HAL open science

Using carbonate absorbance peak to select the most suitable regression model before predicting soil inorganic carbon concentration by mid-infrared reflectance spectroscopy

Cécile Gomez, Tiphaine Chevallier, Patricia Moulin, Dominique Arrouays, Bernard G. Barthès

To cite this version:

Cécile Gomez, Tiphaine Chevallier, Patricia Moulin, Dominique Arrouays, Bernard G. Barthès. Using carbonate absorbance peak to select the most suitable regression model before predicting soil inorganic carbon concentration by mid-infrared reflectance spectroscopy. Geoderma, 2022, 405, pp.115403. 10.1016/j.geoderma.2021.115403 . hal-03332031

HAL Id: hal-03332031

https://hal.inrae.fr/hal-03332031v1

Submitted on 7 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

Using carbonate absorbance peak to select the most suitable regression model before predicting soil inorganic carbon concentration by mid-infrared reflectance spectroscopy

Cécile Gomez [1,2], Tiphaine Chevallier [3], Patricia Moulin [4,5], Dominique Arrouays[6], Bernard G. Barthès [3]

[1] LISAH, Univ. Montpellier, IRD, INRAE, Institut Agro, 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, Institut Agro, 34060 Montpellier, France

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

[5] LMI IESOL, ISRA-IRD Bel-Air Center, Dakar, Senegal

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

Keywords: Soil Inorganic Carbon; Mid-Infrared reflectance spectroscopy; Partial least squares regression; Linear Regression; National dataset.

Abstract

Mid-Infrared reflectance spectroscopy (MIRS, 4000–400 cm^{-1}) is being considered to provide accurate estimations of soil inorganic carbon (SIC) contents, based on prediction models when the test dataset is well represented by the calibration set, with similar SIC range and distribution and pedological context. This work addresses the case where the test dataset, here originating from France, is poorly represented by the calibration set, here originating from Tunisia, with different SIC distributions and pedological contexts. It aimed to demonstrate the usefulness of *1)* classifying test samples according to SIC level based on the height of the carbonate absorbance peak at 2510 cm^{-1}, and then *2)* selecting a suitable prediction model according to SIC level. Two regression methods were tested: Linear Regression using the height of the carbonate peak at 2510 cm^{-1}, called *Peak-LR* model; and Partial Least Squares Regression using the entire MIR spectrum, called *Full-PLSR* model. First, our results showed that *Full-PLSR* was *1)* more accurate than *Peak-LR* on the Tunisian validation set (R^2_{val} = 0.99 vs. 0.86 and $RMSE_{val}$ = 3.0 vs. 9.7 g kg^{-1}, respectively), but *2)* less accurate than *Peak-LR* when applied on the French dataset (R^2_{test} = 0.70 vs. 0.91 and $RMSE_{test}$ = 13.7 vs. 4.9 g kg^{-1}, respectively). Secondly, on the French dataset, predictions on SIC-poor samples tended to be more accurate using *Peak-LR*, while predictions on SIC-rich samples tended to be more accurate using *Full-PLSR*. Thirdly, the height of the carbonate absorbance peak at 2510 cm^{-1} might be used to discriminate SIC-poor and SIC-rich test samples (< 5 vs. > 5 g kg^{-1}): when this height was > 0, *Full-PLSR* was applied; otherwise *Peak-LR* was applied. Coupling *Peak-LR* and *Full-PLSR* models depending on the carbonate peak yielded the best predictions on the French dataset (R^2_{test} = 0.95 and $RMSE_{test}$ = 3.7 g kg^{-1}). This study underlined the interest of using a carbonate peak to select suitable regression approach for predicting SIC content in a database with different distribution than the calibration database.

