Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset

Journal: European Journal of Soil Science

Manuscript ID: EJSS-060-13.R1

Manuscript Type: Original Manuscript

Date Submitted by the Author: 24-Aug-2013

Complete List of Authors: Guerrero, César; Universidad Miguel Hernandez de Elche, Agroquímica y Medio Ambiente
Stenberg, Bo; Swedish University of Agricultural Sciences, Soil and Environment
Wetterlind, Johanna; SLU, Department of Soil and Environment
Viscarra Rossel, Raphael; CSIRO, Land and Water
Maestre, Fernando; Universidad Rey Juan Carlos, Departamento de Biología y Geología
Mouazen, Abdul Mounem; National Soil Science Institute, Natural Resources Department
Zornoza, Raul; Universidad Politécnica de Cartagena, Departamento de Ciencia y Tecnología Agraria
Ruiz-Sinoga, Jose Damian; Universidad de Málaga, Departamento de Geografía
Kuang, Boyan; Cranfield University, Natural Resources Department

Keywords: SOC assessment, soil sensing, near infrared spectroscopy, spiking, extra-weighting
Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset

C. Guerrero, B. Stenberg, J. Wetterlind, R.A. Viscarra Rossel, F.T. Maestre, A.M. Mouazen, R. Zornoza, J.D. Ruiz-Sinoga & B. Kuang

Departamento de Agroquímica y Medio Ambiente, Universidad Miguel Hernández de Elche, E-03202, Spain, Department of Soil and Environment, SLU, Skara, Sweden, CSIRO Land and Water, Canberra, Australia, Departamento de Biología y Geología, Universidad Rey Juan Carlos, Spain, Department of Environmental Science and Technology, Cranfield University, UK, Departamento de Ciencia y Tecnología Agraria, Universidad Politécnica de Cartagena, Spain, and Departamento de Geografía, Universidad de Málaga, Spain.

Correspondence: César Guerrero. E-mail: cesar.guerrero@umh.es

Running title: Spiking and extra-weighting to improve soil organic carbon predictions with NIR
Summary

Spiking is a useful approach to improve the accuracy of regional or national spectroscopic calibrations when they are used to predict at local scales. To do this, a small subset of local samples (spiking subset) is added to recalibrate the regional or national calibration. If the spiking subset is small in comparison with the size of the initial calibration set, then the spiking subset could have little noticeable effect and only a small improvement can be expected. For these reasons, we hypothesised that the accuracy of the spiked calibrations can be improved when the statistical relevance of the spiking subset is given extra-weight. We also hypothesised that the spiking subset selection and the initial calibration size were relevant, and could affect the accuracy of the recalibrated models. To test these hypotheses, we evaluated different strategies to select the best spiking subset, with and without extra-weighting, to spike three initial calibrations of different sizes. These calibrations were used to predict the soil organic carbon (SOC) content in samples from four target sites. Our results confirmed that spiking improved the prediction accuracy of the initial calibrations. We observed differences in accuracy depending on the spiking subset used. The best results were obtained when the spiking subset contained local samples evenly distributed in the spectral space, regardless of the initial calibration’s characteristics. The accuracy was significantly improved when the spiking subset was extra-weighted. For medium- and large-sized initial calibrations, the improvement due to extra-weighting was larger than that caused by the increase in spiking subset size. This result is interesting because extra-weighting the spiking subset is an inexpensive task. Similar accuracies were obtained using small- and large-sized initial calibrations, suggesting that incipient spectral libraries could be useful if the spiking subset is properly selected and extra-weighted. When small-sized spiking subsets were used, the predictions results were
more accurate than those obtained with ‘geographically local’ models. Overall, our results indicate that we can minimise the efforts needed to effectively use near-infrared (NIR) spectroscopy for SOC assessment at local scales.

**Keywords:** SOC assessment, soil sensing, near infrared spectroscopy, spiking, extra-weighting.

**Introduction**

Using near-infrared (NIR) spectroscopy to estimate soil properties is rapid, non-destructive and relatively inexpensive compared to conventional laboratory analyses, particularly when processing many samples. For NIR spectra to be quantitatively useful, we need to develop and use a soil spectral database or library to derive spectroscopic models (calibrations) that relate the spectra to analytical data, e.g. soil organic carbon (SOC). When assessing soil properties at a local scale, we can develop site-specific or ‘geographically local’ calibrations (Wetterlind et al., 2010) that are generally very accurate because smaller areas tend to be less variable in terms of the dependent variable (Stenberg et al., 2010), and the samples used to develop the calibration and those used for prediction share similar characteristics, such as mineralogy and organic matter quality (Reeves et al., 1999; Janik et al., 2007; Guerrero et al., 2010; Wetterlind et al., 2010). A disadvantage of these models is that they are only valid for the local area, which could be an expensive strategy when evaluating multiple areas. Another option is to use regional, national or global calibrations, but they should represent the variability of the soils being analysed. This has caused a trend to develop larger-scale calibrations with a very large number of samples to ensure that the local samples fall...
within the model’s domain (Shepherd & Walsh, 2002; Brown et al., 2006; Viscarra Rossel, 2009; Grinand et al., 2012; Viscarra Rossel & Webster, 2012), although this cannot be guaranteed because soils have such variable characteristics, even at a regional scale. Furthermore, a set of samples comprising a large-scale calibration should be considered heterogeneous, but the local samples could be considered as a homogeneous set that is located in a small area of the overall calibration domain. This could be the reason for inaccurate (biased) results observed by some authors when using regional and national calibrations to make predictions at local scales (Brown et al., 2005; Brown, 2007; Janik et al., 2007; Christy, 2008; Sankey et al., 2008; Guerrero et al., 2010; Stenberg et al., 2010; Wetterlind & Stenberg, 2010), even when the local samples fall within the model domain and are not recognised as outliers. This could also explain why better results are obtained with local (spectrum-specific) models (Genot et al., 2011; Gogé et al., 2012), where a subset of library samples that are similar to the unknown sample is used to construct the calibration (Pérez-Marín et al., 2007). However, local methods are expensive because a large spectral library is needed to find sufficient similar samples for the calibrations.

