# Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library

Cécile Gomez, Tiphaine Chevallier, Patricia Moulin, Imane Bouferra, Kaouther Hmaidi, Dominique D. Arrouays, Claudy C. Jolivet, Bernard Barthès

# Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library

Cécile Gomez [1,2], Tiphaine Chevallier [3], Patricia Moulin [4], Imane Bouferra [3], Kaouther Hmaidi [3,5], Dominique Arrouays [6], Claudy Jolivet [6], Bernard G. Barthès [3]

[1] LISAH, University of Montpellier, INRAE, IRD, Montpellier SupAgro, 34060 Montpellier, France. Corresponding author. cecile.gomez@ird.fr

[2] Indo-French Cell for Water Sciences, IRD, Indian Institute of Science, Bangalore 560012, India

[3] Eco&Sols, University of Montpellier, CIRAD, INRAE, IRD, Montpellier SupAgro, 34060 Montpellier, France

[4] US IMAGO, IRD, BP1386, Dakar, Senegal

[5] UR Pédologie, Faculté des Sciences de Tunis, El Manar Tunis, Tunisia

[6] INRAE, US 1106 InfoSol, F45000, Orléans, France

**Keywords:** Soil Organic and Inorganic Carbon; Mid-Infrared reflectance spectroscopy; natural logarithm transformation; Local and Global calibration; Partial least squares regression (PLSR); National dataset.

## Abstract

Mid-infrared reflectance spectroscopy (MIRS, 4000–400 $cm^{-1}$) is being considered to provide accurate estimations of soil properties, including soil organic carbon (SOC) and

soil inorganic carbon (SIC) contents. This approach has mainly been demonstrated by using datasets originating from the same area *A*, with similar geopedological conditions, to build, validate and test prediction models. The objective of this study was to analyse how MIRS performs when applied to predict SOC and SIC contents, from a calibration database collected over a region *A*, to predict over a region *B,* where *A* and *B* have no common area and different soil and climate conditions. This study used a French MIRS soil dataset including 2178 topsoil samples to calibrate SIC and SOC prediction models with partial least squares regression (PLSR), and a Tunisian MIRS topsoil dataset including 96 soil samples to test them. Our results showed that when using the French MIRS soil database, *i)* the SOC and SIC of French validation samples were successfully predicted using global models ($R^2_{val}$ = 0.88 and 0.98, respectively), *ii)* the SIC of Tunisian samples was also predicted successfully both using a global model and using a selection of spectral neighbours from the French calibration database ($R^2_{test}$ of 0.96 for both), *iii)* the SOC of Tunisian samples was predicted moderately by global model ($R^2_{test}$ of 0.64) and a transformation by natural logarithm of the calibration SOC values significantly improved the SOC prediction of Tunisian samples ($R^2_{test}$ of 0.97), and *iv)* a transformation by natural logarithm of SOC values provided more benefit than a selection of spectral neighbours from the French calibration database for predicting Tunisian SOC values. Therefore, in the future, MIRS might replace conventional physico-chemical analysis techniques, or at least be considered as an alternative technique, especially when optimally exhaustive calibration databases will become available.

## 1. Introduction

Soil is the largest reservoir of continental carbon and participates in the global carbon cycle (Jacobson et al., 2000; Scharlemann et al., 2014). Soil emits carbon dioxide ($CO_2$) through autotrophic and heterotrophic respiration and acts as a sink of atmospheric $CO_2$

through photosynthesis; then, organic decomposition products are integrated into the soil as organic matter, which is composed of approximately 50% to 58% of carbon (Gregorich et al., 1994; Pribyl, 2010). Moreover, soil organic carbon (SOC) has a long-acknowledged and key role in soil physical, chemical and biological fertility (Reeves, 1997). Thus, understanding the dynamics of soil carbon is a major issue for soil fertility and for climate change mitigation.

Soil carbon is not only found in organic form. At the global scale, approximately one-third of the total soil carbon is inorganic (soil inorganic carbon, SIC) (Batjes, 1996), and calcareous soils cover more than 30% of the earth's land surface (Chen and Barak, 1982; Romanyà and Rovira, 2011). SIC is mainly in the form of calcium carbonate ($CaCO_3$), and high SIC contents are often localised in dry areas where SOC stocks are low. In these soils, SIC levels can be 2 to 10 times higher than SOC levels (Bernoux and Chevallier, 2014). SIC is made of primary minerals derived from the fragmentation of carbonate bedrock (lithogenic carbonates) or secondary minerals from inorganic carbon precipitation in soil pores or around roots (pedogenic carbonates). Pedogenic carbonates have various forms—nodules, lamellae or crystals—and have varying solubility.

Several analytical methods have been developed to quantify SIC and SOC contents in soils, but they are tedious and/or costly. In particular, SOC determination in carbonated soils often requires hazardous reagents. SIC content has usually been measured by calcimetry (ISO, 1995b) but can also be measured by dry combustion with a CNH elemental analyser equipped with a specific module ($CO_3$-C module) after phosphoric acid dissolution of the SIC (Mc Crea, 1950; Hannam et al. 2016). Quantifying SOC in calcareous soils has either been carried out directly, by wet oxidation (Walkley and Black, 1934) or dry combustion after removing SIC by acid pretreatment (Harris et al., 2001), or indirectly, by subtracting the SIC, measured by calcimetry, from the total carbon content determined by dry combustion. Due to indirect determination, incomplete oxidation and the

use of hazardous reactants with both of these methods, alternative methods based on the thermal lability of SIC and SOC have also been tested (Wang et al., 2012; Apesteguia et al., 2018).

For one to two decades, visible near-infrared (Vis-NIR, 400–2500 nm) and mid-infrared reflectance spectroscopy (MIRS, 4000–400 $cm^{-1}$), which measures diffuse reflectance, have been proposed as alternative methods to these physico-chemical analytical methods (e.g., Viscarra Rossel et al., 2006; Cécillon et al., 2008; Bellon-Maurel and McBratney, 2011). MIRS is based on the study of absorption bands corresponding to fundamental molecular vibrations, and NIRS is based on the study of absorption bands corresponding to overtones and combinations of fundamental vibrations (Williams and Norris, 1987). Several studies highlighting the potential of Vis-NIR and/or MIRS for predicting various soil attributes, including SIC and SOC, have been listed by Viscarra Rossel et al. (2006) and then by Soriano-Disla et al (2014).

In the MIR range, carbonates may be identified by strong and numerous absorption bands, for instance, bands at approximately 820-750 $cm^{-1}$, 1800 $cm^{-1}$, 2520 $cm^{-1}$ and 2900-2990 $cm^{-1}$ (e.g., Tatzber et al., 2007; Du and Zhou, 2009, Comstock et al., 2019). In the NIR range, carbonates may be identified by peaks at 2341 and 2480 nm (Lagacherie et al., 2008; Barthès et al., 2016). In the MIR range, organic carbon may also be identified by numerous absorption peaks, for instance, peaks at approximately 2920 and 1230 $cm^{-1}$ (Grinand et al., 2012). In the NIR range, organic carbon may be identified by absorption peaks at approximately 1910 nm (Viscarra Rossel et al., 2006; Viscarra Rossel and Webster, 2012) and 2050–2150 nm (Workman and Weyer, 2008). Finally, MIRS has generally been reported to provide more accurate performance in terms of SOC and SIC predictions compared to NIRS (e.g., McCarty et al., 2002; Reeves, 2010; Bellon-Maurel & McBratney, 2011; Clairotte et al., 2016). Nevertheless, better NIRS than MIRS predictions of SOC have been reported in tropical and Mediterranean regions due to the overlap of

104 absorption regions related to metal oxides and organic compounds (Rabenarivo et al.,
105 2013; Barthès et al., 2016).

106 Following the emergence of infrared spectroscopy technologies for soil
107 characterisation, soil spectral libraries covering extensive areas have recently been
108 developed for estimating soil properties, especially SOC; most libraries are in the Vis-NIR
109 range at the national scale (e.g., in Australia, Viscarra Rossel and Webster, 2012; in
110 Denmark, Knadel et al., 2012; in France, Gogé et al., 2014; in China, Shi et al., 2015),
111 continental scale (in Europe, Stevens et al., 2013) and even global scale (Brown et al.,
112 2006; Viscarra Rossel et al., 2016), but libraries also exist in the MIR range (e.g., in
113 France, Grinand et al., 2012, and Clairotte et al., 2016; in the US, Wijewardane et al.,
114 2018, Comstosck et al., 2019 and Dangal et al., 2019). All these libraries use spectra
115 collected using dried and ground samples in laboratory conditions.

116 Most of the studies dealing with large soil spectral databases (national or
117 continental) have aimed to calibrate prediction models with samples from a region *A* to
118 predict the properties of soil samples from the same region *A*. Therefore, the soil and
119 climate conditions are similar between the calibration and validation databases. McCarty et
120 al. (2002) calibrated SOC and SIC prediction models by using two-thirds of 257 soil
121 samples collected from 14 geographically diverse locations over eight states in the west
122 central US and validated the models by using the remaining one-third of soil samples,
123 obtaining good accuracy and low bias. Stevens et al. (2013) used the LUCAS database,
124 which includes samples from 23 member states of the European Union, to predict SOC. In
125 their study, each calibration dataset and associated validation dataset was composed of
126 samples from a similar soil type (organic or mineral) and land use (cropland, grassland or
127 woodland), and the most spectrally representative samples were selected as calibration
128 samples. Shi et al. (2015) used a Chinese database, which included samples from 20
129 provinces, to predict SOC. In their study, the calibration dataset was selected to minimise

both the spectral distance and geographical distance to each validation sample, which also belonged to the database (region *A*). Clairotte et al. (2016) successfully calibrated prediction models by using subsets of the French national database and then tested these models on a test subset (10% of this French database, either spectrally representative or not; samples were not selected based on their pedo-climatic context).

Some studies have focused on potential and limitation analysis of prediction models calibrated with a large database composed of samples collected over a region *A* to predict properties for soil samples collected from a small region *b* within *A*. McCarty et al. (2002) calibrated SOC and SIC prediction models by using 257 soil samples collected from 14 geographically diverse locations over 8 states in the west central US and obtained biased predictions when their models were applied to 16 independent soil samples collected in another state (Nebraska) that was also in the west central US. Gogé et al. (2014) calibrated local prediction models from the French national database, collected over 550 000 km² (region *A*), to predict soil properties in soil samples collected from a small French area of 24 km² (Occitanie region, south of France; this region *b* was included in *A* but under-represented). Additionally, Comstosck et al. (2019) calibrated prediction models by using a US national database (region *A*) to predict carbonate in soil samples collected from two states (New York and Iowa, regions *b*) that were poorly represented in the US database.

Finally, to the best of our knowledge, few studies have focused on potential and limitation analysis of prediction models calibrated with a large database composed of samples collected over a region *A* to predict properties for soil samples collected over a region *B,* where *A* and *B* have no common area and hence have potential differences in soil and climatic conditions. Only Jauss et al. (2017) and Ahmed et al. (2017) used MIR spectroscopy as a routine method for predicting pyrogenic carbon and for predicting Total

Carbon and SOC, respectively, on United States soils (region *B*), from a soil spectral database of Australia (region *A*).

The objective of this study was to analyse how MIRS may be used to predict SOC and SIC contents when using a national database collected over a region *A* to predict values for soil samples collected over a region *B,* where *A* and *B* have no common area and have different soil and climate conditions. This study used a French MIRS soil dataset (region *A*) including 2178 topsoil samples collected from an area of 550 000 km² (French metropolitan territory, composed of temperate and Mediterranean soils) to calibrate SIC and SOC prediction models. These models were tested on a Tunisian MIRS soil dataset (region *B*), including 96 soil samples collected from an area of approximately 80 000 km² (northern half of Tunisia, mainly Mediterranean and arid soils).

## 2. Materials and methods

### 2.1. Soil datasets

#### 2.1.1. The French national soil collection

The national soil collection provided by the French national soil quality monitoring network (RMQS; Arrouays et al., 2002) and called *DB_RMQS* was used in this study to calibrate the SIC and SOC models. This RMQS collection is composed of 2178 soil samples representing all main soil types encountered over the sampled 552 000 km² of the French metropolitan territory (Corsica included): Cambisols, Calcosols, Luvisols, Leptosols, Andosols, Albeluvisols, Podzosols, *etc.* (IUSS Working Group WRB, 2014). The latitude of sample sites ranges from 41 to 51°N, and their longitude ranges from 5.0°W to 9.5°E. The sampling design was based on a square grid with 16-km spacing. At the centre of each square, 25 individual core samples were taken from 0 to 30 cm depth using an unaligned

sampling design within a 20 × 20 m area. Core samples were bulked to obtain a composite sample for each site (Arrouays et al., 2002).

#### 2.1.2. The Tunisian soil samples

Ninety-six soil samples were used as the test set, called *DB_Tunisia*. These samples were collected from 45 localities, covering approximately 80 000 km² (from 35 to 37°N and 08 to 11°E), with the aim of representing the main soil types and land uses of the northern half of Tunisia. This was done based on previous studies carried out at the Tunis El Manar University, without particular design. Field samples within the same locality were kilometres apart and under different land uses. Soil samples were collected at 0-10 cm using a spade, and the sampling campaign was carried out within a few months in late 2010. This Tunisian set was previously studied in Barthès et al. (2016).

