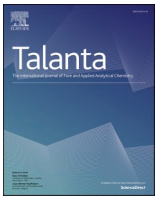
Talanta 251 (2023) 123749

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Monitoring the concentrations of Cd, Cu, Pb, Ni, Cr, Zn, Mn and Fe in cultivated Haplic Luvisol soils using near-infrared reflectance spectroscopy and chemometrics

S. Krzebietke [a], M. Daszykowski [b,*], H. Czarnik-Matusewicz [c], I. Stanimirova [b], L. Pieszczek [b], S. Sienkiewicz [a], J. Wierzbowska [a]

[a] Department of Agricultural and Environmental Chemistry, Faculty of Agriculture and Forestry, University of Warmia and Mazury in Olsztyn, 8 Oczapowskiego Street, 10-719, Olsztyn, Poland
[b] Institute of Chemistry, University of Silesia in Katowice, 9 Szkolna Street, 40-006, Katowice, Poland
[c] Department of Clinical Pharmacology, Faculty of Pharmacy, Wrocław Medical University 211a Borowska Street, 50-556, Wrocław, Poland

ARTICLE INFO

Keywords:
NIR
Soil analysis
Soil monitoring
Proximity soil sensing
Precision agriculture
Chemometrics

ABSTRACT

This study illustrates the successful application of near-infrared reflectance spectroscopy extended with chemometric modeling to profile Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe in cultivated and fertilized Haplic Luvisol soils. The partial least-squares regression (PLSR) models were built to predict the elements present in the soil samples at very low contents. A total of 234 soil samples were investigated, and their reflectance spectra were recorded in the spectral range of 1100–2500 nm. The optimal spectral preprocessing was selected among 56 different scenarios considering the root mean squared error of prediction (RMSEP). The partial robust M-regression method (PRM) was used to handle the outlying samples. The most promising models were obtained for estimating the amount of Cu (using PRM) and Pb (using the classic PLS), leading to RMSEP expressed as a percentage of the response range, equal to 9.63% and 11.5%, respectively. The respective coefficients of determination for validation samples were equal to 0.86 and 0.58, respectively. Assuming similar variability of model residuals for the model and test set samples, coefficients of determination for validation samples were 0.94 and 0.89, respectively. Moreover, the favorable PLS models were also built for Zn, Mn, and Fe with coefficients of determinations equal to 0.87, 0.87, and 0.79.

1. Introduction

In recent years, there has been a growing interest in using different spectroscopic techniques to measure reflectance spectra. The development of instruments and technological progress has enabled reflectance spectra to be recorded in a relatively wide range of electromagnetic radiation (EMR) from the visible (Vis), near (NIR), short (SWIR), to the medium infrared (MIR) range, i.e., from about 350 nm to about 25,000 nm. EMR interacts with matter differently depending on the spectral range, and these interactions propagate into the recorded spectrum and are manifested as specific bands. The relatively low cost of the measurements, their speed and the possibility to directly process samples without preparation have accelerated interest in reflectance spectra. Therefore, many applications of spectroscopic techniques have been described in the literature.

Spectroscopic techniques are often used for proximity sensing. When combined with the chemometric modeling of spectra, they can replace the classic and time-consuming laboratory measurements and provide an innovative and automatic framework for high-throughput monitoring [1,2] and for precision agriculture (an approach to farm management that uses information technology, including spectroscopic and chemometric methods, to ensure that crops and soil receive what they need for optimal health and productivity) [3]. In addition, the most desired opportunity arises when remote sampling technologies are used [4–6]. Measuring devices can be installed on an aircraft, e.g., a drone, airplane, balloon or satellite, thus enabling remote or even orbital sampling. Flying over a specific area to collect spectral data and analyzing them in real-time is a tempting research trend that opens the possibility for efficient monitoring, which cannot be handled based solely on classic sampling and routine analytical procedures. These

* Corresponding author.
E-mail address: michal.daszykowski@us.edu.pl (M. Daszykowski).

https://doi.org/10.1016/j.talanta.2022.123749
Received 26 May 2022; Received in revised form 14 July 2022; Accepted 15 July 2022
Available online 21 July 2022

solutions are desired when sampling large areas over a long period of time.

Not surprisingly, spectroscopic techniques are frequently used to map the type of soil [7] and to study the properties and conditions of the soil at specific sites [8–15], large land fields, entire farmlands or are used to develop comprehensive spectral libraries [16]. Spectroscopy in the visible and NIR range can often be used with great success to assess many physicochemical properties that reflect the vital conditions of soil and its composition [17,18]. Soil is a medium of crucial importance for plants, animals and human beings. It acts as a natural water filter, provides nutrients to plants, delivers food products, stores carbon, reduces greenhouse gas emissions, and contributes to climate change [16, 19]. Therefore, one expects generations to protect this common natural resource and use it wisely.

Soil is a porous and complex mixture containing inorganic elements and compounds formed in different biological and geological processes and anthropogenic pollutants. Its condition and properties result directly from its chemical composition. Furthermore, it significantly affects further soil usage, the optimal yield, and local and global economies. Therefore, knowledge about the chemical content and monitoring of the soil condition is of great importance not only for its sustainable and optimal maintenance but also for evaluating its health, the progress of degradation, the remediation of land, and post-industrial areas (see, e.g., Refs. [20–22]).

In the mid-1960s, Bowers and Hanks published a pioneering work on systematic studies that illustrated the impact of moisture, organic matter and the size of particles on the reflectance spectra of soil samples [23]. Dalal and Henry modeled the moisture, organic carbon and total nitrogen content in air-dried soil samples [24]. Ben-Dor and Banin further extended the scope of the NIR applications. They studied the relationship between the reflectance spectra and the clay content, specific surface area, cation exchange capacity, hygroscopic moisture and organic matter in soil [25]. Since then, many researchers have acknowledged reflectance spectroscopy and extensively explored the possibility of associating spectra with fundamental soil properties such as the content of many metals, their binding forms, and the soil type. These studies were primarily driven by the considerable sensitivity of the reflectance spectra to small variabilities in the concentration of organic matter, minerals and elements that are adsorbed on soil particles. As a result, the reflectance spectra are often regarded as spectroscopic fingerprints that have the potential to characterize soil samples uniquely. Moreover, the absorption feature in the spectrum can be attributed to specific binding forms. For instance, toxic elements such as Cd, Pb and Hg mainly bind to organic matter, while Cr, Cu and As are mostly retained by iron oxides, clays and organic matter [26].

On the other hand, the relevant information contained in the spectra is challenging to access. Thus, the spectral data and libraries are explored and modeled using chemometric techniques. These are, for instance, exploratory methods like principal components analysis and clustering techniques that help visualize multivariate data, assess similarities among spectral profiles, find groups of samples and outlying samples, simplify the data representation and improve interpretation of the results [27,28]. Another group includes supervised methods such as calibration, discriminant and classification techniques, for instance, multiple linear regression (MLR), partial least-squares regression (PLS), multivariate adaptive regression splines (MARS) and support vector machines (SVM). Their applications to model and interpret the diffuse reflectance spectra of soil samples have been discussed, e.g. Refs. [29–31]. Modeling concerns estimating a few fundamental properties directly related to spectral fingerprints, for instance, moisture, the total organic content, mineralogy and particle size distribution. In addition, it is also possible to model properties that correlate with the fundamental ones (an indirect relationship with the spectral fingerprints). These are, for instance, pH of the soil, the concentrations of different macronutrients (e.g., Ca, Mg, K, N, P, and S), micronutrients (e.g., Fe, Mn, Zn, Cu, and B), heavy metals and metalloids (e.g., As, Cd, Cu, Ni, Pb, Zn, Hg, and Cr) [32].

The relatively large number of parameters that can be monitored in the field, including potentially toxic elements, has considerably increased the interest in NIR and other spectroscopic techniques [33]. For instance, a long-term monitoring program of soils was initiated in 2013 for the Saxon region in Germany with a long mining history (the Saxon Permanent Monitoring Soil Program) [34]. The aim was to evaluate the spatial and temporal changes in soil properties, assess the pollution levels and identify different pollution sources. This monitoring network focuses on tracing the content of metals and metalloids (Al, As, Ca, Cu, Fe, K, Mn, Na, Ni, Pb and Zn), the total organic carbon and soil pH using Vis-NIR (350–2500 nm) and MIR (2500–25,000 nm) techniques, while the chemometric techniques support the data modeling. Over time, similar studies have been published focused on monitoring selected elements, e.g. Refs. [35–38].