Spiking is an alternative method proposed to improve the accuracy of regional or national calibrations for use at local scales (Viscarra Rossel et al., 2009; Guerrero et al., 2010; Stenberg et al., 2010; Wetterlind & Stenberg, 2010; Kuang & Mouazen, 2013). Spiking—sometimes referred to as ‘augmentation’ (Brown et al., 2006; Brown, 2007; Sankey et al., 2008) and other names—involves three main steps (Janik et al., 2007). First, analyse a few samples from the target site in the laboratory using the reference method; then add these samples to the initial calibration matrix; and then recalibrate the model. This procedure usually increases the accuracy of the predictions in the rest of the samples from the target site (Brown et al., 2005; Sankey et al., 2008; Wetterlind &
Stenberg, 2010). The higher the number of local samples in the spiking subset, the higher the accuracy in the prediction set (Brown, 2007; Guerrero et al., 2010), but a large spiking subset decreases the advantages of NIR spectroscopy as a quick and low-cost analytical method. To increase the relative proportion of the spiking subset, Guerrero et al. (2010) suggested decreasing the number of samples in the initial calibration set because they obtained higher accuracies when small-sized calibrations were spiked, where the spiking subset had a larger influence. However, the selection of a small number of calibration samples can reduce the amount of important information for modelling, and lead to less robust calibrations. For this reason, we proposed an alternative approach to increase the relevance of the spiking subset in the NIR calibrations. The approach is to increase the statistical weight of the spiking subset by adding several copies of the subset to the calibration matrix. These extra-weighted samples are more important than other samples used to form the statistical model (Capron et al. 2005; Stork & Kowalski 1999), which forces the calibration to better fit the extra-weighted samples. If these samples were similar to the overall prediction set, the model should provide more accurate predictions. We also evaluated different strategies to select the best spiking subset. Since each local sample is different to the others, we hypothesised that the selection of a spiking subset would influence the accuracy of the spiked models, and the selection would be more influential if fewer samples were used for spiking.

The spiking approach tries to gain benefits from a previously developed or initial large-scale calibration set. It is reasonable to assume that results obtained could be affected by the characteristics of the initial calibration, as some authors observed (Guerrero et al., 2010; Wetterlind & Stenberg, 2010). For this reason, we included different initial calibrations in this study and evaluated their influence on the spiking
process. Our first objective was to evaluate how local samples should be selected as a spiking subset for optimal spiking. To do this, we compared thirteen different strategies to select the samples for the spiking subset. Our second objective was to evaluate whether an extra-weighted spiking subset increased the prediction accuracy. In addition, we compared geographically local models that used three different sized spiking subsets. We selected SOC as the soil property for prediction, and we used the coefficient of determination ($R^2$), root mean square error of prediction (RMSEP), standard error of prediction (SEP) and ratio of performance to deviance (RPD) to evaluate the prediction performance for four different target sites.

2. Material and methods

2.1. National samples and initial calibrations

A national soil library (n = 2836) of soils from different sites across Spain (predominantly southeastern Spain) was randomly split into three subsets. These subsets were used to create three initial calibrations of different sizes, representing three different stages or efforts to develop the spectral library: small (IC#1; n = 192), medium (IC#2; n = 365) and large (IC#3; n = 2279). The soils in the soil library were collected under forest and agricultural land uses. Most of these soils developed over sedimentary (mostly calcareous) lithologies. The soil samples were air-dried and sieved (< 2 mm), and the NIR spectra (12 000–3800 cm$^{-1}$) were obtained by FT–NIR diffuse reflectance spectroscopy (MPA, Bruker Optik GmbH, Germany). The scale of the spectra was transformed to nanometers (830–2630 nm), and re-sampled to 1 nm resolution. The SOC concentration (%) was determined using the Walkley & Black (1934) method. The different initial calibrations, relating the SOC to the NIR spectra, were constructed
using partial least squares (PLS) regression (PLS-1 algorithm) (see section 2.6 for details). Key characteristics of the initial calibrations are shown in Table 1.

2.2. Target sites

We selected four independent target sites from four regions with spectral characteristics that differed from each other and from those observed in the initial calibrations (Figure 1; Appendix 1). Each target site is a relatively small area of dense sampling, from several hectares to a few square kilometres in size. A different number of local samples were collected at each target site (Table 2). One site was located in Sweden (TS1), two in Spain (TS2, TS3) and one in the United Kingdom (TS4). As with the initial calibration samples, the soil samples from the target sites were air-dried and sieved (< 2 mm), and the NIR spectra and SOC content were obtained. Most of the spectra were collected using a FT–NIR (MPA, Bruker Optik GmbH, Germany), except the TS1 samples, which were scanned using a vis–NIR (ASD FieldSpec Pro Fr, USA). The scale of the FT–NIR spectra was transformed from cm\(^{-1}\) to nanometers, and resampled to 1 nm. For details about FT–NIR and vis–NIR scanning, see Guerrero et al. (2010) and Wetterlind & Stenberg (2010), respectively.

2.3. Calibration types

Different types of calibrations relating SOC and NIR spectra were obtained using PLS as a regression method (see section 2.6), and were used to predict the SOC contents in the target site samples.

Initial calibrations: three different-sized initial calibrations (IC#1, IC#2 and IC#3, described in section 2.1) that did not contain any samples from the target sites; referred to as unspiked initial calibrations (Figure 2a; section 2.6).
Spiked calibrations: the three initial calibrations modified by adding a spiking subset (n = 8) (Figure 2b). We used 13 different spiking subsets to spike each of the initial calibrations (see section 2.4). In each initial calibration, we obtained 13 subtypes of spiked calibrations, and we repeated this procedure for each of the four target sites.

Spiked calibrations with extra-weighting: in each of the different spiked calibrations, the spiking subset was extra-weighted. To do this, we added 24 copies of each spiking subset sample to the calibration set (Figure 2c), and then recalibrated the model (see section 2.6). Each of the eight spiking subset samples appears 25 times in the calibration matrix, becoming 24 times more influential than the soil library samples because we have modified their leverage (Stork & Kowalski, 1999). We selected 24 copies because the leverage of the target site samples followed an asymptotic pattern after the addition of 15–20 copies (data not shown).

