Statistical comparison of predictive models for quantitative analysis and classification in the framework of LIBS spectroscopy: A tutorial

Ludovic Duponchel, Cécile Fabre, Bruno Bousquet, Vincent Motto-Ros

Statistical comparison of predictive models for quantitative analysis and classification in the framework of LIBS spectroscopy: a tutorial

Ludovic Duponchel[†*], Cécile Fabre[Ⅱ], Bruno Bousquet[⊥], Vincent Motto-Ros[‡]

[†] Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, F-59000, France.

[Ⅱ] GeoRessources UMR 7359, Université de Lorraine-CNRS, 54000 Vandoeuvre les Nancy, France.

[⊥] Institut de Chimie de la Matière Condensée de Bordeaux (ICMCB) UMR5026, CNRS, Univ. Bordeaux, Bordeaux INP, 33600 Pessac, France.

[‡] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon 69622 Villeurbanne, France.

Corresponding Authors

* ludovic.duponchel@univ-lille.fr

1. **Keywords :** Laser-Induced Breakdown Spectroscopy (LIBS), Quantitative analysis, classification, model comparison, statitiscal test, significance, chemometrics.

ABSTRACT: Laser-Induced Breakdown Spectroscopy (LIBS) is a widely accepted technique used for both classification and quantification purposes considering complex and heterogeous samples. Based on a set of training spectra acquired from diverse and representative samples within a specific application domain, it becomes possible to apply various data processing techniques and modeling methods to construct the predictive model in question. Naturally the complexity of both the laser-matter and the laser-plasma interactions and the heterogeneity of natural samples often requires the development of various predictive models, which are then compared based on figures of merit such as

the RMSEP (Root Mean Square Error of Prediction) value for quantification or the classification rate for qualitative analysis. Our ultimate goal is, of course, to select the model that appears to be the most accurate, which ultimately boils down to searching for the lowest RMSEP value or the highest classification rate. This is precisely where the whole problem lies because even if we observe a different level of error for two models, for example, this difference is not necessarily statistically significant. In such a case, we are therefore not allowed to say that the lower error indicates the best predictive model to consider. The purpose of this article is to provide a tutorial on introducing a statistical model comparison procedure, whether they are quantitative or qualitative. Two LIBS data sets have been used to illustrate the principles of the proposed method.

INTRODUCTION

Laser-Induced Breakdown Spectroscopy (LIBS) is nowadays a well-established technique of elemental analysis. It has been successfully applied for both quantification and classification purposes, utilizing microscopes [1], handheld devices [2,3], and standoff instruments [4]. High-quality LIBS spectra are now routinely captured, and diverse data processing and modelling strategies are employed to achieve optimal analytical performance. Quantification is achieved, for example, through univariate regression (namely classical calibration curves), as well as through multivariate approaches such as Partial Least Squares (PLS) regression [5–7] and Artificial Neural Network (ANN) [8], among others. Classification is typically accomplished using other multivariate tools that rely on methods like K-Nearest Neighbors (KNN) [9,10], Soft Independent Modeling of Class Analogy (SIMCA) [11], and Support Vector Machines (SVM) [12], among others. Beyond the choice of the predictive tool, the analyst must also consider the entire spectral correction pipeline, which may include, but is not limited to, denoising, baseline correction, or even normalization. It is evident that no one can claim to be able to select, from scratch, an optimal spectral correction and predictive algorithm for the given dataset. That is why we develop different combinations of models and spectral data treatments, hoping to be able to obtain a good one. The evaluation and comparison of predictive models in LIBS analysis pose unique challenges. The performance of a predictive model is typically assessed using figures of merit

such as the Root Mean Square Error of Prediction (RMSEP) for quantification tasks or the classification rate for qualitative analysis. These metrics provide valuable insights into the predictive capability of a model. However, the simple comparison of error values or classification rates between models may not be sufficient to draw meaningful conclusions. The crux of the problem lies in determining the statistical significance of the differences observed in model performance. Merely observing a lower error value does not necessarily imply that the corresponding model is superior. It is imperative to account for the inherent variability in LIBS data, as well as the uncertainties associated with the model fitting process. Robust statistical methods and procedures are therefore required to enable researchers to make confident decisions regarding model selection and performance evaluation. This article aims to address this challenge by providing a comprehensive tutorial on introducing a statistical model comparison procedure for LIBS data analysis, encompassing both quantitative and qualitative aspects. The tutorial will provide step-by-step guidance on implementing this procedure and interpreting the results. Furthermore, it will utilize two LIBS data sets, carefully selected to exemplify the principles and practical implementation of the proposed method. By the end of this tutorial, readers will gain a solid understanding of the challenges associated with model comparison in LIBS analysis and will be equipped with the knowledge and tools necessary to evaluate and select the most appropriate predictive model for their specific applications.

MATERIAL AND METHODS

Statistical strategy to compare two quantitative models

Assume two predictive models (denoted $model_1$ and $model_2$) that have been developed to predict a y value (which is often a concentration) from LIBS spectral information. Both of these models may of course have been developed in the univariate framework (i.e. predicting a quantity from the emission at a single wavelength), the first using a linear regression and the other using a non-linear one. Two univariate regressions each using a different wavelength could also be considered. From a chemometric point of view (i.e. a multivariate analysis one), these two models can also have been developed with very different strategies such as a multilinear method like the well-known PLS regres-

sion or a non-linear one like a neural network and much more. We can also consider a unique multivariate method like PLS regression using for example different data preprocessing for the two models to be compared. Moreover, these two models can also have used different spectral information as for example all the wavelengths of the spectral domain or a selection of sub-spectral domains and/or wavelengths. These two models could also have used the same or two different calibration data sets. From this description, we can see that we can consider all the experimental conditions for these two models that we wish to compare, and why not compare a univariate model and a multivariate one if we wish. We could even go further by comparing, for example, the predictive power of a first model utilizing data from a specific spectroscopic technique and a second model based on another one (such as LIBS vs X-ray fluorescence). In a way, we can say that the only constraint imposed on the models is to exploit the spectral data acquired on a given sample to predict the concentration of a given element of interest. On the other hand, we have much less freedom regarding the procedure we propose in this article to compare the predictive capabilities of these two quantitative models. Nevertheless, we will see that this constraint is minimal and quite logical. Indeed, we have first to use the same test set of *n* samples. It is obvious that these samples must never have been used by the models during their development. Classically, we can evaluate the predictive power of the two models by calculating their Root Mean Square Error of Prediction (*RMSEP*$_1$ and *RMSEP*$_2$) defines as:

