-2
Fungal Saprotrophs-Symbiotrophs	1.480	CNN _{CWT,PCA(CP)}	1-6
Fungal Symbiotrophs	0.124	CNN _{CWT,PCA(CP)}	1
Fungal Pathotrophs	0.892	CNN _{CWT,CP}	1-5
Fungal Pathotrophs-Symbiotrophs	0.490	CNN _{CWT}	-
Nematodes observed number of OTUs	8.801	CNN _{CWT,CP}	1-5

TARGET PROPERTY	MAE	BEST MODEL	BEST CP INPUT
Nematodes Shannon index	0.254	CNN _{CWT,CP}	1-5
Nematodes absolute number of individuals	1.057.984	MLP	1-4
Nematodes <i>Pratylenchus sp</i>	140.706	CNN _{CWT,CP}	1-5
Nematodes <i>Meloidogyne sp</i>	96.255	MLP	1-6
Nematodes <i>Globodera / Heterodera sp</i>	5.549	CNN _{CWT,PCA(CP)}	1
Nematodes <i>Trichodoridae</i>	1.578	CNN _{CWT,CP}	1-2
Nematodes <i>Xiphinema / Longidorus sp</i>	0.297	CNN _{CWT}	-
Nematodes <i>Ditylenchus sp</i>	1.051	CNN _{CWT,CP}	1
Nematodes <i>Paratylenchus sp</i>	28.607	CNN _{CWT,CP}	1-6
Nematodes <i>Rotylenchus sp</i>	10.494	CNN _{CWT,CP}	1
amoA gene	1,98E+13	CNN _{CWT,PCA(CP)}	1-3
nirK gene	5,22E+15	MLP	1-3
cbbL gene	6,65E+12	CNN _{CWT,CP}	1
GH7 gene	396.500.414	CNN _{CWT,PCA(CP)}	1-6
Prokayotes Observed number of OTUs	237.889	CNN _{CWT,CP}	1-3
Prokayotes Shannon index	0.147	CNN _{CWT}	-
Prokayotes Simpson index	0.001	PLS	1
Prokayotes InvSimpson index	54.390	MLP	1-6
Prokayotes Fisher index	52.901	CNN _{CWT,CP}	1-3
Firmicutes abundance	1.218	CNN _{CWT,CP}	1-5
Actinomycota abundance	0.588	CNN _{CWT,CP}	1-6
Gram positive bacteria abundance	1.577	CNN _{CWT,CP}	1-6
Gram negative bacteria abundance	1.643	CNN _{CWT,CP}	1-6
Arbuscular mycorrhiza fungi abundance	0.211	CNN _{CWT,CP}	1-6
Total bacteria abundance	3.119	CNN _{CWT,CP}	1-5
Zygomycota abundance	0.432	CNN _{CWT,PCA(CP)}	1-2
Total fungi abundance	0.785	CNN _{CWT,CP}	1-5
Total microbial biomass (total PLFAs)	4.430	CNN _{CWT,CP}	1-6

Interestingly, when examining the cases where the best CP input is the group with priority 1—i.e., when the most easily measurable climatic and soil physicochemical properties are used as the only SP input—we find that this configuration appears as the best CP input in six properties in the RMSE table, six properties in the MAE table, and five properties in the R^2 table. This consistency suggests that a subset of climatic physicochemical variables is sufficient to achieve optimal predictive performance across multiple evaluation metrics. The recurrence of this input configuration highlights its potential for computational efficiency, especially in scenarios where minimizing preprocessing complexity is a priority.

Following these results, Figure 3.2 shows the frequency of each configuration of model-CP input as the one providing the best result for each evaluation metric.