### 2.2. Laboratory Analysis

#### 2.2.1. Physico-chemical analyses

The 2178 RMQS samples were air dried, 2-mm sieved and then finely ground (< 0.25 mm) using mortar and pestle. The SIC content of the RMQS samples was calculated as 0.12 times the soil calcium carbonate content, which was determined using these finely ground (< 0.25 mm) air-dried samples using a Bernard calcimeter according to the standard procedure ISO 10693 (ISO, 1995a). The carbonate content was calculated after calibration with a pure calcium carbonate standard and was expressed as the equivalent calcium carbonate content. The SIC content of the RMQS samples (*DB_RMQS_SIC*) ranged from 0 to 103.9 g $kg^{-1}$, averaged 6.4 g $kg^{-1}$, and had a median of 0 g $kg^{-1}$ and a skewness value close to 3.1 (Table 1).

Then, the SOC content of the RMQS samples was calculated as the difference between the total carbon (TC) and inorganic carbon contents. The TC content was

206 determined by dry combustion with an elemental analyser (Thermo Fisher Scientific CHN
207 NA2000, Waltham, MA, US) using approximately 25–30-mg aliquots of finely ground
208 (< 0.25 mm) air-dried soil samples that were sealed into tin capsules, according to the
209 standard procedure ISO 10694 (ISO, 1995b). The SOC content of the RMQS samples
210 (*DB_RMQS_SOC*) ranged from 0.6 to 411.3 g $kg^{-1}$, averaged 25.8 g $kg^{-1}$, and had a
211 median of 19.6 g $kg^{-1}$ and a skewness value close to 4.9 (Table 1).
212
213 [Table 1]
214
215
216 The 96 Tunisian soil samples were also air-dried, sieved to 2 mm and then finely ground
217 (< 0.2 mm) using mortar and pestle. The SOC content of the Tunisian samples was
218 analysed by dry combustion after decarbonisation using chlorhydric acid, following the
219 standard procedure ISO 10694 (1995b), with the same elemental analyser as that used for
220 RMQS samples but using silver capsules. Soils were decarbonated prior to SOC
221 determination: 10 mL of water were added to 1 g of soil and 0.5 M HCl solution was then
222 dripped onto the sample until there was no more effervescence; then the samples were
223 washed in water until pH reached 7. The SOC concentration was then determined on
224 finely ground 25–30 mg aliquots by dry combustion using an elemental analyser (Thermo
225 Fisher Scientific CHN NA2000, Waltham, MA, USA). The SOC content of the Tunisian
226 samples (*DB_Tunisia_SOC*) ranged from 2.0 to 121.0 g $kg^{-1}$, averaged 20.1 g $kg^{-1}$, and
227 had a median of 14.6 g $kg^{-1}$ and a skewness value close to 3 (Table 1).
228 The soil inorganic carbon content of the Tunisian samples was calculated as the
229 difference between the TC (determined by dry combustion using the same CHN analyser
230 as that used for the RMQS samples) and SOC contents. The SIC content of the Tunisian

samples (*DB_Tunisia_SIC*) ranged from 0.0 to 92.9 g $kg^{-1}$, averaged 43.3 g $kg^{-1}$, and had a skewness value close to -0.2 (Table 1).

As the SIC contents of the RMQS and Tunisian samples were analysed by two different methods, 29 test samples from the Tunisian set (i.e., 30% of the set) were re-analysed following the same method as that used for the RMQS samples (i.e., directly, by calcimetry; ISO, 1995a). The Pearson correlation coefficient (R), root mean square error (RMSE) and bias between the SIC values determined by both approaches (calcimetry vs. difference between TC and SOC) were 0.997, 2.4 g $kg^{-1}$ and 1.1 g $kg^{-1}$, respectively (Figure 1). The values of SIC calculated by both approaches could thus be considered equivalent.

[Figure 1]

#### 2.2.2. Mid-infrared spectroscopy

Mid-infrared spectroscopic analysis was performed following the same procedure for both spectral libraries. First, air-dried, 2-mm sieved, and 0.2-mm ground samples were oven-dried at 40°C for twelve hours. Reflectance spectra were acquired using a Fourier transform Nicolet 6700 spectrophotometer (Thermo Fischer Scientific, Madison, WI, US) in the MIR region. Reflectance was acquired at 934 wavenumbers between 4000 and 400 $cm^{-1}$ with a 3.86 $cm^{-1}$ spectral resolution. This spectrophotometer is equipped with a silicon carbide source, a Michelson interferometer as a dispersive element, and a deuterated triglycine sulfate detector. Soil samples were placed in a 17-well plate. The soil surface was flattened with the flat section of a glass cylinder, and samples were then scanned using an auto-sampler (soil surface area scanned: ca. 10 $mm^2$). Each spectrum resulted from 32 co-added scans, and the body of the plate (next to the wells) was used as

a reference standard and scanned once per plate (i.e., every 17 samples). Twenty wavenumbers were removed due to frequent noise in the spectrum, so MIR spectra in the range from 4000 to 478 $cm^{-1}$, with 914 wavenumbers, were used.

### 2.3 PLSR model calibration

All procedures were performed using R software (R Core Team, 2012), and both the ade4 (Dray and Dufour, 2007) and pls packages (Mevik and Wehrens, 2007) were used.

#### 2.3.1. Dataset preparation

Both *DB_RMQS_SIC* and *DB_RMQS_SOC* were divided into a calibration set (3/4 of the dataset) and a validation set (1/4 of the dataset). The samples of each dataset were ranked according to ascending reference value (observed SIC or SOC). The sample with the lowest reference value was put in the calibration set, the next sample was put in the validation set, and then the next three samples were put in the calibration set. The procedure was continued by alternately placing the next sample in the validation set and the following three samples in the calibration set. Following this process, the distributions of the *DB_Calib_RMQS_SOC* calibration and *DB_Valid_RMQS_SOC* validation datasets were similar. As well, the distributions of the *DB_Calib_RMQS_SIC* calibration and *DB_Valid_RMQS_SIC* validation datasets were similar.

As the SOC values of the RMQS dataset followed a non-normal distribution (Table 1), the SOC values of the *DB_Calib_RMQS_SOC* calibration dataset were transformed with natural logarithm (ln(SOC)) to reach a normal distribution, giving rise to new dataset for calibration *DB_Calib_RMQS_lnSOC*. The SIC distribution in the French dataset was also non normal, but due to the very large number of null values, ln-transformation was hardly possible (Bellon-Maurel et al., 2010; Terra et al., 2015).

The reflectance was converted into "absorbance" ($\log_{10}$ [1/reflectance]), and a standard normal variate correction was applied to remove additive and multiplicative effects (Barnes et al., 1989). Spectral outliers, which are defined as samples spectrally different than the rest of the samples (e.g., Pearson, 2002), were removed from the calibration datasets. The spectral outliers were identified by applying the Mahalanobis distance (Mark and Tunnell, 1985) to data condensed by principal component analysis (PCA). In the present study, a Mahalanobis distance of 3.5 was selected as the threshold for the identification of spectral outliers.

### 2.3.2. Partial least squares regression

Partial least squares regression (PLSR) is a multivariate approach that specifies a linear relationship between a dependent (response) variable (Y-variable, i.e., SIC or SOC content in the present case), and a set of predictor variables (X-variables, i.e., MIR spectra in the present case; Tenenhaus, 1998). The general concept of PLSR is to extract a small number of orthogonal variables (called the latent variables) that account for the maximum variation in the X-variables. A detailed description of the PLSR procedure can be found in Wold et al. (2001). This method is commonly used for NIRS or MIRS prediction of soil properties (e.g., Viscarra-Rossel et al., 2006; Bellon-Maurel et al, 2010).

The maximum number of latent variables of PLSR was defined as 30. A leave-one-out cross-validation (LOOCV) procedure was adopted to verify the prediction capability of the PLSR model for the calibration set. Each time, n – 1 samples were used to build a regression model, which was applied to the sample not used in developing the model. This procedure was repeated for all n samples, resulting in predictions for all n samples.

### 2.3.3. Global models

Global calibration is a common calibration procedure where all calibration samples are used to build a unique prediction model that is applied identically to all validation or test samples. One global prediction model (denoted $GM_{SIC}$) was built for SIC prediction based on *DB_Calib_RMQS_SIC,* validated on *DB_Valid_RMQS_SIC* and tested on *DB_Tunisia_SIC.* As well, a global prediction model (denoted $GM_{SOC}$) was built for SOC prediction based on *DB_Calib_RMQS_SOC*, validated on *DB_Valid_RMQS_SOC* and tested on *DB_Tunisia_SOC*. Finally, a global prediction model (denoted $GM_{lnSOC}$) was built for SOC prediction based on *DB_Calib_RMQS_lnSOC,* applied to spectra of *DB_Valid_RMQS_SOC* and *DB_Tunisia_SOC*, and the output predictions ln(SOC) were back-transformed into SOC values using exp(ln(SOC)).

The optimal number of latent variables of $GM_{SIC}$, $GM_{SOC}$ and $GM_{lnSOC}$ was determined using prediction residual error sum of squares (PRESS) analysis of LOOCV results to avoid under- and over-fitting. Then, all calibration samples were used to build the prediction model with the appropriate number of latent variables, and this model was applied to validation and test sets.

### 2.3.4. Local models

A local regression approach was implemented based on PLSR to predict the SOC and SIC content of Tunisian samples. Given a sample $p_i$ from *DB_Tunisia* to predict:

1- The Pearson coefficient of correlation between the spectrum of the Tunisian sample $p_i$ and each RMQS calibration spectrum (from *DB_Calib_RMQS)* was calculated;

2- The *N* samples from *DB_Calib_RMQS* with spectra that correlated to the spectrum of $p_i$ beyond a cut-off value of 0.95 were considered spectral neighbours of the Tunisian sample $p_i$, without maximum limit for *N*.

3- A PLSR model was built using the *N* spectral neighbours of the Tunisian sample $p_i$.

If a Tunisian soil sample had less than 30 spectral neighbours among the *DB_Calib_RMQS* set, this soil sample was not predicted.

One local prediction model (denoted $LM_{SIC}$) was built for SIC prediction based on *DB_Calib_RMQS_SIC,* validated on *DB_Valid_RMQS_SIC* and tested on *DB_Tunisia_SIC.* One local prediction model (denoted $LM_{SOC}$) was built for SOC prediction based on *DB_Calib_RMQS_SOC,* validated on *DB_Valid_RMQS_SOC* and tested on *DB_Tunisia_SOC*. Finally, one local prediction model (denoted $LM_{lnSOC}$) was built for SOC prediction based on *DB_Calib_RMQS_lnSOC,* applied to spectra of *DB_Valid_RMQS_SOC* and *DB_Tunisia_SOC*, and the output predictions ln(SOC) were back-transformed into SOC values using exp(ln(SOC)).

As the calibration sets for SOC and SIC predictions did not include the same samples (*DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_SIC*, respectively), the nearest calibration neighbours of a given French validation and Tunisian sample were not the same for SOC and SIC predictions. The optimal number of latent variables was finally determined using PRESS analysis of LOOCV on the selected spectral neighbours to avoid under- and over-fitting. Regardless of the type of local model, the spectral outliers were not investigated because the selection of nearest neighbours was considered an implicit rejection of outliers.

### 2.4 PLSR model evaluation

The performance of global models was evaluated according to figures of merit described in Bellon Maurel et al. (2010), from cross-validation, validation and test databases.

The coefficient of determination of cross-validation ($R^2_{cv}$) and root mean square error of cross-validation (*RMSECV*) for *DB_Calib_RMQS* were used. $R^2_{cv}$ was computed as 1-ESS/TSS, where ESS is the error sum of squares and TSS the total sum of squares.

The coefficient of determination and root mean square error of prediction for *DB_Valid_RMQS*, $R^2_{val}$ and $RMSE_{val}$ respectively, were used. $R^2_{val}$ was also computed as 1-ESS/TSS. The ratio of performance to deviation in *DB_Valid_RMQS* ($RPD_{val}$), which is the ratio between the standard deviation in *DB_Valid_RMQS* and $RMSE_{val}$, was calculated. The ratio of performance to interquartile range of *DB_Valid_RMQS* ($RPIQ_{val}$), which is the ratio between interquartile range (difference between the third and first quartiles) of *DB_Valid_RMQS* and $RMSE_{val}$, was also calculated. This parameter has been proposed for variables with non-normal distributions (Bellon-Maurel et al., 2010). And the bias, which is the mean difference between observations and predictions, was calculated for *DB_Valid_RMQS* ($bias_{val}$).

The coefficient of determination and root mean square error of prediction for *DB_Tunisia*, $R^2_{test}$ and $RMSE_{test}$ respectively, were used. $R^2_{test}$ was computed as 1-ESS/TSS. The ratio of performance to deviation in *DB_Tunisia* ($RPD_{test}$), which is the ratio between the standard deviation in *DB_Tunisia* and $RMSE_{test}$, was calculated. The ratio of performance to interquartile range of *DB_Tunisia* ($RPIQ_{test}$), which is the ratio between the interquartile range of *DB_Tunisia* and $RMSE_{test}$, was also calculated. And the bias, which is the mean difference between observations and predictions, was calculated for *DB_Tunisia* ($bias_{test}$).

