GEOMATICS AND ENVIRONMENTAL ENGINEERING • Volume 15 • Number 1 • 2021

https://doi.org/10.7494/geom.2021.15.1.63

Stanisław Gruszczyński[1]

An Evaluation of Some Machine Learning Algorithms as Tools for Predicting Soil Characteristics Based on Their Spectral Response in the Vis-NIR Range

Abstract: Using the Land Use and Coverage Frame Survey (LUCAS) database of European soil surface layer properties, statistical and machine learning predictive models for several key soil characteristics (clay content, pH in $CaCl_2$, concentration of organic carbon, calcium carbonates and nitrogen and exchange cations capacity) were compared on the basis of processing their spectral responses in the visible (Vis) and near-infrared (NIR) parts. Standard methods of relationship modeling were used: stepwise regression, partial least squares regression and linear regression with input data obtained from principal components analysis. Using the inputs extracted by statistical algorithms various machine learning algorithms were used in the modeling. The usefulness of the models was analyzed by comparison with the values of the determination coefficients, the root mean square error and the distribution of residual values. The mean square error of estimation in the cross-validation procedure for the stack model using the multilayer perceptron and the distributed random forest were as follows: for clay content – ca. 4.5%; for pH – ca. 0.35; for SOC – ca. 7.5 g/kg (0.75% by weight); for $CaCO_3$ content – ca. 19 g/kg; for N content – ca. 0.50 g/kg; and for CEC – ca. 3.5 cmol(+)/kg.

Keywords: machine learning, soil properties, near infrared spectral response, stacked regression models

Received: 1 September 2020; accepted: 30 September 2020

[1] AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering, Krakow, Poland,
email: sgrusz@agh.edu.pl, ORCID ID: https://orcid.org/0000-0002-8811-0954

1. Introduction

The need for large amounts of soil data, when confronted with high costs and time-consuming implementations, justifies the search for alternative testing or survey methods: namely, aerial or satellite level spectrum analyses, and laboratory spectrum analysis [1]. Spectral response, as a source of information on the chemical and physical properties of soils, has been studied since the 1970s [2–5]. Progress in this field has been driven by development of multispectral registration technology for the Earth's surface from the satellite level, which requires improving and refining methods of interpretation of spectral responses to extract useful information on the state of Earth surface. In parallel, another branch of research using the spectral response of soil samples determined under laboratory conditions is rapidly developing, whose aim is to indirectly determine their characteristics obtained by standard, referencing methods requiring complex, expensive and time-consuming preparation. A fair number of research findings have been published demonstrating the usefulness of this both spectral response approach for modeling soil characteristics or indicating significant limitations to it [6].

There is a relatively extensive body of literature addressing the use of soil spectral response in the visible and near infrared range (Vis-NIR) for predicting various soil properties: clay fraction content, organic carbon content, reaction, macro- and micro-elements contents, and cation exchange capacity (CEC), among others [6]. The conclusions are varied, being generally positive with respect to farm scale trials [7], yet generally more cautious with respect to areas with more diverse soil formation conditions. These results, however, are very difficult to compare, since those studies used differing ranges of recorded spectral data, sampling steps, and number of samples. In some cases, the use of models based on principal component analysis (PCA) and partial least squares regression (PLSR) gives satisfactory results for the prediction of soil characteristics; in others, a very complex and computationally demanding procedure of transformation and selection of variables can provide very good results as well [8].

From numerous research works it is reasonable to conclude that when tasked with interpreting the spectral response in the NIR range for the indirect determination of soil variables, various ways of data preprocessing, extraction of useful information from the extensive measurement data, and modeling algorithms of the relationships sought are attempted. The widest range in data transformation and selection is characterized by the PARACUDA II engine [9]. Second place should go to the PLSR algorithm, which could also be used (next to PCA) to select input data from other regression models (e.g., M5, Cubist, MARSpline, random trees, random forests, and other statistical and machine learning methods). An intermediate solution between the two potential approaches would make full use of soil spectral response vectors, as allowed by deep learning models, mainly through convolutional networks. The Convolutional Neural Network (CNN) represents a model

of deep machine learning [10]. Its main application is image analysis, but with some modifications it can be used for regression tasks, also done with 1D inputs. Using CNN has several advantages. First, it does not require selecting spectral data, because the model inputs can be the whole vector of absorbance or reflectance (or some transformation of them), without the need to extract information by means of an additional algorithm (i.e., PCA, PLSR, genetic algorithm). Second, CNN allows for the simultaneous prediction of several variables; this possibility also applies to other neural models in regression applications, such as multilayer perceptron – MLP or radial basis function – RBF, similarly to PLSR). Third, fragments of the data vector that significantly affect the values of the modeled traits are identified in the course of learning. Fourth, the principle underpinning CNNs is based on detecting, in images (CNNs are currently the basic tool for image analysis) or in data vectors, patterns differentiating objects (classification) or data values associated with vectors (regression). Thus, the CNN architecture, apart from the elementary arithmetic limitations of the size of processing layers, is very flexible. Yet this significantly increases uncertainty in the choice of its processing parameters. The CNN models used in practice are generally large in terms of their number of parameters, optimization time, and the risk of overfitting (despite applying techniques to reduce these trends). Understandably, they require testing different architectures, optimization algorithms, and sample sizes. An important feature of modeling is the uncertainty as to the construction of an algorithm suitable for solving the prediction problem. This particularly applies to machine learning (ML) algorithms, which are characterized by considerable freedom in the selection of the number of optimized parameters. Equally important may be the problem of multivariate data acquisition, represented in the case of the spectral response by large sequences recorded at points with a specific reflectance wavelength. Under such conditions, an experimental selection of the model architecture and the method of extraction from data useful in modeling are inevitable.

The aim of this current work was to experimentally assess the acquisition of a large set of spectral data obtained from laboratory testing and the application of different algorithms for predicting soil characteristics based on spectral responses of soil samples, as collected by the LUCAS project, which includes the testing of topsoil surface layers in more than 20 European Union (EU) countries [11, 12].

2. Materials and Methods

The LUCAS database, made publicly available for research purposes by the European Soil Data Center, contains the results of more than 20,000 laboratory tests of soil surface layer samples from 23 EU countries [11]. In addition to data identifying sampling location, land use, soil type and topography, this database includes the values of 12 soil characterization variables (determined by methods considered

as standard) and results of reflected spectrum recordings in the 400–2,500 nm range, in 0.5-nm steps. The data contained in the database were obtained in the same laboratory, according to uniform methodology. Hence, we may presume the factors differentiating the properties of these samples arise mainly from the spatial variability of soils.

The 17,216 mineral soil sample records collected in the LUCAS database were divided, without any soil-use differentiation, at random, into a training (teaching) part for the optimization of prediction models of 12,898 sites. For the model evaluation of independent data, it used 4,318 data sites. This data included the results of the determination of the surface properties of mineral soils' surface layer. For the statistical models, the whole training set was used, but for machine learning models, depending on the applied learning algorithm, an additional test set (15% of training set data) were separated from the training set, in order to apply the early stopping principle. In some of the experiments, the k-folds validation was used (indicated in the text accordingly).

Soil data included, inter alia, the following characteristics: clay, silt and sand content, reaction (pH in $CaCl_2$ and pH in H_2O), organic carbon content (SOC), carbonate content ($CaCO_3$), N, P and K content and cation exchange capacity – CEC [11]. Considering these traits individually, their values had statistical distributions that differed from normal distributions (Tab. 1, Fig. 1): they are characterized by positive skewness – except for slightly negative skewness for pH – and by extremely high potassium and $CaCO_3$ contents, with a strong data concentration at the distributions' mean and a relatively large range of data lying distant from its median, especially for carbonates, phosphorus, and potassium contents. The soil properties listed in Table 1 were those used for modeling based on their spectral response; hereon called "soil variables".

Table 1. Statistics for the distribution of training and validation set variables

Properties	Training set						Validation set					
	Mean	SD	Min	Max	Me	IQR	Mean	SD	Min	Max	Me	IQR
Clay [%]	18.8	12.9	0	79.0	17.0	18.0	18.9	13.0	0	77.0	17.0	19.0
pH (in $CaCl_2$)	5.74	1.35	2.6	9.2	5.8	2.5	5.77	1.36	2.7	8.3	5.9	2.5
SOC [g/kg]	25.2	19.2	0	165.7	18.9	20.0	25.4	19.6	0	160.3	18.8	19.6
$CaCO_3$ [g/kg]	55.0	127.8	0	944	1.0	18.0	58.8	136.2	0	909.0	1.0	20.0
N [g/kg]	1.95	1.22	0	13.6	1.6	1.3	1.97	1.26	0	10.0	1.6	1.2
P [mg/kg]	29.2	29.9	0	532.8	22.1	31.6	30.0	31.0	0	402.7	22.1	3.5
K [mg/kg]	191.0	226.0	0	7342.0	130.0	173.3	190.0	219.0	0	6861.0	134.0	173.0
CEC [cmol/kg]	13.6	9.7	0	137.0	11.3	11.9	13.7	9.7	0	80.1	11.4	11.5

SD – standard deviation, Min – minimum, Max – maximum, Me – median, IQR – interquartile range. Training crop size 12,898 cases, validation crop size 4,318 cases.

