Assessing the Information Potential of MIR Spectral Signatures for Prediction of Multiple Soil Properties Based on Data from the AfSIS Phase I Project

Stanisław Gruszczyński and Wojciech Gruszczyński

AGH University of Science and Technology, Faculty of Geo-Data Science, Geodesy and Environmental Engineering, Al. Mickiewicza 30, 30-059 Krakow, Poland
* Correspondence: wgrusz@agh.edu.pl

Abstract: The aim of the study was to assess the predictive potential of mid-infrared (MIR) spectral response in the estimation of 60 soil properties. It is important to know the accuracy limitations in estimating various soil characteristics using various models in conditions of high spatial variability of the environment. To fully assess this potential, three types of algorithms were used in modeling, i.e., partial least squares (PLSR), one-dimensional convolutional neural network (1DCCNN), and generalized regression neural network (GRNN). The research used data from 19 sub-Saharan African countries collected as part of the Africa Soil Information Service (AfSIS) Phase I project. The repositories provide 18,250 MIR reflectance recordings and nearly two thousand analytical data records from the determination of many soil properties by reference methods. The modeled subset of these properties included texture (three variables), bulk density, moisture content at soil water characteristic curves (SWCC, 4 variables), total and organic C and total N content (3 variables), total elemental content (32 variables), elemental content in bioavailable forms (12 variables), electrical conductivity, exchangeable acidity, exchangeable bases, pH, and phosphorus sorption index. It is not possible to indicate a universal optimal prediction model for all soil variables. The best prediction results are provided by all regression models for total and organic C, total Fe, total Al and bioavailable Al content, and pH. For bulk density, total N and total K content satisfactory results are provided by specific model type. Many other properties, i.e., texture, SWCC, total Ga, Rb, Na, Ca, Cu, Pb, Hg content, and bioavailable Ca and K content, can be predicted with accuracies sufficient for some less demanding tasks.

Keywords: mid-infrared spectrum; soil properties prediction; partial least squares regression; 1D convolutional neural network; generalized regression neural network

1. Introduction

The interpretation of short-range spectral recordings under field or laboratory conditions attracts a lot of attention [1–4]. The importance of such methods is related to the three-dimensional variability of soils and their profile variation, which are impossible to recognize by recording spectral reflectance of the ground surface. Most studies to date relating to spectral analysis of soils have used the visible near-infrared (Vis-NIR) range covering the 350–2500 nm range of electromagnetic radiation [3,5,6]. Currently, there is increasing interest in investigating the suitability of the mid-infrared (MIR) spectrum covering the 2500–25,000 nm range [7–9].

The analysis of the usefulness of MIR signatures in soil research is of increasing interest. Margenot et al. analyzed the technical aspects of using this spectral range in the estimation of soil properties, mainly on the basis of a literature review [10]. They indicate the need to standardize the method of considering the impact of soil moisture on disturbances in the estimation of soil organic matter (SOM). Nath et al. analyze the usefulness of MIR signatures in estimating biological properties [11], pointing to the high usefulness of this
approach in determining the concentration of C and N. They believe that the estimation of the concentration of enzymes and other organic components requires more sophisticated estimation methods. In a study of alfisols in eastern India [12], using a relatively small (approximately 330 samples) sample set, the authors concluded that the PLSR and SVM model provide a better estimate of SOC, pH, and clay than the random forest model. They point to the need to strengthen these conclusions on the basis of more territorially extensive research. An extensive analysis of the estimates of over 100 soil properties, based on the MIR Kellogg Laboratory’s spectral library, is contained in the publication by Ng et al. [8]. The results obtained confirmed the usefulness of MBL models for the estimation of many properties of soils from the territory of the USA.

A very insightful analysis of prediction results using NIR and MIR was conducted by Bellon-Maurel and McBratney [5]. The review analyzed the effect of the method of measurement implementation, data preprocessing, and algorithms used on soil organic carbon (SOC) prediction. According to one of the review’s conclusions, NIR and MIR allow the prediction of SOC with similar accuracy, with a slight advantage to MIR as it provides better reproducibility, with a lesser error (10–40% when compared to NIR). In addition, it was found that, regardless of the range of the spectrum, homogeneity of data for calibration and validation is essential. It should be noted that comparison of modeling results based on spectral response is justified only for soil samples with similar properties, geology, and climatic conditions. The importance of the range of variability of the data, the number of cases for calibration, preprocessing, and the type of model should also be emphasized.

In order to use spectral analysis in the soil survey, it is necessary to know the potential possibilities and limitations of this approach in estimating the values of various soil characteristics. The problem is complex. The model error, apart from its architecture, depends on the distribution and differentiation of data, the range of their variability, the number of observations, the method of validation, etc. The work of Nduwamungu et al. [13] summarizes the results of modeling tests based on the NIR spectrum of soil properties such as texture, SOC, pH, and cation exchange capacity (CEC) presented in the work of many authors, whose statistics differed dramatically: the $R^2$ spread was between 0.01 and 0.99. The availability of possibly extensive data with high variability increases the credibility of the information potential of spectral data under various soil conditions. The knowledge of the potential usefulness of a spectrum for determining various soil properties is of significant practical importance for the selection of variables used to estimate soil quality indices [14,15].

Soils are described by a number of properties subject to evaluation from the point of view of their cultivation, economic, and engineering suitability. As evaluation criteria, depending on the use under consideration, various properties of soils are treated as more or less important. Classification of soils, in the context of their use in agriculture or forestry, requires characterization of basic physical (texture, water holding capacity, compactness, permeability, and many other potential physical indicators) and chemical (concentration of macro- and micronutrients, pH, sorption capacity, buffering capacity, and many others) properties. The list of elements to be assessed is constantly growing; for example, by including rare earth elements in the assessment of soils [16,17]—important because of their role as micronutrients and potential pollutants. Given this fact, it is important to appreciate the possibility of assessing the information potential contained in the spectral response of soil samples as a source to support, at least roughly, the assessment of environmentally or economically significant properties. It is important to note that the quality of models linking the spectral response to individual soil properties depends largely on the type of prediction algorithm used as well as the statistical distribution of the modeled properties. For this reason, it is advantageous to confront modeling approaches and validation datasets in order to identify features, prediction methods, and conditions for their use.

The aim of this research is to determine prediction accuracy for many soil variables based on the analysis of the spectral response of soils in the proximal sensing mode in the MIR range. The literature generally presents attempts to model several variables—usually
The aim of this research is to determine prediction accuracy for many soil variables from samples taken in highly varied soil and climatic conditions. Determined validation statistics allow for the assessment of the usefulness and limitations of MIR models and data for estimating the values of variables characterizing soils.

2. Materials and Methods

2.1. AfSIS Phase I Data

AfSIS Phase I data were collected as part of a project financed by the Melinda and Bill Gates Foundation, implemented with the participation of five research institutions: The Tropical Soil Biology and Fertility Institute of The International Center for Tropical Agriculture, The International Soil Reference and Information Center—World Soil Information, The Center for International Earth Science Information Network, The Earth Institute at Columbia University, and World Agroforestry (Center for Research in Agroforestry—ICRAF). These data include the results of analyses of samples from soil profiles collected at points determined according to the methodology of the Land Degradation Surveillance Framework [20] in 19 countries of sub-Saharan Africa: Angola, Botswana, Burkina Faso, Cameroon, Ethiopia, Ghana, Guinea, Kenya, Madagascar, Malawi, Mali, Mozambique, Niger, Nigeria, South Africa, Tanzania, Uganda, Zambia, and Zimbabwe. More details regarding the AfSIS database can be found in Vågen et al. [21].

The data provided by the repositories (Amazon AWS, GitHub, and the Agroforestry portal) came from topsoil (0–20 cm) and subsoil (20–50 cm) layers. Soil sampling was carried out in accordance with the Land Degradation Surveillance Framework (LDSF) procedure, taking into account the local and spatial characteristics of the environment. The sampling areas represent a random stratification of the African landscapes south of the Sahara desert (Figure 1). The samples were taken in the areas surrounding the cluster center, which allowed consideration of the spatial variability of soils [22]. A detailed method for determining the location of sampling areas, sampling, and analytical procedures is included in the report written by Leenaars et al. [23].

AfSIS data were the basis for studies on the analysis of the spatial variability of soil chemical composition [24] and the characteristics of soil profiles [23]. On this foundation, cartographic documentation of soils with a resolution of 30 m was developed and...
prepared [25]. The repositories provide 18,250 MIR reflectance recordings [26,27]. A single spectral signature covers 1749 recorded reflectance values in the range from 2500 to 16,666 nm (wavelength 4000 to 600 cm$^{-1}$ with a step of 1.9 cm$^{-1}$). The average density of the spectral registration points was about 8.1 nm, and it ranged from 1.2 nm to 53 nm. The GitHub repository also provides nearly two thousand analytical data records from determinations of soil samples by reference methods. Georeferenced data on the location of sampling points are also available. The analytical data of the original dataset are split at the highest level into two groups, i.e., DryChemistry and WetChemistry. The DryChemistry subset includes the results of X-ray determinations of the content of 32 elements. The WetChemistry subset [28] includes data, which include soil texture, water characteristics of samples, and many soil chemical properties associated with the use of the Mehlich 3 methodology [29], from three different laboratories.

