sciendo
DOI: 10.2478/ssa-2019-0027

SOIL SCIENCE ANNUAL
Vol. 70 No. 4/2019: 298–313

STANISŁAW GRUSZCZYŃSKI*

AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering,
Department of Environmental Management and Protection
Al. Mickiewicza 30, 30-059 Kraków, Poland

Prediction of soil properties with machine learning models based on the spectral response of soil samples in the near infrared range

Abstract: One of the basic methods for soil analysis time and cost reduction is using soil sample spectral response in laboratory conditions. The problem with this method lies in determining the relationship between the shape of the soil spectral response and soil physical or chemical properties. The LUCAS soil database collected by the EU's ESDAC research centre is good material to analyse the relationship between the soil properties and the near infrared (NIR) spectral response. The modelling described in the paper is based on these data. The analysis of the impact of soil properties configuration on absorbance levels in various NIR spectrum ranges was conducted using the stepwise regression models with the properties, properties squared and products of properties being explanatory variables. The analysis of partial correlation of soil properties values with absorbance values and absorbance derivative in the entire spectral range was conducted in order to evaluate the impact of the absorbance transformation (the first derivative of absorbance vector) on the change of significance of relationship with properties values. The Multi Layer Perceptron (MLP) models were used to estimate the absorbance relationship with single soil features. Soil property modelling based on the selection and transformation algorithm of raw values and first and second absorbance derivatives was also conducted along with the suitability evaluation of such models in building digital soil maps. The absorbance is affected by a limited number of tested soil features like pH, texture, content of carbonates, SOC, N, and CEC; P and K contents have, in case of this research, a negligible impact. The NIR methodology can be suitable in conditions of limited soil variation and particularly in development of thematic soil maps.

Keywords: LUCAS database, near-infrared spectroscopy, machine learning models, soil properties prediction

INTRODUCTION

The soil property analysis using indirect methods, omitting "wet methods", allows for making the soil observation points more dense, which is necessary, for example, to make digital soil maps more detailed (McBratney et al. 2002; McBratney et al.2003; Brevik and al. 2016). One of the basic method for time and cost reduction of soil analysis is using soil samples spectral response. USGS provides a comprehensive library of spectral responses of natural and artificial materials (Kokaly et al. 2017). The problem with this method lies in determining the relationship between the shape of the soil spectral response and soil physical or chemical properties. A vector composed of thousands of reflectance or absorbance data items reflects the action of one or many soil characterising factors in individual spectrum fragments. This means that the desired information should be extracted based on the spectrum shape analysis and cleared of noise caused by the action of other factors that are not being observed.

There have been numerous attempts to develop suitable models to link the spectral response with the properties set with various results. When the research area is relatively homogeneous in terms of geology and physiography, the attempts are successful, at least in terms of SOC and N concentration (Shi et al. 2015; Kania and Gruba 2016). In conditions of a larger mineralogical variability (Fang et al. 2018), the modelling results are worse. The sources of approximation models based on determination of the NIR absorbance connected with the vector dimensionality reduction are being searched for in statistical methods (regression with PCA transformation, stepwise regression, partial least-square regression PLSR – a dominating approach in the spectral response analysis) and in machine learning methods (MultiLayer Perceptrons-MLP, Radial Basis Functions-RBF, Support Vector Machines – SVM models, stack autoencoders, convolution networks in regression applications, random trees, including the so-called random forests). In addition to raw data (readings of reflectance or absorbance vectors in the NIR spectrum), the input data include transformed vectors like those subjected to PCA reduction, first and second derivatives of spectral response vectors and data filtered by autoen-

* *Prof. dr hab. inż. S. Gruszczyński, sgrusz@agh.edu.pl*

http://ssa.ptg.sggw.pl/issues/2019/704

coders (Fuentes et al. 2012; Qiu et al. 2014; Shi et al. 2015; Veres et al. 2015; Zhang et al. 2016; Conforti et al. 2018; Mohamed et al. 2018). It is not possible to indicate objectively the best approach. Note that in the very nature of things, the absorbance vector values are modulated due to a simultaneous action of various factors (SOC content, texture N content, pH, carbonate content, CEC, etc.) as indirectly shown by the results of the studies.

The LUCAS soil database collected by the EU's ESDAC research centre (Tóth et al. 2013; Ballabio et al. 2016; Orgiazzi et al. 2017) is a good material to analyse the relationship between the soil properties and the NIR spectral response (Stevens et al. 2013). The modelling described in the paper is based on these data. The paper aims at analysing the impact of determined soil samples properties on the NIR spectral response using a large and varied database from different regions of Europe, and determining the usefulness of machine learning (ML) models for predicting some soils properties treated individually and as vector of features. Statistical indicators of the model deviation from validation data have been used as evaluation criteria.

MATERIALS AND METHODS

The LUCAS soil sample data (17272 data records with full analytical information) from 23 EU countries were transformed to the 0–1 range. In addition to the sample location and typological data, the LUCAS database includes: texture (clay, silt, sand content), pH in w $CaCl_2$, pH in H_2O, SOC, $CaCO_3$, N, P, K as well the CEC. The raw and transformed data statistics are presented in Table 1.

The extent of the study, covering most EU countries, results in a very differentiated dataset: variation exceeds 40%, and in some cases even exceeds 100%, which is particularly visible in the case of concentration of carbonates.

The LUCAS dataset, except to the soil properties value data, contains data vector of the absorbance of all samples in the 400–2500 nm spectral range mea-sured at 0.5 nm. The research team found that, in the spectral region below 500 nm, distortions referred to as "instrumental artifacts" had occurred (Stevens et al. 2013), which justified the exclusion of this segment of the spectrum from modeling. Absorbance is a commonly used logarithmic conversion of reflectance, designed to linearize the interdependences of the spectral response and chemical characteristics of the samples.

TABLE 1. Basic statistics (arithmetic mean, standard deviation and coefficient of variation) properties/transformed to 0–1 range soil samples, from the LUCAS database

Properties	Mean	StdDev	v%
Clay [%]	18.8/0.2385	12.9/0.1633	68.5
Silt [%]	38.3/0.4158	18.4/0.2001	48.1
Sand [%]	42.9/0.4276	26.1/0.2667	62.4
pH in $CaCl_2$	5.75/0.4727	1.35/0.2044	43.2
pH in H_2O	6.35/0.4394	1.29/0.1944	44.2
SOC [g kg^{-1}]	25.24/0.1524	19.31/0.1166	76.5
$CaCO_3$ [g kg^{-1}]	55.8/0.0592	129.8/0.1376	232.4
N [g kg^{-1}]	1.96/0.1443	1.23/0.0905	62.7
P [mg kg^{-1}]	29.4/0.0554	30.2/0.0568	102.5
K [mg kg^{-1}]	190.9/0.026	224.2/0.0305	117.3
CEC [cmol(+) kg^{-1}]	13.6/0.0998	9.6/0.0706	70.7

In the calculations, the absorbance values of the samples and the values of their first and second derivatives were used. The derivatives were calculated using the *diff* function available in the MATLAB system. The *diff* function is calculating differences between adjacent elements of vector or matrix.

Absorbance values, as well as their first and second derivatives, are large vectors of strongly correlated data. This justifies the necessity of extracting relevant information combined with reducing the size of the vector. To reduce the size of the input vector (absorbance, the first and second absorbance derivative) of the machine learning models, the MATLAB system plsregress function implementing the least squares partial regression algorithm was used. In the MATLAB environment (using the SIMPLS algorithm being a variation of PLS), a matrix of coefficients defining a linear combination of PLS components approximating prediction variables was obtained in relation to the matrix of values of soil properties.

In order to determine the effect of soil characteristics differentiation on the absorbance values at individual points of the NIR spectrum, an analysis was carried out using a stepwise regression algorithm to determine the occurrence of individual properties, squares and products of their values in regression models. In the calculations the stepwisefit function available in the MATLAB environment was used, which is an implementation of the stepwise regression algorithm in this software package. This required calculating the parameters of 4000 regression equations, corresponding to the values of absorbance at particular points in the spectrum.

In order to determine the possibility of predicting the values of soil features based on the determination of the absorbance at individual points of the NIR spectrum, and its first and second derivative, an analysis of the effectiveness of MLP models *(Multi Layer Perceptron)* in this task was carried out. The calculations were carried out in the MATLAB package environment using the Neural Network Toolbox module (currently Deep Learning Toolbox). The MLP model algorithm is a universal approximator (Bengio 2009) ensuring an approximation of any continuous, monotonic and limited function. Thus, it potentially allows for a better illustration of the relationships between variables than statistical models, which are linear by assumption. Due to the strong correlations between absorbance, the cycle included every 10th vector value (17272 absorbance values in each case) and a corresponding soil variable value. The cycle was repeated for the absorbance vector subjected to numerical differentiation in the MATLAB package. The hidden layer of MLP models was composed of 20 tangensoidal processing units, random process initiation and Levenberg-Marquard (LM) learning algorithm. For each of the 10 soil features. The MLP architecture is 1: 20: 1 (Input = local absorbance value, Output = property value). For each of the features (clay, silt, sand, pH etc.) the procedure was carried out training (successively) 400 MLP models (every absorbance vector value) with given architecture, modeling cycle repeated 3 times. Only the coefficient of determination was calculated to assess the impact of the local absorbance on the soil attribute value. An "early stopping" rule was applied. The data set (about 17,000) was randomly divided into training (70%), test (15%) and validation (15%). Only obtained curves of R^2 values, Significance-Absorbance relationships for particular points of the NIR spectrum (step: 5 nm). 12,000 MLP models were optimized together, and the result did not show a strong enough connection at any point in the spectrum. The results are presented in figures 5–9. The machine learning optimisation algorithms have some randomness which depends on the optimisation starting point. Moreover, some algorithms (including LM) have a tendency during the weights optimisation to get bogged down in local minima.

In the calculations of machine learning (ML) models for the prediction of individual features, based on the reduced of absorbance vectors and its first and second derivative using the PLSR algorithm, the modules made available on the H_2O platform (Qiu et al. 2014, Website 1) were used. H_2O is an open source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform that allows building machine learning models on big data and provides easy productionalization of those models in an enterprise environment. These results were obtained using the program module of the H_2O platform called AutoML, implementing the "stacked ensemble" model. This algorithm consists in building manymodels available on the platform (Linear Models, Gradient Boosting Machine, Distributed Random Forest, Deep Learning Network) with the decision algorithm in the form of MLP. The regression models available on the H2O platform can have only one output.

