Geoderma 396 (2021) 115069

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Use of composite samples and NIR spectroscopy to detect changes in SOC contents

César Guerrero [a,*], Romina Lorenzetti [b]

[a] Department of Agrochemistry and Environment. Universidad Miguel Hernández de Elche. E-03202. Elche, Alicante, Spain
[b] CREA - Council for Agricultural Research and Agricultural Economy Analysis, Florence, Italy

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Composites
NIR spectroscopy
4p1000
Soil organic carbon
Spiking
Minimum detectable difference

ABSTRACT

Depending on the land management, soils can behave as sources or sinks of carbon. Thus, monitoring the changes in the SOC (ΔSOC) is of significance. However, this monitoring is often challenging because the value of ΔSOC is typically small in comparison with the spatial variability, and a large sample size is required, which may incur prohibitive costs. In the absence of the necessary sample size, hypothesis testing may be performed under deficient statistical conditions, which may result in misleading conclusions. In this study, we compared several cost-effective approaches to solve the aforementioned problems: i) using composites, as the number of samples to be analysed is smaller, and the costs using a reference method do not represent a substantial amount; ii) using NIR spectroscopy, which is a fast and cheap technique, and allows the analysis of massive number of samples; and iii) the combination of these approaches. In particular, the approaches were compared by evaluating the corrected minimum detectable difference (MDD_C), which corresponds to the MDD after the errors of the methods have been accounted for. Bulking into the composites reduced the analysis costs, and the variance was reduced as the sample size reduced. However, a larger penalisation of the random measurement errors increased the MDD_C. Biased predictions were obtained when an NIR spectroscopy based national scale model was used to predict the SOC at the local scale, resulting in extremely high values of the MDD_C. By using composites to adapt that model to the local conditions through spiking with the extra-weighting technique, the bias was considerably reduced, along with the MDD_C. In addition to their role as a spiking subset, the composites exhibited a desirable effect when combined with the NIR spectroscopy technique, since they allowed to find the model that was predicting SOC contents with the lowest bias, and hence measuring the ΔSOC with the highest accuracy. This analysis was highly specific and cannot be applied using any other spiking subsets, and allowed to identify a model that obtained an MDD_C smaller than that obtained using only the composites. Therefore, the combination of the composites and NIR spectroscopy can be used to monitor ΔSOC since it is inexpensive, accurate and robust, and the spatial distribution of the SOC contents is not lost. Moreover, the cost efficiency of this combined approach probably improves along successive rounds in the monitoring process.

1. Introduction

Soil organic carbon (SOC) plays a central role in soil quality (Bünemann et al., 2018), and increasing the SOC contents can help facilitate climate change mitigation (Minasny et al., 2017). Land management can affect the SOC contents (Lal, 2004; González-Sánchez et al., 2012). Consequently, it is necessary to measure the exact influence of the land management on the SOC and identify the agricultural practices that can help enhance or maintain the SOC levels (Smith et al., 2020). Although experimental plots provide valuable information that can be used to understand the effect of certain agricultural management practices on the SOC levels, the local conditions (soil type, climatic conditions, etc.) likely modulate the response once implemented at the local scale. Therefore, it is necessary to measure, report and monitor the changes in the SOC contents (ΔSOC) to support the implementation of the practices and verify the supposed changes (Jandl et al., 2014; Viscarra Rossel and Brus, 2018). However, in several cases, the ΔSOC induced by the implementation of land management practices, or during two assessment periods of a monitoring programme, is small in comparison to the considerably large spatial variation (Conant and Paustian, 2002; Goidts

[*] Corresponding author.
E-mail addresses: cesar.guerrero@umh.es (C. Guerrero), romina.lorenzetti@crea.gov.it (R. Lorenzetti).

https://doi.org/10.1016/j.geoderma.2021.115069
Received 5 February 2020; Received in revised form 28 February 2021; Accepted 2 March 2021
Available online 10 April 2021
0016-7061/© 2021 The Authors.

et al., 2009; Rawlins et al., 2009). Consequently, a large sample size is required, which may exceed the size that can be obtained within the available budget. In such scenarios, the comparisons performed with an statistical power that is lower than the pre-established desired value, allows to a higher occurrence of Type II errors, which may lead to hypothesis testing with misleading conclusions (Saby et al., 2008; Schrumpf et al., 2011). In this context, several researchers have raised concerns regarding the robustness, capacity and credibility of the protocols and methodology to monitor the ΔSOC with a high accuracy and efficiency (Jandl et al., 2014; Saby et al., 2008; Smith et al., 2020).

This problem can be alleviated or avoided by increasing the period between two assessments until the ΔSOC is sufficiently large to be detected using an affordable number of samples. In this way, the ΔSOC to be detected can be over the minimum detectable difference (MDD). However, this strategy is adequate only for cumulative changes and may lead to a considerable delay in certain cases (Smith, 2004; Necpálová et al., 2014). The use of an adequate sampling design can help, as a paired approach typically needs a lower sample size than that required in the unpaired approach (Conant et al., 2003; Heim et al., 2009; Lark, 2009), because the implied variance is smaller. However, this design cannot be applied over independent plots in a chronosequence (started from the implementation of a different land management strategy) when are compared in an arbitrary time (i.e., synchronous design). Moreover, in some scenarios (e.g., carbon auditing), the paired approach (static sampling design) must be avoided since it can be easily gamed, for example, by applying the preferential management only in the sampling plots and not in the whole area (which is expensive), thereby leading to misleading conclusions (Allen et al., 2010; de Gruijter et al., 2016). In certain other scenarios, the second round sampling cannot be performed by revisiting the same sampling locations (static) because the first round sampling may have led to considerable disturbances (e.g., extracting large cores with heavy machinery); hence, the need to displace the sampling points together with a large short-scale spatial variation may introduce an unacceptably large error (Lark, 2012; Chappell et al., 2013). The efficiency can also be improved by using cost-efficient methods, such as by using composite samples. When using composites, the laboratory costs are drastically reduced because the individual samples are bulked into a few composites, and the SOC contents in only these mixtures are analysed. However, the spatial information of the SOC contents in the collected samples is lost, although this situation is acceptable when the research objective pertains to the changes in the mean values (de Gruijter et al., 2006; Patil et al., 2011). Another drawback is that the sample size is reduced (Viscarra Rossel and Brus, 2018), which can adversely influence the hypothesis testing due to the higher influence of the analysis method's random error (de Gruijter et al., 2006, 2016; Patil et al., 2011). The use of low-cost methods such as diffuse reflectance near infrared (NIR) spectroscopy is another interesting alternative as these methods allow to increase the sample size in a relatively inexpensive manner (Bellon-Maurel and McBratney, 2011; Nocita et al., 2015). However, the quality of the data measured using such low-cost methods is considerably lower in comparison to that achievable using reference methods such as the Walkley–Black, or elemental analysers (Bellon-Maurel and McBratney, 2011; Viscarra Rossel et al., 2016). Thus, it remains unclear whether analysing a large set by using a method such as the NIR spectroscopy technique is more efficient that analysing a smaller set (composites) by using a reference method. Consequently, the main objective of this study was to identify the approach that corresponds to the lowest MDD. Three approaches were considered: (1) The use of composites, which involves a small sample set that can be analysed using a reference method with a low cost; (2) the use of NIR spectroscopy, which can enable an inexpensive and rapid analysis while handling a larger number of samples, or (3) the combination of both techniques (composites and NIR spectroscopy). An experiment was performed in which data were obtained with relatively similar costs (same field effort and laboratory analysis). The errors of the considered methods were determined to calculate a *corrected* MDD value (MDD_C) to allow a fairer and more realistic comparison of the approaches.

2. Material and methods

2.1. Study site

Soil samples were collected in December 2017 in two agricultural fields located in Tarazona de la Mancha (Albacete, Spain). The distance between the fields was less than 2 km, and thus, they exhibited similar characteristics such as the soil type (Calcixerept over old river terraces), meteorological conditions (mean temperature of 14 °C; mean annual rainfall of 500 mm), slope (<2%), parent material (calcareous and fluvial deposits), rotation crops (wheat, maize, etc.) and irrigation system (centre pivot irrigation system). Approximately 20 years ago, the management of one of the fields was changed from conventional to no-tillage (NT20), whereas the other field remained with conventional tillage (CON). The size of the CON and NT20 fields was approximately 60 and 45 ha, respectively, and both the fields were circular irrigation plots. Each field (CON and NT20) was divided into 12 strata (k=12) of similar size, which were defined by compact geographical stratification. This stratification method was selected due to the lack of available ancillary information regarding the SOC variability (Chappell et al., 2013; de Gruijter et al., 2016).

2.2. Samples

2.2.1. Individual samples

Six samples were randomly collected in each stratum. As result, 72 individual samples were collected in CON and 72 in NT20. Each individual sample consisted of four subsamples (cores) located approximately 1 m from the sampling point, and the subsamples were collected using a manual auger at a depth of 0–10 cm. The soil samples were air-dried in laboratory conditions (25 °C) for two weeks and later sieved (<2 mm). The NIR spectra were obtained by Fourier Transform (FT)–NIR diffuse reflectance spectroscopy (MPA, Bruker Optik GmbH, Ettlingen, Germany). Each spectrum was composed of 64 scans, and two spectra per sample were acquired and averaged. The x-scale of the spectra was transformed to nanometres (830–2630 nm) and resampled to 1-nm resolution (OPUS 7.0 software, BrukerOptik GmbH, Ettlingen, Germany). The SOC content (%) in each soil sample was analysed using the Walkley–Black method with laboratory replicates: All the 144 samples were analysed in duplicate, except 72 randomly selected samples, for which the SOC contents were analysed using triplicates. These values were used to obtain the mean SOC content of the 72 samples collected in CON ($\overline{X}_{CON}$) and the mean SOC content of the 72 samples collected in NT20 ($\overline{X}_{NT20}$). The difference in SOC contents due to different agricultural management, denoted as ΔSOC, was obtained with Eq. (1):

$$\Delta SOC = \overline{X}_{NT20} - \overline{X}_{CON} \tag{1}$$

These values were also used to determine the pooled variance (s_p^2), as described in Eq. (2):

$$s_p^2 = \frac{(n_{CON} - 1) \times s_{CON}^2 + (n_{NT20} - 1) \times s_{NT20}^2}{n_{CON} + n_{NT20} - 2} \tag{2}$$

where s^2_{CON} and s^2_{NT20} denote the sample variance observed in samples collected in the CON and NT20 fields, respectively, and n_{CON} and n_{NT20} denote the corresponding sample sizes (here, $n_{CON}= n_{NT20}$).

2.2.2. Composites

In each field, composites samples were obtained by using aliquots from the individual samples. Two main types of composites were obtained: physical and virtual composites. Regardless of the type, the field effort (f) was fixed, corresponding to the 72 individual samples (f=72) collected in each field. Moreover, in all the cases, f=k·n, where k is the number of aliquots contributing to each composite, and n is the final number of composites obtained. For each case, the s_p^2 of the composites was calculated with the s^2_{CON} and s^2_{NT20} observed in the composites.