There are slight variations in the amount of data from different laboratories, which resulted in the need for separate modeling for individual parts of the set. Modeling of chemical data for which the reference data were less than 500 records were omitted. In addition, properties for which there are strict relationships allowing for their direct calculation on the basis of the modeled values were omitted in the modeling. Moreover, modeling of soil organic N was not carried out, as the method of determination of reference values was different from the commonly accepted standard. Each of the group of features for which modeling was performed included at least 1860 soil samples data, of which a randomly selected 70% was used for training and the remaining 30% was used for validation. The training and validation sets were kept the same for each model throughout the study.

A total of 60 variables were modeled, with the data divided into groups:

- “Texture” (3 variables, 1307 training, and 560 validation samples), separated into three particle-size fractions, i.e., sand, silt, or clay;
- “Water” (5 variables, 1306 training, and 560 validation samples), i.e., saturation at 0 (Sat. at 0), pH2.0, pH2.5, pH4.2, and bulk density (B. Den.);
- “CN” (3 variables, 1331 training, and 567 validation samples): total N, total C, and SOC;
- “Elements” (32 variables, 1333 training, and 570 validation samples), i.e., concentrations of: Na, Al, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, Ga, Se, Rb, Sr, Y, Zr, Ba, La, Ce, Pr, Nd, Sm, Hf, Ta, W, Hg, Pb, Bi, and Th;
- “Mehlich” (17 variables, 1333 training, and 571 validation samples), i.e., electrical conductivity (EC), exchangeable acidity (Ex. Ac.), exchangeable bases (Ex. Bas.), pH, phosphorus sorption index (PSI), and bioavailable forms of elements extracted with use of Mehlich 3 (M3) methodology, namely, M3 Al, M3 B, M3 Ca, M3 Cu, M3 Fe, M3 K, M3 Mg, M3 Mn, M3 Na, M3 P, M3 S, and M3 Zn.

For a description of the reference methods, in addition to standard operating procedures, see Vågen et al. [22]. The texture of the soil was determined with laser diffraction particle size analysis. For AfSIS reference samples, soil moisture release curves are determined using soil fines in pressure plate apparatus. High throughput total x-ray fluorescence spectroscopy (TXRF) was used in the analysis of the chemical composition (macro and microelements). Total and organic C and total N were determined with combustion method. Extractable Al, Ca, Mg, P, K, Na, S, Fe, Mn, Zn, Cu, B, Mo, H, and other bases were determined with use of ICP analysis of Mehlich 3 extracts. P sorption index was measured by single-point P addition.

Figure 2 and Appendix A (Tables A1–A5) present selected statistics for all variables in these data groups. For each data group, prediction models were created for all soil properties included in the group.
Figure 2. Box plots for modeled soil properties divided into separate data groups, i.e., Elements, Texture, Water, CN, and Mehlich. Blue boxes represent values between first and third quartile, red line represents median value and red crosses represent values outside the range of first quartile minus 1.5 times interquartile range and third quartile plus 1.5 times interquartile range.

2.2. Models

Three different groups of algorithms were used in this study: partial least squares regression (PLSR), one-dimensional convolutional neural network (1DCNN), and generalized regression neural network (GRNN).

The PLSR algorithm has been the most widely described and applied to prediction of the state of the environment from remotely sensed data. The deterministic PLSR algorithm [30,31] produces a linear model preceded by an orthogonalization operation of the input data taking into account their effect on the variance of the outputs. The experiments used a module available in the MATLAB package [32]. Depending on the number of model outputs, the number of PLS coefficients was set in range from 25 to 100. The single input vector consisted of 1749 reflectance values recorded for one soil sample.

A one-dimensional convolutional neural network [33] represents the so-called adaptive deep learning by processing the input data through successive layers of the algorithm. All conducted computational experiments used a network with two convolutional layers, counting 32 and 64 filters, or 64 and 128 filters with widths of three, six, and ten input features separated by MaxPooling layers. A fully connected layer consisted of 15 to 50 units. The number of outputs corresponded to the number of modeled properties. The spectra were standard normal variate (SNV)-transformed [34] and, like the outputs, were
normalized to the numerical interval $<0; 1>$ according to the methodology described in Gruszczyński and Gruszczyński [35]. The calculations associated with this algorithm were performed in the Python language environment using the Keras system and TensorFlow libraries [36,37].

A generalized regression neural network is a memory-based local nonparametric linear and nonlinear regression algorithm [38]. The GRNN algorithm is susceptible to the curse of dimensionality; therefore, the spectral data from the AfSIS project were transformed into vectors of 50 components each using the PCA algorithm, considering all available spectra (18,250). In order to optimize the spread of RBF units, a modeling cycle was performed with its values between 0.5 and 2.5 (with a step of 0.05). Figure 3 shows the dependence of linear correlation coefficients of data and prediction as a function of spread. In the construction of models, the spread generating the highest average linear correlation coefficient of the prediction with the observed ($r$) data was used. Optimal spread values ranged from 0.75 for the “Mehlich” data group to 2.0 for variables C and N. Calculations associated with the GRNN model were performed in the MATLAB environment [32].

![Figure 3. Dependence of the linear correlation coefficient between the reference and modeled values as a function of the spread value.](image)

The evaluation criteria for each model were the following validation statistics: coefficient of determination ($R^2$), root mean square error (RMSE), ratio of performance to interquartile distance (RPIQ), bias, standardized bias ($Sb$), and Lin's concordance correlation coefficient (LCCC).

The coefficient of determination was calculated according to the equation:

$$R^2 = 1 - \frac{SSR_v}{SST_v}$$  \hspace{1cm} (1)

where $SSR_v$ is a sum of squares of differences of modeled value and its prediction in the validation set, and $SST_v$ is a sum of squares of differences of modeled value and its mean in the validation set. The coefficient of determination is considered to be a numerical estimate of the degree to which the variance in a modeled variable is explained.

Root mean square error was calculated according to the equation:

$$RMSE = \sqrt{\frac{SSR_v}{n}}$$  \hspace{1cm} (2)

where $n$ is the size of the validation set.
Ratio of performance to interquartile distance was calculated according to the equation:

\[ RPIQ = \frac{IQR}{RMSE} \]  

(3)

\[ IQR = Q_3 - Q_1 \]  

(4)

where \( IQR \) is an interquartile distance, and \( Q_3 \) and \( Q_1 \) are third and first quartile of validation data.

The \( bias \) is calculated according to equation:

\[ bias = \frac{1}{n} \sum (x_i - y_i) \]  

(5)

where \( x_i \) and \( y_i \) are \( i \)th observed value and its prediction.

Standardized bias is calculated according to equation:

\[ Sb = \frac{bias}{IQR} \]  

(6)

Lin’s concordance correlation coefficient \( LCCC \) [39] is calculated according to equation:

\[ LCCC = \frac{(2\cdot r \cdot s_x \cdot s_y)}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \]  

(7)

where \( s_x, s_y \) are standard deviations of observed values and their predictions, and \( \bar{x} \) and \( \bar{y} \) are the average of the observed values and their predictions. The \( LCCC \) value indicates the degree of agreement between the observed data and their prediction.

2.3. Hierarchical Summary of Prediction Positions

The relatively large number of modeled variables and the use of three different modeling algorithms justify the need to hierarchize the prediction accuracy of the variables in order to compare their practical utility. The hierarchization procedure was implemented using the technique for order of preference by similarity to ideal solution (TOPSIS) methodology [40]. This technique is used to rank objects described by multidimensional criteria through determination of their distance in multidimensional space from the potentially most favorable and least favorable variants. These variants are determined by identifying, for each criterion, the highest and lowest rated values. As a result, two (not necessarily existing in the dataset) variants are created with the best and the worst properties. The distance, in the multidimensional space, from both of them is a measure of the quality of prediction of the variable, according to the formula:

\[ TIndex = \frac{DistWORST}{DistWORST + DistBEST} \]  

(8)

where \( TIndex \) is a ranking coefficient, \( DistWORST \) is a euclidean distance in multidimensional space from the worst variant, and \( DistBEST \) is a euclidean distance in multidimensional space from the most favorable variant.

The criteria used in TOPSIS method to characterize the prediction quality should be independent of the values scale. Thus, \( R^2, RPIQ, Sb, \) and \( LCCC \) statistics scaled to the interval \( <0; 1> \) were used. The prediction accuracy of the model is highly dependent on the data number, distribution, and range of variability. For this reason, the hierarchization performed is of relative importance, depending on the type and homogeneity of the data.

2.4. Importance of Spectral Bands to Soil Property Estimation

The multidimensionality of spectral data is a significant obstacle in some modeling algorithms. The reduction of the dimensionality of the input data is a component of linear
models (PCA, PLS). It is always associated with determining the parts of the spectrum that are important for obtaining a good estimation of the sought relationship between the spectral response and soil properties.