Our primary motivation was to monitor efficiently the long-term chemical changes of Haplic Luvisol soil induced by different fertilization schemes in the well-designed experimental system. Luvisols account for ca. five percent of the total continental land area (500–600 million ha of land). Luvisols are found in central Europe, west-central Russia, the United States, the Mediterranean basin, and southern Australia. In Poland, Luvisols represent about 40–44% of all soils. This type of soil offers optimal physical and chemical properties in favorable climate conditions. Therefore, it can be used to grow demanding plants (rapeseed, sugar beet, wheat) and less demanding ones (triticale, rye or plants from the faba family) and obtain a good harvest of potato, grass, and maize. Considering the potential of NIR reflectance spectroscopy, we took an advantage of the chemometric modeling of spectra by constructing classic and robust PLS models (i.e., insensitive to outliers) for eight chemical elements present at relatively low concentrations in cultivated Haplic Luvisol soil. In addition, we intended to support the emerging field of precision agriculture.

2. Materials and methods

2.1. Experiment and sampling site

The experimental field in Balcyny near Ostroda, Poland (53° 35′ 34.045″ N; 19° 50′ 54.671″ E) has been cultivated since 1986. The experimental design considered randomly selected blocks with two stripes and four repetitions. The overall experimental plan is presented in Supplementary Table S1. The soil belts were fertilized with manure (40 t ha^{-1}) every two years for sugar beet and maize and combined with mineral fertilization, which was applied in the second year. Such fertilization schemes enable diversified mineral soil supplementation and lead to eight groups of samples (see Supplementary Table S2). The plants investigated during the experiment were sugar beet, spring barley, maize for green fodder and spring wheat. The plot area was 35 m^2. The experimental field contains Haplic Luvisol soil. According to the particle size distribution, the soil was classified as sandy loam. The proportions of nine fractions in the soil samples collected at four depths are presented in Supplementary Table S3. In 1986, the chemical content of the soil was examined. Then, a 1 kg sample of soil contained 100 mg of K, 53.2 mg of Mg, 41.3 mg of P, 7.9 g of total organic carbon (C_{org}) and 0.79 g of total nitrogen (N_{tot}); pH (1 mol dm^{-3} KCl) = 6.2. The chemical content of the manure was described in the refs. [39,40].

2.2. Sampling and sample preparation

The soil samples were collected in 2017 and 2018 after the plant vegetation period (after the harvest). In August 2017, the samples were collected after the winter wheat harvest in the second year after the manure had been applied (the end of the 8th yield rotation) from two sampling horizons: 0–30 cm and 30–60 cm. In October 2018, the samples were collected after the sugar beet harvest from a 0–30 cm horizon one year after the manure had been applied. The samples were collected

from each plot using an Egner soil cane, thus obtaining ca. 1 kg of an integrated sample from which 0.5 kg was used for further analysis. The samples were kept under air-dry conditions and then sieved (mesh diameter of 2 mm).

The dry soil was mixed from each fertilizer combination in equal weight proportions 1:1. After the replicates were eliminated, there were 48 samples (2018 – 16 samples with manure and without manure from a 0–30 cm layer after the sugar beet harvest; 2017 – 32 samples with manure and without manure from 0 to 30 cm and 30–60 cm layers after the spring wheat harvest); in 2018, from the 64 samples that represented each sugar beet plot, five samples were collected from the belts surrounding the sampling field. Then, the soil samples were unified concerning the fraction size by grinding them in a planetary mill (Planetary Ball Mill PM 100 – Retsch).

In this study, 234 samples (original and sieved) were used to confirm the feasibility of the NIR reflectance spectroscopy for monitoring the content of *Cd, Cu, Pb, Ni, Cr, Zn, Mn and Fe* in cultivated Haplic Luvisol soils.

2.3. Chemical analysis of Cd, Cu, Pb, Ni, Cr, Zn, Mn and Fe

The content of micronutrients (Zn, Mn, Cu and Fe), and potentially toxic heavy metals (Pb, Cd, Cr and Ni), were examined using a reference method. The bioavailable forms of Cd, Cu, Pb, Ni, Cr, Zn, Mn and Fe were extracted from the soil using a 1 mol dm^{-3} HCl solution. A 5 g sample of the sieved soil (a soil fraction with particles less than 2 mm in diameter) was flooded with 50 cm^3 of 1 mol dm^{-3} HCl, mixed on a rotary evaporator for 30 min and filtered through a hard filter. The eight elements were determined directly from the solution using atomic absorption spectrometry (Shimadzu AA-6800). The precision of the determination (expressed in percentage), which was referred to certified material (Trace Metals-Sevage Sludge 4, Sigma-Aldrich RTC, Inc.), was as follows: Cd – 96.7%, Cu – 94.5%, Pb – 99.5%, Ni – 98.7%, Cr – 97.2%, Zn – 86.7%, Mn – 88.8% and Fe – 111.4%.

2.4. Registering the NIR reflectance spectra

The NIR spectra of the soil samples were recorded in the reflectance mode, log(1/R), within a spectral range of 1100–2500 nm at a 1 nm step using a SpectraStar XL RTW (Rotating Top Window) near-infrared (NIR) diffuse reflectance spectrometer equipped with a pre-dispersive scanning monochromator with a nominal bandwidth of 10 nm (Unity Scientific, Brookfield, CT, USA). Samples were analyzed in an air-conditioned room at 22 °C and relative humidity of 45%. The control of the instrument and the management of the spectral files were performed using InfoStar software (version 3.11.1, Unity Scientific, Brookfield, CT, USA). The NIR spectra of the soil samples (1 cm thickness) were measured in a closed powder cup (US-MPCP-0001, 3.5 cm diameter) equipped with compaction control. The cup was inserted (window facing down) in the rotating accessory. The NIR spectrum of each soil sample was averaged over 24 scans (two cup rotations). The instrument was calibrated automatically every 30 min using the internal standard (US-STDS-0001).

2.5. Construction and validation of the calibration models

The soil sample spectra were split into the model and test sets according to the uniform design of the response variable using the Kennard and Stone algorithm [41,42]. Eighty and thirty-seven soil sample spectra from each soil fraction (fine and coarse) were selected for the model (calibration) and test (validation) sets, respectively. The model set sample spectra were used to construct the multivariate calibration models, while the test set sample spectra served for their validation. Each response variable was modeled separately using the partial least-squares regression.

Prior to constructing a model, different preprocessing methods and scenarios were evaluated, including detrending, standard normal variate (SNV), multiplicative scatter correction (MSC), extended multiplicative scatter correction (EMSC), inverse scattering correction (ISC), extended inverse scattering correction (EISC) and normalization of the spectra to the unit standard deviation. In addition, the transformed spectra were further preprocessed using the first and second derivatives combined with the Savitzky-Golay smoothing. For each response variable, 56 models were constructed with one up to fifteen PLS factors (in total 448 models). The goodness of the model fit was judged by the root mean square error of calibration (RMS), which was calculated based on samples from the respective model set. The prediction power of each PLS model was expressed as the root mean square error of prediction (RMSEP) calculated for the test set samples. These two figures of merit were also expressed as the percentages of the observed calibration range of a given response variable. We have also calculated the coefficient of determination (R^2) for calibration and test set samples. The optimal models were selected based on the prediction errors obtained for test set samples presented as a function of the number of PLS factors. A collection of prediction error curves obtained for models describing the content of Pb in soil samples for samples from the model test sets are shown in Fig. 1.

Moreover, partial robust M-regression (PRM), i.e., insensitive to outlying samples, was used [43]. During the iterative model construction process, samples are weighted according to their distances from the robust center of the multivariate data (computed in the space of the robust latent factors) and the response residuals from the PRM model. In this way, the impact of any outlying samples is reduced and the final model explains well the overall linear trend for data majority. Since the PRM model is robust in the statistical sense, the score distances computed in the space of the robust latent factors and model residuals, which are visualized in a so-called distance-distance plot, can reveal outlying samples depending on their location in the multivariate data space regardless of any swamping and masking effects.

The authors programmed all spectral preprocessing methods and performed all necessary calculations in the MATLAB environment (version 9.0, R2016a) operating under Microsoft Windows 10 Pro Version 10.0. The classic and robust versions of PLS are included in the freely available toolbox for multivariate calibration techniques (TOMCAT) [44].