2.2.2.1. Physical composites.

In each field, the physical composites (n=6) were obtained by bulking one aliquot from each stratum (k=12) defined by the compact geographical stratification (Viscarra Rossel and Brus, 2018). Thus, 12 aliquots from the individual samples (with each sample located in a different stratum) contributed to each composite. In this manner, the 72 individual samples of each field were used to obtain 6 physical composites. Therefore, 6 physical composites were obtained in CON and 6 in NT20. These physical composites represent a case of the composites (f=72, k=12 and n=6) constrained by the strata. The SOC content of each physical composite was analysed in duplicate by using the Walkley–Black method.

2.2.2.2. Virtual composites.

Several types of virtual composites were obtained to provide additional and complementary data that may help explain certain results, especially those related with the materialization and the s_p^2 observed in physical composites. The virtual composites were obtained without generating the physical mixture (i.e., no materialisation occurred). It was assumed that the SOC content of each virtual composite was the arithmetic mean of the SOC contents analysed in the k individual samples that contributed to that composite (de Gruijter et al., 2006; Patil et al., 2011). First, we created virtual composites by using the same combination followed to create the physical composites (see previous section). In the interest of conciseness and to differentiate from other virtual composites, these composites are referred to as theoretic composites hereinafter. The SOC contents of the theoretic composites represented the expected values of the physical composites when it was assumed a perfect materialization and the absence of errors in the analysis the SOC contents.

In order to evaluate if the combination followed to create the physical composites was a particular case or close to an expected case, the s_p^2 of the theoretic composites was compared with the mean s_p^2 obtained in 1,000 different combinations of virtual composites with similar characteristics (f=72, k=12 and n=6), wherein the allocation of each aliquot to a composite was random but constrained by the strata (i.e., one aliquot from each strata). Additionally, another set of 1,000 virtual composites of similar characteristics (f=72, k=12 and n=6) but not constrained by the strata were created to evaluate (by comparison) the relevance of the compact geographical stratification on the s_p^2.

Finally, we obtained virtual composites by varying the number of aliquots (k) contributing to the composite. Since the field effort in each plot was constant (f=72) and f=k·n, the variation in k influenced the final number of composites n, as indicated in Table 1. In each field, 1,000 combinations of the configurations were simulated, as listed in Table 1, in which the k aliquots that contributed to form a composite were randomly selected and not constrained by the compact geographical strata. Thus, the strata defined by the compact geographical stratification did not participate to the allocation of the aliquots into the composites. Subsequently, we computed the s_p^2 in each of these 1,000 simulations for each configuration. The last two rows of Table 1 were not included in this study.

Table 1
Virtual composites in each field.

Number of individual samples (k) contributing to each composite	Final number of composites to be analysed in the laboratory (n)	Field effort (f) (in each field); f = k·n
1 (=not bulking)	72[a]	72 = 1·72
2	36	72 = 2·36
3	24	72 = 3·24
4	18	72 = 4·18
6	12	72 = 6·12
8	9	72 = 8·9
9	8	72 = 9·8
12[c]	6[b]	72 = 12·6
18	4	72 = 18·4
24	3	72 = 24·3
36	2	72 = 36·2
72	1[d]	72 = 72·1

[a] Individual samples.

[b] The physical composites correspond to this combination.

[c] For the physical composites, this k corresponds to the strata as well.

[d] Lowest costs for the laboratory analysis, albeit impractical condition in hypothesis testing (no variance).

2.3. NIR spectroscopy

2.3.1. NIR models

We calibrated different NIR models to estimate the SOC content in the 144 individual samples. The generation of these different models has been based on a modification of the spiking with extra-weighting approach (Guerrero et al., 2014). In all the cases, the PLS-regression algorithm and OPUS software were used to calibrate the models (OPUS 7.0, BrukerOptik GmbH, Ettlingen, Germany).

- UNS: A national scale model was calibrated using a set of 3606 samples collected across Spain. This national scale model was applied to local conditions (i.e., downscaling) without any adaptation. To ensure consistent terminology across all the models, this model was labelled as the unspiked model (UNS).
- SPC: The national scale model was spiked with 12 physical composites (six from each field). The SOC value assigned to each composite of the spiking subset was not the mean of its duplicates, but rather one value randomly selected among the duplicates. This condition thus mimicked the case in which the SOC contents in the samples of the spiking subset were analysed using the reference method without using the laboratory replicates (this aspect is further discussed in Section 2.5.1).
- SEW: To develop this model, the spiking subset of the SPC model was extra-weighted as in Guerrero et al. (2016). To this end, 300 copies of the 12 samples were added to the calibration set. Consequently, the size of the spiking subset (once extra-weighted) was approximately equal to that of the national set (n=3606).
- SEW01 to SEW21: A set of models were derived from the SEW model as follows: (1°) The starting point was the SEW model. Owing to the application of the extra-weighting approach, the extra-weighted samples were fitted preferentially by the model at the expense of a poorer fit over other samples included in the calibration set (i.e., the national samples with different characteristics). Consequently, certain national samples were likely to be identified as outliers. The OPUS software was used to automatically generate a list of potential outliers (concentration outliers) according to the difference between the fitted and actual values and the difference in the residuals among different samples by using the F ratio with a probability level of 99%. (2°) The national samples identified by the software as outliers were removed, and a new model was calibrated without these outliers. (3°) If other samples appeared as new outliers in the new model, they were removed, and a new model was calibrated (i.e., step 2° was repeated). This loop of eliminating the outliers and re-calibrating was stopped when no more national samples were displayed as new outliers or until all the national samples were cleared. In this manner, a number of models was generated during these cycles. The models were successively numbered as SEW01, SEW02, and so on, until SEW21, as the loop finished after 21 cycles.
- SEW-best: All the developed models were used to predict the SOC content in the 144 individual samples. Subsequently, the 144 predictions obtained using a particular model were averaged into 12 mean values in the same manner (same combination) as the physical composites generation. Next, these 12 arithmetic means were compared with the 12 SOC values that were measured in the

composites with the reference method. This comparison directly measures the bias of the prediction, since the bias computed with the 12 values (arithmetic means) and that computed with the 144 values (individual samples) is the same. Therefore, this simple analysis allows to identify which model predicts with less bias the mean SOC in CON ($\overline{\mathrm{X}}_{\mathrm{CON}}$) and the mean SOC in NT ($\overline{\mathrm{X}}_{\mathrm{NT20}}$), and hence, the model that yields the most accurate measure of the change (ΔSOC). Once identified, this model was also labelled as "SEW-best".

2.3.2. *Predictions of SOC*

The SOC contents were estimated in the 144 individual samples with the previously described models, and then the quality of the predictions was assessed by using the classical performance parameters such as the determination coefficient (R^2), standard error of prediction (SEP), bias, root mean square error of prediction (RMSEP) and ratio of performance to the interquartile range (RPIQ). To compute these parameters, the SOC contents predicted using the NIR model were compared with those analysed using the reference method with laboratory replicates.

2.4. *Minimum detectable difference (MDD)*

Given a pre-established acceptable rates of Type I and II errors, the sample variability and its size control the smallest difference between the two means that can be detected, i.e., the MDD (Heim et al., 2009; Petrokofsky et al., 2012; Harcum and Dressing, 2015). In this study, the difference in SOC contents due to different agricultural management, denoted as ΔSOC, is the target. The MDD was computed using Eq. (3), in which the MDD is related to the sample variance and sample size for an arbitrary value of the significance level (α) and power (1 − β), which are parameters related to the Type I and Type II errors, respectively:

$$\mathrm{MDD} = \sqrt{\frac{\left(t_{(\alpha,\nu)} + t_{(\beta,\nu)}\right)^2 \times 2({s_p}^2)}{n}} \tag{3}$$

Here, $t_{(\alpha,\nu)}$ denotes the critical t value (one tail) that is tabulated and can be obtained as a function of alpha (α=0.05 in this study) and the degrees of freedom ν. In this case, a one tail approach was used, which allows a higher power than two tails. If two tails must be considered, then $t_{(\alpha/2)}$ must be determined. $t_{(\beta,\nu)}$ is a critical t value (dependent on the degrees of freedom ν), which is related to the power, and its value was preset as 95% in this study (1−β=0.95; β=0.05). This value can be decreased at the expenses of an increase in the rate of Type II errors. However, this decrease is not recommended due to the importance of this error type. In this scenario, the sample variance was the pooled variance (s_p^2), because it corresponded to a comparison of two means from independent groups. n denoted the sample size (per group). The MDD in composites was obtained with Eq. (3), although the values of s_p^2 and the sample size (n) were different.

2.5. *Correcting the minimum detectable difference*

Eq. (3) implies that the data are obtained using an errorless method. Therefore, the analysis method is expected to yield perfect measurements of the ΔSOC and s_p^2 in the hypothesis testing. This aspect represents an ideal yet unrealistic situation, since all analytical methods contain certain errors. Two methods with substantial differences in their accuracies cannot provide the same MDD for an arbitrary similar sample size (Zimmerman et al., 1993; Zimmerman and Zumbo, 2015). Therefore, the error of the method must be considered in the MDD computation to enable the realisation of a fair and realistic comparison of the MDD when a parameter is obtained using different analysis methods (NIR, Walkley–Black, etc.). Hence, the reliability of the methods is used to produce a fairer and more realistic version of the MDD formulation.

2.5.1. *Reliability of methods used to measure the ΔSOC and pooled variance (s_p^2)*

The reliability refers to the repeatability and consistency of a measure, which in this case refers to the capacity of measuring the ΔSOC and s_p^2. The approaches to quantify the reliability of NIR spectroscopy and Walkley–Black on the different sample supports were slightly different, but in all cases were obtained empirically.

2.5.1.1. NIR spectroscopy. The SOC contents predicted with NIR on the 72 individual samples collected in CON were used to calculate the $\overline{\mathrm{X}}_{\mathrm{CON}}$ and the $s^2{}_{\mathrm{CON}}$. Similarly, the SOC contents predicted with NIR on the 72 individual samples collected in NT20 were used to calculate the $\overline{\mathrm{X}}_{\mathrm{NT20}}$ and the $s^2{}_{\mathrm{NT20}}$. Then, these values were used to obtain the ΔSOC and s_p^2 values, which were compared with the *true* values, allowing the quantification of the errors $\in$ΔSOC and $\in s_\mathrm{p}^2$, respectively:

$$\in\Delta\mathrm{SOC} = |\Delta\mathrm{SOC}\ (\mathrm{NIR}) - \Delta\mathrm{SOC}\ (\mathit{true})| \tag{4}$$

$$\in s_\mathrm{p}^2 = |s_\mathrm{p}^2\ (\mathrm{NIR}) - s_\mathrm{p}^2\ (\mathit{true})| \tag{5}$$

The *true* values of ΔSOC and s_p^2 were considered to be those obtained when the reference method (Walkley–Black) was used to measure the SOC in the individual samples and using laboratory replicates (see Section 2.2.1). We are well aware that a measurement obtained with Walkley–Black is not the *true* value, even if several laboratory replicates are considered; however, here we are considering it as the "gold standard".

As the SOC contents were predicted using 24 different NIR models, 24 different values of $\in$ΔSOC and $\in s_\mathrm{p}^2$ were obtained.