$$RMSEP_1 = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i,1}-y_i)^2}{n}} \quad \text{and} \quad RMSEP_2 = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i,2}-y_i)^2}{n}} \qquad \text{(eq. 1)}$$

where $\hat{y}_{i,1}$ and $\hat{y}_{i,2}$ are respectively the predicted concentration for the sample *i* by *model*$_1$ and *model*$_2$, y_i the reference value for the same sample *i* and *n* the total number of sample in the test set. In all statistical rigor, we cannot directly compare these two RMSEP values without considering the potential bias observed on each model. The bias is the systematic error made by the model at the time of a prediction which must be very small or ideally zero. We can thus calculate the bias of each model by the equations:

$$bias_1 = \frac{1}{n}\sum_{i=1}^{n} e_{i,1} \quad \text{with} \quad e_{i,1} = \hat{y}_{i,1} - y_i \qquad \text{(eq. 2)}$$

$$bias_2 = \frac{1}{n}\sum_{i=1}^{n} e_{i,2} \quad \text{with} \quad e_{i,2} = \hat{y}_{i,2} - y_i \qquad \text{(eq. 3)}$$

where $e_{i,j}$ is the residual (i.e. the prediction error) of the sample *i* given by the model *j*. There is no specific threshold that allows us to determine if a given model's bias is significant, but typically, the experimenter compares it to the RMSEP value, which is, in turn, compared to the range of concentration of the product of interest. This is also an opportunity to introduce the Standard Error of Prediction (SEP) which takes this bias into account for the two models:

$$SEP_1 = \sqrt{\frac{\sum_{i=1}^{n}(e_{i,1}-bias_1)^2}{n-1}} \quad \text{and} \quad SEP_2 = \sqrt{\frac{\sum_{i=1}^{n}(e_{i,2}-bias_2)^2}{n-1}} \qquad \text{(eq. 4)}$$

The methodology we propose here is a pairwise comparison of the two models. We have not invented anything since this methodology was proposed by Pitman in 1939 [13]. This methodology has also been applied in other areas of analytical chemistry [14]. Readers interested in more statistical details are invited to read the book by Snedecor and Cochran, for example [15]. From a statistical standpoint, it is not possible to directly compare the RMSEP values of two models if biases are present. The model comparison procedure will therefore be broken down into two steps. First, we will try to find out if there is a statistically significant difference between the biases of the two models and in a second step we will do the same for the two SEP values. In order to compare the two biases, we will use a *t* confidence interval of paired samples. Thus the difference between the two biases $bias_1 - bias_2$ has a standard deviation S_d :

$$S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i-\overline{d})^2}{n(n-1)}} \qquad \text{(eq. 5)}$$

with $d_i = e_{i,1} - e_{i,2}$ the difference between the two errors for the sample *i* and $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$. It is then possible to define a 95 % confidence interval for the true difference in biases:

$$(biais_1 - biais_2) \pm t_{n-1,0.025} \times S_d \qquad \text{(eq. 6)}$$

with $t_{n-1,0.025}$ the upper 2.5% point of a *t* distribution on *n* - 1 degrees of freedom. You can find this value in any statistics book that offers tables of *t*-values. Then this interval would be expected to include the true difference in biases 95% of the time if we repeat the prediction. As a consequence, if the interval we have just calculated includes zero, then the biases are considered as not significantly different at the 5% level. Generally speaking, we build good models with good quality data which

potentially do not present any real bias and which therefore quite often will not be significantly different. However, this is the occasion to insist on the interest of checking the bias of a constructed model because a systematic error is not acceptable. So we understand that the most important step is now the comparison of the SEP values of the two models. The idea is again to find a confidence interval. To do this, we need to calculate the parameters *K* and *L*:

$$K = 1 + \frac{2(1-r^2)t^2_{n-2,0.025}}{n-2} \quad \text{and} \quad L = \sqrt{K + \sqrt{K^2 - 1}} \qquad \text{(eq. 7)}$$

with r the correlation coefficient between the two sets of prediction errors and $t_{n-2,0.025}$ the upper 2.5% point of a *t* distribution on *n* - 2 degrees of freedom. Then, we can say that the lower and the upper limits of a 95% confidence interval for the ratio of the two SEP values are respectively $\frac{SEP_1}{SEP_2} \times \frac{1}{L}$ and $\frac{SEP_1}{SEP_2} \times L$. So if we calculate this interval and it contains the value 1, we can say that SEP_1 and SEP_2 are not significantly different at the 5% level. In these conditions, we will not be able to say that one model is better than the other, even if we observe two different numerical error values. A matlab code (named *quanti_sig.m*) is provided in supplementary material in order to statistically test differences between two quantitative models.

Statistical strategy to compare two classification models

Assume now two classification models (denoted $model_1$ and $model_2$) that have been developed to predict a class membership of a sample from LIBS spectral information. As for the previous section, these two classification models may have been developed using very different supervised methods such as k-NN, SIMCA, PLS-DA, shallow neural networks, deep neural networks and much more. Of course these two classification models can also have been developed with the same method but using for example different parameters or even different preprocessing. Additionally, these two classification models may have used different spectral information. In a natural way, the evaluation of a classification model accuracy is obtained by predicting the class of the samples of a test set having of course never been seen by the model during its development. A common figure of merit we use to estimate the potential of a classification model is the classification rate, which is the percentage of correct

classifications for all the samples in the test set. It then seems logical to directly compare the classification rates resulting from the analysis of the same test set for the two models we consider. We will see now that it is not exactly what we will do with the proposed McNemar's test in order to compare two classification models [16]. This statistical test was introduced by Qinn McNemar in 1947 in order to compare paired nominal data which is the case here for the prediction obtained from two classification models on the same test set. It is based on a χ^2 test with one degree of freedom. The null hypothesis here is that $model_1$ and $model_2$ have the same percentage of well-predicted samples i.e. the same classification rate. As a first step, we present the prediction results of both $model_1$ and $model_2$ through a single contingency table as described in table 1.