3. Results and discussion

3.1. Exploratory analysis of the spectral data and eight response variables

The major challenge in modeling diffuse reflectance spectra arises from the considerable heterogeneity of the soil samples. The shape and intensity of the spectra are mainly affected by the chemical composition, mineralogical content, soil structure and diameter of soil particles. The possible variations can be relatively large and blur the underlying relationships. In our study, the diameter of the soil particles was the greatest source of variability. The different components of the soil samples and the diameter of the soil particles increased the scattering, as shown in Fig. 1a (a systematic variation among the spectra of the original and sieved samples). Black spectra represent the original soil samples, while the red ones the samples after sieving. The original samples absorbed more EMR, while the sieved ones, containing particles with a diameter below 2 mm, absorbed less EMR due to more extensive scattering (see Fig. 2a). Fig. 2b presents the projection of the soil samples onto the first two principal components obtained from the PCA. The black and red dots refer to the original and sieved soil samples. The scattering effect was modeled by the first principal component explaining above 99% of the total spectral variability.

The differences related to the fertilization schemes are not readily visible on the projections of the first two principal components. However, a sampling effect is revealed by the second principal component, explaining 0.30% of the data variability. The soil samples collected up to

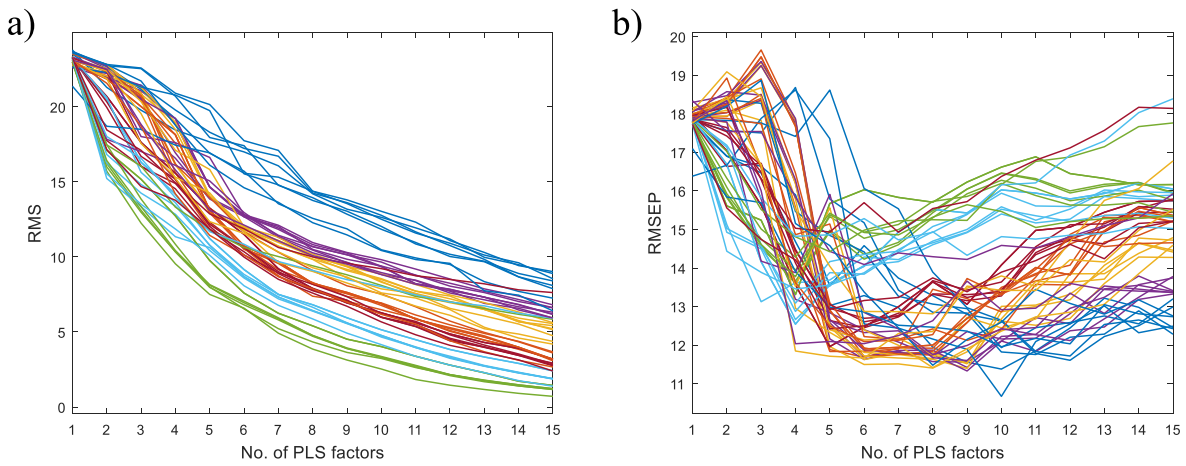


Fig. 1. Predictions, expressed as the root mean squared errors of prediction, obtained from the PLS models for the Pb content in the Haplic Luvisol model (RMS) and independent test (RMSEP) soil samples. Models were built with up to fifteen PLS factors using 56 differently preprocessed NIR spectra.

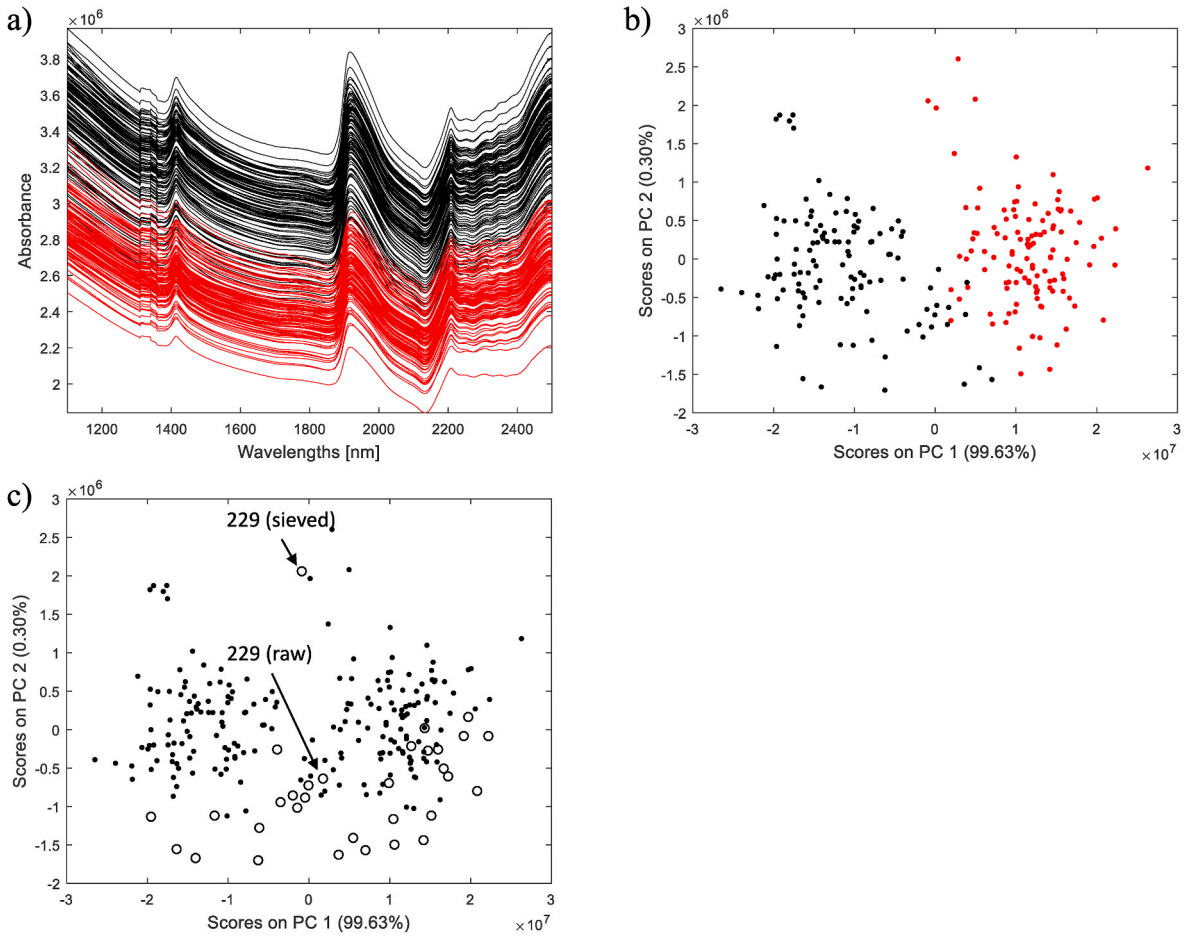


Fig. 2. a) The original near-infrared reflectance spectra that were obtained from two fractions of 117 Haplic Luvisol soil samples before (black lines) and after sieving (red lines), b) projection of the soil samples before (black dots) and after sieving (red dots) on the space of the first two principal components that were obtained from the principal component analysis (PCA) and c) projection of the soil samples that had been collected at depths of 0–30 cm (black dots) and 30–60 cm (black circles) on the space of the first two principal components. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

a 30 cm depth had larger score values than those collected from a deeper depth (see Fig. 2c). Sample no. 229 (sieved soil collected from a depth of 30–60 cm in 2017 only after the mineral fertilization according to scheme 8 - see Table S2) did not follow the discussed trend. Surprisingly, the same soil sample analyzed before sieving (sample no. 112) was close to the samples from the same group.

Some basic statistical parameters for eight modeled elements, including their minimal and maximal values, ranges, means, medians and standard deviations, are presented in Table 1, and corresponding histograms are included in Supplementary Fig. S1.

Table 1
Basic statistics that describe the content of the eight elements (Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe) expressed as mg·kg^{-1} in the Haplic Luvisol soil samples (minimal and maximal values, ranges, mean and median and standard deviation). The content distribution for each element is visualized in the histograms presented in Supplementary Fig. S1.

No.	Statistics	Cd	Cu	Pb	Ni	Cr	Zn	Mn	Fe
1	Minimal value	0.040	1.120	6.980	0.050	0.180	5.650	92.600	778.200
2	Maximal value	0.250	6.140	16.750	3.070	5.660	35.140	509.200	3995.200
3	Range	0.210	5.020	9.770	3.020	5.480	29.490	416.600	3217.000
4	Mean value	0.158	2.464	12.333	1.663	2.244	17.887	323.132	2681.541
5	Median	0.170	2.070	12.860	1.750	2.120	17.550	348.900	2699.900
6	Standard deviation	0.044	1.069	2.167	0.695	0.956	6.541	106.224	819.963

It is also interesting to evaluate the colormap in which the colors of the pixels and their hues express the degree of the pairwise correlations among eight response variables (see Fig. 3). The cold colors indicate poor correlations, while the hot colors show higher correlation values then 0.6? A good relationship was observed between the concentrations of Pb and Mn (correlation coefficient was equal to 0.7938) and Zn and Ni, which had a relatively good correlation of 0.6904. Correlations between Zn and Mn (0.6582) and Zn and Pb (0.6306) are also worth mentioning.