2.5.1.2. Reference method (Walkley-Black) on individual samples. The SOC content in each individual sample was analysed using the Walkley–Black method with laboratory replicates (see Section 2.2.1), allowing to obtain the ΔSOC and s_p^2 in two different ways: (1) when the SOC content assigned to each sample was the mean value of its laboratory replicates, or (2) when the SOC content assigned to each sample was one of its laboratory replicates. As in the previous section, it was assumed that the first case represented the *true* values of ΔSOC and s_p^2, since laboratory replicates were used to measure the SOC ("gold standard"). The second case represented routine conditions, since the SOC was measured with the reference method although without laboratory replicates. Here, the differences between the *true* values of ΔSOC and s_p^2 and those obtained under routine conditions were used to assess the reliability of the Walkley–Black method ($\in$ΔSOC and $\in s_\mathrm{p}^2$).

As the computation of ΔSOC and s_p^2 is obtained with 144 individual samples and at least two laboratory replicates were conducted in each sample, there are billions of different combinations potentially valid to represent empirically the routine conditions, each of which provides plausible but different values of ΔSOC and s_p^2, and hence also billions of different values of $\in$ΔSOC and $\in s_\mathrm{p}^2$. Therefore, to estimate values for $\in$ΔSOC and $\in s_\mathrm{p}^2$ that may be closer to the expected values during routine conditions, the mean values observed in 10,000 different configurations were calculated. These 10,000 different configurations of the dataset were generated by randomly changing the laboratory replicate selected as the SOC content to be assigned to each sample.

The following steps were employed:

1°) For a given arbitrary configuration i, the SOC content assigned to each of the 144 samples was not the mean of the laboratory replicates, but the value of a randomly selected replicate.

2°) The values of $\overline{\mathrm{X}}_{\mathrm{CON}}\ i$ (n = 72), $\overline{\mathrm{X}}_{\mathrm{NT20}}\ i$ (n = 72), $s^2{}_{\mathrm{CON}}\ i$ (n = 72) and $s^2{}_{\mathrm{NT20}}\ i$ (n = 72) were computed for i. Next, the ΔSOCi and $s_\mathrm{p}^2 i$ for i was computed using Eqs. (6) and (7), respectively:

$$\Delta\mathrm{SOC}i = \overline{\mathrm{X}}_{\mathrm{NT20}}i - \overline{\mathrm{X}}_{\mathrm{CON}}i \tag{6}$$

$$s_{\mathrm{p}}^{2}i = \frac{(n_{\mathrm{CON}}-1)\times s^{2}_{\mathrm{CON}}i + (n_{\mathrm{NT20}}-1)\times s^{2}_{\mathrm{NT20}}i}{n_{\mathrm{CON}}+n_{\mathrm{NT20}}-2} \tag{7}$$

3°) Then, the values of the given configuration were compared with the *true* values using Eqs. (8) and (9):

$$\in\Delta\mathrm{SOC}i \quad = |\Delta\mathrm{SOC}i - \Delta\mathrm{SOC}\ (true)| \tag{8}$$

$$\in s_{\mathrm{p}}^{2}i = \left|s_{\mathrm{p}}^{2}i - \ s_{\mathrm{p}}^{2}(true)\right| \tag{9}$$

4°) Steps 1° to 3° were repeated for each of the 10,000 configurations (m), and the mean values were obtained:

$$\in\Delta\mathrm{SOC} = \ \frac{1}{m}\times\sum_{i=1}^{m}(\in\Delta\mathrm{SOC}i) \tag{10}$$

$$\in s_{\mathrm{p}}^{2} = \frac{1}{m}\times\sum_{i=1}^{m}\left(\in s_{\mathrm{p}}^{2}i\right) \tag{11}$$

Thus, the reliability was based on the repeatability (reproducibility) to produce accurate estimations of the ΔSOC and s_{p}^{2}.

2.5.1.3. Reference method (Walkley-Black) on physical composites.. Two laboratory replicates were measured in each physical composite. These two SOC values allowed the generation of 4096 different combinations (2^{12}=4096), each of which represented a plausible case wherein the SOC content in each physical composite was analysed with the reference method without laboratory replicates (i.e., routine conditions). The ΔSOC and s_{p}^{2} were obtained for each combination and then were compared with the *true* values to obtain the ∊ΔSOC and $\in s_{\mathrm{p}}^{2}$. Here it was assumed that the *true* values of ΔSOC and s_{p}^{2} were those obtained with the theoretic composites, because the SOC content of a theoretic composite is the arithmetic mean of 12 individual samples that were analysed using the reference method and laboratory replicates.

As in the previous section, since different values of ∊ΔSOC and $\in s_{\mathrm{p}}^{2}$ were measured for each combination, the mean values were computed. In general, the steps and equations were similar to those used for the individual samples (Eq. (6) to Eq. (11)), and the main difference corresponded to the parameter calculation ($\overline{\mathrm{X}}_{\mathrm{CON}}$, $\overline{\mathrm{X}}_{\mathrm{NT20}}$, s^{2}_{CON} and s^{2}_{NT20}) in step 2°. In this case, the calculations were based on six composites each in CON and NT20, whereas in the previous case, the calculations were based on 72 individual samples each in CON and NT20.

2.5.2. Minimum detectable difference corrected (MDD$_{C}$)

Here, the lack of reliability (or errors) of the methods used to determine the ΔSOC and s_{p}^{2}, namely ∊ΔSOC and $\in s_{\mathrm{p}}^{2}$ respectively, was considered to obtain a realistic MDD, which was denoted as *corrected* MDD (MDD$_{\mathrm{C}}$). This correction was performed in two stages. During the first stage, the $\in s_{\mathrm{p}}^{2}$ was used to *partially* correct the MDD, resulting in MDD$_{\mathrm{Cp}}$, which was obtained with the formula shown in Eq. (12):

$$\mathrm{MDD}_{\mathrm{Cp}} = \sqrt{\frac{\left(t_{(\alpha,\nu)}+t_{(\beta,\nu)}\right)^{2}\times 2\left(s_{pc}{}^{2}\right)}{n}} \tag{12}$$

Eq. (12) is similar to Eq. (3), except for s_{p}^{2} being replaced by s_{pc}^{2}, which represents the *corrected* pooled variance, obtained as follows:

$$s_{\mathrm{pc}}^{2} = s_{\mathrm{p}}^{2} + |\in s_{\mathrm{p}}^{2}| \tag{13}$$

where s_{p}^{2} is the *true* pooled variance, and $\in s_{\mathrm{p}}^{2}$ is the error of the method (Walkley–Black, NIR, etc.; see Section 2.5.1) when s_{p}^{2} is measured. n is the sample size (per group), with a value of 72 for the individual samples and six for the composites. In this manner, the error ($\in s_{\mathrm{p}}^{2}$) was added to the *true* pooled variance (s_{p}^{2}). The deviation against the *true* value could be positive or negative if the values were overestimated or underestimated, respectively. A negative value would decrease the MDD$_{\mathrm{Cp}}$ with respect to the uncorrected MDD, leading to an apparent (false) improvement. Clearly, a method involving a larger error cannot provide improved results, that is, it cannot yield a lower MDD$_{\mathrm{Cp}}$. Therefore, this deviation was considered to be positive, leading to a higher s_{pc}^{2}, which increased the MDD$_{\mathrm{Cp}}$ with respect to the MDD. Consequently, a larger error incurs a higher penalty, which increases the MDD$_{\mathrm{Cp}}$. In other words, a less accurate method possesses less statistical power with the same number of samples (Zimmerman and Zumbo, 2015). In a less formally way, the corrected pooled variance s_{pc}^{2} included in Eqs. (12) and (13) is the pooled variance observed with the method under evaluation (Walkley–Black without laboratory replicates, NIR, etc.).

The second stage of the correction can be expressed as follows:

$$\mathrm{MDD}_{\mathrm{C}} = \mathrm{MDD}_{\mathrm{Cp}} + |\in\Delta\mathrm{SOC}| \tag{14}$$

The value of ∊ΔSOC in Eq. (14) indicates the error of an arbitrary method used to measure the ΔSOC. Therefore, the second correction included in Eq. (14) corresponds to the ability of the method to measure the ΔSOC (∊ΔSOC). As in Eq. (13), this error was directly added (with a positive sign) in Eq. (14) to penalise the MDD$_{\mathrm{C}}$ (i.e., to increase the MDD$_{\mathrm{C}}$) (Chappell and Baldock, 2016). In this manner, both the errors (∊ΔSOC and $\in s_{\mathrm{p}}^{2}$) were included in the expressions to compute the MDD$_{\mathrm{C}}$: larger errors corresponded to a higher MDD$_{\mathrm{C}}$ and to an inferior capacity in hypothesis testing.

3. Results

3.1. Descriptive values of SOC in the fields

The mean SOC content in the 72 samples collected in CON field ($\overline{\mathrm{X}}_{\mathrm{CON}}$) was 1.14 %SOC (Table 2), and the standard deviation was 0.11 % SOC. The values were ranging from 0.90 to 1.34 %SOC. The mean SOC content in the 72 samples collected in NT20 field ($\overline{\mathrm{X}}_{\mathrm{NT20}}$) was 1.66 % SOC (Table 2), and the standard deviation was 0.31 %SOC. The values were ranging from 1.02 to 2.66 %SOC. It was assumed that these fields belonged to a chronosequence, and thus, both fields had similar SOC contents before the change in the management strategy. The accumulation rate (0.53 %SOC in 20 years) was within the range observed by other authors under similar conditions (González-Sánchez et al., 2012; Álvaro-Fuentes et al., 2014).

There was a high correlation (r=0.99) between the SOC contents measured in the physical composites (using the reference method in duplicate) and the theoretic SOC contents of its virtual composites, which were obtained as the arithmetic means of the SOC measured in the individual samples forming the composites. However, despite the high correlation, a slope close to 1 (1.07) and small offset (0.072), the materialisation of the composite was not an perfect procedure because of the inherent errors in the sub-sampling and mixing the aliquots. Moreover, the influence of the measurement random error was expected to be large since the SOC content in a theoretic composite was the arithmetic mean of the 12 values measured in 12 individual samples, whereas only one value was obtained when the SOC content of the physical composite was analysed. As consequence, the mean SOC content measured in the physical composites from CON was 1.15 %SOC (n=6), 1.71 %SOC in those from NT20 (n=6), and their difference (ΔSOC) was 0.55 %SOC, being slightly different respect to that observed with the individual samples (ΔSOC=0.53 %SOC).

3.2. Variances in individual samples and composites

The sample variance observed in the 72 individual samples collected in CON (s^{2}_{CON}) and those 72 collected in NT20 (s^{2}_{NT20}) was 0.012 % SOC^{2} and 0.093 %SOC^{2}, respectively. Therefore, when 72 individual samples from each field were used, then the pooled variance s_{p}^{2} was 0.053 %SOC^{2} (Fig. 1). As expected, the variance in the composites was substantially lower than the observed in the individual samples. Fig. 1 shows the s_{p}^{2} obtained with virtual composites formed at different values of k, and therefore not constrained by the compact geographical strata.

Table 2
Main parameters measured using the reference method Walkley–Black (WB), and predicted with four different NIR models (UNS, SPC, SEW and SEW-best). The SOC content was expressed in %.

