SCIENTIFIC REPORTS
nature research

OPEN

Machine Learning Allows Calibration Models to Predict Trace Element Concentration in Soils with Generalized LIBS Spectra

Received: 11 February 2019
Accepted: 17 July 2019
Published online: 06 August 2019

Chen Sun[1], Ye Tian[2], Liang Gao[1], Yishuai Niu[3,4], Tianlong Zhang[5], Hua Li[5,6], Yuqing Zhang[1], Zengqi Yue[1], Nicole Delepine-Gilon[7] & Jin Yu[1]

Determination of trace elements in soils with laser-induced breakdown spectroscopy is significantly affected by the matrix effect, due to large variations in chemical composition and physical property of different soils. Spectroscopic data treatment with univariate models often leads to poor analytical performances. We have developed in this work a multivariate model using machine learning algorithms based on a back-propagation neural network (BPNN). Beyond the classical chemometry approach, machine learning, with tremendous progresses the last years especially for image processing, is offering an ensemble of powerful and constantly renewed algorithms and tools efficient for the different steps in the construction of a spectroscopic data treatment model, including feature selection and neural network training. Considering the matrix effect as the focus of this work, we have developed the concept of generalized spectrum, where the information about the soil matrix is explicitly included in the input vector of the model as an additional dimension. After a brief presentation of the experimental procedure and the results of regression with a univariate model, the development of the multivariate model will be described in detail together with its analytical performances, showing average relative errors of calibration (*REC*) and of prediction (*REP*) within the range of 5–6%.

Soil test occupies a particularly important place in environment-related activities, such as agriculture, horticulture, mining, geotechnical engineering, as well as geochemical or ecological investigations[1]. It becomes also crucial when an area needs to be decontaminated with respect to human activity-caused pollutions[2]. Such test may often concern elements, especially metals, since a number of them are considered as essential nutrients for plants and animals[3] and some others, heavy metals for example, are determined as toxic, even highly poisonous, in large amounts or certain forms for any living material[4]. It is therefore of great importance to develop techniques and methods for an efficient access to the elemental composition of soils. Established atomic spectroscopy techniques often offer good performances for quantitative elemental analysis in soils. Atomic absorption spectroscopy (AAS) offers limit of quantification (LOQ) in the order of ppm for soil samples prepared in solution[5]. Similar performances can be realized with inductively coupled plasma-optical emission spectrometry (ICP-OES)[6], while inductively coupled plasma-mass spectrometry (ICP-MS) presents for digested soil solutions, lower LOQ below 100 ppb for most of the elements found in soils[7]. Beside the abovementioned techniques which can rather be considered as laboratory-based ones characterized by the need of sample pretreatment with a certain degree of complexity, other techniques have been developed with significantly less requirement of sample preparation, so being better suited for *in situ* and online detections and analyses. Among them, X-ray fluorescence (XRF) allows determining concentrations of major and trace elements in soils[5,8]. A better performance has been demonstrated with total reflection X-ray fluorescence spectroscopy (TXRF)[9]. Techniques based on plasma emission spectroscopy,

[1]School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai, 200240, China. [2]Optics and Optoelectronics Laboratory, Ocean University of China, 266100, Qingdao, China. [3]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, 200240, China. [4]SJTU-Paristech Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, 200240, China. [5]College of Chemistry & Material Science, Northwest University, Xi'an, 710069, China. [6]College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an, 710065, China. [7]Institut des Sciences Analytiques, UMR5280 Université Lyon 1-CNRS, Université de Lyon, 69622, Villeurbanne, Cedex, France. Correspondence and requests for materials should be addressed to J.Y. (email: jin.yu@sjtu.edu.cn)

such as spark-induced breakdown spectroscopy (SIBS), have been developed to enhance the analytical capability of light elements like carbon[10]. Recent developments focus on laser-induced breakdown spectroscopy (LIBS), a laser ablation-based plasma emission spectroscopy[11]. The general attractive features of LIBS include direct laser sampling and excitation without need of complex sample pretreatment, stand-off excitation and detection capabilities, and high sensitivity for multi-elemental detection and analysis, for heavy as well as light elements.

LIBS analysis of soil has contributed to several important aspects of the soil test. Total carbon quantification in soil has been reported with portable LIBS systems for CO_2 leakage from underground storage of greenhouse gases[12,13] and for carbon cycle study in Amazonian forest[14,15]. Analysis of soil nutrients and fertilizer-related soil pollutions is another area covered by LIBS with the analysis of relevant elements such as P, N, S, Mg, Ca, K, Zn, Cu, Fe, Mn, Na[16–18]. LIBS technique and associated data treatment methods have also been developed for monitoring and analyzing metals, especially heavy metals in polluted soils, showing good performances for elements such as Ba, Co, Cr, Cu, Mn, Ni, Pb, V, and Zn[19–22]. Although the importance of the targeted applications leaves no room for doubt, the above demonstrations have not yet today led to large scale applications in real situations. The limited measurement precision and accuracy[23] that can be offered by a LIBS instrument would represent a bottleneck issue for applications of the technique, especially in the case of soil test. Indeed, the quantitative analysis capability of LIBS is still considered as its Achilles' heel[24].

In particular, for soil test, the precision and the accuracy of the measurements can be affected by a mediocre sample-to-sample repeatability of different measurements on (i) samples of a given type of soil and (ii) samples from different types of soil. Such repeatability is greatly influenced by the complex nature of laser-soil interaction, which depends upon both the laser characteristics and the physical and chemical properties of the analyzed soil sample[25]. In a measurement, any uncontrolled change of the conditions of laser-sample interaction (laser pulse energy, laser pulse focusing, laser pulse space and/or temporal profile …) can lead to changes in the property of generated plasma (ablation rate, atomization yield, excitation temperature…), causing the so-called emission source noise[26]. Furthermore, the inhomogeneity of a soil sample, even prepared in pellet after being ground into fine particles (of about 100 µm in size), can also induce changes in the plasma property and thus contribute to the emission source noise[27], when different positions on the surface of a soil sample pellet are ablated by laser. In other frequent cases of analyzing different types of soils, the change of plasma physical property under the same experimental condition because of the change of sample matrix, more specifically refers to the matrix effect, which leads the emission intensity of a given element to change according to its compound speciation in the sample and the composition of the soil[28]. Although the matrix effect represents a general issue in LIBS[29,30], its influence in analysis of soil becomes much more pronounced because of the complex physical and chemical properties and the corresponding wide range of different types of soils[31,32].

It is therefore crucial, for LIBS analysis of soils, to reduce and correct fluctuations of spectral intensity caused by the emission source noise and the matrix effect. Judicious sample preparation and correct use of internal reference may lead to significant improvement of the repeatability of a LIBS instrument, thus the precision and the accuracy of the measurements[33], although such preparation is not always possible nor efficient in the case of soil analysis because of the abovementioned complexity of soil and the practical constraints related to *in situ* and/or online measurements. Post-acquisition data treatment remains often the only efficient way for analytical performance improvement. Multivariate regressions based on chemometry, principally partial least-squares regression (PLSR) and neuronal networks analysis (NNA), have been demonstrated being able to provide robust calibration models for soil samples, with furthermore a reduced dependence of such models on the specific soil physical and chemical properties[34–39]. These demonstrations certainly leave rooms for improvements.

In this work, we used machine learning approach to significantly improve the data processing of LIBS spectra of soils, with a particular concern in the establishment of a soil-independent calibration model able to efficiently take into account samples from different types of soil, and in the same time to significantly reduce the influence of emission source noise. A key point is the introduction the concept of generalized spectrum which includes usual spectral intensities and additional parameters containing the information about the sample (type of soil, sample preparation method…). Machine learning, with the tremendous progresses recently made especially in image processing, is offering an ensemble of powerful and constantly renewed algorithms and tools efficient for the different steps in the construction of a model for spectroscopic data treatment, including feature selection and neural network training. Our results obtained in this work demonstrate the efficiency of such approach. In the following, we will first present the raw experiment data and the analytical performances with a univariate calibration model. The principle of the developed multivariate data processing method is then presented in detail. The analytical performances with the developed multivariate calibration model are described and compared with those of the univariate model. By such comparison, we emphasize the satisfactory and impressive reductions of the matrix effect and the emission source noise allowed by the developed machine learning-based multivariate calibration model, before we deliver the conclusions of the paper.

Experimental Data and Analytical Performances with a Univariate Calibration Model

Laboratory-prepared reference samples. In the experiment, LIBS measurements were performed on 4 types of soils: NIST 2710 (called N1 in this paper), NIST 2587 (N2), collected 1 (U1) and collected 2 (U2). The corresponding powders were first spiked using a standard reference solution of silver in order to prepare a set of laboratory-prepared reference samples for each soil with 7 different Ag (as the analyte) concentrations in the range from 20 to 840 ppm weight. Pellets were then prepared with doped powders for every Ag concentration (Co_{ti}) of each of the 4 soil types (t). Table 1 shows the concentrations in ppm weight of the pellets and their role in the construction and validation of the calibration model. More details about sample preparation are provided in the section "Methods".

Soil type t	Calibration sample set (5 concentrations in ppm weight each soil, Co_{ti})	Validation sample set (2 concentrations in ppm weight each soil, Co'_{ti})
	(i) 1, 3, 4, 5, 7	(i) 2, 6
NIST 2710 (N1) *initially containing 40 ppm weight of Ag*	60, 140, 240, 440, 840	90, 640
NIST 2587 (N2)	20, 100, 200, 400, 800	50, 600
Collected 1 (U1)		
Collected 2 (U2)		

Table 1. Silver concentrations in ppm weight of the prepared sample pellets with their roles in the construction and validation of the univariate calibration model.

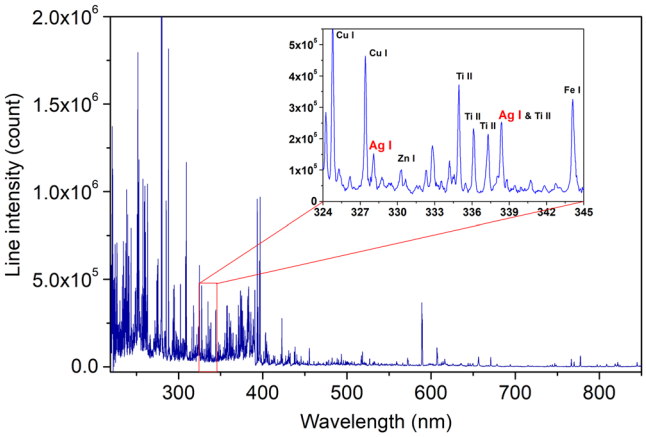

Figure 1. Typical replicate-averaged spectrum of soil sample. In the inset, the detailed spectrum around the Ag I 328.1 nm line is shown. Sample used to obtain the spectrum: t = N1, initially containing the following elements: Cu (3420 ppm), Zn (4180 ppm), Ti (3110 ppm), Fe (43200 ppm), and Ag (40 ppm), 400 ppm of Ag was additionally spiked into the sample (Co_{ti} = 440 ppm).