In PLSR models, the significance of individual parts of the spectrum for the estimation of a soil variable can be determined by calculating the indicator referred to as (VIP scores) variable importance in projection \[30,32\]. VIP scores, as a vector of numerical values with a length equal to the number of spectral registration points, can be calculated a posteriori. The sum of squares of the VIP scores for the spectrum is equal to the number of spectrum components. VIP scores less than 1.0 indicate a lower than average importance of a particular input value. Components with VIP scores greater than 1.0 are important for modeling. To assess the importance of spectral bands, calculations of VIP scores for 11 selected soil features (SOC, total N, Ca, Al, Fe, Cr, Cu, Pb, Zn, and Hg contents) were performed during the creation of PLSR models for these properties.

An independent analysis of the significance of spectral bands was performed with the neighborhood components analysis (NCA) algorithm. This algorithm, not related to a specific modeling method, is used to select the number of input variables a priori, especially in the case of a limited number of data \[32,41–43\]. In order to compare the results of this approach with VIP scores, calculations were performed on the same dataset.

3. Results

3.1. Results of Modeling of Texture

The search for an effective model to estimate soil texture is a frequent topic in soil science research \[44\]. The estimates of the content of the different fractions vary in terms of RMSE values (Table 1). Validation statistics for all algorithm groups analyzed are relatively weak \(R^2 < 0.75\) and \(RPIQ < 2.0\); \(R^2\), RPIQ, and LCCC signal significant agreement between predictions and observations, but RMSE is relatively high. In our study, the individual models unanimously indicate that sand has the largest prediction error, while silt and clay have much smaller errors. This is an indirect indication that when soil texture needs to be determined, it is safest to determine it using these two variables. However, it should be assumed that this result is specific to the AfSIS dataset and it results from a much larger concentration spread of Sand compared to the range of silt and clay. With similar \(R^2\) values, the RMSEs for silt and clay fractions are significantly smaller. Moreover, the distribution of sand content is characterized by negative asymmetry.

Table 1. Summary of validation statistics for models in the “Texture” group. The best prediction result for particular variables is marked in bold.

| Property | PLSR | 1DCNN | GRNN |
|----------|------|-------|------|
|          | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb |
| Sand [%] | 0.62 | 8.83 | 1.08 | 0.79 | 0.05 | 0.60 | 9.02 | 1.05 | 0.73 | -0.02 | 0.54 | 9.65 | 0.98 | 0.72 | -0.01 |
| Silt [%] | 0.54 | 4.04 | 1.41 | 0.73 | 0.00 | 0.37 | 4.71 | 1.21 | 0.59 | 0.31 | 0.32 | 4.90 | 1.16 | 0.56 | 0.03 |
| Clay [%] | 0.64 | 5.72 | 0.93 | 0.80 | -0.04 | 0.64 | 5.74 | 0.93 | 0.77 | 0.28 | 0.66 | 5.54 | 0.97 | 0.79 | 0.03 |

3.2. Results of Modeling of Properties from “Water” Group

The complicated and lengthy process of laboratory determination of water characteristics of soils using sand box and ceramic plates justifies the search for alternative means of estimating them. The validation statistics of the prediction models for the “Water” group indicate that the models based on MIR provide a reasonably good basis for estimation of these quantities (Table 2). RPIQs exceeding 2.0 signal a relatively good approximation of all relevant properties. Saturation at 0 and pF2.0 were estimated with the least error by the GRNN model; soil moisture at pF2.5 and pF4.2, by the 1DCNN model; while the volumetric density estimate with the highest accuracy is provided by the PLSR model.
Table 2. Summary of model validation statistics for the “Water” group. The best prediction result for particular variables is marked in bold.

| Property | PLSR | 1DCNN | GRNN |
|----------|------|-------|------|
|          | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb |
| Sat. at 0 \([g/g]\) | 0.61 | 0.09 | 1.64 | 0.77 | 0.04 | 0.61 | 0.09 | 1.66 | 0.77 | -0.05 | 0.70 | 0.08 | 1.69 | 0.82 | 0.07 |
| pH2.0 \([g/g]\) | 0.55 | 0.08 | 2.05 | 0.73 | 0.05 | 0.60 | 0.08 | 2.17 | 0.76 | 0.00 | 0.62 | 0.08 | 2.24 | 0.77 | 0.05 |
| pH2.5 \([g/g]\) | 0.56 | 0.08 | 2.20 | 0.74 | 0.05 | 0.62 | 0.07 | 2.38 | 0.77 | -0.01 | 0.62 | 0.07 | 2.38 | 0.77 | 0.05 |
| pH4.2 \([g/g]\) | 0.49 | 0.08 | 1.94 | 0.67 | 0.06 | 0.59 | 0.07 | 2.15 | 0.73 | 0.01 | 0.56 | 0.07 | 2.10 | 0.72 | 0.06 |
| B. Den. \([g/cm^3]\) | 0.76 | 0.08 | 2.96 | 0.86 | 0.01 | 0.74 | 0.08 | 2.83 | 0.85 | 0.03 | 0.73 | 0.08 | 2.76 | 0.84 | 0.00 |

3.3. Results of Modeling of C and N Contents

C and N concentrations in soils are the soil properties most commonly modeled, based on NIR and MIR spectral response [2,3]. This is due to the important role of both elements in shaping other soil properties, the role of N as a nutrient, and the importance of C sequestration in soils under conditions of increasing CO\(_2\) content in ambient air. Validation statistics (Table 3, Figure 4) signal that, in the AfSiS dataset, the concentration of total C and N, as well as SOC, can be determined with relatively high accuracy, even when the compound content of both elements in soils is low. RMSE estimates of 0.33% for C and 0.03% for N provide good approximations of the true values. Validation results indicate that (among those tested) the best model for estimating C and N is 1DCNN. This provides reason to believe that the relationship between the spectrum and SOC concentration (and associated N) is nonlinear, as indicated by the weaker result of the PLSR model. The result of the GRNN model is also relatively weak.

Table 3. Summary of model validation statistics for C and N concentration in soils. The best prediction result for particular variables is marked in bold.

| Property | PLSR | 1DCNN | GRNN |
|----------|------|-------|------|
|          | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb | \(R^2\) | RMSE | RPIQ | LCCC | Sb |
| Total N \([\%]\) | 0.84 | 0.03 | 2.03 | 0.91 | 0.01 | 0.91 | 0.03 | 2.68 | 0.95 | 0.00 | 0.80 | 0.04 | 1.85 | 0.88 | -0.01 |
| Total C \([\%]\) | 0.92 | 0.38 | 2.95 | 0.96 | 0.00 | 0.93 | 0.34 | 3.25 | 0.97 | 0.05 | 0.87 | 0.47 | 2.37 | 0.93 | -0.01 |
| SOC \([\%]\) | 0.91 | 0.37 | 3.02 | 0.96 | 0.00 | 0.93 | 0.33 | 3.42 | 0.96 | 0.05 | 0.86 | 0.47 | 2.39 | 0.92 | -0.02 |

Figure 4. Scatter plots of C and N content prediction validation data using the 1DCNN model. When the points are on the red line it is an indication of perfect prediction, deviations from red line indicate prediction errors.

3.4. Results of Modeling of Properties from “Elements” Group

Table 4 contains validation statistics for the prediction models of the 32 elements. The validation results vary widely, although there are several elements whose concentrations
can be estimated with relatively high accuracy. The RPIQ value indicates that a small relative error characterizes the prediction of Fe, Al, Ga, and Rb (RPIQ > 2.0). However, analyzing the RMSE level shows that, despite low values of RPIQ or R², the content of other macroelements can also be predicted with sufficient accuracy, taking into account the scale of data variability (data spread). Comparison of the validation results from the three models indicates that the GRNN model dominates in terms of prediction quality. There are six exceptions to this rule: Na, K, Ca, and Se, for which better prediction results were obtained with the PLSR method; and Se and Bi, for which a slightly better result was obtained with the 1DCNN model.