	Walkley–Black (WB)		NIR models			
	True (WB with replicates)[a]	Routine (WB without replicates)[b]	UNS	SPC	SEW	SEW-best
$\overline{X}_{CON}$	1.14	1.14	1.49	1.45	1.07	1.15
$\overline{X}_{NT20}$	1.66	1.66	2.57	2.52	1.84	1.70
ΔSOC	0.525	0.525	1.081	1.065	0.762	0.551
∊ΔSOC	0[c]	0.0082[d]	0.5559[e]	0.5402[e]	0.2370[e]	0.0256[e]
s^2_{CON}	0.012	0.014	0.029	0.028	0.017	0.019
s^2_{NT20}	0.093	0.099	0.343	0.337	0.222	0.115
s_p^2	0.053	0.056	0.186	0.183	0.120	0.067
ϵs_p^2	0[c]	0.0043[f]	0.1335[g]	0.1300[g]	0.0670[g]	0.0140[g]

[a] The SOC content assigned to each sample was the mean of its laboratory replicates.
[b] The SOC content assigned to each sample was one of its laboratory replicates.
[c] Since WB is used with replicates, these values are assumed as *true* values.
[d] Obtained using Eq.(10).
[e] Obtained using Eq. (4).
[f] Obtained using Eq. (11).
[g] Obtained using Eq. (5).

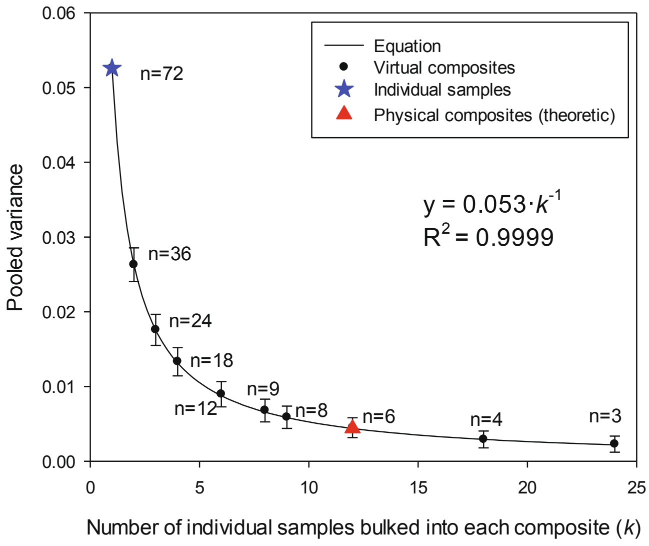

Fig. 1. Pooled variance (s_p^2) in composites: The points denote the s_p^2 obtained using the virtual composites at different values of k for a fixed field effort f (f = 72), with $f = k{\cdot}n$. The number located near the symbol (n) denotes the final number of samples to be analysed in the laboratory (i.e., sample size from each field). The error bars represent the standard deviation observed in 1,000 simulations. The star denotes the s_p^2 of the individual samples (k = 1; not bulked), and the triangle denotes the theoretic value of the physical composites (k = 6). SOC content is expressed in %.

In all the cases, f=72 and f=$k{\cdot}n$ (Table 1); therefore, the final number of composites (n) decreased as k increased. The values of s_p^2 plotted in Fig. 1 indicate averages obtained after 1,000 simulations, and hence, they represent the expected values. The results of these simulations were in agreement with the theoretical pattern, wherein the variation in the f individual samples, s_p^2, reduced by k times if k aliquots were pooled to obtain n composites. For instance, at k=12, the s_p^2 obtained using the six composites each from the CON and NT20 fields was 0.0044 %SOC^2, which was approximately 12 times lower than that observed using the individual samples (s_p^2=0.053 %SOC^2). However, the reduction rate of size k was the expected value, and the exact value in a particular combination depends on the aliquots mixed (Patil et al., 2011). As in stratification, if the aliquots contributing to a composite were collected across different well-defined homogeneous strata, a large reduction in the sample variance and an increase in the precision can be expected (de Gruijter et al., 2016). Conversely, bulking within homogeneous strata (i.e., when generating internally homogeneous composites) may even increase the sample variance (Patil et al., 2011). Therefore, according to the trend shown in Fig. 1 (neutral effect of stratification and compositing), the s_p^2 in the composites was expected to be approximately k times lower than that observed in the individual samples, especially at larger values of k.

When the formation of the virtual composites was constrained by the strata (defined by compact geographical stratification) the s_p^2 was, on average of 1,000 simulations, 0.0043 %SOC^2. This value is very close to the observed with composites not constrained by the strata (0.0044 %SOC^2) when k=12. This result is suggesting that compact geographical stratification has not been particularly efficient in the stratification of the SOC contents, since in both the cases, these numbers were similar to the expected rate of 12 (with k=12 in both the cases), indicating a neutral effect. Thus, the distribution of the SOC in the fields was not directly related to the strata.

The combination followed to create the physical composites represents one of the several plausible combinations wherein the formation of the composites was constrained by the strata defined by the compact geographical stratification. The s_p^2 computed using the corresponding theoretic composites, i.e., the virtual composites developed using the same combination used to obtain the physical composites, was also 0.0043 %SOC^2. Thus, the combination followed to form the physical composites can be considered as a representative case and not a particular case. Nevertheless, this value (s_p^2=0.0043 %SOC^2) represents a perfect materialisation of this combination, and therefore, it serves as the theoretic reference. When the composites were physically materialised, and the SOC contents were analysed in the mixtures using the reference method (with duplicates), the s_p^2 of the physical composites was 0.0101 %SOC^2. This value was different than that computed using the theoretic composites (0.0043 %SOC^2), likely because of the errors in the materialisation process and the SOC determination (Lancaster and Keller-McNulty, 1998; de Gruijter et al., 2006).

3.3. MDD obtained using individual samples and virtual composites

Fig. 2 shows the MDD obtained using Eq. (3) with the data from the individual samples (72 per field; represented by a star), from the virtual composites having different k (represented by points) and from the theoretic SOC contents of the physical composites (6 per field; represented by triangles). The values of the MDD computed using the individual samples (n=72) and theoretic (virtual) composites (n=6) were very close. Despite the large differences between the sample size and s_p^2 (Fig. 1), both the parameters decreased by k times in the case of

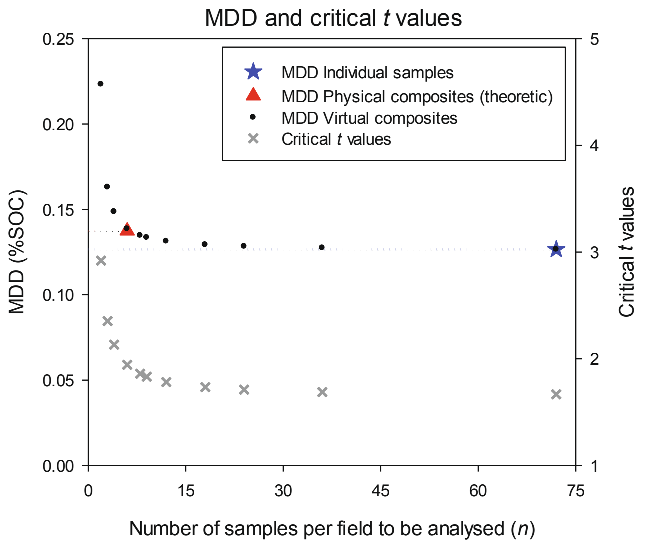

Fig. 2. Left y-axis: Minimum detectable difference (MDD) obtained using individual samples (star), virtual composites (points) and physical composites (theoretic) (triangle). The results were obtained assuming ideal but unrealistic conditions, in which the materialisation (in the case of composites) and the analysis of the SOC contents were perfect procedures, and therefore, the difference in the means and pooled variance was determined without errors. The values of α and β were 0.05 each. Right y-axis: Critical *t* values (denoted as grey crosses) in the computation of the MDD (one tail, when $\alpha = 0.05$). These critical *t* values were also valid for β in Eq. (3) since the power $(1 - \beta)$ was fixed at 95% $(\beta = 0.05)$.

composites. Consequently, it was considered that bulking into the composites exerted an almost neutral effect on the MDD. However, the MDD for the composites was slightly larger than that obtained using the individual samples (Fig. 2). This difference was only due to the degrees of freedom (ν), which affected the critical *t* values. The critical *t* values increased with the decrease in degrees of freedom. In our pre-established conditions (one tail, α=0.05), the value of $t_{(\alpha)}$ was 1.8125 for the composites (ν=10 when n=6 per field) and 1.6556 for the individual samples (ν=142 when n=72 per field). These values were also used for $t_{(1-\beta)}$, because the power was pre-established at 95% (1−β=0.95). Therefore, in a virtual plane, the use of composites is an interesting approach to avoid large costs, since it does not produce a substantial negative impact on the MDD. Indeed, the difference in MDD respect to the obtained with the individual samples was small, whereas the laboratory costs were decreased by 12 times (k=12). The differences in MDD along the different configurations (f=k·n) of the virtual composites occurred only due to the variations in the critical *t* values (denoted as grey crosses in Fig. 2), and assuming a neutral effect of the stratification (i.e., the variance diminishes at rate *k*).

3.4. Predictions

Fig. 3 shows the predicted SOC contents in the 144 individual samples in comparison with the *true* values (measured with the reference method and laboratory replicates). The predictions obtained using the national scale model (UNS) were biased (Fig. 3a). This pattern has been widely observed when a national-scale model has been directly used to predict at the local scale (downscaling) without previous adaptation to the local conditions (Bellon-Maurel and McBratney, 2011; Guerrero et al., 2016). Although its precision is adequate, the accuracy of this unspiked model is extremely low owing to large bias. Similar results were obtained when the national scale model was spiked with 12 composites (SPC), indicating an extremely small effect of spiking (Fig. 3b). This result was expected because the impact of 12 samples on a large calibration set was almost null, as noted by Guerrero et al. (2014, 2016). The accuracy was improved substantially when the spiking subset was extra-weighted (SEW), primarily owing to the decrease in the bias and somewhat also due to the decrease in the SEP (Fig. 3c).

The preferential fitting over the extra-weighted samples was performed at the expense of a poorer fit over the remaining samples. Therefore, certain samples from the national set appeared as concentration outliers after applying the extra-weighting step (data not shown). After the outliers were removed, a new model was fitted with the remaining samples; however, this new fitting involved a new set of outliers, which were removed as well, leading to another new different model (data not shown). This loop was stopped after 21 cycles, when no more outliers were identified. Therefore, a total of 21 new models were generated.

Fig. 4a shows the RMSEP obtained when the different models were used to predict the SOC contents in the 144 individual samples. The point denotes the RMSEP computed using the 144 predictions obtained in the individual samples. As expected, the accuracy of these 21 models varied, since their calibration sets were different in size (as the number of cycles increased, the size of the calibration set decreased). This variation implied the existence of an optimal value. However, under practical conditions, this optimal value cannot be known since the SOC contents in the 144 individual samples are not available. The star in Fig. 4a denotes the RMSEP when the 144 predictions were averaged as the 12 composites. This information is available under practical (realistic) conditions because the SOC contents in the composites are known, since they were analyzed to build the spiking subset. A correlation exists between the two approaches to compute the RMSEP, although the corresponding values are different, as expected. Consequently, the RMSEP obtained using the averages provides only partial information, and the pattern of the RMSEP in the 144 individual samples cannot be inferred.

