Application of a series of artificial neural networks to on-site quantitative analysis of lead into real soil samples by laser induced breakdown spectroscopy

Josette El Haddad, Delphine Bruyère, Amina Ismaël, G. Gallou, Valérie Laperche, Karine Michel, Lionel Canioni, Bruno Bousquet

Application of a series of artificial neural networks to on-site quantitative analysis of lead into real soil samples by laser induced breakdown spectroscopy

J. El Haddad [a], D. Bruyère [b], A. Ismaël [c], G. Gallou [c], V. Laperche [b], K. Michel [b], L. Canioni [a], B. Bousquet [a,*]

[a] Univ. Bordeaux, LOMA, CNRS UMR 5798, F-33400 Talence, France
[b] BRGM, Service Métrologie, Monitoring et Analyse, 3 av. C. Guillemin, B.P 36009, 45060 Orléans Cedex, France
[c] IVEA Solution, Centre Scientifique d'Orsay, Bât 503, 91400 Orsay, France

ARTICLE INFO

Keywords:
LIBS
Artificial neural network
Matrix
Lead
Soil

ABSTRACT

Artificial neural networks were applied to process data from on-site LIBS analysis of soil samples. A first artificial neural network allowed retrieving the relative amounts of silicate, calcareous and ores matrices into soils. As a consequence, each soil sample was correctly located inside the ternary diagram characterized by these three matrices, as verified by ICP-AES. Then a series of artificial neural networks were applied to quantify lead into soil samples. More precisely, two models were designed for classification purpose according to both the type of matrix and the range of lead concentrations. Then, three quantitative models were locally applied to three data subsets. This complete approach allowed reaching a relative error of prediction close to 20%, considered as satisfying in the case of on-site analysis.

1. Introduction

Laser-induced breakdown spectroscopy is recognized to have high potential for geochemical applications since this technique is able to achieve rapid and multi-elemental on-site analysis with very little sample preparation [1–5]. In the framework of a collaborative project, our objective was to quantify heavy metals in soil samples by LIBS analyses. In this paper, a special focus on the analysis of lead is presented. Generally speaking, when considering a series of samples related to a single matrix, the common normalization by an internal standard has been applied to the LIBS data [6–8]. Unfortunately, in the specific case of soil samples, the LIBS signal is known to be highly dependent of the matrix [9] and consequently different matrices must be taken into account. Thus, the basic univariate approach, which consists in building the so-called calibration curve [10] becomes inappropriate in this case, even after several attempts of normalization, and advanced data treatment is required.

Soils are natural samples that are not easy to simply describe. However, it was considered in this work that the two main matrices of soils are i) the silicate matrix ($SiO_2 + Al_2O_3$) and ii) the calcareous matrix ($CaO + MgO$). Matrix effects have already been reported in the case of LIBS analyses of heavy metals in soil samples [9,11] but no major element was encountered under constant concentration, preventing the application of normalization by internal standard [12]. To overcome this problem when dealing with quantitative analysis, two opposite strategies were proposed: i) the Calibration-Free method [13], not discussed in this paper, and ii) the use of multivariate approach known as chemometrics [14,15]. Multivariate analyses have already been successfully applied to the treatment of LIBS data from soil samples [16]. More precisely, several multivariate methods such as PCA, SIMCA, LDA and PLS-DA have been applied to classify soil or geomaterial samples [16–21]. Regarding quantitative LIBS analysis, the most common technique of chemometrics is the partial least square (PLS) regression [22,23]. This method has been exploited for soil analysis by Essington et al. [24] who discussed the difficulty to achieve quantitative analysis with acceptable relative error of prediction. Moreover, PLS has been used to quantify both major and trace elements from the LIBS signals provided by the ChemCam instrument on Mars [25,26]. In this latter work, in order to obtain better prediction ability, the authors suggested to eliminate outliers. Moreover, they used independent component analysis (ICA) to efficiently identify the elements present in the samples. They finally demonstrated that despite the complexity of the samples, univariate analysis provided better results than PLS for trace elements. Then, in order to take into account potential nonlinearities contained into the LIBS spectra, the method of artificial neural networks – hereafter called ANN – has been efficiently applied [16,27,28]. More precisely, in a previous work of our group, we have used ANN to predict the concentrations of major elements such as calcium, aluminum and iron and also those of trace elements as copper. In this

* Corresponding author. Tel.: +33 540002870; fax: +33 540006970.
E-mail address: bruno.bousquet@u-bordeaux.fr (B. Bousquet).

past work, we highlighted the importance of taking into account spectral lines from the matrix in addition to those of the analyte as input data of the ANN in order to improve the prediction ability of the model [28].

In the present work, we focus on the analysis of lead contained in soil samples from three different geological sites. In this case, the lead concentrations ranging between 250 and 147 000 ppm induced some difficulties for a direct treatment. As a consequence, we demonstrate in this paper that applying a series of ANN models for both classification and quantification purposes allowed to obtain satisfying results.

2. Experimental

The experimental setup, sample preparation and LIBS measurements have been already described in [28] so only a brief description is given in this section. The LIBS system dedicated to the on-site LIBS measurements of soils was the MobiLIBS III from IVEA SAS, including a Nd:YAG laser at 266 nm–20 Hz–5 ns, a focusing system providing 43 GW/cm^2 and an Echelle spectrometer coupled to an ICCD camera. The light emitted by the laser-induced plasma was collected with a patented achromatic telescope and injected in a 3-meter fused silica optical fiber of 550 μm diameter. The software AnaLIBS (IVEA SAS) was used to control the experimental parameters. The full system was integrated in a van, as a mobile laboratory, in order to allow on-site LIBS measurements.