### Table 4. Summary of model validation statistics for the “Elements” group. The best prediction result for particular variables is marked in bold.

| Property | PLSR | I1DCNN | GRNN |
|----------|------|--------|------|
| Na [mg/kg] | 0.56 | 0.64 | 0.46 |
| Al [mg/kg] | 0.78 | 0.87 | 0.87 |
| Fe [mg/kg] | 0.77 | 0.86 | 0.86 |
| Ca [mg/kg] | 0.85 | 0.98 | 0.85 |
| Ti [mg/kg] | 0.41 | 0.69 | 0.36 |
| V [mg/kg] | 0.51 | 0.72 | 0.51 |
| Mn [mg/kg] | 0.44 | 0.67 | 0.54 |
| Cu [mg/kg] | 0.55 | 0.58 | 0.55 |
| Zn [mg/kg] | 0.41 | 0.12 | 0.41 |
| Ga [mg/kg] | 0.68 | 0.82 | 0.68 |
| Rb [mg/kg] | 0.71 | 0.72 | 0.71 |
| Sr [mg/kg] | 0.42 | 0.65 | 0.42 |
| Y [mg/kg] | 0.42 | 0.66 | 0.42 |
| Ba [mg/kg] | 0.39 | 0.64 | 0.39 |
| La [mg/kg] | 0.48 | 0.62 | 0.48 |
| Ce [mg/kg] | 0.36 | 0.67 | 0.36 |
| Pr [mg/kg] | 0.26 | 0.44 | 0.26 |
| Nd [mg/kg] | 0.42 | 0.69 | 0.42 |
| Sm [mg/kg] | 0.34 | 0.46 | 0.34 |
| Ho [mg/kg] | 0.14 | 0.54 | 0.14 |
| Ta [mg/kg] | 0.14 | 0.70 | 0.14 |
| W [mg/kg] | 0.27 | 0.57 | 0.27 |
| Hg [mg/kg] | 0.02 | 0.05 | 0.01 |
| Pb [mg/kg] | 0.03 | 0.33 | 0.02 |
| Bi [mg/kg] | 0.01 | 0.35 | 0.02 |
| Th [mg/kg] | 0.65 | 0.74 | 0.65 |

3.5. Results of Modeling of Properties from “Mehlich” Group

Table 5 shows the prediction statistics of chemical properties of soils from the “Mehlich” group. Apart from Al content and pH value, the RPIQ of other characteristics does not exceed a value of 2.0. A relatively good validation result is shown by the model of Ca and K content (PLSR), and Mg content (1DCNN).
Table 5. Summary of model validation statistics for the “Mehlich” group. The best prediction result for particular variables is marked in bold.

| Property          | PLSR     | 1DCNN   | GRNN    |
|-------------------|----------|---------|---------|
|                   | $R^2$    | RMSE    | RPIQ    | LCCC    | Sb   | $R^2$    | RMSE    | RPIQ    | LCCC    | Sb   | $R^2$    | RMSE    | RPIQ    | LCCC    | Sb   |
| EC [dS/m]         | 0.19     | 0.28    | 0.29    | 0.47    | -0.05 | 0.38    | 0.25    | 0.34    | 0.64    | -0.18 | 0.31    | 0.26    | 0.32    | 0.50    | 0.16 |
| Ex. Ac. [cmol/kg] | 0.40     | 0.45    | 1.15    | 0.58    | 0.02  | 0.65    | 0.39    | 1.33    | 0.73    | 0.02  | 0.55    | 0.48    | 1.09    | 0.57    | 0.07 |
| Ex. Bas. [cmol/kg]| 0.82     | 0.99    | 1.10    | 0.91    | 0.01  | 0.88    | 0.05    | 1.26    | 0.94    | 0.05  | 0.90    | 0.75    | 1.35    | 0.95    | 0.02 |
| M3 Al [mg/kg]     | 0.84     | 0.89    | 0.92    | 0.92    | 0.01  | 0.86    | 1.99    | 3.63    | 0.92    | 0.01  | 0.85    | 199.46  | 3.48    | 0.92    | 0.00 |
| M3 B [mg/kg]      | 0.39     | 1.33    | 0.26    | 0.55    | 0.11  | 0.70    | 0.93    | 0.37    | 0.80    | 0.23  | 0.37    | 1.35    | 0.62    | 0.20    | 0.21 |
| M3 Ca [mg/kg]     | 0.82     | 17.52   | 0.76    | 0.90    | 0.02  | 0.82    | 1.74    | 0.87    | 0.90    | 0.04  | 0.89    | 139.4    | 0.96    | 0.94    | 0.02 |
| M3 Cu [mg/kg]     | 0.19     | 0.90    | 0.98    | 0.98    | 0.00  | 0.49    | 0.41    | 0.54    | 0.63    | 0.04  | 0.39    | 72.11    | 1.22    | 0.56    | 0.07 |
| M3 Fe [mg/kg]     | 0.42     | 1.90    | 0.26    | 0.57    | 0.04  | 0.65    | 0.81    | 1.34    | 0.63    | 0.04  | 0.65    | 147.94   | 0.87    | 0.80    | 0.04 |
| M3 K [mg/kg]      | 0.18     | 227.55  | 0.57    | 0.55    | -0.13 | 0.39    | 0.41    | 0.66    | 0.56    | 0.18  | 0.65    | 232.25   | 1.29    | 0.81    | 0.01 |
| M3 Mg [mg/kg]     | 0.68     | 234.54  | 1.28    | 0.82    | -0.03 | 0.74    | 209.04  | 1.44    | 0.86    | 0.02  | 0.68    | 232.25   | 1.29    | 0.81    | 0.01 |
| M3 Mn [mg/kg]     | 0.47     | 75.63   | 1.54    | 0.64    | 0.02  | 0.58    | 67.78   | 1.72    | 0.73    | 0.04  | 0.51    | 73.30    | 1.59    | 0.72    | 0.04 |
| M3 Na [mg/kg]     | 0.00     | 1003.1  | 0.03    | 0.32    | 0.18  | 0.11    | 0.28    | 0.25    | 0.35    | 1.14  | 0.89    | 325.35   | 0.99    | 0.83    | 0.85 |
| M3 P [mg/kg]      | 0.02     | 28.66   | 0.25    | 0.15    | -0.15 | 0.25    | 25.06   | 0.28    | 0.39    | 0.35  | 0.19    | 25.99    | 0.27    | 0.38    | 0.14 |
| M3 S [mg/kg]      | 0.12     | 199.34  | 0.04    | 0.20    | 0.16  | 0.28    | 180.35  | 0.05    | 0.50    | 0.12  | 0.49    | 151.66   | 0.05    | 0.44    | 1.62 |
| M3 Zn [mg/kg]     | 0.13     | 1.50    | 0.64    | 0.30    | -0.15 | 0.22    | 1.42    | 0.68    | 0.31    | 0.08  | 0.11    | 1.52    | 0.64    | 0.34    | 0.02 |
| P3I               | 0.75     | 45.09   | 0.89    | 0.85    | 0.02  | 0.72    | 47.35   | 1.86    | 0.83    | 0.14  | 0.78    | 41.79    | 2.11    | 0.88    | 0.04 |

3.6. Relative Prediction Positions

Table 6 contains hierarchical summaries of the relative prediction positions of soil variables using the TOPSIS method considering the $R^2$, RPIQ, LCCC, and Sb criteria. A high position in the hierarchy indicates a prediction result more consistent with the observations. Hierarchization takes into account all validation results. A higher Tindex value for the same predicted element indicates an algorithm that more accurately approximates the variables. Despite some differences in the position of individual variables, a relatively narrow group of data can be predicted with the greatest precision: SOC, Al and M3 Al, Fe, K, pH, and total N. Ranks of prediction of soil texture elements and water properties are relatively low. However, this highly formalized hierarchical procedure does not invalidate the usefulness of lower-ranked models, which should be further evaluated. This applies, for example, to the prediction of Ca or pF, especially at higher values of these variables.

Spectral response is used for the estimation of soil properties in digital mapping of the environment as well as in the management of field operations. The use of spectral analysis in laboratory conditions should be included in a similar category as it allows rapid, inexpensive, and approximate estimation of some soil properties in the whole soil profile [45,46]. An important problem is that of identifying the optimal modeling algorithm, as well as the limiting prediction precision, taking into account the minimum documentation requirements.

3.7. Impact of Relationships between Properties

It can be assumed that in the field of spectral analysis used for quantitative estimation of soil properties, there are three categories of variables: those that affect the spectral response directly to the extent that they can be quantitatively identified; those that are correlated with properties that affect the spectral response and thus can be quantitatively estimated; and those that do not affect the spectral response and are not correlated with properties that affect the spectral response. Results of previous studies indicate that, for soils, the first category includes carbon organic compounds [47,48], when using both NIR and MIR spectra. For other features, the possibility of belonging to each category has to be evaluated experimentally.
Table 6. Hierarchical summary of relative predictive position of soil variables.