1. Introduction

In recent decades, a huge number of studies have focused on carbon storage evaluations locally, regionally, and globally, and five main global Carbon pools have been identified (oceanic, 38 10^3 Pg; geological, 5 10^3 Pg; soil, 2.5 10^3 Pg; atmospheric, 0.76 10^3 Pg; and biotic, 0.56 10^3 Pg; Lal, 2004). The soil carbon pool comprises the Soil Organic Carbon (SOC) and Soil Inorganic Carbon (SIC) pools, and represents 1.7 times more carbon than atmospheric and vegetation Carbon combined (Yang et al., 2012). At global scale, about one third of total soil carbon is inorganic (Batjes, 1996) and calcareous soils cover more than 30% of the continental surface (Romanyà and Rovira, 2011). The SIC is the dominant form of carbon in arid and semiarid areas (e.g., Mi et al., 2008), with a potential reservoir from 2 to 10 times larger than that of SOC (e.g., Bernoux and Chevallier, 2014). Quantifying SOC in calcareous soils often requires determining SIC, SOC being then calculated by difference between total carbon and SIC. Even if more attention is paid to SOC estimation, as SOC is a key determinant of major soil functions (e.g., carbon sequestration, Lal, 2004, Minasny et al., 2017; erosion, Lal, 2003), SIC pool quantification and an understanding of SIC pool dynamics are important for global carbon budget (Zamanian et al., 2021).

Several analytic methods have been developed to quantify SIC contents in soils (Apesteguia et al., 2018). SIC content has usually been measured by calcimetry (ISO10693, 1995a), but can also be measured by dry combustion with a CNH elemental analyzer equipped with a specific module (CO_3-C module) after phosphoric acid dissolution of SIC (Mc Crea, 1950; Hannam et al., 2016). Mid-Infrared reflectance spectroscopy (MIRS, 4000–400 cm^{-1}), based on the study of absorption bands corresponding to fundamental molecular vibrations (Williams and Norris, 1987), has been proposed as an alternative method to SIC analytic methods (e.g., Viscarra Rossel et al., 2006; Wijewardane et al., 2018). In the MIR range, strong absorbance peaks related to carbonate and associated with fundamental vibrational states have been identified at 700, 880, and 1450 cm^{-1} (Legodi et al., 2001). Several additional bands, for instance at 3000–2900 cm^{-1} due to overtones, and at 2600–2500 and 1830–1760 cm^{-1}, due to combinations of fundamental vibrations, have also been attributed to carbonates (Nguyen et al., 1991; Legodi et al., 2001; Comstock et al., 2019). Legodi et al. (2001) reported high coefficients of correlation (between 0.985 and 0.993) between four absorbance peaks centred at 2510, 1799, 876 and 800 cm^{-1} and soil sample SIC content, with highest correlation for the peak centred at 2510 cm^{-1}.

Several methodologies have been successfully tested to link MIR soil spectra to SIC content. Initially, some developments were done using simple linear regression (LR) using specific absorbance peaks (e.g., Legodi et al., 2001). Currently, the most common MIRS calibration method for building SIC prediction models is based on partial least squares regression (PLSR; e.g., Tasbek et al., 2010, Grinand et al., 2012; Wijewardane et al., 2018; Comstock et al. 2019; Barthès et al., 2020) using entire spectra. Some other calibration methods, part of the machine learning algorithms, have been more recently tested such as artificial neural networks, support

vector machines, Cubist (i.e. regression and decision trees), random forest and memory-based learning (e.g., Wijewardane et al., 2018; Dangal et al., 2019).