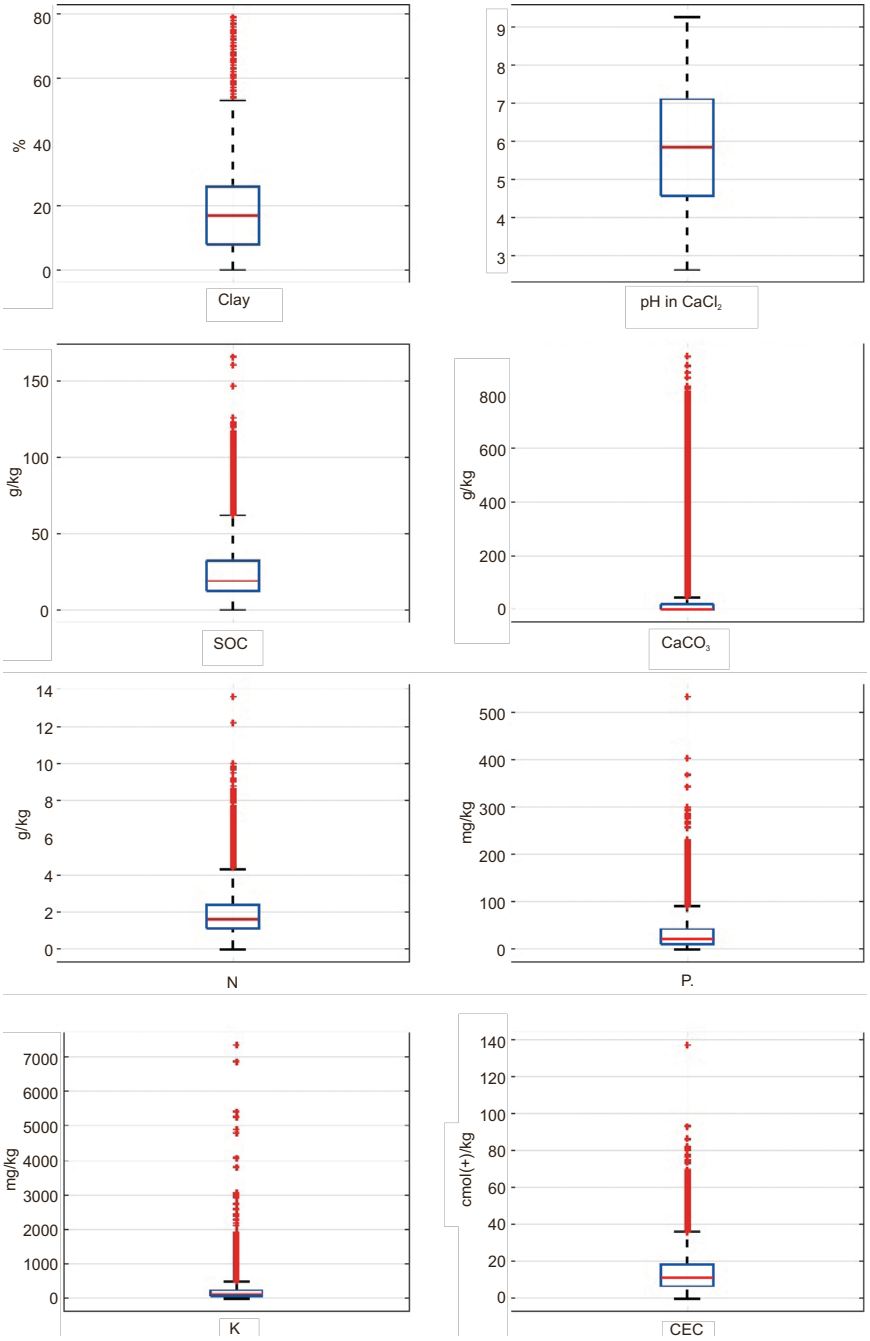

Fig. 1. Boxplots of soil variables. Rectangles indicate the interquartile range (IQR), and red points indicate the elements of the variable set of values distant from the first or third quartile limit (by more than 1.5 × IQR)

Figure 1 shows boxplots of the eight selected soil variables. It indicates, in general, mostly asymmetrical distributions in the data for each. According to the usual interpretation, these boxplots indicate a considerable presence of outlier observations in the data. They are represented in the diagrams by points marked with red crosses, located at a distance of more than 1.5 interquartile distance (IQR), below the lower or above the upper quartile.

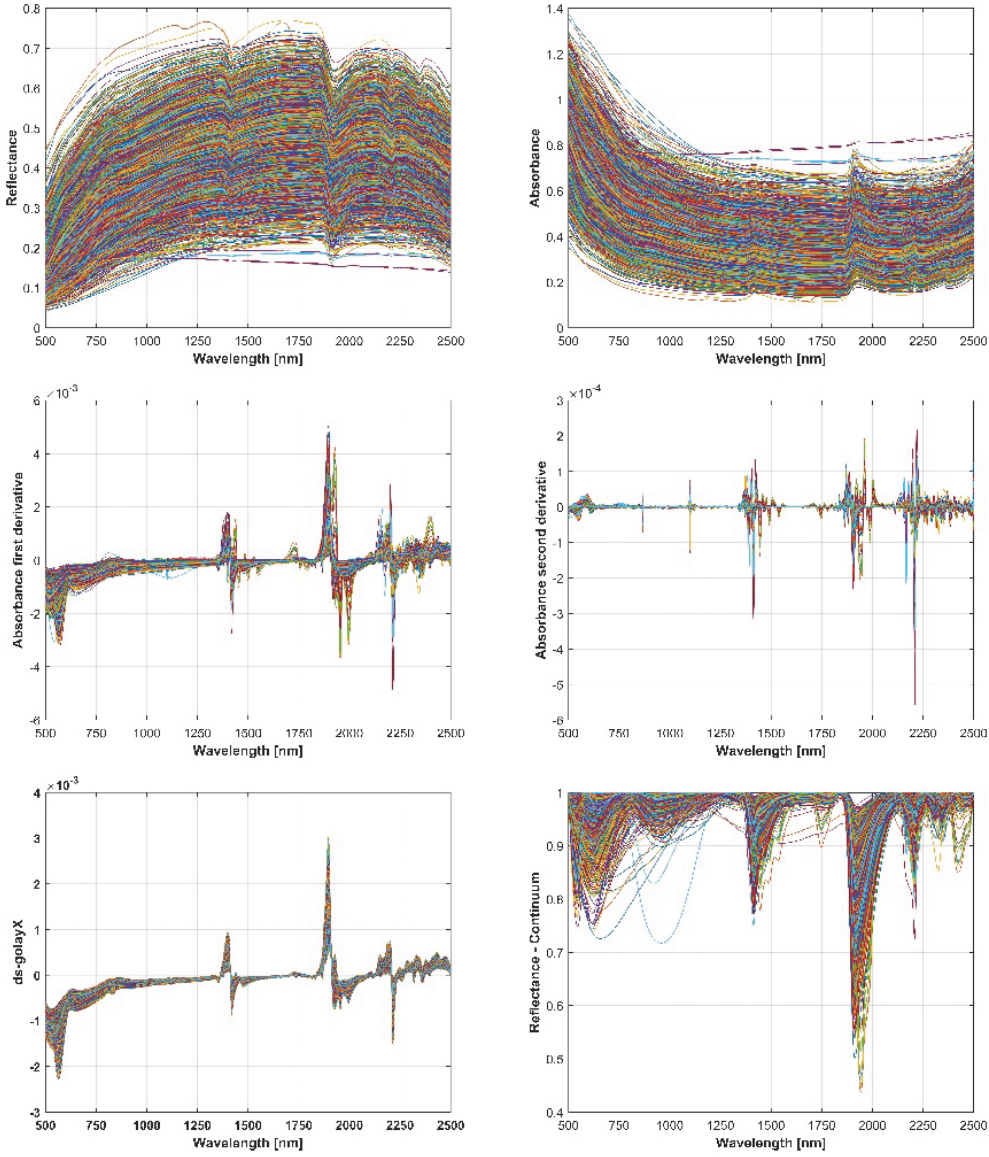

Fig. 2. Absorbance and reflectance curves of the validation set extracted from LUCAS, and transformed curves of the same data

According to usual practice, this would entail their exclusion from modeling; however, this would require the suction: samples with a clay fraction exceeding 53%, a silt and sand fraction content and a distribution of both pH-values would be fully within the acceptable variation, whereas samples with an SOC content greater than 62 g/kg (6.2% by weight), as well as samples with carbonate content greater than 45 g/kg (4.5% by weight), 4.35 g/kg N and CEC greater than 36 cmol(+)/kg, would be discarded. Presumably the presence of outliers in the modeled variables may come from measurement errors not excluded by chance, or represent the actual data distribution. Due to the diligence of data acquisition [11], it can be assumed that this soil data correctly reflects the distribution of each soil variable, while the removal of the indicated data would distort the actual picture of soil conditions' variation. Doing the latter would run against the aim of the experiment, which was to assess the possibility of obtaining a universal prediction model, which may be characterized by a specific, even significant, prediction error for the soil variables.

In the modeling tests, in addition to the spectral absorbance data of the samples in the range 500–2500 nm (matrix of absorbance of soil samples marked with an X symbol in this paper), pre-processed data were used to adapt them to the format applied in the spectral response analysis (refer to [6]): that is, first absorbance derivative (matrix dX), second absorbance derivative (matrix d2X), sample reflectance (matrix RefX), first absorbance derivative after Savitzky–Golay filtering (matrix dsgolayX, framelength of 11, order 4) and continuum removal [13] reflectance (matrix CRX). Transformations, based on absorbance values provided by ESDAC, were performed in the MATLAB environment [14]. Figure 2 shows the transformed spectral data vectors of the validation set extracted from LUCAS.

In testing models for their predictions of soil sample characteristics on the basis of spectral response in the Vis-NIR range, experiments were carried out using statistical algorithms, machine learning, and their combinations, as follows:

1. **Linear and MLP models with inputs obtained by stepwise regression**. Input variables of the model were JX matrix, horizontal concatency matrices X, dX, d2X, RefX, ds-golayX, and CRX (a total of 23,996 columns). The variables modeled were clay fraction content (Clay), pH (pH in $CaCo_3$), SOC content (SOC), $CaCO_3$ content ($CaCO_3$), N, P, K contents (respectively: N, P, K), and cation exchange capacity (CEC). Subsequently, separate models were developed by stepwise regression method for individual soil variables (forward regression, *p-enter* = 0.05, *p-remove* = 0.10). In the second stage, the variables extracted through stepwise regression were used as inputs in the machine learning models (i.e., individual MLP models for soil variables). The calculations were performed in the MATLAB [14] environment.
2. **Linear and MLP models with inputs obtained by PLS**. The input to the eight PLSR models was the JX matrix (similarly as in the stepwise regression). The outputs were soil variables: Clay, pH in $CaCO_3$, SOC, $CaCO_3$, N, P, K, and CEC. The accepted number of components of the PLSR regression

model was 150. This value is considered excessive by convention but owing to the input vector lengths, and to assess the maximum value of the determination coefficient, this solution was adopted here. Subsequently, MLP models were developed with PLSR components as inputs and soil variables as outputs. The calculations were performed in the MATLAB [14] environment.

3. **Linear and MLP models with inputs of PCA components**. The first 150 PCA components extracted from the JX data vector were used to develop linear regression models for the eight soil variables. The same components were then used as input vectors of MLP models (number of hidden units in the samples was 2, 4, 10, 20) for each soil variable. The calculations were performed in the MATLAB [14] environment.
4. **Linear and MLP models for clustered data**. Using the Kohonen SOM clustering algorithm [15], the elements of the JX matrix were divided into nine homogeneous groups. The algorithm groups multidimensional variables, by using the criterion of similarity of their vectors. For individual clusters (i.e., homogeneous groups) and modeled variables (Clay, pH, SOC, $CaCO_3$, N, P, K, CEC.), separate machine learning models (MLP) were developed, whose inputs were components extracted through stepwise regression. The calculations were performed in the MATLAB [14] environment.
5. **Cubist regression models with PLSR components as inputs**. This included the following data processing steps: (1) for some soil variables (Clay, pH in $CaCl_2$, SOC, $CaCO_3$, N, CEC) and spectral data (X, dX, d2X, RefX, ds--golayX, CRX), 25 components of the partial least squares regression (6 × 25 component vectors per soil variable) were calculated); (2) the vectors from the previous step, for each soil variable separately, were combined to form an input vector of 150 input components; and (3) separate regression models were developed for the soil variables using the Cubist algorithm [16] implemented in the R statistical platform [17].
6. **Stacking regression: 0-level data – PLSR components of the spectral data matrices, 0-level models – MLP models, 1-level data – horizontal concatenation of MLP models prediction results, 1-level models – various machine learning algorithms**. This approach [18, 19], was obtained in four steps. (1) For selected soil features (Clay, pH in $CaCl_2$, SOC, $CaCO_3$, N, CEC) and spectral data (X, dX, d2X, RefX, ds-golayX, CRX), 25 components of the least squares partial regression were calculated (25 components for each data matrix and soil variable, 0-level data). (2) For each of the matrices of the PLSR components created in step one, three MLPs having 5, 10, and 20 units in the hidden layer per soil variable were created (18 MLPs of the models in total per soil variable, composing a collection of 0-level models). (3) The horizontal concativity of 0-level prediction models in the matrix of predictions (108 estimates = 18 × 6 soil variable), were 108 elementary 1-level data, consisting of independent estimates of soil variables.