Raw experimental data. Six replicate (j) spectra were taken for each pellet. A spectrum was an accumulation of 200 laser shots distributed over 10 distinguished sites on the sample surface ablated each by 20 successive laser pulses. An individual spectrum can thus be notated by $\overrightarrow{I_{ij}^{t}}$ (for the j^{th} replicate measurement on the sample with analyte concentration Co_{ti} prepared with the soil type t). A typical replicate-averaged spectrum is presented in Fig. 1, showing in particular the emission line chosen for Ag emission intensity measurement, the Ag I 328.1 nm line.

Quantitative analysis performances with univariate calibration. Calibration curves based on univariate regression are constructed by representing the intensities of Ag I 328.1 nm line, $\overrightarrow{I_{ij}^{t}}$ (Ag I 328.1 nm), as a function of the Ag concentrations of the corresponding calibration samples. As we can see in Fig. 1, this line is enough intense and seems free of interference with other lines, its intensity is still not particularly high to avoid significant self-absorption to occur. A linear regression of the line intensities as a function of the Ag concentrations for a given soil type results in a soil-specific calibration curve for each of the 4 analyzed soils as shown in Fig. 2. The error bars in the figure are standard deviations ($\pm\sigma_{I_i^t}$) of the intensities calculated for the 6 replicate measurements performed for each sample pellet. Large error bars associated to the line intensities indicate the effect of the emission source noise as we discussed above. The same noise leads to the reduced values of the determination coefficient r^2 with respect to the unit for each of the 4 soils. In addition, the slopes of the calibration curves are significantly different, showing an obvious matrix effect in LIBS analysis of the 4 soils. By merging the line intensities from the calibration samples of the 4 types of soil, a soil-independent calibration curve can be established. For this purpose, the intensity data from the 4 soils are plotted as a function of Ag concentrations in a same figure as shown in Fig. 3, and a linear regression of the data leads to the soil-independent univariate calibration curve. We can see in this figure that a large dispersion of emission intensities for samples with a given Ag concentration due to the matrix effect leads to a much reduced r^2 value for the soil-independent calibration curve.

Line intensities from the validation samples in Table 1 are then used to evaluate the accuracy and the precision of prediction using the established calibration models. These intensities are represented by crosses in Figs 2 and 3. Table 2 sums up the figures of merit[35,40] of quantitative analysis performances using the univariate model with both the soil-specific and the soil-independent calibration curves, where *REC*(%) is average relative error of calibration, *REP*(%) average relative error of prediction, *RSD*(%) relative standard deviation of the predicted concentrations, and *LOD*(ppm) limit of detection. The definitions of the above quantities are given in the section "Methods".

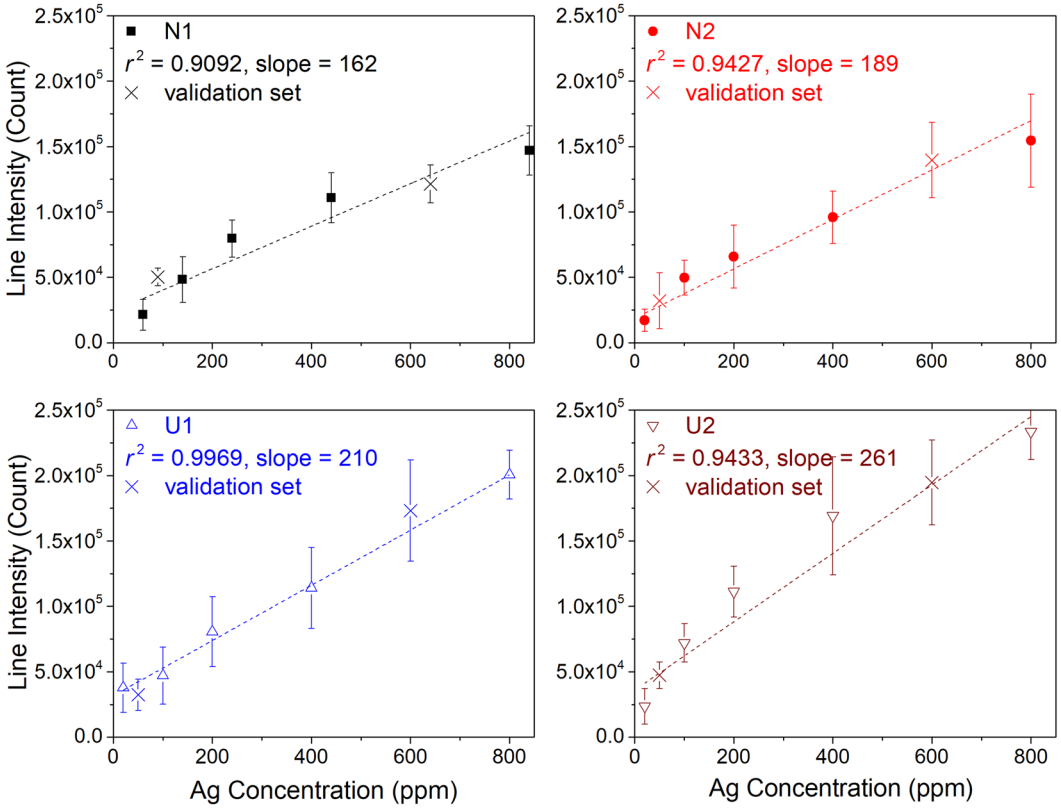

Figure 2. Intensities of Ag I 328.1 nm line of the calibration samples as function of Ag concentrations and soil-specific univariate calibration curve (dashed lines in the figures) of Ag with this line respectively for the 4 analyzed soils. Line intensities from the validation set are represented by crosses, they do not participate in the construction of the calibration models. The error bars are calculated for each line intensity with the standard deviation among the 6 replicate measurements ($\pm\sigma_{I_i^t}$).

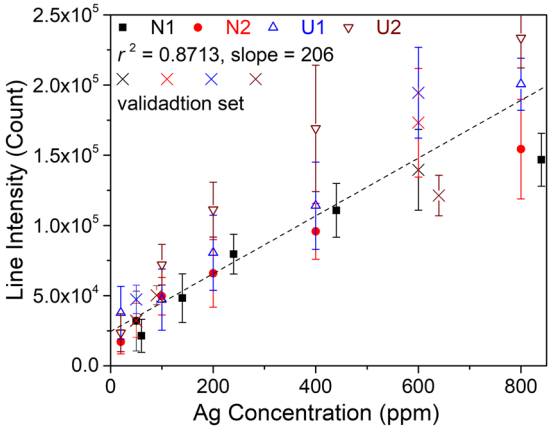

Figure 3. Similar presentation of the experimental data as in Fig. 2, but with line intensities from all the 4 soils merged in a same figure and a soil-independent univariate calibration curve (dashed line).

We can see in Table 2 that the soil-specific calibration curves have fair r^2 values, while their slopes are significantly different, as also shown in Fig. 2, indicating significant influences of both emission source noise and matrix effect. The accuracies of calibration and prediction, indicated respectively by *REC* (mean value 18.28%) and *REP* (mean value 37.07%) of the soil-specific calibration curves are not satisfactory for quantitative analysis. This is due to a limited measurement repeatability from one sample to another. On the other hand, a limited repeatability of replicate measurements leads to an unsatisfactory prediction precision (mean value of $RSD = 42.35\%$), as well as a quite high *LOD* (mean value of $LOD = 24.23$ ppm), compared to standard LIBS measurement performances for solid samples.

Calibration type	Soil	Calibration model				Validation	
		r^2	Slope	*REC*(%)	*LOD*(ppm)	*REP*(%)	*RSD*(%)
Soil-specific	N1	0.9092	162	26.35	27.57	30.75	34.95
	N2	0.9427	189	15.15	18.47	23.99	82.29
	U1	0.9969	210	16.71	31.90	54.43	25.58
	U2	0.9433	261	14.91	18.96	39.09	26.56
	Mean	0.9480	206	18.28	24.23	37.07	42.35
Soil-independent	N1	0.8713	206	32.15	23.83	15.42	27.59
	N2					30.33	55.86
	U1					36.50	24.37
	U2					42.54	50.32
	All					31.20	51.20

Table 2. Figures of merit of quantitative analysis of the univariate calibration model with the both soil-specific and soil-independent calibration curves. The soil-independent calibration curve is constructed using the spectra of the calibration samples from all the 4 soils. It is validated by the validation spectra of specific soils as well as by the validation spectra of all the 4 soils.

When the soil-independent calibration curve is considered, we can find a degraded r^2 value, indicating a significant influence of matrix effect. The determined slope of the calibration curve logically corresponds to the average value of the slopes of the 4 soil-specific calibration curves. We can also see that the mentioned matrix effect also degrades the accuracy of calibration (*REC*) due to a larger dispersion of the line intensities participating in the construction of the soil-independent calibration curve. The degraded calibration accuracy becomes comparable to the prediction accuracy (*REP*) when all the soil types are included for model validation, since both of them are directly influenced by the matrix effect. Due to the same influence, the prediction precision (*RSD*) with all the soil types is degraded compared to the mean value of the soil-specific calibration curves. At the same time, the limit of detection (*LOD*) does not record significant change with respect to the soil-specific calibration models, since it is more sensitive to the fluctuation of replicate measurements due to the emission source noise.