One of the features of machine learning algorithms is the ability to construct models with multiple regression outputs. In some cases, this improves the approximation results, especially after adding the "denoise" stage of the result. The calculations of this model were carried out in the following steps: in step 1 – MLP regression model with many outputs (corresponding to modelled variables) and inputs in the form of transformed absorbance variables (as in the first path) which, in step 2 generated evaluations of variables denoised with the MLP algorithm (inputs – evaluations from step 1, outputs – observation data of modelled features).

One of the possibilities to improve the quality of predictions (statistical and machine learning models) is to expand the list of input variables. It can be expected that strengthening the quality of prediction of soil properties will improve the inclusion of texture information in the list of input variables. Assuming the availability of information on the soil texture class (qualitative variable) from the cartographic soil documentation, this seems to be the potentially cheapest way to improve the machine learning model. This possibility was used to build another MLP denoised model for forecasting the vector of soil properties.

All statistics for the assessment of soil property prediction models relate to the validation set (4330 examples drawn from the 17272 data set).

RESULTS AND DISCUSSION

In order to determine the effect of soil properties differentiation on the absorbance values at individual points of the NIR spectrum, an analysis was carried out using a stepwise regression algorithm. The vector of explanatory variables corresponding to the respective absorbance values of the spectral range consisted of: numerical values of individual soil sample properties (11 variables), numerical values of individual soil sample properties squared (11 variables) and products of numerical values of individual soil sample properties (55 variables). The stepwise regression algorithm indicates, indirectly, significant and insignificant components of the vector 77 of explanatory variables in the models of linear absorbance values at particular points in the spectrum. The coefficients of regression equations corresponding to the spectrum range points in which the spectral response of soil samples was recorded were calculated.

Figure 1 presents the values of the coefficients of determination of 4000 models corresponding to the individual points of registration of the spectral response calculated according to the formula:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where:

R^2 – determination coefficient,
SSE – sum of squares of models residues,
SST – sum of squares of data deviation from the mean.

The observed R^2 values are in the 0.49–0.71 range and the curve shape is similar to the shape of absorbance curves. The relations of root-mean-square error (RMSE) to local standard deviation corresponding to the model absorbance value is a mirror image of the determination coefficients curve. This indicates a rather weak variation of the models error root (0.06–0.07).

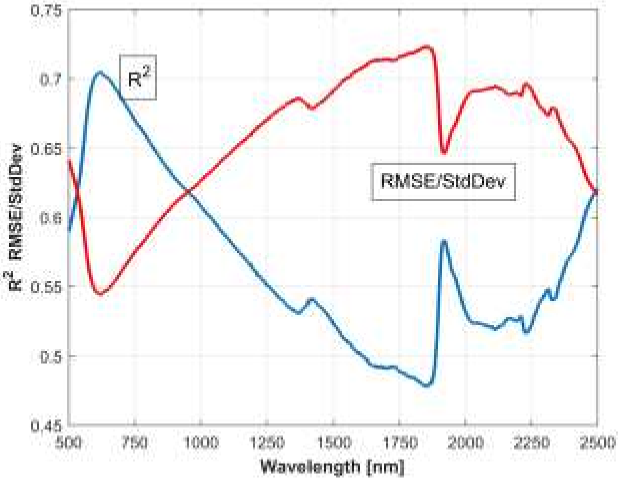


FIGURE 1. Curves of determination coefficients of relation between stepwise regression models (inputs: the vector with 77 soil variables) and absorbance values and corresponding ratios of observance estimation root-mean-square error to standard deviation from the sample

The stepwise regression algorithm accounts for variables that have a significant impact on the explained variable. Figures 2–4 indicate significant and insignificant components of the explanatory data vector. According to the diagram in Figure 2, significant for the absorbance values are: silt, SOC and $CaCO_3$ content as well as the CEC. The remaining variables appear as explanatory variables only in some spectrum ranges. The image of explanatory variables squared in the model (Figure 3) is similar, although the role of clay and N content is more significant.

The interpretation of Figure 4 indicates the significance of combined spectral NIR action of products: clay content (with pH in $CaCl_2$, N and K content), silt content (with SOC, $CaCO_3$, N and K content), sand content (with pH in $CaCl_2$, SOC, N, P content), pH in $CaCl_2$ (with N content), pH w H_2O (with CEC), SOC content (with P content and CEC), $CaCO_3$ content (with N and P content, CEC) and N content (with P and K content).

The obtained results lead to the conclusion that some soil features affect the NIR spectral response, being potential sources of noise in the determination range of other features (Stenberg et al. 2010; Wetterlind et al. 2013). Information extraction algorithms

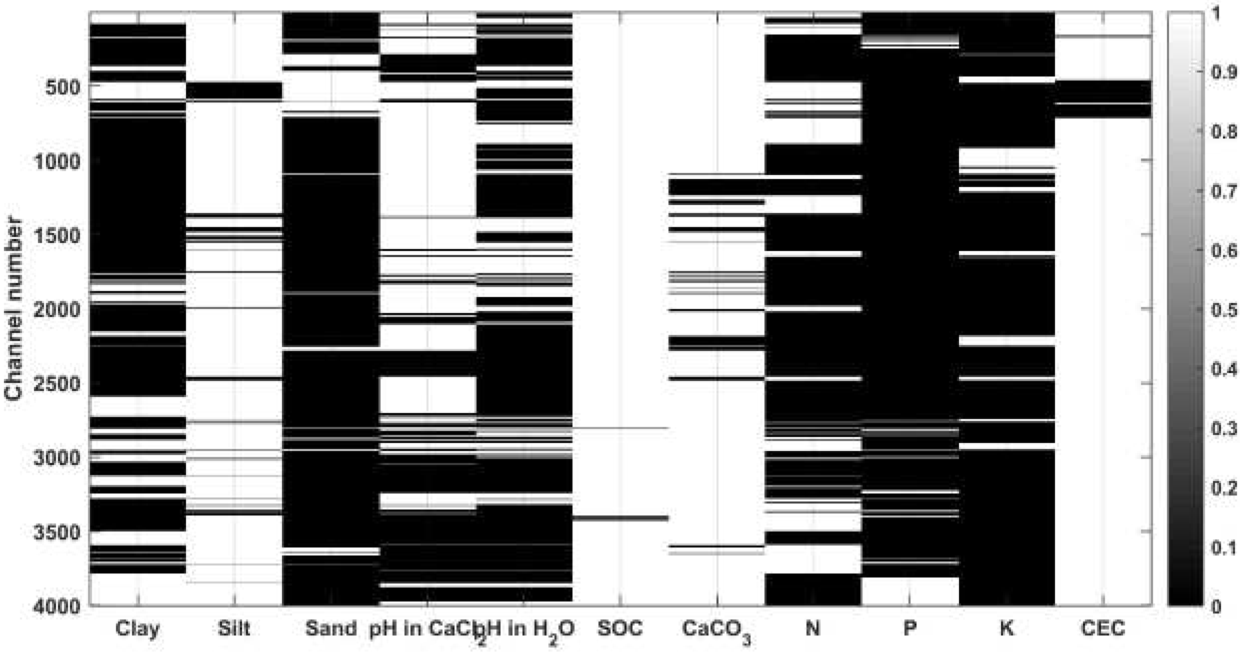


FIGURE 2. Diagram of the significance of explanatory variables relative to absorbance values in individual spectrum ranges; bright fields: variable present in the range, black fields: variable absent in the regression equation

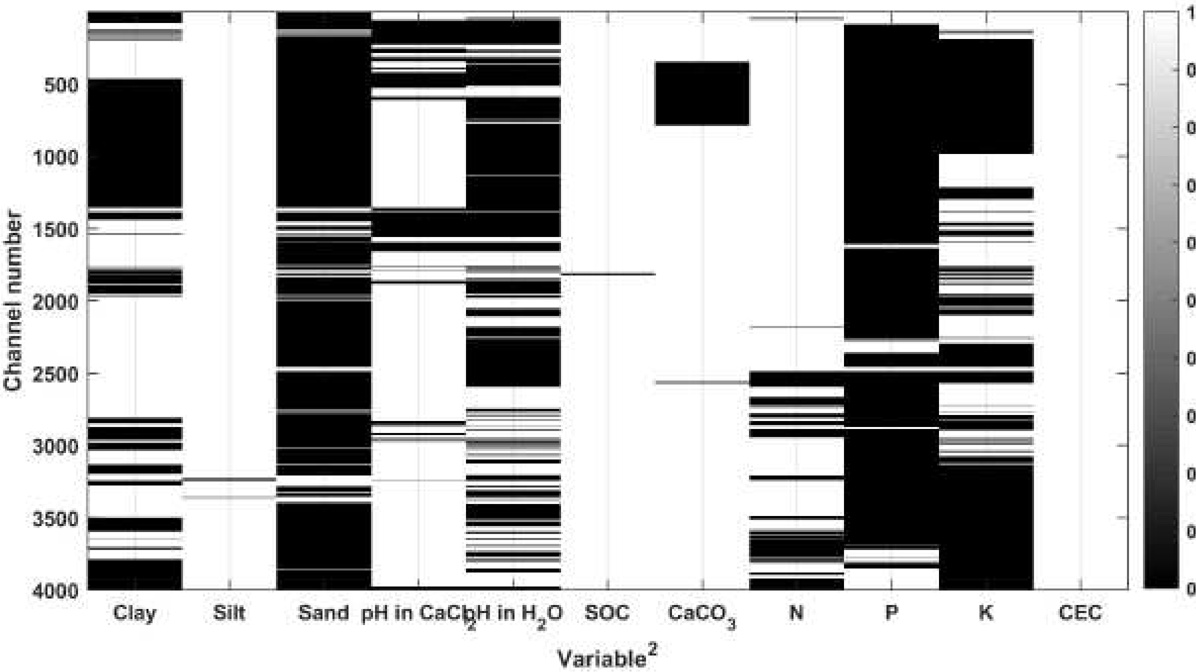


FIGURE 3. Diagram of the significance of explanatory variables squared relative to absorbance values in individual spectrum ranges; bright fields: variable present in the range, black fields: variable absent in the regression equation