Table 1: the contingency table used in a McNemar statistical test.

	Number of test samples well classified by $model_2$	Number of test samples misclassified by $model_2$
Number of test samples well classified by $model_1$	a	b
Number of test samples misclassified by $model_1$	c	d

The McNemar's test will only use two values in this table i.e. c the number of test samples misclassified only by the $model_1$ and b the number of test samples misclassified only by the $model_2$. We thus see that it does not focus on the predictive character of each of the classification models but on the differences that they could present during prediction. It is then written as follows:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \qquad \text{(eq. 8)}$$

This calculated χ^2 value is then compared with the χ^2 critical value with a 5% level of significance which is 3.8414. Then if the calculated χ^2 value is greater than 3.8414, the null hypothesis is false and the two classification models are significantly different. If the calculated χ^2 value is lower than 3.8414, the null hypothesis is true and the two classification models are not statistically different with a risk of 5% and this even if their classification rates seem to have different values. A matlab code (named

classi_sig.m) is also provided in supplementary material in order to statistically test differences between two classification models.

Datasets description for quantitative analysis

The comparison of quantitative models will be done as an example within the context of lithium determination in rocks. Indeed, LIBS analysis is one of the unique techniques that can provide identification and quantification of light elements such as lithium [17,18]. For your information, all the spectra used in this study for developing and validating the models are accessible within an open-source database [17]. They have been acquired using a handheld LIBS instrument [3,19,20] from homogeneous Li-bearing minerals such as spodumene ($LiAlSi_2O_6$), petalite ($LiAlSi_4O_{10}$), amblygonite or montebrasite ($LiAl(PO_4)(F,OH)$), lepidolite ($K_2(Li,Al)_{5-6}(Si_{6-7}Al_{2-1}O_20)(OH,F)_4$), zinnwaldite ($KLiFeAl(AlSi_3)O_{10}(OH,F)_2$) or altered Li-minerals, from natural rocks enriched in lithium and powder pellets. A wavelength calibration of the spectra has also been implemented to ensure robust elemental identification [17,21]. To account for the heterogeneity of the samples, we considered a mean spectrum obtained from analyzing 9 to 15 regions for each of them. The Li reference concentrations have been estimated using Inductively Coupled Plasma - Optical Emission Spectrometry (ICP-OES) and expressed in Li wt %. Thus, 76 mean spectra constituted the calibration set, and 21 others were used for the test set.

Datasets description for classification

In the context of utilizing artificial neural networks within the LIBS imaging framework, the focus will be on comparing classification models specifically for characterizing archaeological lime mortar [8]. Such materials result from the hardening of a mixture composed of a binder, various types aggregates and water. They are by nature highly heterogeneous and complex materials but rich in information for building archaeology. In this study, we have defined a total of 9 classes to represent all the material types that can be found in the majority of mortars. This includes the binder, as well as various aggregates used in the production process (carbonate, quartz, aluminosilicate, coal and tile). Besides, a class hole has also been added as well as a class resin to take into account the sample preparation. To take into account possible micro recrystallization, a postbinder class has also been

added. The reference spectra for each class were produced on a corpus of 27 samples. Several raw and resin-embbeded mortars were analyzed as well as raw materials, such as ceramic (x3), quartz (x2), epoxy resin (x1), coal (x2), limestone (x2), marble (x4), speleothem (x1) and shell (x4). For more information about the context of such study and the experimental procedures, the authors can referes to the folowing recent publications [8,22]. Table s1 also gives the distribution of spectra by class and by sample used for this neural network. Among the 1447 reference spectra associated to the 9 classes, 80% were used for training (i.e. 1160) and 20% for the model test (i.e. 287).

RESULTS AND DISCUSSION

The first part of this section will focus on the statistical comparison of quantitative models within the framework of lithium dosage. Due to the natural heterogeneity of rock samples and the presence of potential matrix effects, it is quite logical to consider multivariate models such as Partial Least Squares regression (PLS), which is widely used in the LIBS community. As a reminder, PLS regression is based on the search for factors whose number needs to be optimized in order to achieve the best accuracy. This optimization was performed through a venitian blind cross-validation in this study. As we frequently do in chemometrics, it seemed interesting to us to develop an initial PLS model (denoted $model_1$) using the 76 spectra from the calibration set without any corrections (i.e. no spectral pre-processing). Cross-validation then indicated an optimal number of 2 PLS factors. This model was then used to predict lithium concentrations for the 21 samples from the test set, resulting in a $RMSEP_1$ value of 0.4464 Li wt %. At first glance, this error may seem significant given the concentration range of the calibration set, which extends from 0.10 to 3.54 Li wt %. However, this is not the case because these rock samples are particularly heterogeneous and are additionally analyzed using a handheld LIBS instrument that naturally has degraded characteristics compared to a benchtop one. The complexity of our samples and the laser-matter interaction often alter the relationship that exists between the measured signals and the concentration of an element of interest. That is why we often apply spectral preprocessing prior to using multivariate regression methods. We have thus developed a second PLS model (denoted $model_2$) using the 76 calibration spectra corrected with the Standard Normal Variate (SNV) method, which is a type of normalization. Cross-validation indicated in this

case an optimal number of 5 PLS factors, which was then tested on the 21 samples of the test set, resulting in a $RMSEP_2$ value of 0.4832 Li wt %. Without a statistical perspective, and as we can see in the vast majority of studies, almost all of us would conclude that $model_1$ is the best since 0.4464 Li wt % is lower than 0.4832 Li wt %. The whole question boils down to whether we are justified in saying that $model_1$ is indeed the best simply because its RMSEP value is the lowest, if we only consider the strict numerical value. The whole purpose of this publication is to show that a statistical approach is needed in order to determine whether these two models are different and, consequently, if one of them is indeed superior. As already explained in the Materials and Methods section, it is not possible to propose a statistical test that allows for comparing the RMSEP values of the two PLS models. We must first examine the calculation of bias in both models and then ensure that there are no statistically significant differences betwwen them. Based on the reference concentration values from the test set and the predicted values from both models, it is easy to calculate their biases (eq. 2 and 3) and then determine a 95 % confidence interval for the true difference in biases. The figure 1a provides a graphical representation of this bias comparison between the two PLS models resulting from the use of the MATLAB code provided in the supplementary material (named *quanti_sig.m*).