2.4. Strategies to select the spiking subset from the target site samples

For each target site, we used 13 strategies to select the different types of spiking subsets. We hypothesised that each strategy had different advantages. The strategies were designed and grouped on the basis of (i) the SOC values of target site samples, (ii) the spectral characteristics of the target site samples and (iii) the spectral relationships between the initial calibrations and the target site samples using the Mahalanobis distance values. The first group of five strategies was designed on the basis of the SOC content of target site samples. These strategies have a strictly theoretical value for interpreting some results because the SOC contents of the target site samples would be unknown in a real scenario, and thus these strategies would not be useful in practice.

Strategy 1 (OC low): select eight target site samples with the lowest SOC values (left tail of SOC histogram). Samples with low SOC contents will show more clearly the
spectral features of the inorganic constituents, which are the most important factors impeding the use of a calibration from one site to another. Moreover, these samples could be useful to correct the bias in target site samples with low SOC contents.

Strategy 2 (OC high): select eight target site samples with the highest SOC values (right tail of SOC histogram). These samples mask the inorganic spectral features, and clearly show the SOC spectral features in the local samples. Moreover, these samples can be useful to correct bias in target site samples with high SOC contents.

Strategy 3 (OC tails): select four samples with the lowest SOC values (from the left tail of the SOC histogram) and four with the highest SOC values (from the right tail). These samples can be useful to correct bias because the low and high SOC contents are well established. Since low and high values are well described, the offset should be also corrected.

Strategy 4 (OC centre): select eight target site samples with SOC values around the median SOC value of the set.

Strategy 5 (OC distrib): select eight target site samples at regular intervals over the entire range of SOC values (samples evenly distributed across the SOC values). These samples should also be adequate for bias and offset correction.

To apply the three strategies in the second group, we performed a principal component analysis (PCA) of the target site samples (NIR spectra pre-processed with Savitzky–Golay first derivative). The scores of the first, second and third principal components (i.e. the first three) are represented in a scatter-plot.

Strategy 6 (PC periph): select eight target site samples located at the periphery of the principal component spectral space defined by the first three principal components.
Strategy 7 (PC centre): select eight target site samples located at the centre of the principal component spectral space defined by the first three principal components. These are the most similar samples to the mean spectrum of the target site spectra.

Strategy 8 (PC distrib): select eight target site samples evenly distributed across the principal component spectral space defined by the first three principal components. This is the most intuitive strategy to uniformly cover the spectral diversity. This selection was made using the ‘Automatic selection subset’ option in OPUS (version 6.5 software; BrukerOptik GmbH, Ettlingen, Germany), which selects samples in a similar fashion to the Kennard–Stone algorithm (Kennard & Stone, 1969).

The third group of five strategies was based on the Mahalanobis distance values of the target site samples. The Mahalanobis distance values were calculated with respect to the unspiked initial calibrations. Each target site sample had a different Mahalanobis distance depending on the initial calibration used (i.e. IC#1, IC#2 or IC#3).

Strategy 9 (MD low): select eight target site samples with the lowest Mahalanobis distance values (left tail of Mahalanobis distance histogram). These target site samples are the closest to the initial calibration samples and the overall target site samples, and could become a ‘bridge’ between both sets.

Strategy 10 (MD high): select eight target site samples with the highest Mahalanobis distance values (right tail of Mahalanobis distance histogram). These samples are the first recognised as outliers. In some schemes of calibration maintenance (Shepherd & Walsh; 2002), it has been suggested the addition of this type of samples when calibrations must be updated. These target site samples are the most effective decreasing the Mahalanobis distance of the overall target site set (Capron et al., 2005).
Strategy 11 (MD tails): select four target site samples with the lowest Mahalanobis values and four with the highest Mahalanobis distance values.

Strategy 12 (MD centre): select eight target site samples with Mahalanobis distance values around the median Mahalanobis distance value.

Strategy 13 (MD distrib): select eight target site samples at regular intervals over the entire range of Mahalanobis distance values (samples evenly distributed across the Mahalanobis distance values).

2.5 Experimental design and statistical analysis

For this study, a repeated measures factorial design was established. The between-subject factors were ‘initial calibration’, with three levels (i.e. three initial calibrations of different sizes, IC#1, IC#2 and IC#3) and ‘strategy’, with 13 levels (i.e. 13 spiking subset selection strategies). The within-subject factor was ‘extra-weighting’, with two levels (i.e. without and with extra-weighting). For each combination of factors, we calculated the $R^2$, RMSEP, SEP and RPD to compare the actual SOC content of the target site samples with the SOC predicted by the different calibrations. This design was applied separately to the four target sites. The prediction performance parameters obtained in each target site were considered as replicates. We used RMSEP to inform us about accuracy and SEP about precision. The RPD (the ratio between the standard deviation of the prediction set and the RMSEP) allowed us to compare the accuracy obtained in prediction sets with different standard deviations.

The differences in RMSEP, SEP and RPD were analysed using a repeated measures ANOVA. We excluded the strategies based on the SOC values (strategies 1–5) from the statistical analysis because they are not useful in practice. In this way, the repeated measures ANOVA was performed using eight levels of spiking subset selection strategy
and three levels of initial calibration as the between-subject factors, and two levels of extra-weighting as the within-subject factor. Homoscedasticity and normality was checked using Levene and Kolmogorov–Smirnov tests, respectively; the original variables were transformed to meet with the ANOVA assumptions when appropriate. The $R^2$ was excluded from this statistical analysis because it did not meet the assumptions. The assumption of sphericity was not violated when using the Mauchly's test of sphericity. The software IBM SPSS Statistics version 20 (IBM, Armonk, NY) was used for statistical analyses. We also obtained predictions using the unspiked initial calibrations, but these results were not included in the statistical analysis.

2.6. Development of calibrations with PLS-regression

The models relating the NIR spectra with the SOC contents in soils were obtained with PLS-regression (PLS-1 algorithm; OPUS version 6.5 software; BrukerOptik GmbH, Ettlingen, Germany). We selected the number of PLS-vectors through leave-one-out cross-validation. Before calibration, the SOC contents were transformed by the square root but predicted SOC data were back-transformed before we compared them with actual SOC and calculated the prediction performance parameters. NIR-spectra were transformed by the first derivative (Savitzky–Golay, 25 points). The number of PLS-vectors in the spiked calibrations was set to the same number as in the corresponding initial calibration. In TS1, we used the spectral range 1000–2500 nm to meet a common range with a similar noise to the spectra collected with the FT-NIR instrument.