Three geological sites located in France were analyzed. The first one – hereafter called SLM – was characterized by high concentrations of lead, zinc, barium and calcium. The two other sites – hereafter called ME and SEB – were characterized by the silicate matrix with much lower concentrations in ores and in calcium than the ones measured on the first site. Smart sampling of the sites was achieved in order to sample the most relevant soils from the ground, avoiding redundancy and taking advantage of the whole range of concentrations that one can observe on each site. This smart sampling was achieved by the use of a portable XRF device (Niton XL3t800, Thermo Scientific). Each soil sample extracted from the ground was sifted at 2 mm maximum grain size and split into two parts, one dedicated to direct LIBS analysis and the other one to later laboratory ICP-AES analysis in order to provide the reference values of concentration. It should be pointed out that in the case of environmental monitoring, sampling is of major importance and could strongly affect the analytical results. However, thorough considerations about sampling are out of the scope of the present paper and consequently, the analytical performances given hereafter may be criticized. The best example to illustrate this point is that, for a given soil sample, two separate amounts of matter were prepared, one for the LIBS analysis and the other for the ICP-AES analysis. They were assumed to be two perfect replicates but this point was not fully assessed. It should be emphasized that the values obtained after ICP-AES analysis were considered as reference values and consequently they had to be measured by reliable and robust method. The ICP-AES measurements were performed at BRGM and based on the international standardization ISO 14869-2:2002. Briefly, it consists in grinding the soil powder at 80 μm, then making the soil fusion by sodium peroxide in an oven at 450 °C and then achieving a dissolution with hydrochloric acid prior to the ICP-AES measurement. During this lab analysis, 10% of the samples were duplicated for the mineralization step in order to evaluate the analysis.

For LIBS analysis, the soil samples were finally dried with the use of a microwave oven since it has been reported that the higher the moisture level the lower the LIBS signal [5]. Finally the dried soils were prepared as pressed pellets of 13 mm diameter by applying 8 tons/cm^2 during 2 min with a manual press. To optimize to signal-to-noise ratio, it was decided that each LIBS spectrum would be the result of 25 laser shots accumulated at the same point of the sample, with a gate delay of 300 ns and a gate width of 3 μs. And to reduce the effects of heterogeneity, 25 spectra were acquired for each sample. One single average spectrum was calculated for each sample and used for quantitative analysis. Indeed, side experiments allowed verifying that averaging over 25 locations of the laser spot at the sample surface was sufficient to correctly take into account the sample's heterogeneity.

Finally, statistics were calculated by running five times each ANN model. Each calculation starting with different initial random values of weights, the same input LIBS data generated five slightly different output values. Thus, the results of the ANN calculations are given by the average value and the RSD value over five repetitions.

3. Results

3.1. Description of the samples within a ternary diagram

Soil is considered to be amongst the most complex samples and consequently the most difficult to analyze by LIBS due to the high diversity of matrices. Thus, prior to quantitative analysis, it is highly recommended to have a strong understanding of the matrix related to the sample under study. Indeed, this could allow selecting the more efficient model of calibration.

As a first observation, let us have a look at the data provided by ICP-AES. From the values of concentrations, three values were calculated in order to highlight the type of matrix, namely to determine if the sample should be related to a silicate, calcareous or ore matrix. The three values calculated from ICP-AES data are given hereafter.

For the silicate matrix:

$$V1 = ([Si] + [Al])/([Si] + [Al] + [Ca] + [Mg] + [Ba] + [Zn] + [Pb]) \quad (1)$$

For the calcareous matrix:

$$V2 = ([Ca] + [Mg])/([Si] + [Al] + [Ca] + [Mg] + [Ba] + [Zn] + [Pb]) \quad (2)$$

For the ore matrix:

$$V3 = ([Ba] + [Zn] + [Pb])/([Si] + [Al] + [Ca] + [Mg] + [Ba] + [Zn] + [Pb]) \quad (3)$$

It should be emphasized that Si and Al are frequently strongly correlated in geological samples (more precisely SiO_2 and Al_2O_3), and thus, they were considered together to calculate the value V1 characterizing the silicate matrix. In the same way, Ca and Mg both contribute to the calcareous matrix (more precisely CaO and MgO) and consequently both were taken into account in the calculation of the value V2 characterizing the calcareous matrix. Finally, Ba, Zn and Pb were selected to represent the ore matrix (value V3) regarding the range of concentrations provided by the ICP-AES analysis. High values of concentration for Pb should be related to the natural ore of Galena (PbS) while the presence of Zn could be related to two types of natural ores, namely sphalerite (ZnS) and calamine (ZnCO). Finally, high concentrations in Ba should be associated to a soil rich in barite ($BaSO_4$).

Eqs. (1) to (3) illustrate that the values V1, V2 and V3 were normalized to 1 so that each of them expressed a percentage. Consequently, any soil sample could be described through these three values. As an example, a soil sample characterized by the values V1 = 0.8, V2 = 0.15 and V3 = 0.05 should be considered as a soil with a matrix 80% silicate, 15% calcareous and 5% ores.

Soil samples from three geological sites and thus potentially three different matrices were studied: 27 samples from the site SEB, 30 from the site ME, and 60 from the site SLM. Finally, Fig. 1 displays these 117 soil samples inside a ternary diagram based on the calculation of the three values V1, V2 and V3, namely based on the three types of matrices: silicate, calcareous and ore. It should be emphasized that the values reported in Fig. 1 were retrieved from the values of concentrations provided by ICP-AES.

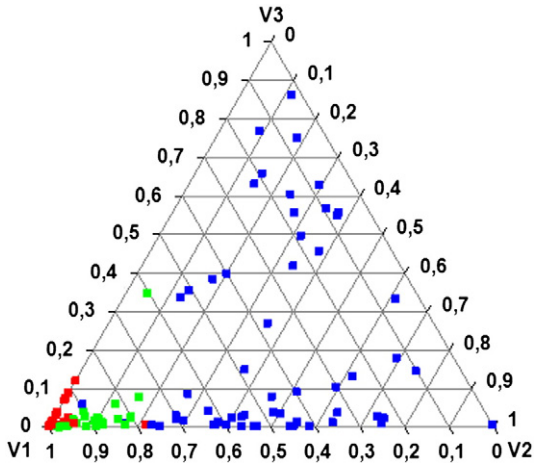

Fig. 1. Ternary diagram based on the values of concentrations measured by ICP-AES presenting the soil samples of three different sites (blue: SLM, red: SEB and green: ME). The location of each point is given by the calculation of the values V1, V2 and V3 given in the text.