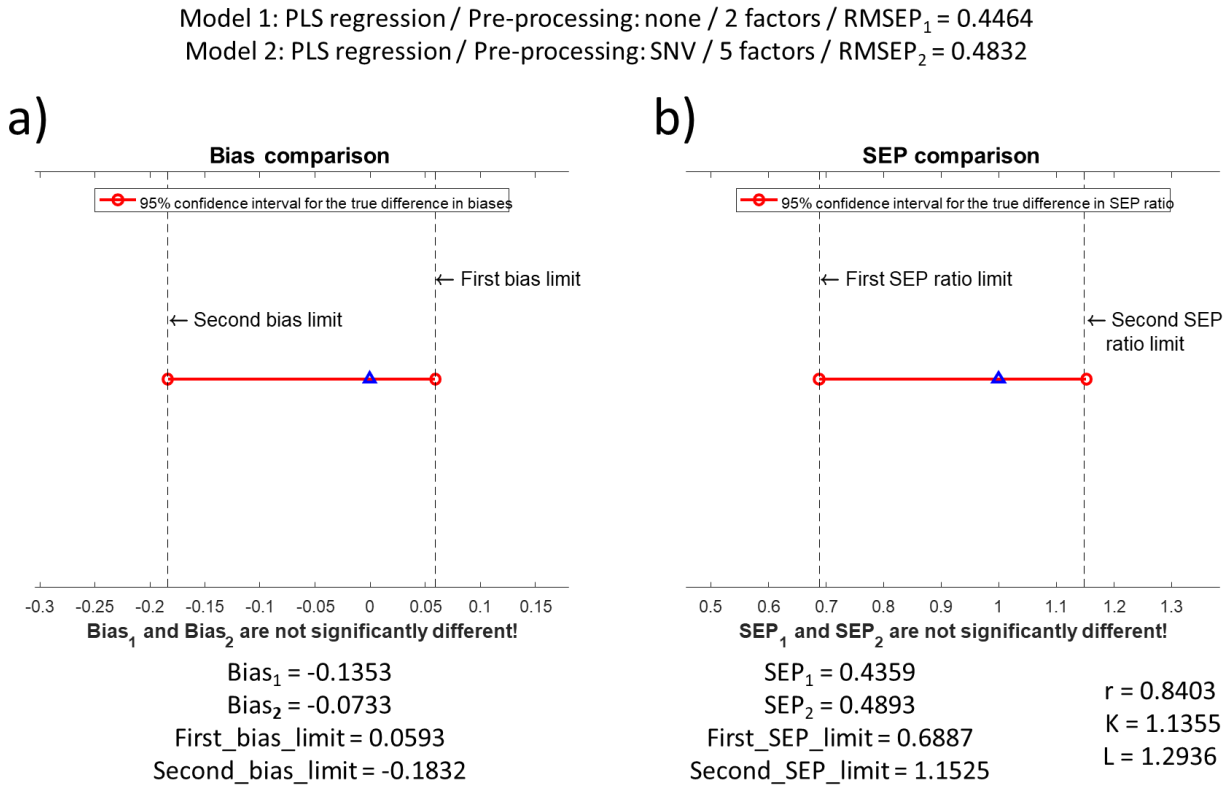

Figure 1: statistical tests for a) bias comparison and b) SEP comparison considering the two PLS regressions

We thus observe weak biases for $models_1$ and $models_2$ (-0.0135 and -0.0733 Li wt % respectively). The 95 % confidence interval for the true difference in biases (represented in red in Figure 1a) has then lower and upper limits of -0.1832 and 0.0593 Li wt % respectively. Since this interval contains the value zero (represented in blue in the figure), then $bias_1$ and $bias_2$ are considered as not significantly different at the 5% level. Based on this assessment, we can now move on to the second step, which involves comparing the standard error of prediction errors (*SEP*) of the two models (Figure 1b). We observe SEP_1 and SEP_2 values equal to 0.4359 and 0.4893 Li wt %, respectively. Based on an estimated correlation coefficient between the two sets of prediction errors from the two models (r = 0.8403), we can then calculate the parameter L, which is equal to 1.2936 in this case (see eq. 7). This final parameter ultimately allows calculating the lower and the upper limits of a 95% confidence interval for the ratio of the considered SEP values, which have values of 0.6887 and 1.1525 in this case. This interval is represented in red in Figure 1b. With a value of 1 contained within it (represented in blue in the figure), there is no significant difference between SEP_1 and SEP_2. From a statistical standpoint, we are therefore not allowed to say that $model_1$ is better than $model_2$. The question that naturally arises then is, what do we do now because we must indeed select a final model to use. The basic rule in statistics is to remember that the most robust model is usually the least complex one. Therefore, we should prioritize the use of the PLS model with the fewest components, namely the first one using only two PLS factors. It is now interesting to know under what conditions we would have a new $model_2$ that would become significantly different from $model_1$. Assuming that the correlation coefficient between the two models remains equal to 0.8403, it would then be necessary for the SEP_2 value to be greater than 0.5639 Li wt % (i.e. $SEP_1 \times$ L). We can see that the value of L is crucial in this search for significance of differences. Thus, the value of L must be as small as possible if we want to highlight a significant difference between two close *SEP* values. It is therefore appropriate to examine the evolution of L as a function of the number *n* of observations present in the test set and the correlation coefficient *r* between the two models being compared. Figure 2 provides a representation of this evolution calculated from equation 7. For a given correlation coefficient, we observe that L decreases as

the number *n* increases. Therefore, we should always strive to have the maximum number of samples in the test set to properly evaluate models. Under these conditions, we will effectively be able to demonstrate that there is a significant difference, even if the SEP values of the two models are close. We also observe that as the correlation becomes weaker, the value of L increases for a fixed number of samples. However, this is a parameter that we do not control when comparing two models.