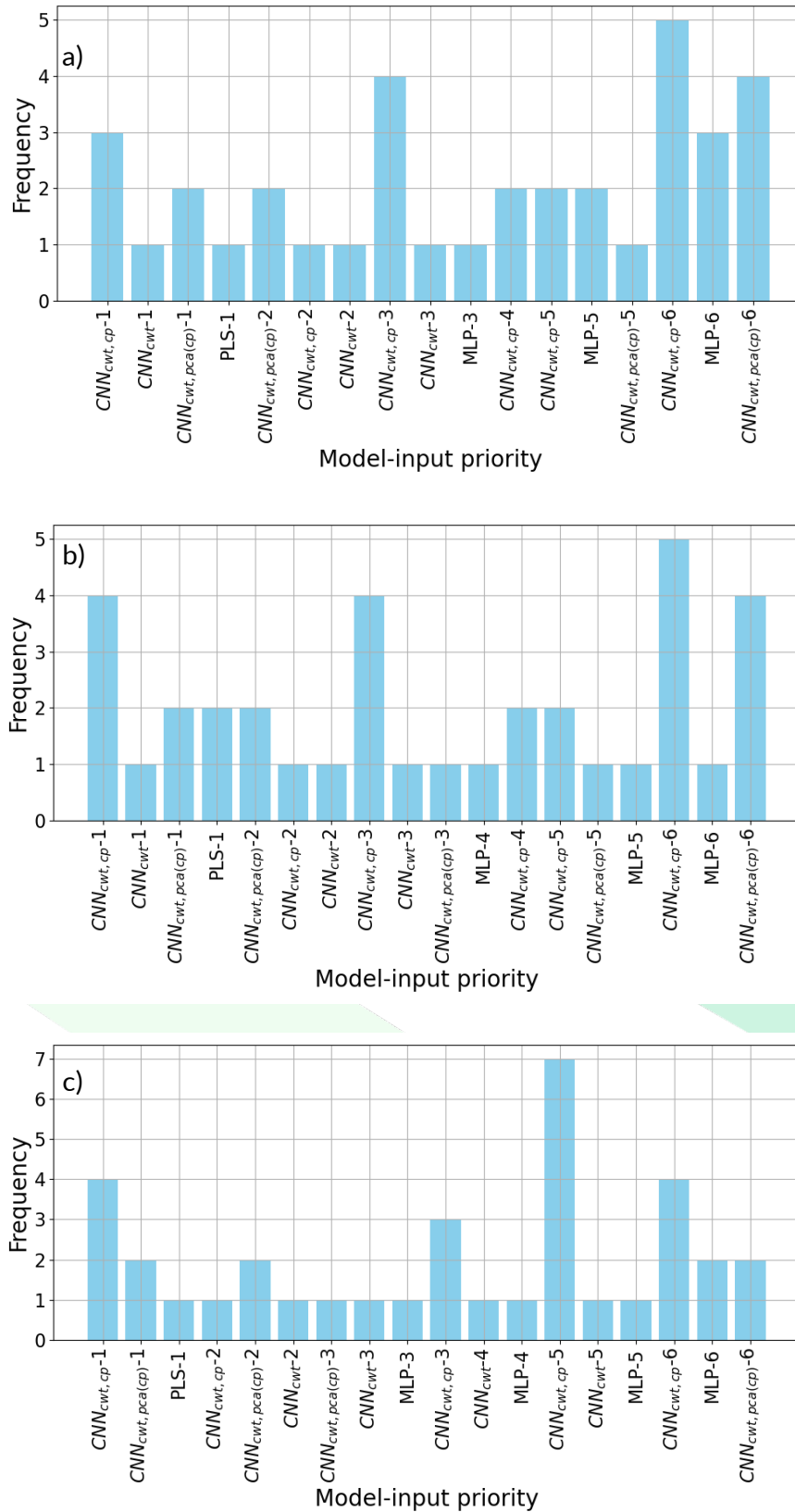


Figure 3.2. Frequency of the tuples (model, CP input) as the best combination according to the three evaluation metrics: R² (a), RMSE (b) and MAE (c).