On can observe in Fig. 1 that all the samples from the SEB and ME sites are displayed in a small area located very close to the pole silicate, related to V1 = 1. At the opposite, the samples from the SLM site are spread over the entire diagram, indicating that there is not a single matrix that characterizes this geological site. It should be pointed out that the SLM site was the place of mining activities many years ago, and this is why the high concentrations in ores (V3) are not surprising. Moreover, all the values reported here were obtained from natural soils, without any doping.

Once the values of concentrations provided by ICP-AES were reported into the ternary diagram, it was interesting to study if such a diagram could be drawn also from the LIBS data in order to check the ability of LIBS to provide relevant information about the matrix directly on site. Thus, a 3-layer artificial neural network (ANN) was fed with few selected LIBS data. The output layer contained 3 neurons. A learning step consisted in training the ANN model to retrieve the three values V1, V2 and V3 provided by the calculation of the ICP-AES data. Once the training step is completed, the ANN was ready to calculate the three output values for any unknown soil sample.

Regarding the most relevant elements inside the soil samples under study, it was decided to select 35 values as input data for the ANN model. These 35 data extracted from the LIBS spectra were the intensity values of the 35 spectral lines reported in Table 1. Of course, among the 35 selected lines, some of them were related to the two elements Al and Si to describe the silicate matrix, some others were related to Ca and Mg to represent the calcareous matrix and the last ones were related to Zn, Pb and Ba to describe the ore matrix. The 117 soil samples were split into three subsets in order to properly build, validate and test the ANN model: 76 samples into the calibration set, 21 into the validation set and 20 into the test set. It should be mentioned that, taking into account the number of input data and the architecture of the 3-layer artificial neural network, the number of weights to be retrieved was very large.

Table 1
Wavelength (nm) of the 35 spectral lines selected for the LIBS analyses.

Element	Wavelength (nm)
Si	250.689; 251.611; 251.920; 252.41; 252.85; 288.157
Zn	307.589; 319.631; 330.258; 334.501; 472.215; 481.053; 636.234
Ca	612.221; 442.544; 558.875; 610.272; 616.217; 643.907; 646.256
Al	309.215; 309.271; 394.407; 396.152
Pb	261.417; 283.305; 363.956; 368.346; 405.780
Ba	652.735; 659.532; 669.384; 705.994; 728.029
Mg	285.212

As a consequence, a large number of samples should have been analyzed by ANN in order to avoid any risk of overfitting. However, the number of samples analyzed during the on-site LIBS campaigns reported here was reduced and thus the risk of overfitting was always present. To address this point, all the results obtained by ANN were evaluated via a series of five repetitions of the ANN calculation starting from different random weight values, and also via the Y-randomization test.

Based on the method of external validation and using data from both the calibration and the validation sets, the optimized parameters for the ANN were found to be: number of neurons into the hidden layer: 4, learning rate: 0.01, momentum: 0.1, and number of iterations: 19 000. The three output values calculated by the ANN model ranged between 0 and 1 and thus could be directly interpreted as percentage values without any additional treatment. Fig. 2 displays the predicted versus reference values of the quantities V1 (a), V2 (b) and V3 (c) defined by Eqs. (1)–(3). The points being displayed very close to the ideal line

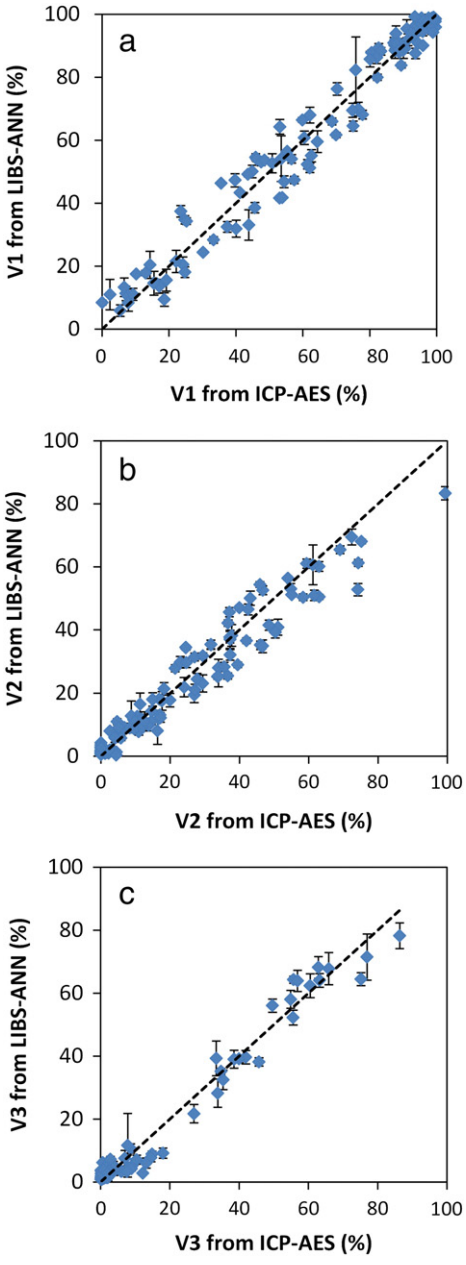

Fig. 2. Correlation plots of the factors V1 (a), V2 (b) and V3 (c) defined in the text displaying the values (%) predicted by LIBS-ANN against the reference values (%) obtained from ICP-AES. Errors bars correspond to the standard deviation of five repetitions of the ANN calculation. Dashed lines correspond to the equations y = x.

given by the equation y = x, one can conclude that LIBS data based on on-site measurement and processed by ANN provide very good correlation for each value V1, V2 and V3. As a consequence, one can expect from the LIBS-ANN treatment a very similar distribution of points inside the ternary diagram as the one resulting from ICP-AES analysis. This very interesting result demonstrates the ability of on-site LIBS analysis to correctly describe a series of samples according to their matrix. To go further, we decided to assess this process of locating samples into the ternary diagram by comparing the values resulting from ICP-AES to the ones resulting from ANN applied to LIBS data, as displayed in Fig. 2. Thus, for samples of the calibration, validation and test sets, we separately calculated the correlation coefficient R^2 in order to check the correlation between the LIBS and ICP-AES values for each pole of the ternary diagram. In addition, we calculated the root mean square error (RMSE) for each set of data and for each pole, i.e. for each value V1, V2 and V3. Table 2 displays the values of R^2 and RMSE for the calibration (C), validation (V) and test (T) sets. One single ANN model has been run, providing three output values, one for the silicate pole (V1), one for the calcareous pole (V2) and one for the ores pole (V3). It should be noticed that RMSE is given in percent in Table 2 because the values of concentrations are also given in percent.