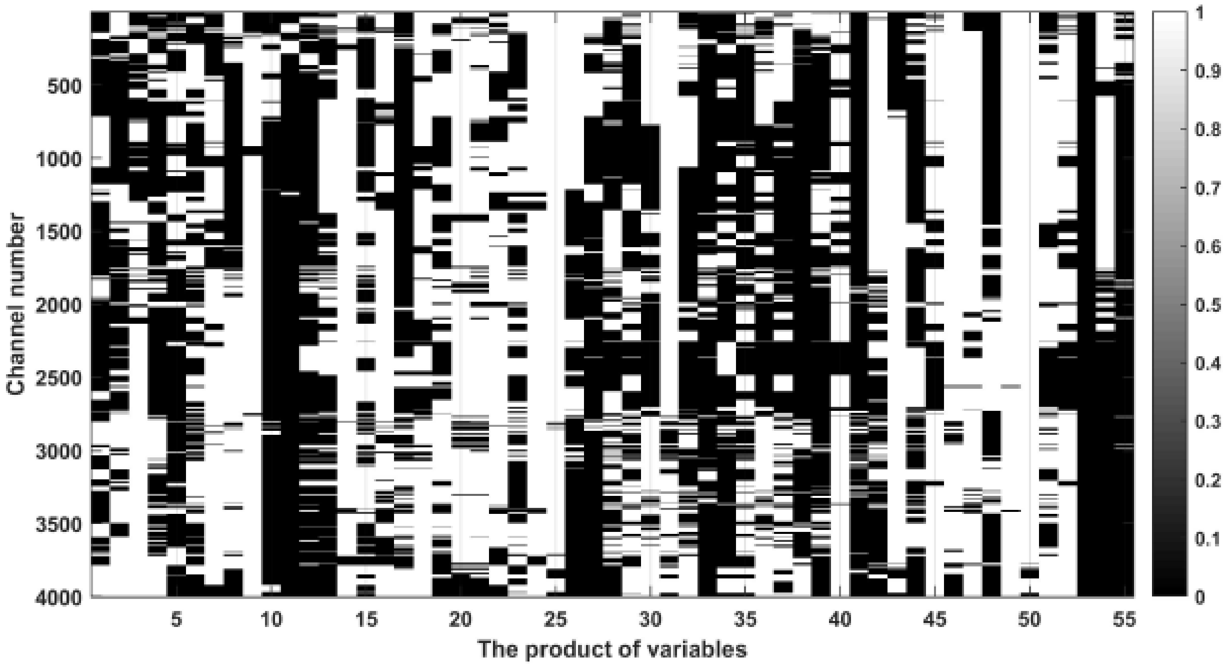


FIGURE 4. Diagram of the significance of products of explanatory variables relative to absorbance values in individual spectrum ranges; bright fields: variable present in the range, black fields: variable absent in the regression equation. Product identifiers: 1– (Clay × Silt), 2 – (Clay × Sand), 3– (Clay × pHCaCl$_2$), 4 – (Clay × pHH$_2$O), 5 – (Clay × SOC), 6 – (Clay × CaCO$_3$), 7 – (Clay × N), 8 – (Clay × P), 9 – (Clay × K), 10 – (Clay × CEC), 11– (Silt × Sand), 12 – (Silt × pHCaCl$_2$), 13 – (Silt × pHH$_2$O), 14 – (Silt × SOC), 15 – (Silt × CaCO$_3$), 16 – (Silt × N), 17 – (Silt × P), 18 – (Silt × K), 19 – (Sil × CEC), 20 – (Sand × pHCaCl$_2$), 21 – (Sand × pHH$_2$O), 22 – (Sand × SOC), 23 – (Sand × CaCO$_3$), 24 – (Sand × N), 25 – (Sand × P), 26 – (Sand × K), 27 – (Sand × CEC), 28– (pHCaCl$_2$ × pHH$_2$O), 29 – (pHCaCl$_2$ × SOC), 30 – (pHCaCl$_2$ × CaCO$_3$), 31 –(pHCaCl$_2$ × N), 32 – (pHCaCl$_2$ × P), 33 – (pHCaCl$_2$ × K), 34 – (pHCaCl$_2$ × CEC), 35– (pHH$_2$O × SOC), 36 – (pHH$_2$O × CaCO$_3$), 37 – (pHH$_2$O × N), 38 – (pHH$_2$O × P), 39 –(pHH$_2$O × K), 40 – (pHH$_2$O × CEC), 41 – (SOC × CaCO$_3$), 42 – (SOC × N), 43 – (SOC × P), 44 – (SOC × K), 45 – (SOC × CEC), 46 – (CaCO$_3$ × N), 47 – (CaCO$_3$ × P), 48 – (CaCO$_3$ × K), 49 – (CaCO$_3$ × CEC), 50 – (N × P), 51 – (N × K), 52 – (N × CEC), 53 – (P × K), 54–(P × CEC), 55 – (K × CEC)

from the extensive data vector are necessary. The literature includes two approaches to the information extraction from the spectral data vector. The first is an approach based on raw data from laboratory measurements or their linear transformation. The second is an approach based on derivatives of absorbance (or reflectance) curves.

In order to determine the impact of absorbance vector transformation on the quality of the potential model used to determine the values of soil sample features, optimisation (training) of multilayer perceptron models (MLP) in regression version was performed, in which the input variable was the absorbance value at a given vector point, first derivative of the absorbance and second derivative of absorbance. The output variable was a corresponding soil feature value.

Figures 5–8 show the coefficient of determination values at a specific spectrum point for two versions of input variables: absorbance and first derivative of absorbance. The following conclusions can be drawn from the diagrams:

- a specific absorbance value at a specific spectrum point is not sufficient to build a prediction model of the NIR-soil feature relation in any case,
- the silt and $CaCO_3$ content models characterize, in the entire spectral range, higher values of the determination coefficients with inputs in the form of absorbance than absorbance derivatives.
- seven sets of models (variables: clay content, sand content, pH (measured both, in H_2O and $CaCl_2$), SOC content, N content and CEC) are characterized by higher values of determination coefficients for transformed inputs (absorbance derivatives) than raw data,
- two variables (P and K content) show no significant relationship with the absorbance values.

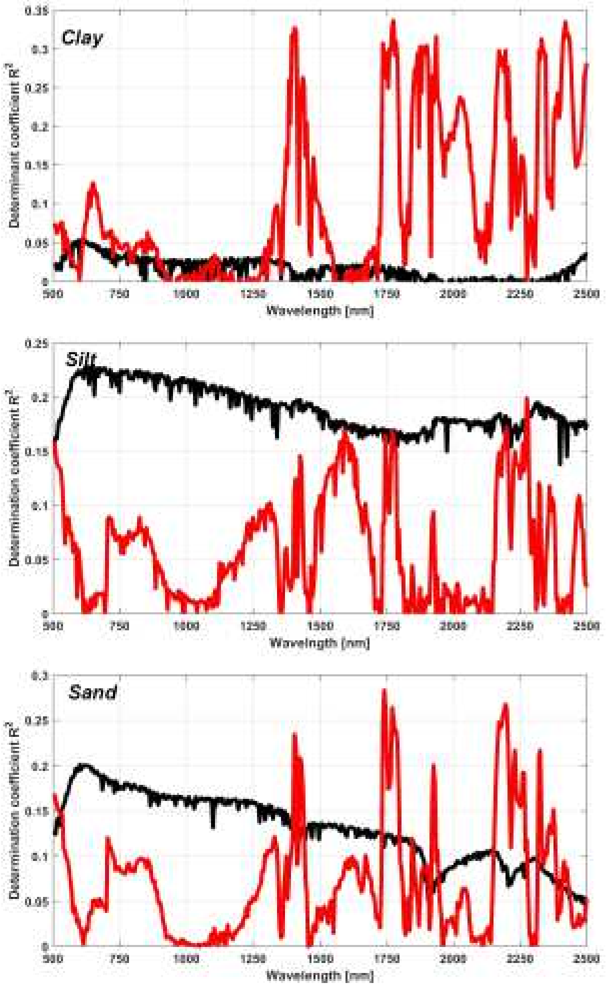


FIGURE 5. Determination coefficients for MLP models with single-input: absorbance value (black line) and absorbance derivative value (red line); output value: variables Clay, Silt, Sand

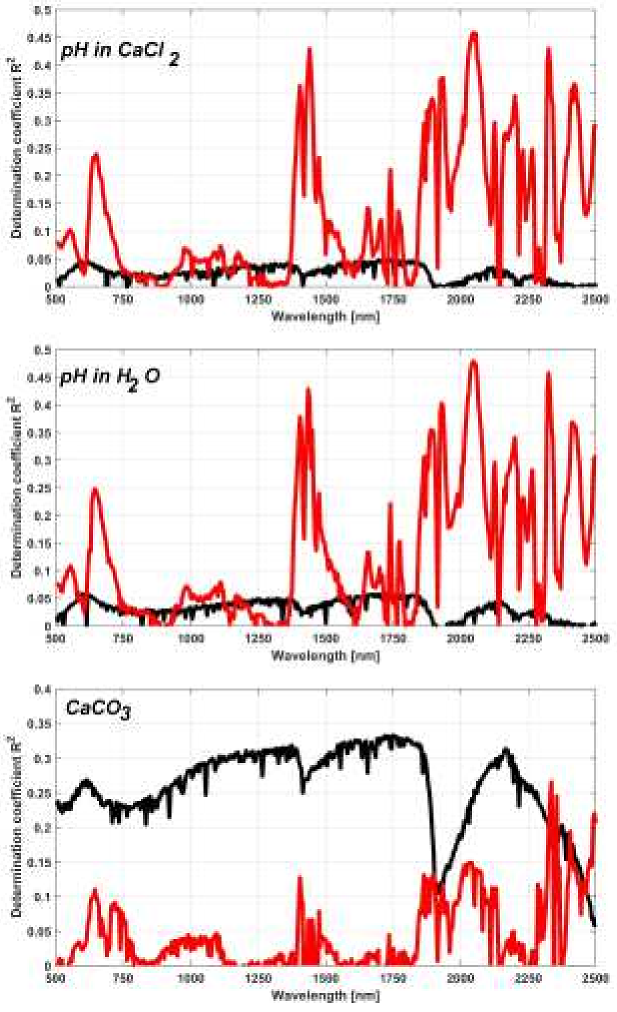


FIGURE 6. Determination coefficients for MLP models with single-input: absorbance value (black line) and absorbance derivative value (red line); output value: pH in $CaCl_2$, pH in H_2O, $CaCO_3$