Finally, a wavelength was considered a significant contributor in a global model when the values of both the regression coefficient and variable importance in the projection (VIP) were sufficiently large: the threshold for the VIP was set to 1 (Chong and Jun, 2005; Wold et al., 1993, 2001), and the thresholds for the regression coefficients were their standard deviations (Viscarra-Rossel et al., 2008).

The performances of local models were based on the same figures of merit as those used in global calibration, calculated on *DB_Valid_RMQS and DB_Tunisia.*

382 Concerning the models built from the *DB_Calib_RMQS_lnSOC,* independent validation
383 and test statistics were calculated from back-transformed data.
384
385
386 **4. Results**
387 **4.1. Preliminary analysis of soil properties and spectra**
388 The RMQS soil sampling covered the French territory, and ranges of SOC and SIC
389 contents in the *DB_RMQS_SOC* and *DB_RMQS_SIC*, respectively, are large (Table 1).
390 According to the French soil classification, 33 soil reference groups were sampled, with a
391 dominance of Cambisols (IUSS Working Group WRB, 2014; 27% of the sample set),
392 calcareous soils (Calcosols, 22%) and Luvisols (16%). High SIC values are mainly located
393 in three French areas: 1) the southeast (Prealps), mainly with Leptosols and Calcosols, 2)
394 the northeast (chalk Champagne) also mainly with Leptosols, and 3) a transect from west
395 (the Aquitanian Basin) to south (Mediterranean Sea), mainly with Calcosols (Figure 2A1).
396 High SOC values are mainly located in 1) mountain areas (Alps in the southeast, Pyrenees
397 in the extreme southwest, Massif Central in the south-centre, Jura in the centre-east), 2)
398 cool regions covered by forests and pastures (centre-east), and 3) intensive livestock
399 production areas (northwest; Figure 2B1).
400 The Tunisian soil samples covered the northern half of Tunisian territory, and the
401 SIC contents range in the *DB_Tunisia_SIC* is as large as the one in the *DB_RMQS_SIC*,
402 whereas the SOC contents range in the *DB_Tunisia_SOC* is lower than the one of the
403 *DB_RMQS_SOC* (Table 1). The sampled Tunisian soils were mainly Calcaric Cambisols
404 and Regosols, Kastanozems, and Chromic and Vertic Cambisols.
405
406 [Figure 2]
407

Principal component analyses were performed on pre-treated spectra of *DB_Calib_RMQS_SIC* and *DB_Calib_RMQS_SOC*, respectively, and pre-treated Tunisian spectra were projected onto the plans made by the first and second components. Most Tunisian spectra overlapped a subset of RMQS spectra for SIC (Figure 3a) and for SOC (Figure 3b). So most Tunisian soil samples had similar spectral signatures than a subset of RMQS spectra used for calibrating prediction models.

[Figure 3]

### 4.2. Global models

#### a. Soil inorganic carbon content

For SIC prediction, 52 spectral outliers were identified within the initial calibration dataset, so 1582 RMQS samples were ultimately kept and constituted the *DB_Calib_RMQS_SIC* dataset. The SIC content of these 1582 RMQS samples contained in *DB_Calib_RMQS_SIC* ranged from 0.0 to 103.9 g $kg^{-1}$, averaged 6.4 g $kg^{-1}$, and had a skewness value close to 3.1 (Table 1). The SIC content of the 544 RMQS samples contained in *DB_Valid_RMQS_SIC* ranged from 0.0 to 95.2 g $kg^{-1}$, averaged 6.3 g $kg^{-1}$, and also had a skewness value close to 3.1 (Table 1).

The $GM_{SIC}$ was built from the *DB_Calib_RMQS_SIC* dataset using an optimal number of 15 latent variables, validated on the *DB_Valid_RMQS_SIC* dataset and then tested on the *DB_Tunisia_SIC* dataset. The performance of the $GM_{SIC}$ prediction model was accurate, with an $R^2_{cv}$ of 0.97 and *RMSECV* of 2.8 g $kg^{-1}$ in the calibration step (Table 2) and an $R^2_{val}$ of 0.98 and $RMSE_{val}$ of 2.1 g $kg^{-1}$ in the validation step (Table 2, Figure 4b). When applied to the *DB_Tunisia_SIC* dataset, this $GM_{SIC}$ prediction model provided accurate and unbiased predictions ($R^2_{test}$ = 0.96, $RMSE_{test}$ = 5.2 g $kg^{-1}$ and $bias_{test}$ = 0.2 g $kg^{-1}$; Table 3, Figure 4c).

A total of 96 spectral bands might be considered significant based on the analysis of the VIP and regression coefficients of $GM_{SIC}$ (Figure 4d). Among the 96 significant spectral bands, those at approximately 2500, 1800 and 860 $cm^{-1}$ had regression coefficients higher than 3 times the standard deviation and therefore might be considered the most significant ones.

[Figure 4]

[Table 2]

[Table 3]

**b. Soil organic carbon content**

For SOC prediction, 48 spectral outliers were identified within the initial calibration dataset, so 1586 RMQS samples were ultimately kept and constituted the *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets. The SOC contents of the 1586 RMQS samples of *DB_Calib_RMQS_SOC* ranged from 0.6 to 411.3 g $kg^{-1}$, averaged 25.6 g $kg^{-1}$, and had a skewness value close to 5.5 (Table 1). The ln(SOC) values of the 1586 RMQS samples of *DB_Calib_RMQS_lnSOC* ranged from -0.5 to 6 ln(g $kg^{-1}$), averaged 3 ln(g $kg^{-1}$), had a median of 3 ln(g $kg^{-1}$) and a skewness value close to 0.3. The SOC content of the 544 RMQS samples contained in *DB_Valid_RMQS_SOC* ranged from 1.5 to 159 g $kg^{-1}$, averaged 25.5 g $kg^{-1}$, and had a skewness value close to 2.5 (Table 1).

The $GM_{SOC}$ was built from the *DB_Calib_RMQS_SOC* dataset using 23 latent variables, validated on the *DB_Valid_RMQS_SOC* dataset and then tested on the *DB_Tunisia_SOC* dataset. The performance of the $GM_{SOC}$ prediction model was modest, with an $R^2_{cv}$ of 0.80 and *RMSECV* of 9.9 g $kg^{-1}$ in the calibration step (Table 2) and an $R^2_{val}$

460 of 0.88 and $RMSE_{val}$ of 7.2 g kg$^{-1}$ in the validation step (Table 2, Figure 5b). When applied
461 to the Tunisian test set, this $GM_{SOC}$ prediction model provided low accuracy ($R^2_{test}$ = 0.64,
462 $RMSE_{test}$ = 16.0 g kg$^{-1}$) and biased predictions ($bias_{test}$ = -5.2 g kg$^{-1}$) (Table 3, Figure 5c).
463 A total of 92 spectral bands might be considered significant based on the analysis of
464 the VIP and regression coefficients of $GM_{SOC}$ (Figure 5d). Nevertheless, among these 92
465 spectral bands, none was associated to very high regression coefficients.
466
467 [Figure 5]
468
469 A $GM_{lnSOC}$ prediction model was built from the *DB_Calib_RMQS_lnSOC* dataset using 10
470 latent variables and this model was applied to spectra of both *DB_Valid_RMQS_SOC* and
471 *DB_Tunisia_SOC* datasets. Finally, the ln(SOC) predictions were back-transformed into
472 SOC values for calculating the figures of merit. The performance of the $GM_{lnSOC}$ prediction
473 model was accurate, with an $R^2_{cv}$ of 0.89 and *RMSECV* of 0.2 g kg$^{-1}$ in the calibration step
474 (Table 2) and an $R^2_{val}$ of 0.90 and $RMSE_{val}$ of 6.6 g kg$^{-1}$ in the validation step (Table 2,
475 Figure 6b). When applied to the Tunisian test set, the $GM_{lnSOC}$ prediction model provided
476 high accuracy ($R^2_{test}$ = 0.97, $RMSE_{test}$ = 4.2 g kg$^{-1}$) and very slightly biased predictions
477 ($bias_{test}$ = 0.7 g kg$^{-1}$) (Table 3, Figure 6c).
478 A total of 95 spectral bands might be considered significant based on the analysis of
479 the VIP and regression coefficients in the $GM_{lnSOC}$ (Figure 6d). Among these 95 significant
480 spectral bands, those at approximately 2915, and 1800 cm$^{-1}$ had regression coefficients
481 higher than 3 times the standard deviation and therefore might be considered the most
482 significant ones.
483
484 [New Figure 6]
485

### 4.3. Local models

#### a. Soil inorganic carbon content

All validation soil samples had more than 30 spectral neighbours within the *DB_Calib_RMQS_SIC* dataset, so SIC could be predicted from $LM_{SIC}$ for all samples of *DB_Valid_RMQS_SIC.* The $LM_{SIC}$ provided very accurate and unbiased SIC predictions in validation ($R^2_{val}$ = 0.99 and $RMSE_{val}$ = 1.8 g kg$^{-1}$; Table 2). Therefore, this $LM_{SIC}$ provided validation performance slightly better than that of $GM_{SIC}$ (Table 2).

All Tunisian soil samples had more than 30 spectral neighbours within the *DB_Calib_RMQS_SIC* dataset, so SIC could be predicted from $LM_{SIC}$ for all Tunisian samples. The $LM_{SIC}$ provided accurate and slightly biased SIC predictions on Tunisian samples ($R^2_{test}$ = 0.96, $RMSE_{test}$ = 5.6 g kg$^{-1}$ and $bias_{test}$ = 1.7 g kg$^{-1}$; Table 3, Figure 7a). Therefore, $LM_{SIC}$ provided test performance slightly lower than that of $GM_{SIC}$, mainly due to bias (Table 3). The number of latent variables selected for $LM_{SIC}$ on Tunisian samples varied depending on the sample predicted and followed a relatively normal distribution centred at approximately 13, which was close to the optimal number of latent variables selected by the $GM_{SIC}$ (Figure 7c).

The number of spectral neighbours of Tunisian samples varied from 65 to 1293 (Figure 7b). Only a slight trend was observed between the number of neighbours and the prediction error, with a lower error when the number of neighbours increased (Figure 7b). This trend could be expected, as the use of a higher number of neighbour samples for calibration should result in more accurate predictions. The spectra from *DB_Calib_RMQS_SIC* used for building the 96 local individual Tunisian models $LM_{SIC}$ were selected from 0 to 87 times. So no spectrum from *DB_Calib_RMQS_SIC* was systematically selected, whereas only 1.6% of spectra from *DB_Calib_RMQS_SIC* were never selected. The frequently selected samples were mainly located in SIC-richest areas (Calcosols and Leptosols) such as the southeast (Prealps), the northeast (chalk

512 Champagne) and the transect from west (the Aquitanian Basin) to south (Mediterranean
513 Sea) (Figure 2A1 and 2A2).
514
515 [Figure 7]
516
517 **b. Soil organic carbon content**
518 As *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets contained same
519 predictors X-variables (MIR spectra), spectral neighbours of validation samples were the
520 same for $LM_{SOC}$ and $LM_{lnSOC}$ models. All validation soil samples had more than 30 spectral
521 neighbours within the *DB_Calib_RMQS_SOC* dataset and *DB_Calib_RMQS_lnSOC*
522 datasets, so SOC and ln(SOC) could be predicted from $LM_{SOC}$ and $LM_{lnSOC}$, respectively,
523 for all samples of *DB_Valid_RMQS_SOC.* The $LM_{SOC}$ provided accurate and very slightly
524 biased SOC predictions in validation ($R^2_{val}$ = 0.93, $RMSE_{val}$ = 5.4 g kg$^{-1}$ and $bias_{val}$ = -0.7
525 g kg$^{-1}$; Table 2). Therefore, $LM_{SOC}$ provided validation performance higher than that of
526 $GM_{SOC}$ (Table 2). The $LM_{lnSOC}$ provided accurate and unbiased SOC predictions in
527 validation ($R^2_{val}$ = 0.92, $RMSE_{val}$ = 5.7 g kg$^{-1}$ and $bias_{val}$ = -0.1 g kg$^{-1}$; Table 2). Therefore,
528 $LM_{lnSOC}$ provided validation performance higher than that of $GM_{SOC}$ and almost similar to
529 that of $LM_{SOC}$ (Table 2).
530 All Tunisian soil samples also had more than 30 spectral neighbours within the
531 *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets, so SOC and ln(SOC)
532 could be predicted from $LM_{SOC}$ and $LM_{lnSOC}$, respectively, for all Tunisian samples. The
533 $LM_{SOC}$ provided accurate and very slightly biased SOC predictions on Tunisian samples
534 ($R^2_{test}$ = 0.89, $RMSE_{test}$ = 6.9 g kg$^{-1}$ and $bias_{test}$ = 0.5 g kg$^{-1}$; Table 3, Figure 8a). Therefore,
535 this $LM_{SOC}$ provided test performance markedly higher than that of $GM_{SOC}$ (Table 3). The
536 number of latent variables selected by this $LM_{SOC}$ depending on the sample followed a
537 bimodal distribution centred at approximately 16 and 22 (Figure 8c). The second peak of

number of latent variables, approximately 22, was close to the number of latent variables selected by the $GM_{SOC}$ (Figure 8c).