(4) Using a selection algorithm, called 'Boruta' [20], for variables relevant to variable prediction, were selected as inputs; specifically, the vector of combined predictions created input variables of the below machine learning models served as a 1-level model in the stacking regression, based on indication vectors of 108 weak regression algorithms (corresponding to the 0-level of the stacking model). The following algorithms, as 1-level model were used: (5a) variants of MLP models, with 3, 6, 7 and 10 units in their hidden layer, for which the model with the smallest RMSE values for test data was selected; (5b) on the basis of the same input variables, individual models for soil variables were developed according to Quinlan's M5 decision tree algorithm [16], by implementing M5P of the WEKA [21] calculation package; (5c) the M5 algorithm implementation named Cubist was used [17, 22]; (5d) decision module Gradient Boosting Machine (GBM) random tree algorithm, run in the software platform H2O [23]; (5e) the Distributed Random Forest (DRF) decision module [23, 24], also implemented in H2O platform. The Cubist and M5P packages are an implementation of Quinlan's M5 algorithm combines a conventional decision tree with the possibility of linear regression functions at the nodes. Data assigned to the nodes of a random tree are used to build a linear model instead of averaging [16]. GBM is a machine learning algorithm in which a set (random forest) of multiple weak regression (or classification) models is used to iteratively optimize the prediction. The algorithm is iterative incrementing the regression tree architecture taking into account the value of the residuals from the previous state of the model [25]. DRF is a machine learning algorithm that iteratively optimizes a random forest structure by generating trees that use data from random subsets of training data [24].

7. **Convolutional Neural Network (CNN)** with a 1D input and a regression output consisting of six units corresponding to the eight soil variables examined (clay, pH in $CaCl_2$, SOC, $CaCO_3$, N, CEC). The CNN network architecture was used with a 1D input in the Keras environment (implemented in the R Studio as well as in Anaconda–Python language environment), with a TensorFlow library. Reflectance vectors and SNV-transformed absorbance vectors [13] ere used as inputs to the model. The CNN architecture consisted of four convolution 1D layers with a 50–75-unit wide filter (in first convolutional layer), and 20-units filters (subsequent layers), with, respectively 32, 64 128 and 256 of filters. The data from the convolution layers were then transferred to the MaxPooling layers, while the model output included the layers "flatten" (transformation of data matrix into vectors), "dropout" (removal of small-scale connections), and two "fully connected" layers, which are equivalent to an MLP network with an output of six linear units. For the processing layers (other than the output per se) the ReLU transfer function was used. When not able to determine, *a priori*, the appropriate architecture

of the CNN network, many attempts were made regarding the size of evolutionary filters, their number, as well as the size of fully-connected layers. Three optimization algorithms were used: Adam, Adadelta, RMSprop, in addition to the size of the "batch" (batch) data and number of optimization cycles [26–29].

The following statistics were used to evaluate the data distributions and prediction results obtained:

- determination coefficient $R^2 = 1 - SSR/SST$, where SSR is the sum of squares of the model's residuals, SST is the sum of squares of deviations of the modeled variable from the average;
- root mean square error: $RMSE = \sqrt{SSR / n}$, where n is number of observations;
- interquartile range: $IQR = Q_3 - Q_1$, where Q_1 and Q_3, respectively: first and third quartiles of distribution;
- lower quartile, median and upper quartile residue: RQ-0.25, RQ-0.5, RQ-0.75;
- interquartile range of the distribution of the residue: RIQR = (RQ-0.75) − (RQ-0.25);
- RPIQ (Ratio of Performance to InterQuartile distance), $RPIQ = Q_3 - Q_1/RMSE$;
- RPD (Residual Prediction Deviation), RPD = SD/RMSE, where SD is the standard deviation.

3. Results

In the scientific literature on the use of NIR in soil research, the most frequently modelled feature is the organic carbon content (SOC). It can be assumed that the monotonic dependence of the spectral response and the feature identifies the spectrum range best for quantifying the feature value.

The Figure 3 graphs show the curves of Spearman's correlation coefficients values (*rho*) for the transformed absorbance vectors and organic carbon content (SOC) in the soil samples. (Note that the curves relate to the SOC variable and are only relevant for this relationship.) From their analysis, the following conclusions may be drawn: (1) For the range of SOC concentrations, it is not possible to distinguish that part of the spectrum where the spectral response would allow an error-free estimation of the SOC content, on this basis; hence, it can be assumed that the linear model may show a lower usefulness in this range. (2) Both absorbance and reflectance are relatively weak predicators of SOC: a flat course of their curves ϱ (curves are mutually symmetrical with respect to *rho* = 0) suggests mainly the role of the ordinates of reflected or absorbed signals, as an indicator of SOC concentration. (3) Absorbance derivatives, filtered and unfiltered, and the removal of the continuum, jointly indicate the presence of spectrum fragments, for which the influence of SOC concentration – or factors correlated with it – upon the growth of reflected or absorbed signals is conveyed. (4) The curves' form may indicate the need for the simultaneous

use of different transformations of absorbance and reflectance vectors as input information for soil variable prediction models. Given the size of the input vectors, it was necessary to extract the most useful information, best achieved by statistical algorithms: stepwise regression, partial least squares regression, and of principal components analysis.

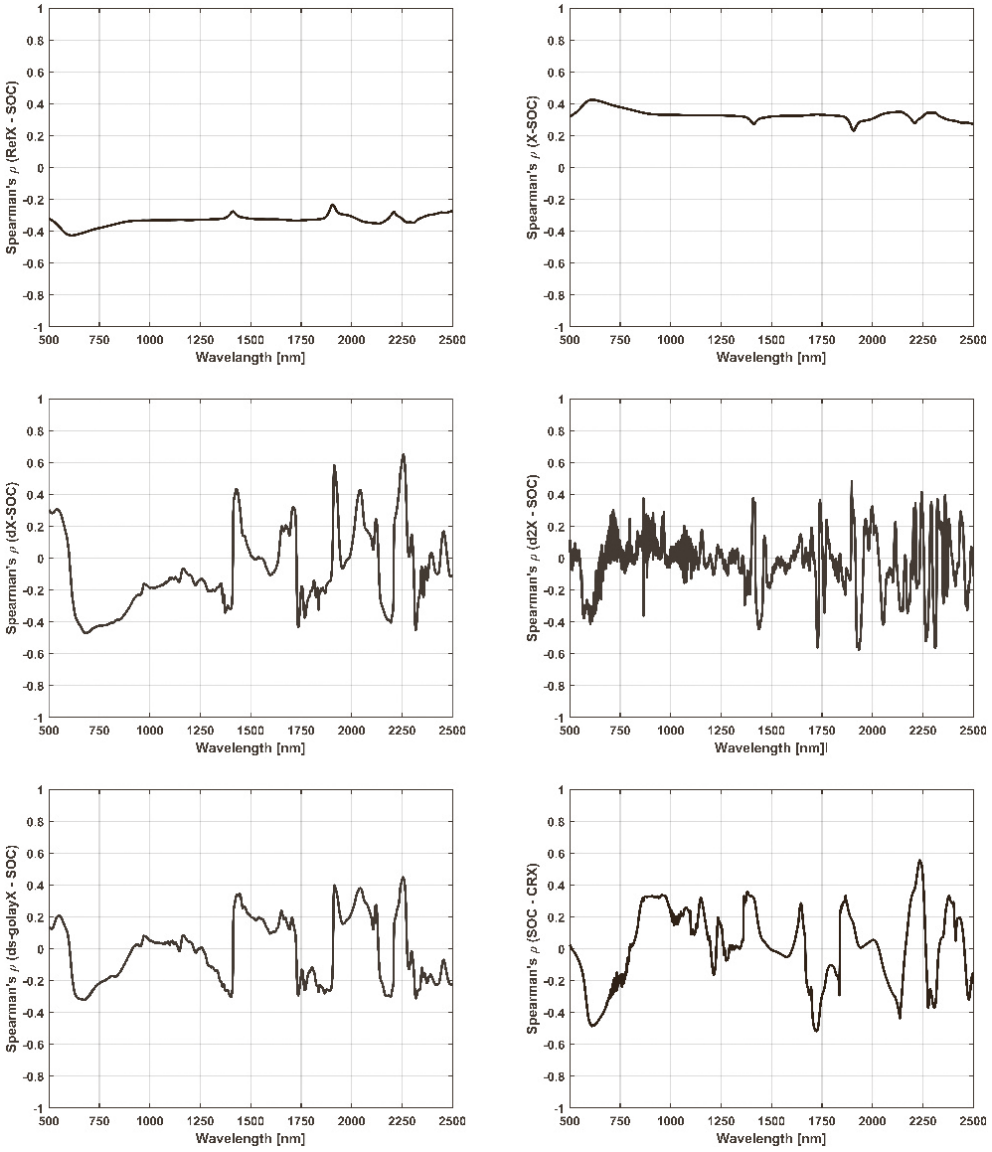

Fig. 3. Curves of Spearman correlation coefficient values (*rho*) between the concentrations of SOC in the samples and the six transformed absorbance values of the soil samples

3.1. Linear and MLP Models with Inputs Obtained by Stepwise Regression, PLSR and PCA (JX Datamatrix)

Table 2 contains the statistics for all seven models of the eight soil variables whose input vector was the JX vector of combined data. The step regression model especially the comparison of RMSE errors from training and validation data, indicates the differentiation among models. Because differences in empirical distribution parameters (i.e., mean and standard deviation) of the training and validation data are very similar, we should presume they are differentiated by distributions in the spectral response as well as random interactions of other soil elements not included in the models. It should be noted that the step regression algorithm extracted from almost 24,000 variables (combined vectors of transformed reflectance values) several hundred (from 344 to 558) variables that significantly impacted the prediction of selected properties. Yet the evaluation of regression models varied. The error of clay fraction content estimation, both for the training and validation sets (4.8% and 5.7%, respectively), may be deemed acceptable, leading to a possible texture classification error of no more than one group.

Table 2. Statistics of prediction models for soil variables based on the JX vector

Model	Statistic	Clay	pH in $CaCl_2$	SOC	$CaCO_3$	N	P	K	CEC
StepReg (T)	Ncomp	557	524	469	436	548	344	558	504
	R^2	0.87	0.93	0.80	0.95	0.81	0.46	0.67	0.84
	RMSE	4.81	0.36	8.69	29.32	0.54	22.21	132.78	3.99
StepReg (V)	R^2	0.81	0.90	0.73	0.94	0.73	0.34	0.34	0.80
	RMSE	5.65	0.41	10.21	33.69	0.65	25.19	177.58	4.32
	RPD	2.30	3.30	1.91	4.04	1.93	1.22	1.23	2.24
MLP (SR-V)	Hidden	3	3	2	2	3	2	3	3
	R^2	0.83	0.94	0.74	0.96	0.78	0.38	0.45	0.82
	RMSE	5.29	0.34	9.99	27.87	0.60	24.45	163.38	4.16
PLSR (V)	R^2	0.82	0.91	0.75	0.94	0.75	0.37	0.41	0.81
	RMSE	5.52	0.41	9.73	33.48	0.63	24.69	168.29	4.20
MLP (PLSRcoef)	R^2	0.84	0.91	0.78	0.95	0.76	0.32	0.40	0.78
	RMSE	5.11	0.40	9.18	31.96	0.61	25.61	170.02	4.50
LReg (PCA-150)	R^2	0.76	0.84	0.69	0.92	0.68	0.28	0.35	0.75
	RMSE	6.32	0.53	10.81	39.57	0.71	26.31	176.67	4.84

Ncomp – step regression components number; Hidden – hidden units number; RPD – ratio of performance to deviation; StepReg (T) – step regression with training data; StepReg (V) – step regression with validation data; MLP (SR-V) – MultiLayer Perceptron with input variables extracted by step regression (validation set); PLSR (V) – partial least squares regression (validation set); MLP (PLSRcoef) – MultiLayer Perceptron with inputs of PLS (validation set); LReg (PCA-150) – linear regression with 150 PCA components as inputs; MLP (PCA-150) – MultiLayer Perceptron with 150 PCA components as inputs. The best validation results are marked in bold type.