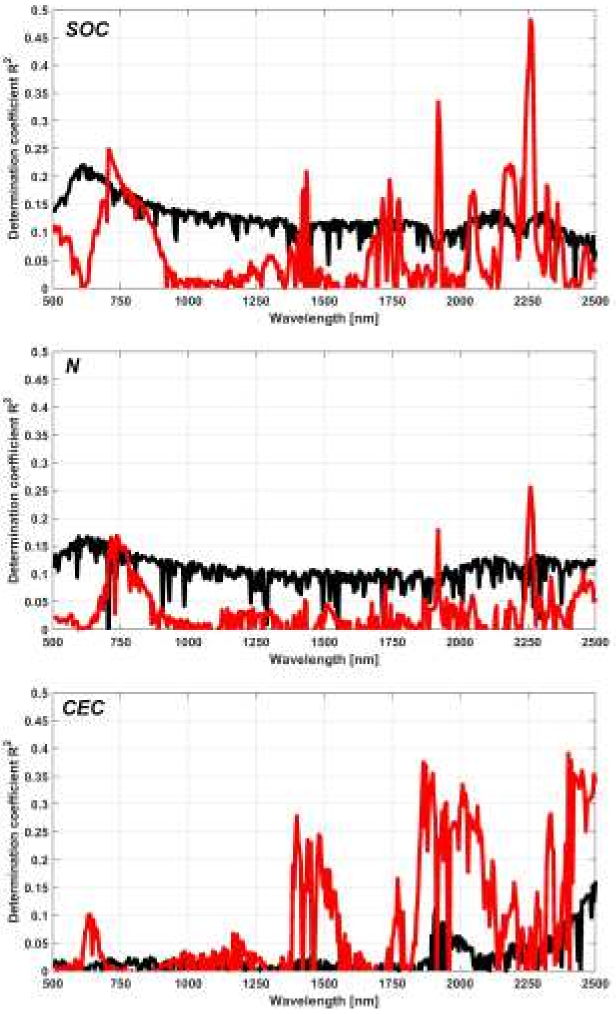


FIGURE 7. Determination coefficients for MLP models with single-input: absorbance value (black line) and absorbance derivative value (red line); output: SOC, N, CEC

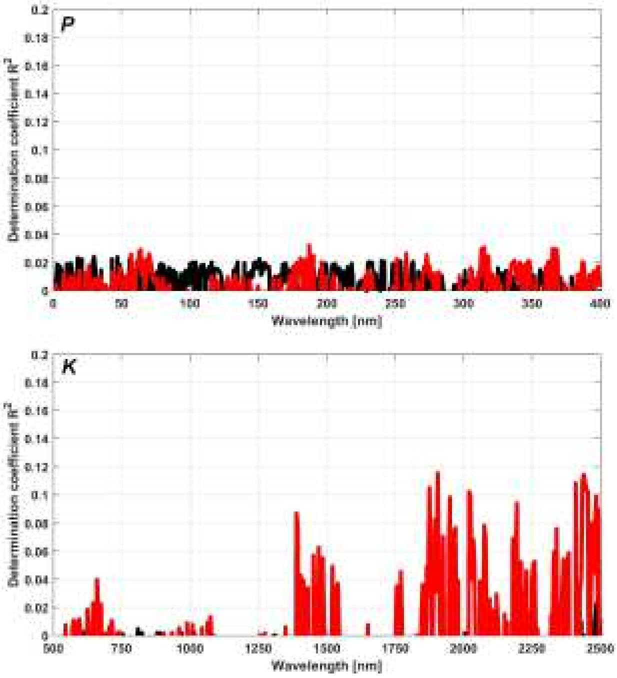


FIGURE 8. Determination coefficients for MLP models with single-input: absorbance value (black line) and absorbance derivative value (red line); output: P and K content

Analysis of the diagrams indicates that the statistical relationship between the NIR spectrum absorbance and soil features is visible more or less clearly at various spectrum transformation methods. Figure 9. and Table 2 illustrate the maximum determination coefficients of the MLP models based on raw data, first derivatives and second derivatives.

The presented results confirm that at varied soil mineralogy, chemical composition and soil texture there are no band f NIR spectral range where the absorbance or reflectance level would unambiguously approximate their relationship with a specific soil indicator (soil feature). Over the entire NIR range, the properties recorded in the LUCAS database impact the spectral response to a lesser or greater degree. In addition, it is difficult to evaluate to what extent the spectral response depends on two or more statistically significant indicators. The analysis of obtained results allows for formulating the following conclusions:

- most analysed soil features have the strongest relationship with the absorbance vector derivative (at different spectrum sections); the exceptions include: $CaCO_3$ and K content (stronger relationship with the second absorbance vector derivative), and Silt (absorbance value),
- P and K contents are weakly correlated with absorbance values,
- the issue of separating the impact of N and SOC content on the spectral response is not totally clear; one can suppose that a potential interpretation of N

TABLE 2. List of maximum determination coefficients of "NIR-feature" MLP models, and wavelength corresponding to the maximum value

Feature	Raw data		First derivative		Second derivative	
	R^2	W(max) [nm]	R^2	W(max) [nm]	R^2	W(max) [nm]
Clay	0.055	610	0.336	1775	0.010	1435
Silt	0.228	630	0.199	2275	0.185	1045
Sand	0.201	600	0.283	1740	0.250	1415
$pHCaCl_2$	0.049	1735	0.458	2045	0.436	1910
pHH_2O	0.058	1725	0.479	2045	0.419	1910
SOC	0.220	615	0.482	2260	0.337	1735
$CaCO_3$	0.333	1750	0.266	2335	0.488	2340
N	0.168	595	0.256	2260	0.146	2280
P	0.024	650	0.032	1435	0.022	540
K	0.023	2485	0.115	1905	0.125	1910
CEC	0.161	2500	0.391	2400	0.381	1895

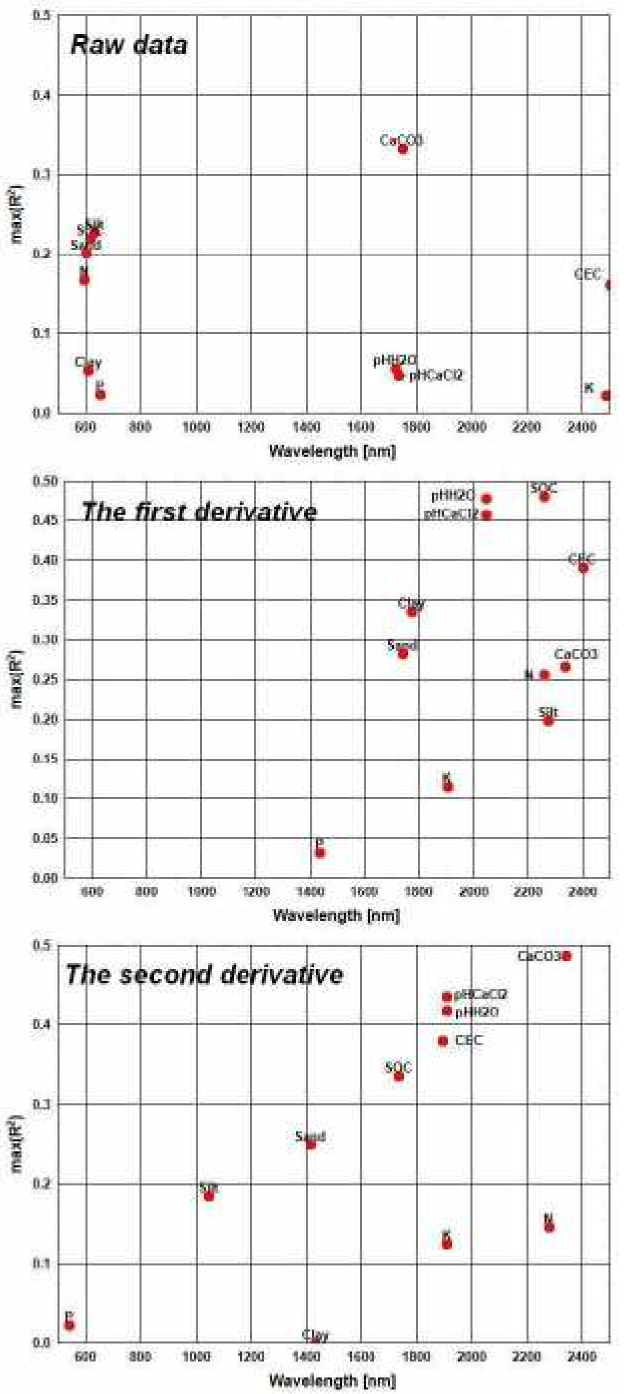


FIGURE 9. Maximum determination coefficients of MLP models, and corresponding wavelength for the raw data, first derivative and second derivative

concentration based on the NIR analysis can be justified by a rather close relationship between SOC and N content.

The most frequently used methods of extracting explanatory variables from the spectral reflection vector include reduction by the stepwise regression algorithm, PCA and PLSR. The following variable extraction steps were taken in order to create a potentially strong set of explanatory variables:

- it was assumed that modelling can be applied to soil features that have the strongest relationship with the absorbance vector: pH in H_2O, SOC, $CaCO_3$ and N content as well as the CEC; the input dataset was optimised relative to this set of features (Figure 10),
- based on the variation diagrams of PLSR transformation resultant components (Figure 11) the following were included in the explanatory dataset: 5 first components obtained with the use of PLSR algorithm from raw data; 15 first components from absorbance derivative data and 5 first components from absorbance second derivative data.

The set of 25 values thus obtained included transformed elements from the data linear transformation that affect the five selected soil features to the greatest degree. The analyzed models were: model ensemble to predict individual soil properties, model MLP prediction of whole properties vector and model with noise reduction of the output vector. Relatively, the best quality indicators were characterized by the model combined with noise reduction of the output vector (Table 3).

The full assessment of these approaches requires that the RMSE be expressed in units not transformed to the 0–1 scale. This recalculation indicates that:

- pH evaluation is modelled with error equal to 0.48 pH unit,
- SOC concentration is modelled with error equal to 9.46 g kg^{-1}

TABLE 3. Statistics of machine learning models with linked inputs: raw data inputs reduced with PLSR algorithm (5 variables), first derivative (15 variables) and second derivative (5 variables). Raw data transformed to 0–1 range

Individual models				
Feature	R^2	RMSE/MAE	St. Dev.	RMSE(MAE)/StDev
pH	0.80	0.088/0.067	0.194	0.45(0.35)
SOC	0.67	0.068/0.044	0.117	0.58(0.38)
$CaCO_3$	0.91	0.041/0.023	0.138	0.30(0.17)
N	0.65	0.054/0.039	0.091	0.60(0.43)
CEC	0.69	0.038/0.027	0.071	0.53(0.38)
Multi-output MLP model				
pH	0.84	0.078/0.060	0.194	0.40(0.31)
SOC	0.74	0.061/0.041	0.117	0.52(0.35)
$CaCO_3$	0.92	0.038/0.024	0.138	0.28(0.17)
N	0.69	0.050/0.035	0.091	0.55(0.38)
CEC	0.72	0.037/0.026	0.071	0.53(0.37)
Denoised MLP model				
pH	0.86	0.073/0.051	0.194	0.37(0.26)
SOC	0.75	0.058/0.039	0.117	0.49(0.33)
$CaCO_3$	0.95	0.031/0.014	0.138	0.23(0.10)
N	0.72	0.048/0.033	0.091	0.53(0.36)
CEC	0.75	0.035/0.024	0.071	0.50(0.34)