Fig. 4b shows the bias. The bias calculated using the 144 individual samples (points in Fig. 4b) and the 12 averages (stars in Fig. 4b) was the same owing to the data configuration (as indicated in Table 1), wherein all the individual samples were averaged into mean values with equal sizes (i.e., all the composites were created using *k* aliquots). Therefore, the model producing less biased predictions could be identified under all those practical conditions in which the SOC contents in the 12 composites are known, such as in the SPC, SEW and the 21 newly derived models (SEW01 to SEW21).

Similar to the bias, the ∊ΔSOC computed using the 144 individual samples (points in Fig. 5a) and computed using the 12 averages (stars in Fig. 5a) was the same. The sequential elimination of the outliers (together with the model recalibration as new model) resulted in a decrease in the ∊ΔSOC, especially during the first five cycles, followed by a stabilising trend. The model calibrated in cycle #10, i.e., model SEW10, was the "best model" as it corresponded to the minimum ∊ΔSOC (Fig. 5a), and therefore, it was also labelled as "SEW-best" (Fig. 3d). As discussed, the information displayed as points cannot be obtained under realistic conditions, because the actual SOC contents of the 144 samples are unknown, and are only available under experimental conditions. Hence, the SEW-best model cannot be identified. In contrast, SEW-best can be determined by inspecting the ∊ΔSOC computed using the predictions averaged as composites (stars, Fig. 5a), because under practical conditions, the actual SOC contents in the 12 composites are known.

3.5. Reliability of methods used to determine the ΔSOC and pooled variance (s_p^2)

3.5.1. Using NIR spectroscopy (n=72 per field)

Table 2 lists the $\overline{X}_{CON}$ and s^2_{CON} of the 72 samples collected in the CON field, and the $\overline{X}_{NT20}$ and s^2_{NT20} of the 72 samples collected in the NT20 field, depending on the NIR model used to estimate the SOC. This table also summarises the capacity of the NIR predictions to measure the ΔSOC and s_p^2. Despite the high R^2 values (above 0.83), the capacity of the UNS and SPC models to measure the ΔSOC was low due to the large bias,

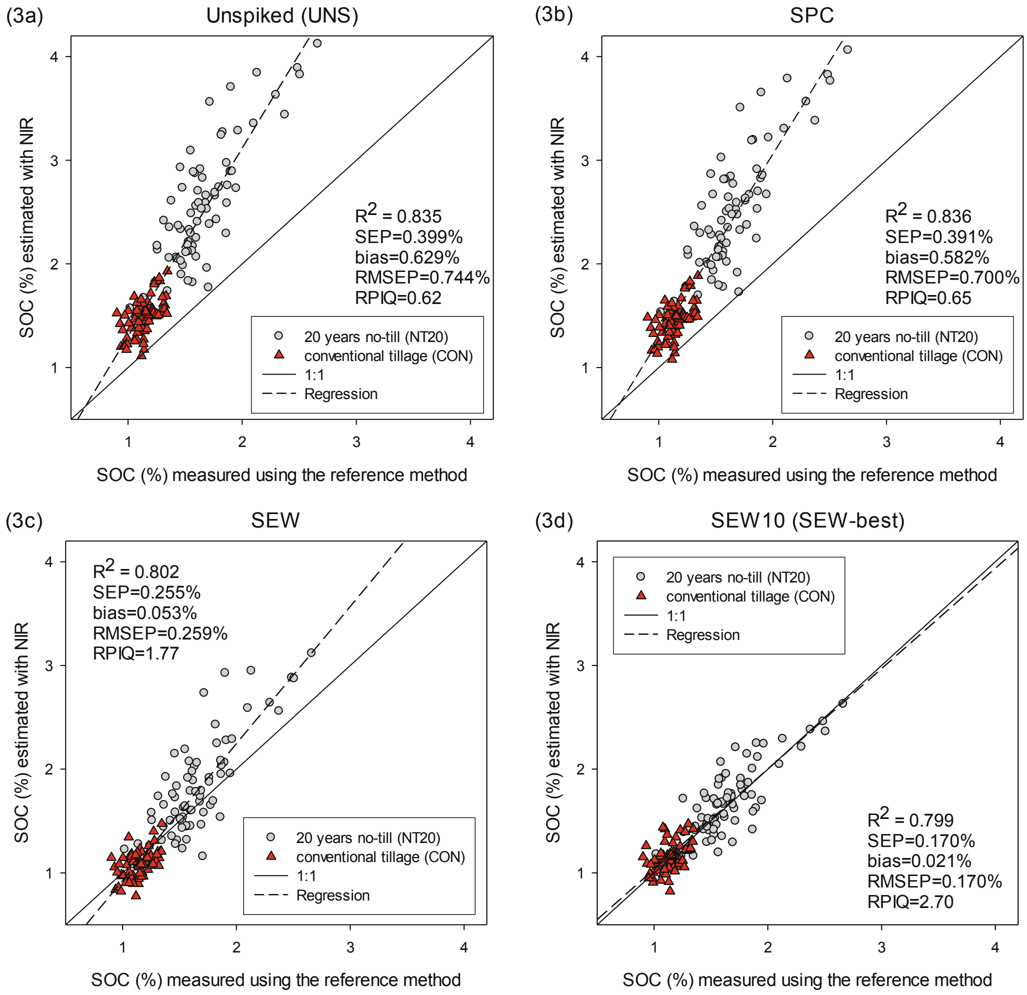

Fig. 3. Predicted SOC contents in the 144 individual samples against the values measured using the reference method (Walkley–Black). The values predicted using the unspiked model (UNS), spiked model (SPC), SEW model, and SEW10 (also labelled as "best") are shown in 3a, 3b, 3c and 3d, respectively.

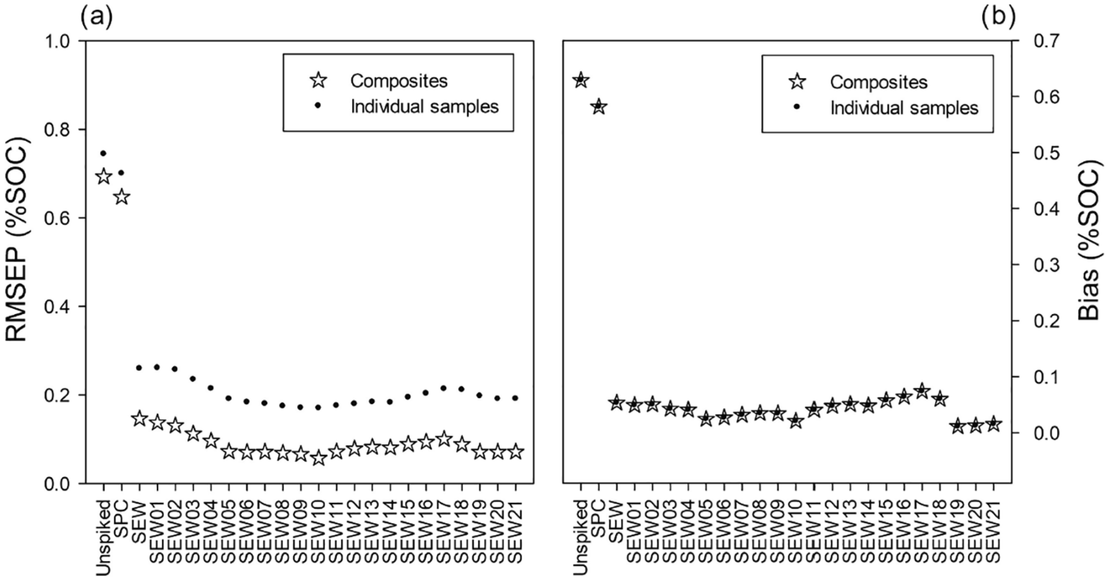

Fig. 4. Values of RMSEP (4a) and bias (4b) of predictions obtained with different models. The point denotes the value computed with the 144 individual samples, whereas the star denotes the value when it was computed with 12 averages of these 144 individual samples.

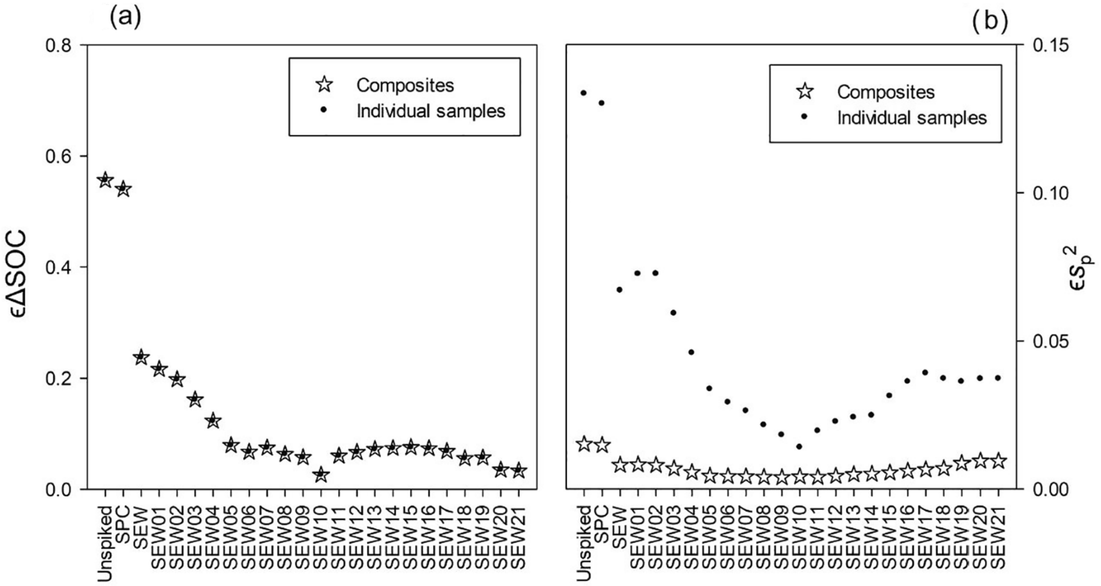

Fig. 5. Values of ϵΔSOC (5a) and ϵs_p^2 (5b) obtained with different models. The point denotes the value computed with the 144 individual samples, whereas the star denotes the value when it was computed with 12 averages of these 144 individual samples.

and the values of ϵΔSOC were substantial (>0.54). The predictions obtained after extra-weighting (SEW) halved that error due to the effect of the extra-weighting in the bias. As discussed, when the values of the composites are known, the model that yields the lowest ϵΔSOC can be determined. This identification is crucial, as in this study, the minimum ϵΔSOC was one order of magnitude lower than that observed using the SEW model (Table 2). As shown in Fig. 3d, the SEW-best model predicted the SOC with an extremely small bias (0.021%).

Unlike the ϵΔSOC, the ϵs_p^2 values computed using the 144 individual values (points; Fig. 5b) were notably different than those computed with the values averaged as composites (stars; Fig. 5b). Therefore, the model yielding the minimum ϵs_p^2 value could not be identified under realistic conditions. Nevertheless, the different accuracy of those spectroscopic models was also noted when the predictions were used to estimate the s_p^2, and the largest and smallest error ϵs_p^2 occurred when the SOC was predicted using the unspiked model (UNS) and SEW-best model, respectively (Fig. 5b).

