Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy

Véronique Bellon-Maurel, Elvira Fernandez-Ahumada, Bernard Palagos, Jean-Michel Roger, Alex McBratney

Near-infrared (NIR) and mid-IR spectroscopy applied to soil compositional analysis started to develop markedly in the 1990s, taking advantage of earlier advances in instrumentation and chemometrics for agricultural products. Today, NIR spectroscopy is envisioned as replacing laboratory analysis in certain applications (e.g., soil-carbon-credit assessment at the farm level). However, accuracy is still unsatisfactory compared with standard laboratory procedures, leading some authors to think that such a challenge will never be met.

This article investigates the critical points to be aware of when accuracy of NIR-based measurements is assessed. First is the decomposition of the standard error of prediction into components of bias and variance, only the latter being reducible by averaging. This decomposition is not used routinely in the soil-science literature. Contrarily, a log-normal distribution of reference values is very often encountered with soil samples [e.g., elemental concentrations (e.g., carbon)] with numerous small or zero values. These very skewed distributions make us take precautions when using inverse regression methods (e.g., principal component regression or partial least squares), which force the predictions towards the centre of the calibration set, leading to negative effects on the standard error prediction – and therefore on prediction accuracy – especially when log-normal distributions are encountered. Such distributions, which are very common for soil components, also make the ratio of performance to deviation a useless, even hazardous, tool, leading to erroneous conclusions.

We propose a new index based on the quartiles of the empirical distribution – ratio of performance to inter-quartile distance – to overcome this problem.

Keywords: Accuracy; Bias; Calibration; Chemometrics; Figure of merit; Near infrared; NIR spectroscopy; Non-Gaussian population; Soil; Uncertainty

Véronique Bellon-Maurel*,
Elvira Fernandez-Ahumada,
Bernard Palagos,
Jean-Michel Roger
1- Montpellier Supagro-Cemagref
UMR ITAP, BP 5095, 34033
MONTPELLIER Cedex 1,
France

Alex McBratney
ACPA, Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney NSW 2006, Australia

*Corresponding author.
Tel.: +61 450 16 84 80;
Fax: +33 4 67 04 63 06;
E-mail: bellonv@supagro.inra.fr

1. Introduction

Near-infrared (NIR) and mid-IR (MIR) spectroscopy are more and more commonly used in soil science for measuring various soil attributes mainly related to chemical composition {e.g., various forms of carbon, N, P, K contents, cation-exchange capacity (CEC) and pH} but also to some extent, related to physical parameters {e.g., texture (clay, sand and silt contents), structure, porosity or bulk density} [1,2]. It is even thought possible to replace conventional soil analysis with NIR/MIR spectroscopy, provided reliability is satisfactory.

Although the first serious attempts to measure soil properties were carried out in the early 1990s [3], NIR/MIR spectroscopy applied to soil science really began to take off much more recently: about half the papers published on NIR/MIR applied to soil science have been published in the past three years. Research started with MIR spectroscopy, because of the legacy of mineralogical studies by MIR, but soil

0165-9936/$ - see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2010.05.006

scientists have preferred to take up NIR spectroscopy, because of its ease of use, portability, and lesser demand for sample preparation. In the following, we therefore focus on examples from NIR spectroscopy, but all the concepts described below are similarly applicable to MIR spectroscopy.

As described by Bellon-Maurel [4], NIR spectroscopy applied to soil science really began in earnest at the turn of the millennium, whereas NIR spectroscopy applied to food and agricultural products boomed in the mid 1980s [5]. Nevertheless, NIR spectroscopy applied to soil science has benefited from the research carried out on these commodities and from subsequent advances in chemometrics, particularly multivariate analytical methods {e.g., partial least squares (PLS), which was made popular by Martens and Naes [6]}. However, blindly applying methodologies developed for agricultural products to soil issues is potentially hazardous and may not lead to optimal use of NIR spectroscopy. Indeed, agricultural products and soils, although sharing some comparable traits (e.g., high optical scattering) are far from having generally similar properties or constraints.

The main difference between agricultural products and soil materials is that biological samples (e.g., agricultural commodities) are constrained by biological genetics. This means that, whatever the location or the conditions of production of a biological commodity, its composition is very stable, so all the major components expected in each commodity are known (there are no unexpected peaks), the concentration of each component ranges between expected maximum and minimum values, the Gaussian distribution applies within this range and some components may be correlated (e.g., for sugar and acidity in ripening fruits, the more sugar the less the acidity). Soils do not match these features (e.g., the distributions of components of interest are generally highly skewed) so methods that are appropriate for biological products should be avoided for soils or used with special care. This has been outlined by Brown et al. [7]: ''Reliable calibrations for materials like wheat grains and forages can be constructed with just a few thousand samples, but these materials are compositionally constrained by plant genetics. Soil composition is, unfortunately, not so constrained, which makes the problem of VNIR-DRS (visible near infrared diffuse reflectance spectroscopy) soil characterization both different and more challenging than that of grain or forage analysis.''