The $LM_{lnSOC}$ model provided accurate and very slightly biased SOC predictions on Tunisian samples ($R^2_{test}$ = 0.93, $RMSE_{test}$ = 5.8 g kg$^{-1}$ and $bias_{test}$ = -0.6 g kg$^{-1}$; Table 3, Figure 9a). The $LM_{lnSOC}$ provided test performance markedly higher than that of $GM_{SOC}$ but lower than that of $GM_{lnSOC}$ (Table 3). The number of latent variables selected by this $LM_{lnSOC}$ depending on the sample followed a relatively normal distribution centred at approximately 11 (Figure 9c), which was close to the optimal number of latent variables selected by the $GM_{lnSOC}$.

As *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets contained same predictors X-variables (MIR spectra), spectral neighbours of Tunisian samples were the same for $LM_{SOC}$ and $LM_{lnSOC}$ models. The number of spectral neighbours of Tunisian samples varied from 65 to 1292 (Figure 8b and 9b), and no clear trend was observed between the number of neighbours and the prediction error obtained by $LM_{SOC}$ and $LM_{lnSOC}$ models (Figure 8b and 9b, respectively). As for SIC prediction with $LM_{SIC}$, the spectra from *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets used for building the 96 individual local Tunisian models were selected from 0 to 87 times. So no spectrum from *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets was systematically selected, whereas 1.7% of spectra from *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC* datasets were never selected. The frequently selected samples were mainly located in SOC-poor areas (Calcosols and Leptosols) in the northeast (chalk Champagne) and on the transect from west (Aquitanian Basin) to south (Mediterranean Sea) and in soils richer in SOC in the southeast (Prealps) (Figure 2B1 and 2B2).

[Figure 8]

564 [New Figure 9]

565

566

567 **5. Discussion**

568 ***5.1. Global models built on region A for application to region A.***

569 Before being applied to the Tunisian database, the global models were calibrated with an
570 RMQS subset (*DB_Calib_RMQS_SIC, DB_Calib_RMQS_SOC* and
571 *DB_Calib_RMQS_lnSOC*) and validated on an RMQS subset (*DB_Valid_RMQS_SIC* and
572 *DB_Valid_RMQS_SOC*), which means that the global models were calibrated using soil
573 samples collected over a region *A* to predict values for soil samples collected over this
574 same region *A*.

575 The validation performance of $GM_{SIC}$ was in accordance with results reported in the
576 literature. Grinand et al. (2012) obtained similar performances using the same RMQS
577 MIRS database, with $R^2_{val}$ and $RPD_{val}$ values of 0.97 and 7.6, respectively, when 3/4 of the
578 set, selected at random, was used for calibration and 1/4 was used for validation. Barthès
579 et al. (2016) obtained similar performances for SIC using the Tunisian MIRS database for
580 both calibration and prediction, with $R^2_{cv}$ and $RPD_{cv}$ values of 0.98 and 7.8, respectively.
581 Mc Carty et al. (2002) obtained similar performances using another MIRS database
582 collected in the US, with $R^2_{val}$ = 0.98 ($RPD_{val}$ was not mentioned and could not be
583 calculated). The most significant spectral bands for $GM_{SIC}$, located at 2500, 1800 and 860
584 $cm^{-1}$ (Figure 4d), might be attributed to stretching or bending vibrations in carbonate
585 molecules, as suggested by Du and Zhou (2009) and then by Grinand et al. (2012).

586 The validation performance of the $GM_{SOC}$ was also in accordance with some
587 literature results. Clairotte et al. (2016) obtained very similar performances using the full
588 RMQS MIRS database (including two depth layers, 0-30 and 30-50 cm, instead of one in
589 the present study), with $R^2_{val}$ and $RPD_{val}$ values of 0.88 and 2.7, respectively. Barthès et al.

590 (2016) obtained better performances using the Tunisian MIRS database with cross-
591 validation, with $R^2_{cv}$ and $RPD_{cv}$ values of 0.95 and 4.3, respectively. Moreover, Mc Carty et
592 al. (2002) obtained slightly higher performances using a MIRS database collected in the
593 US for both SOC calibration and validation, with $R^2_{val}$ = 0.94 ($RPD_{val}$ was not mentioned
594 and could not be calculated).

595 Following previous researches dealing with non-normal distribution of soil properties
596 (e.g., Waruru et al., 2014; Dangal et al., 2019), natural logarithm transformation was
597 applied to the highly skewed SOC values of the RMQS database to reach a normal
598 distribution in the calibration dataset (*DB_Calib_RMQS_lnSOC)*. Thanks to this normal
599 distribution, the performance of the $GM_{lnSOC}$ on *DB_Valid_RMQS_SOC* was slightly better
600 than the one of the $GM_{SOC}$.

601 Finally, validation performance was higher for SIC with $GM_{SIC}$ ($R^2_{val}$ and $RPD_{val}$
602 values of 0.98 and 7.6, respectively; Table 2) than for SOC prediction with $GM_{SOC}$ ($R^2_{val}$
603 and $RPD_{val}$ values of 0.88 and 2.7, respectively; Table 2) and $GM_{lnSOC}$ ($R^2_{val}$ and $RPD_{val}$
604 values of 0.90 and 2.9, respectively; Table 2), confirming that MIRS allows markedly more
605 accurate predictions of SIC than SOC as also shown by McCarty et al. (2002), Grinand et
606 al. (2012) and Barthès et al. (2016).

607

608 ***5.2. Models built on region A for application to region B.***

609 The models were calibrated by using a RMQS subset (*DB_Calib_RMQS_SIC,*
610 *DB_Calib_RMQS_SOC* and *DB_Calib_RMQS_lnSOC*) and tested on the Tunisian
611 dataset, which means that the models were calibrated by using soil samples collected over
612 a region *A* to predict values for soil samples collected over a region *B*, where *A* and *B* had
613 no common area, so the soil and climate conditions were different between the calibration
614 and test datasets. Our results showed that $GM_{SIC}$ provided accurate test performance
615 ($R^2_{test}$ and $RPD_{test}$ values of 0.96 and 4.9, respectively; Table 3), which was however lower

when applied to region *B* than to region *A* (Figure 4c, Tables 2 and 3). The $GM_{SOC}$ also provided markedly lower performance ($R^2_{test}$ and $RPD_{test}$ values of 0.64 and 1.3, respectively; Table 3) when applied to our region *B* than to our region *A* (Figures 5c, Table 2 and 3).

The RMQS spectra used for SIC and SOC predictions by local models were selected using a similarity measure (Pearson's coefficients of correlation), following the same approach than Shenk et al. (1997) and Nocita et al. (2014). So the driver of neighbours selection was the spectral similarity between Tunisian and French spectra. The $LM_{SIC}$ did not improve the SIC prediction accuracy compared to the $GM_{SIC}$ (Table 3). Therefore, $GM_{SIC}$ seemed robust and did not need to be adjusted to spectral particularities of region *B*. So rather than spectral similarity between French and Tunisian samples, the main reason for accurate SIC predictions in region *B* seemed to be the strong spectral features of SIC in the MIR region, as suggested by Gogé et al. (2014). The $LM_{SOC}$ improved SOC prediction accuracy compared to the $GM_{SOC}$ (Table 3). Therefore, $GM_{SOC}$ seemed poorly robust and the calibration over region *A* needed to be adjusted to spectral particularities of region *B* using spectral neighbours (e.g., Shenk et al., 1997; Nocita et al., 2014). The increase in the performance of MIRS-based SOC prediction when shifting from global to local PLSR is in accordance with literature (e.g. Ramirez-Lopez et al., 2013; Shi et al., 2015; Clairotte et al., 2016 and Dangal et al., 2019). Finally, the frequently selected spectral neighbours of Tunisian samples by the $LM_{SOC}$ were mainly located in SOC-poor areas in the northeast (chalk Champagne) and on the transect from west (Aquitanian Basin) to south (Mediterranean Sea) and in soils richer in SOC in the southeast (Prealps) (Figure 2B1 and 2B2). So these frequently selected spectral neighbours of Tunisian samples by SOC local model were not located only over the more similar climatic and pedological contexts such as the Mediterranean context (southeast of France).

### *5.3. Impact of the SOC ln-transformation on models built on region A for application to region B.*

The global model calibrated on region *A* with log-transformed SOC values ($GM_{lnSOC}$) provided accurate performance on region *B* ($R^2_{test}$ and $RMSE_{test}$ values of 0.97 and 4.9 g $kg^{-1}$, respectively; Figure 6c, Table 3). The local model provided better performance when calibrated on region *A* with log-transformed SOC values ($LM_{lnSOC}$) than with SOC values ($LM_{SOC}$), but less accurate than $GM_{lnSOC}$ on region *B* ($R^2_{test}$ and $RMSE_{test}$ values of 0.93 and 3.6 g $kg^{-1}$, respectively; Figure 9c, Table 3). So whatever the model using log-transformed SOC data in calibration database ($GM_{lnSOC}$ or $LM_{lnSOC}$), the SOC predictions on region *B* were improved compared to models using highly skewed SOC values in calibration database ($GM_{SOC}$ or $LM_{SOC}$; Table 3), as showed by Jaconi et al. (2019).

Finally, and unexpectedly, $GM_{lnSOC}$ provided better performance than $LM_{SOC}$ when applied on Tunisian samples. So a transformation of calibration SOC values improved SOC model performance more clearly than spectral selection of calibration samples (neighbours). Therefore SOC model performance was more sensitive to the distribution of the explained variable of the calibration samples than to spectral similarity between calibration and test spectra.

### *5.4. Perspectives*

This study, which used MIRS to predict SOC and SIC contents by using a database collected over a region *A* to predict values over a region *B*, where *A* and *B* have no common area, could be continued with a study to develop predictions based on selected spectral bands. Indeed, spectral band selection remains to be explored, as several studies testing such an approach have obtained different results. Viscarra Rossel and Lark (2009) successfully used wavelets and a variable selection technique to improve SOC calibration using Vis–NIR and MIRS data. Additionally, Volhand et al. (2016) outperformed SOC

predictions based on Vis–NIR spectra by using band selection. However, Stevens et al. (2013) tested recursive feature elimination based on the random forest approach and obtained no overall increase in the accuracy of soil property prediction using the LUCAS (European) Vis-NIR soil database compared to models using all spectral bands. In addition, Yang et al. (2019) tested a generic algorithm for spectral band selection but obtained no overall increase in SOC prediction accuracy. As previously tested by, e.g., Guerrero et al. (2014) and Guy et al. (2016), spiking could be another useful approach to improve prediction accuracy when applying large-scale calibrations to small regions. Spiking consists of adding a small subset of samples from region *B* (spiking subset) to the dataset from region *A* to recalibrate a model.

As well, this study could be continued with an impact analysis of the selection of spectral neighbours. Both the number of spectral neighbours and the procedure to select them could be analysed. Several approaches are available for selecting representative calibration samples (Shetty et al., 2012) and could also be tested. For example, to analyse the spectral similarity between calibration and test spectra, the Pearson correlation coefficient between spectra could be replaced by the Mahalanobis distance between spectra (Nocita et al., 2014) or the Pearson correlation coefficient distance based on Fast Fourier Transform of spectra (Gogé et al., 2012). Finally, some covariates could be added in local regression to improve prediction accuracy, as previously tested by Nocita et al. (2014) who used clay contents of samples as covariates to predict SOC content.

## 6. Conclusion

This work highlighted that, as expected, the SOC and SIC contents of French samples were successfully predicted from the French MIRS soil database using a global model based on PLS regression. Predictions of SIC and SOC are accurate when the calibration

and validation samples come from same pedologic and climatic contexts. This work also highlighted that when the calibration and validation samples come from different pedologic and climatic contexts, the SOC prediction performance over validation samples decreases, whereas the SIC prediction performance remains accurate. Finally, this work showed that prediction models were more sensitive to the distribution of the explained variables of calibration samples than to the spectral similarity between calibration and test spectra. This study confirmed the very high applicability of MIRS for SIC determination and the robustness of SIC prediction models, even when the calibration and validation samples come from different contexts.

**Acknowledgment**

RMQS soil sampling and physico-chemical analyses were supported by the GIS Sol, which is a scientific group of interest on soils involving the French Ministry for ecology and sustainable development, the French Ministry of agriculture, the French National institute for geographical and forest information (IGN), the French Environment and Energy Management Agency (ADEME), the French agency for biodiversity, French Institute for Research and Development (IRD) and the French Institute for Agronomic Research (INRA, which is a French public research organisation dedicated to agriculture, food and environment). We thank all the people involved in sampling RMQS sites and in samples preparation. D.A. is coordinator of the Research Consortium GLADSOILMAP, supported by LE STUDIUM Loire Valley Institute for Advanced Studies. And we acknowledge the contribution of three anonymous reviewers for their help in improving the manuscript.