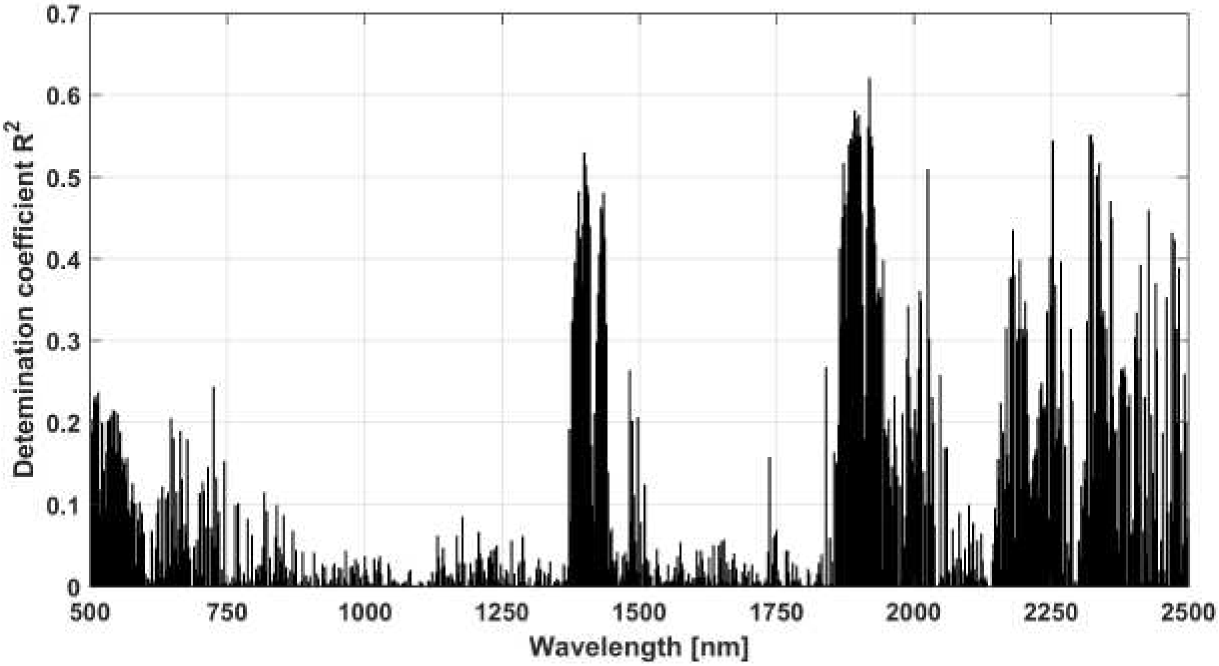


FIGURE 10. Determination coefficient R^2 of MLP models with soil features (pH in H_2O, SOC content, $CaCO_3$ content, N content, CEC) as inputs, and absorbance value at a given wavelength as output

- evaluation of carbonates content ($CaCO_3$) is modelled with error equal to 29.8 g kg^{-1} (in a traditional evaluation method it reaches 3% of weight),
- N content is modelled with error equal to 0.65 g kg^{-1},
- CEC evaluation is modelled with error equal to 4.84 cmol(+) kg^{-1}.

Figure 12 presents histograms of the feature evaluation error distribution by the hybrid model, expressed in standard deviation units and in the same arrangement. The error distributions are generally symmetrical ($CaCO_3$ content is an exception, as it shows a positive asymmetry) in relation to the normal distribution. Figures 13–17 show matrix diagrams of the observed and obtained distributions; they can be considered a rather good illustration of the observed variation of the features distribution.

The analysis of the final modelling result indicates that relatively flexible approximators as machine learning algorithms did not give a fully satisfactory model, which confirms other attempts made in this area (Stevens et al. 2013). The possible reason is probably a mutual disturbance of spectral response by various soil factors, including the factors that are not included in the observation (other structural, mineralogical or chemical features). A thesis can be presented that potentially it is possible to obtain a model slightly better than the one presented in this paper or described in the literature. Thus, it should be confirmed that the NIR analysis – at least presently – is not an alternative to classic soil analysis, which does not mean that it should be disregarded as a source of geospatial data on soils. The cartographic soil image is – at least for now – very generalised (thematic maps), and any attempt to make data nodes more dense, whether in a screen image or in continuous and diffuse images, can be of a significant informational value. It is difficult to overestimate the awareness of spatial soil heterogeneity that, in the form of databases, can be used for diffuse presentation of soils closer to the nature than discrete images of thematic maps.

The spatial range of samples from the LUCAS database includes most European Union countries. Diversity in terms of geology, climate, land morphology and use justifies the supposition that a model based on such data would be universal on a European scale. Estimation errors (RMSE) generated by the models must, however, be considered large, particularly in the lower sections of their potential range. A 1% SOC concentration error, when the SOC content does not exceed 1% (as it is appropriate for most Polish soils), disqualifies the estimation as an insufficiently accurate method. A similar situation is with CEC, and even seemingly better estimation of carbonates content or pH.

LUCAS data were used for methodological studies, mainly related to the prediction of soil organic carbon content, based on NIR analysis. The preliminary research report (Tóth et al. 2013) contains information on the SOC concentration prediction error (RMSE) of 3.6 g kg^{-1} (cropland), 7.2 g kg^{-1}

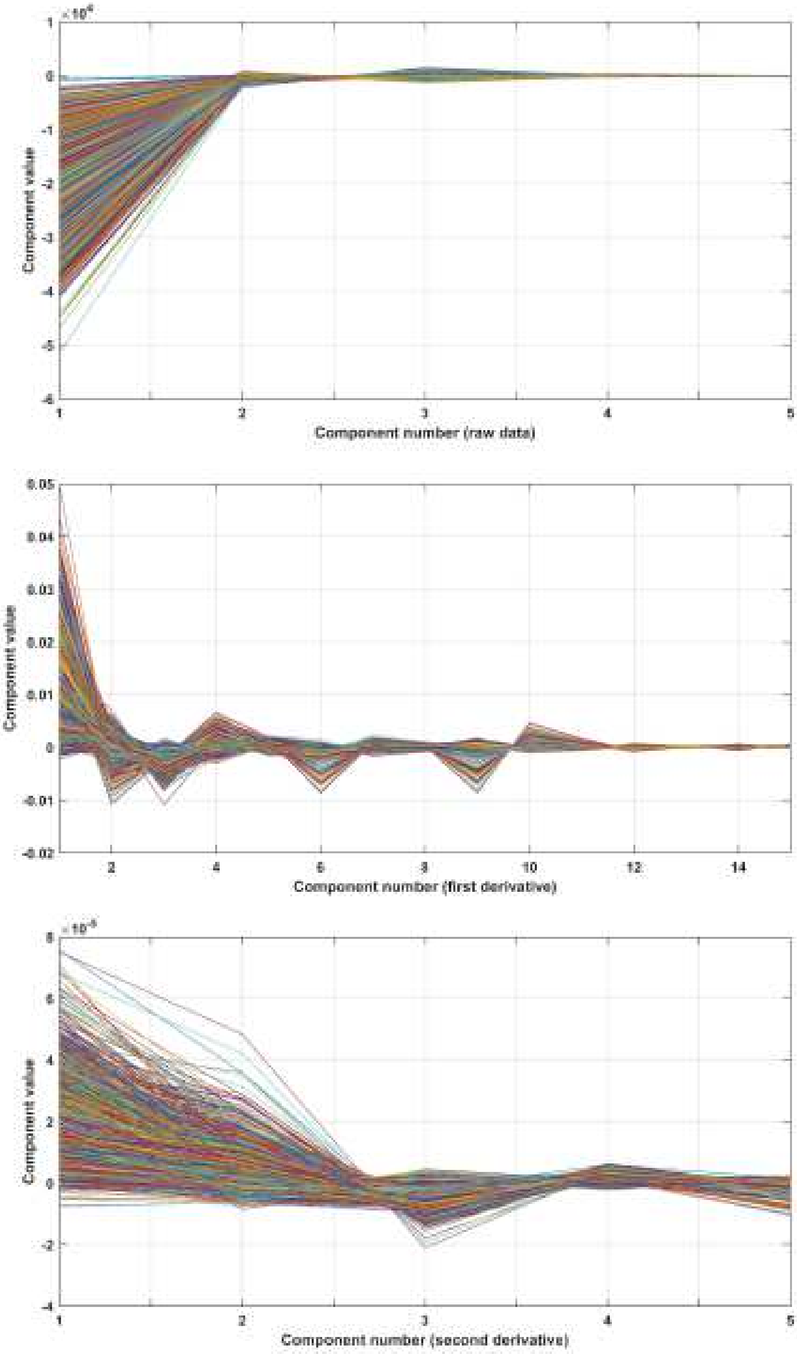


FIGURE 11. Values of the NIR absorbance vector components (5 components), first absorbance vector derivative (15 components) and second absorbance vector derivative (5 components) for LUCAS data as a result of variable transformation with PLSR algorithm

(grassland) to 11.9 g kg^{-1} (woodland). The modeling method used was the combination of the LOCAL algorithm (Shenk et al. 1997) and the methodology of interval partial last squares (I–PLS). In another work using the same data (Stevens et al. 2013), SOC prediction errors of 4.0 g kg^{-1} (cropland), 6.4 g kg^{-1} (grassland), 10.3 g kg^{-1} (woodland) and 7.3 g kg^{-1} (in mineral soils, regardless of land use type) were found. The model of the support vector machines (SVM) were used, with the selection of variables in accordance to the recursive feature elimination. In the paper (Liu et al. 2018) LUCAS data were used to model the content of clay fraction in soils, using a one-dimensional, convolutional neural network (1D–CNN). The RMSE statistics were 8.62% of the clay content, while the RPD statistics = 1.54.

