INFRASTRUKTURA I EKOLOGIA TERENÓW WIEJSKICH
INFRASTRUCTURE AND ECOLOGY OF RURAL AREAS

No II/1/2019, POLISH ACADEMY OF SCIENCES, Cracow Branch, pp. 95-106
Commission of Technical Rural Infrastructure

DOI: https://doi.org/10.14597/INFRAECO.2019.2.1.008

INFORMATION POTENTIAL OF THE SPECTRAL RESPONSE OF POLISH SOILS, IN THE NIR RANGE, IN THE LIGHT OF LUCAS DATABASE ANALYSES. SOIL PROPERTIES VECTOR MODEL

Stanisław Gruszczynski
AGH University of Science and Technology in Krakow

Abstract

The paper presents simple machine learning models used for prediction of some soil properties based on the NIR spectral response. Data on mineral soils from Poland were taken from the LUCAS dataset. Machine learning model was used that is included in the category of so-called multilayer perceptron (MLP). The MLP model input was a vector of combined, transformed inputs made by means of the PLSR (partial last squares regression) algorithm (45 inputs in total). The output was a vector of properties, reduced to 9 components due to poor modelling effects of the P and K components. The estimation errors for texture, soil organic carbon (SOC) and carbonates can be considered acceptable from the point of view of their suitability in the development of cartographic documentation. It can be supposed that further regionalization will improve these results.

Keywords: near infrared spectroscopy, soil properties prediction, machine learning model

INTRODUCTION

The quantitative evaluation of environment components using indirect methods requires interpretation keys that allow extracting the sought-after infor-

mation from remote observations, or alternatives to the "wet" laboratory observations. With the popularization of laboratory and field spectrometers, attempts are being made in many countries to use spectral measurements instead of traditional, more costly and time consuming laboratory technique. For many years the United States Geological Survey (USGS) has been recording and publishing data on reflectance of various materials, natural and artificial, in various spectral ranges, obtained as a result of laboratory, field and aero-photography studies (Kokaly *et al*. 2017). Thousands of materials were subjected to spectrophotometric analysis: minerals, soils, liquids, organic substances, structural and biological materials, etc., creating an extensive spectral data library with indexes and quantitative chemical characteristics of analysed materials. Identification and quantitative determination of sample features based exclusively on graphical analysis of reflected spectrum has not been possible to date, mainly due to the lack of deterministic models of the shaping of the spectral response by various materials. Thus, the appropriate spectrum interpretation algorithms have to be searched for empirically, allowing an indirect determination of sample features without the use of wet methods. The effectiveness of such modelling depends on heterogeneity of physical and chemical properties of materials which is a potential source of spectral response disturbance in terms of a property being determined.

Soils are mixtures of materials of different grain distribution, organic parts of varying chemical composition, soil solutions, etc. Probably, the soils are the most difficult materials for spectral analysis. Attempts are being made to develop models allowing the cost and labour intensity reduction of soil features determination (Conforti *et al.* 2018, Fuentes *et al.* 2012, Kania and Gruba 2016, Stenberg *et al.* 2010). A success in this field would give a tool allowing a substantial increase of soil sampling, leading to a real, continuous picture of soils variation instead of a discrete image (McBratney *et al.* 2002, McBratney *et al.* 2003). Most often, the spectral analysis range covers the wavelengths from 700 to 2500 nm, including the near infrared (NIR). The spectrometers analysing the 400-2500 nm spectrum are also used (a part of the visible range and NIR (VIS-NIR). The idea of using the spectral response relation models and soil properties vector results from the aim to replace the contour soil thematic maps with their continuous versions, more closely reflecting the actual variation of soil environment (Mohamed *et al*. 2018). Theoretically, one can consider the universal models of the aforementioned relation, and also – more probably – regional or local models. The goal of the paper is to estimate, based on the samples from Poland included in the LUCAS database, the possibility of developing a useful model for prediction of some mineral soil properties, based on the transformed samples absorbance spectrum in the near infrared and to evaluate the possible use of such models to estimate the soil properties for cartographic documentation.