**References**

720 Apesteguia, M., Plante, A.F., Virto, I., 2018. Methods assessment for organic and
721 inorganic carbon quantification in calcareous soils of the Mediterranean region.
722 Geoderma Regional 12, 39–48.
723 Ahmed, Z.U., Woodbury, P.B., Sanderman, J., Hawke, B., Jauss, V., Solomon, D.,
724 Lehmann, J., 2017. Assessing soil carbon vulnerability in the Western USA by
725 geospatial modeling of pyrogenic and particulate carbon stocks. Journal of
726 Geophysical Research: Biogeosciences 122, 354–369.
727 Arrouays, D., Jolivet, C., Boulonne, L., Bodineau, G., Saby, N., Grolleau, E., 2002. A new
728 initiative in France: a multi-institutional soil quality monitoring network. Comptes
729 Rendus de l'Académie d'Agriculture de France 88, 93–105.
730 Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and
731 de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy. 43,
732 772–777.
733 Barthès, B.G., Kouakoua, E., Moulin, P., Hmaidi, K., Gallali, T., Clairotte, M., Bernoux, M.,
734 Bourdon, E., Toucet, J., Chevallier, T., 2016. Studying the physical protection of soil
735 carbon with quantitative infrared spectroscopy. Journal of Near Infrared Spectroscopy
736 24, 199–214.
737 Batjes, N. H., 1996. Total carbon and nitrogen in the soils of the world. European Journal
738 of Soil Science 47, 151–163.
739 Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010.
740 Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric
741 indicators commonly used for assessing the quality of the prediction. Trends in
742 Analytical Chemistry (TRAC) 29(9), 1073–1081.
743 Bellon-Maurel, V., McBratney, A.B., 2011. Near-infrared (NIR) and mid-infrared (MIR)
744 spectroscopic techniques for assessing the amount of carbon stock in soils — Critical
745 review and research perspectives. Soil Biology and Biochemistry 43, 1398–1410.

Bernoux, M., Chevallier, T., 2014. Carbon in Drylands. Multiple Essential Functions. Les Dossier Thématiques du CSFD. N°10. CSFD/Agropolis International, Montpellier, France.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.

Cécillon, L., Barthès, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., Brun, J.J., 2009. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). European Journal of Soil Science 60(5), 770–784.

Chen, Y., Barak, P., 1982. Iron nutrition of plants in calcareous soils. Advances in Agronomy 35, 217–240.

Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, 103–112.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M., Barthès, B.G., 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma 276, 41–52.

Comstock, J.P., Sherpa, S.R., Ferguson, R., Bailey, S., Beem-Miller, J.P., Lin, F., Lehmann, J., Wolfe, D.W., 2019. Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. PLoS ONE 14(2): e0210235.

Dangal, S.R.S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Systems 3, 11.

Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22, 1–20.

771 Du, C., Zhou, J., 2009. Evaluation of soil fertility using infrared spectroscopy: a review.
772 Environmental Chemistry Letters 7, 97–113.
773 Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample
774 selection step of local regression for quantitative analysis of large soil NIRS
775 database. Chemometrics and Intelligent Laboratory Systems 110, 168–176.
776 Gogé, F., Gomez, C., Jolivet, C., Joffre, R., 2014. Which strategy is best to predict soil
777 properties of a local site from a national Vis–NIR dataset? Geoderma 213, 1–9.
778 Gregorich E.G., Carter M.R., Angers D.A., Monreal C.M. et Ellert B.H., 1994. Towards a
779 minimum data set to assess soil organic matter quality in agricultural soils. Canadian
780 Journal of Soil Science 74, 367–385.
781 Grinand, C., Barthès, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G.,
782 Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a
783 national scale (France) using mid-infrared reflectance spectroscopy (MIRS).
784 European Journal of Soil Science 63(2), 141–151.
785 Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen,
786 A.M., Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic
787 carbon at local scale with spiked NIR calibrations: effects of selection and extra-
788 weighting on the spiking subset. European Journal of Soil Science 65, 248–263.
789 Guy, A.L., Siciliano, S.D., Lamb, E.G., 2015. Spiking regional vis-NIR calibration models
790 with local samples to predict soil organic carbon in two High Arcticpolar deserts using
791 a vis-NIR probe. Canadian Journal of Soil Science 95, 237–249.
792 Hannam, K.D., Kehila, D., Millard, P., Midwood, A.J., Neilson, D., Neilson, G.H., Forge, T.
793 A., Nichol, C., Jones, M.D., 2016. Bicarbonates in irrigation water contribute to
794 carbonate formation and $CO_2$ production in orchard soils under drip irrigation,
795 Geoderma 266, 120–126.

796 Harris, D., Horwath, W.R., van Kessel, C., 2001. Acid fumigation of soils to remove
797 carbonates prior to total organic carbon or carbon-13 isotopic analysis. Soil Science
798 Society of America Journal 65, 1853–1856.
799 ISO (International Organization for Standardisation), 1995a. ISO 10693:1995 –
800 Determination of Carbonate Content – Volumetric Method. ISO, Geneva.
801 ISO (International Organization for Standardisation), 1995b. ISO 10694:1995 - Soil Quality
802 - Determination of Organic and Total Carbon after Dry Combustion (Elementary
803 Analysis). ISO, Geneva.
804 IUSS Working Group WRB, 2014. International Union of Soil Sciences, Working Group
805 World Reference Base for Soil Resources. World Reference Base for Soil Resources
806 2014. International Soil Classification System for Naming Soils and Creating Legends
807 of Soil Maps. FAO, Rome.
808 Jacobson, M., Charlson, R.J., Rodhe, H., Orians, G.H., 2000. Earth System Science: From
809 Biogeochemical Cycles to Global Changes. Academic Press, London.
810 Jaconi, A., Poeplau, C., Ramirez-Lopez, L., Van Wesemael, B., Don, A., 2019. Log-ratio
811 transformation is the key to determining soil organic carbon fractions with near-
812 infrared spectroscopy. European Journal of Soil Science, 70, 127–139.
813 Janik, L.J., Skjemstad, J.O., Merry, R.H., 1998. Can mid infrared diffuse reflectance
814 analysis replace soil extractions? Australian Journal of Experimental Agriculture 38,
815 681–696.
816 Jauss, V., Sullivan, P.J., Sanderman, J., Smith, D.B., and Lehmann, J., 2017. Pyrogenic
817 carbon distribution in mineral topsoils of the northeastern United States. Geoderma
818 296, 69–78.
819 Knadel, M., Deng, F., Thomsen, A., Greve, M.H., 2012. Development of a Danish national
820 vis-NIR soil spectral library for soil organic carbon determination. In: Minasny, B.,

821 Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. CRC
822 Press, Boca Raton, FL.
823 Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., Robbez-Masson, J.-M., 2008.
824 Estimation of soil clay and calcium carbonate using laboratory, field and airborne
825 hyperspectral measurements. Remote Sensing of Environment 112(3), 825–835.
826 McCrea, J.M. , 1950. On the isotopic chemistry of carbonates and a paleotemperature
827 scale. Journal of Chemical Physics 18, 849–857.
828 Mark, H.L., Tunnell, D., 1985. Qualitative near-infrared reflectance analysis using
829 Mahalanobis distances. Analytical Chemistry 57, 1449–1456.
830 McCarty, G.W., Reeves III, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-
831 infrared and near-infrared diffuse reflectance spectroscopy for soil carbon
832 measurement. Soil Science Society of America Journal 66: 640–646.
833 Mevik, B.-H., Wehrens, R., 2007. The pls Package: Principal Component and Partial Least
834 Squares Regression in R. Journal of Statistical Software 18, 1–24.
835 Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014.
836 Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a
837 local partial least square regression approach. Soil Biology & Biochemistry 68, 337–
838 347
839 Pearson, R.K., 2002. Outliers in process modeling and identification. IEEE Transactions on
840 Control Systems Technology 10(1), 55–63.
841 Pribyl, D.W., 2010. A critical review of the conventional SOC to SOM conversion factor.
842 Geoderma 156, 75–83.
843 R Development Core Team, 2012. R: A Language and Environment for Statistical
844 Computing. R foundation for Statistical Computing, Vienna. http://www.R-project.org/
845 Rabenarivo, M., Chapuis-Lardy, L., Brunet, D., Chotte, J.-L., Rabeharisoa, L.,
846 Barthès, B.G., 2013. Comparing near and mid-infrared reflectance spectroscopy for

847 determining properties of Malagasy soils, using global or LOCAL calibration. Journal
848 of Near Infrared Spectroscopy 21(6): 495–509.
849 Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T.
850 2013. The spectrum-based learner: A new local approach for modeling soil vis–NIR
851 spectra of complex datasets. Geoderma 195–196, 268–279.
852 Reeves, D.W., 1997. The role of soil organic matter in maintaining soil quality in
853 continuous cropping systems. Soil & Tillage Research 43, 131–167.
854 Reeves III, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil
855 analysis emphasizing carbon and laboratory versus on-site analysis: where are we
856 and what needs to be done? Geoderma 158 (1-2), 3–14.
857 Romanyà, J., Rovira, P., 2011. An appraisal of soil organic C content in Mediterranean
858 agricultural soils. Soil Use and Management 27, 321–332.
859 Scharlemann, J.P., Tanner, E.V., Hiederer, R., Kapos, V., 2014. Global soil carbon:
860 understanding and managing the largest terrestrial carbon pool. Carbon
861 Management 5, 81–91.
862 Shenk, J., Westerhaus, M., Berzaghi, P., 1997. Investigation of a LOCAL calibration
863 procedure for near infrared instruments. J. Near Infrared Spectrosc. 5 (1), 223-232.
864 Shetty, N., Rinnan, A., Gislum, R., 2012. Selection of representative calibration sample
865 sets fornear-infrared reflectance spectroscopy to predict nitrogen concentration in
866 grasses. Chemometrics and Intelligent Laboratory Systems 111, 59–65.
867 Shi, Z., Ji, W., Viscarra-Rossel, R.A., Chen, S., Zhou, Y., 2015. Prediction of soil organic
868 matter using a spatially constrained local partial least squares regression and the
869 Chinese vis–NIR spectral library. European Journal of Soil Science 66, 679–687.
870 Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J.,
871 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy

for prediction of soil physical, chemical, and biological properties. Applied Spectroscopy Reviews 49(2), 139–186.

Stevens, A, Nocita, M, Tóth, G, Montanarella, L, van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS ONE 8(6): e66409.

Tatzber, M., Mutsch, F., Mentler, A., Leitger, E., Englich, M., Gerzabek, M., 2010. Determination of organic and inorganic carbon in forest soil samples by mid-infrared spectroscopy and partial least squares regression. Applied Spectroscopy 64(10):1167–75.

Tenenhaus, M., 1998. La Régression PLS. Editions Technip, Paris.

Terra, F.S., Demattê, J.A.M., Viscarra Rossel R.A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. Geoderma 255–256 (2015) 81–93

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 2008. Using a legacy soil sample to develop a mid-IR spectral library. Soil Research 46, 1–16.

Viscarra Rossel, R.A., Walvoort, D.J.J., Mc Bratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.

Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. European Journal of Soil Science 63, 848–860.

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., Brodsky, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero,

C., Hedley, C.B., Knadel, M., Morras, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M. Rufasto, 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-Science Reviews 155, 198–230.

Viscarra Rossel, R.A., Lark, R.M., 2009, Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. European Journal of Soil Science 60, 453–464.

Vohland, M., Ludwig, M., Harbich, M., Emmerling, C., Thiele-Bruhn, S., 2016. Using variable selection and wavelets to exploit the full potential of visible–near infrared spectra for predicting soil properties. Journal of Near Infrared Spectroscopy 24, 255–269.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Science 37(1), 29–38.