Prediction of the reaction (pH in $CaCl_2$), could also be considered acceptable, but only for a rough estimate, while the SOC estimation error (8.7 and 10.2 g/kg respectively, against the median value ca. 19 g/kg) was relatively high, especially for soils with low organic carbon content (e.g., sandy soils). The high determination factor of the model for carbonate content, given the asymmetry and significant dispersion of this characteristic in the sample, generated a relatively high RMSE estimation error (ca. 30 g/kg, against the median value ca. 1), which may disqualify it for the vast majority of soils having low carbonate content. The nitrogen (N) content was estimated with relative accuracy analogous to SOC, probably due to the correlation of these two components. The concentrations of P and K were estimated with significant error, while some error in estimating CEC is acceptable for soils with a higher clay fraction.

The decision on the number of components used in the PLSR model is arbitrary, although it may be determined *a posteriori*. Figure 4 illustrates the increase in the determination factor as the number of PLS components is increased. Most of these graphs show ca. 20 to 50 such components are sufficient to model output variables close to the maximum accuracy offered by this algorithm. The graphs also indicate that predicted N, P, and K values are relatively imprecise. Comparing the results (R^2 and RMSE) of prediction data validation between the step regression algorithm and PLSR suggests the latter (partial least squares regression) is the better tool for building a soil prediction model (though differences in prediction errors are nevertheless quite small).

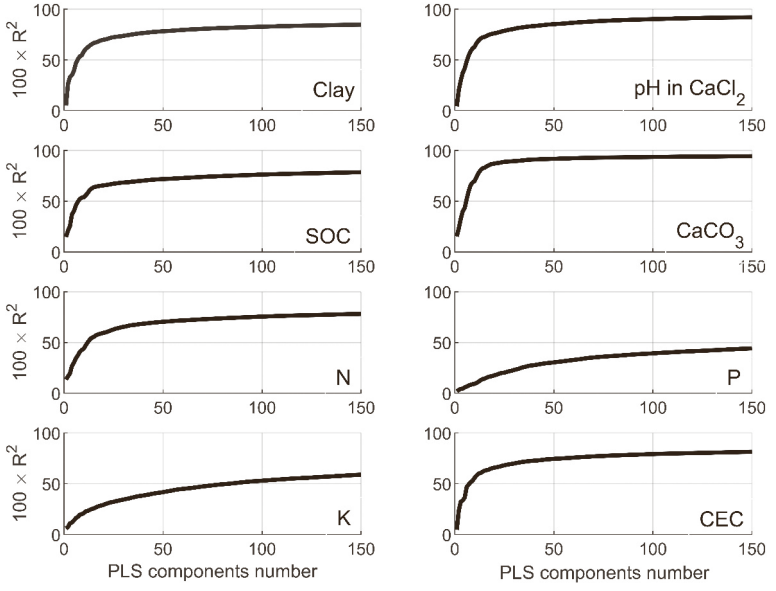

Fig. 4. Values of 100 × R^2 (percentage of variability explained by the PLSR model), plotted as function of the number of PLS algorithm components, for the eight individual soil variables

The construction of neural prediction models, in which inputs are explanatory variables selected in a step-by-step regression procedure, or components obtained in the partial least squares regression procedure, may improve prediction quality since this includes natural non-linearity of the machine learning algorithm architecture in the model. But this must be done iteratively, because, despite a significant reduction in the number of explanatory variables (down from >20,000 to several hundred), the number of input variables remains relatively high. It should be noted that in the case of architectures with non-local transfer functions, increasing the number of hidden layer units multiplies the number of parameters that must be optimized in the learning process, thereby elevating the risk of overfitting and greater validation error (decreasing the ability to generalize the model). For MLP models developed here for soil variables contained in the LUCAS database, with inputs from step regression or PLSR components, exceeding the size of the hidden layer by 2 or 3 units (values determined in a multiple-trial process) significantly increased their validation error. Generally, the validation errors of MLP were smaller than those of statistical models (step regression and PLSR), while mostly smaller for those models whose inputs were determined via step regression procedures.

3.2. Linear and MLP Models for Clustered Data

One way to increase the accuracy of the predictions worth considering is to group the vectors of the input fields according to their similarity, and then create separate spectral response libraries. Greater forecasting accuracy may result from clustering the inputs to such sets for which separate forecasting models can be constructed, based on input data that are more homogeneous than for the whole population. In clustering the JX vectors, the fragment covering the first derivative of absorbance was used, since the experiments showed this had the largest quantitative share of the set of inputs emerging from the stepwise regression process.

Figure 5 shows the distribution of the vectors of the first absorbance derivatives on the SOM map. The starting point of the SOM clustering process is random, making the final clusters' layout on the map unalike in different attempts, though still maintaining the spatial relationships and mutual distances between the clusters. For each cluster, prediction models were created with inputs obtained in the stepwise regression procedure of JX vectors belonging to that cluster; in each, the prediction model was the MLP algorithm, trained by Bayesian regularization in three repetitions, having 2, 3 and 4 units in the hidden layer. The model with the lowest MSE value for test data was then selected. Using this latter model, the values of soil variables in the validation set were predicted.

Figure 6 shows the distribution of mean square error elements from prediction models for soil variables in SOM clusters. Highlighted by bold values are those clusters in which RMSE values of the validation set are smaller than those obtained on the basis of MLP models without division into clusters, with inputs obtained in a stepwise regression procedure and PLSR.

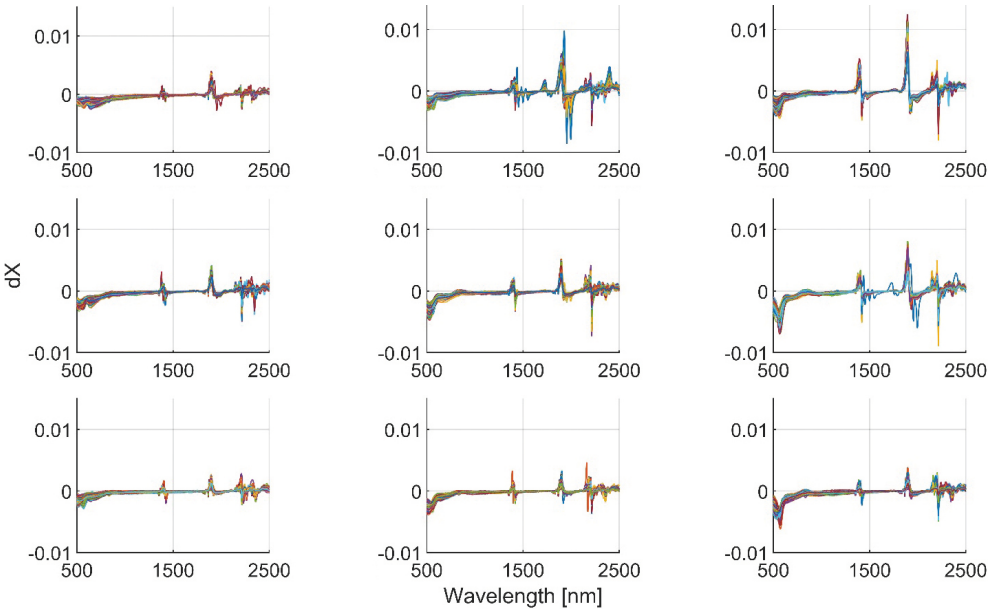

Fig. 5. Distributions of dX vectors on the Kohnen Self Organizing Map (SOM)

Clay	**4.90**	**4.90**	6.80
pH	0.44	0.44	0.53
SOC	13.31	**5.90**	11.70
$CaCO_3$	88.30	36.00	88.20
N	**0.58**	**0.42**	0.71
P	**23.20**	**24.70**	27.70
K	**185.00**	**139.00**	216.00
CEC	**3.44**	**3.66**	7.97
n%	3.0	8.6	13.7

Clay	5.90	7.80	7.90
pH	0.41	0.44	0.37
SOC	13.80	**5.90**	**5.00**
$CaCO_3$	101.00	35.60	52.90
N	0.92	**0.50**	0.88
P	33.10	25.10	25.70
K	406.00	283.00	273.00
CEC	6.43	6.08	9.08
n%	3.4	9.9	16.5

Clay	**4.70**	8.40	8.30
pH	0.35	0.35	0.54
SOC	17.90	**5.40**	**4.90**
$CaCO_3$	43.80	52.00	37.50
N	0.99	**0.48**	**0.43**
P	30.30	**23.90**	**22.70**
K	**105.00**	319.00	208.00
CEC	5.74	7.11	7.88
n%	13.7	8.6	3.0

Fig. 6. Distribution of RMSE values of the validation set variables forecast in groups distributed on the SOM map. Variables whose RMSE value of the validation set was lower than that of the MLP model validation – with inputs of PLS coefficients and variables chosen via stepwise regression algorithm without clustering – are highlighted by bold values

In addition, the validation for all soil variables did not give better results than those in Table 2. For several soil variables, in some clusters the RMSEs were smaller than those of the overall models. Yet, apart from cases of strong local relationships between the spectral response and a soil variable, there is also the phenomenon of clusters featuring very weak predictive models but very low dispersion of a soil variable. Such circumstance arose in the case of the third cluster and SOC variable (Fig. 7), where the variables' dispersion is quite small, generating a low RMSE value with a low $R^2 = 0.52$.