The value of R^2 was always found to be higher than 0.94, which reveals a good correlation between the values calculated by the ANN and based on the LIBS data and the values provided by ICP-AES and considered as reference values. Moreover, the values of RMSE were found to be quite small, indicating that the values given by ANN after LIBS analysis were very close to the ones provided by ICP-AES. Indeed, the value of RMSE was always lower than 5.8%. Such performance was very satisfying. It reveals that, after a learning step, the 3-layer ANN model described here and providing three output values is perfectly adapted to process LIBS data in order to describe the soil samples according to their types of matrix. This result is consequently very important, since it demonstrates that a single ANN model applied to LIBS data could allow predicting the type of matrix for any unknown soil sample, directly on-site.

3.2. Quantitative analysis of lead

In this section, lead was chosen to demonstrate the ability of ANN to process LIBS data in the case of concentration values highly varying from one sample to another. Thus, a three-layer ANN with only one neuron in the output layer was applied. The output value was the lead concentration.

It was decided to introduce 10 input values into the ANN: 5 of them were related to LIBS spectral lines of lead (261.418; 283.305; 363.957; 368.346; 405.781 nm) and the other 5 were related to spectral lines of major elements expected to represent the matrix, namely Ca (612.222 nm), Fe (382.043 nm), Ti (399.864 nm), Ba (652.731 nm) and Al (309.271 nm). Actually, we demonstrated in a previous study the importance of introducing into the ANN data not only data directly linked to the analyte but also additional data presenting no link with the analyte but offering a good representation of the matrix, in order to correctly quantify the analyte [28]. The 117 soil samples provided by the 3 campaigns of on-site LIBS measurements were split in the same way as for the previous study, namely 76 samples into the calibration set, 21 into the validation set and 20 into the test set. It should be pointed out that the concentration of lead ranged from 250 ppm to 147 000 ppm, according to the ICP-AES values.

The ANN model was first built from data included in the calibration set. Then it was applied to the data contained in the validation set in order to check overfitting and to optimize the parameters of the ANN model. This external validation allowed converging to the best model, namely the model giving the lowest RMSE values. Finally, the 20 samples of the test set were exploited afterward in order to post-evaluate the model. It should be emphasized that the three datasets were not built randomly but rather in a way to include the larger range of concentrations of lead in each of them, for good learning, good validation, and good test. Based on the external validation procedure, the intrinsic parameters of the ANN were optimized, namely, the number of nodes in the hidden layer, the learning rate, the momentum and the number of iterations during the iterative algorithm. After optimization of those parameters, we obtained the results displayed in Table 3.

The relatively good correlation between the predicted and reference values of concentration of lead was evidenced by values of R^2 between 0.90 and 0.95 for the three datasets. Moreover, despite relatively high values of Q^2, namely between 0.89 and 0.94, the relative errors (RE) were found to be higher than 100%, which demonstrated that this ANN model was not able to achieve good prediction for the concentrations of lead in the case of the soil samples under study. Moreover, the RMSE value was found to be as high as 9800 ppm for the calibration set and looking closer at the reference values of concentrations provided by ICP-AES, the very large range of concentrations of lead from 250 ppm to 147 000 ppm was considered to be the probable origin of the poor ability of the ANN model to achieve quantitative analysis in this case.

To go further with the quantitative analysis of lead, it was decided to split the original data set into two subsets, the first one containing samples with lead concentrations above 10 000 ppm and the second one containing samples with lead concentrations below this value. Thus two new ANN models were built, one for each subset. The model applied to the lead concentrations higher than 10 000 ppm and called hereafter ANN1 was built from 23 samples (calibration set 1) and optimized by using 6 samples (validation set 1). It was finally tested a posteriori with 6 samples (test set 1). In the same way, ANN2 was the model optimized for the analysis of lead concentrations below 10 000 ppm. It was built from 53 samples (calibration set 2) and optimized with 15 samples (validation set 2). Finally it was tested a posteriori with 14 samples (test set 2).

Once the samples were split into two groups, namely above and below the threshold value of 10 000 ppm for the lead concentration, the two ANN models were separately optimized and the results

Table 2
Performance of the ANN model optimized to provide the relative values V1, V2 and V3 described in the text and related to the three poles of the ternary diagrams (cf. Figs. 1 and 2). C, V and T stand for calibration, validation and test set, respectively.

		Silicate	Calcareous	Ores
R^2	C	0.97	0.94	0.98
	V	0.96	0.96	0.97
	T	0.98	0.96	0.95
RMSE (%)	C	5.6	5.8	3.2
	V	5.7	4.5	3.2
	T	4.8	5.5	3.5

Table 3
Performance of the ANN model applied to the whole set of LIBS data in the case of quantitative analysis of soil samples.

Number of iterations		5000
Learning rate		0.1
Momentum		0.1
Number of neurons in the hidden layer		5
Calibration set	R^2	0.90
(76 samples)	Q^2	0.89
	RE (%)	158
	RMSE (ppm)	9800
Validation set	R^2	0.95
(21 samples)	Q^2	0.94
	RE (%)	151
	RMSE (ppm)	6700
Test set	R^2	0.92
(20 samples)	Q^2	0.89
	RE (%)	88
	RMSE (ppm)	6700

obtained after optimizing are presented in Table 4. It should be noticed that the number of samples involved in the model ANN1 was equal to 35 while this number was equal to 82 for the model ANN2. Regarding R^2 and Q^2, the model ANN1 displayed higher correlation and predictive ability than the model ANN2. Moreover, ANN1 provided a maximum value of relative error of about 21% while this value was about 32% for ANN2. As a consequence, splitting the whole dataset into two subsets and applying two separate ANN models allowed retrieving results much more satisfying than the ones obtained with a single ANN model (cf. Table 3). However, the value of the mean relative error obtained by the model ANN2 was still too high regarding the expectation of on-site measurement. There was consequently a pending challenge to better quantify lead concentrations in the case of values below 10 000 ppm.