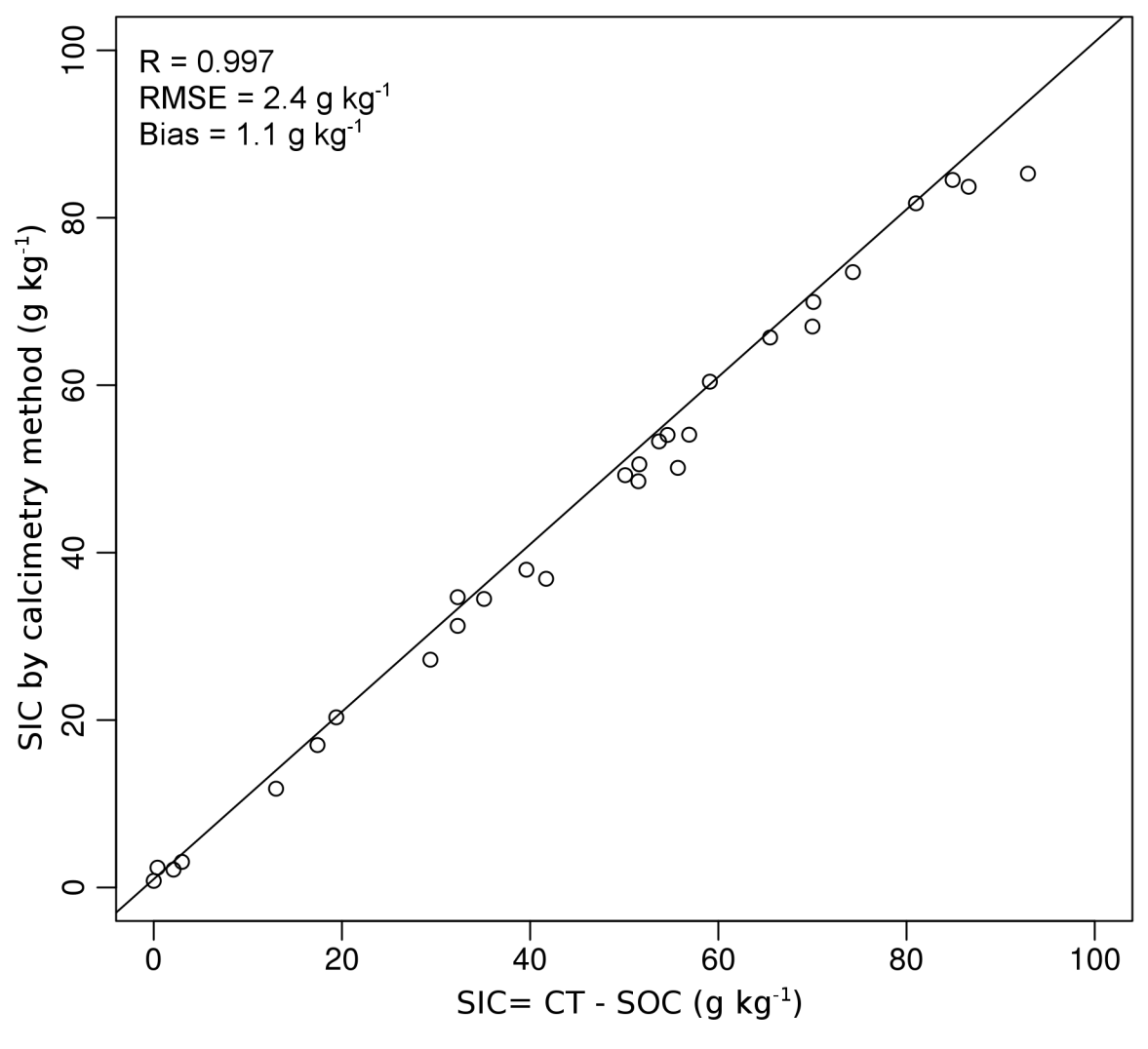
Wang, X., Wang, J., Zhang, J. (2012) Comparisons of three methods for organic and inorganic carbon in calcareous soils of northwestern China. PLoS ONE 7(8): e44334.

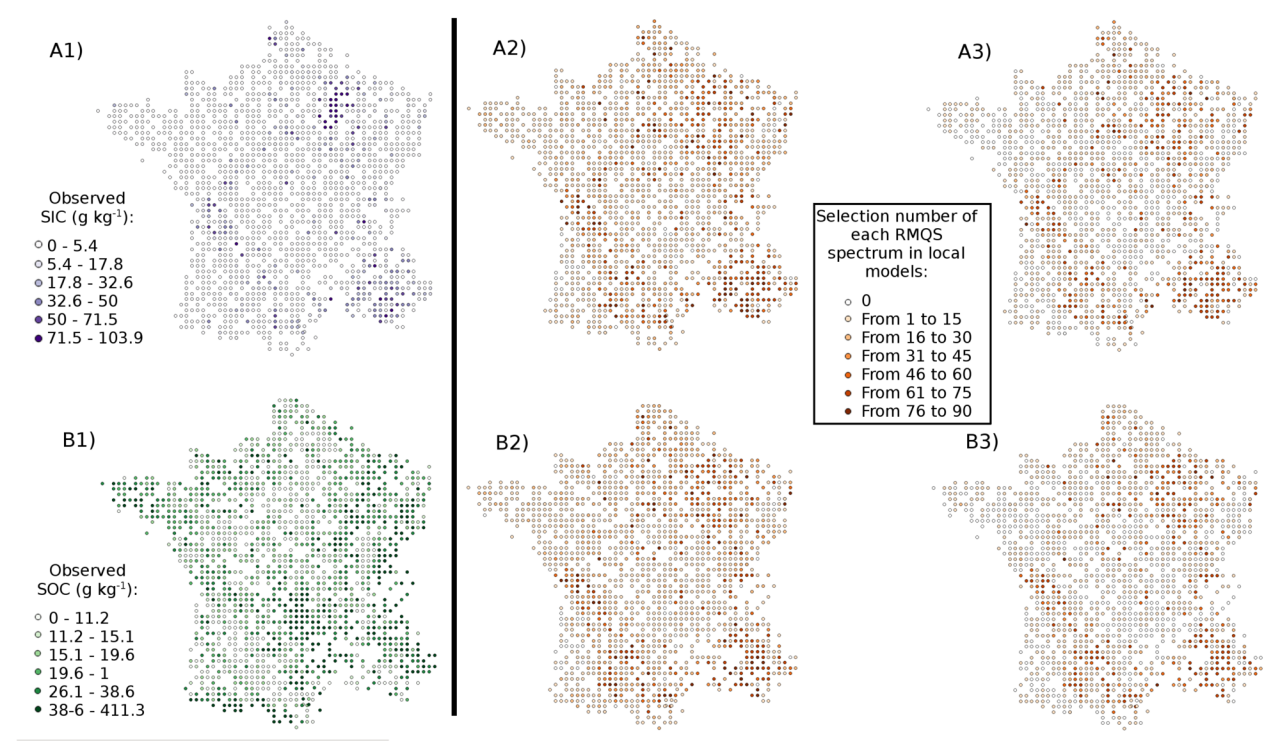
Wijewardane, N.K., Ge, Y., Wills, S., Libohova, Z., 2018. Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. Soil Science Society of America Journal 82, 722–731

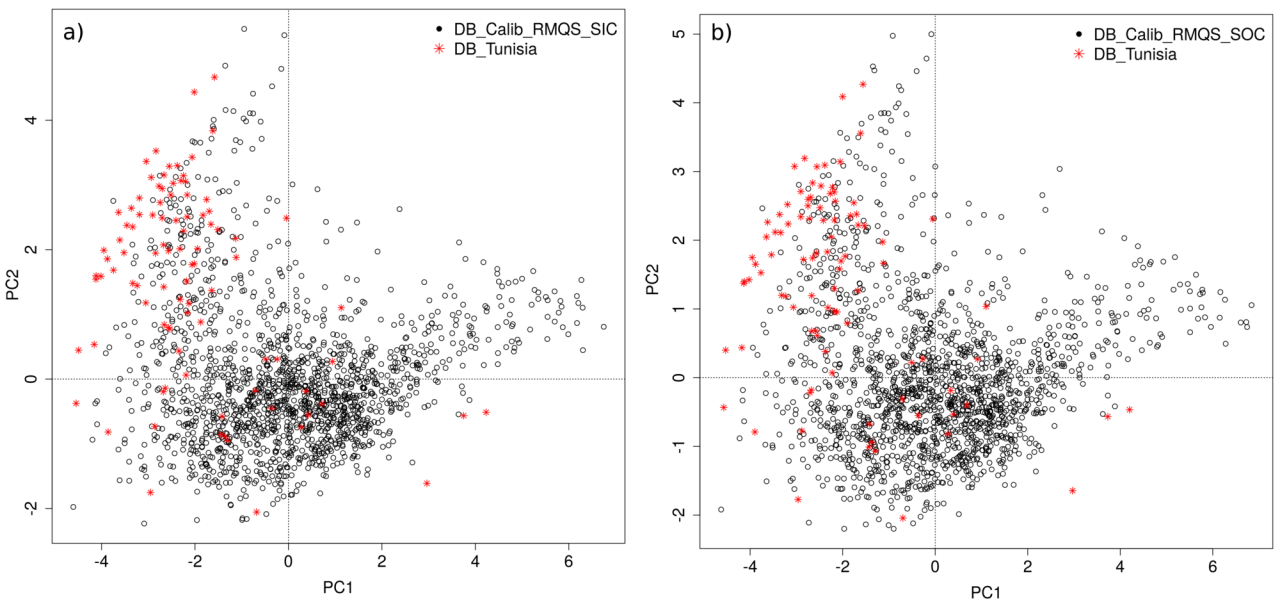
Williams, P.C., Norris, K.H., 1987. Qualitative applications of near-infrared reflectance spectroscopy. In: Williams, P., Norris, K. (Eds.), Near-Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, St. Paul, MN, pp. 241–246.

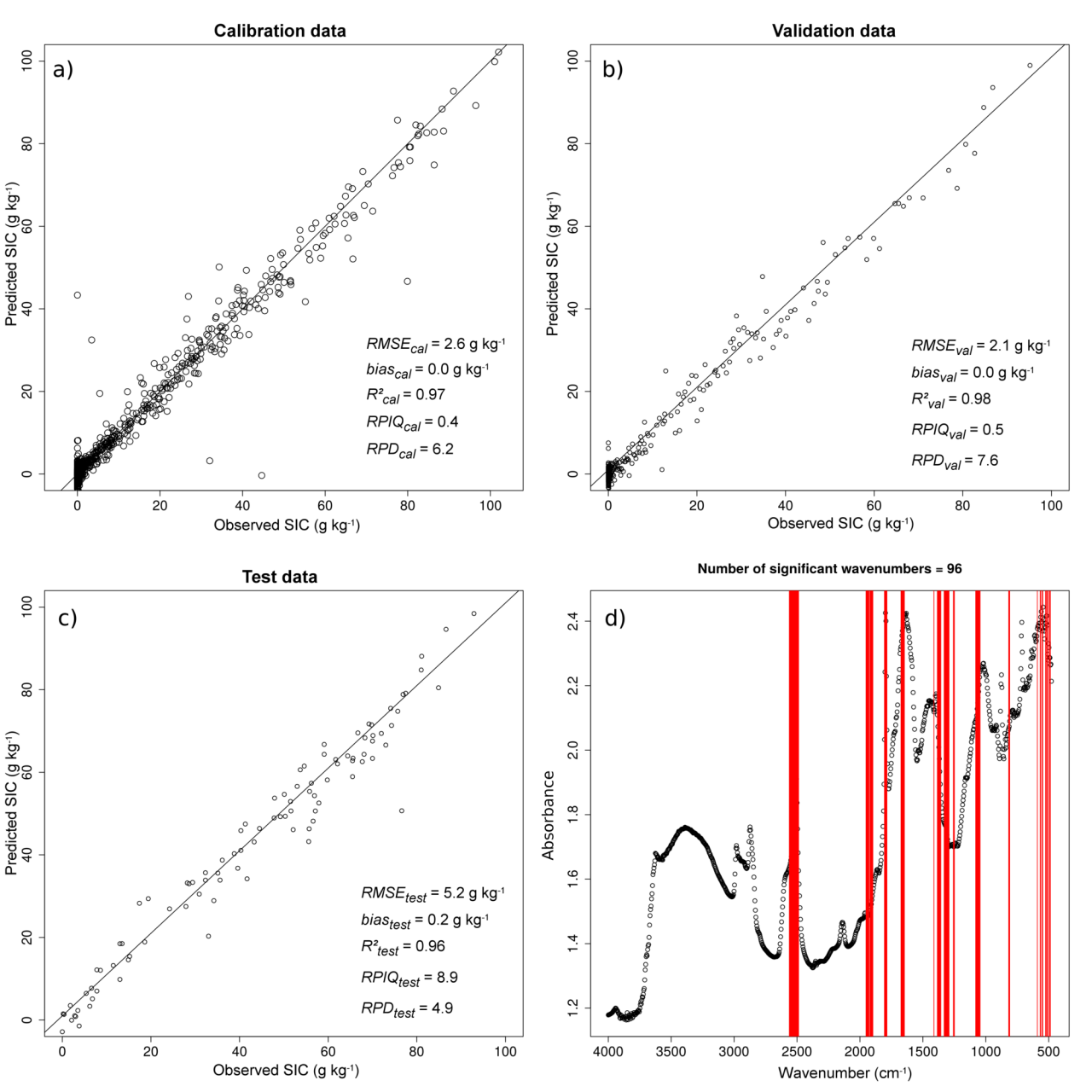
Wold, S., Johansson, E., Cocchi, M., 1993. PLS - partial least squares projections to latent structures. In: Kubinyi, H. (Ed.), 3D-QSAR in Drug Design, Theory, Methods, and Applications. ESCOM Science Publishers, Leiden, pp. 523–550.