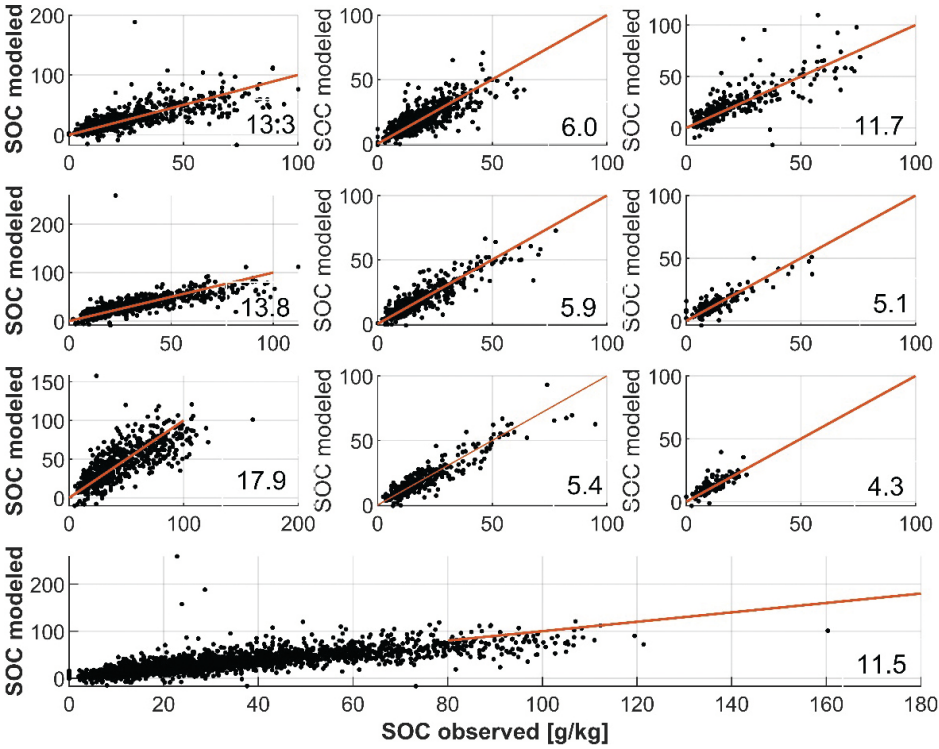

Fig. 7. Scatterplots of observed and modeled (predicted) SOC values for SOM clusters. The bottom panel depicts all points together in a single graphic. Each panel above it shows RSMP data values for a given cluster

3.3. Cubist Regression Models with PLSR Components as Inputs

Committees of classification and regression models provide better quality prediction in many cases. Nevertheless, the models themselves can generate relatively large forecasting errors. Combining their estimates, either using separate data or coming from different types of models, can improve the ability to better generalize. A key issue here is choosing the way the team infers this. The classic solution is to create what is called a stack system. This consists of many models with different properties

(differing inputs, architectures, and ways of optimization) and a decision algorithm, optimized in relation to the same outputs, providing a solution based on the indications of 1-level models [18]. Table 3 contains statistics on validations of six stacked models whose inputs were the vector of combined (concatected) PLS components obtained from spectral data X, dX, d2X, RefX, dsgolayX, and CRX along with relevant soil variables from the PLSR procedures. Qualitatively, these models were characterized by higher RMSE values for the validation set than those models created solely on the basis of PLS components extracted from the whole JX vector (included in Table 2). The presented models, however, are not a strict realization of the idea of stack regression, in that its inputs are not estimates of modeled variables but rather are values of PLS coefficients.

Table 3. Statistics of Cubist models whose inputs were connected by the first 25 PLSR coefficients from vectors to X, dX, d2X, RefX, CRX, ds-golayX outputs (soil variables). Cubist algorithm employed the "boosting" option, and predictions with tuning took into account the averaging of five nearest-neighbors

Properties	Clay	pH in $CaCl_2$	SOC	$CaCO_3$	N	CEC
R^2	0.84	0.91	0.77	0.96	0.77	0.81
RMSE	5.25	0.41	9.41	27.47	0.61	4.25
RIQR	4.97	0.38	6.40	2.04	0.49	4.01
RQ-0.25	–2.14	–0.20	–3.16	–1.21	–0.21	–2.56
RQ-0.50	0.17	–0.01	–0.03	–0.22	0.03	–0.58
RQ-0.75	2.84	0.18	3.24	0.83	0.28	1.45
RPIQ	3.62	6.14	2.08	0.73	1.97	2.71

3.4. Stacking Regression

It can be assumed that the inclusion of nonlinear models (e.g. MLP or other machine learning algorithms) will improve the prediction quality of the stack regression algorithm. Moreover, concatenating the predictions of all 0-level models (for all soil variables) and thus creating a 1-level data may improve the quality of the prediction by taking into account the interrelationships between soil variables. The 1-level model serving as a decision module remains an open issue. Table 4 contains the statistics of a stacked regression model with various inference modules, in which inputs were estimations of soil variables made by the MLP model set. For a given soil variable, differences between the element values of the root mean square error were relatively small: 5.02–5.38 for clay content, 0.39–0.41 for pH, 8.5–9.4 for SOC, 26.4–29.6 for $CaCO_3$, 0.58–0.61 for N, 3.9–4.2 for CEC; however, the relationships between the values of determination factors and RPIQ were similar. Machine-learning models are, in part, random, and decisions made by their designer are subjective and devoid of objective premises for using a particular architecture. It can only be stated that among the examined models, a specific architecture has the best properties in terms of a specific evaluation criterion (Fig. 8).

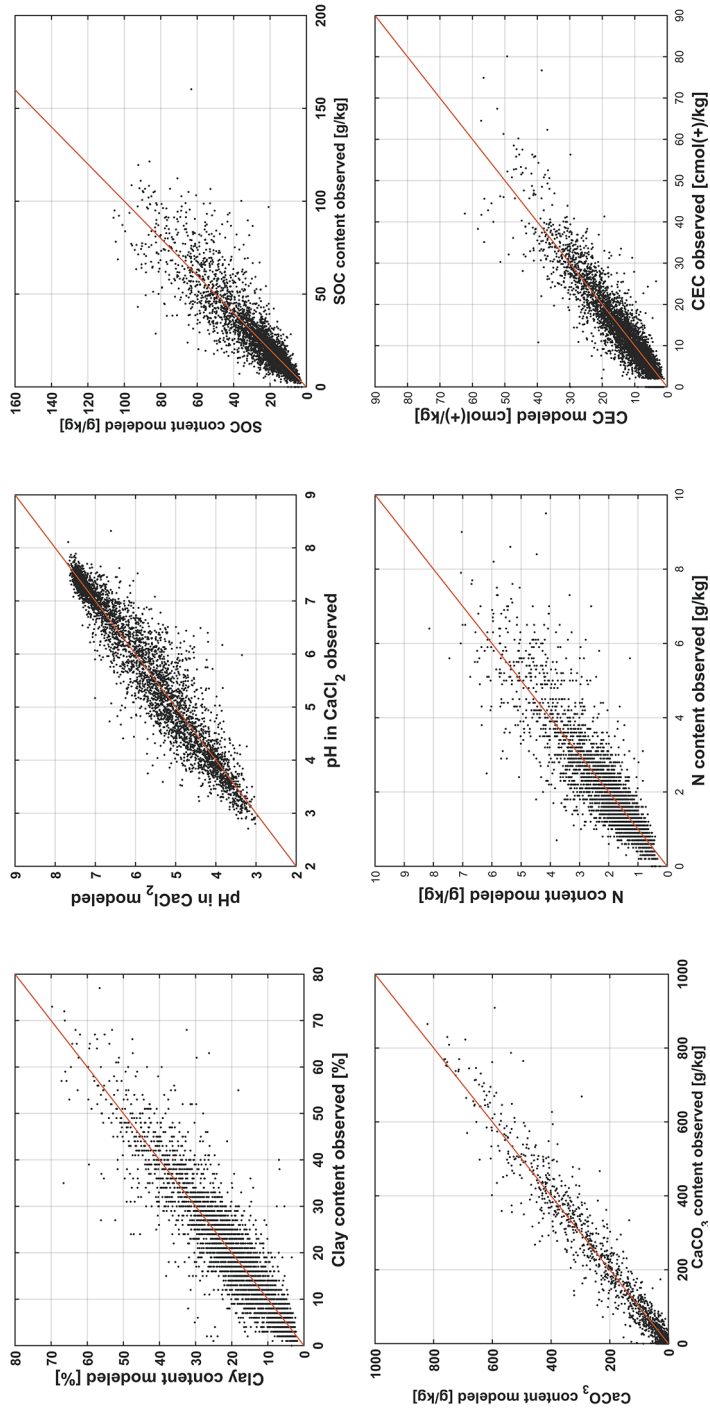

Fig. 8. Scatterplots of relationships between the values of observed soil variables and their estimation by applied models incorporating the DRF decision module, based on validation data

From the testing, in terms of R^2 and RMSE, the model with the DRF decision module is best, although differences among models' predictions were relatively small. Quality indicators of models are their RPIQ values and observed distribution parameters of residual values (leads).

Table 4. Statistics of stacked regressions incorporating different 1-level models, and corresponding statistics on the distribution of residual values of the validation data set

1-level model	Properties	Clay	pH (in $CaCl_2$)	SOC	$CaCO_3$	N	CEC
MLP	R^2	0.83	0.91	0.77	0.95	0.77	0.82
	RMSE	5.38	0.41	9.44	29.62	0.61	4.15
	RIQR	4.97	0.40	6.60	9.22	0.50	3.92
	RQ-0.25	−2.55	−0.21	−3.67	−5.12	−0.26	−2.11
	RQ-0.75	2.42	0.18	2.93	4.10	0.24	1.81
M5P	R^2	0.83	0.91	0.77	0.95	0.77	0.82
	RMSE	5.38	0.41	9.44	29.62	0.61	4.15
	RIQR	4.91	0.45	6.41	1.42	0.59	3.91
	RQ-0.25	−2.54	−0.24	−3.52	−0.71	−0.40	−2.10
	RQ-0.75	2.37	0.21	2.89	0.71	0.19	1.81
	RPIQ	3.53	6.19	2.07	0.67	1.96	2.77
Cubist	R^2	0.84	0.91	0.77	0.96	0.77	0.82
	RMSE	5.22	0.40	9.33	27.33	0.60	4.09
	RIQR	4.81	0.36	6.39	1.67	0.49	3.91
	RQ-0.25	−1.99	−0.18	−2.95	−0.33	−0.19	−2.52
	RQ-0.75	2.82	0.18	3.44	1.35	0.30	1.38
	RPIQ	3.64	6.36	2.10	0.73	1.99	2.81
GBM	R^2	0.85	0.92	0.79	0.96	0.78	0.83
	RMSE	5.07	0.39	9.01	27.24	0.59	3.96
	RIQR	4.94	0.37	6.60	1.93	0.50	3.89
	RQ-0.25	−2.60	−0.20	−3.74	−1.61	−0.27	−2.17
	RQ-0.75	2.34	0.17	2.86	0.32	0.23	1.72
	RPIQ	3.74	6.44	2.17	0.73	2.02	2.91
DRF	R^2	0.85	0.92	0.79	0.96	0.78	0.83
	RMSE	5.02	0.39	8.52	26.35	0.58	3.94
	RIQR	4.81	0.36	6.44	2.06	0.49	3.87
	RQ-0.25	−2.56	−0.20	−3.79	−1.41	−0.28	−2.19
	RQ-0.75	0.25	0.16	2.64	0.65	0.22	1.68
	RPIQ	3.77	6.47	2.19	0.77	2.03	2.92

Relatively high RPIQ values were recorded for the prediction of clay content (RPIQ = 3.77), pH (RPIQ = 6.47), and CEC (RPIQ = 2.92), whereas that SOC model had a relatively low value (RPIQ = 2.19), likewise for N (RPIQ = 2.03). For $CaCO_3$ content, this statistic was at its lowest (RPIQ = 0.77). Statistics for the remaining models (i.e., using validation data) indicated a very strong concentration of values in their distributions, as evinced by their corresponding RIQR (interquartile ranges of the residuals) and kurtosis parameters. Kurtosis, a statistic characterized by high inherent variance [30], is not considered a significant indicator of a distribution's shape, but in this case, it concerns sets of equal numbers and similar forms of distributions (Tab. 5).