2.7. Additional comparisons: extra-weighting effect versus the increase of the spiking subsets size and versus geographically local models

These comparisons were made only with spiking subsets selected by the ‘PC distrib’ strategy, which was one of the most effective selection strategies in terms of increasing
accuracy. We compared the extra-weighting effect against the increase of the spiking subsets size. To do this, we spiked the three initial calibrations with 8, 16 and 32 spiking subset samples selected by the ‘PC distrib’ strategy. Similar to the procedure described in section 2.3, we obtained spiked calibrations by adding 24 copies of the spiking subset (denoted as EW_24). For each target site, we used these calibrations to predict the SOC contents in the target site samples. In all cases, the 32 spiking subset samples were not used in the RMSEP computation, to allow a fair comparison of accuracy regardless of the size of the spiking subset. The RMSEP values were analysed with a repeated measures ANOVA, where two levels of extra-weighting (with and without extra-weighted) acted as the within-subject factor, and three levels of the spiking subsets size (8, 16 and 32 samples) acted as the between-subject factor. Due to the large differences between the sizes of the initial calibrations, we also used a different approach to calculate the number of copies to add, which was the ratio between the initial calibration size and the spiking subset size. In this way, more copies are added when the initial calibration size is larger or when the spiking subset size is smaller. The extra-weighting effect obtained using the initial calibration-to-spiking subset ratio (denoted as EW_ratio) was evaluated using repeated measures ANOVA, as for the EW_24 approach. The data used in these statistical analyses did not violate the ANOVA assumptions (homocedasticity and normality) or the condition of sphericity. For each target site, three geographically local or site-specific models were constructed using the 8, 16 and 32 spiking subsets selected by the ‘PC distrib’ strategy.

3. Results

3.1. Effect of spiking (without extra-weighted)
The predictions obtained with the unspiked initial calibrations for each target site were inaccurate, with large prediction errors (Figure 3). For the 12 cases (three initial calibrations applied to four target sites), the RPD values ranged from < 0.10 to 1.44, which clearly indicated poor predictions. Figure 4 shows the $R^2$, RMSEP, SEP and RPD values obtained with the unspiked and spiked calibrations, where each value shown is the mean value of those obtained for the four target sites. The unspiked IC#1 provided very low quality predictions, with $R^2 = 0.33 \pm 0.34$ (mean ± standard deviation) and RPD = 0.52 ± 0.21 (Figure 4a). Once spiked, we observed a drastic and positive change in all the parameters related to the quality of predictions (Figure 4a), and bias was substantially decreased. There were differences in accuracy for the spiked calibrations depending on the strategy used to select the spiking subset. For example, the RMSEP values obtained with the IC#1 spiked using the ‘OC low’ (worst) and ‘PC distrib’ (best) strategies were 0.70 ± 0.16% and 0.37 ± 0.15% SOC, respectively, both of which were clearly better than the RMSEP for the unspiked IC#1 of 1.86 ± 1.77% SOC (Figure 4a). Similarly, spiking of IC#2 (Figure 4b) caused a noticeable improvement in prediction accuracy, mostly due to improvement of bias. Interestingly, the worst (‘OC low’) and best (‘PC distrib’) strategies for IC#2 were the same as those observed for IC#1. A substantial improvement in accuracy was also obtained when IC#3 was spiked, due to a strong decrease in bias (Figure 4c). In this case, the worst and best strategies (in terms of accuracy) were not the same as for IC#1 and IC#2. In general, the best accuracies were obtained using IC#1 (the calibration with the smallest size) and the worst accuracies were obtained with IC#3 (the calibration with the largest size). To illustrate the effect of spiking with different spiking subsets, individual results for the four target sites obtained with the ‘MD centre’ and ‘PC distrib’ selection strategies are shown in Figure 3.
3.2. Effect of extra-weighting on the spiking subset selection strategies

The addition of several copies of the spiking subset (i.e. extra-weighting) in the spiked calibrations caused a significant improvement ($P < 0.001$) in the RMSEP, SEP and RPD (Table 3). The effect of extra-weighting on these parameters was similar across the spiking subset selection strategies (extra-weighting × strategy, $P > 0.05$; Table 3), and also similar in the three different initial calibrations evaluated (extra-weighting × initial calibration, $P > 0.05$; Table 3), although the extra-weighting effect on the $R^2$ was greater in IC#3 (Figure 4).

We observed that accuracy differed depending on the strategy used to select the spiking subset (Figure 4). Indeed, all the parameters evaluated showed significant differences across the strategies (Table 3). The differences between strategies were similar in the three initial calibrations evaluated, as suggested by the non-significant interaction between the ‘strategy’ and the ‘initial calibration’ ($P > 0.05$; Table 3). In two strategies (‘OC low’ and ‘OC high’), extra-weighting had a negative effect through an increase in bias (Figure 4). The ‘OC low’ strategy was worst for IC#2 and IC#3, and second worst for IC#1. When extra-weighting was applied, ‘PC distrib’ was the best performing strategy in the three initial calibrations, and clearly improved the accuracy due to decrease in bias, but also due to a decrease in SEP (Figure 3 & Figure 4). In IC#1 and IC#2, the combined use of the spiking subset (‘PC distrib’) and extra-weighting increased the RPD by 1.5 units compared to the unspiked initial calibrations, allowing RPD values to exceed 2 (Figure 4). The results obtained with the ‘MD centre’ and ‘PC distrib’ strategies (without and with extra-weighting) for each target site illustrate the extra-weighting effects (Figure 3).