Analytical Performances with Multivariate Calibration Model

Principle of the developed multivariate calibration model. From the above results with univariate calibration models, we can see that the analytical performances are affected by the both matrix effect and emission source noise. The developed multivariate model is therefore designed to deal with different types of soil, and in the same time, such model should efficiently reduce the dispersion of analytical results due to any change and fluctuation of experimental condition. The idea is to explicitly include the information about soil type in the input variables for the training and validation of the calibration model. More specifically, for a given reference sample with known analyte concentration Co_{ti} prepared from the soil type t, the j^{th} replicate LIBS measurement generates a spectrum which can be presented as a vector in the form of $\overrightarrow{I_{ij}^t} = (I_{ij1}^t, I_{ij2}^t, \ldots I_{ijk}^t, \ldots I_{ijM_1}^t)$, where M_1 is the dimension of the spectrum (pixel number of a raw spectrum or the number of contained intensities in a pretreated spectrum). Such physical spectrum can be concatenated with an ensemble of M_2 variables, $(Ma_1^t, Ma_2^t, \ldots Ma_{M_2}^t)$, representing the properties of the sample. The result is a generalized spectrum with $M_1 + M_2$ generalized intensities,

$$\overrightarrow{I_{ij}^{t,Genral}} = (I_{ij1}^t, I_{ij2}^t, \ldots I_{ijM_1}^t, Ma_1^t, Ma_2^t, \ldots Ma_{M_2}^t). \quad (1)$$

Such spectrum is considered in the method as a vector in a hyperspace of $M_1 + M_2$ dimensions. A generalized module, $|\overrightarrow{I_{ij}^{t,General}}|_{General}$, can thus be attributed to it for formally representing the concentration of the analyte (silver for instance) in the corresponding sample. Such module cannot be calculated using a simple mathematical function. The physical correlation between the generalized spectrum and the concentration of the analyte can only be expressed as a mathematical relation of mapping:

$$f: \mathbb{R}_+^{M_1+M_2} \rightarrow \mathbb{R}_+, \ \overrightarrow{I_{ij}^{t,General}} \mapsto Co_{ti} = \left|\overrightarrow{I_{ij}^{t,General}}\right|_{General}. \quad (2)$$

In our experiment, a machine learning algorithm is used through a training process, to establish the mapping between the collection of generalized spectra and the ensemble of element concentrations of the corresponding reference samples. The result of such training process leads to a calibration model which is able to predict the concentration of the analyte in a validation sample when its generalized spectrum is used as the input of the model. The physical basis of the mapping between the generalized spectra and the elemental concentrations is the interaction between the different species in a laser-induced plasma, which leads to the correlation of the concentration of a specific element contained in the plasma to the whole plasma emission spectrum. In our experiment, tests have been performed to establish the importance of the additional dimension in a generalized LIBS spectrum for the training of the model and the effectiveness of this dimension for the correction of the matrix effect. More detailed information is provided in the section "Methods".

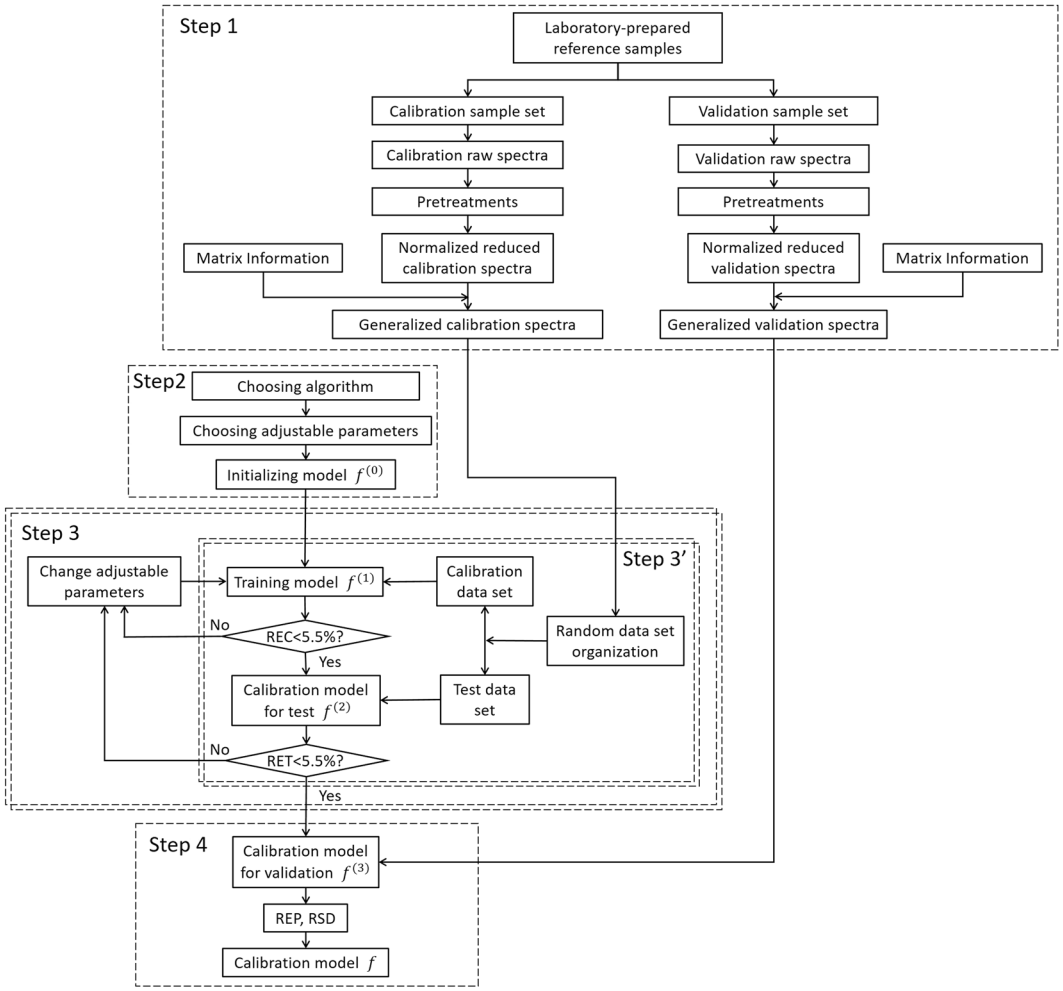

Figure 4. Flowchart for the buildup of the multivariate calibration model. The steps contained in double dashed line rectangles are repeated within a conditional loop.

Soil type t	Calibration sample set (6 concentrations in ppm weight each soil, Co_{ti}) (i) 1, 2, 3, 5, 6, 7	Validation sample set (1 concentration in ppm weight each soil, Co'_{ti}) (i) 4
NIST 2710 (N1)	60, 90, 140, 440, 640, 840	240
NIST 2587 (N2)	20, 50, 100, 400, 600, 800	200
Collected 1 (U1)		
Collected 2 (U2)		

Table 3. Organization of the experimental data for the buildup of the multivariate calibration model.

Implementation of the method: flowchart of training and validation of the calibration model.

Figure 4 shows the flowchart of the developed multivariate calibration method. Several steps can be distinguished in a successive way.

Step 1. Data set organization, pretreatment and formatting. The experimental data are organized in this step in the way shown in Table 3, where we can see that for each soil type, 6 pellets with different analyte concentrations are selected for the calibration sample set ($i \in \{1, 2, 3, 5, 6, 7\}$), and the rest one ($i \in \{4\}$) for the validation sample set. In order to have a clear vision of the structure of the experimental data, they are presented within a rectangular parallelepiped as shown in Fig. 5. An individual raw spectrum, $\overrightarrow{I_{ij}^{t}} = \left(I_{ij1}^{t}, I_{ij2}^{t}, \ldots I_{ijk}^{t}, \ldots I_{ijM_0}^{t}\right)$, is represented in the rectangular parallelepiped by a cube with a set of given values of (t, i, j), here the index k is used to indicate a pixel in the spectrum: $1 \leq k \leq M_0$, $M_0 = 21915$ is the pixel number of a raw spectra, which physically corresponds to the spectral range of the used spectrometer, $220\,\text{nm} \leq \lambda \leq 850\,\text{nm}$.

Pretreatment is performed on the raw spectra, which consists in (i) normalization and (ii) feature selection. The normalization, applied to all the raw spectra of the laboratory-prepared reference samples, is a simple operation which transforms the intensity rang of each pixel of all the raw spectrum into the interval between 0 and 1:

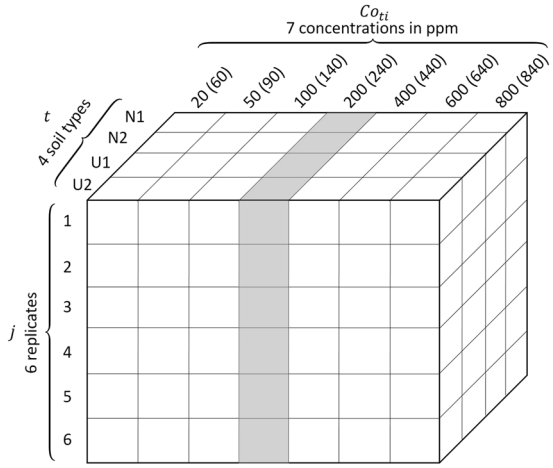

Figure 5. Structure of the experimental data with 4 soil types (t), 7 analyte concentrations (Co_{ti}) for each soil type and 6 replicate LIBS measurements (j) for a sample pellet of given soil type and analyte concentration. The samples with 200 (240) ppm analyte concentration are chosen as the validation sample set, the rest as the calibration sample set.

$$I_{ijk}^{t,norm} = \frac{I_{ijk}^{t} - I_{k}^{min}}{I_{k}^{max} - I_{k}^{min}} \text{ for } 1 \leq k \leq M_0, \tag{3}$$

where $i \in \{1, 2, 3, 4, 5, 6, 7\}$, I_k^{min} and I_k^{max} are respectively the minimum and the maximum of the pixel k among the same pixels of all the individual spectra ($4 \times 7 \times 6 = 168$ spectra). Such normalization reduces the contrast among the pixel intensities of a raw spectrum, which can initially exceed one order of magnitude for a large part of the pixels as shown in Fig. 1. Since one could expect smaller variations among the intensities of the different individual spectra for a given pixel, unless a physical reason, variation of the analyte concentration for example, makes them to change in a correlated way. After the normalization, all the pixels of an individual spectrum, whatever their initial physical intensities, should contribute in a more statistically equivalent way, to characterize it with respect to the other ones.