Table 5. Kurtosis of distributions of residual values of soil feature prediction stacked models. 0-level model: DRF

1-level model	Clay	pH	SOC	$CaCO_3$	N	CEC
MLP	12.2	7.7	15.1	45.1	13.5	16.8
M5	16.8	53.7	17.3	66.9	11.4	14.6
Cubist	12.1	7.6	14.1	37.3	11.8	12.9
GBM	9.4	6.9	14.2	36.8	12.3	11.1
DRF	9.0	6.6	13.6	36.0	11.5	11.1

It should be emphasized that a high kurtosis value for a residual distribution, with similar values of RMSE and R^2, indirectly shows it has more extreme spacing. Table 6 lists the ranks according to Spearman's *rho* and Kendall's *tau* between the values of predictions from applied models and observations. This revealed a significant difference between ranks statistics in comparison with the determination coefficients of some models, especially those used to estimate the $CaCO_3$ concentration in soil. The very high R^2-values are mainly due to very high variance of modeled values' distribution, which may be of lesser importance for non-parametric statistical tests.

Table 6. Spearman's *rho* and Kendall's *tau* ranked correlation coefficients between the results for estimated variables from the applied model incorporating a DRF module and the observed validation values

Statistic	Clay	pH	SOC	$CaCO_3$	N	CEC
Spearman's *rho*	0.92	0.96	0.90	0.78	0.88	0.91
Kendall's *tau*	0.78	0.82	0.73	0.66	0.72	0.75

An important criterion by which to evaluate a model is the relation of its estimation error to actual values of given variables. A greater tolerance for errors associated with larger values of modeled variables may be justified, in certain cases, by practical aspects of model use. The graphs in Figure 9 show boxplots of the model's residuals in arbitrarily determined classes of values of the observed variables. Except for the pH model, the remaining distributions indicate an increase in residuals with higher values of observed variables. For smaller values of the variables, their estimation error is less than at larger real values. For example, when analyzing the $CaCO_3$ model validation residuals of 0–100 g/kg (0–10%), its 'local' determination coefficient was 0.66 and the 'local' RMSE value was 9.9 (0.99%), set against an RMSE for the whole variation range of ca. 27 (2.7%), while the local value of the inter-quartile spacing of residuals RIQR = 1.58. It is thus easy to see that higher residuals values at higher variable values are due to non-linearity in the relationship between modeled and observed values.

Selecting a fixed validation set, especially when there is high variability of input values, entails a certain weakening of the model, associated with omitting a certain portion of input data's variability constituting the fixed validation set. The procedure of model implementation, after determining its suboptimal architecture, was based here on the use of all available data, supported by cross-validation statistics, which incorporates all available data in subsequent optimization cycles (i.e., *k*-folds validation). Statistics of the final models – their training set error statistics and cross-validation averages – developed in this way are presented in Table 7. From these Figure 10 follows that a prediction algorithm trained on a full set of data, due to its flexibility, would provide much better forecasting results. This makes it difficult to differentiate data and discern the influence of non-observable factors interfering with model outputs by modifying the spectral response. It should be stressed, however, that all algorithms and data pre-processing methods used here provide predictions having similar statistical characteristics.

Table 7. Statistics of stacked regression models with DRF decision module, optimized for the whole data set: 5-fold validation and cross-validation (CV) ratios, RMSE values and RMSE ratios of cross-validation and training data estimates

Properties	Training set				Cross-validation data		
	R (T)^2	RMSE (T)	SLOPE	CONST	R (CV)^2	RMSE (CV)	RMSE (CV)/RMSE (T)
Clay	0.98	1.75	1.03	−0.6	0.88	4.4	2.51
pH	0.99	0.14	1.02	−0.1	0.93	0.36	2.57
SOC	0.98	2.93	1.04	−1.1	0.85	7.5	2.56
$CaCO_3$	0.99	7.5	1.00	−0.4	0.98	19.6	2.61
N	0.97	0.20	1.04	−0.1	0.83	0.5	2.50
CEC	0.98	1.44	1.04	−0.5	0.85	3.7	2.57

RMSE (T) – root mean squared error stacked regression model with DRF decision module for training data in the *k*-folds cross-validation procedure, RMSE (V) – root mean squared error stacked regression model with DRF decision module for validation data in the *k*-folds cross-validation procedure, in the last column relations RMSE (V)/RMSE (T).

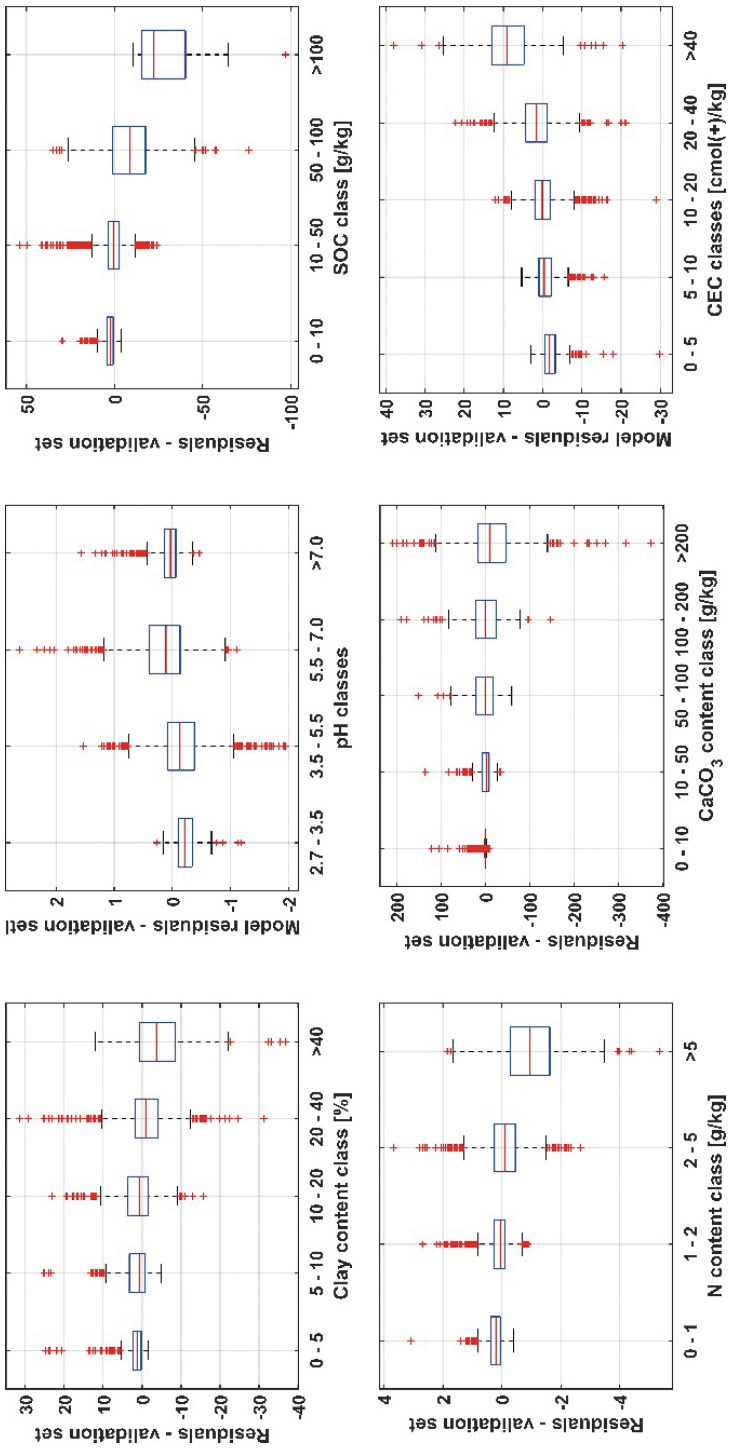

Fig. 9. Boxplots of the validation residuals' distributions in subclasses of the observed values of six soil variables

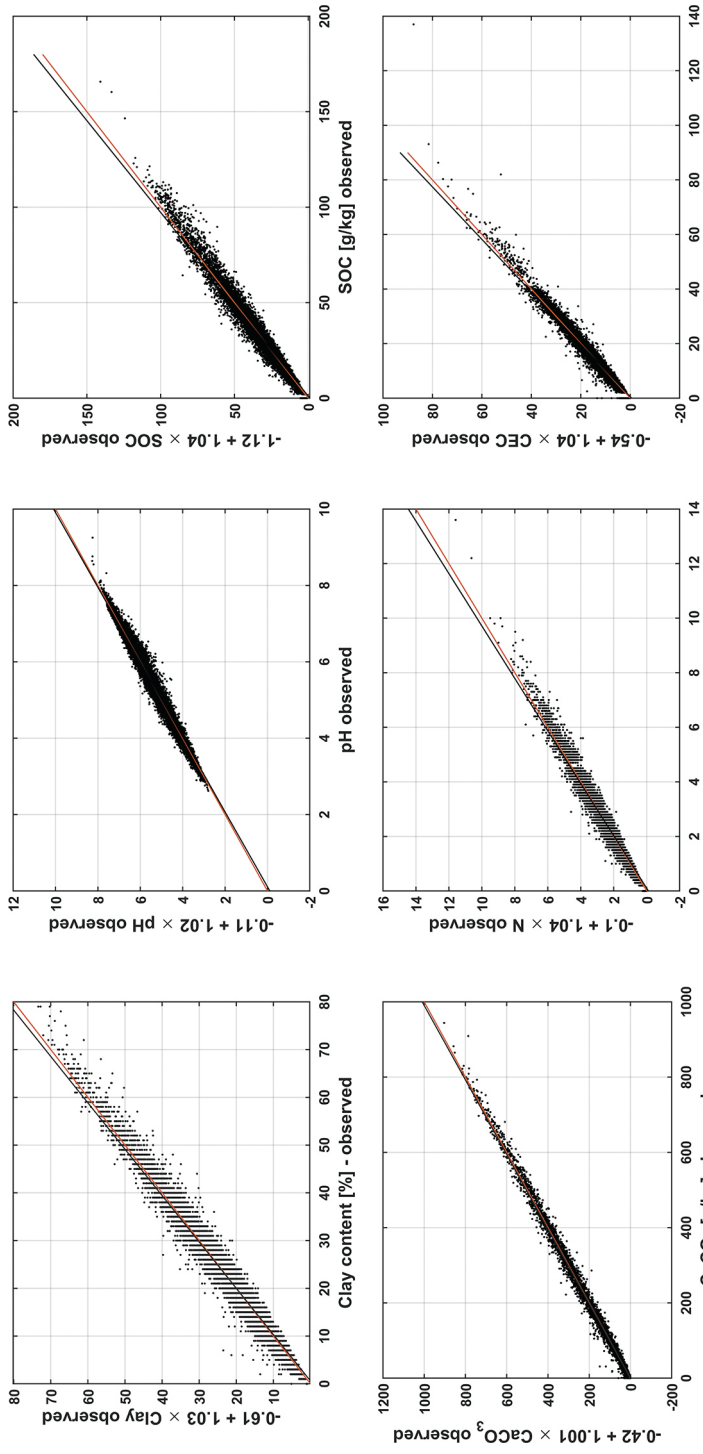

Fig. 10. Scatterplots of linear relationships between the observed values of soil variables and those forecasted, for models based on the whole data set: stacked regression with DRF as 1-level model

3.5. Convolutional Neural Network (CNN)

The results of CNN modeling were comparable to predictions made using the previously described stack models with different 1-level models. Prediction statistics for training and validation data, with two versions of 1D input data, namely the reflectance and absorbance vectors transformed in the SNV procedure [31] are given in Table 8.