In an attempt to make NIR spectroscopy a routine analytical technique for soil, one has to be careful about avoiding the pitfalls of NIR-based analysis. Such a study has already been proposed by various authors [8,9] for general analytical purposes. But, in the case of soil, new pitfalls and opportunities are specific to the objective of the measurement (e.g., giving an average value over a field, or even over a whole farm, or region) or to the asymmetric distributions of some soil components. It is therefore of primary interest to study how the various parameters classically used for assessing the quality of the analysis (i.e. the prediction of new values and model comparison) would be adequate for the purpose of soil analysis. The parameters generally found in the literature to express the goodness of NIR-calibration models are: the standard error of prediction (SEP), the ratio of performance to deviation (RPD) and sometimes, but not very often, the bias (found in only 25% of the cases in a bibliometric survey we carried out on carbon analysis by NIR/MIR spectroscopy).

The aim of this article is to investigate the critical points to be aware of when accuracy of NIR-based measurements is assessed, with a special focus on soil analysis. The ambition is not to propose ways to overcome the past difficulties of the NIR analysis of soils but to help NIR users in soil science to develop a good metrology for NIR-based routine analysis of soil properties. To do so, we:

(1) explain how the model-performance parameters can, or cannot, be used to express the uncertainty of future predictions in the context of routine soil analysis;
(2) explain how the non-Gaussian distribution of soil properties can bias these indices; and, finally,
(3) suggest how to improve NIR-based prediction of soil components by the good use of statistics.

2. Preliminary assumptions

To avoid extensive discussion about model quality, we assume in the following that a multi-linear model exists between spectral data and the component of interest; this means that the prediction errors show a certain level of homoscedasticity or, in other words, that the average prediction error of replication experiments is independent of **y**. We also suppose that, as in most procedures for least-squares regression, the variables are centered.

3. The issue of bias and final accuracy

3.1. The two components of the standard error of prediction

The SEP or root mean square error of prediction (RMSEP) is the parameter commonly used in the NIR spectroscopy literature to describe the prediction ability of a model. SEP^2 is computed as the sum of squares of the differences between the predicted and the actual values of **y** for a test sample set, which is independent from the calibration value:

$$SEP^2 = \sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{m} \quad (1)$$

where $\hat{y}$ is the predicted value and y the true value, and m the number of observations or samples in the validation set.

SEP therefore appears as an averaged error recorded on the validation-sample set. Authors generally compare the SEP to the expected accuracy (EA) or, more exactly its dual (i.e. the expected error) to decide whether or not a method is acceptable as an alternative analytical method. However, it is worth going deeper into the SEP in order to overcome the issue of accuracy by posing the question ''How can we reduce the SEP?''

The SEP value can be decomposed as follows [10]:

$$SEP^2 = Bias^2 + SEP^2c \quad (2)$$

where:

$$Bias = \sum_{i=1}^{m} \frac{\hat{y}_i}{m} - \sum_{i=1}^{m} \frac{y_i}{m} = \bar{\hat{y}} - \bar{y} \quad (3)$$

and

$$SEP^2c = \sum_{i=1}^{m} \frac{(\hat{y}_i - Bias - y_i)^2}{m} \quad (4)$$

N.B. Another common way of estimating SEP^2c is to use the standard deviation of errors which has (m-1), instead of m, as the denominator; whatever the choice of computation (SEP decomposition or standard deviation), in practice, the results show little to negligible difference.

This means that SEP is made up of two quantities:

- The bias, which is the difference of the mean of the predicted versus the mean of the true **y** values. It is also called the error of means. The bias comes from systematic errors (e.g., due to the instrument, methodology of analysis, and even discrepancies in the reference analysis). Most of these sources of errors can be mitigated (instrument, methodology, reference analysis), but others (e.g., the lack of fit of model) really depend on the robustness of the model with regard to these new conditions.
- A deviation term, SEPc (for ''SEP corrected for bias''), which is the square root of the quadratic sum of the error of the predicted versus the true value, once the predicted value has been corrected for bias. The SEP^2c is also called the residual variance. This error is a random value and has a mean equal to zero, provided there is no slope effect (i.e. that the slope remains equal to 1), which is often the case in soil applications. It accounts for the dispersion around the 1:1 line in the predicted *versus* true value graph, once the bias has been removed (see Fig. 1). SEPc and the bias are independent.

3.2. How to reduce the SEP

The issue of ''how to reduce the SEP'' can therefore be split into two parts, involving reducing the deviation and the bias.

The deviation comes from random errors, so it can be reduced by averaging the measured outputs: if value $\hat{y}_i$ is the average of a k-tuplet, then SEP^2c is reduced by a factor of $1/k$ compared with a single measurement. Making replications is particularly relevant for soil, because the aim is generally to predict the composition of large areas, so multiplying the samples is not a major difficulty. As NIR analysis is fast and easy to run, the cost of a k-tuplet NIR analysis will not be too high (the major cost will almost inevitably come from field-sampling operations).