Comparing the values of lead concentration provided by ICP-AES (not reported here) and the ternary diagram presented in Fig. 1, it was easy to conclude that all the soil samples with lead concentrations below 10 000 ppm were also characterized by a percentage value of ore matrix below 10%. Moreover, the points corresponding to this condition in Fig. 1 are spread over the entire range of concentrations along the silicate-calcareous axis. Thus, taking into account the density of points in Fig. 1, it was finally decided to set a threshold value at 80%-silicate and thus to separate the dataset into two subsets in order to further analyze each subset with a separate ANN model. Consequently, the ANN model previously called ANN2 was cast-off and two new ANN models were exploited instead. Thus, the ANN model hereafter called ANN3 was built for samples characterized by percentage values of silicate matrix below 80%. In this case, 21 samples were included into the calibration set, 7 into the validation set and 6 into the test set. In the same way, a second ANN model hereafter called ANN4 was built for samples characterized by percentage values of silicate matrix above 80%. In this latter case, 32 samples composed the calibration set, 8 composed the validation set and 8 composed the test set. It should be recalled that all these samples, in both cases, had lead concentrations below 10 000 ppm and also percentage values of ore matrix below 10%.

After optimization, the results obtained by the two ANN models (ANN3 and ANN4) are displayed in Table 5. Based on these two ANN models, the mean relative error was decreased to 24% or less instead of the 32% obtained with the model ANN2 and reported in Table 4. The values of RMSE reported in Table 5 also confirm the advantage of splitting the dataset (<10 000 ppm lead) into two subsets, namely <80% silicate and >80% silicate.

In conclusion, to quantify the lead concentrations of any soil sample from one of the three geological sites under study, namely SLM, ME and SEB, one should select the ANN model that takes into account both the range of lead concentrations and the type of matrix. In our study, this can be summarized as follows:

- ANN1 should be selected in the case of lead concentrations higher than 10 000 ppm
- ANN3 should be selected in the case of lead concentration lower than 10 000 ppm and a matrix type such that the percentage of silicate is lower than 80%
- ANN4 should be selected in the case of lead concentration lower than 10 000 ppm and a matrix type such that the percentage of silicate is higher than 80%

To confirm this conclusion, we decided to verify that the performance achieved by the three ANN models (ANN1, ANN3 and ANN4) was not obtained by abnormal chance. Consequently, we applied to each model the well-known method of Y-randomization [29]. This method consists in keeping unchanged the input data and to calculate a new model after random reorder of the reference values exploited during the learning step. In this work, this process of random reorder was repeated 25 times and finally, the average values of the parameters introduced early were calculated. These values are displayed in Table 6. As expected, the consequence of the Y-randomization procedure was to drastically decrease the predicting ability of the ANN models. The negative impact of Y-randomization on the ANN models was thus clearly

Table 4
Performance of the two quantitative ANN models applied to the LIBS data in the case of quantitative analysis of soil samples; ANN1 for lead concentrations above 10 000 ppm and ANN2 for lead concentrations below this value.

Model		ANN1	ANN2
Range of lead concentrations		>10 000 ppm	<10 000 ppm
Number of iterations		9000	8000
Learning rate		0.1	0.15
Momentum		0.1	0.1
Number of neurons in the hidden layer		5	5
Calibration sets	R^2	0.96	0.95
	Q^2	0.93	0.94
	RE (%)	21	32
	RMSE (ppm)	9100	600
Validation sets	R^2	0.96	0.79
	Q^2	0.93	0.78
	RE (%)	12	32
	RMSE (ppm)	6400	1100
Test sets	R^2	0.94	0.85
	Q^2	0.87	0.84
	RE (%)	15	27
	RMSE (ppm)	6700	900

Table 5
Performance of the two quantitative ANN models applied to the LIBS data in the case of quantitative analysis of soil samples and for lead concentrations above 10 000 ppm. The model ANN3 was optimized for analyzing samples characterized by a matrix <80% silicate and the model ANN4 was optimized for samples with >80% silicate.

Model		ANN3	ANN4
Range of lead concentrations		<10 000 ppm	<10 000 ppm
Matrix type		<80% silicate	>80% silicate
Number of iterations		17000	7000
Learning rate		0.15	0.1
Momentum		0.1	0.2
Number of neurons in the hidden layer		5	7
Calibration set	R^2	0.99	0.93
	Q^2	0.99	0.93
	RE (%)	24	21
	RMSE (ppm)	200	700
Validation set	R^2	0.89	0.95
	Q^2	0.87	0.94
	RE (%)	21	19
	RMSE (ppm)	900	500
Test set	R^2	0.91	0.91
	Q^2	0.83	0.91
	RE (%)	18	19
	RMSE (ppm)	800	700

Table 6
Performance of the three ANN models calculated after Y-randomization. The values reported in the table are average values calculated after 25 repetitions (see text for details). The letters c, v, and t stand for calibration, validation and test sets, respectively.

Model	ANN1	ANN3	ANN4
Range of lead concentrations	>10 000 ppm	<10 000 ppm	<10 000 ppm
Matrix type		<80% silicate	>80% silicate
R^2c	0.73	0.73	0.93
R^2v	0.00	0.02	0.04
R^2t	0.01	0.00	0.00
Q^2c	0.73	0.73	0.92
Q^2v	−5.36	−6.75	−7.70
Q^2t	−11.97	−31.65	−8.48
REc (%)	29	73	32
REv (%)	96	208	186
REt (%)	115	229	256
RMSEc (ppm)	13 100	1000	600
RMSEv (ppm)	41 000	2800	3600
RMSEt (ppm)	44 300	3200	4000

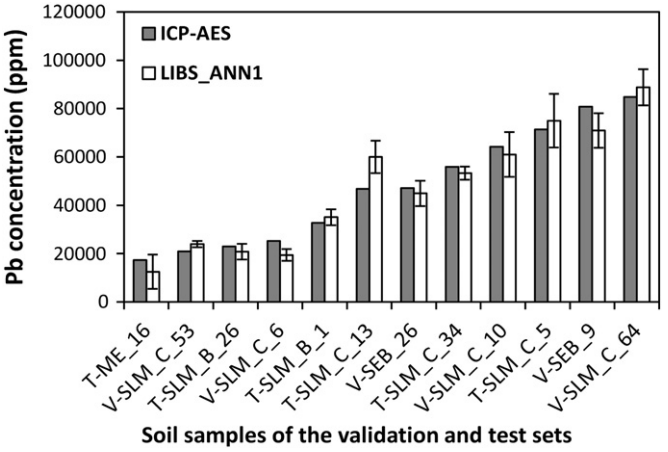

Fig. 3. Comparison chart of the lead concentrations (ppm) of soil samples measured by ICP-AES (gray) and those obtained by applying the model ANN1 (white) to the LIBS data, for both the validation and the test sets. Errors bars correspond to the standard deviation of five repetitions of the ANN calculation.

revealed by values of R^2 and Q^2 much lower than 1 as well as values of relative error and RMSE much higher than the ones obtained in the case of the optimized models (cf. Tables 4 and 5).