Feature selection is performed by applying the SelectKBest algorithm[41] to the normalized spectra of the calibration sample set. The principle consists in selecting and keeping in an individual spectrum for the further processing, pixel intensities with high enough correlation with the series of analyte concentrations of the calibration sample set. Such correlation is calculated in the algorithm with a score function $Score(k_0)$:

$$Score(k_0) = \mathcal{D}\frac{Corr(k_0)^2}{1 - Corr(k_0)^2}, \text{ for } 1 \leq k_0 \leq M_0, \tag{4}$$

$$Corr(k_0) = \frac{Cov\left(\left\{I_{ijk_0}^{t,norm}\right\}, \{Co_{ti}\}\right)}{\sqrt{Var(I_{ijk_0}^{t,norm})Var(Co_{ti})}}, \tag{5}$$

$$Cov\left(\left\{I_{ijk_0}^{t,norm}\right\}, \{Co_{ti}\}\right) = \frac{1}{\mathcal{D}}\sum_{i\ in\ S}\sum_{t=1}^{4}\sum_{j=1}^{6}\left[(I_{ijk_0}^{t,norm} - \overline{I_{k_0}^{norm}})(Co_{ti} - \overline{Co})\right], \tag{6}$$

$$Var(I_{ijk_0}^{t,norm}) = \frac{1}{\mathcal{D}}\sum_{i\ in\ S}\sum_{t=1}^{4}\sum_{j=1}^{6}(I_{ijk_0}^{t,norm} - \overline{I_{k_0}^{norm}})^2, \tag{7}$$

$$Var(Co_i) = \frac{1}{\mathcal{D}}\sum_{i\ in\ S}\sum_{t=1}^{4}\sum_{j=1}^{6}(Co_{ti} - \overline{Co})^2, \tag{8}$$

where $S = \{1, 2, 3, 5, 6, 7\}$, $\mathcal{D} = 6 \times 4 \times 6 = 144$ is the number of the individual spectra in the calibration sample set, $\overline{I_{k_0}^{norm}}$ stands for the mean value of normalized intensity of the pixel k_0 (hence the corresponding wavelength) with respect to the measurement replicates, the soil types and the prepared concentrations of the calibration samples; and $\overline{Co}$ refers to the mean value of the prepared concentrations of the calibration samples. In the case of model training with a given type of soil, the above sums with respect to t reduce to a single corresponding term. The threshold value applied to $Score(k_0)$ for feature selection takes into account the number of individual spectra

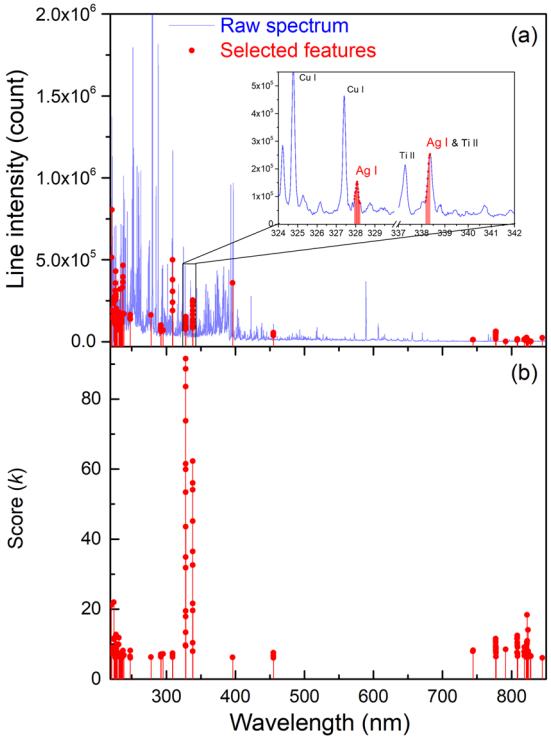

Figure 6. (**a**) Spectrum of the selected features (in red) with in the inset, those corresponding to the 2 Ag I lines, the raw spectrum (in light blue) is also shown for comparison; (**b**) Spectrum of the SelectKBest scores.

included in the calibration sample set, for instance $\mathcal{D} = 6 \times 4 \times 6 = 144$. In this work, 150 pixels were selected over the initial 21915 ones, so that the reduced normalized spectrum have a dimension of $M_1 = 150$. Such dimension is comparable to the total number of spectra used in the calibration sample set, reducing thus the risk of overfitting.

The spectrum of selected features is shown in Fig. 6. We can see that pixels (or equivalently wavelengths) receiving the highest scores are concentrated in the spectral range from 327 nm to 339 nm as shown in Fig. 6(b), and correspondingly in Fig. 6(a). We can identify 2 neutral silver lines: Ag I 328.1 nm with a NIST relative intensity[42] of 55000 and Ag I 338.3 nm with a smaller NIST relative intensity of 28000. The experimental spectrum shows however $I(\text{at } 338.3\,\text{nm}) > I(\text{at } 328.1\,\text{nm})$. A detailed inspection in the NIST Atomic Spectra Database[42] shows the presence of a relatively intense titanium ionic line, Ti II 338.4 nm line, with a NIST relative intensity of 7100. Since titanium represents an important trace element in soil, this line can therefore significantly interfere with the Ag I 338.3 nm line. This is why the pixels in the Ag I 328.1 nm line receive higher scores, while a part of the pixels corresponding to the Ag I 338.3 nm line receive lower scores, and those pixels are all situated in the low frequency side of the intensity peak around 338.3 nm. A second zone where high score features are found extends from 750 nm to 850 nm, where we can remark the correspondence between the selected features and the lines emitted by oxygen and nitrogen atoms, which are mainly contributed by the ambient gas. Correlation between the elements from the ambient gas (especially O and N) and an element to be detected in the sample has been studied in our previous work[33]. Clear physical interpretation of the selected spectral features demonstrates the significance of the used SelectKBest algorithm.

In our experiment, the type of soil is the only significant information which distinguishes the 4 soils (with the same preparation procedure), it is thus concatenated with the normalized and reduced spectrum to form generalized spectrum: $\overrightarrow{I_{ij}^{t,general}} = \left(I_{ij1}^{t,norm}, I_{ij2}^{t,norm}, \ldots I_{ij150}^{t,norm}, Ma_1^t\right)$. Numerical values of $Ma_1^t = 1, 2, 3$ and 4 are arbitrarily chosen for representing the 4 soil types, N1, N2, U1 and U2 respectively.

Step 2. Model initialization. Back-propagation neural network (BPNN)[43] is chosen in this work to provide the algorithm which maps the generalized spectra and the corresponding analyte concentrations. Such choice is motivated by the fact that BPNN corresponds rather to an algorithm in machine learning than a specific neural network (NN). The back-propagation procedure (i.e., the application of the chain rule to calculate derivatives of composite functions) allows training a neural network using gradient-type optimization algorithms such as stochastic gradient descent (SGD), which is one of the most successful and widely used training algorithms in machine learning. The number of hidden layers *n_layers*, the number of nodes in a hidden layer *n_nodes*, the learning rate and the maximum epochs of BPNN are selected as the externally adjustable parameters to optimize the performance of the model. The model starts with its default state denoted by $f^{(0)}$.

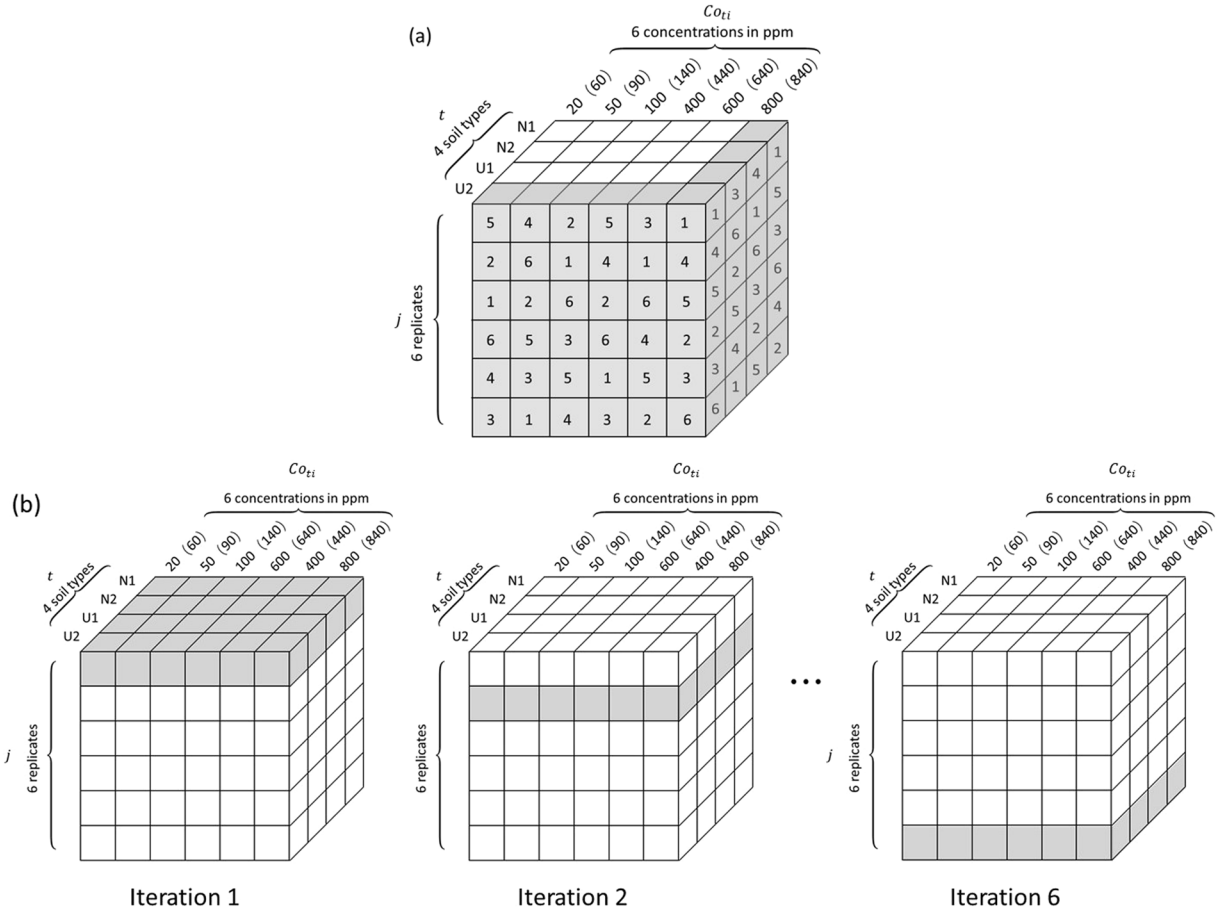

Figure 7. (**a**) A randomly and independently arranged data configuration among $(6!)^{24}$ possible and statistically equivalent ones; (**b**) For a given randomly and independently arranged data configuration, illustration of a 6-fold cross-validated training iteration, with the cubes in grey representing the test data set.

Step 3. Model built-up loop: training of the algorithm and optimization of its externally adjustable parameters. This is the central body of the model construction process which comprises an internal and an external loop as shown in Fig. 4.