A high bias means a low trueness of the measurement. Trueness is a metrological term, which means closeness of the average of values obtained by replicate to the true value. The problem is that the bias can be computed only if both predicted values and real values of the samples are known, which is not the case in routine analysis. In routine analysis, the bias is unknown. The bias cannot be reduced by averaging because averaging will retain the systematic error, so the only way to improve the trueness of the measurement is to reduce the sources of discrepancy. This means attempting to build models that are as robust as possible, {i.e., insensitive [e.g., to the origin of the samples, or to external parameters (e.g., particle size, moisture)]}, standardizing procedures as much as possible, in particular reference measurement, and setting up adequate calibration-transfer procedures. Discussion of how to improve model robustness is beyond the scope of this article, but can be found [11–15].

The issue of SEP reduction can be illustrated by an example. Let two models for carbon prediction in soil (say model A and model B) be fitted on the same calibration set and validated on the same validation-sample sets. Let us assume that the SEP is 2%C for these two models. Let us assume that the EA of the NIR carbon analysis is 1%C. We would conclude from the SEP value obtained that these two models have the same predictive performance and that they do not satisfy the expected accuracy. However, if model A has a bias of 1.8%C and model B a bias of 0.2%C, then, in practice, these models have very different prediction abilities. Let us assume that we compute the SEP, hereafter called SEP^*, after averaging k replicates. As shown in Table 1, replication will greatly improve model B, whereas they will have a small effect on model A. With 10 replications, SEP drops from 2%C to 1.82%C and 0.66%C, for models A and B, respectively. In that case, model B fulfils expectations about accuracy.

More generally, if bias >EA, then the model will never satisfy the EA. If bias <EA <SEP, then it is possible to achieve a final SEP^* <EA by making replicate measurements. The number of replications, k, has to be such that:

$$SEP^{*2} < EA^2 \quad \text{with} \quad SEP^{*2} = Bias^2 + \left(\frac{SEP^2 - Bias^2}{k}\right)$$

so the condition on k is that:

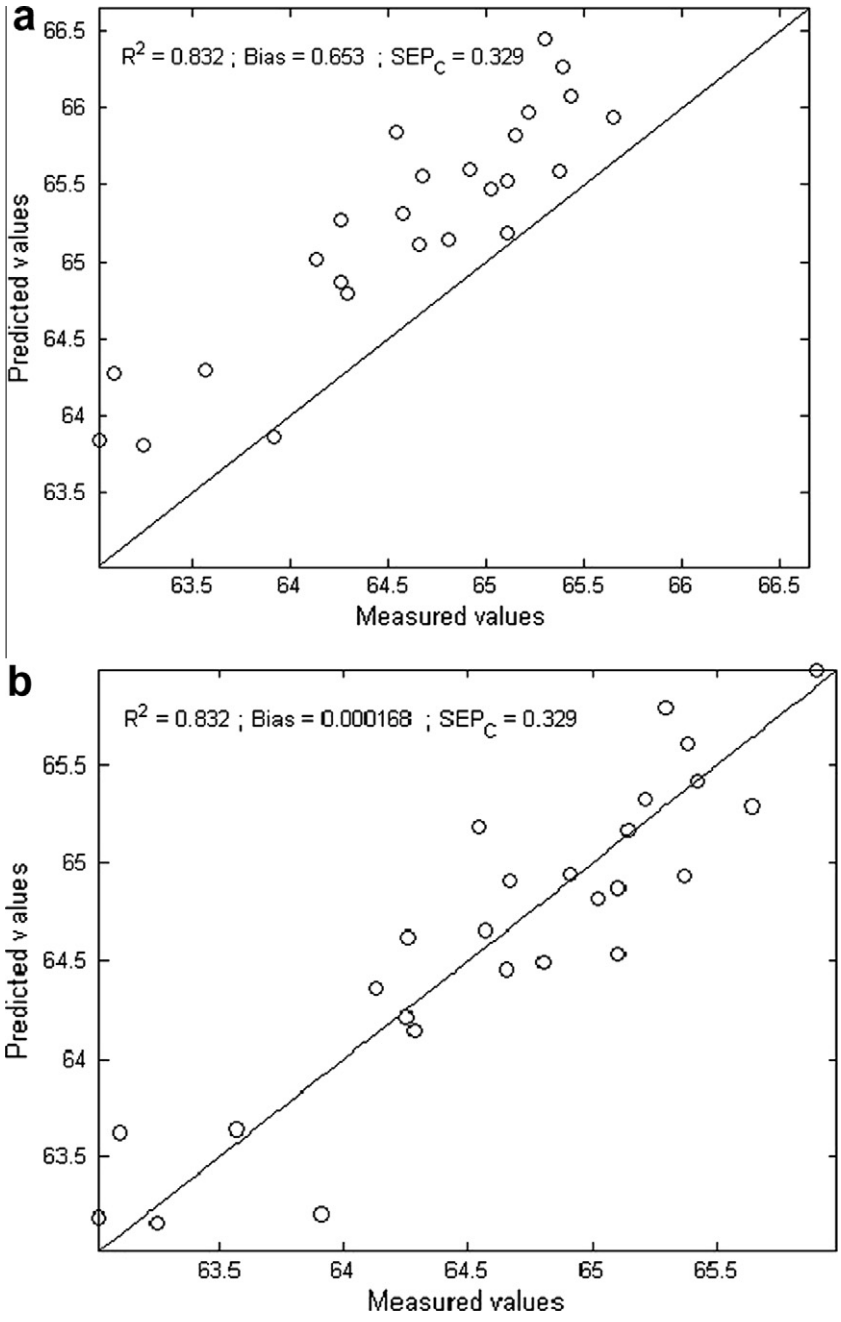

Figure 1. The effect of bias (SEP = 0.73): (a) a bias of 0.65 has been found; and, (b) after removal of the bias, only the dispersion remains (SEP = 0.33).

$$k > \frac{SEP^2 - Bias^2}{EA^2 - Bias^2} \quad (5)$$

However if bias is close to EA, the number of replications may very well be too large to be feasible.

In conclusion, in the context of soil analysis where replications can be carried out, if the SEP is not low enough to match the EA, it is of prime interest also to know the bias really to assess the future utility of the technique and, if needed, the number of replications necessary to achieve the EA.

4. Uncertainties of the predicted values, and relations with SEP and the population distribution

With the aim of developing a new analytical technique, one must be able to deliver predicted value $\hat{y}$ with an

Table 1. The effect of averaging the predicted values on the SEP for two models of same SEP but different bias

Model	1 replicate					10 replicates				
	SEP^2	SEP	$Bias^2$	Bias	$SEPc^2$	SEP^{*2}	SEP	$Bias^2$	Bias	$SEPc^2$
A	4	2	3.24	1.8	0.76	3.316	1.82	3.24	1.8	0.076
B	4	2	0.04	0.2	3.96	0.436	0.66	0.04	0.2	0.396

associated uncertainty limit (given a confidence level). To compute it, it is worth understanding the relationships between the standard error of calibration, the confidence intervals of a newly predicted value, $\hat{y}$, the standard error of prediction and the distribution of the validation set.

As a first step, SEP can be used to give the accuracy of the model. Its main advantage is that, provided it has been computed on a really independent sample set, it will account for the effects of differences in acquisition conditions (e.g., samples of different origin, different day, and maybe different operator). But several drawbacks pertain to SEP:

(1) As shown above, it can contain a bias, which is unknown; in the following, we shall consider that the bias is null.
(2) It does not directly provide the uncertainty of each individual prediction; the latter changes along the range of the measure and Y; as stated by Olivieri et al. [16]: ''Although RMSEP is a correct summary statistic for guiding the model-selection process (e.g., optimal data pretreatment), it cannot lead to prediction intervals with good coverage probabilities''.
(3) As we will see below, it is very dependent on the statistical distribution of the values of the validation-sample set.

It is therefore necessary to come back to the evaluation of the uncertainty for each new prediction to understand the relationships between the individual uncertainties, the SEP and the distribution of the sample set.

4.1. The expression of the individual prediction uncertainties

As Zhang and Garcia-Munoz [17] remind us: ''the general procedure to estimate the uncertainty for y-response $\hat{y}$ of a new individual observation with predictor row-vector x' consists of 2 steps''.

The first step deals with estimating the standard deviation of the prediction error, σ_p. The second step is to compute the confidence interval of a newly predicted sample, assuming that the estimated error follows t-statistics:

$$\textit{Confidence Intervals} \quad CI = Y_p \pm t_{1-\alpha/2,df}\sigma_p \tag{6}$$

where: t is Student t distribution, α is the significance level of the interval and df degrees of freedom; in MLR, $df = N - p - 1$, where p is the number of variables used for the calibration; in PLS, as latent variables are computed by using **y** and X, the true df is unknown and lies somewhere between the number of factors and the number of wavelengths.

The first step is therefore to obtain σ_p, with $\sigma_p{}^2 = \mathrm{Var}(\hat{y})$.

The expressions of $\mathrm{Var}(\hat{y})$ are diverse, depending on whether we consider a multiple linear regression (MLR) model or a multivariate model based on projections on latent variables (PCR, PLS). For MLR analysis, Zhang and Garcia-Munoz [17] give the following expression of $\mathrm{Var}(\hat{y}_i)$:

$$Var(\hat{y}_i) = \frac{s^2}{N} + h_i s^2 + s^2 = s^2\left(\frac{1}{N} + h_i + 1\right) \tag{7}$$

where: N is the number of calibration samples, h_i, the leverage of sample i [see Equation (8)], s, the standard deviation of the residuals, e_i, which is approximated by SEC.Note that this equation is proposed under the simplifying assumption that Var(X) = 0.