3.3. Increase of spiking subsets size versus extra-weighting, and comparison with geographically local models
We compared the effects of increasing the spiking subset size with extra-weighting for the ‘PC distrib’ selection strategy. There was a positive effect on the accuracy when the spiking subset size was increased (Figure 5), although this effect was not significant ($P > 0.05$; Table 4). Regardless of the spiking subset size, there was a significant improvement in the accuracy when the spiking subsets were extra-weighted ($P < 0.001$, Table 4). These results were similar for the two approaches followed to select the number of copies to add for extra-weighted (Table 4, Figure 5). It is worth highlighting that in IC#2 and IC#3, the improvement of the accuracy due to extra-weighting was clearly higher than the duplication of the spiking subset size (Figure 5), and even higher than the quadruplication of the spiking subset size in IC#3 (Figure 5). The extra-weighting effect in IC#1 was smaller because spiking was enough to cause the saturation of the improvement, mainly due to its smaller size. When the spiking subset was not extra-weighted (black bars in Figure 5), the best results were obtained with the small-sized initial calibration (IC#1), and results obtained with IC#2 and IC#3 were less accurate than those obtained with the geographically local models. Once the spiking subset was extra-weighted, the differences between initial calibrations practically disappeared, especially when the number of copies added was selected according to the ratio of the initial calibration to the spiking subset (EW_ratio; light grey bars in Figure 5). When this approach was used for extra-weighting (EW_ratio), the spiked initial calibrations were more accurate than the geographically local models. When a large number of local samples (32) were considered as spiking subset size ($SS = 32$), and also as ‘n’ of the geographically local models ($n = 32$), scarce differences between both approaches were observed, except for the reduced robustness obtained with the geographically local models (Figure 5).
4. Discussion

4.1. Effect of spiking

The predictions obtained using the unspiked initial calibrations had a low accuracy. The bias was the main problem, representing more than 50% of the error, as some authors observed (e.g. Bellon-Maurel & McBratney, 2011). These results were expected, and clearly demonstrate how we cannot safely use calibrations do not cover the characteristics of the target sites. As for any model, the spectroscopic calibrations are valid only for samples with similar characteristics as those used in the calibration (Viscarra Rossel et al., 2008). For these reasons, there is a trend to develop large spectral libraries (Shepherd & Walsh, 2002; Brown et al., 2006; Viscarra Rossel, 2009; Grinand et al., 2012; Viscarra Rossel & Webster, 2012). But the accuracy of the calibrations improved drastically when only eight local samples were added to spike the initial calibrations. Once the calibrations contained relevant information for the target site, the predictions became more accurate. The improved accuracy was mostly due to the decrease in bias, in accordance with previous studies (e.g. Stork & Kowalski, 1999; Bricklemyer & Brown, 2010; Guerrero et al., 2010; Stenberg et al., 2010; Wetterlind & Stenberg, 2010), but also by an improvement in precision. Many factors affect soil genesis, and soils present an extraordinary variation in composition and characteristics compared with other environmental materials. This makes it difficult to construct a calibration containing the immense variation found in soils, even at a regional scale (Sudduth & Hummel, 1996; Sankey et al., 2008; Minasny et al., 2009; Reeves & Smith, 2009). In this way, a large calibration does not guarantee accurate predictions. In fact, several authors observed inaccurate predictions when calibrations were used in samples from independent sites (Christy, 2008; D’Acqui et al., 2010; Wetterlind & Stenberg, 2010; Bellon-Maurel & McBratney, 2011). Thus, trying to include all the soil’s
variation is an immense and probably unnecessary effort. Spiking could be an attractive
and economical alternative, avoiding the need for large spectral libraries, since we
observed the best results when the small-sized initial calibration was spiked. As
Guerrero et al. (2010) observed, the new information added (i.e. the spiking subset) was
more influential on a small-sized initial calibration than on a large-sized one, which
eexplains why better predictions were obtained after spiking the small-sized initial
calibration (IC#1).

4.2. Effects of extra-weighting on the spiking subset selection strategies

To directly increase the significance or relevance of the added information, several
copies of the spiking subset were included in the spiked initial calibrations. The addition
of several copies increased their weight and influence on the model (Stork & Kowalski,
1999). Under these circumstances, the calibration was forced to fit preferentially to
these samples. Consequently, if the extra-weighted samples are representative of the
overall prediction set (i.e. the target site), then the calibration must provide reliable
predictions for that set. Indeed, extra-weighting caused a significant improvement
($P < 0.001$) on all the parameters related to the quality of predictions. It is interesting to
highlight that the effects on the precision (SEP) and accuracy (RMSEP) were similar for
the three initial calibrations evaluated, suggesting a robustness of that pattern, since the
three initial calibrations were different to each other. So, extra-weighting is a simple,
fast and inexpensive task that we recommend when spiking calibrations. The extra-
weighting caused a strong decrease in the leverage of the spiking subset (Stork &
Kowalski, 1999; Capron et al., 2005). Consequently, the extra-weighting could be
considered as a manipulation of the spectral space, since it causes a displacement of the
calibration centroid toward the extra-weighted samples. In this sense, the extra-
weighting is a frequent approach used in samples that are added for updating
calibrations to new conditions, especially when their number is relatively low in comparison with the overall calibration set (Stork & Kowalski, 1999), as in our scenarios (especially in IC#2 and IC#3).

The improvement in the RMSEP, SEP and RPD was dependent on the strategy used to select the spiking subset, as Capron et al. (2005) also observed. The differences found between strategies were similar in the three initial calibrations used, as revealed by the non-significant interaction \( P > 0.05 \) between the ‘strategy’ and ‘initial calibration’ factors. These results suggest that the effects exerted by the added samples (spiking subset) are not totally controlled by the characteristics of the initial calibration. The soil samples within a local set are different from each other, the information provided by each sample is different (Naes, 1987; Isaksson & Naes, 1990; Shetty et al., 2012), and consequently, the improvement in the accuracy of the spiked calibration should also vary. In this sense, using an inadequate spiking subset could be one of the reasons explaining why some authors have found a scarce effect of spiking (Bricklemyer & Brown, 2010; Guerrero et al., 2010). Thus, the identification of a successful strategy to select the most adequate spiking subset is clearly relevant. For these reasons, we evaluated strategies aimed to cover a wide range of different types of spiking subset. Since large bias values have been the most common problem observed (Stork & Kowalski, 1999; Janik et al., 2007; Bellon-Maurel & McBratney, 2011), we suspected that using a spiking subset containing strategic SOC values could be adequate to improve the bias, and consequently the accuracy. In fact, we observed that the ‘OC tails’ and ‘OC distrib’ selection strategies offered better predictions than the ‘OC centre’, ‘OC high’ and ‘OC low’ strategies, since they were adding information in several strategic spaces related with the bias, slope and offset. But it is important to note
that the strategies based on the SOC values are not useful in practice, and they were included in the experiment for conceptual evaluation and comparison.