The internal loop (Step 3′ in Fig. 4) is devoted to train the algorithm in such way for an input individual generalized spectrum $\overrightarrow{I_{ij}^{t,General}}$, the resulted generalized module becomes as close as possible to the targeted analyte concentration Co_{ti}. In considering the statistical equivalence and experimental fluctuation among the replicate measurements (j is a dummy index) and the matrix effect due to different soils, and in order to fulfill the requirements for the model to tackle both the experimental fluctuation and the matrix effect, the training process is implemented in the following way:

(i). Randomly permuting among j of all the data columns of given t and Co_{ti}, in order to randomly and independently fix the arrangement of all the 24 columns of replicate spectra as shown in Fig. 7a, with the arrangements visible for all the columns in the two surfaces of $t = U2$ and $Co_{ti} = 800(840)$ ppm of the date cube;

(ii). For one of the data configurations (in total $(6!)^{24}$ possible and statistically equivalent ones) generated in the above way, performing a dynamic cross validation training process of 6 iterations. In each of these iterations, successively one layer of the data, for example the top layer, then the second, then the third…, up to the bottom one, is considered as the test data set, while the rest as the calibration data set as shown in Fig. 7b. In such iteration, the algorithm corresponding to a training model, $f^{(1)}$, is trained, with the calibration data set, in order for the output generalized modules of the individual generalized spectra to be as close as possible to the silver concentrations of the corresponding targets. These iterations generate 6 different BPNNs.

(iii). In the end of the above 6-fold iterative training and cross validation process, another randomly and independently arranged data configuration is generated for a new 6-fold iterative cross validation training of the algorithm. In the experiment, we fixed the considered number of randomly and independently arranged data configurations to 10, because a larger number of data configurations would not significantly enrich useful information that we can extract from the given ensemble of raw experimental spectra. In the

end of the 10-data-configuration training, 60 different BPNNs are generated.

(iv). The average relative error of calibration (*REC*) is calculated. If the value is larger than the fixed threshold, the process goes back to the training step of $f^{(1)}$. Otherwise a calibration model for test, $f^{(2)}$, is generated.

(v). $f^{(2)}$ is then tested by the test data set in a similar way as the above training process. The average relative error of test (*RET*) is calculated.

(vi). The resulted *REC* and *RET* are compared to the fixed threshold values. If they, or one of them, are larger than the threshold value(s), the process goes to the external loop. Otherwise a calibration model for validation, $f^{(3)}$, is generated.

In this experiment, the threshold values were fixed for 10-data-configuration resulted *REC* and *RET* both at 5.50%. This value was chosen to minimize the average relative error of prediction (*REP*) calculated in the validation process of the calibration model for validation, $f^{(3)}$, using generalized validation spectra, which were not involved in the model training process. Numerical experiments were thus necessary to determine these thresholds, even though the values could be intrinsically smaller if only the model training process in the step 3 is concerned.

The detailed definitions of *REC*, *RET* and *REP* for the assessment of the multivariate model are given in the section "Methods".

Then only one type of soil is under consideration, *t* takes a fixed valued among *N*1, *N*2, *U*1, *U*2.

The external loop of this step is aimed to optimized the externally adjustable parameters of the algorithm, BPNN for instance. The used method is grid-search parameter tuning, which is known as an efficient method of optimization for constructing a calibration model. In this method, for given ranges of the selected adjustable parameters, the performance of the model is evaluated for all the possible combinations of the adjustable parameters in an exhaustive way. The combination generating the best performance is retained. In our experiment, the ranges of the 2 externally adjustable parameters, *n_layers* and *n_nodes*, both positive integer, were respectively fixed being 1 to 2 and 3 to 8, 12 combinations were therefore evaluated.

When the values of *REC* and *RET* are simultaneously smaller than 5.50%, the iteration in the external loop stops. A calibration model for validation $f^{(3)}$ is generated as the output of the step 3. In our experiment, $f^{(3)}$ is obtained with the externally adjustable parameters of $n_{layers} = 1$, $n_{nodes} = 5$, learning rate = 0.2, and maximum epochs = 10000. The fact that the final BPNN structure is optimized with a single hidden layer corresponds well to the universal approximation theorem which states that a BPNN with 3 layers (the input, the hidden and the output layers) under mild assumptions on the activation function is enough for fitting any continuous function over a finite dimension compact set[44–46].

Step 4. Model validation with an independent validation sample set. The output model of the step 3, $f^{(3)}$, is validated in this step using generalized validation spectra obtained from the validation samples which is not involved in the model training process. Average relative error of prediction (*REP*) and average relative standard deviation (*RSD*) are calculated for individual generalized validation spectra of the validation sample set to respectively evaluate the prediction accuracy and precision of the model. After the validation, the final calibration model, *f*, is generated with the corresponding *REP* and *RSD*, indicating its performances.

Results and Discussions

Soil-specific and soil-independent calibration curves are respectively shown in Figs 8 and 9. We use here a similar presentation as in Figs 2 and 3 to easy the comparison with the univariate models. And the parameters showing the analytical performances of the multivariate models are presented in Table 4.

We can see that the soil-specific calibration curves exhibit all a r^2 value very close to the unit. This means that the multivariate models efficiently reduce the experimental fluctuation from a reference sample to another. The fluctuation from a replicate measurement to another for a given sample is also significantly reduced, which leads to a very small error bar on each predicted concentration. A direct consequence of such reduced fluctuations is a significant improvement of *LODs* from several tens ppm to around ppm. In coherence with the high r^2 values, the calibration accuracy is greatly improved and reaches now an impressive level of around 1% for *REC* and *RET*. The prediction capacity of the soil-specific calibration models is clearly reinforced with order-of-magnitude reduction for both *REP* and *RSD* compared to those of the univariate model. Such performance clearly fulfills the requirements of precise and accurate quantitative analyses.

When the calibration spectra from all the soils are used to build a calibration model, the soil-independent calibration curve is obtained as shown in Fig. 9. We can see a r^2 value very close to the unit as in the case of soil-specific multivariate calibration curves. This does not only mean an efficient improvement of the repeatability from a sample to another for a given type of soil, but more importantly shows the ability of the multivariate model to take into account the specific matrices between the different soils and to reduce the matrix effect. In fact, the data from the different types of soil can be fitted with a unique linear model with a determination coefficient r^2 very close to those of soil-specific calibration curves. The *LOD* allowed by the soil-independent multivariate model remains quite low in the order of 5 ppm. A slight increase of this value with respect to those of soil-specific calibration curves would indicate a residual matrix effect. The same residual matrix effect should contribute to slightly reduce the calibration accuracy compared to the soil-specific calibration curves, as indicated by the values of *REC* and *RET* in the order of 5% for the soil-independent model. Compared to the univariate model, the performance of the calibration curve is greatly improved with a matrix effect reduced within an acceptable level.

Concerning the prediction capacity of the multivariate soil-independent model, great improvements can be observed with respect to the univariate model for the accuracy as well as for the precision, although degradations are observed compared to the multivariate soil-specific models. Such degradations should be related to the above mentioned residual matrix effect, which would lead to, sometimes, unexpected large values of *REP* and *RSD* when

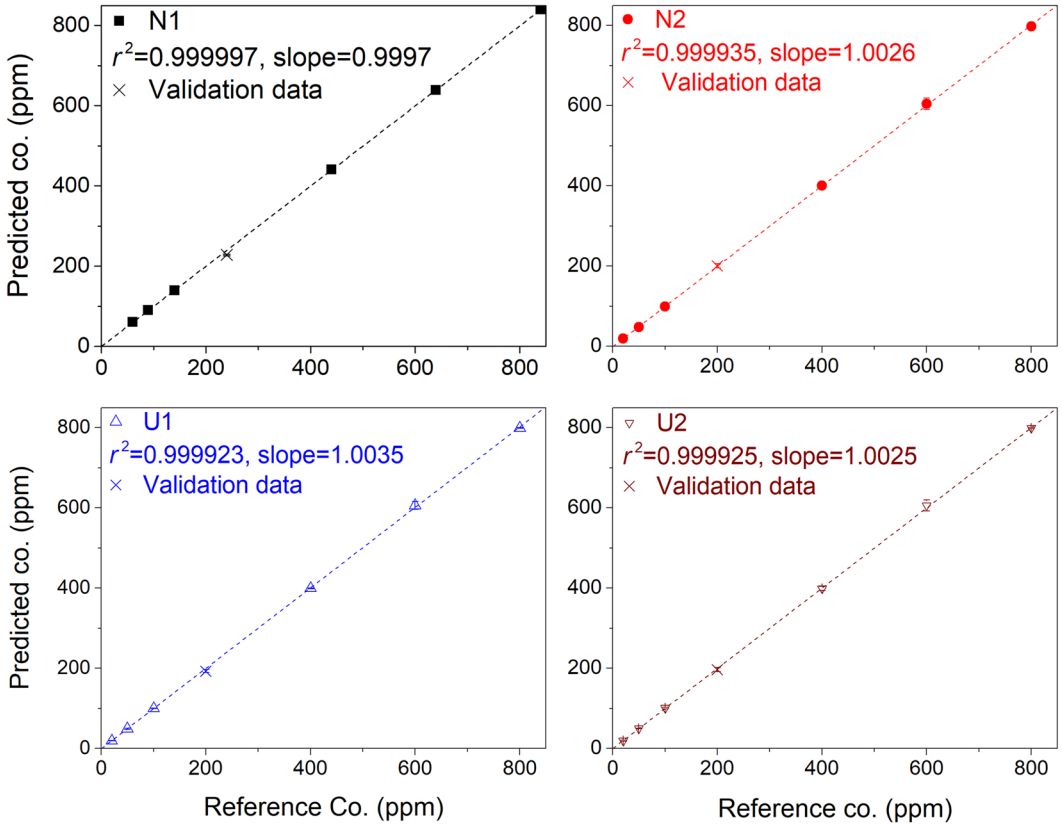

Figure 8. Model-predicted Ag concentrations as function of the prepared ones and soil-specific calibration curves for Ag concentration based on the multivariate calibration models. Validation data are represented in the figures with crosses.