To conclude, the results presented in Table 6 demonstrate that the performance of the ANN models displayed in Tables 4 and 5 was not obtained by abnormal chance but really by optimizing the models. The advantage of using three ANN models instead of one was clearly demonstrated. However, it should be pointed out that the parameters reported in Tables 4 and 5 are average values. To go further, a more detailed view of the prediction ability of the ANN models is given below for each model. Fig. 3 displays the chart obtained for the model ANN1 (lead concentrations higher than 10 000 ppm), and for the samples of the validation and test sets. The error bars on the LIBS-ANN data represent the standard deviation of 5 repetitions of the ANN calculation starting with different random values of weights. One can conclude that the prediction ability of ANN1 was quite satisfying for most of the soil samples. However, for the sample T-SLM_C_13, the predicted value was slightly overestimated and this might reveal an incomplete description of the related matrix.

Fig. 4 displays the results obtained by LIBS_ANN when the model ANN3 was applied and those obtained by ICP-AES and considered as reference values. In this case, the lead concentration was lower than 10 000 ppm and the percentage of silicate was lower than 80%. One can observe that most of the concentrations were correctly predicted. However, the model ANN3 clearly under estimated the values of concentration higher than 5000 ppm. This may be due to the very small number of samples, i.e. 2, inside the calibration set in the range 5000–10 000 ppm. More generally, one can expect increasing the performances of the ANN models by increasing the number of samples into the calibration set, since ANN is a supervised method that needs a learning step based on a wide variety of samples.

Finally, Fig. 5 displays the results obtained by LIBS_ANN when the model ANN4 was applied and those obtained by ICP-AES and considered as reference values. For this model, the lead concentrations were always lower than 10 000 ppm and the percentage of silicate was higher than 80%. One can observe that most of the concentrations were correctly predicted except for the sample T-SEB_12, which was underestimated by ANN. Nevertheless, it is interesting to notice that this sample was analyzed a second time by ICP-AES and the result was different from the one obtained during the first analysis. It would be worth considering in more details the heterogeneity of this sample in order to reach a better understanding of the residual discrepancy between the LIBS and ICP-AES results.

The results presented here were obtained on the basis of 10 input data including 5 spectral lines of lead and 5 spectral lines related to matrix elements. If only the 5 spectral lines related to lead had been injected into the ANN models, the performances would have been clearly decreased, especially for the low-concentration range. This highlights the importance of introducing not only spectral lines from lead but also spectral lines from other significant matrix elements. This result was not exclusive to the quantitative analysis of lead but rather general as it was already discussed in [28]. It clearly reveals that most of the matrix effects can be taken into account by introducing data from the matrix as input data of the ANN.

In conclusion, three ANN models have been applied, depending on the values of lead concentration on one hand and on the type of matrix on the other hand. But the last step consisted in choosing the right ANN model among the three when analyzing any unknown soil sample. This choice could be simply based on the position of the sample inside the ternary diagram, which can be retrieved from the correlation plots displayed in Fig. 2. However, since the models ANN1 on one hand and ANN3 and ANN4 on the other hand were based on the values of lead concentrations, above and below the threshold of 10 000 ppm respectively, we decided to investigate also the ability of a new ANN model to achieve the classification of a series of soil samples according to a given threshold value of concentration.

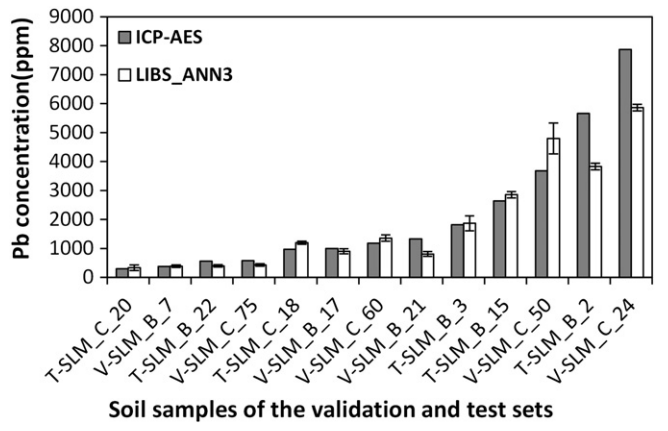

Fig. 4. Comparison chart of the lead concentrations (ppm) of soil samples measured by ICP-AES (gray) and those obtained by applying the model ANN3 (white) to the LIBS data, for both the validation and the test sets. Errors bars correspond to the standard deviation of five repetitions of the ANN calculation.

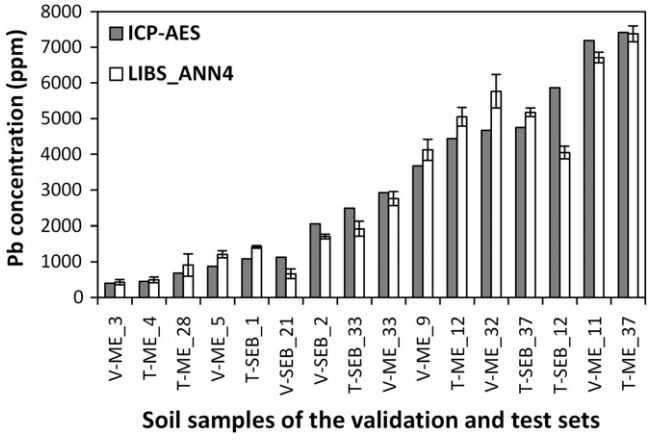

Fig. 5. Comparison chart of the lead concentrations (ppm) of soil samples measured by ICP-AES (gray) and those obtained by applying the model ANN4 (white) to the LIBS data, for both the validation and the test sets. Errors bars correspond to the standard deviation of five repetitions of the ANN calculation.