The calibrations spiked with samples evenly distributed in the principal component spectral space (‘PC distrib’) gave better predictions than those spiked with samples evenly distributed along the concentration values (‘OC distrib’). Both strategies select different local samples because the SOC content is not uniquely responsible for the spectral variation within a target site. Compared to texture and mineralogy composition, SOC typically has a fairly small influence on spectra (Stenberg et al., 1995; Islam et al., 2005; Stenberg et al., 2010). This result is interesting since only the spectral information is available in a real situation (Kusumo et al., 2008; Mora & Schimleck, 2008). The predictions obtained with calibrations spiked with a spiking subset selected using the ‘PC centre’ strategy were less accurate than those selected with ‘PC periph’. The samples selected with the ‘PC centre’ strategy are those more similar to the mean spectrum of the target site. In contrast, those selected with ‘PC periph’ are more dissimilar to the mean spectrum, but they represent greater diversity. The strategies that included most of the spectral diversity were ‘PC distrib’ and ‘PC periph’, and they were two successful strategies, especially the latter. Indeed, there are several methods for optimal sample selection based on spectral characteristics (Naes, 1987; Puchwein, 1988; Isaksson & Naes, 1990; Shenk & Westerhaus, 1991; Kusumo et al., 2008) but two of the most commonly used are the Kennard–Stone algorithm (Kennard & Stone, 1969; Mora & Schimleck, 2008; Shetty et al., 2012), which covers the experimental region uniformly (as in ‘PC distrib’), and the D-optimal procedure (Olsson et al., 2004; Rodionova & Pomerantsev, 2007; Brandmaier et al., 2012), which selects objects located on the periphery (most extreme) of the experimental region (as in ‘PC periph’).
There were scarce differences between the selections made using the Mahalanobis distance. The values of Mahalanobis distance were extremely high, and all the local samples were always classified as outliers. Consequently, these sets are not sensitive to the Mahalanobis distance criterion. This criterion would probably be relevant when samples from the target sites are more similar to those comprising the initial calibration (Puchwein, 1988; Capron et al., 2005).

4.3. Increase of spiking subset size versus extra-weighting, and comparison with geographically local models

When the ‘PC distrib’ strategy was used to select the spiking subset, extra-weighting was preferred over the increase in spiking subset size. This was a very interesting result, since extra-weighting caused a significant improvement inaccuracy without any analytical effort. In contrast, the increase of the spiking subset size implies efforts in terms of time and money, and the improvement of the RMSEP was not statistically significant. The non-significant improvement of the RMSEP was probably due to the high efficiency of the ‘PC distrib’ strategy to select the most representative samples. Consequently, a further addition of samples would prove scarcely useful, since the new added samples would be redundant (in comparison with the first ones selected). These results agree with those obtained by other authors (Naes, 1987; Puchwein, 1988; Isaksson & Næs, 1990; Capron et al., 2005; D'Acqui et al., 2010; Grinand et al., 2012; Shetty et al., 2012), where only a small subset of samples properly selected can offer a similar accuracy than a larger set. In this context, extra-weighting the spiking subset is an efficient approach, which can avoid the need of large-sized spiking subsets.

The influence of spiking was greater in the small-sized initial calibrations than in the large-sized ones (Guerrero et al., 2010). When the extra-weighting was made using the same number of copies regardless of the initial calibration size (EW_24), this
pattern was still present, but clearly to a lesser degree. When the extra-weighting was based on the initial calibration to spiking subset ratio (EW_ratio), more copies were included in the large-sized initial calibration (IC#3) than in the smaller-sized initial calibrations (IC#1 and IC#2). However, even under these conditions, the results obtained for the three initial calibrations were similar. This result was very interesting because it suggests that small-sized initial calibrations could offer a similar accuracy than large-sized initial calibrations. Consequently, this approach can be considered as a strong alternative to the need to develop large spectral libraries. In addition, in those circumstances where only a few local samples can be analysed by the reference method (i.e. 8–16 samples), this approach offered more accurate results than the geographically local (or site-specific) models. When a larger number of local samples were analysed (32 local samples), small differences in accuracy were observed between both approaches, although the geographically local models were less robust, indicating the difficulty to develop consistent spectroscopic calibrations when the number of samples is low.

More studies are needed to evaluate if extra-weighting can outperform local models (spectrum-specific models), where a dedicated model is calibrated for an individual unknown sample (Pérez-Marín et al., 2007), or other approaches where a partition of the spectral information is used (Viscarra Rossel & Webster, 2012). It is interesting to highlight that local methods (spectrum-specific) can be used only when the spectral library contains similar samples to the target site samples, which is not the case for sets evaluated in this paper. In contrast, spiking with a properly selected spiking subset, together with extra-weighting, can overcome this problem, allowing the extrapolation of the initial calibrations applicability.
Conclusions

The addition of a small spiking subset (eight local samples) to spike the calibrations improved the accuracy of the SOC predictions. There were, however, important differences in accuracy, which were dependent on the strategy used to select the spiking subset. The best results were obtained when the calibrations were spiked with local samples that were evenly distributed across the space defined by the first three principal components (spiking subset selected with the ‘PC distrib’ strategy). In addition, extra-weighting was an effective way to improve the accuracy of the spiked calibrations. Extra-weighting of the spiking subset accentuates the spiking effect, giving an acceptable level of accuracy when predictions of SOC are needed at local scale, and when using small-sized spiking subsets. Large-sized calibrations are probably not needed when these approaches are considered, since similar results were obtained with the small- and large-sized calibrations, and it suggests that incipient spectral libraries could be useful if they are properly spiked and extra-weighted. Consequently, extra-weighting is a simple, fast and inexpensive task that we highly recommend when calibrations are spiked, and can avoid the need to develop geographically local models. Overall, our results indicate that the efforts needed to use NIR spectroscopy for SOC assessment at local scales can be minimised.