The leverage value of sample i is the distance of sample i to the centre of the calibration set, with regard to the basis in use (i.e. the independent variables in MLR):

$$x_i^{*\prime} C x_i^* = h_i \tag{8}$$

where: x_i^* is the centered spectrum of new sample i, and $C = \mathbf{(X'X)}^{-1}$ or $\mathbf{(R'R)}^{-1}$ with X the matrix of centered spectra and R the matrix of scores for, respectively. MLR or PLS/PCR analysis.

The importance of Equation (7) is therefore that it applies to both MLR and latent projection methods (e.g., PLS and PCR).

Other relationships have been proposed in the literature [18,19] for expressing $\mathrm{Var}(\hat{y})$. They vary with regard to the assumptions that have been made (e.g., about the variance of X, the S/N ratio, and the variance of the reference measurement). It is beyond the scope of this article to discuss all of these in detail. They all express $\mathrm{Var}(\hat{y})$ by the leverage h_i of the new object (i.e. its distance to the centre of the calibration sample set) as a coefficient of the SEC, in addition to other terms.

The uncertainty of new prediction $\hat{y}_i$ therefore follows Student t distribution and the confidence intervals are:

$$CI = \hat{y}_i \pm t_{1-\alpha/2,df} SEC \sqrt{\frac{1}{N} + h_i + 1} \quad (9)$$

4.2. How the validation-population distribution can affect the SEP

The uncertainty of each new sample prediction therefore partly depends on the leverage, h, i.e.:

(i) on the distance of the new sample spectrum to the centre of the calibration-sample set in the variable space (in the case of MLR) or the latent variable space (for a PLS/PCR); and,

(ii) on the size and dispersion of this calibration space.

Of course, the larger the calibration-sample set and the smaller the number of independent variables, the lower the leverage value and therefore the lower the influence of the leverage. This leads us to make two important remarks with regard to computation of uncertainty.

4.2.1. Remark 1. The smallest variance for $\hat{y}$ is obtained for an object i at the centre of the calibration space. In that case, the leverage is minimal (i.e. can be approximated by 0) and, based on Equation (9), the uncertainty becomes:

$$\pm t_{1-\alpha/2,df} SEC \sqrt{\frac{1}{N} + 1}$$

If the number of calibration samples is high, then $\frac{1}{N} << 1$, and we can approximate t by a normal distribution, i.e.:

$$t_{1-\alpha/2,df} \rightarrow N(0, 1-\alpha/2)$$

i.e. t can be approximated by 2 (precisely 1.96), if $\alpha = 5\%$. This means that, in the best case (i.e., no bias, new sample close to the centre of the calibration set, large calibration set) and in this case only, the first approximation of the confidence interval is $\pm 2SEC$.

Consequently, for the opposite conditions (i.e. with samples with high leverage), the uncertainty increases. Let us consider a leverage sample that is not an outlier [i.e. which fits the regression line (e.g., a sample showing simultaneously extreme values for X and for **y**)]. The confidence interval of the prediction of such a sample will be large due to high leverage: in Equation (9), h_i becomes no longer negligible, so the confidence interval becomes $\pm t_{1-\alpha/2,df} SEC \sqrt{1+h_i}$. This comes because, in PCR or PLS regressions, also called inverse regressions: ''predictions are biased towards the mean of the distribution of the reference values in the training (i.e. calibration) set'' [20]. This is particularly a problem for low-concentration samples because the relative error attached to the prediction (i.e. error e_i divided by the true value of y_i may be huge).

4.2.2. Remark 2. The second remark comes from the fact that leverage h_i not only depends on new sample i but also on the size and the dispersion of the calibration set. If the size of the calibration set is small or if the dispersion is small, h_i increases. This means that, in the calibration phase, one should take as many samples as possible at the borders of the sample space, and avoid taking too many close to the centre to reduce the leverage. The choice of the samples has already been extensively discussed [5]. The authors advocated not taking calibration samples ''as they went'', because biological samples would follow a Gaussian distribution; this would automatically increase h_i. But this advice should be followed only in the case where calibration samples can be chosen by the operator. If the calibration samples are already available, they must all be used, even if they show a Gaussian distribution. Achieving a Gaussian distribution by throwing away samples at the centre would be an error: it does not improve the leverage situation for future samples, because it reduces the number of calibration samples.

The conclusions of this analysis are:

- The uncertainty of the predicted samples is not constant along the whole concentration range: higher and lower concentration samples will have increased uncertainties due to higher leverage towards the calibration set. This can be particularly problematic for low-concentration samples (huge relative error).
- SEP is the root mean square of prediction errors e_i for the m samples to be predicted. As described in Equation (8), the error e_i of a new prediction follows a Student t distribution with zero mean and variance given by:

$$SEC^2(1 + h_i + \frac{1}{N})$$

Consequently, SEP also depends on the leverage of the new samples with regard to the calibration-sample set, so SEP not only reflects the robustness of the model but is also influenced by the validation-set distribution. A validation set having a normal population centered on the calibration-set average will generate a smaller SEP than a uniform distribution or, even worse, a lognormal one, even if centered on the calibration-set average. Once again, this is because PLS optimizes predictions for a validation set with a normal distribution with mean and variance equal to the training set. As far as soil is concerned, this can be a main issue because many chemical and physical attributes of soil show lognormal distributions with numerous samples having negligible concentrations.