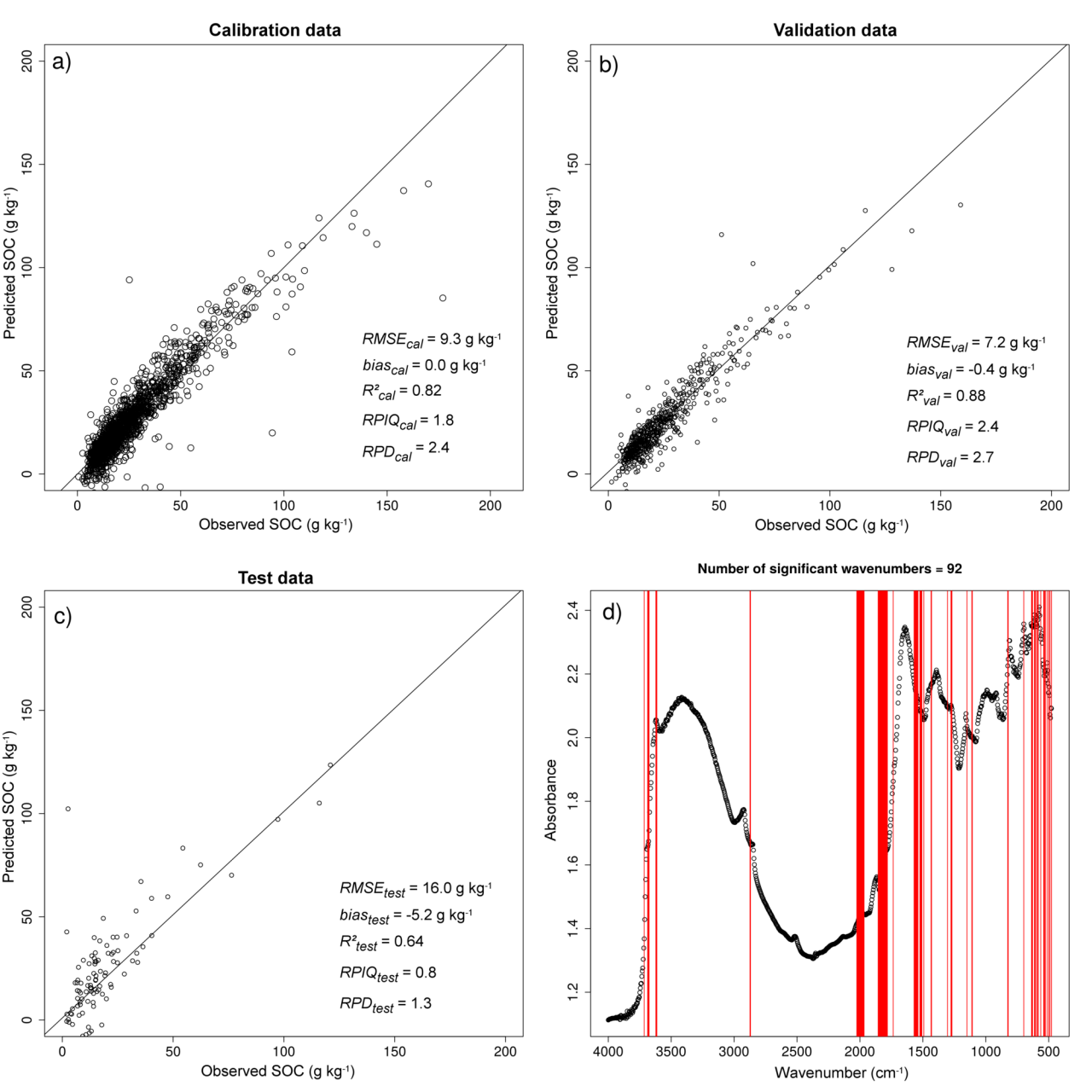
923 Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics.
924 Chemometrics and Intelligent Laboratory Systems 58, 109–130.
925 Workman, J. Jr.,, Weyer, L., 2008. Practical Guide to Interpretive Near-Infrared
926 Spectroscopy. CRC Press, Boca Raton, FL.
927 Waruru, B.K., Shepherd K.D., Ndegwa G.M., Kamoni P.T., Sila A.M. 2014. Rapid
928 estimation of soil engineering properties using diffuse reflectance near infrared
929 spectroscopy. Biosystems Engineering, 121, pp. 177-185.
930 Yang, M., Xu, D., Chen, S., Li, H., Shi, Z., 2019. Evaluation of machine learning
931 approaches to predict soil organic matter and pH using vis-NIR spectra. Sensors 19,
932 263.

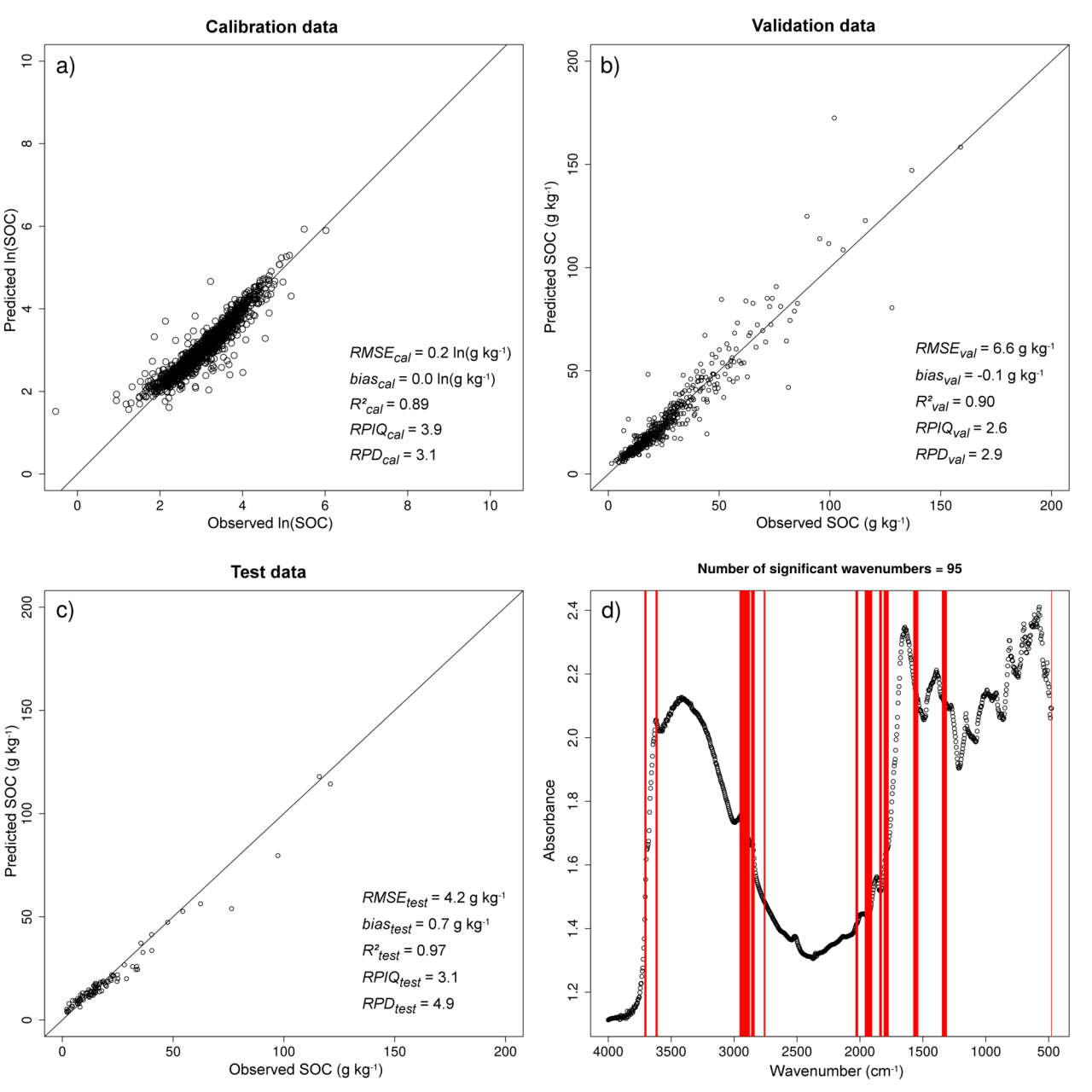

R = 0.997
RMSE = 2.4 g kg-1
Bias = 1.1 g kg-1
SIC by calcimetry method (g kg-1)
SIC= CT - SOC (g kg-1)
0
20
40
60
80
100

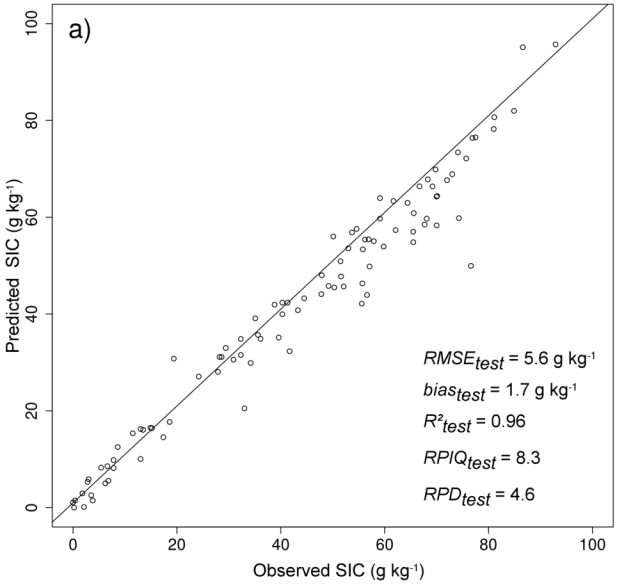

A1)
Observed SIC (g kg-1):
0 - 5.4
5.4 - 17.8
17.8 - 32.6
32.6 - 50
50 - 71.5
71.5 - 103.9
A2)
A3)
Selection number of each RMQS spectrum in local models:
0
From 1 to 15
From 16 to 30
From 31 to 45
From 46 to 60
From 61 to 75
From 76 to 90
B1)
Observed SOC (g kg-1):
0 - 11.2
11.2 - 15.1
15.1 - 19.6
19.6 - 1
26.1 - 38.6
38-6 - 411.3
B2)
B3)

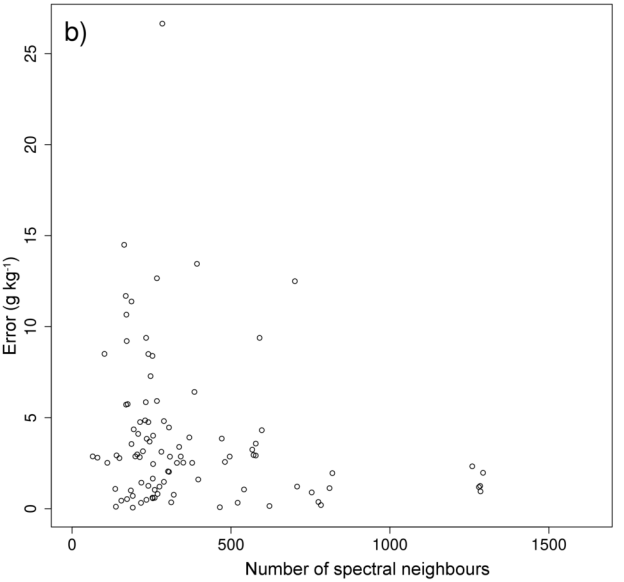

a)
DB_Calib_RMQS_SIC
DB_Tunisia
PC2
PC1
b)
DB_Calib_RMQS_SOC
DB_Tunisia
PC2
PC1

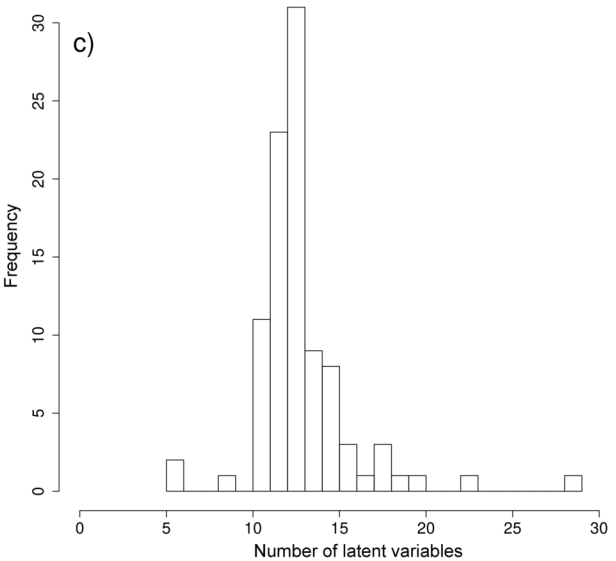

Calibration data
a)
Predicted SIC (g kg-1)
Observed SIC (g kg-1)
RMSEcal = 2.6 g kg-1
biascal = 0.0 g kg-1
R²cal = 0.97
RPIQcal = 0.4
RPDcal = 6.2
Validation data
b)
Predicted SIC (g kg-1)
Observed SIC (g kg-1)
RMSEval = 2.1 g kg-1
biasval = 0.0 g kg-1
R²val = 0.98
RPIQval = 0.5
RPDval = 7.6
Test data
c)
Predicted SIC (g kg-1)
Observed SIC (g kg-1)
RMSEtest = 5.2 g kg-1
biastest = 0.2 g kg-1
R²test = 0.96
RPIQtest = 8.9
RPDtest = 4.9
Number of significant wavenumbers = 96
d)
Absorbance
Wavenumber (cm-1)