Acknowledgements

This work was part of a research project (Ref. CGL2011-27001) sponsored by the Spanish Government Ministerio de Economía y Competitividad, and C. Guerrero gratefully acknowledges this financial support. C. Guerrero acknowledges the Spanish Government Ministerio de Educación for a travel grant (ref. JC2011-0342). F.T. Maestre was supported by the European Research Council through the European
Commission's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 242658 (BIOCOM).


FIGURE CAPTIONS

Figure 1 Projections of the NIR spectra from the target sites (TS) into the principal component space defined by the first two principal components, in each initial calibration (IC). Grey stars denote the national samples of the initial calibrations and black dots denote target site samples.

Figure 2 Schematic description of the experimental setup: a) initial calibration (IC) unspiked, constructed only with national samples (NS); b) initial calibration spiked with a spiking subset (SS) selected by strategy #1; c) initial calibration spiked with spiking subset selected by strategy #1, where an extra-weighting was applied to the spiking subset. This scheme only shows one of the 13 strategies of spiking subset selection and one of the three initial calibrations. This scheme was used with four different target sites (TS). Dashed and double lines denote spiking and the use of the calibration for obtaining predictions (ŷ), respectively.

Figure 3a Representative illustration of predictions obtained in each target site (TS) with the different calibrations conducted. Left: predictions obtained with the unspiked IC#1 (white stars; dotted line). Centre: predictions obtained with IC#1 spiked with the spiking subset selected with the ‘MD centre’ strategy (white circles, dashed line) and spiking subset extra-weighted (EW) (black circles, solid line). Right: predictions obtained with IC#1 spiked with the spiking subset selected with the ‘PC distrib’ strategy (white circles, dashed line) and spiking subset extra-weighted (black circles, solid line).
Figure 3b Representative illustration of predictions obtained in each target site (TS) with the different calibrations conducted. Left: predictions obtained with the unspiked IC#2 (white stars; dotted line). Centre: predictions obtained with IC#2 spiked with the spiking subset selected with the ‘MD centre’ strategy (white circles, dashed line) and spiking subset extra-weighted (EW) (black circles, solid line). Right: predictions obtained with IC#2 spiked with the spiking subset selected with the ‘PC distrib’ strategy (white circles, dashed line) and spiking subset extra-weighted (black circles, solid line).

Figure 3c Representative illustration of predictions obtained in each target site (TS) with the different calibrations conducted. Left: predictions obtained with the unspiked IC#3 (white stars; dotted line). Centre: predictions obtained with IC#3 spiked with the spiking subset selected with the ‘MD centre’ strategy (white circles, dashed line) and spiking subset extra-weighted (EW) (black circles, solid line). Right: predictions obtained with IC#3 spiked with the spiking subset selected with the ‘PC distrib’ strategy (white circles, dashed line) and spiking subset extra-weighted (black circles, solid line).

Figure 4 Predictions obtained with unspiked and spiked calibrations (without and with extra-weight) using the 13 different strategies to select the spiking subset. Strategies in spiked calibrations (with and without extra-weighting) are arranged by RMSEP. a) IC#1; b) IC#2; c) IC#3. In all cases, n = 4 (from the four target sites studied). The two horizontal dark grey lines are displaying values of RMSEP = 0.4% soil organic carbon (SOC) and RMSEP = 0.8% SOC to facilitate visual comparisons.
**Figure 5** Values of the root mean square error of prediction (RMSEP) obtained with the three initial calibrations (IC) spiked with a spiking subset (SS) of size 8 (SS8), 16 (SS16) and 32 (SS32), without extra-weight (black bars), and with extra-weight (EW; grey bars). Dark-grey bars are used when 24 copies of the spiking subset were added for extra-weighting (EW_24), and light-grey bars are used when the numbers of copies were added in proportion of the initial calibration to spiking subset ratio (EW_ratio). White bars and horizontal lines were used to show the RMSEP obtained with geographically local models, constructed uniquely with 8 (horizontal dotted line), 16 (horizontal dashed line) or 32 local samples (horizontal solid line). In all the cases, the local samples were selected by the ‘PC distrib’ strategy. In all the cases n = 4 (from four target sites). The error bars are denoting one standard deviation.
Figure 1
Figure 2
Figure 3a
Figure 3b
Figure 3c
Figure 4
Figure 5
**Tables**

**Table 1** Characteristics of the three subsets used for the development of the different Initial Calibrations (ICs), and the coefficient of determination ($R^2$) and root mean square error (RMSE) obtained in the cross-validations (RMSECV). All the results refer to soil organic carbon (in %).

<table>
<thead>
<tr>
<th></th>
<th>IC #1</th>
<th>IC #2</th>
<th>IC #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>192</td>
<td>365</td>
<td>2279</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.32</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>Maximum</td>
<td>8.97</td>
<td>14.49</td>
<td>14.62</td>
</tr>
<tr>
<td>Mean</td>
<td>2.35</td>
<td>5.07</td>
<td>1.54</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.87</td>
<td>3.59</td>
<td>2.14</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.05</td>
<td>0.41</td>
<td>3.20</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.95</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>RMSECV</td>
<td>0.40</td>
<td>0.67</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Table 2 Characteristics of the four target sites used. Data refer in all cases to soil organic carbon (SOC; %).

<table>
<thead>
<tr>
<th>Target site 1</th>
<th>Target site 2</th>
<th>Target site 3</th>
<th>Target site 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinates</td>
<td>55°41’N, 13°19’E</td>
<td>38°32’N, 0°49’W</td>
<td>37°09’N, 2°35’W</td>
</tr>
<tr>
<td>Site (country)</td>
<td>Sjöstorp (Sweden)</td>
<td>Sax (Spain)</td>
<td>Gergal (Spain)</td>
</tr>
<tr>
<td>Parent material</td>
<td>Sandy till (25%) and sedimentary clay with elements of chalk (75%)</td>
<td>Gypsum</td>
<td>Mica schists</td>
</tr>
<tr>
<td>Method SOC</td>
<td>LOI(^a) (900°C)</td>
<td>Elemental Analyser</td>
<td>Walkley &amp; Black</td>
</tr>
<tr>
<td>Spectral range / nm</td>
<td>1000-2500</td>
<td>834-2650</td>
<td>834-2650</td>
</tr>
<tr>
<td>n</td>
<td>125</td>
<td>95</td>
<td>60</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.20</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>Maximum</td>
<td>3.87</td>
<td>4.04</td>
<td>6.70</td>
</tr>
<tr>
<td>Mean</td>
<td>1.83</td>
<td>1.80</td>
<td>1.23</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.50</td>
<td>0.71</td>
<td>1.05</td>
</tr>
</tbody>
</table>