Table 8. Statistics for the CNN model validation

Input	Statistic	Clay	pH	SOC	$CaCO_3$	N	CEC
Reflectance	R^2	0.73	0.79	0.68	0.84	0.61	0.62
	RMSE	6.66	0.62	10.91	50.39	0.76	5.95
SNV(Abs)	Epochs	400	1400	1400	600	1400	1600
	R^2	0.82	0.90	0.77	0.95	0.76	0.77
	RMSE	5.46	0.42	9.39	31.78	0.62	4.58

Epochs are the number of training epochs after which the result was obtained. SNV(Abs) – Standard Normal Variate transformation of absorbance vectors.

Tests may use an architecture similar to that described in [10], with specific modifications regarding the number of processing layers, optimization algorithm, validation method, and the number of examples that constitute a single "batch" of network training.

In this testing, relatively worse results were obtained with inputs in the form of reflectance vectors than with the transformation of absorbance vectors by the SNV algorithm. The use of models with multiple outputs inevitably complicates the problem of completing their optimization. According to Table 8 the optima – lowest values of validation errors of particular soil features – occurred in different optimization periods (from 400 to 1,600 epochs).

Continuing the optimization, especially with smaller 'batch' values, the number of examples taken into account at the same time in a single optimization step, resulted in an over-adjustment of the model: its systematic correction of training errors with an unchanged validation error.

One criterion of robust multidimensional models is the maintenance of quantitative relationships between the modeled variables.

Table 9 presents the Pearson linear correlation coefficients calculated for soil data as determined by reference methods and obtained from the predictions for validation data.

Table 9. Comparison of Pearson's linear correlation coefficients (r) of soil variables observed (validation set) to the forecasted data

Properties	Data	pH	SOC	$CaCO_3$	N	CEC
Clay	Obs	0.52	0.02	0.31	0.27	0.75
	SR	0.58	−0.07	0.32	0.20	0.78
	PCN	0.62	−0.08	0.34	0.23	0.86
pH	Obs	–	−0.36	0.51	−0.08	0.57
	SR	–	−0.40	0.53	−0.09	0.63
	PCN	–	−0.46	0.58	−0.12	0.69
SOC	Obs	–	–	−0.11	0.83	0.19
	SR	–	–	−0.13	0.82	0.11
	PCN	–	–	−0.18	0.86	0.07
$CaCO_3$	Obs	–	–	–	−0.06	0.26
	SR	–	–	–	−0.09	0.27
	PCN	–	–	–	−0.12	0.32
N	Obs	–	–	–	–	0.43
	SR	–	–	–	–	0.38
	PCN	–	–	–	–	0.39

The coefficients observed in the population and those calculated for predictions between soil variables are generally similar in term of the rankings of their values.

4. Discussion and Conclusion

The idea of using the spectral response in the Vis-NIR range as a source of information on soil properties has an extensive literature, with varied conclusions having been drawn on this approach's suitability and utility for characterizing soil states. The digitization of soil environment documentation places high demands on the number of soil properties determinations. Cheap and quick methods of obtaining reliable soil data are of great importance. For many years, indirect remote sensing tools have been used to obtain information about the state of the components of the environment. The use of indirect methods in soil research in the laboratory cannot be underestimated due to the importance of determining the characteristics of the soil profile, impossible from the aerial or satellite level.

A literature review on the possibilities and limitations of this methodology [4, 6, 32] cited the extreme values of statistics (R^2 and RPD) of 'spectral response-soil property' relationship models. According to that investigation, in studies by various

authors, models' determination coefficients were obtained that ranged from 0.01 to 0.99 for clay fraction content, from 0.66 to 0.96 for SOC content, from 0.68 to 0.98 for N content, from 0.22 to 0.88 for pH, and from 0.07 to 0.93 for CEC. Similarly, large differences were found between their model's RPD values. Practically then, the models can range being from completely useless to close to ideal when it comes to relying on data-based algorithms. The reasons for this great variation are obvious: namely, the number of samples and variation in soil properties, the methodology for extracting useful information from the spectral data, the model design, the validation method used, among other objective and subjective factors. Such studies typically use numerous transformations of spectral response vectors, differing dimensional reduction algorithms, and different inference models and thus it would be difficult to uncover a regression algorithm that would be omitted by them. Additionally, using concepts of so-called "auxiliary variables" is often presented – such as the type of land use, geographical location, soil texture – which, in addition to the spectral response, may be used to improve the prediction of other characteristics. In sum, lacking a dominant and widely recognized algorithm of inference based on the spectral response of soils, different concepts of soil formation ostensibly must compete with each other.

The basic modeling methodology that combines dimensional reduction and extraction of spectral response fragments relevant for modeling is dimensional reduction by PCA and PLSR [5, 32], especially for small datasets which, due to their low soil variability, provide acceptable values for prediction errors. For years, the results of various machine learning methods have been published in combination with using such statistical methods (PCA or PLSR) for extracting input variables. This concept is perhaps best known as the PARACUDA II procedure [9, 33], which combines multiple sampling of a data set, dimensional reduction, and sets of prediction models. This framework is characterized by a high quality of prediction (RMSE: 0.17% in SOC prediction, 5.4 cmol/kg CEC, 5.8% for $CaCO_3$ content), although it applies to relatively small sized data sets (sample size of ca. 100).

For ca. 30 years, deep processing has been the focus of interest in image analysis, offering an alternative to so-called "shallow" models. The use of CNN models for regression and classification tasks is now becoming common. Among other things, the problem of the number of training sets necessary to obtain useful results using CNNs, which usually entail large processing structures, is being resolved. According to [10], the CNN model can improve prediction quality with an increase in the number of learning examples, while for shallow models, an increase in their number of learning sets beyond a certain limit is not accompanied by improved predictions. According to that study [10], a set of ca. 8000 examples (data from Brazil) used to optimize the CNN network gives better forecasting results than either the PLSR or Cubist algorithm. Specifically, the RMSE value of clay fraction content validation data was 6.7% for the CNN model (for the PLSR and Cubist: 7.3% and 6.9%, respectively), the CEC value for CNN was 1.3 cmol/kg (for PLSR and Cubist: 1.7 cmol/kg),

and organic matter content for CNN was 3.8 g/kg (for PLSR and Cubist: 5.0 g/kg and 4.8 g/kg, respectively). These values, however, cannot be compared with estimates of errors for other data, since they depend on the distribution and variability of the modeled soil variables.

Since its initial release, the LUCAS database has been widely used to develop various models to predict soil properties. One of the first studies using it was a publication on soil organic carbon modeling [34], in which the SVM model (RMSE = 8.9 g/kg) gave the best result for mineral soils (no use distinguished). In the models developed for particular types of land use, the results were not alike: for arable land, grassland, and forest areas the RMSE values obtained were respectively 4.9 g/kg, 9.3 g/kg and 15.0 g/kg (SVM and Cubist models); the addition of an auxiliary variable to the input variables (content of sandy or clay fraction in the sample) significantly improved their prediction. In the case of mineral soils (again, without distinction of use), this addition reduced the RMSE to 7.3 g/kg. In another study [35], LUCAS data were analyzed for modeling based on spectral response, SOC, N and clay fraction content. The modeling algorithm included the extraction of variables via PLSR (the number of factors considered experimentally ranged from 42 to 78). The corresponding RMSE values for the SOC model, for arable land, permanent grassland, and forests were respectively 6.0 g/kg, 10.9 g/kg, and 13.8 g/kg. Likewise, the clay fraction of those habitats was modeled with RMSE (in the same order) of 5.5%, 6.2%, and 5.4%, with corresponding values of 0.42 g/kg, 0.82 g/kg and 0.74 g/kg for nitrogen content.

Advanced models of machine learning (an MLP network, Boltzmann's restricted machine, and CNN) were used by [35–37], in their modeling based on the LUCAS soil database. Those modeling results are difficult to interpret as the authors provided RMSE values for standardized variables. For five modeled variables (sandy fraction content, pH, SOC, $CaCO_3$, and P content), the mean RMSE value was 0.42 (CNN and a combination of Boltzmann's machine and the convolutional network), which corresponds to an RPD = 2.36. When recalculated to comply with LUCAS data (as declared by the authors), the RMSE values for a hybrid model (limited Boltzmann machine and convolutional network) were 7.4 g/kg for SOC, 0.41 for pH, and 25.2 for $CaCO_3$ content.

Work by Liu et al. [36] reported on the practical use of a pre-trained CNN network's properties. The construction of this network, to model the clay fraction content on the basis of its spectral response, was also based on LUCAS data for mineral soils. The trained network was first used to predict the clay fraction content in organic samples, after refining it on the basis of a small number of organic samples (RMSE = 7.07%). This was then used to estimate the clay content of soils base on multispectral imaging, after fine tuning the pre-treated network on a small number of samples (RMSE = 8.62).

Spectral data and spectral data in combination with the auxiliary predicators of the LUCAS collection were also used in the advanced three-level MKL (Multiple

Kernel Learning) model for SOC prediction [38]. The prediction variable was logSOC, the decimal logarithm of the value (SOC+1), for which the $R^2 = 0.86$ and RMSE = 0.13. Much smaller error characterized the model when an auxiliary variable was incorporated (RMSE = 0.1). It should be noted that, in the current study, the conversion of RMSE value for the stack model with DRF output (Tab. 7) to the logSOC scale indicates that it is ca. 0.115, which is slightly less than that (without the auxiliary variable) reported by Tsakiridis et al. [38].

Recently, Tsakiridis et al. [39] published the results of their research on the usefulness of 1D multi-channel 1D CNN's for the prediction of clay fraction content, SOC and N content based on the Vis-NIR spectral response using LUCAS database. The results presented in this study (RMSE for clay content about 4.8%, SOC about 10.96 and N about 0.66) are similar to those presented in our study. This applies to CNN models, while the stacked regression model has even slightly better characteristics.

Against the background of the results available in the literature, the applied regression models presented in the paper, in relation to the LUCAS database, can be considered comparable with the best prediction algorithms. The square roots mean square error estimation in the cross-validation procedure were as follows: for clay content, ca. 4.4%; for pH, ca. 0.36; for SOC, ca. 7.5 g/kg (0.75% by weight); for $CaCO_3$ content, ca. 19 g/kg; for N content, ca. 0.50 g/kg; and for CEC, ca. 3.7 cmol(+)/kg. These results do not compete with laboratory tests, but are sufficiently accurate to dense point observations of soil condition for digital mapping purposes. Unsatisfactory low $CaCO_3$ prediction accuracy. It should be noted, however, that the prediction error of this component is positively correlated with its content in the soil and is relatively small at its low concentration. In Polish soils (except for rare cases of soils rich in Ca), exceeding of $CaCO_3$ concentration 20–30 g/kg is observed in limited cases.

The LUCAS database contains soil sample data from areas with high geological, climatic, and land use diversity. In the testing with statistical models, machine learning models (MLP, trees and random forests) with and without data grouping, applied models and deep learning models, relatively good results were obtained when using applied models coupled to PLS inputs processed with MLP algorithms. These modeling results are consistent with those reported in the literature, provided that the results from different areas are truly comparable. Furthermore, the poor results obtained here for the effects of stratification of spectral data were probably caused by methodology, specifically establishing similarity on the principle of Euclidean distance of vectors, which was significantly influenced by fragments not related to the modeled soil variables.