To cope with this issue of very non-normal distributions, Fearn et al. [20] recently introduced a very interesting alternative way to carrying out predictions for such populations. The method is based on a Bayesian approach, in which the prior distribution of the samples to be predicted is explicitly used. Using this approach on a very non-normal distribution (a bimodal one) of percentage of wheat in feedstuff samples, they created a model that made SEP decrease from 5.33% to 0.98%.

As a result, whereas SEP appears a good index to compare calibration models validated using the same validation-sample set, its use is more questionable when models fit using inverse regression have to be compared based on different calibration/validation-sample sets. This is particularly sensitive in the case of soils, because population distributions can vary greatly and can be very non-normal (lognormal distributions). A study is being carried out to evaluate the feasibility of Fearn's new approach for soil samples.

5. Is RPD relevant in NIR analysis of soil?

5.1. The use of RPD in NIR spectroscopy

Another concern with comparing the SEP values computed on different validation populations is that the SEP value generally increases when the measurement range of this parameter – or the mean of this range – increases. This issue is well known and has been addressed so far by standardizing the SEP to remove any range effect. The standardization deals with building up a ratio of the SEP and of any statistical index representing the population. The most popular indices are the coefficient of variation (i.e. the ratio of the SEP to the mean of the validation population) (CV% = SEP/Mean), and the RPD, the ratio of performance to deviation (i.e. the ratio of the SD to the SEP) (RPD = SD/SEP). RPD is the one in most common use.

RPD has been used for several years by NIR scientists working on agricultural products [5] and has been widely appropriated by soil-science researchers since the paper by Chang et al. [21]. Several authors refer to Chang et al. [21], in which three quality categories were defined to account for the model reliability:

(1) excellent models, with RPD >2;
(2) fair models, with 1.4 < RPD < 2; and,
(3) non-reliable models, with RPD <1.4.

However, no statistical basis was used to determine these thresholds, and other researchers [5] gave quite different (i.e. much higher) thresholds.

Reeves III and Smith [22] disagreed with these fixed thresholds. According to them, considering that ''many researchers found calibrations to be useful with RPD values considerably lower than the proposed standards, it is up to the reader to evaluate all the statistic provided and decide if similar calibrations would be useful for their needs [...]''.

We concur and, especially in soil science, RPD, which has been developed for biological samples showing normal distributions, may be inappropriate.

First, we have seen in the previous section that SEP, when computed using inverse regression (PCR, PLS) on validation sets with very non-normal distributions – especially lognormal ones – can be problematic. We next show how the use of SD for standardization is also questionable, for the same reason (i.e. the lognormal distribution of the validation set).

5.2. The use of SD for standardizing SEP obtained on soil data

The standard deviation of soil population is not very helpful because soil-sample sets may present a highly skewed distribution, with many low values. Let us take an example with real data given by the US Geol. Surv. Open-File Rep. to which the work of Reeves and Smith refers [22]. Table 2 shows some of the statistical parameters provided by Reeves and Smith for organic carbon in the samples they used for their NIR studies. Because the selection of these samples is not explicitly described by the authors, we approximated the data by a lognormal distribution with meanlog = 0.0234 and SDlog = 1.03. Fig. 2 shows this as the example of real data. For simulated data based on this distribution, a mean of 1.85%C and a standard deviation of 2.7%C were obtained. As the reported SEP for the NIR analysis is 2.7%C, the RPD is 2.7/1.94 = 1.39.

When one looks at RPD as a performance index, a calibration of the same RPD (e.g., same SD and same SEP) but built up on a normally distributed sample set would have been ranked the same as the Reeves and Smith example.

However, due to the difference of distributions, the SD has absolutely not the same interpretation in terms of the range of values. In a normal population, a 2.SD interval around the mean includes 66% of the population (i.e. 66% of the population is located at +/− 1 SD of the mean); however, it represents 93% of the population spread in the case of a lognormal distribution. That means that it accounts for a much larger range. The

Table 2. Statistics for organic carbon based on real data given by [22] (SEC and SEP have been carried out on independent test samples), on simulated data with a lognormal law to match the distribution LN (0.0243;1.03) and on simulated data following a normal law with the same SD as the real data, but with a mean of 8.1 to ensure that 99.9% of the population is over zero [i.e. N (8.1; 2.7). RPIQ = (Q3 − Q1)/SEP with SEP = 1.94%]