\(^a\) LOI: loss on ignition
Table 3 Results of the repeated measures ANOVA to evaluate the effects of extra-weighting, initial calibration and strategy on the different prediction performance parameters: root mean square error of prediction (RMSEP), standard error of prediction (SEP) and ratio of performance to deviance (RPD).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEP</td>
<td>Between-subjects Initial Calibration (IC)</td>
<td>0.605</td>
<td>2</td>
<td>0.302</td>
<td>11.84</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Strategy</td>
<td>0.701</td>
<td>7</td>
<td>0.100</td>
<td>3.918</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>Between-subjects IC × Strategy</td>
<td>0.078</td>
<td>14</td>
<td>0.005</td>
<td>0.220</td>
<td>0.9985</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Error</td>
<td>1.840</td>
<td>72</td>
<td>0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-subjects Extra-weighting (EW)</td>
<td>0.668</td>
<td>1</td>
<td>0.668</td>
<td>81.90</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC</td>
<td>0.015</td>
<td>2</td>
<td>0.007</td>
<td>0.956</td>
<td>0.3890</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × Strategy</td>
<td>0.045</td>
<td>7</td>
<td>0.006</td>
<td>0.794</td>
<td>0.5940</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC × Strategy</td>
<td>0.085</td>
<td>14</td>
<td>0.006</td>
<td>0.751</td>
<td>0.7165</td>
</tr>
<tr>
<td></td>
<td>Within-subjects Error (EW)</td>
<td>0.587</td>
<td>72</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>Between-subjects IC</td>
<td>1.872</td>
<td>2</td>
<td>0.936</td>
<td>6.593</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Strategy</td>
<td>3.760</td>
<td>7</td>
<td>0.537</td>
<td>3.782</td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>Between-subjects IC × Strategy</td>
<td>0.420</td>
<td>14</td>
<td>0.030</td>
<td>0.211</td>
<td>0.9988</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Error</td>
<td>10.22</td>
<td>72</td>
<td>0.142</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW</td>
<td>2.125</td>
<td>1</td>
<td>2.125</td>
<td>60.76</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC</td>
<td>0.126</td>
<td>2</td>
<td>0.063</td>
<td>1.801</td>
<td>0.1725</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × Strategy</td>
<td>0.235</td>
<td>7</td>
<td>0.033</td>
<td>0.959</td>
<td>0.4673</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC × Strategy</td>
<td>0.306</td>
<td>14</td>
<td>0.021</td>
<td>0.626</td>
<td>0.8346</td>
</tr>
<tr>
<td></td>
<td>Within-subjects Error (EW)</td>
<td>2.518</td>
<td>72</td>
<td>0.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>Between-subjects IC</td>
<td>3.209</td>
<td>2</td>
<td>1.604</td>
<td>7.372</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Strategy</td>
<td>3.716</td>
<td>7</td>
<td>0.531</td>
<td>2.439</td>
<td>0.0266</td>
</tr>
<tr>
<td></td>
<td>Between-subjects IC × Strategy</td>
<td>0.417</td>
<td>14</td>
<td>0.029</td>
<td>0.137</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>Between-subjects Error</td>
<td>15.67</td>
<td>72</td>
<td>0.217</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW</td>
<td>3.543</td>
<td>1</td>
<td>3.543</td>
<td>81.90</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC</td>
<td>0.082</td>
<td>2</td>
<td>0.041</td>
<td>0.956</td>
<td>0.3890</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × Strategy</td>
<td>0.240</td>
<td>7</td>
<td>0.034</td>
<td>0.794</td>
<td>0.5940</td>
</tr>
<tr>
<td></td>
<td>Within-subjects EW × IC × Strategy</td>
<td>0.454</td>
<td>14</td>
<td>0.032</td>
<td>0.751</td>
<td>0.7165</td>
</tr>
<tr>
<td></td>
<td>Within-subjects Error (EW)</td>
<td>3.114</td>
<td>72</td>
<td>0.043</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a Log transformed

*b Ln transformed
Table 4. Results of the repeated measures ANOVAs to evaluate the effects of the spiking subset size (SS-size), and those of the extra-weighting (EW) on the root mean square error of prediction (RMSEP) obtained with spiked calibrations. (a) Results obtained when 24 copies were used for EW (EW_24). (b) Results obtained when the number of copies to add for EW was equal to the ratio between the IC size and the SS size (EW_ratio).

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Between-subjects</td>
<td>SS-size</td>
<td>0.0696</td>
<td>2</td>
<td>0.0348</td>
<td>2.328</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>0.4936</td>
<td>33</td>
<td>0.0149</td>
<td></td>
</tr>
<tr>
<td>Within-subjects</td>
<td>EW_24</td>
<td>0.1341</td>
<td>1</td>
<td>0.1341</td>
<td>21.28</td>
</tr>
<tr>
<td></td>
<td>EW_24 × SS-size</td>
<td>0.0087</td>
<td>2</td>
<td>0.0043</td>
<td>0.695</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>0.2079</td>
<td>33</td>
<td>0.0063</td>
<td></td>
</tr>
<tr>
<td>(b) Between-subjects</td>
<td>SS-size</td>
<td>0.0649</td>
<td>2</td>
<td>0.0324</td>
<td>3.117</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>0.3437</td>
<td>33</td>
<td>0.0104</td>
<td></td>
</tr>
<tr>
<td>Within-subjects</td>
<td>EW_ratio</td>
<td>0.1578</td>
<td>1</td>
<td>0.1578</td>
<td>18.45</td>
</tr>
<tr>
<td></td>
<td>EW_ratio × SS-size</td>
<td>0.0119</td>
<td>2</td>
<td>0.0059</td>
<td>0.695</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>0.2821</td>
<td>33</td>
<td>0.0085</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary content: **Appendix I.** Representative NIR spectra of the national samples included in the initial calibrations (top), and two representative NIR spectra of each of the four target sites (bottom).