This way, Figure 3.2a shows that the family of models CNNcwt,cp was the one providing the best results according to the R^2 metric. As a matter of fact, the CNNcwt,cp trained with the groups 1-3 of CP properties was the best one for 4 different biological properties. This frequency increases to 5 when CP groups 1-6 are used as input. Figure 3.2b shows a very similar pattern where CNNcwt,cp using CP properties from group 1 to 6 provided the most accurate prediction in 6 different biological properties. Last, Figure 3.2c shows that the CNNcwt,cp trained with the groups CP groups 1-5 obtained the lowest MAE in 7 different biological properties. The same model trained with group 1 and trained with groups 1-6 was also the one with the lowest MAE for 4 different biological properties respectively.

To sum up, Figure 3.3 shows that CNNcwt,cp model was the most suitable one as it provided the higher R^2 score for 18 BPs, and the lowest MAE and RMSE for the same number of BPs. It is also interesting to observe that the version of the CNN using PCA to reduce the CP dimensionality, CNNcwt,pca(cp), was the best model in 9, 10 and 8 BPs according to the R^2 (Fig. 3.3a), RMSE (Fig. 3.3b) and MAE (Fig. 3.3c). The fact that both models relied on a combination of CWT and CPs confirms the combination of both inputs improved accuracy of the prediction.

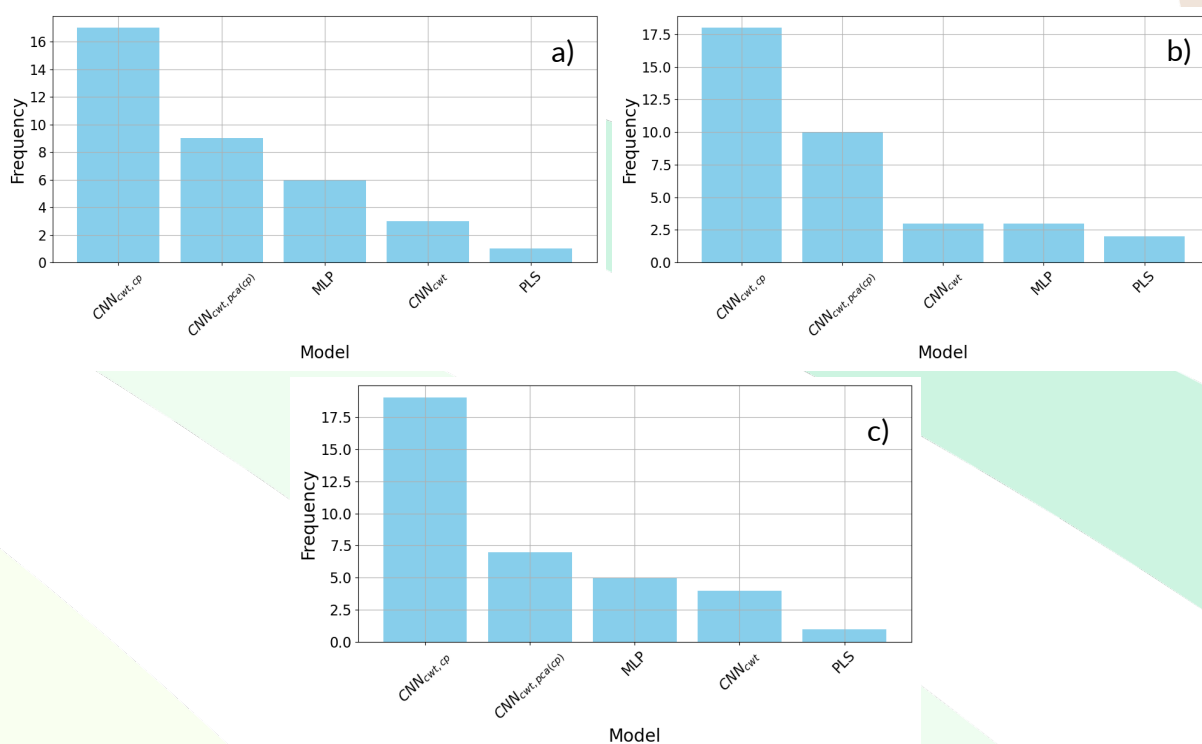


Figure 3.3. Frequency of the target models as the best one according to the three evaluation metrics: R^2 (a), RMSE (b) and MAE (c).