Data model	Min	Max	Mea	Med.	Q1	Q3	SD	Skew	Kurtosis	RPIQ
Real data	0.04	34.2	1.8				2.6	5.56	53.55	
LN Sim (0.024; 1.03)	0.05	34.17	1.85	1.11	0.57	2.05	2.7	6.16	58.66	0.76
N (8.1; 2.7)			8.1	8.1	6.3	9.9	2.7			1.85

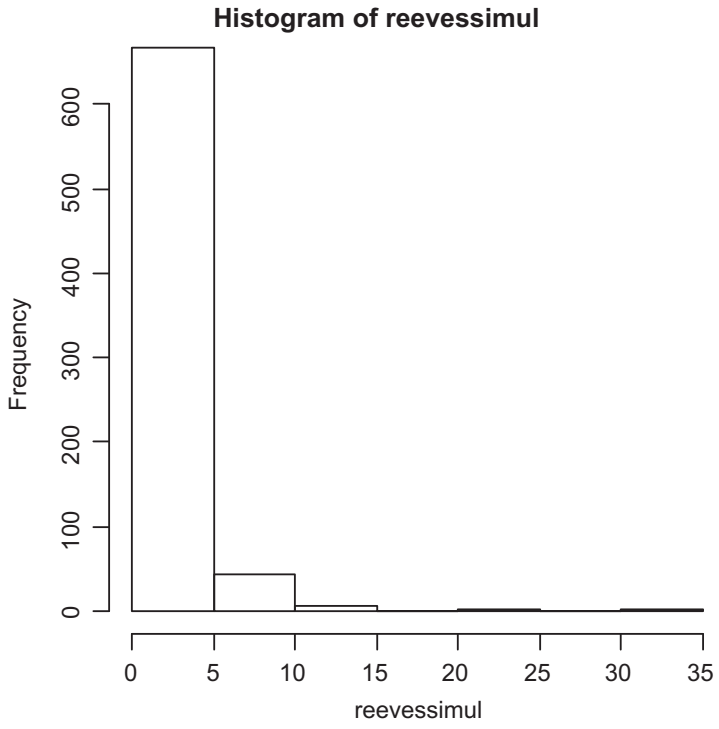

Figure 2. Simulated data for organic-carbon content of soil. The simulated data are built using data from [1].

RPD standardization for a normal and a lognormal distribution are therefore not directly comparable.

5.3. Proposal for a more robust index

Instead of using RPD = SD/SEP, we propose using a new index, based on quartiles, which better represents the spread of the population. The quartiles are milestones in the population range:

- Q1 is the value below which we can find 25% of the samples;
- Q3 is the value below which we find 75% of the samples; and,
- Q2, commonly called the median, is the value under which 50% of samples are found.

The quartiles are therefore useful to determine equivalent ranges of population spread. For example, interquartile distance IQ (= Q3 – Q1) gives the range that accounts for 50% of the population around the median. A new index can be devised using IQ, instead of SD, as the numerator. Let us call it RPIQ (ratio of performance to IQ).

In our case, with the lognormal distribution, the IQ is equal to 1.48%C and the corresponding RPIQ is then 1.48/1.94 = 0.76. A normal population of the same RPD would have an IQ of 3.6%, so RPIQ would be 1.89. RPIQ is therefore more than double for the normal distribution, whereas RPD would have been the same. This means that the RPD value of a lognormal distribution gives an artificially good performance compared with the lognormal distribution. Other simulations made up with data found in the literature [23] show the same trend (i.e. a more-than-a-doubling of RPIQ for a normal population with respect to a lognormal population of the same RPD).

As a conclusion, especially for soil-sample sets, which often show a skewed distribution, the RPD is not a good way for standardizing the SEP with respect to population spread. It does not correctly represent the spread of the population, because the assumptions on normal distributions are generally not fulfilled as they are with biological samples. The RPIQ index, in which SD is replaced by IQ (= Q3 – Q1), accounts much better for the spread of the population.

6. General conclusion

The purpose of this article is to study the metrological issues related to the assessment of soil-sequestrated carbon content. The originality of this NIR-based measurement with regard to the classical measurement of compounds in biological products is because:

- carbon concentrations in soils show highly skewed distributions (lognormal distributions);
- the target of soil-carbon assessment is to deliver a value for a whole field or even a whole farm;
- the issue of this global assessment is definitely accuracy (i.e. closeness of the estimated and true concentration values) and low-cost but, if necessary, replications are allowed.

The main focus has been on the SEP and on the RPD, which are the main indices used to qualify the accuracy and the quality of a NIR-based measurement method.