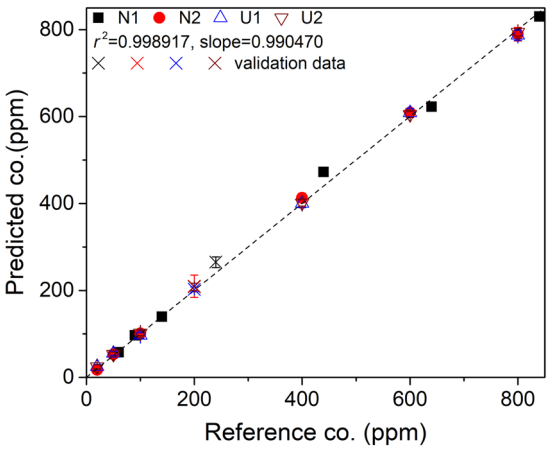

Figure 9. Model-predicted Ag concentrations as a function of the prepared ones and soil-independent calibration curve for Ag concentration based on the multivariate calibration model. Validation data are represented in the figure with crosses.

the model is validated by the spectra from specific soils, which is the cases for the validations with *N*1 (specified with an informative Ag initial concentration of 40 ppm weight from NIST) with a large *REP* and *N*2 with a large *RSD*. Nevertheless, when the model is validated by the validation spectra of all the soils, the prediction capacity exhibits an excellent level, as indicated by the values of corresponding *REP* and *RSD* in the range of 5–6%, which is order-of-magnitude improved compared to the univariate model. The degradations observed for the soil-independent model with respect to the soil-specific models seem suggesting possible improvements with a better correction of matrix effect, which might need an enlarged number of soil types used for the training of the multivariate model.

Calibration type	Soil	Calibration model					Validation	
		r^2	Slope	*REC*(%)	*RET*(%)	*LOD*(ppm)	*REP*(%)	*RSD*(%)
Soil-specific	N1	0.999997	0.9997	0.084	0.176	1.158	4.97	0.89
	N2	0.999935	1.0026	0.671	1.215	1.405	0.54	2.63
	U1	0.999923	1.0035	0.447	0.991	0.710	3.37	1.88
	U2	0.999925	1.0025	0.389	0.914	2.386	1.49	2.76
	Mean	0.999945	1.0021	0.398	0.824	1.415	2.59	2.04
Soil-independent	N1	0.998917	0.9882	3.705	5.09	4.962	22.13	3.96
	N2						0.25	12.24
	U1						3.19	2.76
	U2						0.38	3.96
	All						5.20	5.90

Table 4. Figures of merit of quantitative analysis of the mulitivariate calibration models with both the soil-specific and the soil-independent calibration curves. The soil-independent calibration model is trained using the calibration spectra from all the 4 soils. It is validated by the validation spectra of a specific soil as well as by the ensemble of validation spectra of all the 4 soils.

Conclusions

In this work, a multivariate calibration model has been developed with machine learning algorithms. The purpose is to strength the quantitative analysis ability of LIBS by efficiently correcting fluctuations due to the emission source noise and deviations due to the matrix effect. In an application case as important as soil analyses, such fluctuations and deviations prevent a univariate calibration model from being sufficient for precise and accurate quantitative analysis of the contained trace elements. The multivariate calibration model has been therefore designed for taking into account the specificities of different soils and in the same time, efficiently reducing data dispersions due to experimental fluctuations. A key point is to introduce the concept of generalized spectrum, in which the information about sample matrix is explicitly included. BPNN has been used to map a generalized spectrum to the corresponding analyte concentration. A training process, including data pretreatment, model initiation, model training loops and model validation, has been implemented within the framework of Python programing language. In the data pretreatment, a feature selection with the SelectKBest algorithm reduces the dimension of a spectrum to a value compatible with the number of the raw spectra, limiting thus the risk of overfitting, and in the same time efficiently extracts the most significant features for characterizing the spectrum.

The resulted multivariate model shows great improvements with respect to the univariate one. The fluctuation over the replicates is efficiently reduced, leading to very small error bars on the predicted concentrations. The improvement of sample-to-sample repeatability for a given soil type further allows the soil-specific calibration curves exhibiting a r^2 value exceeding 0.9999, a calibration accuracy reaching 1% level, and a *LOD* being down to the order of ppm. When being validated by independent samples, the prediction capacity of the soil-specific models presents high performance in terms of accuracy (mean *REP* = 2.59%) as well as precision (mean *RSD* = 2.04%). When the soil-independent calibration model is considered, the result of matrix effect correction is impressive with order-of-magnitude improvements with respect to the univariate model. Thereby, the accuracy and the precision of the predictions are both improved into the range of 5–6%. Our works have demonstrated the effectiveness and the advantage of applying machine learning to treat LIBS spectra of soils. The perspective to generalize the developed method to LIBS analysis of other materials, and furthermore to other spectroscopies is certainly worth to be mentioned here. Such generalization indeed allows spectroscopic techniques benefiting from the tremendous progresses realized today in machine learning, and opens wider application perspectives.

Methods

Soil samples and their preparation. Four different soils were analyzed in the experiment. Two of them were standard reference materials (SRM) from National Institute of Standards and Technology (NIST): NIST 2710 (https://www-s.nist.gov/srmors/view detail.cfm?srm = 2710a) and NIST 2587 (https://www-s.nist.gov/srmors/view detail. cfm?srm = 2587), and respectively named as N1 and N2 in this work. The other 2 soil samples were collected from 2 different places near Lyon in France, one near a river (sand-like soil) and another in an agriculture field (yellow colored soil) with unknown elemental compositions and named as U1 and U2 in this work. The 2 NIST samples were provided in fine and uniform powder of particle size $<75\,\mu$m (200 mesh). The 2 collected samples were first dried, separated from small stones and organic materials, ground and then sequentially sieved through stainless steel sieves of 100, 200 and 400 mesh, assisted by an electromagnetic vibratory shaker, finally resulting particles with sizes of $<38\,\mu$m. In each type of soil powders, silver (Ag) as analyte was added in different concentrations, by mixing the soil powders with Ag solutions at different concentrations obtained by dilution with deionized water of an Ag standard solution (2% nitric acid solution at an Ag concentration of 1000 mg/L from SPEX CertiPrep). Notice that the initial content of Ag in the collected soils was negligible (under the limit of detection of the experimental setup). For the 2 NIST samples, the N1 sample was specified with an informative initial silver concentration of 40 mg/kg (40 ppm weight). For the N2 sample, there was no specification of silver concentration. Doped powders were prepared in pellets of different soils and different Ag concentrations. For the preparation of a pellet, 0.2 g soil powder was pressed without binder under a pressure of 667 MPa (6.8 t/cm^2) for 5 min to form a pellet with a diameter of 13 mm.

Experimental setup and measurement protocol. The experimental setup used to produce the LIBS spectra has been described in detail elsewhere[33,47]. The following experimental parameters were used for the spectrum acquisition in this experiment: laser wavelength 1064 nm; laser pulse energy 60 mJ; diameter of the focused laser spot on the sample surface ~300 µm, estimated laser fluence on the sample surface 85 J/cm^2 and ablation under the atmospheric ambient. The emission from a zone around the symmetry axis of the plasma situated at a height of 1.3 mm from the sample surface was captured and coupled to an Echelle spectrometer (Andor Technology Mechelle 5000). The spectral range of the spectrometer was 220–850 nm, with a resolution power of $\lambda/\Delta\lambda \approx 5000$. The intensified CCD camera (ICCD) coupled to the spectrometer was triggered by laser pulse and set with a delay of 1 µs and a gate width of 2 µs. A gain of 60 (maximum 250) was applied of the intensifier of the ICCD for all the measurements. For each sample pellet of given Ag concentration of each type of soil, 6 replicate spectra were taken. Each spectrum was an accumulation of 200 laser shot distributed over 10 sites ablated each by 20 consequent laser pulses. Between 2 neighbor ablation sites, a translation stage displaced the pellet over a distance of 600 µm in order to avoid overlapping between the sites.

Assessment of univariate calibration model. For a given type t among the T types of soil ($T = 4$ in this experiment), an ensemble of laboratory-prepared reference samples with different analyte concentrations are separated into a calibration sample set and a validation sample set[35,40]:

n (n'): number of reference samples with different concentrations Co_{ti} (Co'_{ti}) prepared for the calibration (validation) sample set of soil type t, $1 \le i \le n$ ($1 \le i \le n'$), $n = 5$ ($n' = 2$) for the univariate model.

J: number of replicate measurements j, performed per calibration (or validation) sample, $1 \le j \le J$, $J = 6$ for the univariate model.

I_{ij}^t(Ag I 328.1 nm) ($I_{ij}^{t'}$(Ag I 328.1 nm)): experimentally recorded analyte emission intensity (called intensity for simplicity) from replicate measurement j performed on a sample with concentration Co_{ti} (Co'_{ti}).

I_i^t: mean value of experimental intensity corresponding to calibration sample Co_{ti}, $I_i^t = \frac{1}{J}\sum_{j=1}^{J} I_{ij}^t$(Ag I 328.1 nm).

I_m: mean experimental intensity for the calibration sample set of the T soil types:

$$I_m = \frac{1}{n \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} I_i^t,$$

where $S = \{1, 3, 4, 5, 7\}$ referring to the calibration sample set.

$\hat{I}_i^t$: calculated intensity with the calibration model for a calibration sample with prepared concentration Co_{ti}.

$\widehat{Co}_{tij}$ ($\widehat{Co'}_{tij}$): predicted concentration with the calibration model (reverse calibration) for the experimental replicate intensity I_{ij}^t(Ag I 328.1 nm) ($I_{ij}^{t'}$(Ag I 328.1 nm)).

$\widehat{Co}_{ti}$ ($\widehat{Co'}_{ti}$): predicted concentration with the calibration model (reverse calibration) for the experimental mean intensity I_i^t ($I_i^{t'}$).

- Determination coefficient r^2 (the square of the correlation coefficient r), a usual criterion of the performance of a calibration model:

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{9}$$

where SS_{tot} is the sum of squares of the experimental intensities corrected by their mean value, SS_{res} is the sum of squares of the residuals with respect to the calibration model:

$$SS_{tot} = \sum_{i\ in\ S} \sum_{t=1}^{T} (I_i^t - I_m)^2, \tag{10}$$

$$SS_{res} = \sum_{i\ in\ S} \sum_{t=1}^{T} (I_i^t - \hat{I}_i^t)^2, \tag{11}$$

where $S = \{1, 3, 4, 5, 7\}$ referring to the calibration sample set.

- Average relative error of calibration REC(%) for calibration accuracy evaluation:

$$REC(\%) = \frac{100}{n \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} \left| \frac{\widehat{Co}_{ti} - Co_{ti}}{Co_{ti}} \right|, \tag{12}$$

  where $S = \{1, 3, 4, 5, 7\}$ referring to the calibration sample set.
- Average relative error of prediction REP(%) for prediction accuracy evaluation:

$$REP(\%) = \frac{100}{n' \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} \left| \frac{\widehat{Co'}_{ti} - Co'_{ti}}{Co'_{ti}} \right|, \tag{13}$$

where $S = \{2, 6\}$ referring to the validation sample set.