Figure 3.4 shows the frequency of the CP groups defined in Table 2.1 when they were used as input by the most accurate predictor. As we can see, the usage of the 6 CP groups unsurprisingly provided the best prediction for 12 BPs based on the R^2 score (Fig. 3.4a), 10 based on the RMSE (Fig. 3.4b), and 8 considering the MAE (Fig. 3.4c). These differences highlight the distinct aspects of prediction error captured by each metric: while RMSE penalizes larger errors more heavily due to the squaring of residuals, thus being more sensitive to outliers or extreme deviations, MAE

treats all errors equally and offers a more balanced view of average performance across all predictions. Therefore, the fact that slightly different sets of BPs achieve the best score under RMSE and MAE suggests that, in some cases, predictions with small frequent errors (favored by MAE) might not coincide with those minimizing occasional large deviations (penalized more strongly by RMSE). Interestingly, the usage of solely the 5 CPs properties included in group one allowed to obtain the best prediction around 7 or 9 BPs, indicating that a reduced set of relevant variables may still capture significant patterns for a substantial number of target variables, even when evaluated through metrics that emphasize different types of error.

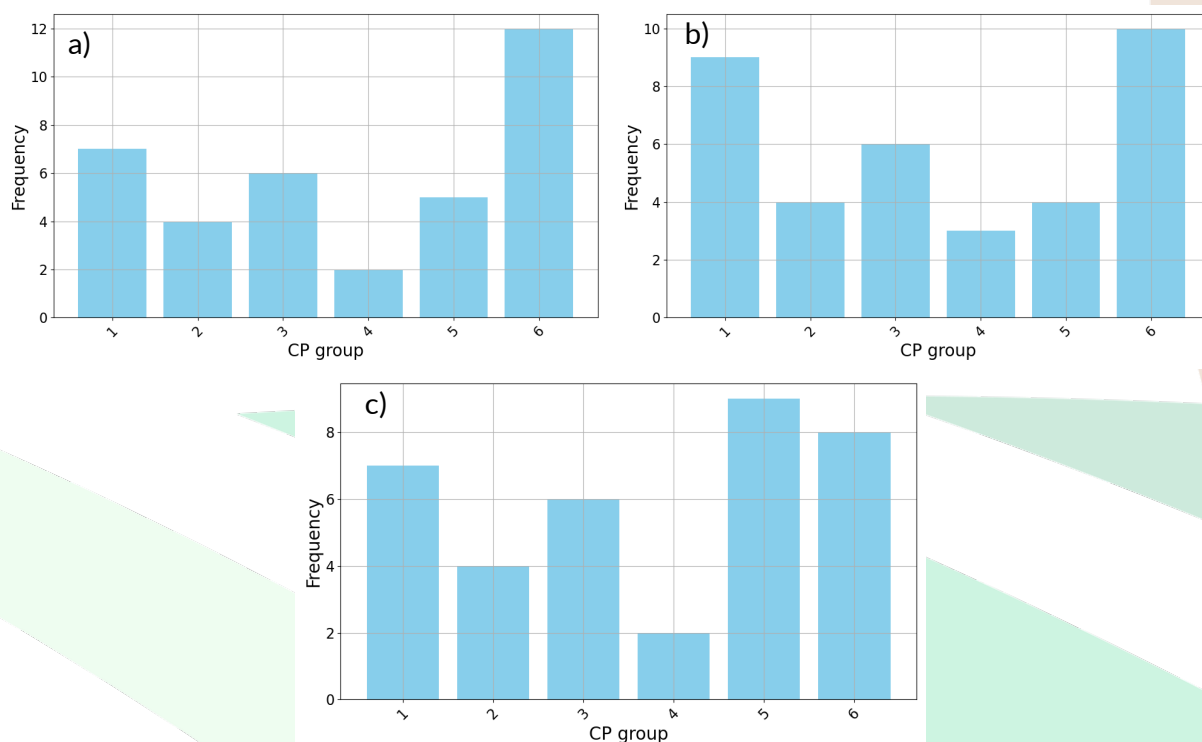


Figure 3.4. Frequency of the CP groups as the best input according to the three evaluation metrics: R^2 (a), RMSE (b) and MAE (c).