3.5.2. Using the Walkley–Black method in the individual samples (n=72 per field)

Although Walkley–Black is a reference method, it involves errors, and its reliability depends on how the method is applied during routine conditions (for example, if laboratory replicates are used). As described, in this section we examined the reliability of the Walkley–Black method to measure the ΔSOC and s_p^2 when only one determination per sample was used (i.e., without laboratory replicates), which were considered as the routine conditions. To resemble the scenario in which the SOC is analysed without laboratory replicates, we simulated 10,000 different configurations of the data. In each of these configurations, the SOC content assigned to each sample was not the mean of its laboratory replicates, but the SOC content of one of its randomly selected replicates. Consequently, the values of the parameters ($\overline{X}_{CON}$, $\overline{X}_{NT20}$, ΔSOC, s^2_{CON}, s^2_{NT20} and s_p^2) varied in each particular random configuration, owing to the differences in the replicates (i.e., repetitiveness of the method).

The values of the $\overline{X}_{CON}$ (n=72) and $\overline{X}_{NT20}$ (n=72) observed in these 10,000 configurations are shown in the x-axis in Fig. 6. The mean ΔSOC value was 0.525 %SOC (Table 2), and the minimum and maximum ΔSOC values were 0.485 and 0.560 %SOC, respectively. The mean ϵΔSOC was 0.0082, and this value was considered as the expected error when the ΔSOC was computed under routine conditions (i.e., using the Walkley–Black method to measure the SOC content in the samples, without laboratory replicates). Similarly, the values of the s_p^2 differed in each of the 10,000 simulations. The values of s^2_{CON} (n=72) and s^2_{NT20} (n=72) obtained in each simulation are shown on the y-axis in Fig. 6. The mean s_p^2 was 0.056 %SOC^2. The mean ϵs_p^2 was 0.0043, and this value was considered as the expected error when the s_p^2 was computed under the given routine conditions (that is, using the Walkley–Black without laboratory replicates).

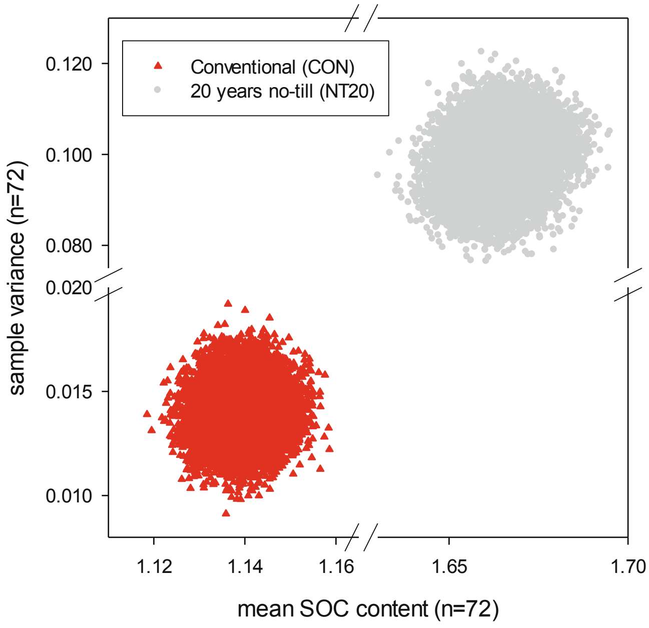

Fig. 6. Values of the mean SOC content and sample variance in the CON (n = 72) and NT20 (n = 72) fields, observed in 10,000 different configurations of data. Each configuration represents a case in which the SOC content in the 144 individual samples was measured using the reference method (Walkley–Black) without laboratory replicates (routine conditions). SOC content is expressed in %.

3.5.3. *Using the Walkley–Black method in the composites (n=6 per field)*

The capacity to provide a reliable value of ΔSOC and s_{p}^2 was different in the case of composites than in the case of individual samples, despite the method for the SOC analysis and its laboratory conditions being the same (Walkley–Black without laboratory replicates). When only one replicate was used to analyse the SOC content in the composites, ∊ΔSOC and ∊s_{p}^2 were 0.0321 and 0.0078, respectively. These errors were higher than those obtained in the individual samples (Table 2), and the lower reliability could be partially attributed to the presence of additional errors linked to the imperfect materialisation of the composites, such as those pertaining to the sub-sampling and weighting of the aliquots, which was performed 12 times, and the mixing. However, the main reason for the lower reliability was the larger contribution of the random measurement error in the composites, because for the computation of each parameter ($\overline{\mathrm{X}}_{\mathrm{CON}}$, $\overline{\mathrm{X}}_{\mathrm{NT20}}$, ΔSOC, $s^2{}_{\mathrm{CON}}$, $s^2{}_{\mathrm{NT20}}$ and s_{p}^2), the individual samples involved 12 times more measurements than those pertaining to the composites.

Assuming ideal conditions wherein the analytical method measures without error, the variance of the estimated sample mean depends on the spatial variation (sample variance) and the sample size. In such a case, the variance of the estimated sample mean in the CON and NT20 fields by using the individual samples is expected to be $s^2{}_{\mathrm{CON}}/n$ and $s^2{}_{\mathrm{NT20}}/n$, respectively. When the composites are used, these values are the same, because $s^2{}_{\mathrm{CON}}$ and $s^2{}_{\mathrm{NT20}}$ decreased k times (in the same proportion as the decrease in the sample size, k). Interestingly, the process is k times cheaper, since the sample size to be analysed is k times smaller, and neutral (in terms of the precision) because the sample variance in the composites is approximately k times smaller than the sample variance of the individual samples. However, the measurements are not exact even when using a reference method, and the measurement error also contributes to the variance of the estimated sample mean. Therefore, when the individual samples were used, the sample size was k times larger than that in the case of the composites, and hence, the random measurement error was expected to exert a k times smaller influence on the variance (Patil et al., 2011; de Gruijter et al., 2016).

3.6. *MDD*$_c$

The results described in the previous section indicated that the methods cannot perfectly measure the ΔSOC and s_{p}^2 owing to the differences in the reliability measuring SOC contents (Table 2). These differential reliabilities must be considered when comparing the different methods and conditions. To enable a fairer and more realistic comparison, we computed the corrected MDD (MDD$_{\mathrm{C}}$), which takes into account the errors. To this end, we used Eq. (14), according to which, an increase in the errors results in an increase in the MDD$_{\mathrm{C}}$ (i.e., larger penalisation).

The MDD$_{\mathrm{C}}$ values are shown in Fig. 7. When the errors of the Walkley–Black method were considered, the MDD$_{\mathrm{C}}$ calculated using the 144 individual samples (72 per field) was 0.1399 %SOC, which is slightly larger than the "ideal" MDD (0.1266 %SOC), represented by the white segment in Fig. 7. The penalisation owing to ∊ΔSOC was more important than that of ∊s_{p}^2, as the size of the black segment was considerable larger than the grey segment, almost imperceptible in Fig. 7. The SOC contents in these 144 samples might be analysed using laboratory replicates to reduce the errors ∊ΔSOC and ∊s_{p}^2 and consequently the MDD$_{\mathrm{C}}$. However, due to the small differences between the MDD and MDD$_{\mathrm{C}}$, the expected improvement is small. Thus, the use of laboratory duplicates is not worth, as it requires the duplication of the efforts in the laboratory, necessitating the determination of 144 additional values of SOC (288 in total), which may be extremely expensive yet almost ineffectual (Goidts et al., 2009; Rawlins et al., 2009).

The SOC contents in the physical composites were analysed under similar routine conditions as those of the individual samples, that is, by using the Walkley–Black method without laboratory replicates. However, the number of measurements to obtain ΔSOC and s_{p}^2 was 12 times lower in the composites (n=6 per field) than that in the case of individual samples (n=72 per field), and therefore, the random measurement errors were less attenuated. Moreover, additional errors were present in the composites, owing to the imperfect materialisation. Therefore, the difference between the MDD and MDD$_{\mathrm{C}}$ of the composites was substantially higher, since it changed from 0.1374 to 0.2629 %SOC (Fig. 7) and only from 0.1266 to 0.1399 %SOC in the case of the individual samples. A small portion of the difference between the MDD$_{\mathrm{C}}$ in composites and individual samples could be partially attributed to the higher t value (almost similar white segments in Fig. 7), which depends on the degrees of freedom, as explained in Section 3.3. Hence, the contribution of this effect is smaller than that of the errors. Nevertheless, whereas the composites represent a 12-fold cheaper option compared to the individual samples, the associated MDD$_{\mathrm{C}}$ is not 12-fold worse. Therefore, the use of composites may be considered as a cost-effective approach that can be used when the budget for the SOC analysis is limited and a poorer (higher) value of the MDD$_{\mathrm{C}}$ (0.2629 %SOC) is acceptable and still useful (Saby et al., 2008; Minasny et al., 2017; FAO, 2019; Smith et al., 2020).

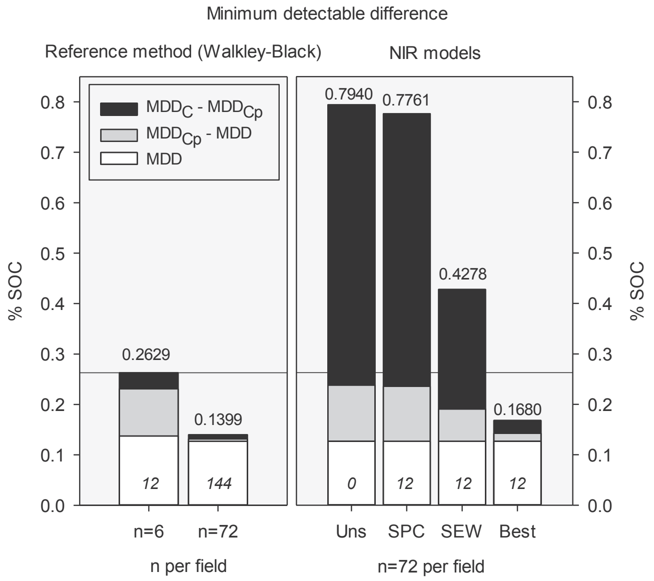

Fig. 7. The white segments represent the MDD under ideal conditions (SOC determined using an exact method; Eq. (3)). The grey and black segments denote the two stages of the MDD correction (see Section 2.5.2), and together with the MDD (white segment), all these segments represent the total size of the MDD$_{\mathrm{C}}$. n denotes the sample size (per group). The number in italics (inside the white segment in the columns: 0, 12, 144) denotes the total number of Walkley–Black determinations required. The number above the bar denotes the MDD$_{\mathrm{C}}$.

The MDD$_{\mathrm{C}}$ obtained using the NIR spectroscopy was strongly dependent on the model used to estimate the SOC content in the 144 individual samples, since the MDD$_{\mathrm{C}}$ values ranged from 0.1680 to 0.7940 %SOC (Fig. 7). These values were notably higher compared to the corresponding MDD (white segments in Fig. 7), which was 0.1266 % SOC in all the cases, owing to the same sample size (72 samples per group). The highest MDD$_{\mathrm{C}}$ was obtained with data obtained from model UNS, which was six times higher than the corresponding MDD, mostly due to the error ∊ΔSOC (Table 2), which represents the 70% of the total MDD$_{\mathrm{C}}$. Similar results and patterns were observed when using the SPC model. Due to the substantial decrease in the bias caused by the extra-weighting, the reliability to measure the ΔSOC was notably improved (Table 2), resulting in a diminution of the MDD$_{\mathrm{C}}$ of the SEW model compared to that of the SPC model. Despite the important attenuation of the bias, ∊ΔSOC was the main factor contributing to the difference between the MDD$_{\mathrm{C}}$ and MDD in the SEW model, representing 55% of the

MDD$_{\mathrm{C}}$.