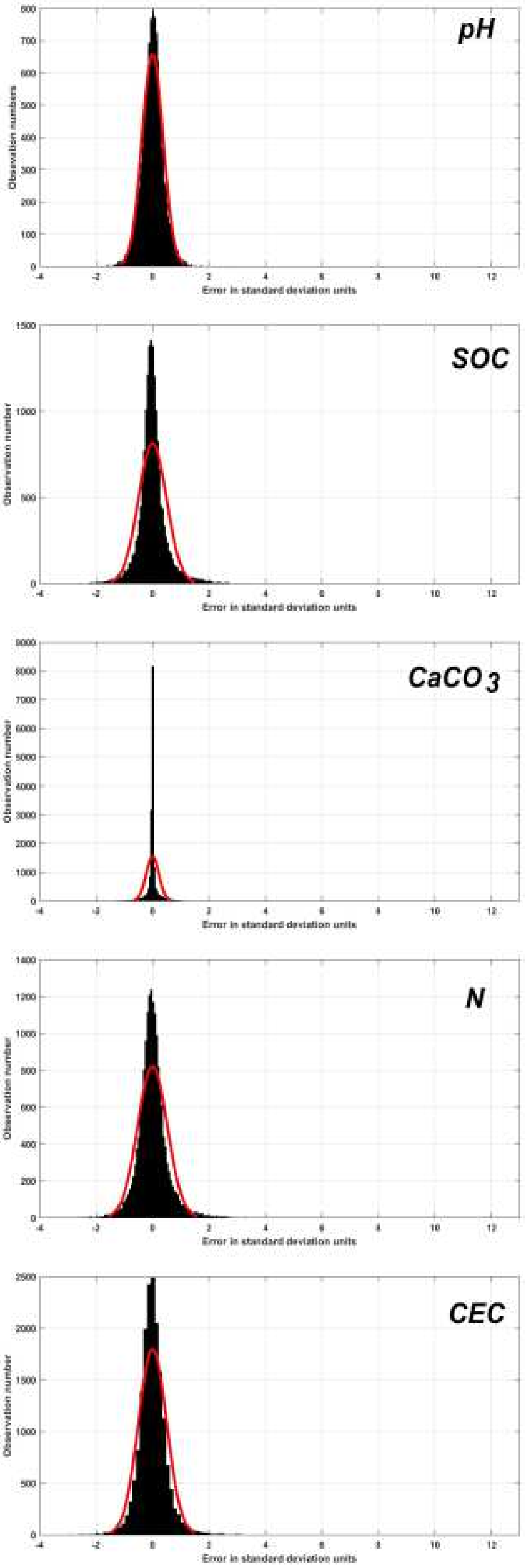
In another work by the same authors (Liu et al. 2017), the combination of PLSR algorithms (variable selection) and decision trees (Gradient Boosting Machine) was used to build a combined model. The SOC prediction model was characterized by RMSE of 6.8 g kg^{-1} (cropland), 10.9 g kg^{-1} (grassland) and 13.31 g kg^{-1} (woodland). The same statistics for the prediction N were: 0.42 g kg^{-1} (cropland), 0.82 g kg^{-1} (grassland) and 0.78 g kg^{-1} (woodland). RMSE prediction of clay content was within 5.1–6.2%.

TABLE 4. Statistics of machine learning models with linked inputs: raw data inputs reduced with PLSR algorithm (5 variables), first derivative (15 variables) and second derivative (5 variables), and with the addition of a quality variable: mechanical group according to FAO. RPD=St.Dev./RMSE

Denoised MLP model + additional input (mechanical group)					
Feature	R^2	RMSE/MAE	St. Dev.	RPD	RMSE(MAE)/ StDev
pH	0.86	0.068/0.054	0.194	2.85	0.35(0.28)
SOC	0.83	0.046/0.031	0.117	2.56	0.39(0.26)
$CaCO_3$	0.96	0.026/0.012	0.138	5.26	0.19(0.09)
N	0.82	0.037/0.026	0.091	2.43	0.41(0.29)
CEC	0.83	0.027/0.019	0.071	2.63	0.38(0.27)

Note that digital cartography is based partially on the existing soil documentation. Some information – usually generalised – is available without extra costs. This kind of information includes soil texture. It can be assumed that adding a qualitative variable texture class to the set of explanatory variables can increase the model quality. The results of such an attempt are presented in Table 4, which in terms Dzień dobry nie brakuje mi of contents is analogous to Table 3. A slight model improvement can be seen,

FIGURE 12. Comparison of distribution of standardised errors of pH in H_2O, SOC, $CaCO_3$ and N and CEC evaluation; red line indicates normal distribution



which can be summarized as the following conclusions:

- pH evaluation is modelled with error equal to 0.45 pH unit,
- SOC concentration is modelled with error equal to 7.5 g kg^{-1} (0.75% of weight in a traditional evaluation method),
- evaluation of carbonates content (CaCO^3) is modelled with error equal to 24.6 g kg^{-1} (in a traditional evaluation method it reaches 2.5% of weight),
- N content is modelled with error equal to 0.5 g kg^{-1},
- CEC evaluation is modelled with error equal to 3.7 cmol(+) kg^{-1}.

Each regression model estimates the conditional explained variable in a specific configuration of explanatory variables. In the presented case, the regression procedure includes five explained variables estimated in two stages: in stage one, the explanatory variables are selected, NIR spectrum absorbance transformed with the PLSR algorithm, first and second derivatives transformed with the PLSR algorithm and a quality variable represented by mechanical group according to the ISO classification. The modelling result can be considered statistically satisfactory, however less satisfactory from the point of view of substantive estimation quality. Very important is the observation that the NIR analysis is not competitive as a universal alternative to classic laboratory analysis, which is a reference point. This does not eliminate its competitiveness in conditions of limited areas, homogenous in terms of geology and climate (Debaene et al. 2014)

The assessment of the obtained results from the point of view of documenting the spatial soil variation – at least in the range of the vector of 5 modelled properties – should be, however, a bit different. The model algorithm generates a total estimation of the properties vector based on transformed NIR vectors and qualitative information on the grain size group. The RMSE values and values of RMSE quotient and observed standard deviation are included in Table 4. Comparing the properties vectors, we can discuss their mutual proximity or distances. Figure 18 shows the statistical distribution of Euclidean distances of observed and modelled soil features expressed in standard deviation units. The solid line approximates the histogram of these distances with log-normal distribution, with a mean of 0.69 and standard deviation of 0.465. Assuming the log-normal distribution

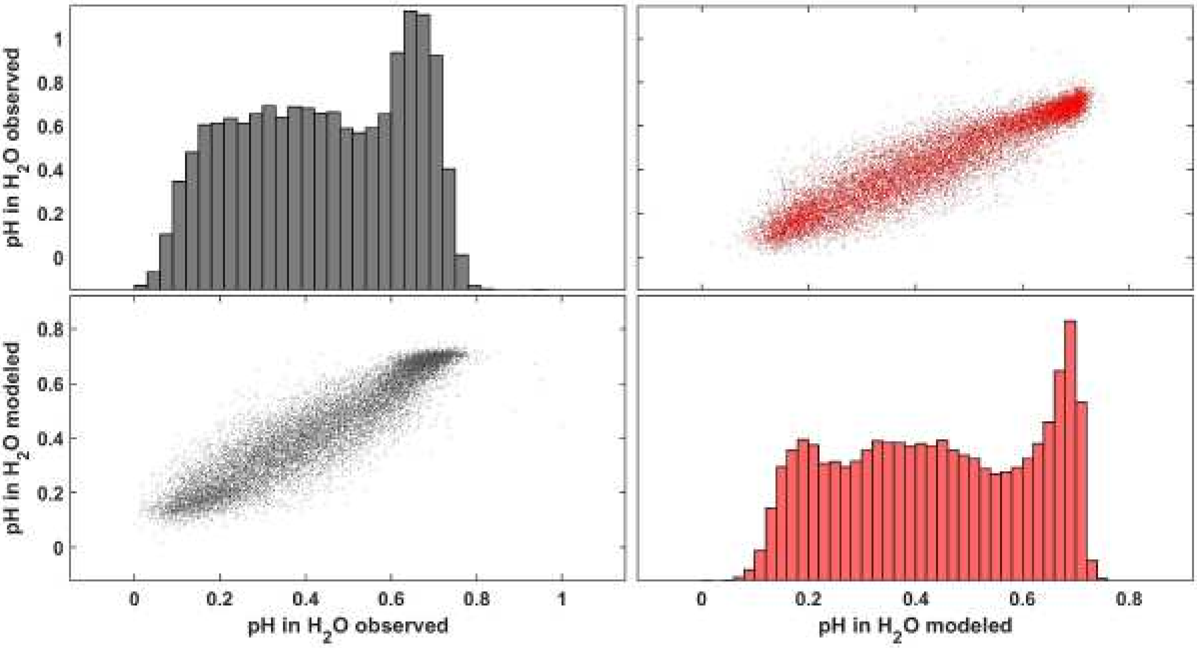


FIGURE 13. Observed and modelled distribution and dot diagram of observed and modelled relationships: pH in H_2O

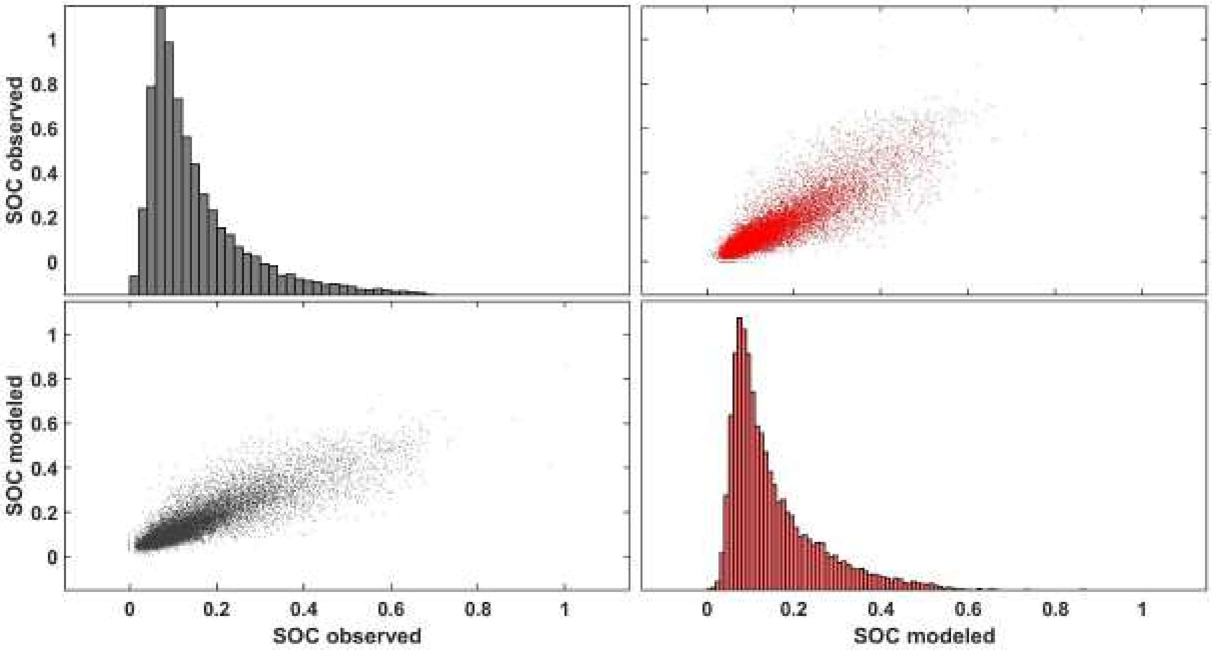


FIGURE 14. Observed and modelled distribution and dot diagram of observed and modelled relationships: SOC concentration

as reliable for the distance of the observed and modelled vectors, and assuming the tolerance for mean property distance at maximum one standard deviation, this condition is satisfied by 98% of the modelled data; when the tolerance is reduced to 1/2 of the standard deviation, the number of sufficiently close cases is limited to 86%; a further tolerance reduction to 1/2 of standard deviation decreases the share to 48%. The decision concerning what tolerance to use should be substantiated in terms of expected credibility of documentation. Considering the total absence of such information in databases, for most of the Poland's territory one can say that even a universal model based on such data would be sufficient for thickening and, in a significant part, documenting of the grid of soil property observations using the NIR. One can expect that development of regional models, accounting for uniformity of geological conditions, would allow a significant limitation of estimation errors of soil