All in all, our results show that there is no silver bullet in which a single model provides the best prediction. On the contrary, our evaluation suggests that it is better to use a family of predictors, each focusing on a particular biological indicator. For the sake of completeness, Table 3.4 presents the models to be used and the target biological indicator for each, based on the RMSE scores provided in Table 3.2. As we can see, 18 different combinations of models and CP groups are required to predict all the target soil biological indicators.

Table 3.4. The palette of predictors to be used to estimate all the target soil biological indicators according to the RMSE score. See table 2.1 for CP inputs.

MODEL	CP INPUT	TARGET SOIL BIOLOGICAL INDICATOR
CNN _{cwt,cp}	1	Fungal Shannon index Fungal Pathotrophs-Saprotrophs
	2	Nematode <i>Trichodoridae</i>
	3	Fungal chao1 index Prokaryotic observed number of OTUs Zygomycota abundance Prokaryotic Fisher index
	4	Nematodes Shannon index Nematodes <i>Xiphinema</i> / <i>Longidorus</i> sp
	5	Fungal Inverse Simpson index Nematodes observed number of OTUs Gram positive bacteria abundance Total bacterial abundance
	6	Nematodes <i>Paratylenchus</i> sp Fimicutes abundance Actinomycota abundance Fungal abundance Total bacteria abundance Total microbial biomass (Total PLFAs)
CNN _{cwt,pca(cp)}	1	Fungal Symbiotrophs Nematodes <i>Globodera</i> / <i>Heterodera</i> sp Arbuscular mycorrhiza fungi abundance
	2	Fungal Pathotrophs-Saprotrophs-Symbiotrophs Fungal Pathotrophs-Symbiotrophs
	3	<i>amoA</i> gene
	5	Nematodes <i>Meloidogyne</i> sp
	6	Fungal Saprotrophs-Symbiotrophs Fungal Pathotrophs Nematodes <i>Ditylenchus</i> sp GH7 gene
CNN _{cwt}	1	Nematodes <i>Pratylenchus</i> sp <i>cbbL</i> gene
	2	Fungal yields
	3	Prokaryotic Shannon index
MLP	4	Nematodes absolute number of individuals
	6	Prokaryotic Inverse Simpson index
PLS	1	<i>nirK</i> gene Prokaryotic Simpson index

4 CONCLUSIONS

The following soil biological properties: Fungal chao1, Fungal yields, Fungal Pathotrophs-Saprotrophs, Fungal Pathotrophs-Saprotrophs-Symbiotrophs, Fungal Saprotrophs-Symbiotrophs, Fungal Symbiotrophs, Fungal Pathotrophs, Fungal Pathotrophs-Symbiotrophs, Nematodes *Globodera* / *Heterodera* sp, *amoA* gene, Prokaryotes Observed number of OTUs, Prokaryotes Shannon index, Prokaryotes Fisher index, Firmicutes abundance, Actinomycota abundance, Zygomycota abundance, Gram positive bacteria abundance, Gram positive negative abundance, Arbuscular mycorrhiza fungi abundance, Total bacteria abundance, Total fungi abundance, and Total microbial biomass (total PLFAs), were accurately predicted by FTIR algorithms ($R^2 > 0.80$), with incorporation of additional environmental and soil physicochemical properties to increase the accuracy of the models. Thus, these results confirm the potential use of FTIR spectra to predict soil biological properties, including outputs from DNA metabarcoding such as prokaryotic, fungal, and nematodes diversity indices and functional guilds, nematodes species abundances visually identified, functional genes obtained by qPCR, and microbial abundance by phospholipid fatty acid (PLFA) analysis. This fact could encourage the measurement of these important indicators of soil health as a routine basis by farmers, with reduction of costs and time of analysis and data treatment. These models should be further improved with the incorporation of new samples from other pedoclimatic regions, cropping systems or even land uses to generalise their use as a routine basis.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