3.3. Classification of soil samples

Classification of samples prior to quantitative analysis is essential, not only for soils but also for any kind of samples. This classification can be achieved by the use of a new ANN model. Actually, ANN can be designed not only for quantitative analysis but also for classification purpose. In the present study, a 3-layer model was used with one single neuron in the output layer. This ANN was designed to provide the output value 1 when the samples belong to the class 1, i.e. the class related to lead concentrations higher than 10 000 ppm. Symmetrically, the output value of the ANN was expected to be 0 otherwise. This second case should naturally correspond to the class 0 related to lead concentrations lower than 10 000 ppm. Thus any unknown sample should be classified either into class 1 or class 0.

The 117 original samples were split into three data subsets: 76 into the calibration set, 21 into the validation set and 20 into the test set. The data introduced into the ANN model were the 10 spectral lines previously discussed in Section 3.2, namely 5 lines from Pb and 5 lines from the matrix elements (Al, Ba, Ca, Fe and Ti). Ideally, the ANN model was expected to provide a pure binary answer: 0 or 1. However, the activation function of each perceptron being the sigmoid function, the actual output value of the ANN was potentially continuously varying between 0 and 1. Consequently, samples related to output values slightly lower than 1 could be misclassified. It was thus necessary to introduce a tolerance value to overcome this difficulty. This means that an output value close to 1 but different than 1 could also drive to classify the related sample into class 1. We decided to set to 0.05 the tolerance value. In this case, all the samples giving output values higher than 0.95 were classified into class 1.

After optimizing the ANN model dedicated to this classification in two classes, the sensitivity was found to be equal to 87% and the specificity to 100%. In other words, all the samples classified in the class 0 were actually belonging to this class. At the opposite, few samples were misclassified in the class 1. For these samples, the risk was to further apply an ANN model for quantification that could be inadequate. To go deeper into details, the consequence is that one may want to apply an ANN model optimized for high concentrations to samples characterized by low concentrations.

Moreover, when the tolerance value was set to 0.2, both the sensitivity and the specificity were obtained to be equal to 100%. It should be pointed out that this excellent result was obtained not only for the samples of the calibration set but also for those of the validation and test sets. This meant that, once the learning step is achieved, the ANN model dedicated to classification and exploited with the tolerance value of 0.2 was 100% efficient for this classification purpose.

4. Conclusion

In the first part, we presented the efficiency of ANN applied to LIBS data to predict the type of soil matrix through the calculation of three values related to the silicate, calcareous and ores poles. Indeed, RMSE values lower than 10% compared to the reference values provided by ICP-AES were obtained. This result was very good considering that the LIBS data were measured directly on-site.

Then we studied the ability of ANN to quantify lead in soil samples presenting a very large variability of matrices as well as a very large range of lead concentrations. We observed that a single ANN model was not sufficient in this case to reach good quantitative results. Consequently, we decided to split the series of samples into three different subsets and then to apply a specific ANN model to each of them. The first splitting was based on the value of lead concentration regarding a threshold of 10 000 ppm. The concentrations of the samples related to the high concentrations were correctly predicted while the other ones were still difficult to quantify. At this point, a second splitting was applied in order to separate the soil samples into two classes regarding their concentration of silicate. Thus a first ANN model was designed for the soil samples rich in silicate (higher than 80%) and a second one was designed for the others.

To summarize, three quantitative ANN models were applied to the LIBS data measured on-site. And despite of the very small amount of matter analyzed by LIBS for each soil sample, typically in the range of hundreds of micrograms, relative error of prediction as low as 20 % was obtained. This result was perfectly satisfying for on-site analysis. Finally, in order to be able to decide which model should be applied to unknown samples, a last ANN model was exploited for classification purpose. We decided to check the ability of ANN to classify the samples into two classes by setting a threshold value of 10 000 ppm for the lead concentration. We obtained both sensitivity and specificity equal to 100%, indicating no error of classification. This result is very important since it allows selecting the right quantitative model after a preliminary classification of the samples.

Finally, it should be pointed out that the original dataset exploited in this study was originating from a collection of soil samples that were analyzed on-site by a mobile LIBS system (Mobilibs III from IVEA SAS). Moreover, the ANN algorithm was rewritten step-by-step, and then implemented into the software Analibs commercialized by IVEA SAS for on-site prediction.

Further work should be dedicated to build a growing database of soils in order to continue to enhance the performance of the ANN models for quantitative LIBS. The strategy consists in building as many ANN models as necessary in order to be able to analyze in the future any soil sample, whatever its matrix.

Acknowledgments

This work was sponsored by the French Environment and Energy Management Agency (ADEME) and by IVEA SAS.