MATERIAL AND METHODS

Within the framework of the soils documentation projects in the EU countries, since 2009 the European Soil Data Centre (ESDAC) has been conducting the Land Use/Cover Area Frame Statistical Survey (acronym LUCAS). In addition to the location of topsoil samples taken from the depth of 0-30 cm, typology, use and other classification details, the dataset made available by ESDAC includes the laboratory determinations of the grain size (percent of coarse fraction, sand, silt and clay) pH in $CaCl_2$, pH in H_2O, soil organic carbon (SOC), $CaCO_3$, N, P and K, and cation exchange capacity (CEC). In addition, each soil sample is described with the absorbance vector in the 400-2500 nm range (VIS-NIR), in 0.5 nm increments (4200 values), laboratory measured under identical conditions (Orgiazzi *et al.* 2017, Toth *et al.* 2013).

Table 1. Statistics of soil samples from Poland included in the LUCAS database

Variable	Statistics	
	Mean	Standard deviation
Clay [%]	8.9	7.68
Silt [%]	26.9	21.42
Sand [%]	64.1	26.43
pH.in.$CaCl_2$	5.1	1.12
pH.in.H_2O	5.7	1.06
SOC [$g \cdot kg^{-1}$]	16.9	14.43
$CaCO_3$ [$g \cdot kg^{-1}$]	3.7	19.08
N [$g \cdot kg^{-1}$]	1.5	1.07
P [$mg \cdot kg^{-1}$]	37.5	27.02
K [$mg \cdot kg^{-}$]	103.4	102.24
CEC [$cmol(+) \cdot kg^{-1}$]	7.9	7.40

Among about 20 thousand soils samples from the EU countries, there are 1589 samples from Poland, mostly from the superficial layer of mineral soils, only 21 samples are from organic formations. Table 1 presents the statistics of 1568 samples of mineral soils that form a substrate of cropland, permanent grassland and woodland. The characteristic features of this dataset include:

- very high variability of properties, except pH; in case of content of clay, silt, SOC, CaCO3, N, P, K and CEC the standard deviation is close (for carbonates, it significantly exceeds) the mean;
- relatively high sand content;
- domination of acidic pH;

- low average carbonates content (less than 0.5% by weight);
- relatively low cation sorption capacity;
- low average organic carbon content.

Table 2. Matrix of linear correlation coefficients between the properties values in the sample; data from samples from Poland (significant coefficients at the α=0.05 significance level are highlighted)

Variables	Clay	Silt	Sand	pH ($CaCl_2$)	pH (H_2O)	SOC	$CaCO_3$	N	P	K	CEC
Coarse	**0.28**	**0.15**	**-0.21**	**0.06**	**0.06**	**0.10**	**0.12**	**0.12**	**-0.07**	**0.12**	**0.13**
Clay		**0.54**	**-0.74**	**0.36**	**0.35**	**0.25**	**0.29**	**0.39**	**-0.11**	**0.46**	**0.58**
Silt			**-0.97**	**0.26**	**0.25**	0.00	**0.08**	**0.10**	**-0.08**	**0.33**	**0.22**
Sand				**-0.32**	**-0.31**	**-0.07**	**-0.15**	**-0.20**	**0.09**	**-0.40**	**-0.35**
pH ($CaCl_2$)					**0.99**	-0.03	**0.29**	**0.14**	0.00	**0.39**	**0.41**
pH (H_2O)						**-0.06**	**0.28**	**0.09**	-0.02	**0.38**	**0.38**
SOC							**0.15**	**0.93**	**-0.08**	0.02	**0.64**
$CaCO_3$								**0.22**	**-0.03**	**0.16**	**0.26**
N									**-0.05**	**0.09**	**0.76**
P										**0.29**	**-0.07**
K											**0.21**

Table 2 presents the linear correlation coefficients between individual properties of soil samples. Some coefficients indicate a rather strong statistical correlation of some properties (Clay-CEC, SOC-CEC, N-CEC, N, SOC), few correlation coefficients are not statistically significant (SOC-pH, P-pH, SOC-K), the majority are statistically significant, also due to a rather large sample size.