| Rank | Property | PLSR Tindex | IDCNN Property | Tindex | GRNN Property | Tindex |
|------|----------|-------------|----------------|--------|---------------|--------|
| 1    | M3 Al    | 0.839       | SOC            | 0.853  | M3 Al         | 0.843  |
| 2    | SOC      | 0.833       | M3 Al          | 0.847  | Al            | 0.800  |
| 3    | Total C  | 0.831       | Total C        | 0.846  | Ga            | 0.780  |
| 4    | K        | 0.809       | pH             | 0.824  | Fe            | 0.777  |
| 5    | pH       | 0.796       | Total N        | 0.808  | Total C       | 0.776  |
| 6    | B. Den.  | 0.790       | Al             | 0.782  | SOC           | 0.776  |
| 7    | Al       | 0.789       | K              | 0.777  | B. Den.       | 0.762  |
| 8    | Ga       | 0.748       | B. Den.        | 0.772  | pH            | 0.760  |
| 9    | Total N  | 0.737       | Fe             | 0.725  | Rb            | 0.733  |
| 10   | Fe       | 0.736       | Ga             | 0.718  | PSI           | 0.729  |
| 11   | Rb       | 0.732       | pF2.5          | 0.692  | K             | 0.728  |
| 12   | PSI      | 0.704       | PSI            | 0.683  | Total N       | 0.713  |
| 13   | Ca       | 0.666       | Rb             | 0.680  | pF2.5         | 0.690  |
| 14   | pF2.5    | 0.652       | Ex. Bas.       | 0.678  | Ex. Bas.      | 0.689  |
| 15   | Ex. Bas. | 0.646       | pF2.0          | 0.667  | Sat. at 0     | 0.682  |
| 16   | pF2.0    | 0.638       | M3 Mg          | 0.660  | pF2.0         | 0.680  |
| 17   | Sat. at 0| 0.630       | pF4.2          | 0.656  | M3 Ca         | 0.655  |
| 18   | M3 Mg    | 0.629       | Sat. at 0      | 0.635  | Nd            | 0.646  |
| 19   | M3 Ca    | 0.626       | M3 Ca          | 0.626  | Ca            | 0.646  |
| 20   | Cu       | 0.600       | M3 Mn          | 0.619  | pF4.2         | 0.642  |
| 21   | pF4.2    | 0.598       | Cu             | 0.589  | Ta            | 0.637  |
| 22   | Sand     | 0.593       | Ex. Ac.        | 0.583  | Ce            | 0.630  |
| 23   | Clay     | 0.592       | Sand           | 0.578  | M3 Mg         | 0.628  |
| 24   | Silt     | 0.588       | Mn             | 0.576  | Ni            | 0.623  |
| 25   | Se       | 0.562       | Clay           | 0.575  | Ti            | 0.621  |
| 26   | Th       | 0.559       | Sr             | 0.569  | La            | 0.621  |
| 27   | M3 Mn    | 0.557       | Ni             | 0.567  | Cr            | 0.620  |
| 28   | V        | 0.554       | M3 B           | 0.557  | Cu            | 0.619  |
| 29   | Nd       | 0.546       | Ce             | 0.550  | Mn            | 0.618  |
| 30   | Ta       | 0.545       | M3 Fe          | 0.547  | V             | 0.608  |

The linkages between soil properties that enable or interfere with the quantitative identification of some components can be traced by confronting the statistics of the prediction results with the cluster system generated by the variable agglomeration algorithm. Figure 5 shows the dendrogram of associations between soil properties obtained by Ward’s agglomeration procedure using the 1–r² metric [49,50]. The algorithm groups statistically related variables using a metric that expresses the similarity or dissimilarity between variables. The value of the association distance indicates the degree of similarity of the variables in the dataset. A cluster composed of mutually correlated variables Ti, Bi, V, Cu, Fe, Th, Pb, and Mn in models using MIR is characterized by coefficients of determination of, respectively: 0.67, 0.35, 0.63, 0.59, 0.84, 0.76, 0.64, 0.62. The highest R² value of the Fe prediction leader gives reason to believe that the variability of its content reflects to some extent the variability of the other elements of the cluster, supporting their prediction. The cluster related to Fe is in turn related to the cluster including M3 Cu, Pr, Ni, Cr, M3 Fe, Ex. Acid., PSI, M3 Al, and Al (R²: 0.50, 0.33, 0.66, 0.68, 0.49, 0.39, 0.78, 0.86, 0.79, respectively). There is also a cluster of variables related to soil pH, including pH, M3 Mg, M3 K, M3 Ca, Ex. Bas., and Ca. A characteristic group is formed by variables shaping water properties including: bulk density, pF2.0, pF2.5, pF4.2, SOC, total C, and total N. In this case, one might suspect that the supporting factor for quantitative identification in this group is total C, or SOC.
The graphs in Figure 6 illustrate the statistical relationships between Fe content and the concentration of other metals (Al, V, Mn, Cu, Pb, and Th). It can be assumed that the model of these elements’ content is to some extent related to the content of Fe, which is well quantified by the models. It cannot be excluded that the chain of connections between particular properties is more complex and may lead to correct estimation of features poorly influencing the spectral response or not influencing it at all.

Figure 5. Dendrogram of associations between variables obtained by Ward’s agglomeration method.

Figure 6. Scatter plots of the content of selected elements as a function of Fe content.
Similar relationships exist between the exchangeable bases and the elements included in them (Figure 7). The exchangeable bases value is shaped by Ca to the largest extent, although the association with K and Mg is also significant. Calcium content, fairly well identified through the MIR spectral response, is nonlinearly related to pH. It is difficult to indicate which of these elements is the primary basis for correct modeling of soil sorption elements.

![Figure 7. Scatter plots of soil variables significantly correlated with the exchangeable bases and Ca content.](image)

3.8. Importance of Spectral Bands in Prediction of Selected Properties

Figure 8 shows the diagrams of squares of VIP scores and NCA score values for 11 soil characteristics, including the concentration of heavy metals. The diagrams reflect the properties of the PLSR and NCA algorithms. In the case of PLSR, most orthogonalized inputs are used in the construction of the model. In the selection, in the NCA procedure, only spectral sections with appropriately high NCA scores are considered. In the analyzed cases, the number of values indicated by this procedure ranged from 12 to 54.

It can be seen that SOC and total N is determined by taking the same spectral segments into account. It can also be concluded that the concentration of Al, and especially Fe, is related to sections of the spectrum of heavy metals.
Figure 8. Distribution diagrams of squares of VIP scores and NCA scores for selected soil properties. The diagrams include index values exceeding 1.0. The diameters of points are proportional to the squares of VIP scores and NCA scores.

4. Discussion

Most soil data, for classification purposes or for making economic decisions, are interpreted in a semi-fuzzy way, or at least a tendency is observed leading to a blurring of the boundaries between classification criteria. The evaluation of modeling results of many soil characteristics must allow for a certain level of acceptable errors. In this context, the estimation of soil texture components, such as silt and clay contents, using the models described in this paper, can be considered as sufficiently precise. Similarly, the correctness of the models of water properties of soils, which are extremely time-consuming to determine in the laboratory using the classical approaches, should be assessed.

The modeling of SOC concentration based on NIR and MIR signatures is understandably of the greatest interest to researchers [2–4,51]. It can be assumed that, as an important component of soil organic compounds, SOC and total N content can be estimated with satisfactory accuracy. In terms of total element prediction, only a few elements can be
estimated with reasonable accuracy: Al, Fe, Ca, and K, along with several elements usually found in soils in low concentrations, namely Se, Rb, Ga, Zr, La, and Th. The contents of Cr, Ti, Mn, Ni, and Pb are estimated with slightly lower accuracy.

In modeling the “Mehlich” data group (17 variables), the 1DCNN (10 variables with the best prediction statistics) and GRNN (7 variables with the best prediction statistics) algorithms performed best. The PLSR model was inferior to the other two in every case. Particularly good prediction scores apply to: M3 Al, M3 B, M3 Mg, and pH (1DCNN model); and to total exchangeable bases, M3 Ca, M3 K, M3 Na, and PSI (GRNN model).

The results presented indicate that, among the examined models, there is no single optimal one that would provide the best predictions for all variables. Each of the models shows advantages in predicting a certain group of variables and at the same time is less useful in predicting others. This has certainly been influenced by the heterogeneity of data from many different countries, which is evident through the very large spread of variable values [24]. Compared to results with less dispersed distributions, more than 100 variables across the United States from the Kellogg Soil Survey Laboratory [8] show, in many cases, similar values of determination coefficients, but are associated, in the case of AfSIS data, with a much larger RMSE values. Comparisons between the two sets of results are difficult, mainly due to the smaller spread—often by several times—of the Kellogg Laboratory data relative to AfSIS, and also due to the slightly wider spectral range they used. Data from the USA were also used by Wijewardane et al. [9], who indicated satisfactory results for prediction of C, N, CEC, pH, and clay content, while poor results were produced for extractable forms of phosphorus and potassium. In this case, the results (RMSE) are similar to those obtained from the AfSIS data, presumably also due to less restrictive data selection conditions that did not limit the spread of the data. AfSIS observational data have been used in very advanced digital mapping attempts. Hengl et al. [25] presented a modeling methodology to generate a map of soil properties, among other things, from data collected in the AfSIS project. They used an ensemble of machine learning regression models whose inputs were topographic, hydrological, meteorological, and vegetation data, as well as spectral reflectance images collected by the Sentinel satellites (Copernicus program). Nineteen variables were modeled, mainly Mehlich 3 extractable components, pH, texture, CEC, C, and N contents. Results of modeling based on spectra recorded from the satellite ceiling were slightly worse than those obtained by proximal sensing. The disadvantage of this approach is the limitation of quantitative identification of variables to the surface layer. The indisputable advantage, however, is the possibility of relatively precise representation of many soil properties with good spatial resolution over a very large area.

If we arbitrarily take \( R^2 \geq 0.75 \) and \( RPIQ \geq 2.0 \) as the criteria of correct prediction, then for PLSR models they are fulfilled for 10 of the studied soil properties. For nine variables, these criteria are met when 1DCNN is used, while among predictions using the GRNN model, the adopted criteria are met for seven variables. For six variables, the assumed criteria are met by all tested algorithms. These variables are SOC, total C, Al, Fe, M3 Al, and pH.