5 CODE AVAILABILITY

The source code of the project is available at:

https://github.com/fterroso/SoildiverAgro_deeplearning

6 REFERENCES

- Antoine, J., Murenzi, R., Vandergheynst, P., & Ali, S. (2004). *Two-Dimensional Wavelets and their Relatives*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511543395>.
- Blanco, M., Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. *Trac-Trends in Analytical Chemistry* 21, 240-250.
- Chodak, M., Khanna, P., Horvarth, B, Beese, F. (2004). Near infrared spectroscopy for determination of total and exchangeable cations in geologically heterogeneous forest soils. *Journal of Near Infrared Spectroscopy* 12, 315-324.
- Chen H., Lin Z., Mo L., Wu H., Wu T., Tan C., (2015) Continuous wavelet transform-based feature selection applied to near-infrared spectral diagnosis of cancer, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 151,286-291
- Cohen, M.J., Prenger, J.P., DeBusk, W.F. (2005). Visible-Near infrared reflectance spectroscopy for rapid, non-destructive assessment of wetland soil quality. *Journal of Environmental Quality* 34, 1422-1434.
- Fernández-Calviño, D., Pérez-Rodríguez, P., Arias-Estévez, M., Gómez-Armesto, A., Soto-Gómez, D., Álvarez-Pousa, S., Zornoza, R., Lloret, E., Ollio, I., Sánchez-Navarro, V., Martínez-Martínez, S., Acosta, J. A., Brandt, K. K., Bo Lassen, S., Iversen, S., Pitkänen, J. M., Peltoniemi, K., Rajala, A. P., Eskola, A., ... Waeyenberge, L. (2023). General soil properties of wheat fields along 9 Pedoclimatic regions in Europe (Versión v1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7682445>.
- Johnson, M.J., Lee, K.Y., Scow, K.M. (2003). DNA fingerprinting reveals links among agricultural crops, soil properties, and the composition of soil microbial communities. *Geoderma* 114, 279-303.
- Karl Pearson, F.R.S. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572. <https://doi.org/10.1080/14786440109462720>
- McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), 861, <https://doi.org/10.21105/joss.00861>.
- Ng, W.; Minasny, B.; Montazerolghaem, M. ; Padarian, J.; Ferguson, R.; Bailey, S.; McBratney, A. B. (2019) Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352, 251-267
- Omer, M.; Idowu, O.J.; Brungard, C.W.; Ulery, A.L.; Adedokun, B.; McMillan, N. (2020). Visible Near-Infrared Reflectance and Laser-Induced Breakdown Spectroscopy for Estimating Soil Quality in Arid and Semiarid Agroecosystems. *Soil Systems* 4, 42. <https://doi.org/10.3390/soilsystems4030042>.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817819

Padarian, J., Minasny, B., and McBratney, A. B. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16, e00198. <https://doi.org/10.1016/j.geodrs.2018.e00198>.

Viscarra Rossel, R. A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* 131, 59-75.

Yang, Y., Shen Z., Bissett, A., and Viscarra Rossel, R. (2022): Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer, *Soil* 8, 223-235

Zhang, Y., Freedman, Z.B., Hartemink, A.E., Whitman, T., Huang, J. (2022). Characterizing soil microbial properties using MIR spectra across 12 ecoclimatic zones (NEON sites). *Geoderma* 409, 115647. <https://doi.org/10.1016/j.geoderma.2021.115647>.

Zornoza, R., Guerrero, C., Mataix-Solera, J., Scow, K.M., Arcenegui, V., Mataix-Beneyto, J. (2008). Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biology & Biochemistry* 40, 1923-1930. <https://doi.org/10.1016/j.soilbio.2008.04.003>.