The main problem of the NIR data-based prediction is the spectral response vector size (radiation absorbance vector in individual points of the VIS-NIR range) (Wetterlind *et al.* 2013). The number of vector components usually exceeds one thousand, and is 4200 in case of the LUCAS database. Regardless of the algorithm and the model architecture, the number of parameters subjected to optimization may exceed the number of observations. Due to a strong linear correlation between the absorbance vector components, the generally used method to reduce the number of vector inputs is the PCA algorithm or the Partial Least Squares Regression (PLSR) (Liu *et al.* 2017). Probably, there are other methods to reduce the dimensionality.

The method of searching for an appropriate soil properties estimation model based on the NIR spectral response is not standardized, mainly because of differentiation of spectral properties of the mixture of minerals, mineral and organic

substances, each of which has its own, usually unknown, radiation absorption characteristics. The problem lies in extraction of input information necessary for prediction. It can be raw data obtained directly as a result of reflectance determination, first derivatives of raw data spectra, or second derivatives (Figure 1.). The generalization of the radiation absorption spectrum with the use of Savitzky-Golay filter is also applied. The dimensionality reduction method widely used in industrial laboratories for the analysis of highly homogenous materials (raw materials, chemical products, pharmaceuticals) is the Partial Least Squares Regression (PLSR). It combines a few advantageous functions, orthogonalizes the input data and reduces their number, maintaining the main components of relationships with the modelled variables. It performs well also in case of low variability of soil material (Kania and Gruba 2016).

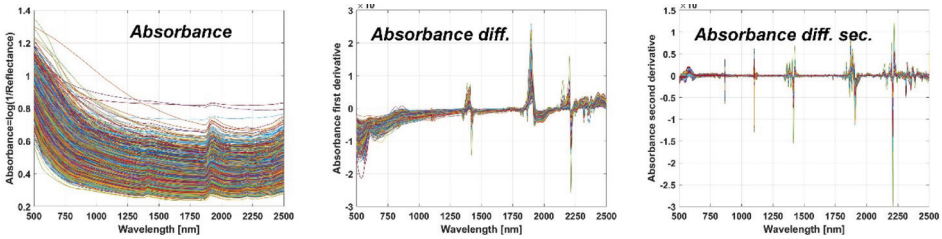

Figure 1. Graphs of absorbance vectors and their transformations used in the spectral response analyses: first and second derivatives. The absorbance is a logarithmic transformation of reflectance in order to linearize the spectral response relationship

The important problem is the possibility of predicting individual soil properties. Some properties do not have a significant impact on the spectral response, or the response is masked by other factors; the possibility of predicting other features can result from the statistical relationships between various properties (e.g. SOC and N, pH and CaCO3, SOC, and clay content and CEC).

The classical, dominating approach to the "NIR-Soil properties" modelling is the use of PLSR – *Partial Least Squares Regression* (Liu *et al.* 2017, Shi *et al.* 2015). The PLSR is an extension of traditional multiple regression, particularly widely applied in case of a smaller amount of data than the number of potential vector inputs (references). The algorithm is included in many statistical and calculation packages. These algorithms usually provide also input data transformation coefficients, involving the data orthogonalization and reduction. These calculations used the algorithm included in the MATLAB software package, reducing the number of input components to 15 (after transformation). The PLSR algorithms inputs included: absorbance vector obtained during the laboratory measurements of samples from Poland, absorbance vector derivatives, and

the second absorbance vector derivatives. The model output was the vector of modelled properties values.

In addition to the PLSR methods, a machine learning model was used that is included in the category of so-called multilayer perceptron (MLP). The MLP model input was a vector of combined, transformed inputs made by means of the PLSR algorithm (45 inputs in total). The output was a vector of properties, reduced to 9 components due to poor modelling effects of the P and K components. The MLP hidden layer consisted of 20 units with a tangent transfer function. The algorithm included in the MATLAB package was used. The results given below concern only the validation data (235 observations) selected from the dataset.

RESULTS

Table 3 includes coefficients of determination (R^2) and root mean square errors (RMSE) of validation models estimations. Figure 2 presents the dot diagrams of the results of models estimations conformity and observed values of properties.

Table 3. Coefficients of determination and root mean square errors from the validation of estimation models of soil properties vector based on the VIS-NIR absorbance vector of soil samples from Poland included in the LUCAS database; the best modelling results are highlighted.