MIR spectra-based regression models calibrated on samples from a region *A* have been reported to provide accurate SIC predictions when applied on test soil samples from the same region *A*, so where soil and climate conditions could be considered similar (e.g., Tasbek et al., 2010; Grinand et al., 2012; Comstock et al., 2019). Regression models calibrated on samples collected on a region *A* have also been reported to provide accurate SIC predictions when applied on test soil samples collected on a region *B*, where *A* and *B* have no common area, so have potential differences in soil and climate (McCarty et al., 2002; Gomez et al., 2020). Several statistical methods have been applied to adapt the prediction model to sample specificities, for example local calibration (e.g., Gogé et al., 2012; Nocita et al., 2014; Gomez et al., 2020), or spiking, which consist to enrich the calibration set with some representative soil samples from the prediction set (e.g., Guerrero et al., 2010; Barthès et al., 2020). A discrimination of soil types before applying a regression model for SOC prediction has been proposed by Liu et al. (2018). Liu et al. (2018) discriminated five soil types through partial least squares discriminant analysis (PLS-DA) and then calibrated a PLSR model using (I) the entire dataset or (II) a subset by soil type. In these previous studies, the test dataset was well represented by the calibration set, with similar soil properties range and distribution and the number of calibration samples was always much larger than the one of the test samples (e.g., Gomez et al., 2020; Barthès et al., 2020).

This work aims at developing an approach for MIRS prediction of SIC when the calibration and test sets have different SIC distributions and originate from different contexts, here Tunisian samples, mostly carbonated, and French samples, mostly non-carbonated, respectively. It proposes to (i) pre-analyze the test soil spectra to separate samples into two classes (a-priori SIC-poor and SIC-rich test samples) based on the height of the carbonate absorbance peak at 2510 cm^{-1}, and then (ii) select suitable prediction model for each test sample depending on its absorbance peak at 2510 cm^{-1}. Two regression methods were tested for SIC prediction: a simple Linear Regression using the height of the carbonate absorbance peak at 2510 cm^{-1} and a Partial Least Squares Regression using the entire MIR spectrum.

2. Materials

2.1. Soil datasets

2.1.1. The Tunisian soil samples

Ninety-six soil samples were collected in the northern half of Tunisia over approximately 80,000 km² (from 35 to 37°N and 08 to 11°E) within few months in late 2010 (Barthès et al., 2016; Gomez et al., 2020). The soil sampling was designed to represent the main soil types and land uses over the studied region based on previous studies carried out at the Tunis El Manar University. The soil sampling covered 45 localities, and the field samples within the same locality were kilometres apart and under different land uses. Each soil sample was collected at 0-10 cm depth using a spade.

The soil samples were mainly Calcaric Cambisols and Regosols, Kastanozems, and Chromic and Vertic Cambisols (IUSS Working Group WRB, 2014). This Tunisian set, called *DB_Tunisia*, was used to calibrate and validate the regression models.

2.1.2. The French national soil collection

This collection included 2178 soil samples collected over the 552,000 km² of the French metropolitan territory (Corsica included) from 41 to 51°N and from 5.0°W to 9.5°E, and constitutes the national soil collection, provided by the French national soil quality monitoring network (RMQS; Arrouays et al., 2002). This RMQS collection is composed of all main soil types encountered over the French metropolitan territory. According to the French soil classification, 33 soil reference groups were sampled, including three dominant major soil types: Cambisols (IUSS Working Group WRB, 2014; 27% of the sample set), calcareous soils (Calcosols, 22%) and Luvisols (16%). As described in Arrouays et al. (2002) and Jolivet et al. (2006), 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. This French national set, called *DB_RMQS*, was used to test the SIC model calibrated on Tunisian samples.

2.2. Laboratory Analyses

2.2.1. Soil inorganic carbon analyses

The 96 Tunisian soil samples were air-dried, sieved to 2 mm and then finely ground (< 0.2 mm) using mortar and pestle. Their SIC content was calculated as the difference between total carbon content (TC) determined by dry combustion using a CHN analyser (Thermo Fischer Scientific CHN NA2000, Waltham, MA, USA) and SOC content determined by dry combustion after decarbonatation using chlorhydric acid, following the standard procedure ISO 10694 (ISO, 1995a).