The decision to use a model based on spectral analysis instead of the reference method is inevitably subjective. Using validation statistics, it is possible to estimate the risk of error based, for example, on \( R^2 \) and \( RMSE \) values. Acceptance or rejection of the model as an approximation for classification or mapping of soils is determined by the mapping scale (spatial resolution) and the way in which the variability of soils is reflected—either as a continuous image or discrete representation with information about the mean value of properties and the range of their variability.

Assuming a symmetrical distribution of deviations from observed values, properties such as sand, clay; total C, N, and SOC contents; and bulk density can be estimated with sufficient precision for large-scale soil mapping. In addition, the Ca content error of 0.5% relative to sample weight is small enough even for assessing the cultivation needs of soils, although the K content (error of 0.7%) is estimated with relatively low accuracy.
Fe compounds rarely show signs of deficiency; therefore, an error of 1% is not significant in practice. Similarly, Al concentration itself does not pose a significant problem; it only becomes important once the soil pH is below 5.0, when the presence of aluminum compounds is a threat to crops. The accuracy of estimates of other variables must be considered depending on the purpose and scale of the documentation required. For example, in the context of investigating threats to soils with high metal contents, even estimates of relatively low accuracy may be sufficient to determine the extent of transformation.

For large soil datasets, especially those from diverse geological settings, a significant asymmetry in the distribution of soil characteristics, usually right-handed, is characteristic. The aforementioned data from the US territory analyzed by Ng et al. [8] possess such a characteristic. Towett et al. [24], characterizing the chemical properties of the AfSIS data, noted the strong skewness of the distributions of many soil variables. Relative to the distribution of US data, the spread of AfSIS observations is, in most cases, much larger (in extreme cases, several tens of times). This has a significant impact on the level of prediction errors, especially at low values of the variables. In such circumstances, one of the solutions used is to exclude extremely high values (outliers) from the dataset. Such an operation, in the context of the real occurrence of high values of variables, is controversial. Figure 9 illustrates, based on the analyzed data, the relationship between the prediction error of the model with the most favorable characteristics (RMSE) and the prediction errors after removing the values considered as outliers, i.e., exceeding the assumed cut-off value (HL):

\[
HL = Q_3 + 1.5 \cdot IQR
\]

The graph shows two versions of this approach: the RMSE ratio after removing extreme values from the validation data after prediction with a model derived from the entire data (RMSE_{OUT}); and the RMSE ratio after removing extreme values from the training and validation data before training the model (RMSE_{modOUT}).

Figure 9. Example RMSE ratios.

A significant variation in the ratios of RMSE values can be seen. As a rule, the error after removing outliers before training is smaller than the RMSE for the full data and the RMSE after removing outliers from the validation data after training on the full data. The exceptions are clay, Sat. at 0, K, M3 Al, and pH, for which limiting the range of variables in
the learning set worsened the validation results. For sand and total N content, the exclusion of extreme feature values did not significantly affect the validation error.

The largest decreases in RMSE are for M3 Na, M3 K, M3 B, Hg, and Zr. Other RMSEs are closer to the variant with removal of outliers after training on the whole data. The problem of outliers for prediction of soil property values is debatable. The increase in accuracy due to truncation of large data values limits the applicability of the model, especially for memory-based models. In extreme cases, the construction of two different models should be considered [21]: one for data close to the modal and mean value, and a separate one for very high values of the variables. However, for models built for high property values, the problem will be the small number of learning examples.

5. Conclusions

Each of the models examined possesses a specific suitability for prediction of different soil variables. It should be assumed that differences in the correctness of estimation of particular variables depend on the type of relationship between the variable and the spectral response. This means that, due to the specificity of the relationship between the MIR spectrum and soil properties, it is not possible to indicate a universal optimal prediction model for all soil variables. Because of the wide range of the variable distributions, this result should be treated as specific for particular geological, morphological, and climatic conditions.

The best prediction results are provided by all regression models for total C, SOC, total Fe, Al, M3 Al, and pH. The clustering and modeling results indicate that many other soil properties do not directly affect the MIR spectral response, but there are correlations of features that can support such a prediction with limited accuracy.

Asymmetry in the distribution of the modeled variables makes it difficult to build a valid regression model. Removing outliers from the training set results in a significant reduction of the prediction error in many cases. However, this operation limits the universality of the model to a range relatively close to the median. Consideration of building two different models for different ranges of data variability is conditioned by the disproportion of the number of examples close to the median and outliers.

Author Contributions: Conceptualization, S.G.; data curation, S.G.; formal analysis, S.G.; funding acquisition, S.G. and W.G.; investigation, S.G.; methodology, S.G. and W.G.; project administration, S.G.; resources, S.G. and W.G.; software, S.G. and W.G.; supervision, S.G.; validation, S.G.; visualization, W.G.; writing—original draft, S.G. and W.G.; writing—review and editing, S.G. and W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a research subvention of AGH University of Science and Technology (Grant number 16.16.150.545).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: AfSIS Phase I was a collaborative projective funded by the Bill and Melinda Gates Foundation, which aimed to provide a consistent baseline of soil information for monitoring soil ecosystem services in sub-Saharan Africa. Partners included: The Tropical Soil Biology and Fertility Institute of The International Center for Tropical Agriculture, The International Soil Reference and Information Center—World Soil Information, The Center for International Earth Science Information Network, The Earth Institute at Columbia University, and World Agroforestry (Center for Research in Agroforestry-ICRAF, previously The International Council for Research in Agroforestry).

Conflicts of Interest: The authors declare no conflict of interest.
Appendix A

Table A1. Descriptive statistics of properties in the “Texture” group.

| Property | Units | Mean    | SD    | Min. | Max.  | IQR  | Skewness |
|----------|-------|---------|-------|------|-------|------|----------|
| Sand     | %     | 84.97   | 14.32 | 0.44 | 100.00 | 9.50 | −2.71    |
| Silt     | %     | 8.30    | 5.94  | 0.00 | 33.92  | 5.71 | 1.65     |
| Clay     | %     | 6.73    | 9.51  | 0.00 | 82.15  | 5.35 | 4.00     |

Table A2. Descriptive statistics of properties in the “Water” group.

| Property | Units | Mean    | SD    | Min. | Max.  | IQR  | Skewness |
|----------|-------|---------|-------|------|-------|------|----------|
| Sat. at 0| g/g   | 0.41    | 0.14  | 0.09 | 1.10  | 0.14 | 1.21     |
| pF2.0    | g/g   | 0.27    | 0.12  | 0.02 | 0.71  | 0.17 | 0.58     |
| pF2.5    | g/g   | 0.23    | 0.12  | 0.01 | 0.67  | 0.18 | 0.69     |
| pF4.2    | g/g   | 0.15    | 0.11  | 0.00 | 0.58  | 0.15 | 1.10     |
| B. Den.  | g/cm³ | 1.03    | 0.16  | 0.58 | 1.41  | 0.23 | −0.02    |

Table A3. Descriptive statistics of properties in the “CN” group.

| Property | Units | Mean    | SD    | Min. | Max.  | IQR  | Skewness |
|----------|-------|---------|-------|------|-------|------|----------|
| Total N  | %     | 0.08    | 0.08  | 0.00 | 0.58  | 0.07 | 2.46     |
| Total C  | %     | 1.21    | 1.31  | 0.09 | 11.19 | 1.11 | 2.69     |
| SON      | %     | 0.08    | 0.08  | 0.00 | 0.56  | 0.06 | 2.46     |
| SOC      | %     | 1.14    | 1.26  | 0.09 | 10.40 | 1.13 | 2.69     |

Table A4. Descriptive statistics of properties in the “Elements” group.