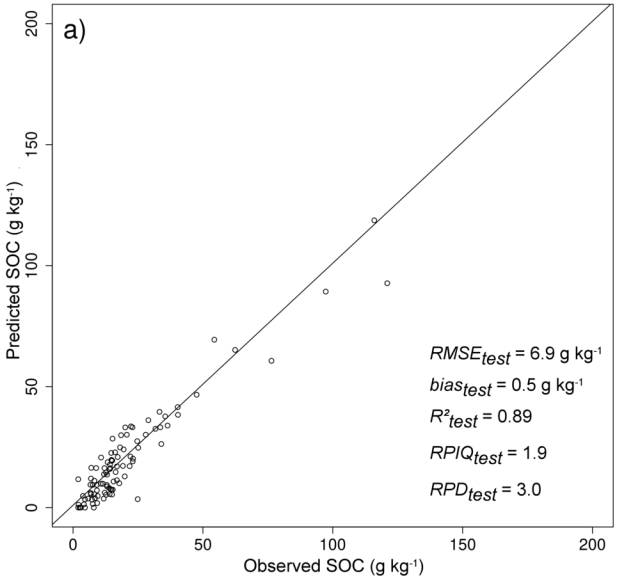

Calibration data
a)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSEcal = 9.3 g kg-1
biascal = 0.0 g kg-1
R²cal = 0.82
RPIQcal = 1.8
RPDcal = 2.4
Validation data
b)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSEval = 7.2 g kg-1
biasval = -0.4 g kg-1
R²val = 0.88
RPIQval = 2.4
RPDval = 2.7
Test data
c)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSEtest = 16.0 g kg-1
biastest = -5.2 g kg-1
R²test = 0.64
RPIQtest = 0.8
RPDtest = 1.3
Number of significant wavenumbers = 92
d)
Absorbance
Wavenumber (cm-1)

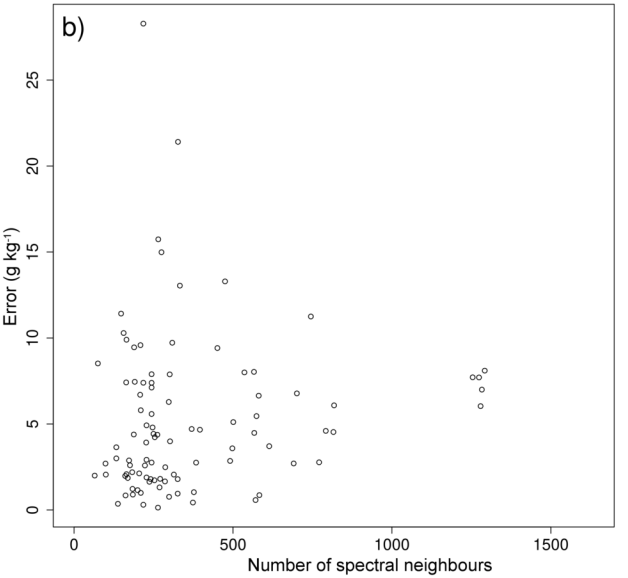

Calibration data
a)
Predicted ln(SOC)
Observed ln(SOC)
RMSE_cal = 0.2 ln(g kg-1)
bias_cal = 0.0 ln(g kg-1)
R²_cal = 0.89
RPIQ_cal = 3.9
RPD_cal = 3.1
Validation data
b)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSE_val = 6.6 g kg-1
bias_val = -0.1 g kg-1
R²_val = 0.90
RPIQ_val = 2.6
RPD_val = 2.9
Test data
c)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSE_test = 4.2 g kg-1
bias_test = 0.7 g kg-1
R²_test = 0.97
RPIQ_test = 3.1
RPD_test = 4.9
Number of significant wavenumbers = 95
d)
Absorbance
Wavenumber (cm-1)

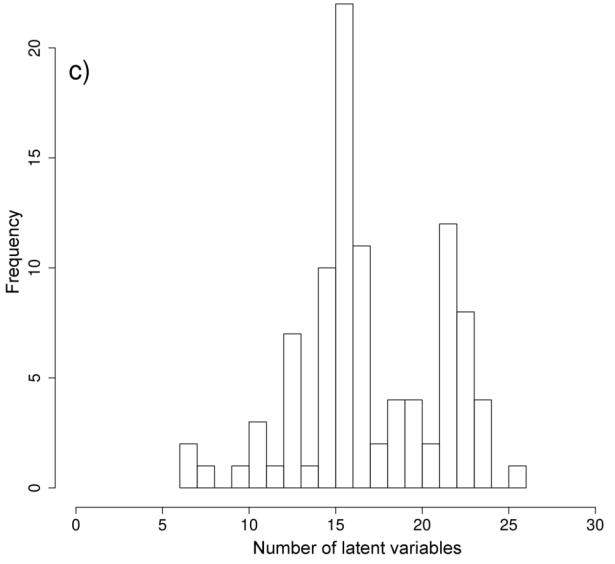

a)
Predicted SIC (g kg-1)
Observed SIC (g kg-1)
RMSEtest = 5.6 g kg-1
biastest = 1.7 g kg-1
R²test = 0.96
RPIQtest = 8.3
RPDtest = 4.6

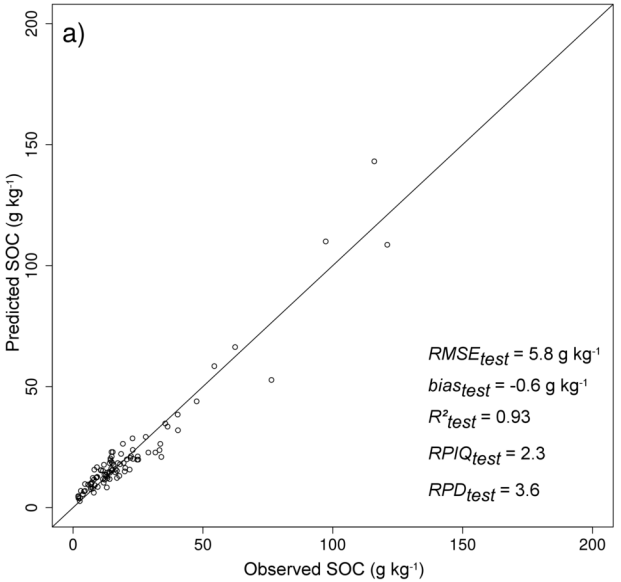

b)
Error (g kg-1)
Number of spectral neighbours

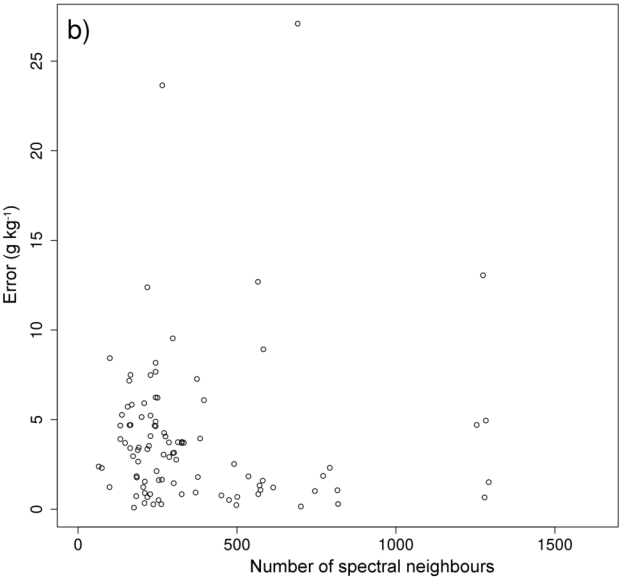

c)
Frequency
Number of latent variables


a)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSEtest = 6.9 g kg-1
biastest = 0.5 g kg-1
R²test = 0.89
RPIQtest = 1.9
RPDtest = 3.0


b)
Error (g kg-1)
Number of spectral neighbours


c)
Frequency
Number of latent variables


a)
Predicted SOC (g kg-1)
Observed SOC (g kg-1)
RMSEtest = 5.8 g kg-1
biastest = -0.6 g kg-1
R²test = 0.93
RPIQtest = 2.3
RPDtest = 3.6


b)
Error (g kg-1)
Number of spectral neighbours

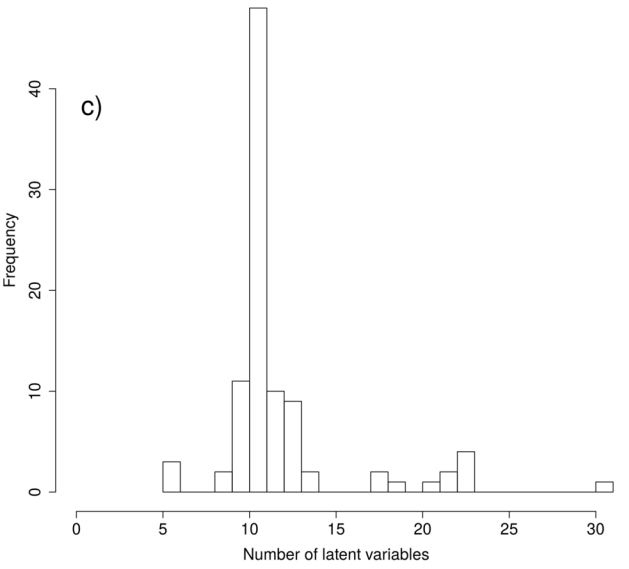

c)
Frequency
Number of latent variables

Table 1: Soil datasets statistics. The SIC values set to zero correspond to values under the laboratory quantification limit (< 0.1 g $kg^{-1}$).

| Dataset | Number of soil samples | Min g $kg^{-1}$ | Max g $kg^{-1}$ | Mean g $kg^{-1}$ | Median g $kg^{-1}$ | SD[a] g $kg^{-1}$ | Skewness |
|---|---|---|---|---|---|---|---|
| ***DB_RMQS_SOC*** | 2178 | 0.6 | 411.3 | 25.8 | 19.6 | 21.8 | 4.9 |
| ***DB_RMQS_SIC*** | 2178 | 0.0 | 103.9 | 6.4 | 0.0 | 16.1 | 3.1 |
| ***DB_Tunisia_SOC*** | 96 | 2.0 | 121.0 | 20.1 | 14.6 | 21.0 | 3.0 |
| ***DB_Tunisia_SIC*** | 96 | 0.0 | 92.9 | 43.3 | 48.5 | 25.6 | -0.2 |
| ***DB_Calib_RMQS_SOC*** | 1586[b] | 0.6 | 411.3 | 25.6 | 19.4 | 22.3 | 5.5 |
| ***DB_Calib_RMQS_SIC*** | 1582[c] | 0.0 | 103.9 | 6.4 | 0 | 16.1 | 3.1 |
| ***DB_Valid_RMQS_SOC*** | 544 | 1.5 | 159.0 | 25.5 | 19.6 | 19.5 | 2.5 |
| ***DB_Valid_RMQS_SIC*** | 544 | 0.0 | 95.2 | 6.3 | 0.0 | 15.7 | 3.1 |

[a] SD: standard deviation

[b] after removing 48 spectral outliers

[c] after removing 52 spectral outliers

Table 2: Figures of merit obtained with global and local models over the French calibration and validation databases.

| **Models** | $R^2_{cv}$ | *RMSECV* (g kg$^{-1}$) | $R^2_{val}$ | $RMSE_{val}$ (g kg$^{-1}$) | $bias_{val}$ (g kg$^{-1}$) | $RPD_{val}$ | $RPIQ_{val}$ |
|---|---|---|---|---|---|---|---|
| ***GM$_{SIC}$*** | 0.97 | 2.8 | 0.98 | 2.1 | 0.0 | 7.6 | 0.5 |
| ***LM$_{SIC}$*** | *nd* | *nd* | 0.99 | 1.8 | 0.0 | 8.8 | 0.6 |
| ***GM$_{SOC}$*** | 0.80 | 9.9 | 0.88 | 7.2 | -0.4 | 2.7 | 2.4 |
| ***LM$_{SOC}$*** | *nd* | *nd* | 0.93 | 5.4 | -0.7 | 3.6 | 3.2 |
| ***GM$_{lnSOC}$*** | 0.89 | 0.2* | 0.90 | 6.6 | -0.1 | 2.9 | 2.6 |
| ***LM$_{lnSOC}$*** | *nd* | *nd* | 0.92 | 5.7 | -0.1 | 3.4 | 3 |

**RMSEcv calculated on ln(SOC)*

*nd: Not determined.*

Table 3: Figures of merit obtained with global and local models over the Tunisian soil samples.

| **Prediction model** | $R^2_{test}$ | $RMSE_{test}$ (g kg$^{-1}$) | $bias_{test}$ (g kg$^{-1}$) | $RPD_{test}$ | $RPIQ_{test}$ |
|---|---|---|---|---|---|
| ***GM$_{SIC}$*** | 0.96 | 5.2 | 0.2 | 4.9 | 8.9 |
| ***LM$_{SIC}$*** | 0.96 | 5.6 | 1.7 | 4.6 | 8.3 |
| ***GM$_{SOC}$*** | 0.64 | 16.0 | -5.2 | 1.3 | 0.8 |
| ***LM$_{SOC}$*** | 0.89 | 6.9 | 0.5 | 3.0 | 1.9 |
| ***GM$_{lnSOC}$*** | 0.97 | 4.2 | 0.7 | 4.9 | 3.1 |
| ***LM$_{lnSOC}$*** | 0.93 | 5.8 | -0.6 | 3.6 | 2.3 |