The 2178 RMQS samples were also air dried, 2-mm sieved and then finely ground (< 0.25 mm) using mortar and pestle. Their SIC content was calculated as 0.12 times the soil calcium carbonate content, which was determined on these finely ground (< 0.25 mm) air-dried samples using a Bernard calcimeter according to the standard procedure NF ISO 10693 (ISO, 1995b). The carbonate content was calculated after calibration with a pure calcium carbonate standard and was expressed as equivalent calcium carbonate content.

As the SIC contents of Tunisian and RMQS samples were analyzed by two different methods (difference between TC and SOC vs. calcimetry, respectively), 30% of the *DB_Tunisia* were re-analyzed following the method used for the RMQS samples, the calcimetry method, to check the correspondence of the SIC values. The close correlation between both SIC determinations (R = 0.997) was presented in Gomez et al. (2020).

The SIC content of the Tunisian samples (*DB_Tunisia*) ranged from 0.0 to 92.9 g kg^{-1}, averaged 43.3 g kg^{-1}, had a median of 48.5 g kg^{-1} and a skewness value close to -0.2 (Table 1).

The SIC content of the RMQS samples (*DB_RMQS*) ranged from 0 to 103.9 g kg^{-1}, averaged 6.4 g kg^{-1}, had a median of 0 g kg^{-1} and a skewness value close to 3.1 (Table 1). Moreover, SIC distributions of both datasets were different (Figure 1A and B): the SIC values of *DB_RMQS* followed a non-normal distribution, due to a very large number of null values (Figure 1A), whereas the SIC values of *DB_Tunisia* followed a bimodal distribution with modes at 0 and 55 g kg^{-1}, respectively (Figure 1B).

[Table 1]

[Figure 1]

2.2.3. Mid-infrared spectroscopy

The Mid-infrared (MIR) spectra were acquired following the same procedure on both soil datasets, *DB_Tunisia* and *DB_RMQS,* using a Fourier transform Nicolet 6700 spectrophotometer (Thermo Fischer Scientific, Madison, WI, US). The spectrophotometer is equipped with a silicon carbide source, a Michelson interferometer as a dispersive element, and a deuterated triglycine sulfate detector. The soil samples were air-dried, sieved to 2-mm, then 0.2-mm ground aliquots were oven-dried at 40°C for twelve hours and then placed in a 17-well plate. The soil samples surface was flattened with the flat section of a glass cylinder, and each sample was 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).

The MIR reflectance was collected at 934 wavenumbers between 4000 and 400 cm^{-1} with a 3.86 cm^{-1} spectral resolution. Due to noise in the spectrum, twenty wavenumbers were removed, resulting in 914 wavenumbers from 4000 to 478 cm^{-1} used in this study. Reflectance was converted into “absorbance” ($\log_{10}$[1/reflectance]), and a standard normal variate (SNV) correction was applied to remove additive and multiplicative effects (Barnes et al., 1989).

3. Methods

The Tunisian dataset (96 samples) was used to calibrate and validate SIC prediction models, and the French dataset (2178 samples) was used to test them. The large SIC range of the Tunisian dataset represented reasonable conditions for calibrating prediction models. The larger size and diversity of the French dataset provided great opportunity to test strategies for optimizing SIC predictions when SIC distribution differs between calibration and test sets. 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. Figure 2 illustrates the process flow.

3.1. Dataset preparation for regression models

The *DB_Tunisia* dataset was divided into a calibration set (3/4 of the dataset) and a validation set (1/4 of the dataset; Figure 2a) according to the following procedure. The samples were ranked according to ascending observed SIC. The sample with the lowest SIC was put in the calibration set, the next sample in the validation set, and then the next three samples 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, distributions of Tunisian calibration and validation datasets (*DB_Calib_Tunisia* and *DB_Valid_Tunisia*, respectively) were similar.