The lowest MDD$_{\mathrm{C}}$ was obtained using the SEW-best model, for which the errors represented less than 25% of the MDD$_{\mathrm{C}}$. Anyway, as expected, the MDD$_{\mathrm{C}}$ in SEW-best (0.1680 %SOC) was higher than that obtained using the Walkley–Black method (0.1399 %SOC) with the same sample size (72 per group; 144 samples). However, the costs associated with the Walkley–Black for 144 samples are likely prohibitive, whereas the analytical cost to process 144 samples using NIR spectroscopy is considerably lower, since the scanning of the samples does not require the use of chemicals and is a rapid process.

Interestingly, the MDD$_{\mathrm{C}}$ obtained using the SEW-best model (0.1680 %SOC) was lower than that obtained using the 12 physical composites (0.2629 %SOC). These approaches differ only in the scanning costs (required for SEW-best). Except for the scanning cost, the approaches are nearly equivalent in terms of the effort, because both the approaches require a similar number of SOC analyses using the reference method. This equivalence occurs since the size of the spiking subset was the same as the number of physical composites (12), n=6 per field in both the cases, and the SOC content was determined without laboratory replicates. Hence, the scanning cost, which is relatively small, offers a net advantage owing to the lower MDD$_{\mathrm{C}}$ and allows the retention of the information regarding the spatial distribution of the SOC contents, which is provided by the SOC predicted using the individual samples and is lost when using composites.

4. Discussion

It is necessary to develop credible, accurate, efficient and affordable methodologies and protocols to verify the changes in SOC contents (Saby et al., 2008; Schrumpf et al., 2011; Petrokofsky et al., 2012; Jandl et al., 2014; Smith et al., 2020). In this regard, it is necessary to identify the optimal methodology, because even small differences in the costs (and accuracy) represent a substantial amount once extended in space and time. In terms of the accuracy, we must assume that all the methods involve errors, and their measurements are more or less reliable. A method with a low reliability produces measurements with errors, and consequently, for any arbitrary sample size, the statistical power is always overestimated compared to that of an ideal exact (non-existent) method (Zimmerman and Zumbo, 2015). Thus, comparing methods using the MDD is not recommended because the errors in the method are not considered in this approach. In contrast, the comparison using the MDD$_{\mathrm{C}}$ is fairer and more realistic, since this value represents the MDD corrected considering the errors in the method. Larger errors correspond to a higher penalisation (and correction) and thus higher MDD$_{\mathrm{C}}$ values. This aspect enables a suitable evaluation of the capacity of the method, thereby allowing a fairer comparison among different approaches.

Composite samples is well-known and widely used sample support used in soil science to reduce the analytical costs (de Gruijter et al., 2006). For a fixed field work, the MDD$_{\mathrm{C}}$ obtained with physical composites is higher than that obtained with individual samples, mostly due to the higher influence of the measurement errors because the number of measurements is lower, the imperfect materialization, and somewhat owing to the higher t values as consequence of the lower degrees of freedom. However, the inferior capacity (higher MDD$_{\mathrm{C}}$) is a minor problem compared to the substantial cost reduction, as this approach is 12 times cheaper. Assuming an annual linear change in the SOC, the MDD$_{\mathrm{C}}$ of the 72 individual samples would allow to revisit every 5.33 years. Using the composites, the second round must be delayed until 10.01 years, which is when the cumulative change (ΔSOC) equals the MDD$_{\mathrm{C}}$. This difference does not represent a substantial delay, whereas the reduction in analytical costs is important, since is 12 times lower. Thus, the use of composites is a competitive alternative in several scenarios, such as in those in which the economic restrictions are stronger than those imposed by the need to increase the revisitation frequency (Smith, 2004; Necpálová et al., 2014; FAO, 2019).

In comparison with a reference method such as the Walkley-Black for SOC, the NIR spectroscopy is faster and cheaper (Bellon-Maurel and McBratney, 2011; O'Rourke and Holden, 2011; Viscarra Rossel and Brus, 2018), thereby allowing a reduction in the costs necessary to analyse an arbitrary sample size (or an increase in the sample size with an arbitrary budget). However, there should be a balance between these advantages (higher sample size; minor costs) and its major drawback, which is the inferior data quality, that may be compromising and limiting its usefulness. An example of useless predictions includes those obtained at the local scale by using models calibrated with samples collected at the national, continental or global scale, since these predictions are typically biased, as observed in this study when the predictions were obtained using the UNS model. Several researchers have proposed different procedures, strategies and approaches to predict at the local scale; for instance, (i) by calibrating site-specific models (Wetterlind et al., 2010), (ii) by using rs-LOCAL models (Lobsey et al., 2017), (iii) through the adaptation of pre-existing larger-scale models to the local characteristics via model transferring (Grunwald et al., 2018; Padarian et al., 2019), or (iv) by adapting the model to the local conditions through spiking with local samples (Guerrero et al., 2014, 2016), which substantially reduces the bias and enables a successful downscaling. However, to apply these procedures, certain local samples must be analysed using the reference method, which requires an investment. When realising the adaptation through spiking with extra-weighting, the local samples are known as the spiking subset. Spiking ensures that the recalibrated model contains a set of samples with the local characteristics, whereas the extra-weighting forces the model to fit preferentially to that set (Guerrero et al., 2016). This preferential fit is expected to lead to better predictions over the samples with similar characteristics. Consequently, representative samples should be used as the spiking subset owing to their resemblance with the overall prediction set. In fact, Guerrero et al. (2014) compared several types of samples as the optimal spiking subset and noted that the highest accuracy was obtained when the spiking subset was composed of representative samples selected using the Kennard–Stone algorithm, with the bias reduction being the main mechanism. In this study, the spiking subset was constituted by the 12 physical composites, a type of samples which were has not previously tested for this purpose (to the best of our knowledge), and once extra-weighted, the quality of predictions was considerably improved, mostly owing to a decrease in the bias too. The efficiency of composites to act as an effective spiking subset can be explained by the fact that they are formed by mixtures of the samples to be predicted, and hence, they can also be considered as representative samples. Despite the considerable improvement, some bias was still present in the SOC predictions obtained using the SEW, and consequently, the MDD$_{\mathrm{C}}$ was larger than the value achieved using the physical composites. Hence, at first glance, it may appear more reasonable to proceed only with the physical composites (without coupling with NIR) because in addition to a lower MDD$_{\mathrm{C}}$, is a more conservative approach (simpler and less risky than NIR). However, the combination (composites and NIR) is recommended, because in addition to act as spiking subset, the composites allow the measurement of the bias and ∊ΔSOC of the predictions. This is an interesting and powerful analysis (Demattê et al., 2019), which is feasible once the predictions have been averaged as the composites, and then these averages are compared against the values measured in the composites. Hence, if several models are available, such as those generated during the cycles of outliers removal, then the "best" model in terms of the smallest ∊ΔSOC or the smallest bias can be identified. The relevance of this identification is not minor, since the MDD$_{\mathrm{C}}$ obtained using the predictions of the "best" model (SEW-best) was considerably smaller than that obtained using the 12 physical composites analysed using the Walkley–Black method. The unique condition needed to check the bias and ∊ΔSOC is that the total number of individual samples (f) should be bulked into n composites of equal size k (as all those combinations included in Table 1, where f=k·n). This aspect affects the sample size; however, this limitation is minor in comparison to the associated benefits. The possibility of determining the bias of the predictions amplifies the level of credibility and confidence on the spectroscopic models, since typically, the bias is the largest and most commonly observed error when large scale models are predicting at the local scale (downscaling), and it represents a key aspect of the prediction quality. Nevertheless, once the SOC predictions have been obtained in the individual samples and averaged as composites, if the bias and ∊ΔSOC are large, then these NIR predictions can be discarded, and the more conservative approach can be employed to obtain the $\mathrm{MDD_C}$ (i.e., using the SOC values measured on the physical composites with the reference method).

Other spiking subsets, such as those composed by samples selected using the Kennard–Stone algorithm, may be probably more efficient removing the bias than the composites (Guerrero et al., 2014); however, they cannot identify the "best" model, which is an exclusive and particular analysis of the composites when are used as spiking subset. Thus, additional studies are needed to evaluate if that identification is an advantage over other spiking subsets.

In this work, the cost of the national model was not considered, since this task should ideally be addressed by institutions operating at the national scale, for whom such development cost is affordable. Moreover, the scanning cost has not been included, although this cost is justified when maps are also required. One disadvantage of using composites is that maps cannot be created. If no maps are required, and the objective is only to measure the change in the mean, the scanning costs should be considered and included in the evaluation to decide the methodology to be employed. In particular, these costs may be relevant in the case of mid-IR spectroscopy, in which the sample preprocessing is more intensive than that in the case of the NIR spectroscopy (Guillou et al., 2015), since the samples must be preferentially scanned after grinding (whereas in NIR spectroscopy, only sieving is required). Nevertheless, in the spectroscopic approach, the total cost decreases in successive rounds, since a new spiking subset does not need to be included owing to the model is already adapted. In other words, the cost of the model adaptation to local conditions is restricted to the first survey (first round). However, in cases in which the SOC quality is expected to change between rounds, which may be a new source of bias, an additional adaptation, and thus an additional spiking subset may be required. Nevertheless, the efforts required for this additional recalibration are likely less stringent than those in the initial adaptation, because the presence of new minerals or drastic textural changes is not expected. Consequently, the cost efficiency of the NIR spectroscopy is expected to increase along the overall monitoring period. Nevertheless, it may be interesting to perform the analysis of a few composites in the successive monitoring rounds, not to play a role as a spiking subset but to verify that the bias and ∊ΔSOC are within a reasonable range. In this case, once the individual samples have been scanned, they can be bulked into a small (affordable) number of composites. However, using extremely few samples is not advised owing to the influence of the measurement error.

According to the results, the use of the NIR spectroscopy in field conditions does not seem to be an adequate strategy because the errors are considerably higher than those under laboratory conditions. Although scanning directly in the field may facilitate the sample acquisition and thereby increase the sample size, the penalisation owing to the errors may be substantial and hamper any positive net effect. A more careful evaluation must be performed that is not restricted to the throughput. In addition, the increase in the sample size does not help correct the bias, which directly influences the ∊ΔSOC, which is often the most important error. An effective strategy must be robust against any bias. Moreover, an increase in the sample size is only effective within a certain range, beyond which, the benefits are almost ineffectual. Thus, the idea of replacing quality with quantity only holds for a certain number of conditions. The replacement of a few precise measurements with lots of imprecise measurements is only feasible if the bias is small.

We encourage authors to evaluate the capacity of their models in a practical way, focussing on their utility in realistic scenarios, such as in hypothesis testing (as in this study), thereby avoiding evaluations based exclusively on the prediction performance parameters such as R^2, RMSEP, RPD or RPIQ values.