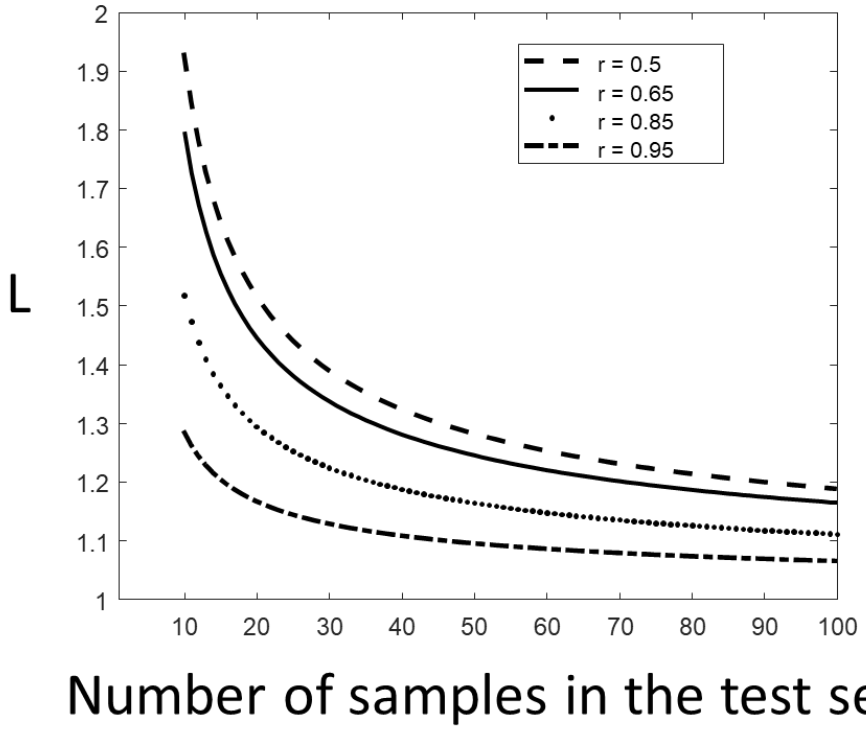

Figure 2: Evolution of L as a function of the number *n* of observations present in the test set and the correlation coefficient r

Regarding the applicability of such a procedure for statistical comparison of quantitative models, it can only be considered with a number of test samples greater than 20. In fact, it would be unreasonable to go below this number to claim a meaningful statistical comparison. In order to continue this discussion regarding the comparison of quantitative models, it seemed interesting to us to introduce a third predictive model (denoted $model_3$), this time more traditional, as it is based on a simple linear regression by exploiting only a specific emission line of lithium located around 610 nm (instead of the whole spectral range used by the two previous PLS models). Based on our pairwise procedure, the idea was to compare, this time, $model_1$ and $model_3$ always on the same test set. Just like in the previous case,

providing reference values for the test samples and their predicted values by these two models, $RMSEP_1$ was always egal to 0.4464 Li wt % when $RMSEP_3$ was significantly larger with a value of 0.6090 Li wt %. This is due to the fact that a simple linear model has little chance of being able to handle matrix effects. As we saw previously, we are not allowed to directly compare these two *RMSEP* values. Therefore, we begin by comparing the biases of these two models (Figure 3a). We once again have completely appropriate biases with $bias_1$ and $bias_3$ equal to -0.1353 and -0.1983 Li wt % respectivly. Again in this case, the 95 % confidence interval for the true difference in biases contain the value zero and therefore the two biases are considered as not significantly different. We can now move on to the step of calculating and comparing the *SEP* values (Figure 3b).