3.2. Construction of the optimal PLS models

Constructing any multivariate calibration model requires a set of representative samples that span the calibration domain. They express the variability sources, determine the potential calibration range and affect the quality of future predictions. Including various variability sources at the calibration stage is strongly recommended because the stability of the final model increases and its maintenance is much longer. On the other hand, certain factors can induce groups in data, e.g., different soil types, soil fractions, etc. Then, the local models will yield smaller prediction errors than the global ones. The choice between these two opposing modeling frameworks is problem- and data-dependent. However, from the perspective of model maintenance, global calibration models are general, and their performance is more stable over time. These aspects are essential for developing effective calibration strategies and constructing calibration models for proximity or distant sampling and soil assessment applications.

The Haplic Luvisol evaluated in this study belonged to a medium-sized soil category. The metals in the soil (cations) were retained mainly due to the exchangeable physicochemical sorption. The more floatable (colloidal) components the soil contains, the greater the capacity of the sorption complex. Hence, heavy soils with a significant amount of silt and dust fractions contain more metals in the topsoil than light or organic soils with a relatively low sorption capacity.

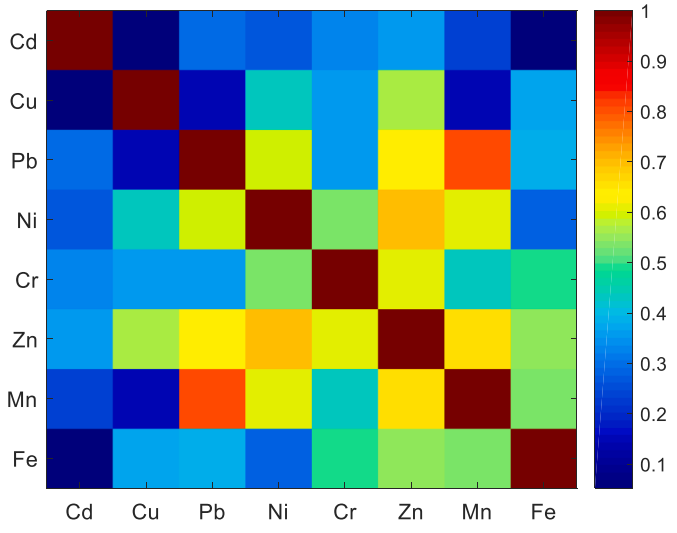


Fig. 3. Color map visualizing the pairwise correlations between the eight parameters that describe the concentrations of Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe in the Haplic Luvisol soil samples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Considering these aspects, in our study, as many sources of variability as possible were included in the modeling. The calibration (model) set contained two fractions of soil samples 'before' and 'after' sieving because when soil samples are analyzed outside the laboratory or for proximity or remote soil measurements, sample preparation is impossible. Additional variability was introduced by including samples collected at two horizon depths in the model set. The source of metals in the soil was mainly soil bedrock, whereas in the arable layer, they were released from the mineral fertilizers, fertilizing waste, plant protection products, dry exposure to the near industrial plants, transport, and air. The metal content also depended on the proportion of mineral, organic or organic-mineral colloids, soil pH, and the origin of the soil formation. Organic soils have a limited ability to accumulate metals. With an increase in soil acidification, the availability of metals for plants and their mobility increases significantly. Additional variability sources incorporated in the modeling corresponded to fertilization type and sampling time.

3.3. Performance of the classic PLS models

The optimal PLS models that described the content of eight elements in the soil samples were built as described in section 3.2. The optimal preprocessing scenarios and figures of merit describing the optimal PLS model are presented in Table 2. Most of them involve derivatives and improve the models (see figures of merit). In addition to analyzing the figures of merit, it is also recommended to display the predictions obtained for model and test set samples as predicted response values obtained from the model versus the observed values (see Fig. 4).

The first PLS model built for estimating Cd content had a narrow calibration range from 0.04 to 0.25 mg kg^{-1} (up to 0.25 ppm). The optimal model included five factors and offered ca. 19.00% and 14.26% errors for the model and test set samples, respectively (see Table 2). The coefficients of determinations that described the model's performance for the model set, R^2m, and for the test set, R^2t, were relatively low and below 0.3, indicating limited prediction accuracy. The distribution of predictions shown in Fig. 4a suggested a possible non-linear relationship. It was mainly observed for the soil samples with low concentrations of Cd (below approx. 0.1 ppm) because all model residuals were positive.

The PLS model describing the Cu content was built for spectra after the ISC followed by the first derivative (see Table 2). The model worked in a calibration range of 1.12–6.14 mg kg^{-1} (ca. 1–6 ppm) and had five factors. In this modeling example, it is worth noting that the distribution of Cu deviates most from the normal distribution compared to the remaining modeled responses (see Supplementary Fig. S1). The content of Cu was large for several samples. They can be easily spotted in Fig. 4b because their residuals from the classic PLS model were large, and samples were located far from the line with a slope that equals one (representing the ideal model with residuals equal to zero). Even though

Table 2
Overview of the optimal PLS models constructed to predict the concentrations of Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe in the Haplic Luvisol soil samples. Along with the optimal preprocessing method, the number of factors included in the PLS model (*f*), and several basic figures of merit are also reported. These are the root mean square error (RMS) that were calculated for the samples in the model set, the root mean square error of prediction (RMSEP) that was calculated for samples in the model set, which are also expressed as the percentage of the calibration range and the respective coefficients of determination (R^2m and R^2t). The modified coefficient of determination calculated for the samples from the test set, R2t*, took into account the range of the responses observed for the samples in the model set.

Model	Preprocessing method(s) used to transform the NIR spectra	*f*	RMS	RMSEP	RMS [%]	RMSEP [%]	R^2m	R^2t	R^2t*
Cd	1st derivative (window = 15, polynomial degree = 2)	5	0.0399	0.0299	19.00	14.26	0.2974	0.2641	0.8170
Cu	ISC + 1st derivative (window = 15, polynomial degree = 2)	5	0.8731	0.5032	17.39	10.02	0.4510	0.2948	0.9156
Pb	1st derivative (window = 11, polynomial degree = 2)	6	1.1690	1.1233	11.96	11.50	0.7473	0.5831	0.8921
	EISC	10	1.1759	1.0425	12.04	10.67	0.7443	0.6409	0.9070
Ni	1st derivative (window = 7, polynomial degree = 2)	6	0.4023	0.4320	13.32	14.31	0.7255	−0.1178	0.8536
	EISC	11	0.4225	0.4042	13.99	13.39	0.6972	0.0214	0.8718
Cr	detrending + 2nd derivative (window = 7, polynomial degree = 2)	1	1.0313	0.6397	18.82	11.67	0.0681	−0.0716	0.8341
Zn	detrending + 1st derivative (window = 7, polynomial degree = 2)	7	2.7777	3.7495	9.42	12.71	0.8437	0.4861	0.8683
Mn	1st derivative (window = 11, polynomial degree = 2)	5	72.6538	62.9890	17.44	15.12	0.5604	0.4991	0.8472
Fe	Detrending	9	521.2990	605.0577	16.20	18.81	0.6133	0.3671	0.7591

predictions for the test set samples were promising (ca. 10% of error), in our opinion, its actual predictive capabilities were also distorted by the representativeness of the samples. The calibration range influenced the coefficient of determination. Thus, the possible discrepancy between the coefficients of determination could have been caused by different variabilities and the effective response range. The relationship between the NIR spectra and Cu concentration was linear, but the PLS model required improvement via robust modeling, resistant to outlying samples (located either in the space of explanatory variables or in the space of response(s)).

Concerning estimation of the Pb content, the PLS model included a calibration range between ca. 7 and 17 mg kg^{-1} (i.e., 7–17 ppm). Modeling of the first derivative spectra led to the PLS model with six factors offering RMS and EMSEP equal to 11.96% and 11.50%, respectively. The respective coefficients of determinations were equal to 0.7473 and 0.5831 (see Table 2). An analysis of the residuals from the model shown in Fig. 4c confirmed the good model performance in terms of the figures of merit with the exception of the coefficient of determination calculated for test samples. Differences between R^2m and R^2t can be clarified by the effective range of the responses for the model and test set samples. Specifically, there were no samples with concentrations of Pb above 15 ppm in the test set. Similar prediction performance was also obtained for the NIR spectra transformed using the EISC method; however, the optimal model required ten PLS factors (see Table 2).