A principal component analysis (PCA) was applied to the spectra of *DB_Calib_Tunisia* to identify and remove spectral outliers, which are defined as samples spectrally different from the rest of the samples (e.g., Pearson, 2002). These spectral outliers were identified by calculating the Mahalanobis distance (Mark and Tunnell, 1985) to data condensed by PCA. Each spectrum with Mahalanobis distance higher than 3.5 was identified as spectral outlier and removed from the calibration dataset.

[Figure 2]

3.2. Carbonate absorbance peak calculation

As Legodi et al. (2001) reported high coefficients of correlation (between 0.985 and 0.993) between four absorbance peaks centred at 2510, 1799, 876 and 800 cm^{-1} and SIC content of soil samples, with the highest correlation for the peak centred at 2510 cm^{-1}, the MIR spectra were analyzed regarding this absorbance peak at 2510 cm^{-1}. In the literature, the peak area centred at 2510 cm^{-1} has been reported to cover different ranges depending on the spectral library, from 2646 to 2462 cm^{-1} (Legodi et al., 2001), 2611 to 2423 cm^{-1} (Tatzber et al., 2010), and 2680 to 2424 cm^{-1} (Comstock et al., 2019), and has been attributed to the combination of symmetric and asymmetric stretching vibrations (Socrates, 2001). In our study and considering our own spectra (Figure 3), we calculated the absorbance peak using the absorbance values at 2510 and 2650 cm^{-1} as follows:

$$Peak_{2510} = (A_{2510}) - (A_{2650}) \qquad (1)$$

where A_{2510} and A_{2650} are absorbance values at 2510 and 2650 cm^{-1}, respectively, after SNV correction. High peak may correspond to high SIC content and conversely, low peak to low SIC content (Figure 3). The spectral domain centered at 2510 cm^{-1} relates to carbonate due to combinations of fundamental vibrations (Nguyen et al., 1991; Reeves and Smith, 2009).

[Figure 3]

3.3. Linear model regression (Peak-LR)

A simple linear regression was built to model the relationship between SIC content (Y-variable; response variable) and the carbonate absorbance peak using the absorbance values at 2510 and 2650 cm^{-1} ($Peak_{2510}$) as calculated in equation (1) (X-variable; predictor variable) on *DB_Calib_Tunisia*, validated on *DB_Valid_Tunisia,* and tested on *DB_RMQS* (Figure 2b). This linear regression was called the *Peak-LR* model.

3.4. Partial least squares regression

A Partial least squares regression (PLSR) was built to model the relationship between SIC content (Y-variable; response variable) and the entire MIR spectra (X-variables; predictor variables) on *DB_Calib_Tunisia*, validated on *DB_Valid_Tunisia,* and tested on *DB_RMQS* (Figure 2c). The general concept of PLSR is to extract a small number of orthogonal variables (called latent variables) that are linear combinations of MIRS absorbance, account for the maximum variation in the X-variables, and have maximum covariance with the Y-variable (Tenenhaus, 1998). A detailed description of the PLSR procedure can be found in Wold et al. (2001).

The maximum possible number of latent variables 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. In LOOCV, n−1 samples are used to build a regression model, which is applied to the sample not used for developing the model; then the procedure is repeated for all n samples, resulting in predictions for all n samples. The optimal number of latent variables of the model was that which minimized the prediction residual error sum of squares (PRESS) of LOOCV, 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 *DB_Valid_Tunisia* and *DB_RMQS*. This PLSR model was called the *Full-PLSR* model.

Finally, a wavelength was considered as significant contributor to the prediction 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).

3.5. Prediction by class according to absorbance peak at 2510 cm^{-1}

The $Peak_{2510}$ values of the RMQS soil samples were used to separate the *N1* spectra with low $Peak_{2510}$ value, hypothesizing they corresponded to SIC-poor samples, and the *N2* spectra with higher $Peak_{2510}$ value, hypothesizing they corresponded to SIC-rich samples (Figure 2e). Five values *M* of $Peak_{2510}$ were tested for separating samples: -0.05, -0.025, 0, 0.025, 0.05. The $N1_M$ test samples having a $Peak_{2510}$ value lower than M were predicted by the *Peak-LR* model while the $N2_M$ test samples having a $Peak_{2510}$ value higher than *M* were predicted by the *Full-PLSR* model.