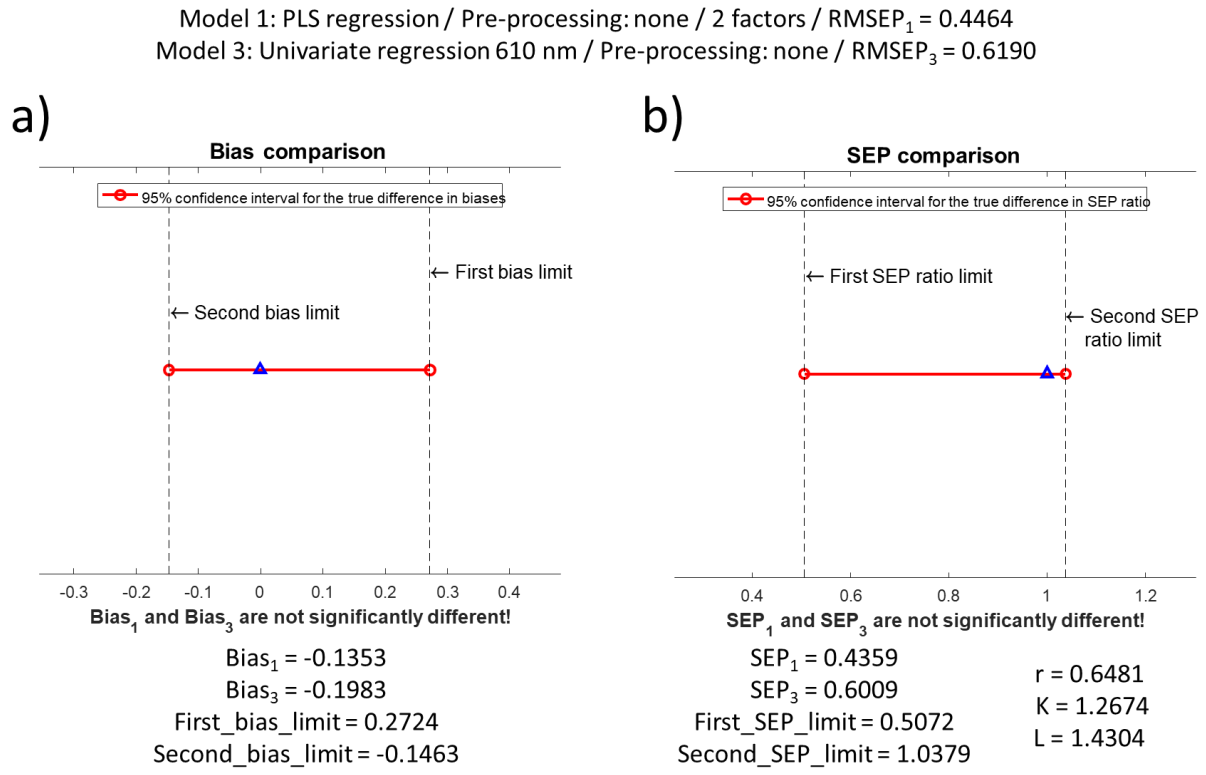

Figure 3: statistical tests for a) bias comparison and b) SEP comparison considering a PLS model and a univariate one.

We observe SEP_1 and SEP_3 values equal to 0.4359 and 0.6009 Li wt % respectively, which intuitively might lead us to believe that such a numerical difference could finally highlight a significant statistical difference. However the 95% confidence interval for the ratio of the considered *SEP* values contains the value of 1 indicating that SEP_1 and SEP_3 are not statistically different. This may seem strange considering the observed results for the two first models ($model_1$ and $model_2$), but we have here a

correlation coefficient *r* that has significantly decreased, as it is now only 0.6481, which has mechanically increased the value of L to 1.4304. From a statistical standpoint, we cannot say that the Partial Least Squares (PLS) model is better than the univariate model using a single wavelength in this particular case. If we take a closer look, it is true that we were not far from highlighting a significant difference between the two *SEP* values. In fact, we can see in Figure 3b that the blue triangle representing the value 1 is very close to the confidence interval limit, but it is not enough. In fact, considering the same correlation coefficient *r*, the univariate model should have had a higher *SEP* value than 0.6235 (i.e. $SEP_1 \times L$) in order to be considered different from the PLS model. In any case, the three models compared here are not statistically different under the chosen experimental conditions. Once again, our quest for robustness would lead us to choose simple linear regression, which corresponds to the least complex model. If such a conclusion were to bother experimenters, the only way for them to demonstrate the superiority of one of these models would be to increase the number of samples in the test set. The entire calculation procedure would, of course, need to be redone for these new conditions.

This second part of the discussion will now focus on the comparison of classification models. We will see that this procedure is a bit faster as it operates in a single step, unlike the comparison of quantitative models. As an example, we will focus here on the use of neural networks, which are known for their strong prediction capabilities, particularly for nonlinear modeling problems. Nevertheless, we must not lose sight of the fact that for the implementation of such an approach, we always need to have a high number of spectra available to train the network while simultaneously attempting to minimize the number of neurons in its architecture (i.e. the number of weights to optimize). In our case, we do have a significant training dataset as it consists of 1160. However, the spectral domain consists of 3x2048 wavelengths (i.e. 3 spectrometers were used simulatenously), and it is not feasible to use all of them as input to a network. Indeed, this would naturally result in $(3 \times (2048 + 9) \times h)$ weights to optimize considering a single hidden layer with h neurons and 9 output ones, which would be too high. As a consequence, the first two neural networks that we trained only used 27 wavelengths from the spectral domain, corresponding to emission lines from specific elements. Details regarding

this list of selected emission lines can be found in [8]. Thus these two networks had 27 neurons in the input layer. The predictive capacity of a neural network is also dependent on the number of neurons in the hidden layer. Therefore, we have developed two different architectures. The first neural network had 50 hidden neurons, and the second one had 200 hidden neurons (referred to as $model_1$ and $model_2$, respectively). The training of these two networks being completed, we then assessed their predictive power by predicting the classes of the 287 spectra in the test set, and comparing them to the reference classes. Thus, $model_1$ accurately predicted the class of 275 spectra out of the total 287, resulting in a percentage of correct predictions equal to 95.82%. On its part, $model_2$ accurately predicted the class of 269 spectra, resulting in a rate of 93.73%. The goal is therefore naturally to determine whether these two models are statistically different, and if so, whether we can say that $model_1$ is better. As we have seen in the Materials and Methods section, the McNemar's test will not specifically address the number of spectra correctly classified by each model. Instead, it will focus on the spectra that are correctly classified by one model and incorrectly classified by the other, and vice versa. The contingency table provided in Table 2 presents the count of these correct or incorrect predictions when considering both models simultaneously.

Table 2: the contingency table used in a McNemar statistical test to compare a first ANN with 50 hidden neurons ($model_1$) and a second one with 200 neurons ($model_2$).

	Number of test samples well classified by $model_2$	Number of test samples misclassified by $model_2$
Number of test samples well classified by $model_1$	268	7
Number of test samples misclassified by $model_1$	1	1

From the values in the contingency table and equation 8, it is possible to calculate a χ^2 value equal to 3.125. This value being lower than the critical χ^2 value of 3.841, we demonstrate that there is no significant difference between the predictions of these two neural networks. At first glance, we could thus choose one of the two networks, but our quest for robustness in prediction compels us to select, as

usual, the less complex model. In the context of neural networks, this means favoring the structure with fewer weights, and therefore here, the model utilizing 50 neurons in the hidden layer (i.e., $model_1$). Having done this, we then wondered if it was possible to have a neural network that could use fewer neurons in the input layer by selecting only 9 wavelengths while keeping the number of hidden neurons at 50. This new model, named $model_3$, would then be compared to our $model_1$. As in the previous step, following the training of this new network, the class of the 287 spectra in the test were predicted. However, the class labels of only 259 spectra in the test set were accurately predicted, resulting in a classification rate of 90.24%. From a strictly numerical perspective, we observe a greater difference between $model_1$ and $model_3$, here, and the question is whether it is significant. From the new contingency table (Table 3), we then calculate a χ^2 value of 12.5 that is now higher than the critical value. The two networks $model_1$ and $model_3$ are therefore significantly different from a statistical point of view, and as a result, we can say that $model_1$ is better than $model_3$.