As was mentioned earlier, the concentration of Pb was correlated with Mn, Zn and Ni to a large extent. Thus, the prediction behavior of these three PLS models should be similar to the PLS model that described the content of Pb. For Ni, this tendency was reflected adequately by the model residuals. Based on the first derivative of the spectra, the optimal PLS model with six factors yielded ca. 13.32% of error for the model samples and its R^2m was equal to 0.7255. However, the model performed worse for these set samples, especially when the concentrations of Ni in the samples were above 2.5 ppm. As is indicated in Fig. 4d, the linear relationship became weaker in the upper calibration range (i.e., above 2.25 ppm). In contrast to the modeling of the concentration of Cu in the soil samples, in this case, the samples that were included in the test set had relatively large response values. Therefore, it affected the calculation of the R^2t value but not the RMSEP. For the discussed PLS model, the discrepancy between R^2m and R^2t values was the largest, i.e., 0.7255 and −0.1178, respectively. This tendency was not observed in Fig. 4d. In the calculation of R^2t, the total variability of the test set was replaced by the total variability of the model set, after which the modified coefficient of determination (R^2t*) improved and was equal to 0.8536. Similar performance of the PLS model was obtained for the NIR spectra after the EISC transformation, but it was more complex (eleven factors).

The PLS model constructed for estimating Cr concentration was the most surprising in terms of its predictive ability. The upper limit of the calibration was ca. 6 ppm. The optimal model, built for spectra after detrending and the second derivative, required one factor (see Table 2). While the prediction error for the test set samples was relatively low, ca. 11.67%, the visual assessment of the model's performance was disappointing (see Fig. 4e). Such an effect is typical for any regression model that minimizes the sum of squared residuals, including PLS, when the calibration data contain outlying samples. As a result, the model fits the outlying samples. Surprisingly, in the case of Cr, the considered figures of merit seemed to ignore the problem, but a disturbing prediction performance was revealed in Fig. 4e. In this situation, a robust PLS model was required.

The optimal PLS model for calibrating the Zn content was constructed for spectra after detrending and the second derivative. It covered a concentration range from approx. 5 ppm up to 35 ppm. When seven factors were used, it offered a good performance in terms of RMS and RMSEP, comparable to the Pb model. RMSEP was equal to 12.71%. This model had the largest R^2m value compared to all of the remaining models (0.8437); however, the R^2t value was much smaller and equaled 0.4861, due to the narrower response range for the samples from the test set compared to the response range of the model set samples. A modified version of R^2t* led to a value equal to 0.8683. In Fig. 4f, the model tended to underestimate the content of Zn for a few samples in the upper calibration range from ca. 30 ppm. Further verification if the linear relationship holds would require an extension of the calibration range by including samples with larger Zn content.

The PLS model that described the Mn content was built for the first derivative spectra. Its range was from ca. 93 ppm to 509 ppm. The PLS model with five factors offered RMS and RMSEP equal to 17.44% and 15.12%, respectively. However, it underestimated the predictions in the upper calibration range, i.e., above 475 ppm (see Fig. 4g). The presence of a few samples with relatively large residuals encourages the construction of a robust model. The Zn concentrations were correlated with the Pb concentrations (the correlation coefficient was equal to 0.7938), but the pattern of residuals obtained from the classic PLS model did not confirm the expected similar modeling behavior (cf. Fig. 4c and g).

Regarding modeling the Fe content in the soil samples, the PLS model had the most extended calibration range from 778 ppm to 3995 ppm. At first glance, the predictions for the spectra after detrending, were unsatisfactory. RMSEP was the largest among all of the models and equaled 18.81% (see Table 2). This can be explained by the scarce spectral information reflecting the impact of increased Fe content on soil samples. Many similar studies have proven modeling potential when spectra include the visible region of EMR. Due to the instrument limitations, it was impossible in our study, but the model was eventually enhanced by removing influential samples.

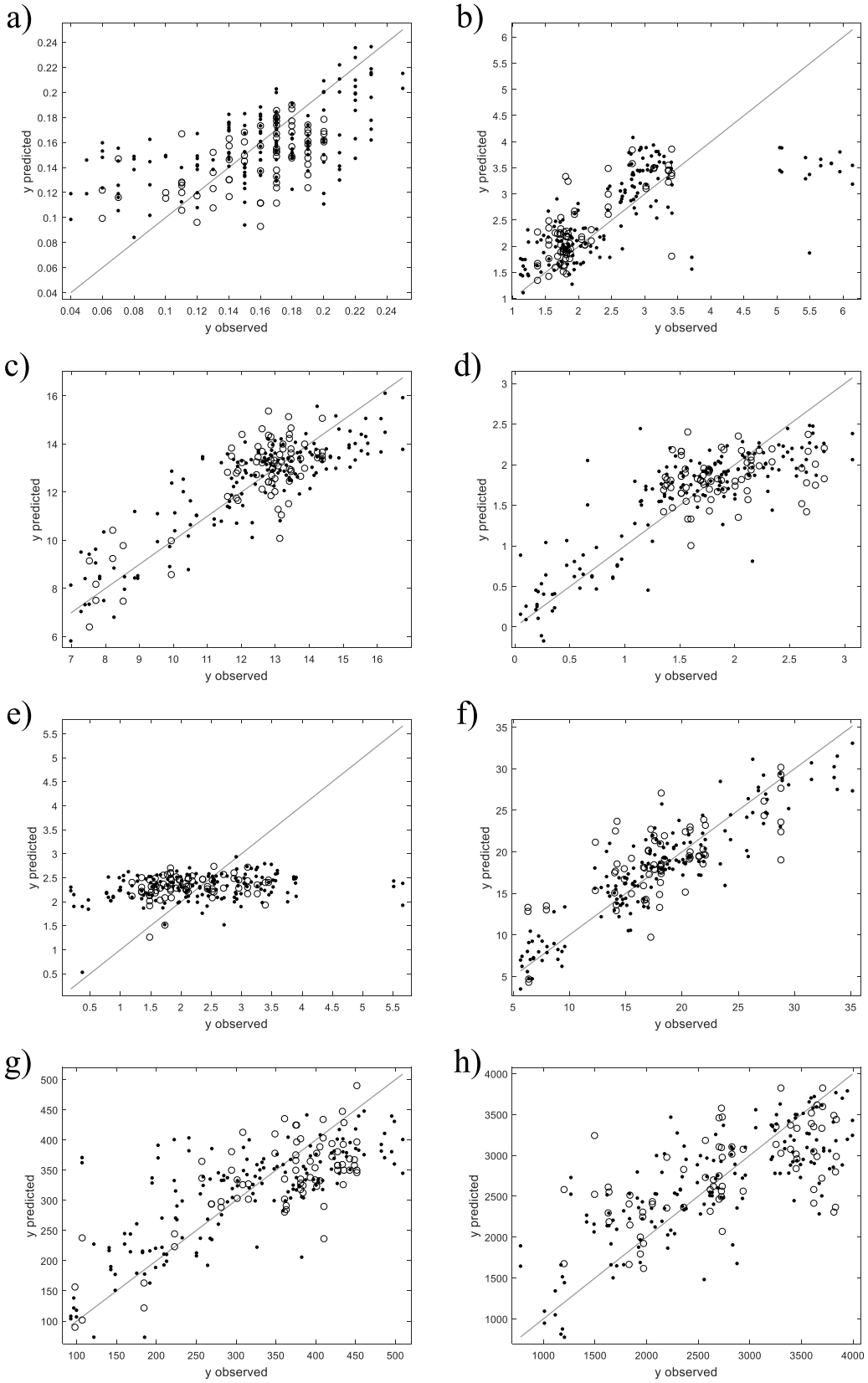


Fig. 4. Partial least-squares models presented as the predicted versus the observed response values for the model set samples (black dots) and the test set samples (black circles), which were constructed to estimate the concentrations of a) Cd, b) Cu, c) Pb, d) Ni, e) Cr, f) Zn, g) Mn and h) Fe. A gray line with a slope equals to one illustrates the ideal situation when residuals from the model are zeros. The concentrations of the elements are expressed in $mg \cdot kg^{-1}$.