| Property | Units | Mean    | SD    | Min. | Max.  | IQR  | Skewness |
|----------|-------|---------|-------|------|-------|------|----------|
| Na       | mg/kg | 23,184.64 | 8764.41 | 9568.58 | 92,079.60 | 7023.58 | 3.60     |
| Al       | mg/kg | 30,499.97 | 18,155.75 | 824.40 | 120,089.47 | 25,022.70 | 0.63     |
| Cl       | mg/kg | 214.66 | 1210.23 | 8.90 | 28,193.55 | 111.00 | 21.85    |
| K        | mg/kg | 11,055.74 | 12,002.48 | 147.40 | 97,368.89 | 14,449.81 | 2.29     |
| Ca       | mg/kg | 7558.74 | 24,037.62 | 55.30 | 426,430.90 | 4772.05 | 11.01    |
| Ti       | mg/kg | 3907.00 | 4195.35 | 2.10 | 56,099.97 | 2970.08 | 5.02     |
| V        | mg/kg | 28.49 | 39.99 | 1.00 | 357.90 | 28.90 | 3.61     |
| Cr       | mg/kg | 61.13 | 84.01 | 1.00 | 1148.79 | 55.30 | 6.08     |
| Mn       | mg/kg | 390.29 | 451.06 | 5.40 | 3863.57 | 408.00 | 2.61     |
| Fe       | mg/kg | 26,233.94 | 27,297.12 | 74.30 | 181,690.38 | 26,458.38 | 2.43     |
| Ni       | mg/kg | 19.40 | 27.15 | 0.30 | 286.00 | 20.70 | 4.43     |
| Cu       | mg/kg | 14.91 | 15.06 | 0.70 | 128.40 | 15.60 | 2.67     |
| Zn       | mg/kg | 25.66 | 27.97 | 0.90 | 349.00 | 27.40 | 4.38     |
| Ga       | mg/kg | 7.18 | 5.95 | 0.20 | 28.20 | 9.20 | 0.95     |
| Se       | mg/kg | 886.63 | 451.06 | 5.40 | 3863.57 | 408.00 | 2.61     |
| Rb       | mg/kg | 23,184.64 | 8764.41 | 9568.58 | 92,079.60 | 7023.58 | 3.60     |
| Sr       | mg/kg | 85.46 | 155.68 | 0.40 | 1984.91 | 77.40 | 5.57     |
| Y        | mg/kg | 10.73 | 11.41 | 0.20 | 105.40 | 11.20 | 2.76     |
| Zr       | mg/kg | 65.63 | 114.80 | 19.70 | 950.41 | 5.10 | 4.03     |
| Ba       | mg/kg | 1788.02 | 2985.26 | 16.50 | 42,287.14 | 2237.29 | 6.50     |
| La       | mg/kg | 1974.23 | 4428.23 | 1.90 | 65,385.60 | 1506.81 | 8.13     |
| Ce       | mg/kg | 48.05 | 48.38 | 1.20 | 413.30 | 58.30 | 1.96     |
| Pr       | mg/kg | 2.04 | 3.41 | 0.50 | 29.70 | 0.60 | 4.00     |
| Nd       | mg/kg | 8.14 | 8.80 | 0.90 | 56.60 | 9.40 | 2.15     |
| Sm       | mg/kg | 148.84 | 737.65 | 0.60 | 9806.80 | 19.20 | 8.40     |
Table A4. Cont.

| Property | Unit | Mean | SD  | Min. | Max. | IQR | Skewness |
|----------|------|------|-----|------|------|-----|----------|
| Hf       | mg/kg| 2.90 | 4.37| 0.20 | 52.90| 3.30| 4.36     |
| Ta       | mg/kg| 4.54 | 6.68| 0.10 | 64.30| 4.10| 5.24     |
| W        | mg/kg| 0.93 | 2.24| 0.20 | 30.70| 0.30| 7.25     |
| Hg       | mg/kg| 1.97 | 4.85| 0.10 | 36.80| 0.40| 4.05     |
| Pb       | mg/kg| 37.51| 67.18| 0.40| 686.80| 34.60| 5.63     |
| Bi       | mg/kg| 1.18 | 3.06| 0.10 | 37.50| 0.90| 6.10     |
| Th       | mg/kg| 39.82| 91.74| 0.20| 818.50| 29.20| 4.89     |

Table A5. Descriptive statistics of properties in the “Mehlich” group.

| Property | Units | Mean | SD  | Min. | Max. | IQR | Skewness |
|----------|-------|------|-----|------|------|-----|----------|
| EC       | dS/m  | 0.13 | 0.31| 0.01 | 3.88 | 0.08| 8.11     |
| Ex. Ac.  | cmol/kg| 0.50 | 0.38| 0.05 | 4.21 | 0.52| 2.39     |
| Ex. Bas. | cmol/kg| 14.03| 23.63| 0.05| 186.95| 10.14| 3.40     |
| M3 Al    | mg/kg | 842.34| 488.75| 3.00| 3041.00| 660.25| 1.00     |
| M3 B     | mg/kg | 0.48 | 1.71| 0.00 | 23.55| 0.35| 8.82     |
| M3 Ca    | mg/kg | 2096.97| 4126.19| 0.00| 35200.00| 1332.25| 4.09     |
| M3 Cu    | mg/kg | 1.74 | 2.11| 0.00 | 23.70| 1.86| 4.65     |
| M3 Fe    | mg/kg | 116.00| 92.59| 1.26| 782.00| 88.28| 2.76     |
| M3 K     | mg/kg | 167.05| 251.24| 5.50| 3432.00| 128.70| 6.02     |
| M3 Mg    | mg/kg | 309.07| 412.41| 0.00| 4740.00| 300.32| 4.19     |
| M3 Mn    | mg/kg | 101.61| 104.46| 1.05| 592.00| 116.33| 1.67     |
| M3 Na    | mg/kg | 130.06| 998.77| 0.00| 23,000.00| 29.75| 21.22    |
| M3 P     | mg/kg | 11.77 | 28.99| 0.00 | 396.00| 7.08| 7.86     |
| M3 S     | mg/kg | 30.03 | 212.38| 0.62| 3940.00| 8.25| 15.07    |
| M3 Zn    | mg/kg | 1.46 | 1.61| 0.00 | 23.80| 0.96| 6.42     |
| pH       |       | 6.19 | 1.09| 3.61 | 9.72 | 1.33| 0.68     |
| PSI      |       | 84.88| 89.50| –116.90| 452.89| 88.27| 1.80     |

References

1. Angelopoulou, T.; Balafoutis, A.; Zalidis, G.; Bochtis, D. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review. *Sustainability* 2020, 12, 443. [CrossRef]

2. Gholizadeh, A.; Boruvka, L.; Saberioon, M.; Vašť, R. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Appl. Spectrosc.* 2013, 67, 1349–1362. [CrossRef] [PubMed]

3. Lazaar, A.; Mouazen, A.M.; Hammouti, K.E.L.; Fullen, M.; Pradhan, B.; Memon, M.S.; Andich, K.; Monir, A. The application of proximal visible and near-infrared spectroscopy to estimate soil organic matter on the Triffa Plain of Morocco. *Int. Soil Water Conserv. Res.* 2020, 8, 195–204. [CrossRef]

4. Viscarra Rossel, R.A.; Lobsey, C.R.; Sharman, C.; Flick, P.; McLachlan, G. Novel proximal sensing for monitoring soil organic C stocks and condition. *Environ. Sci. Technol.* 2017, 51, 5630–5641. [CrossRef] [PubMed]

5. Bellon-Maurel, V.; McBratney, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils—Critical review and research perspectives. *Soil Biol. Biochem.* 2011, 43, 1398–1410. [CrossRef]

6. Ramirez-Lopez, L.; Wadoux, A.M.J.-C.; Franceschini, M.H.D.; Terra, E.S.; Marques, K.P.; Sayao, V.M.; Dematté, J.A.M. Robust soil mapping at the farm scale with vis–NIR spectroscopy. *Eur. J. Soil Sci.* 2019, 70, 378–393. [CrossRef]

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

8. Ng, W.; Minasny, B.; Jeon, S.-H.; McBratney, A. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Secur.* 2022, 6, 100043. [CrossRef]

9. Wijewardane, N.K.; Ge, Y.; Wills, S.; Libohova, Z. Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. *Soil Sci. Soc. Am. J.* 2018, 82, 722–731. [CrossRef]

10. Margenot, A.J.; Calderon, F.J.; Parikh, S.J. Limitations and Potential of Spectral Subtractions in Fourier-Transform Infrared Spectroscopy of Soil Samples. *Soil Sci. Soc. Am. J.* 2016, 80, 10–26. [CrossRef]

11. Nath, D.; Laik, R.; Singh Meena, V.; Pramanick, B.; Kumar Singh, S. Can mid-infrared (mid-IR) spectroscopy evaluate soil conditions by predicting soil biological properties? *Soil Secur.* 2021, 4, 100008. [CrossRef]
12. Hati, K.M.; Sinha, N.K.; Mohanty, M.; Jha, P.; Londhe, S.; Sila, A.; Towett, E.; Chaudhary, R.S.; Jayaraman, S.; Vassanda Coumar, M.; et al. Mid-Infrared Reflectance Spectroscopy for Estimation of Soil Properties of Alfisols from Eastern India. *Sustainability* **2022**, *14*, 4883. [CrossRef]

13. Nduwamungu, C.; Ziadi, N.; Parent, L.-E.; Tremblay, G.F.; Thuries, L. Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: A review. *Can. J. Soil Sci.* **2009**, *89*, 531–541. [CrossRef]

14. Asensio, V.; Guall, S.; Vega, F.A.; Covelo, E.F. A soil quality index for reclaimed mine soils. *Environ. Toxicol. Chem.* **2013**, *32*, 2240–2248. [CrossRef]

15. Klimkowicz-Pawlas, A.; Ukalska-Jaruga, A.; Smreczak, B. Soil quality index for agricultural areas under different levels of anthropopressure. *Int. Agrophys.* **2019**, *33*, 455–462. [CrossRef]

16. Hu, Z.; Haneklaus, S.; Sparovek, G.; Schnug, E. Rare Earth Elements in Soils. Communications in Soil Science and Plant Analysis. *Commun. Soil Sci. Plant Anal.* **2006**, *37*, 1381–1420. [CrossRef]

17. Ramos, S.J.; Dinali, G.S.; Oliveira, C.; Martins, G.C.; Moreira, C.G.; Siqueira, J.O.; Guilherme, L.R.G. Rare Earth Elements in the Soil Environment. *Curr. Pollut. Rep.* **2016**, *2*, 28–50. [CrossRef]

18. Xu, X.; Chen, S.; Ren, L.; Han, C.; Lv, D.; Zhang, Y.; Ai, F. Estimation of Heavy Metals in Agricultural Soils Using Vis-NIR Spectroscopy with Fractional-Order Derivative and Generalized Regression Neural Network. *Remote Sens.* **2021**, *13*, 2718. [CrossRef]

19. Yang, M.; Xu, Y.; Zhang, J.; Chen, H.; Liu, S.; Li, W.; Hao, Y. Near-Infrared Spectroscopic Study of Heavy-Metal-Contaminated Loess Soils in Tongguan Gold Area, China. *Minerals* **2020**, *10*, 89. [CrossRef]

20. Vågen, T.-G.; Winowiecki, L.; Walsh, M.G.; Desta, L.T.; Tondoh, J.E. *Land Degradation Surveillance Framework (LSDF): Field Guide*; International Center for Tropical Agriculture, World Agroforestry Centre, and the Earth Institute at Columbia University: New York, NY, USA, 2010.