- Relative standard deviation $RSD(\%)$ of the predicted concentrations for the validation sample set for prediction precision evaluation:

$$RSD(\%) = \frac{100}{n'} \sum_{i\ in\ S} \sqrt{\frac{1}{T \times J - 1} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(\frac{\widehat{Co'}_{tij} - Co'_{ti}}{Co'_{ti}} \right)^2}, \tag{14}$$

  where $S = \{2, 6\}$ referring to the validation sample set.
- Limit of detection LOD(ppm), deduced by fitting the experimental intensity I_{ij}^{t}(Ag I 328.1 nm) versus prepared concentrations of the calibration sample set, Co_{ti}, by a straight line,

$$I_{ij}^{t}(\text{Ag I } 328.1\text{ nm}) = \text{a} + \text{b} \times \text{Co}_{\text{ti}}, \tag{15}$$

$$LOD(\text{ppm}) = \frac{3\sigma_a}{b}, \tag{16}$$

where σ_a is the standard deviation of a, such variation is due to the dispersion of I_{ij}^{t}(Ag I 328.1 nm). LOD is thus determined by the sensibility of the technique (the slope b) and the repeatability and precision of intensity measurements among the different reference samples and different replicates for given samples (standard deviation of a, σ_a).

In the case of consideration of a specific soil type, the variable t takes the corresponding given value and the concerned sum reduces to a specific term in the above definitions.

Assessment of multivariate calibration model. In the experiment, the multivariate calibration model, in its different training stages $f^{(q)}$, allows deducing a predicted analyte concentration $\widehat{Co}_{tij}^{(q)}$ when an individual generalized spectrum, $\overrightarrow{I_{ij}^{t,General}}$, of a sample with a laboratory-prepared analyte concentration Co_{ti} (targeted concentration) is used as the input variable:

$$f^{(q)}: \mathbb{R}_{+}^{M_1+M_2} \to \mathbb{R}_{+},\ \overrightarrow{I_{ij}^{t,General}} \mapsto \widehat{Co}_{tij}^{(q)} = \left| \overrightarrow{I_{ij}^{t,General}} \right|_{General}. \tag{17}$$

The following parameters are defined to assess the performance of the multivariate model:
$T = 4$: total number of soil type;
$n = 6$: number of different concentrations in the calibration sample set;
$n' = 1$: number of different concentrations in the validation sample set;
$J = 5$: number of replicates in the calibration data set;
$J' = 1$: number of replicates in the test data set;
$J'' = 6$: number of replicates in the validation sample set;
$O = 6$: number of iterations for a given randomly and independently arranged data configuration;
$P = 10$: number of randomly and independently arranged data configurations.

- Average relative error of calibration $REC(\%)$:

$$REC(\%) = \frac{100}{n \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} \left| \frac{\widehat{Co}_{ti}^{(1)} - Co_{ti}}{Co_{ti}} \right|, \tag{18}$$

$$\widehat{Co}_{ti}^{(1)} = \frac{1}{P \times O \times J} \sum_{p=1}^{P} \sum_{o=1}^{O} \sum_{j=1}^{J} \left[\widehat{Co}_{tij}^{(1)} \right]_{(o,p)}, \tag{19}$$

where $S = \{1, 2, 3, 5, 6, 7\}$ referring to the calibration sample set, $\left[\widehat{Co}_{tij}^{(1)} \right]_{(o,p)}$ is the predicted concentration corresponding to the targeted concentration Co_{ti} by $f^{(1)}$ in a given iteration for a given randomly and independently arranged data configuration (o, p):

$$f^{(1)}: \left[\overrightarrow{I_{ij}^{t,Genral}} \right]_{(o,p)} \mapsto \left[\widehat{Co}_{tij}^{(1)} \right]_{(o,p)}. \tag{20}$$

$\widehat{Co}_{ti}^{(1)}$ is the mean predicted concentration by $f^{(1)}$ with respect to the laboratory-prepared reference concentration Co_{ti}.
- Average relative error of test $RET(\%)$:

$$RET(\%) = \frac{100}{n \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} \left| \frac{\widehat{Co}_{ti}^{(2)} - Co_{ti}}{Co_{ti}} \right|, \tag{21}$$

$$\widehat{Co}_{ti}^{(2)} = \frac{1}{P \times O \times J'} \sum_{p=1}^{P} \sum_{o=1}^{O} \sum_{j=1}^{J'} \left[\widehat{Co}_{tij}^{(2)} \right]_{(o,p)}, \tag{22}$$

where $S = \{1, 2, 3, 5, 6, 7\}$ referring to the calibration sample set, $\left[\widehat{Co}_{tij}^{(2)} \right]_{(o,p)}$ is the predicted concentration corresponding to the targeted concentration Co_{ti} by $f^{(2)}$ in a given iteration for a given randomly and independently arranged data configuration (o, p):

$$f^{(2)} : \left[\overrightarrow{I_{ij}^{t,General}} \right]_{(o,p)} \mapsto \left[\widehat{Co}_{tij}^{(2)} \right]_{(o,p)}. \tag{23}$$

$\widehat{Co}_{ti}^{(2)}$ is the mean predicted concentration by $f^{(2)}$ with respect to the laboratory-prepared reference concentration Co_{ti}.

- Average relative error of prediction $REP(\%)$:

$$REP(\%) = \frac{100}{n' \times T} \sum_{i\ in\ S} \sum_{t=1}^{T} \left| \frac{\widehat{Co}_{ti}^{(3)} - Co_{ti}}{Co_{ti}} \right|, \tag{24}$$

$$\widehat{Co}_{ti}^{(3)} = \frac{1}{J''} \sum_{j=1}^{J''} \widehat{Co}_{tij}^{(3)}, \tag{25}$$

Average relative standard deviation $RSD(\%)$ of the predicted concentrations for the validation data set:

$$RSD(\%) = \frac{100}{n'} \sum_{i\ in\ S} \sqrt{\frac{1}{T \times J'' - 1} \sum_{t=1}^{T} \sum_{j=1}^{J''} \left(\frac{\widehat{Co}_{tij}^{(3)} - Co_{ti}}{Co_{ti}} \right)^2}, \tag{26}$$

where $S = \{4\}$ referring to the validation sample set, $\widehat{Co}_{tij}^{(3)}$ is the predicted concentration corresponding to the targeted concentration Co_{ti} by $f^{(3)}$:

$$f^{(3)} : \overrightarrow{I_{ij}^{t,General}} \mapsto \widehat{Co}_{tij}^{(3)}. \tag{27}$$

- Limit of detection LOD(ppm), deduced by fitting with a straight line, the predicted concentrations $\widehat{Co}_{ijt}$ by f for the calibration sample set:

$$f : \overrightarrow{I_{ij}^{t,General}} \mapsto \widehat{Co}_{tij}, \tag{28}$$

where $i \in \{1, 2, 3, 5, 6, 7\}$ referring to the calibration sample set, versus the corresponding prepared concentrations of the calibration sample set, Co_{ti}:

$$\widehat{Co}_{ijt} = a + b \times Co_{ti}, \tag{29}$$

$$LOD(\text{ppm}) = \frac{3\sigma_a}{b}, \tag{30}$$

where σ_a is the standard deviation of a, such variation is due to the dispersion of $\widehat{Co}_{ijt}$. LOD is thus determined by the sensibility of the technique (the slope b) and the accuracy and precision of concentration prediction by the model for the different reference samples and different replicates for a given sample (standard deviation of a, σ_a).

In the case of consideration of a specific soil type, the variable t takes the corresponding given value and the concerned sum reduces to a specific term in the above definitions.

Back-propagation neuronal networks (BPNN). A single hidden layer BPNN used in this work consists of an input layer, a hidden layer, and an output layer as shown in Fig. 10. The tanh function is used as the activation function of the hidden layer. The Stochastic Gradient Descent (SGD)[43] and Mini-batch Stochastic Gradient Descent (MSGD)[48] iterations are used to construct the BPNN model. The batch size of MSGD is 0.2 (i.e., randomly chosen 20% of training samples for each epoch).

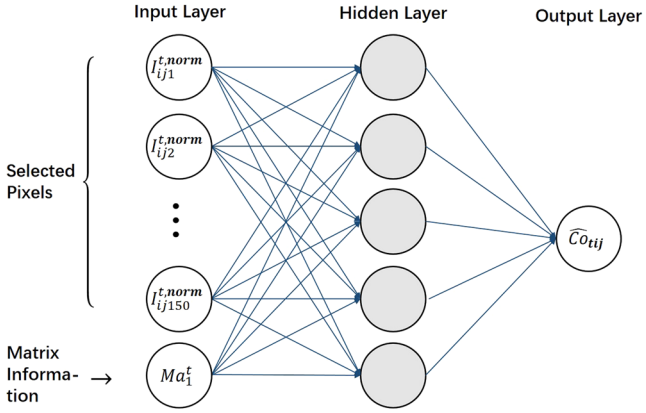

Figure 10. Structure of the used neural networks.

Comparison with the use of original spectra in the training and validation of the soil-independent model. In order to show the importance and effectiveness of the generalized spectra in the training and validation of the soil-independent model, original LIBS spectra instead of generalized spectra are used in the training and validation of the model. The obtained model presents the values of $REC = 3.92\%$ and $RET = 5.81\%$. When it is validated by the validation spectra of all the 4 soils, a $REP = 16.77\%$ is obtained. Comparing with the results shown in Table 4, we can observe a slight degradation of the calibration accuracy when the original LIBS spectra is used for model training. In contrast, an important degradation is observed for the accuracy of prediction, clearly indicating the effectiveness of the use of generalized spectra for the correction of matrix effect. In addition, we have extracted the weights applied to the outputs of the neuron in the input layer corresponding to the soil matrix information in a generalized spectrum. These outputs are respectively connected to the 5 neurons of the hidden layer. Typical values of these weights for a trained NN are 0.838435, −1.6776, −0.39414, 2.200763 and 1.503809, which is orders of magnitude larger than the mean values of the weights applied to 5 outputs of the rest of the input neurons in the same NN, with the typical values of 0.015773, −0.04353, 0.005409, −0.02946 and 0.001227. The importance of the additional dimension in a generalized LIBS spectrum related to the soil matrix is therefore clearly demonstrated together with its effectiveness for the matrix effect correction.