3.4. Improving the calibration models for Cu, Cr, Mn, and Fe by handling outlying samples

Four models estimating Cu, Cr, Mn, and Fe contents were improved by robust calibration using the partial robust M-regression (PRM). The number of PRM factors was selected based on the robust RMSECV estimates, which were obtained from the Monte-Carlo cross-validation (we assumed a maximal fraction of potential outliers equal to 10% of the number of samples in the model set, 40 samples were left out at each Monte-Carlo iteration and 320 iterations were run). Generally, depending on the location of the influential samples in the model space, they affected the performance of the classic PLS model differently. Their impact was revealed in the distance-distance plot illustrating the absolute standardized distances in the space of the PRM factors and the absolute residual distances computed for the samples from the model set. These two distances helped to divide the samples from the model set into four categories according to their influence on the classic least-squares model. The regular samples had small absolute distances in the space of the robust latent factors and low residuals from the model; thus, they did not harm the classic PLS model. The so-called good leverage samples were far from the data majority in the space of the robust factors, but they fit the model well (small residuals). Their presence in the calibration data is beneficial for the model and its future maintenance. They extended the calibration range and increased the model's stability. The high residual samples do not fit the model well; therefore, their residuals are large. Finally, the samples located far from data majority in the space of the PRM factors and with large residuals from the model were the most dominant. They are called bad leverage samples because they can easily distort the underlying relationship. Four categories of samples were detected using the distance-distance plots shown in Fig. 5. In our study, the influential samples were identified and then handled.

In Table 3, a comparison between the classic PLS models and PLS models that were built after removing the most influential samples (identified using the PRM approach) is presented. Generally, one can observe an improved performance in terms of a few figures of merit for all of the robust PLS models presented in paragraph 3.4.

Concerning the robust calibration of Cu, twenty influential samples were found in the model set. One was flagged as bad leverage, while the remaining were high residual samples. The final model offered a superior performance over the classic PLS model. The RMS was nearly 3.5 times, while the RMSEP was twice better. The coefficients of determination computed for the model and test set samples were, in this case, above 0.86 (see Table 3 and Fig. 5a).

For the most controversial PLS model that described the content of Cr, its robust variant fits the data nearly two-fold better in terms of the RMS. Therefore, the overall trend of the model shown in Fig. 5b is appropriate and describes the data well after the fourteen influential samples were removed.

The final PLS model was built for Mn after removing the nineteen high residual samples from the model set. As a result, its fit and prediction abilities for the test samples were considerably improved. This observation was confirmed by all of the figures of merit that are presented in Table 3 and in Fig. 5c.

The same improvement trend was observed for the model that estimated the content of Fe. After removing the sixteen high residual samples from the model set and five high residual samples from the test set, the final performance was better (see Table 3 and Fig. 5d).

3.5. Comparison of the PLS models with the models described in the literature

It is also worthwhile to confront the performance of our PLS models with similar models described in the literature, but it is not straightforward. There are several explanations. Different instruments are used to collect spectra, and their spectral ranges vary. There is a considerable spectral variation due to soil sources and sample preparation. Modeled responses span very differently in calibration ranges. Calibration samples were also selected differently, and the prediction abilities of models are not assessed consistently. Considering these limitations, we compared the models' performance only for illustrative purposes. In Table 4, coefficients of determination are reported for the optimal PLS models constructed in this study (either classic or robust) and the models discussed in the literature (see Table 41.1 in Ref. [32]).

An unsatisfactory calibration model was obtained for Cd; however, in the literature, similar models in terms of coefficient of determination were also reported (see Table 4). For contaminated soils, the linear relationship is reinforced. Therefore, monitoring the content of Cd in cultivated soils using the NIR spectroscopy is challenging. The most spectacular result, compared with those published in the literature, was obtained for Cu (a robust model). Its R^2t considerably exceeds the average value and is larger than the one obtained from the best model (see Table 4). If we assume a comparable variation of residuals from the model observed for the calibration and test set samples, modified R^2t^* is equal to 0.94. Regarding the modeling of the content of Ni, our PLS model has average prediction performance considering the R^2t^* value. The robust model constructed for calibrating Cr content in terms of R^2t^* performs slightly better than most of the models reported in the literature (see Table 4). Moreover, the models built for Zn, Mn, and Fe offered better predictions for validation samples than the average value of the coefficients of determination describing the reference models reported in Table 4.

4. Conclusions

The results of this study indicated that the spectra of Haplic Luvisol soil samples, collected between 1100 and 2500 nm, contained much relevant information which can be used for monitoring Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe at low concentrations. Furthermore, considering the significant representation of Haplic Luvisol soils worldwide (ca. 5% of the total area) and in Poland (up to 44%) and their arable potential, the results are helpful in the international context. In particular, the discussed methodology supports the development of precision agriculture. Modeling spectral data is challenging and requires a careful preprocessing and identifying the influential samples. This can be done automatically by building models for differently preprocessed spectra and evaluating figures of merit presented as a function of model complexity (e.g., Ref. [45]). The influential samples must be detected and handled correctly using robust in the statistical sense calibration method, for instance, PRM. The most promising models were obtained for Cu (using PRM) and Pb (using the classic PLS). Their RMSEP expressed as a percentage of the response range were equal to 9.63% and 11.5%, while their coefficients of determination were 0.8615 and 0.5831, respectively. Moreover, compared to most models reported in the literature, the satisfactory predictions were obtained for Zn, Mn, and Fe, with coefficients of determination over at least 0.79 for validation samples.

Credit author statement

S.K.: Conceptualization, Investigation, Data curation, Writing – original draft, Writing – review and editing, Resources. M.D.: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review and editing, Software, Formal Analysis. H.C-M.: Data curation, Formal Analysis, Writing – original draft. I. S.: Formal Analysis, Writing – original draft; Writing – review and editing, Software, Formal Analysis. L.P.: Software, Validation, Writing – original draft. S.S.: Investigation, Writing – original draft. J.W. Investigation, Writing – original draft