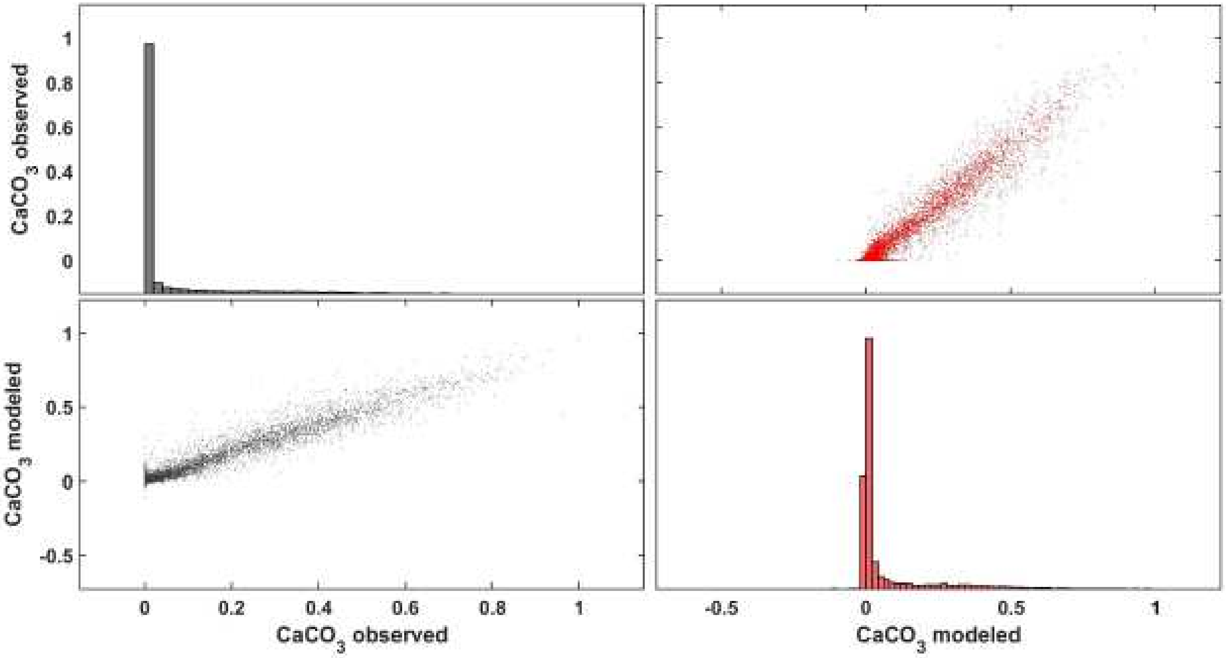


FIGURE 15. Observed and modelled distribution and dot diagram of observed and modelled relationships: $CaCO_3$ concentration

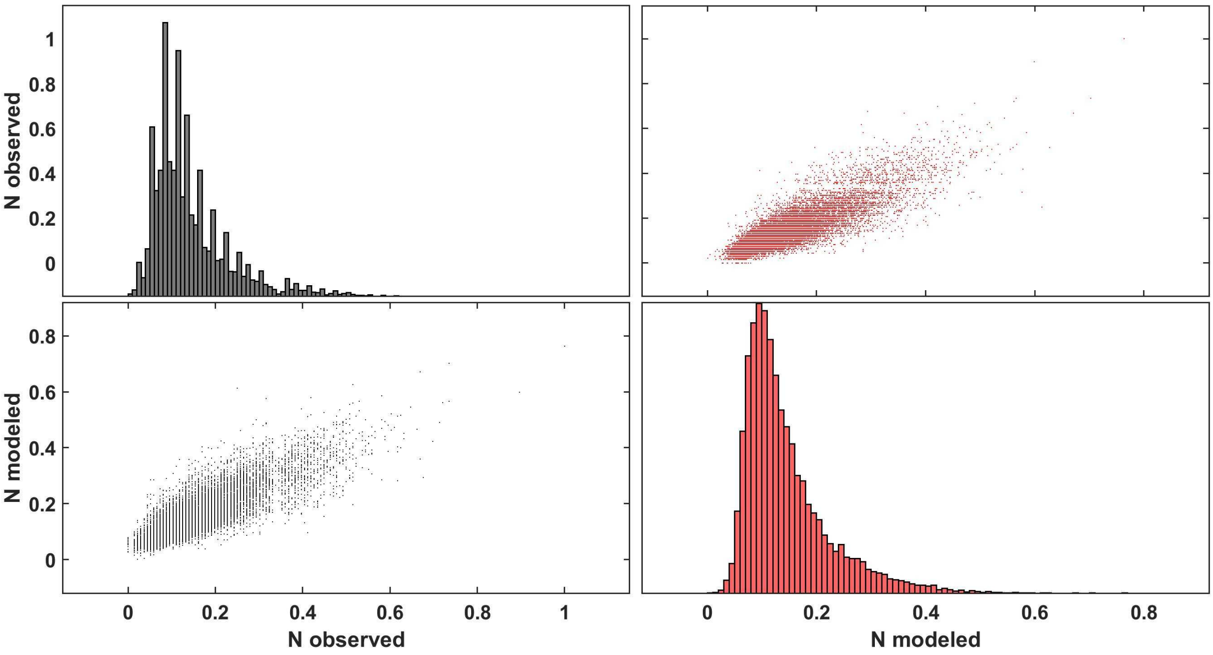


FIGURE 16. Observed and modelled distribution and dot diagram of observed and modelled relationships: N content

property vectors, without the need for full sampling and laboratory analysis from large areas.

Considering the problem of a universal model (limiting it even only to mineral soils), the use of spectral response as a source of information on soil properties, allows their estimation with a relatively significant error. From the comparison of available results, it can be concluded that statistical models and machine learning are characterized by a similar level of uncertainty, while the method of building a prediction model is of less importance. Improvement of prediction occurs as a result of data clustering: creation of libraries of spectra with similar features (LOCAL method), development of separate models for different land management, or inclusion of additional variables (soil texture). The prediction error is important especially in the case of strongly asymmetric distribution of properties: according to LUCAS databa-

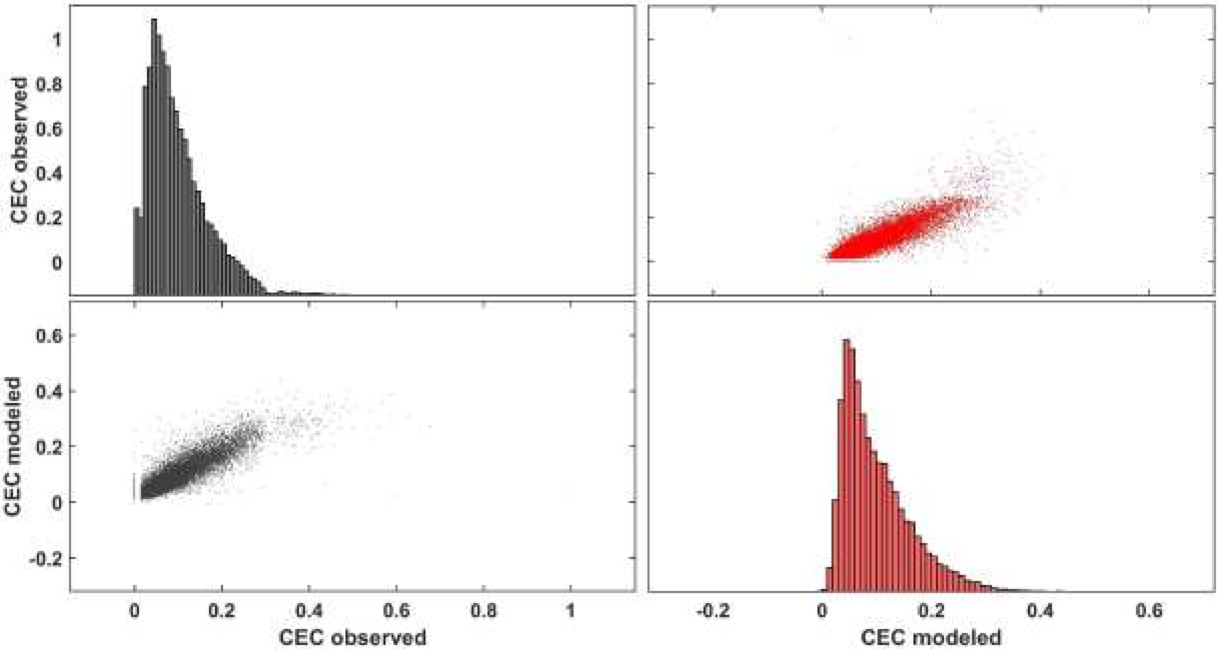


FIGURE 17. Observed and modelled distribution and dot diagram of observed and modelled relationships: CEC values