It should be assumed that the practical use of the Vis-NIR spectral response in prediction of soil properties will be dominated by data-based models. The literature reports present various algorithms of model construction with satisfactory prediction quality. In cases of large and diverse databases, such as LUCAS, PLSR, linear models using PCA or linear step regression do not give fully satisfactory

results. Large datasets require more sophisticated methods to identify correlations between the spectral response and soil variables, such as PARACUDA II, the three--level MKL, or deep models for example the convolution 1D network, which did not give the best results in the tests. The applied stacking model gives good prediction results, probably partly due to the presence of many soil variables estimations in the input vector. This can be considered a disadvantage if the aim is to predict one or two soil variables, however, the approach is relatively flexible and the choice of the decision module needs testing. This approach, requires a large training dataset, which, also applies to the convolutional model. The advantage of the convolution network lies in the possibility to develop a pre-trained model based on a large and diverse data set. The CNN model can be "fine-tuned" using a relatively small set of local conditions.

The comparison of the results of modeling soil characteristics from different regions is difficult due to the differences in the variability of the mineralogy, soil types, sample size and other factors influencing the spectral response and the range of data variability. The similarity of physiographic conditions is likely to have a positive effect on the prediction error due to the homogeneity of potential disturbances in the relationship between the examined features and the reflectance of the samples. In addition, the machine learning algorithms present in modeling are characterized by a great flexibility in fixing optimization parameters. This is a factor that forces the randomness of the prediction result and the uncertainty as to the best possible architecture. For this reason, it is impossible to indicate the objectively optimal prediction model.

The conducted tests lead to the following conclusions:

1. The use of machine learning models reduces the error of the prediction of soil features based on the analysis of the spectral response in the Vis-NIR range in relation to the statistical linear models: stepwise regression or partial least squares regression.
2. The relatively best prediction result was obtained using a stack regression model with 0-level data obtained from many MLP models.
3. Clustering of the input data, due to the presence of variables not related to the modeled dependencies, does not improve the prediction results.
4. The influence of factors disturbing the "Vis-NIR – soil features" relationship is visible in the attempt to use the stack regression algorithm with cross--validation: the RMSE of the model built on the basis of the entire set is much lower than the average RMSE values from the validation.
5. The prediction errors of some features (Clay, SOC, $CaCO_3$, N and CEC) increase monotonically according to their true value.
6. The use of a single prediction algorithm for data from very diverse geological and soil conditions has significant limitations as a method that replaces traditional laboratory analysis. This does not rule out its usefulness as a data source in soil cartography.

Acknowledgements

The LUCAS topsoil data set used in this work was made available by the European Commission, through the European Soil Data Centre managed by the Joint Research Centre (JRC), http://esdac.jrc.ec.europa.eu/.

References

[1] McBratney A., Minasny B., Viscarra Rossel R.: *Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis*. Geoderma, vol. 136, 2006, pp. 272–278. https://doi.org/10.1016/j.geoderma.2006.03.051.

[2] Al-Abbas A.H., Swain P.H., Baumgardner M.F.: *Relating Organic Matter and Clay Content to the Multispectral Radiance of Soils*. Soil Science, vol. 114, 1972, pp. 477–485.

[3] Kokaly R.F., Clark R.N., Swayze G.A., Livo K.E., Hoefen T.M., Pearson N.C., Wise R.A., Benzel W.M., Lowers H.A., Driscoll R.L., Klein A.J.: *USGS spectral library version 7.* Data Series 1035, U.S. Geological Survey, U.S. Department of the Interior, 2017. https://doi.org/10.3133/ds1035.

[4] Stenberg B., Viscara Rossel R.A., Mounem Mouazen A., Wetterlind J.: *Visible and near infrared spectroscopy in soil science*. [in:] Sparks D.L. (ed.), *Advances in Agronomy*, vol. 107, Academic Press, Burlington 2010, pp. 163–215. http://doi.org/10.1016/S0065-2113(10)07005-7.

[5] Wetterlind J., Stenberg B., Viscarra Rossel R.A.: *Soil analysis using visible and near infrared spectroscopy*. [in:] Maathuis F.J.M. (ed.), *Plant Mineral Nutrients: Methods and Protocols*, Methods in Molecular Biology, vol. 953, Humana Press, Springer, New York 2013, pp. 95–107.

[6] Masso C., Ziadi N., Parent L., Tremblay G., Thuries L.: *Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: A review*. Canadian Journal of Soil Science, vol. 89, 2009, pp. 531–541. https://doi.org/10.4141/CJSS08076.

[7] Wetterlind J., Stenberg B.: *Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples*. European Journal of Soil Science, vol. 61, 2010, pp. 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.

[8] Shi Z., Ji W., Viscarra Rossel R.A., Chen S., Zhou Y.: *Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese Vis-NIR spectral library*. European Journal of Soil Science, vol. 66, 2015, pp. 679–687. https://doi.org/10.1111/ejss.12272.

[9] Gholizadeh A., Saberioon M., Carmon N., Boruvka L., Ben-Dor E.: *Examining the performance of PARACUDA-II data-mining engine versus selected techniques to model soil carbon from reflectance spectra*. Remote Sensing, vol. 10(8), 2018, 1172. https://doi.org/10.3390/rs10081172.

[10] Ng W., Minasny B., Montazerolghaem M., Padarian J., Ferguson R., Bailey S., McBratney A.B.: *Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra*. Geoderma, vol. 352, 2019, pp. 251–267.

[11] Tóth G., Jones A., Montanarella L. (eds.): *LUCAS topsoil survey. Methodology, data and results*. JRC Technical Reports, Publications Office of the European Union, Luxembourg 2013. https://doi.org/10.2788/97922.

[12] Orgiazzi A., Ballabio C., Panagos P., Jones A., Fernández-Ugalde O.: *LUCAS Soil, the largest expandable soil dataset for Europe: a review*. European Journal of Soil Science, vol. 69, 2017, pp. 140–153.

[13] Iordache M.-D.: *Matlab code and demo for continuum removal*. 2016. https://doi.org/10.13140/RG.2.1.2885.9285.

[14] *MATLAB., version 9.7.0.1190202 (R2019b)*. The MathWorks Inc., Natick, Massechusetts 2019.

[15] Kohonen T.: *Self-organized formation of topologically correct feature maps*. Biological Cybernetics, vol. 43, 1982, pp. 59–69. https://doi.org/10.1007/BF00337288.

[16] Quinlan J.R.: *Learning with continuous classes*. [in:] Adams A., Sterling L. (eds.), *Ai '92 – Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, World Scientific, 1992, pp. 343–348.

[17] Kuhn M., Quinlan R.: *Cubist: Rule- and Instance-Based Regression Modeling*. R package version 0.2.2. 2018. https://CRAN.R-project.org/package=Cubist [access: 31.01.2020].

[18] Wolpert D.: *Stacked generalization*. Neural Networks, vol. 5(2), 1992, pp. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

[19] Breiman L.: *Stacked regressions*. Machine Learning, vol. 24, 1996, pp. 49–64. https://doi.org/10.1023/A:1018046112532.

[20] Kursa M.B., Rudnicki W.R.: *Feature selection with the Boruta package*. Journal of Statistical Software, vol. 36, 2010, pp. 1–13. https://doi.org/10.18637/jss.v036.i11.

[21] Frank E., Hall M.A., Witten I.H.: *The WEKA Workbench*. Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.

[22] R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria 2019. https://www.R-project.org/ [access: 29.05.2020].

[23] Cook D.: *Practical Machine Learning with H_2O. Powerful, Scalable Techniques for Deep Learning and AI*. O'Reilly Media, 2016.

[24] Geurts P., Ernst D., Wehenkel L.: *Extremely randomized trees*. Machine Learning, vol. 63, 2006, pp. 3–42. https://doi.org/10.1007/s10994-006-6226-1.

[25] Friedman J.H.: *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, vol. 29(5), 2001, pp. 1189–1232.

[26] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp H., Irving G., Isard M., Jozefowicz R., Jia Y., Kaiser L., Kudlur M., Levenberg J., Mané D., Schuster M., Monga R., Moore S., Murray D., Olah C., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viégas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X.: *TensorFlow: Large-scale machine learning on heterogeneous systems*. 2015. https://www.tensorflow.org [access: 29.05.2020].

[27] Chollet F.: *Keras*. 2015. https://keras.io [access: 29.05.2020].

[28] Allaire J.J., Chollet F.: *Keras: R Interface to 'Keras'. R package version 2.2.5.0*. 2019. https://cran.r-project.org/web/packages/keras/index.html [access: 29.05.2020].

[29] Allaire J.J., Tang Y.: *Tensoorflow: Interface to 'TensorFlow'. R package version 2.0.0*. 2019. https://CRAN.R-project.org/package=tensorflow.

[30] Westfall P.H.: *Kurtosis as peakedness, 1905–2014. R.I.P*. The American Statistician, vol. 68(3), 2014, pp. 191–195. https://doi.org/10.1080/00031305.2014.917055.

[31] Rinnan Å., Berg F.V., Engelsen S.B.: *Review of the most common pre-processing techniques for near-infrared spectra*. Trends in Analytical Chemistry, vol. 28, 2009, pp. 1201–1222.

[32] Dunn B., Batten G., Beecher H.G., Ciavarella S.: *The Potential of near-infrared reflectance specytroscopy for soil analysis: a case study from the Riverine Plain of south-eastern Australia*. Animal Production Science, vol. 42, no. 5, 2002, pp. 607–614. https://doi.org/10.1071/EA01172.

[33] Gholizadeh A., Carmon N., Klement A., Ben-Dor E., Borůvka L.: *Agricultural soil spectral response and properties assessment: effects of measurement protocol and data mining technique*. Remote Sensing, vol. 9, 2017, 1078.

[34] Stevens A., Nocita M., Toth G., Montanarella L., Van Wesemael B.: *Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy*. PLoS ONE, vol. 8, 2013, e66409. https://doi.org/10.1371/journal.pone.0066409.

[35] Liu L., Ji M., Buchroithner M.: *Combining partial least squares and the gradient--boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra*. Remote Sensing, vol. 9, 2017, 1299.

[36] Liu L., Ji M., Buchroithner M.: *Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery*. Sensors, vol. 18, 2018, 3169.

[37] Veres M., Lacey G., Taylor G.W.: *Deep Learning Architectures for Soil Property Prediction*. [in:] *CRV 2015: 12th Conference on Computer and Robot Vision: Proceedings: 3–5 June 2015, Halifax, Nova Scotia, Canada*, IEEE, 2015, pp. 8–15. https://ieeexplore.ieee.org/document/7158315.

[38] Tsakiridis N.L., Chadoulos C.G., Theocharis J.B., Ben-Dor E., Zalidis G.C.: *A three-level multiple-kernel learning approach for soil spectral analysis*. Neurocomputing, vol. 389, 2020, pp. 27–41. https://doi.org/10.1016/j.neucom.2020.01.008.

[39] Tsakiridis N.L., Keramaris K.D., Theocharis J.B., Zalidis B.C.: *Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi--channel 1-D convolutional neural network*. Geoderma, vol. 367, 2020, 114208. https://doi.org/10.1016/j.geoderma.2020.114208.