Funding

The results presented in this article were obtained as part of a

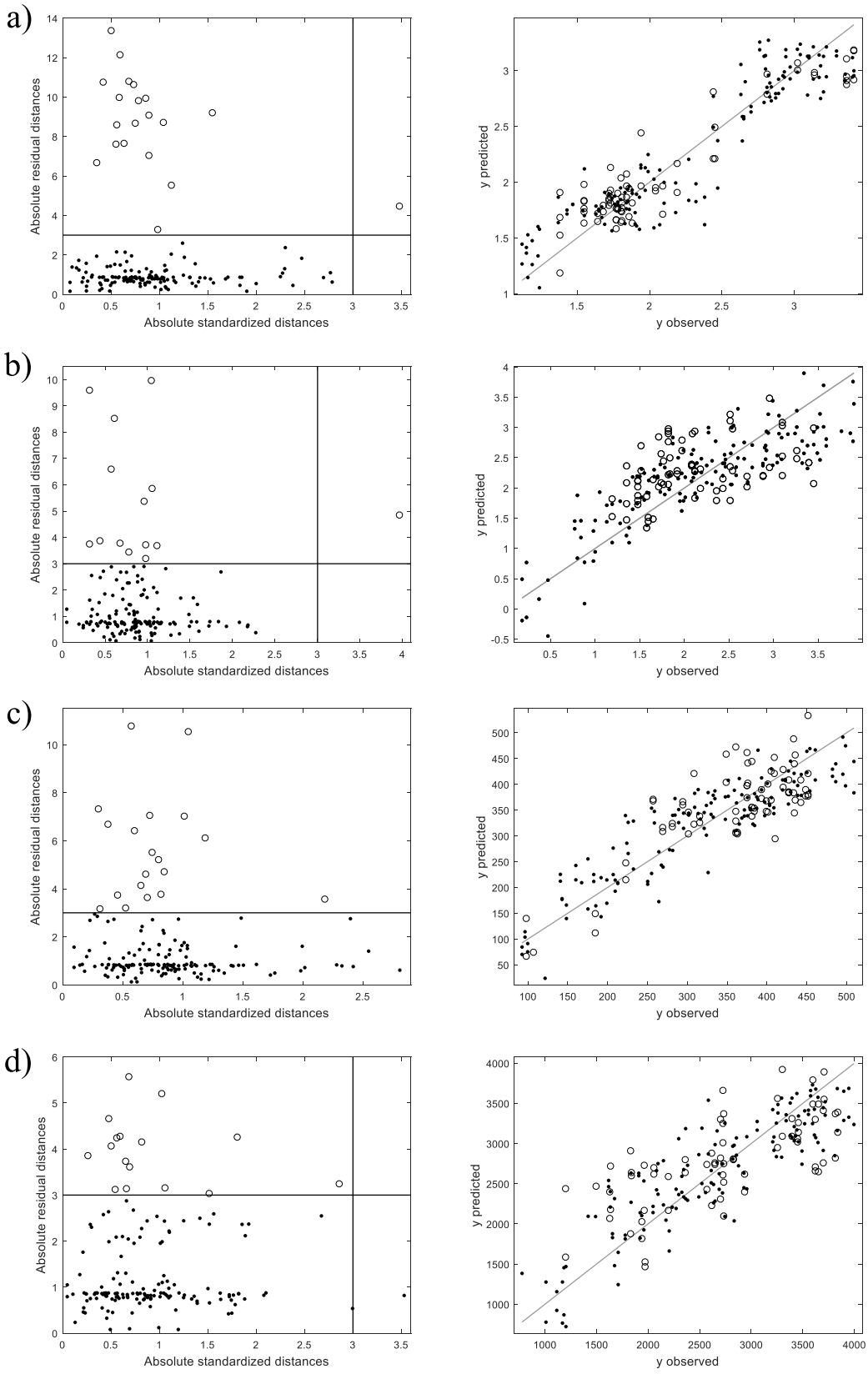


Fig. 5. The so-called distance-distance plots illustrating the influence of the samples from the model set on the model (residuals from the model versus Mahalanobis distance computed in the model space). The distance-distance plots were constructed based on the parameters obtained from the partial robust M-regression (PRM) model built with the optimal number of factors. After removing the samples that were flagged as influential from the model set, the final PLS models that described the concentrations of a) Cu, b) Cr, c) Mn and d) Fe in the Haplic Luvisol soil samples are visualized (the content of the elements is expressed in $mg \cdot kg^{-1}$).

Table 3
Comparison of the PLS models that were constructed to predict the content of Cu, Cr, Mn and Fe in the Haplic Luvisol soil samples before and after (*) the outlying samples were eliminated using partial robust M-regression. Along with the optimal preprocessing method, the number of factors that were included in the PRM and the final PLS models (*f*), several basic figures of merit are also reported. These are the root mean square error (RMS) that was calculated for the samples in the model set, the root mean square error of prediction (RMSEP) that was calculated for the samples in the model set, which are also expressed as the percentage of the calibration range and the corresponding coefficients of determination (R^2m and R^2t). The modified coefficient of determination that was calculated for samples from the test set, R^2t*, took into account the range of the responses that were observed for the samples in the model set.

Model	Preprocessing method(s) used to transform the NIR spectra	*f*	RMS	RMSEP	RMS [%]	RMSEP [%]	R^2m	R^2t	R^2t*
Cu	ISC + 1st derivative (window = 15, polynomial degree = 2)	5	0.8731	0.5032	17.39	10.02	0.4510	0.2948	0.9156
Cu*		6(5)	0.2458	0.2205	10.74	9.63	0.8645	0.8615	0.9455
Cr	detrending + 2nd derivative (window = 7, polynomial degree = 2)	1	1.0313	0.6397	18.82	11.67	0.0681	−0.0716	0.8341
Cr*		5(4)	0.4862	0.6170	13.07	16.59	0.7166	0.0032	0.7686
Mn	1st derivative (window = 11, polynomial degree = 2)	5	72.6538	62.9890	17.44	15.12	0.5604	0.4991	0.8472
Mn*		8(5)	48.1998	54.3216	11.57	13.04	0.8030	0.5922	0.8722
Fe	Detrending	9	521.2990	605.0577	16.20	18.81	0.6133	0.3671	0.7591
Fe*		12(9)	400.3545	536.7548	12.44	16.68	0.7533	0.4786	0.7861

Table 4
Calibration models built for estimating the content of eight elements (Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe) in Haplic Luvisol soil samples and the corresponding coefficients of determination for model (R^2m) and test sets (R^2t and R^2t*). Figures of merit (FOM) are reported for two groups of models: (1) - PLS models obtained in this study and (2) models described in the literature (see Table 41.1 in Ref. [32]). Asterisks denote that a robust calibration model was constructed for a given element.

Models	FOM	Cd	Cu*	Pb	Ni	Cr*	Zn	Mn*	Fe*
(1)	R^2m	0.30	0.86	0.74	0.72	0.72	0.84	0.80	0.75
	R^2t	0.26	0.86	0.58	−0.12	0.00	0.49	0.59	0.48
	R^2t*	–	0.94	0.89	0.85	0.77	0.87	0.87	0.79
(2)	R^2	0.30 ÷ 0.97	0.01 ÷ 0.71	0.01 ÷ 0.74	0.50 ÷ 0.92	0.44 ÷ 0.98	0.09 ÷ 0.93	0.01 ÷ 0.70	0.19 ÷ 0.90
	mean	0.70	0.32	0.67	0.86	0.75	0.62	0.42	0.62

comprehensive study that was financed by the University of Warmia and Mazury in Olsztyn, Faculty of Agriculture and Forestry, Department of Agricultural and Environmental Chemistry.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be available upon request once research plan concerning measured spectral profiles and samples will be completed.

Acknowledgment

MD, IS, and LP are grateful for fruitful discussions within the research team, named "*The impact of long-term pollutant emissions – environmental, health, and socio-cultural effects of non-ferrous metallurgy*", and acknowledge the support by the Research Excellence Initiative (program founded at the University of Silesia in Katowice, Poland).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2022.123749.

References

[1] M. Nocita, A. Stevens, B. van Wesemael, M. Aitkenhead, M. Bachmann, B. Barthès, E. Ben Dor, D.J. Brown, M. Clairotte, A. Csorba, P. Dardenne, J.A.M. Demattê, V. Genot, C. Guerrero, M. Knadel, L. Montanarella, C. Noon, L. Ramirez-Lopez, J. Robertson, H. Sakai, J.M. Soriano-Disla, K.D. Shepherd, B. Stenberg, E.K. Towett, R. Vargas, J. Wetterlind, Soil spectroscopy: an alternative to wet chemistry for soil monitoring, in: Adv. Agron., Elsevier, 2015, pp. 139–159, https://doi.org/10.1016/bs.agron.2015.02.002.

[2] M. Fuentes, C. Hidalgo, I. González-Martín, J.M. Hernández-Hierro, B. Govaerts, K. D. Sayre, J. Etchevers, NIR Spectroscopy: an alternative for soil analysis, Commun. Soil Sci. Plant Anal. 43 (2012) 346–356, https://doi.org/10.1080/00103624.2012.641471.

[3] A. McBratney, B. Whelan, T. Ancev, J. Bouma, Future directions of precision agriculture, Precis. Agric. 6 (2005) 7–23, https://doi.org/10.1007/s11119-005-0681-8.

[4] R. Casa, F. Castaldi, S. Pascucci, A. Palombo, S. Pignatti, A comparison of sensor resolution and calibration strategies for soil texture estimation from hyperspectral remote sensing, Geoderma 197–198 (2013) 17–26, https://doi.org/10.1016/j.geoderma.2012.12.016.

[5] R. Samiei Fard, R.S. Fard, H.R. Matinfar, Capability of vis-NIR spectroscopy and Landsat 8 spectral data to predict soil heavy metals in polluted agricultural land (Iran), Arabian J. Geosci. 9 (2016) 745.

[6] C. Gomez, P. Lagacherie, G. Coulouma, Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis–NIR data, Geoderma 189–190 (2012) 176–185, https://doi.org/10.1016/j.geoderma.2012.05.023.

[7] G. Zhang, D. Brus, F. Liu, X.-D. Song, P. Lagacherie, Digital Soil Mapping across Paradigms, Scales and Boundaries, Springer, 2019.

[8] A. Gholizadeh, L. Borůvka, R. Vašát, M. Saberioon, A. Klement, J. Kratina, V. Tejnecký, O. Drábek, Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: a case study, PLoS One 10 (2015), e0117457, https://doi.org/10.1371/journal.pone.0117457.

[9] X.-L. Xie, X.-Z. Pan, B. Sun, Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a copper smelter, Pedosphere 22 (2012) 351–366, https://doi.org/10.1016/S1002-0160(12)60022-8.

[10] C.M. Pandit, G.M. Filippelli, L. Li, Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression, Int. J. Rem. Sens. 31 (2010) 4111–4123, https://doi.org/10.1080/01431160903229200.

[11] Y. Wu, J. Chen, X. Wu, Q. Tian, J. Ji, Z. Qin, Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils, Appl. Geochem. 20 (2005) 1051–1059, https://doi.org/10.1016/j.apgeochem.2005.01.009.