References

[1] R.S. Harmon, R.E. Russo, R.R. Hark, Applications of laser-induced breakdown spectroscopy for geochemical and environmental analysis: a comprehensive review, Spectrochim. Acta Part B 87 (2013) 11–26.
[2] Anna P.M Michel, Review: applications of single-shot laser-induced breakdown spectroscopy, Spectrochim. Acta Part B 65 (2010) 185–191.
[3] J. Frank, C. De Lucia, J.L. Gottfried, Rapid analysis of energetic and geo-materials using LIBS, Mater. Today 14 (2011) 274–281.
[4] J. Rakovsky, O. Musset, J. Buoncristiani, V. Bichet, F. Monna, P. Neige, P. Veis, Testing a portable laser-induced breakdown spectroscopy system on geological samples, Spectrochim. Acta Part B 74–75 (2012) 57–65.
[5] B. Bousquet, G. Travaillé, A. Ismaël, L. Canioni, K. Michel-Le Pierrès, E. Brasseur, S. Roy, I. le Hecho, M. Larregieu, S. Tellier, M. Potin-Gautier, T. Boriachon, P. Wazen, A. Diard, S. Belbèze, 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 (2008) 1085–1090.
[6] M.A. Ismail, H. Imam, A. Elhassan, W.T. Youniss, M.A. Harith, LIBS limit of detection and plasma parameters of some elements in two different metallic matrices, J. Anal. At. Spectrom. 19 (2004) 489–494.
[7] W.T.Y. Mohamed, Improved LIBS limit of detection of Be, Mg, Si, Mn, Fe and Cu in aluminum alloy samples using a portable Echelle spectrometer with ICCD camera, Opt. Laser Technol. 40 (2008) 30–38.
[8] A. Stankova, N. Gilon, L. Dutruch, V. Kanicky, A simple LIBS method for fast quantitative analysis of fly ashes, Fuel 89 (2010) 3468–3474.
[9] A.S. Eppler, D.A. Cremers, D.D. Hickmott, M.J. Ferris, A.C. Koskelo, Matrix effects in the detection of Pb and Ba in soils using laser-induced breakdown spectroscopy, Appl. Spectrosc. 50 (1996) 1175–1181.
[10] A.W. Andrzej, V. Palleschi, I. Schechter, Laser-induced Breakdown Spectroscopy (Libs): Fundamentals and Applications, 2006.
[11] W.T.Y. Mohamed, A. Askar, Study of the matrix effect on the plasma characterization of heavy elements in soil sediments using LIBS with a portable echelle spectrometer, Prog. Phys. 1 (2007) 46–52.
[12] D. D az, D.W. Hahn, A. Molina, Evaluation of laser-induced breakdown spectroscopy (LIBS) as a measurement technique for evaluation of total elemental concentration in soils, Appl. Spectrosc. 66 (2012) 99–106.
[13] A. Ciucci, M. Corsi, V. Palleschi, S. Rastelli, A. Salvetti, E. Tognoni, New procedure for quantitative elemental analysis by laser-induced plasma spectroscopy, Appl. Spectrosc. 53 (1999) 960–964.
[14] N.C. Dingari, I. Barman, A.K. Myakalwar, S.P. Tewari, M. Kumar Gundawar, Incorporation of support vector machines in the LIBS toolbox for sensitive and robust classification amidst unexpected sample and system variability, Anal. Chem. 84 (2012) 2686–2694.

[15] A.K. Myakalwar, S. Sreedhar, I. Barman, N.C. Dingari, S. Venugopal Rao, P. Prem Kiran, S.P. Tewari, G. Manoj Kumar, Laser-induced breakdown spectroscopy-based investigation and classification of pharmaceutical tablets using multivariate chemometric analysis, Talanta 87 (2011) 53–59.
[16] J.B. Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Tellier, M. Potin-Gautier, I.L. Hecho, Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis, Anal. Bioanal. Chem. 385 (2006) 256–262.
[17] B. Bousquet, J.B. Sirven, L. Canioni, Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples, Spectrochim. Acta Part B 62 (2007) 1582–1589.
[18] P.M. Mukhono, K.H. Angeyo, A. Dehayem-Kamadjeu, K.A. Kaduki, Laser induced breakdown spectroscopy and characterization of environmental matrices utilizing multivariate chemometrics, Spectrochim. Acta Part B 87 (2013) 81–85.
[19] J.L. Gottfried, R.S. Harmon, F.C. De Lucia Jr, A.W. Miziolek, Multivariate analysis of laser-induced breakdown spectroscopy chemical signatures for geomaterial classification, Spectrochim. Acta Part B 64 (2009) 1009–1019.
[20] J.-B. Sirven, B. Salle, P. Mauchien, J.-L. Lacour, S. Maurice, G. Manhes, Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods, J. Anal. At. Spectrom. 22 (2007) 1471–1480.
[21] M.r.J.C. Pontes, J. Cortez, R.K.H. Galvão, C. Pasquini, M.r.C.s.U. Araújo, R.M. Coelho, M. r.K. Chiba, M.n.F. de Abreu, B.t.E.k. Madari, Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain, Anal. Chim. Acta. 642 (2009) 12–18.
[22] M.M. Tripathi, K.K. Srinivasan, S.R. Krishnan, F.-Y. Yueh, J.P. Singh, A comparison of multivariate LIBS and chemiluminescence-based local equivalence ratio measurements in premixed atmospheric methane-air flames, Fuel 106 (2012) 318–326.
[23] J.M. Andrade, G. Cristoforetti, S. Legnaioli, G. Lorenzetti, V. Palleschi, A.A. Shaltout, Classical univariate calibration and partial least squares for quantitative analysis of brass samples by laser-induced breakdown spectroscopy, Spectrochim. Acta Part B 65 (2010) 658–663.
[24] M.E. Essington, G.V. Melnichenko, M.A. Stewart, R.A. Hull, Soil metals analysis using laser-induced breakdown spectroscopy (libs), Soil Sci. Soc. Am. J. 73 (2009) 1469–1478.
[25] M.D. Dyar, M.L. Carmosino, E.A. Breves, M.V. Ozanne, S.M. Clegg, R.C. Wiens, Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples, Spectrochim. Acta Part B 70 (2012) 51–67.
[26] R.C. Wiens, S. Maurice, J. Lasue, O. Forni, R.B. Anderson, S. Clegg, S. Bender, D. Blaney, B.L. Barraclough, A. Cousin, L. Deflores, D. Delapp, M.D. Dyar, C. Fabre, O. Gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P.Y. Meslin, H. Newsom, A. Ollila, R. Perez, R.L. Tokar, D. Vaniman, Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover, Spectrochim. Acta Part B 82 (2013) 1–27.
[27] E.C. Ferreira, D.b.M.B.P. Milori, E.J. Ferreira, R.M. Da Silva, L. Martin-Neto, Artificial neural network for Cu quantitative determination in soil using a portable laser induced breakdown spectroscopy system, Spectrochim. Acta Part B 63 (2008) 1216–1220.
[28] J. El Haddad, M. Villot-Kadri, A. Ismaël, G. Gallou, K. Michel, D. Bruyère, V. Laperche, L. Canioni, B. Bousquet, Artificial neural network for on-site quantitative analysis of soils using laser induced breakdown spectroscopy, Spectrochim. Acta Part B 79–80 (2013) 51–57.
[29] C. Rucker, G. Rucker, M. Meringer, y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345–2357.