Algorithm	Statistic	Clay	Silt	Sand	pH(1)	pH(2)	SOC	$CaCO_3$	N	P	K	CEC
PLSR(A)	R^2	0.71	0.55	0.58	**0.61**	**0.60**	0.62	0.43	0.61	0.23	0.35	0.54
	RMSE	4.15	14.37	17.07	**0.70**	**0.68**	8.88	14.33	0.67	23.67	82.34	5.04
PLSR(A1)	R^2	0.65	0.46	0.51	0.52	0.51	0.51	0.56	0.53	0.25	0.42	0.47
	RMSE	4.55	15.72	18.52	0.77	0.74	10.07	12.70	0.73	23.39	77.75	5.39
PLSR(A2)	R^2	0.65	0.57	0.61	0.57	0.55	0.58	0.23	0.54	0.27	0.63	0.49
	RMSE	4.49	14.00	16.53	0.73	0.71	9.40	16.72	0.72	23.02	62.29	5.31
MLP(J)	R^2	**0.83**	**0.91**	**0.91**	0.45	0.44	**0.87**	**0.96**	**0.76**	-	-	**0.67**
	RMSE	**3.19**	**6.57**	**7.92**	0.82	0.79	**5.13**	**3.91**	**0.52**	-	-	**4.22**

Explanation of abbreviations and symbols: pH(1) – pH in $CaCl_2$, pH(2) – pH in H_2O, R^2 – coefficient of determination, RMSE – root mean square error, PLSR(A) – Partial Least Squares Regression algorithm (15 first components) from the VIS-NIR absorbance vector, PLSR(A1) – Partial Least Squares Regression algorithm (15 first components) from the first absorbance derivative vector, PLSR(A2) – Partial Least Squares Regression algorithm (15 first components) from the second absorbance derivative vector, MLP(J) – multilayer perceptron algorithm with inputs – combined components (15 first components each) of absorbance PLSR, first absorbance derivative, and second absorbance derivative.