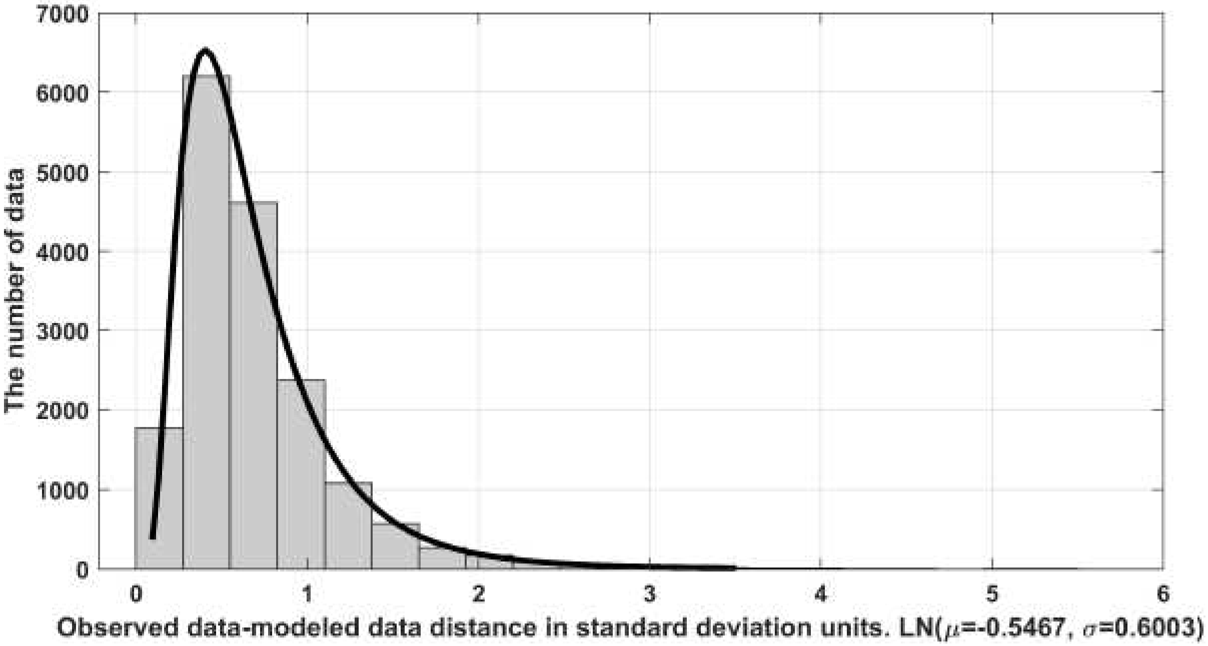


FIGURE 18. Distribution histogram of Euclidean distances of observed and modelled 5 soil features expressed in standard deviation units; the solid line is the log-normal distribution curve

se data, in Poland the average organic carbon content in mineral soils is 16.9 g kg^{-1} (with standard deviation 14.4) and the significant area of the sandy soils contains less than 10 g kg^{-1} organic carbon. In proportion to this value, an estimation error of 5–10 g kg^{-1} is the important problem. The clay average content in Poland's soils is 8.9% (with a standard deviation of about 7.7). Therefore, similar RMSE level makes it difficult to accept estimates of NIR-based models.

Similar quantitative relation are characterized by N content (mean 1.5 g kg^{-1}). The particular situation is related to the prediction of the of $CaCO_3$ concentration, which is scarce in the soils of Poland (average only 3.7 g kg^{-1} with a standard deviation exceeding 19). Coefficient of determination of the R^2 model =

0.96, which, however, results from a very strongly asymmetrical distribution of $CaCO_3$ content, because the model's RMSE reaches 25 g kg^{-1}. It can be assumed that a sufficient level of reliability of the estimation of the pedon's properties based on models based on NIR analysis can be obtained by applying repetitions of determinations in accordance with the acceptable assessment error and local variability of the soil characteristics.

CONCLUSIONS

1. The absorbance is affected by a limited number of tested soil features: pH, texture, content of carbonates, SOC and N, as well as the CEC, while P and K content has a negligible impact.
2. It can be noticed that transformation by absorbance differentiation, selection of variables (e.g. with the use of PLSR) in the case of some features leads to the model's improvement. The prediction result is improved after adding the texture class as a model explanatory variable.
3. Development of an universal model, analogous to methodologies used in chemical or pharmaceutical industry, does not seem possible because of noise created by factors that are not being observed. It is hard to assume that the NIR spectral analysis could become an alternative to "wet" methods without a significant expansion of the list of analysed factors.
4. The NIR methodology can be suitable in conditions of limited soil variation and particularly in development of thematic (legacy) soil maps.
5. None of the tried-and-tested prediction models (algorithms) of soil properties provide sufficient accuracy for their estimation. The low cost of estimating the vector of soil traits, however, allows the use of several repetitions of NIR spectra within the pedon range, which will reduce this error.

ACKNOWLEDGEMENTS

The LUCAS topsoil dataset 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

Ballabio C., Panagos P., Montanarella L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261: 110–123.

Bengio Y., 2009. Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 2 (1): 1–127.

Brevik E. C., Calzolari C., Miller B. A., Pereira P., Kabala C., Baumgarten A., Jordán A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264: 256–274.

Conforti M., Matteucci G., Buttafuoco G., 2018. Using laboratory Vis-NIR spectroscopy for monitoring some forest soil properties. Journal of Soils and Sediments 18(3): 1009–1019

Debaene G., Niedźwiecki J., Pecio A., Żurek A., 2014. Effect of the number of calibraion samples on the prediction of several soil properties at the farm-scale. Geoderma 214–215: 114–125.

Fang Q., Hong H., Zhao L., Kukolich S., Yin K., Wang C., 2018. Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A Review. Journal of Spectroscopy Article ID 3168974, 14 pages, https://doi.org/10.1155/2018/3168974https://doi.org/10.1155/2018/3168974

Fuentes M., Hidalgo C., González-Martín I., Hernández-Hierro J.M., Govaerts B., Sayre K.D., Etchevers J., 2012. NIR spectroscopy: an alternative for soil analysis. Communications in Soil Science and Plant Analysis 43(1–2): 346–356.

Kania M., Gruba P., 2016. Estimation of selected properties of forest soils using near-infrared spectroscopy (NIR). Soil Science Annual 67(1): 32–36.

Kokaly R.F., Clark R.N., Swayze G.A., Livo K.E., Hoefen T.M., Pearson N.C., Wis, R.A., Benzel W.M., Lowers H.A., Driscoll R.L., and Klein A.J., 2017. USGS Spectral Library Version 7. U.S. Geological Survey Data Series 1035: 61 p., https://doi.org/10.3133/ds1035

Liu L., Ji M., Buchroithner M., 2017. Combining Partial Least Squares and the Gradient-Boosting Method for soil property retrieval using visible near-infrared shortwave infrared spectra. Remote Sensing 9: 1299, 10.3390/rs9121299.

Liu L., Ji M., Buchroithner M., 2018. Transfer learning for soil spectroscopy based on Convolutional Neural Networks and its application in soil clay content mapping using hyperspectral imagery. Sensors (Basel) 18(9): 3169, doi:10.3390/s18093169

McBratney A.B., Mendonça Santos M.L., Minasny B., 2003. On digital soil mapping. Geoderma 117(1–2): 3–52.

McBratney A.B., Minasny B., Cattle S. R., Vervoort W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109(1–2): 41–73.

Mohamed E.S., Saleh A.M., Belal A.B., Abd_Allah Gad, 2018. Application of near-infrared reflectance for quantitative assessment of soil properties. The Egyptian Journal of Remote Sensing and Space Science 21(1): 1–14.

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

Shenk J.S., Westerhaus M.O., Berzaghi P.J., 1997. Investigation of a LOCAL calibration procedure for near infra-red instruments. Journal Near Infrared Spectroscopy 5: 223–232.

Shi Z., Ji W., Viscarra Rossel R.A., Chen S., Zhou Y., 2015. 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 66: 679–687. doi:10.1111/ejss.12272

Stenberg B., Viscara Rossel R.A., Mounem Mouazen A., Wetterlind J., 2010. Visible and near infrared spectroscopy in soil science. [In:] Advances in Agronomy (Sparks D.L. Editor),107, Burlington: Academic Press: 163–215.

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

Tóth G., Jones A., Montanarella L., 2013. LUCAS Topsoil Survey. Methodology, data and results. JRC Technical Reports. Luxembourg. Publications Office of the European Union, EUR26102 – Scientific and Technical Research series – ISSN 1831-9424 (online).

Veres M., Lacey, G., Graham W.T., 2015. Deep Learning Architectures for Soil Property Prediction. Proceedings – 2015 12th Conference on Computer and Robot Vision, CRV 2015: 8–15.

Website 1: http://docs.h2o.ai

Wetterlind J., Stenberg B.,Viscarra Rossel R.A., 2013. Soil analysis using visible and near infrared spectroscopy. [In:] Plant Mineral Nutrients: Methods and Protocols. (Maathuis F.J.M. editor), New York: Humana Press, Springer: 95–107.

Qiu X, Zhang L., Ren Y., Suganthan P. N., Amaratunga N., 2014. Ensemble deep learning for regression and time series forecasting. IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL), Orlando, FL, 2014: 1–6.

Zhang Y., MinZan Li, LiHua Zheng, Yi Zhao, Xiaoshuai Pei, 2016. Soil nitrogen content forecasting based on real-time NIR spectroscopy. Computers and Electronics in Agriculture 124: 29–36.

Received: November 10, 2018
Accepted: September 18, 2019
Associated editor: Ł. Mendyk

Predykcja właściwości gleb modelami uczenia maszynowego na podstawie odpowiedzi spektralnej prób glebowych w zakresie bliskiej podczerwieni

Streszczenie: Analiza właściwości gleby przy użyciu metod pośrednich, pozwala na zwiększenie gęstości punktów obserwacji gleby, co jest konieczne, dla uszczegółowienia cyfrowych map gleb. Jedną z podstawowych metod przyspieszania i redukcji kosztów analizy gleby jest wykorzystanie odpowiedzi spektralnej próbek gleby w warunkach laboratoryjnych. Problem w tej metodzie polega na określeniu zależności między kształtem odpowiedzi spektralnej gleby a właściwościami fizycznymi lub chemicznymi gleby. Baza danych gleby LUCAS zebrana przez centrum badawcze ESDAC UE jest dobrym materiałem do analizy zależności między właściwościami gleby a reakcją spektralną NIR. Modelowanie opisane w artykule opiera się na tych danych. Analizę wpływu konfiguracji właściwości gleby na poziomy absorbancji w różnych zakresach spektrum NIR przeprowadzono za pomocą modeli regresji krokowej. Analizę korelacji cząstkowej wartości cech gleb z wartościami absorbancji i pochodnej absorbancji w całym zakresie spektralnym przeprowadzono w celu oceny wpływu transformacji wektora absorbancji (pierwszej pochodnej wektora absorbancji) na zmianę istotności związku z wartościami właściwości. Modele MLP wykorzystano do oszacowania zależności absorbancji z cechami pojedynczej gleby. Przeprowadzono także modelowanie właściwości gleby w oparciu o algorytm selekcji i transformacji wartości surowych oraz pochodnych absorbancji pierwszej i drugiej, wraz z oceną przydatności takich modeli w budowaniu cyfrowych map gleby. Na kształt krzywej absorbancji ma wpływ ograniczona liczba badanych cech gleby: pH, uziarnienie, zawartość węglanów, SOC, N i CEC; Zawartość P i K ma znikomy wpływ. Metodologia NIR może być odpowiednia w warunkach ograniczonej zmienności gleb, a zwłaszcza w opracowywaniu tematycznych map glebowych.

Słowa kluczowe: baza LUCAS, spektroskopia bliskiej podczerwieni, modele uczenia maszynowego, predykcja właściwości gleb