Table 3: the contingency table used in a McNemar statistical test to compare a first ANN using 27 input neurons ($model_1$) and another one using 9 input neurons ($model_3$) both using 50 hidden neurons

	Number of test samples well classified by $model_3$	Number of test samples misclassified by $model_3$
Number of test samples well classified by $model_1$	258	17
Number of test samples misclassified by $model_1$	1	1

CONCLUSION

The aim of this publication was to introduce statistical procedures for comparing the predictive power of quantitative and qualitative models. We felt it was important to provide such a tutorial because we often observe model choices being made based solely on a single observation of the numerical value of an RMSEP error or a classification rate in numerous studies. Thus, even though some readers may initially be intimidated by the implementation of statistical tests on their data, the proposed procedures are actually quite straightforward, especially considering that Matlab codes are made

available to the community. We aim through this work to primarily raise awareness among researchers about the importance of statistically comparing the chemometric models they construct. This step is crucial because the selection of a model believed to be the best inevitably has repercussions on the subsequent progress and utilization of our research. It was also an opportunity to emphasize that the simplest models are the most robust ones when it comes to handling variations in spectroscopic measurements that they may encounter throughout their lifespan, which is a fundamental principle of statistics. Finally, we have also shown that the best way to evaluate a quantitative or classification model is to try to have as many samples as possible in the test set, in order to be able to observe real statistical differences even between small SEP values or classification rates.

AKNOWLEDGEMENTS

This work was partially supported by the French region Rhônes Alpes Auvergne (Optolyse, CPER2016), the French "Agence Nationale de la Recherche" ((ANR-22-CE27-0017) "MEMOar" and ANR-20-CE17-0021 "dIAg-EM). In addition, we gratefully acknowledge Nicolas Herryere, Clothilde Comby-Zerbino, Christine Oberlin, and Anne Schmitt for fruitfull discussions. We would like to extend our thanks to the Greentropism company for also providing us with access to their online PLS calculation.

REFERENCES

[1] D.W. Hahn, N. Omenetto, Laser-Induced Breakdown Spectroscopy (LIBS), Part II: Review of Instrumental and Methodological Approaches to Material Analysis and Applications to Different Fields, Appl. Spectrosc. 66 (2012) 347–419. https://doi.org/10.1366/11-06574.

[2] B. Connors, A. Somers, D. Day, Application of Handheld Laser-Induced Breakdown Spectroscopy (LIBS) to Geochemical Analysis, Appl. Spectrosc. 70 (2016) 810–815. https://doi.org/10.1177/0003702816638247.

[3] G.S. Senesi, R.S. Harmon, R.R. Hark, Field-portable and handheld laser-induced breakdown spectroscopy: Historical review, current status and future prospects, Spectrochim. Acta Part B At. Spectrosc. 175 (2021) 106013. https://doi.org/10.1016/j.sab.2020.106013.

[4] S. Maurice, R.C. Wiens, M. Saccoccio, B. Barraclough, O. Gasnault, O. Forni, N. Mangold, D. Baratoux, S. Bender, G. Berger, J. Bernardin, M. Berthé, N. Bridges, D. Blaney, M. Bouyé, P. Caïs, B. Clark, S. Clegg, A. Cousin, D. Cremers, A. Cros, L. DeFlores, C. Derycke, B. Dingler, G. Dromart, B. Dubois, M. Dupieux, E. Durand, L. d'Uston, C. Fabre, B. Faure, A. Gaboriaud, T. Gharsa, K. Herkenhoff, E. Kan, L. Kirkland, D. Kouach, J.-L. Lacour, Y. Langevin, J. Lasue, S. Le Mouélic, M. Lescure, E. Lewin, D. Limonadi, G. Manhès, P. Mauchien, C. McKay, P.-Y. Meslin, Y. Michel, E. Miller, H.E. Newsom, G. Orttner, A. Paillet, L. Parès, Y. Parot, R. Pérez, P. Pinet, F. Poitrasson, B. Quertier, B. Sallé, C. Sotin, V. Sautter, H. Séran, J.J. Simmonds, J.-B.

Sirven, R. Stiglich, N. Striebig, J.-J. Thocaven, M.J. Toplis, D. Vaniman, The ChemCam Instrument Suite on the Mars Science Laboratory (MSL) Rover: Science Objectives and Mast Unit Description, Space Sci. Rev. 170 (2012) 95–166. https://doi.org/10.1007/s11214-012-9912-2.

[5] N. Rethfeldt, P. Brinkmann, D. Riebe, T. Beitz, N. Köllner, U. Altenberger, H.-G. Löhmannsröben, Detection of Rare Earth Elements in Minerals and Soils by Laser-Induced Breakdown Spectroscopy (LIBS) Using Interval PLS, Minerals. 11 (2021) 1379. https://doi.org/10.3390/min11121379.