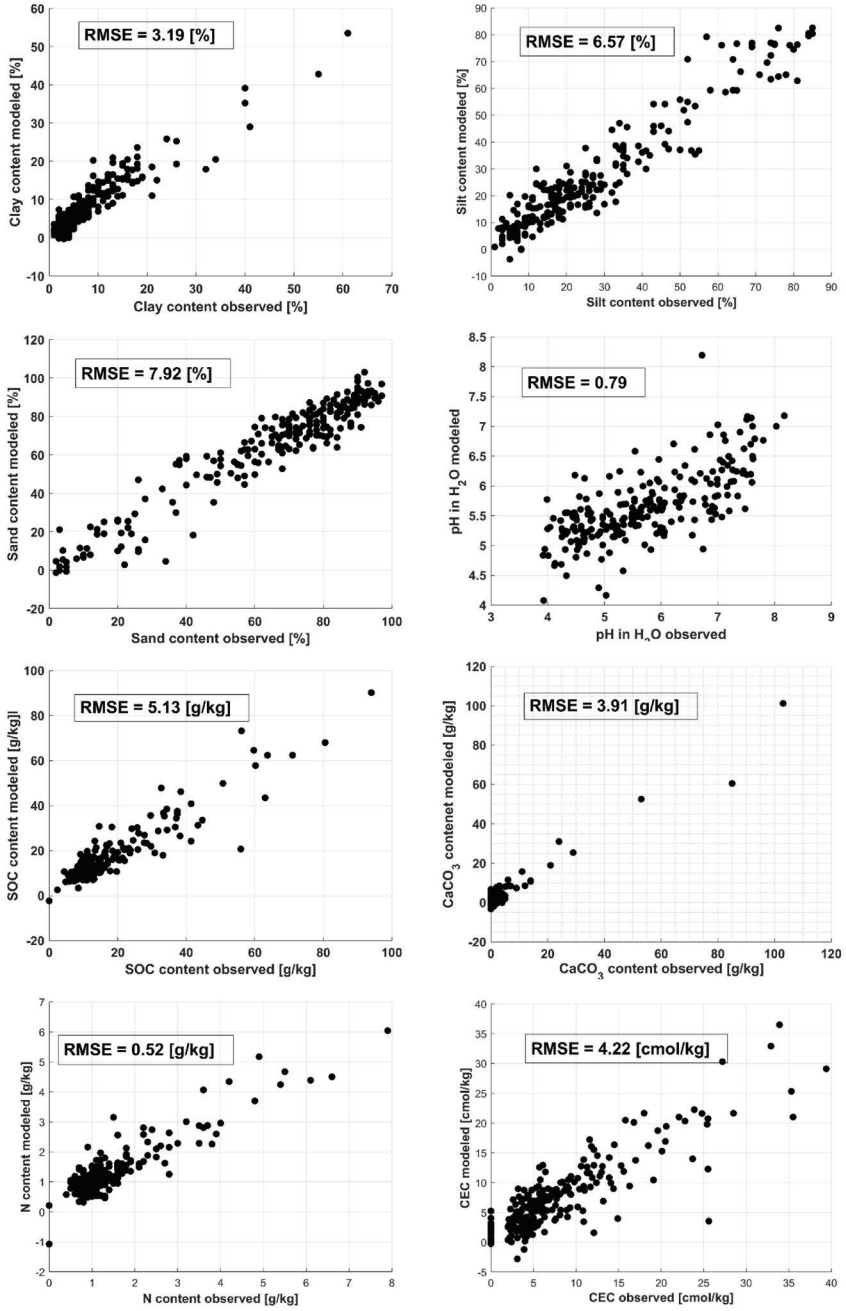

Figure 2. Dot diagrams of conformity between the observed data and the MLP NIR-based model

The LUCAS data were used in methodology studies, related mainly to the prediction of soil organic carbon based on the NIR analysis. The report from preliminary studies (Toth *et al.* 2013) includes information about the SOC prediction error (RMSE) in the range from 3.6 $g \cdot kg^{-1}$ (cropland), 7.2 $g \cdot kg^{-1}$ (grassland) to 11.9 $g \cdot kg^{-1}$ (woodland). The applied modelling method was a combination of the LOCAL algorithm (Shenk *et al.* 1997) and the interval partial least squares (I-PLS). Another paper based on the same data (Stevens *et al.* 2013) indicated the SOC prediction errors of 4.0 $g \cdot kg^{-1}$ (cropland), 6.4 $g \cdot kg^{-1}$ (grassland), 10.3 $g \cdot kg^{-1}$ (woodland) and 7.3 $g \cdot kg^{-1}$ (mineral soils). The model used was the support vector machines (SVM) with selection of variables according to the recursive feature elimination. The paper (Liu *et al.* 2018) used the LUCAS data for modelling the clay content in soils using one-dimensional convolutional neural network (1D-CNN). The RMSE statistic was 8.62% of clay content, and the RPD statistics =1.54. In another paper by the same authors (Liu *et al.* 2017), a combination of PLSR algorithms (selection of variables) and decision trees (Gradient Boosting Machine) was used to build a combined model. The SOC prediction model had the RMSE of 6.8 $g \cdot kg^{-1}$ (cropland), 10.9 $g \cdot kg^{-1}$ (grassland) and 13.31 $g \cdot kg^{-1}$ (woodland). The same statistic for prediction were: 0.42 $g \cdot kg^{-1}$ (cropland), 0.82 $g \cdot kg^{-1}$ (grassland) and 0.78 $g \cdot kg^{-1}$ (woodland). The clay prediction RMSE was 5.1 – 6.2%. Note that these results relate to the whole set of data from areas different in terms of mineral material, climate, methods of use. (Zhang *et al.* 2016) considered an attempt to model the soil nitrogen content based on the NIR analysis in one of the China regions to be successful. The deep learning technique for the NIR-based soil modelling from the LUCAS data is presented in (Veres *et al.* 2015).

The data from Poland are less differentiated in many terms, and more homogenous than the differentiated LUCAS database. This can be considered the main reason for relatively better NIR-based prediction results. The RPD (*Ratio of Performance to Deviation*), often used in comparisons of prediction models, in case of the presented model was from 1.76 (CEC) to 4.9 ($CaCO_3$ content). The RMSE values are more convincing as the model reliability measures. The RMSE values of the grain size estimation accuracy are surprisingly good, certainly competitive to macroscopic analyses. Also the organic carbon and carbonates content prediction errors indicate a sufficient model reliability, particularly for cartographic representation of these two values, especially when the time and costs of NIR determinations are taken into account which favour their repetitions for better understanding of the actual soils variation and reduction of the estimation error of averaged values at the pedon scale. It can be supposed that the near infrared determination of soils spectral response will allow a gradual expansion of cartographic soil documentation in Poland.