[12] M. Todorova, A.M. Mouazen, H. Lange, S. Atanassova, Potential of near-infrared spectroscopy for measurement of heavy metals in soil as affected by calibration set size, Water Air Soil Pollut. 225 (2014) 2036, https://doi.org/10.1007/s11270-014-2036-4.

[13] M. Pietrzykowski, M. Chodak, Near infrared spectroscopy - a tool for chemical properties and organic matter assessment of afforested mine soils, Ecol. Eng. 62 (2014) 115–122, https://doi.org/10.1016/j.ecoleng.2013.10.025.

[14] G. Siebielec, G.W. McCarty, T.I. Stuczynski, J.B. Reeves, Near- and mid-infrared diffuse reflectance spectroscopy for measuring soil metal content, J. Environ. Qual. 33 (2004) 2056–2069, https://doi.org/10.2134/jeq2004.2056.

[15] R.A. Viscarra Rossel, S.R. Cattle, A. Ortega, Y. Fouad, In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy, Geoderma 150 (2009) 253–266, https://doi.org/10.1016/j.geoderma.2009.01.025.

[16] R.A. Viscarra Rossel, T. Behrens, E. Ben-Dor, D.J. Brown, J.A.M. Demattê, K. D. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aïchi, B.G. Barthès, H.M. Bartholomeus, A.D. Bayer, M. Bernoux, K. Böttcher, L. Brodský, C.W. Du, A. Chappell, Y. Fouad, V. Genot, C. Gomez, S. Grunwald, A. Gubler, C. Guerrero, C.

B. Hedley, M. Knadel, H.J.M. Morrás, M. Nocita, L. Ramirez-Lopez, P. Roudier, E. M.R. Campos, P. Sanborn, V.M. Sellitto, K.A. Sudduth, B.G. Rawlins, C. Walter, L. A. Winowiecki, S.Y. Hong, W. Ji, A global spectral library to characterize the world's soil, Earth Sci. Rev. 155 (2016) 198–230, https://doi.org/10.1016/j.earscirev.2016.01.012.

[17] B. Stenberg, R.A. Viscarra Rossel, A.M. Mouazen, J. Wetterlind, in: D.L. Sparks (Ed.), Visible and Near Infrared Spectroscopy in Soil Science, Adv. Agron., Academic Press, 2010, pp. 163–215, https://doi.org/10.1016/S0065-2113(10)07005-7.

[18] J.M. Soriano-Disla, L.J. Janik, R.A. Viscarra Rossel, L.M. Macdonald, M. J. McLaughlin, The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties, Appl. Spectrosc. Rev. 49 (2014) 139–186, https://doi.org/10.1080/05704928.2013.811081.

[19] Welcome to the "4 per 1000" Initiative. https://www.4p1000.org/. (Accessed 20 January 2022).

[20] K.D. Shepherd, M.G. Walsh, Infrared spectroscopy - enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries, J. Infrared Spectrosc. 15 (2007) 1–19, https://doi.org/10.1255/jnirs.716.

[21] S. Chakraborty, D.C. Weindorf, C.L.S. Morgan, Y. Ge, J.M. Galbraith, B. Li, C. S. Kahlon, Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy, J. Environ. Qual. 39 (2010) 1378–1387, https://doi.org/10.2134/jeq2010.0183.

[22] L. Kooistra, R. Wehrens, R.S.E.W. Leuven, L.M.C. Buydens, Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river floodplains, Anal. Chim. Acta 446 (2001) 97–105, https://doi.org/10.1016/S0003-2670(01)01265-X.

[23] S.A. Bowers, R.J. Hanks, Reflection of radiant energy from soils, Soil Sci. 100 (1965) 130–138.

[24] R.C. Dalal, R.J. Henry, Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry, Soil Sci. Soc. Am. J. 50 (1986) 120–123, https://doi.org/10.2136/sssaj1986.03615995005000010023x.

[25] E. Ben-Dor, A. Banin, Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties, Soil Sci. Soc. Am. J. 59 (1995) 364–372, https://doi.org/10.2136/sssaj1995.03615995005900020014x.

[26] Y. Song, F. Li, Z. Yang, G.A. Ayoko, R.L. Frost, J. Ji, Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China, Appl. Clay Sci. 64 (2012) 75–83, https://doi.org/10.1016/j.clay.2011.09.010.

[27] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemometr. Intell. Lab. Syst. 65 (2003) 97–112, https://doi.org/10.1016/S0169-7439(02)00107-7.

[28] K. Drab, M. Daszykowski, Clustering in analytical chemistry, J. AOAC Int. 97 (2014) 29–38, https://doi.org/10.5740/jaoacint.SGEDrab.

[29] I. Barra, S.M. Haefele, R. Sakrabani, F. Kebede, Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: recent advances–A review, TrAC Trends Anal. Chem. 135 (2021), 116166, https://doi.org/10.1016/j.trac.2020.116166.

[30] A.C. Dotto, R.S.D. Dalmolin, A. ten Caten, S. Grunwald, A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra, Geoderma 314 (2018) 262–274, https://doi.org/10.1016/j.geoderma.2017.11.006.

[31] R.A.V. Rossel, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma 158 (2010) 46–54, https://doi.org/10.1016/j.geoderma.2009.12.025.

[32] F. da Silva Terra, R. Rizzo, E. Ben Dor, J.A.M. Dematte, Soil sensing by visible and infrared radiation, in: Handb. -Infrared Anal., fourth ed., CRC Press, 2021.

[33] S. Nawar, S. Cipullo, R.K. Douglas, F. Coulon, A.M. Mouazen, The applicability of spectroscopy methods for estimating potentially toxic elements in soils: state-of-the-art and future trends, Appl. Spectrosc. Rev. 55 (2020) 525–557, https://doi.org/10.1080/05704928.2019.1608110.

[34] F. Riedel, M. Denk, I. Müller, N. Barth, C. Gläßer, Prediction of soil parameters using the spectral range between 350 and 15,000nm: a case study based on the Permanent Soil Monitoring Program in Saxony, Germany, Geoderma 315 (2018) 188–198, https://doi.org/10.1016/j.geoderma.2017.11.027.

[35] T. Kemper, S. Sommer, Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy, Environ. Sci. Technol. 36 (2002) 2742–2747, https://doi.org/10.1021/es015747j.

[36] T. Chen, Q. Chang, J.G.P.W. Clevers, L. Kooistra, Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy, Environ. Pollut. 206 (2015) 217–226, https://doi.org/10.1016/j.envpol.2015.07.009.

[37] T. Chen, Q. Chang, J. Liu, J.G.P.W. Clevers, L. Kooistra, Identification of soil heavy metal sources and improvement in spatial mapping based on soil spectral information: a case study in northwest China, Sci. Total Environ. 565 (2016) 155–164, https://doi.org/10.1016/j.scitotenv.2016.04.163.

[38] A.A. Paltseva, M. Deeb, E. Di Iorio, L. Circelli, Z. Cheng, C. Colombo, Prediction of bioaccessible lead in urban and suburban soils with Vis-NIR diffuse reflectance spectroscopy, Sci. Total Environ. 809 (2022), 151107, https://doi.org/10.1016/j.scitotenv.2021.151107.

[39] S. Krzebietke, E. Mackiewicz-Walec, S. Sienkiewicz, D. Załuski, Effect of manure and mineral fertilisers on the content of light and heavy polycyclic aromatic hydrocarbons in soil, Sci. Rep. 10 (2020) 4573, https://doi.org/10.1038/s41598-020-61574-2.

[40] E. Mackiewicz-Walec, S.J. Krzebietke, Content of polycyclic aromatic hydrocarbons in soil in a multi-annual fertilisation regime, Environ. Monit. Assess. 192 (2020) 314, https://doi.org/10.1007/s10661-020-08252-y.

[41] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.

[42] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, Anal. Chim. Acta 468 (2002) 91–103, https://doi.org/10.1016/S0003-2670(02)00651-7.

[43] S. Serneels, C. Croux, P. Filzmoser, P.J. Van Espen, Partial robust M-regression, Chemometr. Intell. Lab. Syst. 79 (2005) 55–64, https://doi.org/10.1016/j.chemolab.2005.04.007.

[44] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, TOMCAT: a MATLAB toolbox for multivariate calibration techniques, Chemometr. Intell. Lab. Syst. 85 (2007) 269–277, https://doi.org/10.1016/j.chemolab.2006.03.006.

[45] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst 141 (2016) 1060–1070, https://doi.org/10.1039/C5AN01656H.