[6] R.B. Anderson, O. Forni, A. Cousin, R.C. Wiens, S.M. Clegg, J. Frydenvang, T.S.J. Gabriel, A. Ollila, S. Schröder, O. Beyssac, E. Gibbons, D.S. Vogt, E. Clavé, J.-A. Manrique, C. Legett, P. Pilleri, R.T. Newell, J. Sarrao, S. Maurice, G. Arana, K. Benzerara, P. Bernardi, S. Bernard, B. Bousquet, A.J. Brown, C. Alvarez-Llamas, B. Chide, E. Cloutis, J. Comellas, S. Connell, E. Dehouck, D.M. Delapp, A. Essunfeld, C. Fabre, T. Fouchet, C. Garcia-Florentino, L. García-Gómez, P. Gasda, O. Gasnault, E.M. Hausrath, N.L. Lanza, J. Laserna, J. Lasue, G. Lopez, J.M. Madariaga, L. Mandon, N. Mangold, P.-Y. Meslin, A.E. Nelson, H. Newsom, A.L. Reyes-Newell, S. Robinson, F. Rull, S. Sharma, J.I. Simon, P. Sobron, I.T. Fernandez, A. Udry, D. Venhaus, S.M. McLennan, R.V. Morris, B. Ehlmann, Post-landing major element quantification using SuperCam laser induced breakdown spectroscopy, Spectrochim. Acta Part B At. Spectrosc. 188 (2022) 106347. https://doi.org/10.1016/j.sab.2021.106347.

[7] J.-B. Sirven, B. Bousquet, L. Canioni, L. Sarger, Laser-Induced Breakdown Spectroscopy of Composite Samples: Comparison of Advanced Chemometrics Methods, Anal. Chem. 78 (2006) 1462–1469. https://doi.org/10.1021/ac051721p.

[8] N. Herreyre, A. Cormier, S. Hermelin, C. Oberlin, A. Schmitt, V. Thirion-Merle, A. Borlenghi, D. Prigent, C. Coquidé, A. Valois, C. Dujardin, P. Dugourd, L. Duponchel, C. Comby-Zerbino, V. Motto-Ros, Artificial neural network for high-throughput spectral data processing in LIBS imaging: application to archaeological mortar, J. Anal. At. Spectrom. 38 (2023) 730–741. https://doi.org/10.1039/D2JA00389A.

[9] J. Álvarez, M. Velásquez, A.K. Myakalwar, C. Sandoval, R. Fuentes, R. Castillo, D. Sbarbaro, J. Yáñez, Determination of copper-based mineral species by laser induced breakdown spectroscopy and chemometric methods, J. Anal. At. Spectrom. 34 (2019) 2459–2468. https://doi.org/10.1039/C9JA00271E.

[10] S. Müller, J.A. Meima, D. Rammlmair, Detecting REE-rich areas in heterogeneous drill cores from Storkwitz using LIBS and a combination of k-means clustering and spatial raster analysis, J. Geochem. Explor. 221 (2021) 106697. https://doi.org/10.1016/j.gexplo.2020.106697.

[11] P. Pease, V. Tchakerian, Source provenance of carbonate grains in the Wahiba Sand Sea, Oman, using a new LIBS method, Aeolian Res. 15 (2014) 203–216. https://doi.org/10.1016/j.aeolia.2014.06.001.

[12] P.A. Defnet, M.A. Wise, R.S. Harmon, R.R. Hark, K. Hilferding, Analysis of Garnet by Laser-Induced Breakdown Spectroscopy—Two Practical Applications, Minerals. 11 (2021) 705. https://doi.org/10.3390/min11070705.

[13] E.T.G. Pitman, A Note on Normal Correlation, Biometrika. 31 (1939) 9–12.

[14] Y. Roggo, L. Duponchel, C. Ruckebusch, J.-P. Huvenne, Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data, J. Mol. Struct. 654 (2003) 253–262. https://doi.org/10.1016/S0022-2860(03)00248-5.

[15] W.G. Cochran, G.W. Snedecor, Statistical methods, 6th edition, Iowa State University Press, 1967.

[16] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika. 12 (1947) 153–157. https://doi.org/10.1007/BF02295996.

[17] C. Fabre, N.E. Ourti, J. Mercadier, J. Cardoso-Fernandes, F. Dias, M. Perrotta, F. Koerting, A. Lima, F. Kaestner, N. Koellner, R. Linnen, D. Benn, T. Martins, J. Cauzid, Analyses of Li-Rich Minerals Using Handheld LIBS Tool, Data. 6 (2021) 68. https://doi.org/10.3390/data6060068.

[18] C.J.M. Lawley, A.M. Somers, B.A. Kjarsgaard, Rapid geochemical imaging of rocks and minerals with handheld laser induced breakdown spectroscopy (LIBS), J. Geochem. Explor. 222 (2021) 106694. https://doi.org/10.1016/j.gexplo.2020.106694.

[19] J. Rakovský, P. Čermák, O. Musset, P. Veis, A review of the development of portable laser induced breakdown spectroscopy and its applications, Spectrochim. Acta Part B At. Spectrosc. 101 (2014) 269–287. https://doi.org/10.1016/j.sab.2014.09.015.
[20] K. Rammelkamp, S. Schröder, G. Ortenzi, A. Pisello, K. Stephan, M. Baqué, H.-W. Hübers, O. Forni, F. Sohl, L. Thomsen, V. Unnithan, Field investigation of volcanic deposits on Vulcano, Italy using a handheld laser-induced breakdown spectroscopy instrument, Spectrochim. Acta Part B At. Spectrosc. 177 (2021) 106067. https://doi.org/10.1016/j.sab.2021.106067.
[21] C. Fabre, N.E. Ourti, C. Ballouard, J. Mercadier, J. Cauzid, Handheld LIBS analysis for in situ quantification of Li and detection of the trace elements (Be, Rb and Cs), J. Geochem. Explor. 236 (2022) 106979. https://doi.org/10.1016/j.gexplo.2022.106979.
[22] S. Richiero, C. Sandoval, C. Oberlin, A. Schmitt, J.-C. Lefevre, A. Bensalah-Ledoux, D. Prigent, C. Coquidé, A. Valois, F. Giletti, F. Pelascini, L. Duponchel, P. Dugourd, C. Comby-Zerbino, V. Motto-Ros, Archaeological Mortar Characterization Using Laser-Induced Breakdown Spectroscopy (LIBS) Imaging Microscopy, Appl. Spectrosc. 76 (2022) 978–987. https://doi.org/10.1177/00037028211071141.