5. Conclusions

We quantified empirically the errors expected during the measurement of the ΔSOC and s_p^2 using NIR spectroscopy and Walkley–Black as the analysis techniques. These errors were used to correct the MDD to obtain the $\mathrm{MDD_C}$. The $\mathrm{MDD_C}$ was always higher than the MDD because these methods involved measurement errors. The lowest $\mathrm{MDD_C}$ was obtained when the SOC contents of the individual samples were analysed using the reference method (Walkley–Black). However, this approach may be unaffordable because the reference method is typically expensive. In a strictly theoretic plane, this problem might be solved by bulking the individual samples into composites, since the number of analyses to be performed is considerably reduced. However, in a realistic plane, although this approach is clearly cheaper, the $\mathrm{MDD_C}$ may be notably increased because of the higher penalisation of the measurement error.

The composites can be used in combination with the NIR spectroscopy to act as a spiking subset to adapt the NIR model to the local conditions. This approach can avoid the generation of excessively biased predictions that are useless due to the high $\mathrm{MDD_C}$. In comparison with the approach using only composites, the combination with the NIR spectroscopy requires an additional effort to scan the individual samples. However, in this combined approach, the composites can also be used to find the spectroscopic model that predicts with the lowest bias and the smallest ∊ΔSOC. In this study, we could identify a model that obtained an $\mathrm{MDD_C}$ value lower than that obtained using only composites. In addition to the net advantage, the predictions in the individual samples retain the information regarding the spatial distribution, which is lost when using only composites. The costs required to adapt the NIR spectroscopy model to the local conditions are probably restricted to the first round; therefore, the cost efficiency is expected to be improved in the successive monitoring rounds. Therefore, the combination of composites and the NIR spectroscopy exhibits several desirable characteristics, such as low cost, high accuracy, and high robustness, which make it an ideal protocol for SOC monitoring (Demattê et al., 2019).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Sincere acknowledgements to the Spanish Ministerio de Economía, Industria y Competitividad, Agencia Estatal de Investigación, and FEDER for the financial support provided to the project "Monitoring changes in SOC with NIR spectroscopy" (ref. CGL2016-78642-R). Authors sincerely acknowledge to Professor Javier Morales for valuable statistical advices during earlier stages of this study. Authors also acknowledge the cooperation of land owners, and to ASALBAC for their help during the search of fields. The authors would like to thank the University of Molise (International and Innovative PhD in Agriculture Technology and Biotechnology) and CREA for covering the research stay of Romina Lorenzetti at Elche, in the frame of a PhD grant.

References

Allen, D.E., Pringle, M.J., Page, K.L., Dalal, R.C., 2010. A review of sampling designs for the measurement of soil organic carbon in Australian grazing lands. Rangeland J. 32, 227–246.

Álvaro-Fuentes, J., Plaza-Bonilla, D., Arrúe, J.L., Lampurlanés, J., Cantero-Martínez, C., 2014. Soil organic carbon storage in a no-tillage chronosequence under

Mediterranean conditions. Plant Soil 376 (1), 31–41. https://doi.org/10.1007/s11104-012-1167-x.

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. Soil Biol. Biochem. 43 (7), 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019.

Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., Pulleman, M., Sukkel, W., van Groenigen, J.W., Brussaard, L., 2018. Soil quality – a critical review. Soil Biol. Biochem. 120, 105–125. DOI: 10.1016/j.soilbio.2018.01.030.

Chappell, A., Baldock, J.A., 2016. Wind erosion reduces soil organic carbon sequestration falsely indicating ineffective management practices. Aeolian Res. 22, 107–116. https://doi.org/10.1016/j.aeolia.2016.07.005.

Chappell, A., Baldock, J.A., Viscarra Rossel, R.A., 2013. Sampling soil organic carbon to detect change over time. CSIRO, Department of Environment. Australia.

Conant, R.T., Paustian, K., 2002. Spatial variability of soil organic carbon in grasslands: implications for detecting change at different scales. Environ. Pollut. 116, S127–S135. https://doi.org/10.1016/S0269-7491(01)00265-2.

Conant, R.T., Smith, G.R., Paustian, K., 2003. Spatial variability of soil carbon in forested and cultivated sites: implications for change detection. J. Environ. Qual. 32 (1), 278–286.

de Gruijter, J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag.

de Gruijter, J.J., McBratney, A.B., Minasny, B., Wheeler, I., Malone, B.P., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130. https://doi.org/10.1016/j.geoderma.2015.11.010.

Demattê, J.A.M., Dotto, A.C., Bedin, L.G., Sayão, V.M., Souza, A.B.E., 2019. Soil analytical quality control by traditional and spectroscopy techniques: constructing the future of a hybrid laboratory for low environmental impact. Geoderma 337, 111–121. https://doi.org/10.1016/j.geoderma.2018.09.010.

FAO, 2019. Measuring and modelling soil carbon stocks and stock changes in livestock production systems: guidelines for assessment (Version 1). Livestock Environmental Assessment and Performance (LEAP) Partnership. Rome, FAO. 170 pp.

Goidts, E., Van Wesemael, B., Crucifix, M., 2009. Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. Eur. J. Soil Sci. 60 (5), 723–739. https://doi.org/10.1111/j.1365-2389.2009.01157.x.

González-Sánchez, E.J., Ordóñez-Fernández, R., Carbonell-Bojollo, R., Veroz-González, O., Gil-Ribes, J.A., 2012. Meta-analysis on atmospheric carbon capture in Spain through the use of conservation agriculture. Soil Till. Res. 122, 52–60. https://doi.org/10.1016/j.still.2012.03.001.

Grunwald, S., Yu, C., Xiong, X., 2018. Transferability and scalability of soil total carbon prediction models in Florida, USA. Pedosphere 28 (6), 856–872. https://doi.org/10.1016/S1002-0160(18)60048-7.

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M., Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. Eur. J. Soil Sci. 65, 248–263.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil Till. Res. 15, 501–509. https://doi.org/10.1016/j.still.2015.07.008.

Guillou F. Le, Wetterlind W., Viscarra Rossel R. A., Hicks W., Grundy M., Tuomi S., 2015. How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon? Soil Res. 53, 913–921. https://doi.org/10.1071/SR15019.

Harcum, J.B., Dressing S.A., 2015. Technical Memorandum #3: Minimum Detectable Change and Power Analysis. Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA. https://www.epa.gov/sites/production/files/2015-10/documents/tech_memo_3_oct15.pdf.

Heim, A., Wehrli, L., Eugster, W., Schmidt, M.W.I., 2009. Effects of sampling design on the probability to detect soil carbon stock changes at the Swiss CarboEurope site Lägeren. Geoderma 149, 347–354. https://doi.org/10.1016/j.geoderma.2008.12.018.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D.D., Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. Sci. Total Environ. 468–469, 376–383. doi:10.1016/j.scitotenv.2013.08.026.

Lal, R., 2004. Soil carbon sequestration to mitigate climate change. Geoderma 123, 1–22.

Lancaster, V.A., Keller-McNulty, S., 1998. A review of composite sampling methods. JASA 93 (443), 1216–1230. https://doi.org/10.1080/01621459.1998.10473781.

Lark, R.M., 2009. Estimating the regional mean status and change of soil properties: two distinct objectives for soil survey. Eur. J. Soil Sci. 60, 748–756.

Lark, R.M., 2012. Some considerations on aggregate sample supports for soil inventory and monitoring. Eur. J. Soil Sci. 63, 86–95.

Lobsey, C.R., Viscarra Rossel, R.A., Roudier, P., Hedley, C.B., 2017. rs-local data-mines information from spectral libraries to improve local calibrations. Eur. J. Soil Sci. 68 (6), 840–852. https://doi.org/10.1111/ejss.12490.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86. DOI: 10.1016/j.geoderma.2017.01.002.

Necpálová, M., Anex, R.P., Jr., Kxravchenko, A.N., Abendroth, L.J., Grosso, S.J.D., Dick, W.A., Helmers, M.J., Herzmann, D., Lauer, J.G., Nafziger, E.D., Sawyer, J.E., Scharf, P.C., Strock, J.S., Villamil, M.B., 2014. What does it take to detect a change in soil carbon stock? A regional comparison of minimum detectable difference and experiment duration in the north central United States. J. Soil Water Conserv. 69, 517–531. DOI: 10.2489/jswc.69.6.517.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E.B., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E. K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Adv. Agron. 132, 139–159.

O'Rourke, S., Holden, N., 2011. Optical sensing and chemometric analysis of soil organic carbon - a cost effective alternative to conventional laboratory methods? Soil Use Manage. 27, 143–155.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma 340, 279–288. https://doi.org/10.1016/j.geoderma.2019.01.009.

Patil, G.P., Gore, S.D., Taillie, C., 2011. Composite Sampling: A Novel Method to Accomplish Observational Economy in Environmental Studies. Springer US. ISBN 978-1-4419-7627-7.

Petrokofsky, G., Kanamaru, H., Achard, F., Goetz, S.J., Joosten, H., Holmgren, P., Lehtonen, A., Menton, M., Pullin, A.S., Wattenbach, M., 2012. Comparison of methods for measuring and assessing carbon stocks and carbon stock changes in terrestrial carbon pools. How do the accuracy and precision of current methods compare? A systematic review protocol. Environ. Evid. 1, 6. https://doi.org/10.1186/2047-2382-1-6.

Rawlins, B.G., Scheib, A.J., Lark, R.M., Lister, T.R., 2009. Sampling and analytical plus subsampling variance components for five soil indicators observed at regional scale. Eur. J. Soil Sci. 60 (5), 740–747. https://doi.org/10.1111/j.1365-2389.2009.01159.x.

Saby, N.P.A., Bellamy, P.H., Morvan, X., Arrouays, D., Jones, R.J.A., Verheijen, F.G.A., Kibblewhite, M.G., Verdoodt, A., Üveges, J.B., Freudenschuß, A., Simota, C., 2008. Will European soil-monitoring networks be able to detect changes in topsoil organic carbon content? Global Change Biol. 14 (10), 2432–2442.

Schrumpf, M., Schulze, E.D., Kaiser, K., Schumacher, J., 2011. How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories? Biogeosciences 8, 1193–1212.

Smith, P., 2004. How long before a change in soil organic carbon can be detected? Global Change Biol. 10, 1878–1883.

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D.P., Batjes, N.H., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J.E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., Klumpp, K., 2020. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. Global Change Biol. 26 (1), 219–241. DOI: 10.1111/gcb.14815.

Viscarra Rossel, R.A., Brus, D.J., 2018. The cost-efficiency and reliability of two methods for soil organic C accounting. Land Degrad. Dev. 29 (3), 506–520. https://doi.org/10.1002/ldr.2887.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Sci. Rev. 155, 198-230. DOI: 10.1016/j.earscirev.2016.01.012.

Wetterlind, J., Stenberg, B., Söderström, M., 2010. Increased sample point density in farm soil mapping by local calibration of near infrared prediction models. Geoderma 156 (3–4), 152–160. https://doi.org/10.1016/j.geoderma.2010.02.012.

Zimmerman, D.W., Zumbo, B.D., 2015. Resolving the issue of how reliability is related to statistical power: adhering to mathematical definitions. JMASM 14 (2), Article 5. https://doi.org/10.22237/jmasm/1446350640.

Zimmerman, D.W., Williams, R.H., Zumbo, B.D., 1993. Reliability of measurement and power of significance tests based on differences. Appl. Psychol. Meas 17 (1), 1–9. https://doi.org/10.1177/014662169301700101.