CONCLUSIONS

1. Contrary to the results obtained with the application of universal data models from 23 European countries, the limitation of modelling to a single country significantly increases the soil properties prediction results based on the NIR spectral response.
2. Simple machine learning models give satisfactory prediction results for a few soil properties: grain size, content of organic carbon and carbonates.
3. It can be supposed that the NIR determination technique will be helpful in development of the nationwide spatially continuous cartographic documentation of soils.

ACKNOWLEDGMENTS

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

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

Fuentes, M., Hidalgo, C., González-Martín, I., Hernández-Hierro, JM., Govaerts, B., Sayre, KD., Etchevers, J. (2012). NIR Spectroscopy: An Alternative for Soil Analysis. *Communications in Soil Science and Plant Analysis*, 43(1-2): 346-356, DOI: 10.1080/00103624.2012.641471.

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

Kokaly, RF., Clark, RN., Swayze, GA., Livo, KE., Hoefen, TM., Pearson, NC., Wise, RA., Benzel, WM., Lowers, HA., Driscoll, RL., Klein, AJ. (2017), USGS Spectral Library Version 7: U.S. Geological Survey Data Series 1035: 61, DOI: 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 Sens. 9: 1299.

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.

McBratney, AB., Mendonça Santos, ML., Minasny, B. (2003). On digital soil mapping. *Geoderma* 117(1-2): 3-52.

McBratney, AB., Minasny, B., Cattle, SR., Vervoort, W. (2002). From pedotransfer functions to soil inference systems. *Geoderma* 109(1-2): 41-73.

Mohamed, ES., Saleh, AM., Belal, AB., Abd_Allah, G. (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. DOI: 10.1016/j.ejrs.2017.02.001.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O. (2017). LUCAS Soil, the largest expandable soil dataset for Europe: a review, *Eur J Soil Sci* 69: 140-153.

Shenk, JS., Westerhaus, MO., Berzaghi, PJ. (1997). Investigation of a LOCAL calibration procedure for near infra-red instruments. *J. Near Infrared Spectrosc.* 5: 223-232.

Shi, Z., Ji, W., Viscarra Rossel, RA., 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. *Eur J Soil Sci*, 66: 679-687. DOI:10.1111/ejss.12272.

Stenberg, B., Viscara Rossel, RA., Mounem Mouazen, A., Wetterlind, J. (2010). Visible and Near Infrared Spectroscopy in Soil Science. *Advances in Agronomy* (Sparks D.L. Editor) 107: 163-215. DOI: 10.1016/S0065-2113(10)07005-7.

Stevens, A., Nocita, M., Toth, 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 8(6): e66409. DOI: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. DOI: 10.2788/97922.

Veres, M., Lacey, G., Graham, WT. (2015). Deep Learning Architectures for Soil Property Prediction'. Proceedings – 2015 12th *Conference on Computer and Robot Vision,* CRV 2015: 8-15. DOI: 10.1109/CRV.2015.15.

Wetterlind, J., Stenberg, B.,Viscarra Rossel, RA. (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, pp 95-107. Published in serie: Methods in molecular biology, nr. 953.

Zhang, Y., Min-Zan, L., Li-Hua Z., Yi Z., Xiaoshuai P. (2016). Soil nitrogen content forecasting based on real-time NIR spectroscopy. *Computers and Electronics in Agriculture*. 124: 29-36. DOI: 10.1016/j.compag.2016.03.016.

Prof. Stanisław Gruszczyński, PhD, DSc
AGH University of Science and Technology
Faculty of Mining Surveying and Environmental Engineering,
al. Mickiewicza 30
PL 30-059 Krakow
Tel. (+48) 12 617 22 89
Email: sgrusz@agh.edu.pl

Received: 7 May 2019
Accepted: 21 May 2019