Software. The data processing was carried in the framework of Python version 3.6.4. Scikit-learn and NumPy were used. In addition, Origin Pro 8.0 (Origin Lab Corporation, Northampton, MA, USA) was used to design the figures. All processes were run on a PC (CPU: Intel Core i7-7700 @3.60 GHz, RAM: 8.00 GB) under Windows 10.

References

1. Mallarino, A. P. Testing of soils. *in Encyclopedia of soils in the environment*, 143–143 (Elsevier, 2005).
2. McGrath, S. P. Pollution/Industrial. *in Encyclopedia of soils in the environment*, 282–287 (Elsevier, 2005).
3. Kirkby, E. A. Essential elements. *in Encyclopedia of soils in the environment*, 478–485 (Elsevier, 2005).
4. https://en.wikipedia.org/wiki/Heavy_metals.
5. Singh, V. & Agrawal, H. M. Qualitative soil mineral analysis by EDXRF, XRD and AAS probes. *Radiat. Phys. Chem.* **81**, 1796–1803 (2012).
6. Simon, E., Vidic, A., Braun, M., Fábián, I. & Tóthmérész, B. Trace element concentrations in soils along urbanization gradients in the city of Wien, Austria. *Environ. Sci. Pollut. Res.* **20**, 917–92 (2013).
7. Falciani, R., Novaro, E., Marchesini, M. & Gucciardi, M. Multi-element analysis of soil and sediment by ICP-MS after a microwave assisted digestion method. *J. Anal. Atom. Spectrom.* **15**, 561–565 (2000).
8. Dos Anjos, M. J. *et al*. Quantitative analysis of metals in soil using X-ray fluorescence. *Spectrochim. Acat B* **55**, 1189–1194 (2000).
9. Towett, E. K., Shepherd, K. D. & Cadisch, G. Quantification of total element concentrations in soils usi ng total X-ray fluorescence spectroscopy (TXRF). *Sci. Total Environ.* **463–464**, 374–388 (2013).
10. Schmidt, M. S. *et al*. Spark-induced breakdown spectroscopy and multivariate analysis applied to the measurement of total carbon in soil. *Appl. Opt.* **51**, B176–B182 (2012).
11. Noll, R. Laser-induced breakdown spectroscopy, *Fundamental and Applications*. (Springer-Verlag Berlin Heidelberg 2012).
12. Martin, M. Z., Wullschleger, S. D., Garten, C. T. Jr. & Palumb, A. V. Laser-induced breakdown spectroscopy for the environmental determination of total carbon and nitrogen in soils. *Appl. Opt.* **42**, 2072–20776 (2003).
13. Ayyalasomayajula, K. K., Yu-Yueh, F., Singh, J. P., McIntyre, D. L. & Jain, J. Application of laser-induced breakdown spectroscopy for total carbon quantification in soil samples. *Appl. Opt.* **51**, B149–B154 (2012).
14. Nicolodelli, G. *et al*. Quantification of total carbon in soil using laser-induced breakdown spectroscopy: a method to correct interference lines. *Appl. Opt.* **53**, 2170–2176 (2014).
15. Senesi, G. S. & Senesi, N. Laser-induced breakdown spectroscopy (LIBS) to measure quantitatively soil carbon with emphasis on soil organic carbon. A review. *Anal. Chim. Acta* **938**, 7–17 (2016).
16. Díaz, D., Hahn, D. W. & Molina, A. Evaluation of Laser-Induced Breakdown Spectroscopy (LIBS) as a Measurement Technique for Evaluation of Total Elemental Concentration in Soils. *Appl. Spectrosc.* **66**, 99–106 (2012).
17. Dong, D. M., Zhao, C. J., Zheng, W. G., Zhao, X. D. & Jiao, L. Z. Spectral characterization of nitrogen in farmland soil by laser-induced breakdown spectroscopy. *Spectrosc. Lett.* **46**, 421–426 (2013).
18. Nicolodelli, G. *et al*. Double pulse laser induced breakdown spectroscopy: A potential tool for the analysis of contaminants and macro/micronutrients in organic mineral fertilizers. *Sci. Total Environ.* **565**, 1116–1123 (2016).
19. Senesi, G. S. *et al*. Heavy metal concentrations in soils as determined by laser-induced breakdown spectroscopy (LIBS), with special emphasis on chromium. *Environmen. Res.* **109**, 413–420 (2009).

20. Dell'Aglio, M. *et al*. Monitoring of Cr, Cu, Pb, V and Zn in polluted soils by laser induced breakdown spectroscopy (LIBS). *J Environ Monit.* **13**, 1422–1426 (2011).
21. Ferreira, E. C. *et al*. Evaluation of laser induced breakdown spectroscopy for multielemental determination in soils under sewage sludge application. *Talanta* **85**, 435–440 (2011).
22. Bousquet, B. *et al*. Development of a mobile system based on laser-induced breakdown spectroscopy and dedicated to *in situ* analysis of polluted soils. *Spectrochim. Acta Part B* **63**, 1085–1090 (2008).
23. Mermet, J.-M. Limit of quantitation in atomic spectrometry: An unambiguous concept? *Spectrochim. Acta Part B* **63**, 166–182 (2008).
24. Hahn, D. W. & Omenetto, N. Laser-induced breakdown spectroscopy (LIBS), part I: review of basic diagnostics and plasma–particle interactions: still-challenging issues within the analytical plasma community. *Appl. Spectrosc.* **64**, 335A–366A (2010).
25. Hahn, D. W. & Omenetto, N. Laser-induced breakdown spectroscopy (LIBS), part II: Review of Instrumental and Methodological Approaches to Material Analysis and Applications to Different Fields. *Appl. Spectrosc.* **66**, 347–419 (2012).
26. Mermet, J. M., Mauchien, P. & Lacour, J. L. Processing of shot-to-shot raw data to improve precision in laser-induced breakdown spectrometry microprobe. *Spectrochim. Acta Part B* **63**, 999–1005 (2008).
27. Colao, F. *et al*. Investigation of LIBS feasibility for *in situ* planetary exploration: An analysis on Martian rock analogues. *Planet. Space Sci.* **52**, 117–123 (2004).
28. Eppler, A. S., Cremers, D. A., Hickmott, D. D., Ferris, M. J. & Koskelo, A. C. Matrix Effects in the Detection of Pb and Ba in Soils Using Laser-Induced Breakdown Spectroscopy. *Appl. Spectrosc.* **50**, 1175–1181 (1996).
29. Lei, W. Q. *et al*. Comparative measurements of mineral elements in milk powders with ICP and LIBS: assessment and validation of the CF-LIBS procedure. *Anal. Bioanal. Chem.* **400**, 3303–3313 (2011).
30. Gilon, N. *et al*. A matrix effect and accuracy evaluation for the determination of elements in milk powder LIBS and laser ablation/ICP-OES spectrometry. *Anal. Bioanal. Chem.* **401**, 2681–2689 (2011).
31. https://web.archive.org/web/20041205053402/http://clic.cses.vt.edu/icomanth/02-AS_Classification.pdf.
32. Lepore, K. H. *et al*. Matrix Effects in Quantitative Analysis of Laser-Induced Breakdown Spectroscopy (LIBS) of Rock Powders Doped with Cr, Mn, Ni, Zn, and Co. *Appl. Spectrosc.* **71**, 600–626 (2017).
33. Tian, Y. *et al*. Elemental analysis in powders with surface-assisted thin film laser-induced breakdown spectroscopy. *Spectrochim. Acta B* **214**, 16–24 (2016).
34. Martin, M. Z. *et al*. Novel Multivariate Analysis for Soil Carbon Measurements Using Laser-Induced Breakdown Spectroscopy. *Soil Sci. Soc. Am. J.* **74**, 87–93 (2010).
35. Sirven, J.-B. *et al*. Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis. *Anal. Bioanal. Chem.* **385**, 256–262 (2006).
36. Mukhono, P. M., Angeyo, K. H., Dehayem-Kamadjeu, A. & Kaduki, K. A. Laser induced breakdown spectroscopy and characterization of environmental matrices utilizing multivariate chemometrics. *Spectrochim. Acta B* **87**, 81–85 (2013).
37. El Haddad, J. *et al*. Artificial neural network for on-site quantitative analysis of soils using laser induced breakdown spectroscopy. *Spectrochim. Acta B* **79-80**, 51–57 (2013).
38. Yu, K. Q., Zhao, Y. R., Liu, F. & He, Y. Laser-Induced Breakdown Spectroscopy Coupled with Multivariate Chemometrics for Variety Discrimination of Soil. *Sci. Rep.* **6**, 27574 (2016).
39. Yongcheng, J., Wen, S., Baohua, Z. & Dong, L. Quantitative Analysis of Magnesium in Soil by Laser-Induced Breakdown Spectroscopy Coupled with Nonlinear Multivariate Calibration. *J. Appl. Spectrosc.* **84**, 731–737 (2017).
40. Mermet, J.-M. Calibration in atomic spectrometry: A tutorial review dealing with quality criteria, weighting procedures and possible curvatures. *Spectrochim. Acta B* **65**, 509–523 (2010).
41. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill (2001).
42. https://physics.nist.gov/PhysRefData/ASD/lines_form.html.
43. Hinton, G. *et al*. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Proc. Mag.* **29**, 82–97 (2012).
44. Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Math. Control Signals Systems* **2**, 303–314 (1989).
45. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural networks* **2**, 359–366 (1989).
46. Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks* **2**, 183–192 (1989).
47. Xiu, J., Motto-Ros, V., Panczer, G., Zheng, R. & Yu, J. Feasibility of wear metal analysis in oils with ppm and sub-ppm sensitivity using laser-induced breakdown spectroscopy of thin oil layer on metallic target. *Spectrochim. Acta B* **91**, 24–30 (2014).
48. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. 30th International Conference on Machine Learning, ICML 2013 (PART 3), 2176–2184 (2013).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant Nos 11574209, 11805126, and 11601327), the Science and Technology Commission of Shanghai Municipality (Grant No. 15142201000) and China Postdoctoral Science Foundation (Grant No. 2018M641992).

Author Contributions

C.S. wrote the program and performed the calculations for the multivariate model and participated in the paper writing, Y.T. performed the sample preparation, the experimental measurement and the calculations for the univariate model and participated in the paper writing, J.Y. conceived the experiment and the data treatment strategy, interpreted the results and wrote the paper. N.D.G. participated in the conception of the experimental measurement and the sample preparation, L.G., Y.Z., Z.Y. participated in the data treatment, Y.N., T.Z., H.L. participated in multivariate model construction, Y.N. participated in the paper writing. All the authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.