21. Vågen, T.-G.; Winowiecki, L.A.; Tondoh, J.E.; Desta, L.T.; Gumbricht, T. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma* **2016**, *263*, 216–225. [CrossRef]

22. Vågen, T.-G.; Shepherd, K.D.; Walsh, M.G.; Winowiecki, L.A.; Desta, L.T.; Tondoh, J.E. AfSIS Technical Specifications. Soil Health Surveillance. 2010. Available online: [https://worldagroforestry.org/sites/default/files/afsisSoilHealthTechSpecs_v1_smaller.pdf](https://worldagroforestry.org/sites/default/files/afsisSoilHealthTechSpecs_v1_smaller.pdf) (accessed on 10 April 2022).

23. Leenaars, J.G.B.; van Oostrum, A.J.M.; Gonzalez, M.R. *Africa Soil Profiles Database, Version 1.2. A Compilation of Georeferenced and Standardised Legacy Soil Profile Data for Sub-Saharan Africa (with Dataset)*; ISRIC Report 2014/01; Africa Soil Information Service (AfSIS) Project; ISRIC, World Soil Information: Wageningen, The Netherlands, 2014.

24. Towett, E.K.; Shepherd, K.D.; Tondoh, J.E.; Winowiecki, L.A.; Lulsegd, T.; Nyambura, M.; Sila, A.; Vågen, T.o.r.-G.; Cadisch, G. Total elemental composition of soils in Sub-Saharan Africa and relationship with soil forming factors. *Geoderma Reg.* **2015**, *5*, 157–168. [CrossRef]

25. Hengl, T.; Miller, M.; Križan, J.; Shepherd, K.; Sila, A.; Kilibarda, M.; Antonijević, O.; Glušica, L.; Dobermann, A.; Haefele, S.; et al. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Sci. Rep.* **2022**, *11*, 6130. [CrossRef] [PubMed]

26. Vågen, T.-G.; Winowiecki, L.A.; Desta, L.T.; Tondoh, J.E.; Weullow, E.; Shepherd, K., Sila, A. Mid-Infrared Spectra (MIRS) from ICRAF Soil and Plant Spectroscopy Laboratory: Africa Soil Information Service (AfSIS) Phase I 2009–2013. World Agroforestry–Research Data Repository, V1. 2020. Available online: [https://data.worldagroforestry.org/dataset.html?persistentId=doi:10.34725/DVN/QXCPW1](https://data.worldagroforestry.org/dataset.html?persistentId=doi:10.34725/DVN/QXCPW1) (accessed on 10 April 2022).

27. Summerauer, L.; Baumann, P.; Ramirez-Lopez, L.; Barthel, M.; Bauters, M.; Bukombe, B.; Reichenbach, M.; Boeckx, P.; Kearsley, E.; Van Oost, K.; et al. The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis. *Soil* **2021**, *7*, 693–715. [CrossRef]

28. Vågen, T.-G.; Winowiecki, L.A.; Desta, L.T.; Tondoh, J.E.; Weullow, E.; Shepherd, K.; Sila, A.; Dunham, S.J.; Hernández-Allica, J.; Carter, J.; et al. Wet Chemistry Data for a Subset of AfSIS: Phase I Archived Soil Samples. World Agroforestry–Research Data Repository, V1. 2021. Available online: [https://data.worldagroforestry.org/dataset.html?persistentId=doi:10.34725/DVN/66BF0](https://data.worldagroforestry.org/dataset.html?persistentId=doi:10.34725/DVN/66BF0) (accessed on 10 April 2022).

29. Mehlich, A. Mehlich 3 Soil Test Extractant. A Modification of the Mehlich 2 Extractant. *Commun. Soil Sci. Plant Anal.* **1984**, *15*, 1409–1416. [CrossRef]

30. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]

31. Leone, A.P.; Viscarra-Rossel, R.A.; Amenta, P.; Buondonno, A. Prediction of soil properties with PLSR and vis-NIR Spectroscopy: Application to Mediterranean soils of Southern Italy. *Curr. Anal. Chem.* **2012**, *8*, 283–299. [CrossRef]

32. MATLAB, Version 9.13.0 (R2022b); The MathWorks Inc.: Natick, MA, USA, 2022.

33. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]

34. Dhanoua, M.S.; Lister, S.J.; Sanderson, R.; Barnes, R.J. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *J. Near-Infrared Spectrosc.* **1994**, *2*, 43–47. [CrossRef]

35. Gruszczyński, S.; Gruszczyński, W. Supporting soil and land assessment with machine learning models using the Vis-NIR spectral response. *Geoderma* **2022**, *405*, 115451. [CrossRef]
36. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org (accessed on 1 January 2022).

37. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 15 June 2021).

38. Specht, D.F. A general regression neural network. IEEE Trans. Neural Netw. 1991, 2, 568–576. [CrossRef]

39. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989, 45, 255–268. [CrossRef] [PubMed]

40. Hwang, C.-L.; Lai, Y.-J.; Liu, T.-Y. A new approach for multiple objective decision making. Comput. Oper. Res. 1993, 20, 889–899. [CrossRef]

41. Goldberger, J.; Roweis, S.; Hinton, G.; Salakhutdinov, R. Neighbourhood Components Analysis. 2022. Available online: https://www.cs.toronto.edu/~{}hinton/absps/nca.pdf (accessed on 4 November 2022).

42. Sinaice, B.B.; Owada, N.; Saadat, M.; Toriya, H.; Inagaki, F.; Bagai, Z.; Kawamura, Y. Coupling NCA Dimensionality Reduction with Machine Learning in Multispectral Rock Classification Problems. Minerals 2021, 11, 846. [CrossRef]

43. Yang, W.; Wang, K.; Zuo, W. Neighborhood Component Feature Selection for High-Dimensional Data. J. Comput. 2012, 7, 161–168. [CrossRef]

44. Thomas, C.L.; Hernandez-Allica, J.; Dunham, S.J.; McGrath, S.P.; Haefele, S.M. A comparison of soil texture measurements using mid-infrared spectroscopy (MIRS) and laser diffraction analysis (LDA) in diverse soils. Sci. Rep. 2021, 11, 16. [CrossRef] [PubMed]

45. Debaene, G.; Bartmiński, P.; Niedźwiecki, J.; Miturski, T. Visible and Near-Infrared Spectroscopy as a Tool for Soil Classification and Soil Profile Description. Pol. J. Soil Sci. 2017, 50, 1–10. [CrossRef]

46. Xie, X.-L.; Li, A.-B. Identification of soil profile classes using depth-weighted visible–near-infrared spectral reflectance. Geoderma 2018, 325, 90–101. [CrossRef]

47. Francos, N.; Ogen, Y.; Ben-Dor, E. Spectral assessment of organic matter with different composition using reflectance spectroscopy. Remote Sens. 2021, 13, 1549. [CrossRef]

48. Ramirez, P.B.; Calderón, F.J.; Haddix, M.; Lugato, E.; Cotrufo, M.F. Using diffuse reflectance spectroscopy as a high throughput method for quantifying soil c and n and their distribution in particulate and mineral-associated organic matter fractions. Front. Environ. Sci. 2021, 9, 634472. [CrossRef]

49. De Amorim, R.C. Feature relevance in Ward’s hierarchical clustering using the Lp norm. J. Classif. 2015, 32, 46–62. [CrossRef]

50. TIBCO Software Inc. Statistica (Data Analysis Software System), Version 13. 2017. Available online: http://statistica.io (accessed on 15 June 2021).

51. Zhou, W.; Li, H.; Wen, S.; Xie, L.; Wang, T.; Tian, Y.; Yu, W. Simulation of Soil Organic Carbon Content Based on Laboratory Spectrum in the Three-Rivers Source Region of China. Remote Sens. 2022, 14, 1521. [CrossRef]