The main conclusions and advice from our analysis are:

- bias is often totally neglected, whereas it is a major component of SEP, based on SEP decomposition theorem ($SEP^2 = bias^2 + SEP^2c$); the bias is the part of the SEP, which is irreducible by averaging, whereas SEPc may be reduced by averaging; so, in order to know whether an NIR-based measurement of sequestrated carbon can reach the EA, it is absolutely necessary to know the bias; replications can help to reduce SEPc and therefore approach SEP to the adequate EA, provided that the bias is lower than the EA; thorough attention has to be paid to reducing bias and to mitigating bias source;
- whatever the calibration method used (but this is enhanced with inverse regression methods), the SEPc (= SEP, when bias is null) depends not only on the quality of the model but also on the distribution of the validation sample, because inverse regression tends to regress towards the mean of the reference values of the calibration sets; so, if validation sets are very non-normal, the SEP will be negatively affected (i.e. worsened) with regards to a normal distribution of validation samples; therefore comparing models that used different validation sets is hazardous, especially if the validation sets have very diverse distributions;
- because the SEP depends on the range of the reference values of the validation set, RPD = SD/SEP has been proposed as a standardized ratio of quality. Although some authors have begun to criticize this parameter and the quality thresholds associated to it to define the excellent, good, average, ''forget-it'' quality of models, it is very widely spread in NIR spectroscopy applied to soils; however, for the same reason as log-normal distributions of reference values, its use is somehow irrelevant, because SD does not describe correctly the spread of the population in skewed populations; we therefore propose a new index RPIQ = (Q3 − Q1)/SEP to represent the population spread better, regardless of the distribution.

The main goal of this methodological study was to make soil scientists fully aware of the limits and the critical points of classical indicators (e.g., SEP and RPD) in order to overcome current limitations on soil characterization based on NIR measurements. If we were successful, it would make obsolete the following view of Brown et al. [7]: ''Precision and accuracy are elusive goals in soil characterization and yet not often quantified – for both NIR and quantitative methods''.

The RPIQ index would then contribute to paving the way to a reliable, low-cost assessment of sequestrated-carbon content in soil and a fairer trade in carbon credits.

Acknowledgments

This article was written as part of a travelling scholarship supported by the European Commission (IRSES program, IRSES Project Nr. 235108) and the Languedoc Roussillon Council (Regional Platform GEPETOS – ECO-TECH-LR). The authors would like to thank Tom Fearn for his valuable advice.

References

[1] R.A. Viscarra-Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik, J.O. Skjemstad, Geoderma 131 (2006) 59.
[2] L. Cécillon, B.G. Barthès, C. Gomez, D. Ertlen, V. Genot, M. Hedde, A. Stevens, J.J. Brun, Eur. J. Soil Sci. 60 (2009) 770.
[3] K.A. Sudduth, J.W. Hummel, Trans. Am. Soc. Agric. Eng. 34 (1991) 1900.
[4] V. Bellon-Maurel, Pedometron 28 (2009) 27.
[5] P. Williams, K. Norris, Near-Infrared Technology in the Agricultural and Food Industries, Am. Assoc. Cereal Chem, St. Paul, MN, USA, 1987.
[6] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, USA, 1989.
[7] D.J. Brown, K.D. Shepherd, M.G. Walsh, M. Dewayne Mays, T.G. Reinsch, Geoderma 132 (2006) 273.
[8] H.A. Martens, P. Dardenne, Chemometr. Intell. Lab. Syst. 44 (1998) 99.
[9] G.W. Small, Trends Anal. Chem. 25 (2006) 1057.
[10] A. Davies, T. Fearn, Spectrosc. Eur. 18 (2006) 31.
[11] J. Trygg, S. Wold, J. Chemometr. 16 (2002) 119.
[12] J.M. Roger, F. Chauchard, V. Bellon-Maurel, Chemometr. Intell. Lab. Syst. 66 (2003) 191.
[13] F. Chauchard, J.M. Roger, V. Bellon-Maurel, J. NIR Spectrosc. 12 (2004) 199.
[14] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, Trends Anal. Chem. 24 (2005) 437.
[15] Y. Zhu, T. Fearn, D. Samuel, A. Dhar, O. Hameed, S.G. Bown, L.B. Lovat, J. Chemometr. 22 (2008) 130.
[16] A. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Pure Appl. Chem. 78 (2006) 633.
[17] L. Zhang, S. Garcia-Munoz, Chemometr. Intell. Lab. Syst. 97 (2009) 152.
[18] S. De Vries, C.J.F. Ter Braak, Chemometr. Intell. Lab. Syst. 30 (1995) 239.
[19] N.M. Faber, Chemometr. Intell. Lab. Syst. 52 (2000) 123.
[20] T. Fearn, D. Perez-Marin, A. Garrido-Varo, J.E. Guerrero-Ginel, J. NIR Spectrosc. 18 (2010) 27.
[21] C.W. Chang, D.A. Laird, M.J. Mausbach, C.R. Hurburgh Jr., Soil Sci. Soc. Am. J. 65 (2001) 480.
[22] J.B. Reeves III, D.B. Smith, Appl. Geochem. 24 (2009) 1472.
[23] K.S. Lee, D.H. Lee, I.K. Junk, S.O. Chung, K.A. Sudduth, J. Biosyst. Eng. 33 (2008) 260.