This process would allow a coupling of *Peak-LR* and *Full-PLSR* models, depending on the value *M* of $Peak_{2510}$.

3.6. Models evaluation

The performance of SIC prediction models was evaluated according to figures of merit described in Bellon-Maurel et al. (2010), for *DB_Calib_Tunisia*, *DB_Valid_Tunisia* and *DB_RMQS* (Figure 2d). Before calculating the figures of merit, negative predicted SIC values were replaced by 0.

The coefficient of determination (R^2_{cal}) and root mean square error ($RMSE_{cal}$) were used for *DB_Calib_Tunisia*. R^2_{cal} was computed as 1-ESS/TSS, where ESS is the error sum of squares and TSS the total sum of squares. The ratio of performance to deviation in *DB_Calib_Tunisia* (RPD_{cal}) was calculated as the ratio between the standard deviation in *DB_Calib_Tunisia* and $RMSE_{cal}$. The ratio of performance to interquartile range of *DB_Calib_Tunisia* ($RPIQ_{cal}$) was calculated as the ratio between interquartile range (difference between the third and first quartiles) of *DB_Calib_Tunisia* and $RMSE_{cal}$. This parameter has been proposed for variables with non-normal distributions (Bellon-Maurel et al., 2010).

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

The coefficient of determination and root mean square error of prediction were also used for *DB_RMQS*, and denoted R^2_{test} and $RMSE_{test}$, respectively. R^2_{test} was also computed as 1-ESS/TSS. The ratio of performance to deviation in *DB_RMQS* (RPD_{test}) was calculated as the ratio between the standard deviation in *DB_RMQS* and $RMSE_{test}$. The ratio of performance to interquartile range of *DB_RMQS* ($RPIQ_{test}$) was calculated as the ratio between the interquartile range of *DB_RMQS* and $RMSE_{test}$. And the bias was also calculated for *DB_RMQS* ($bias_{test}$). The $RMSE_{test}$ was also calculated for eleven regular sub-ranges of observed SIC (namely < 5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50 and > 50 g kg^{-1}) to study the variation of prediction performances according to SIC level.

The coefficient of determination and root mean square error of prediction calculated on *DB_RMQS* after coupling *Full-PLSR* and *Peak-LR* models, depending on the value *M* of $Peak_{2510}$ (section 3.5), were noted $R^2_{test_M}$ and $RMSE_{test_M}$, respectively.

4. Results

4.1. Preliminary analysis of MIR spectra

A PCA was performed on pre-treated spectra of *DB_Tunisia,* and the pre-treated spectra of *DB_RMQS* were projected onto the plan made by the first and second components. Most of the spectra of RMQS soil samples with high SIC values (orange and red points, Figure 4) overlapped the spectra of *DB_Tunisia* (black stars, Figure 4). Conversely, most of the RMQS soil samples with low SIC values (blue and cyan points, Figure 4) did not overlap the spectra of *DB_Tunisia* (black stars, Figure 4), which is consistent with the scarcity of SIC-poor samples in *DB_Tunisia* (cf. Figure 1). Moreover, RMQS soil samples with low SIC values (blue and cyan points, Figure 4) had no overlapping with those of RMQS soil samples with high SIC values (orange and red points, Figure 4).

[Figure 4]

Soil samples with high SIC content were characterized by a high peak centered at 2510 cm^{-1} due to combinations of fundamental vibrations of carbonates (Nguyen et al., 1991; Reeves et al., 2009), whatever the dataset (red spectra, Figure 3). Conversely, soil samples with low SIC contents were characterized by an absence of peak centered at 2510 cm^{-1} (blue spectra, Figure 3). So the higher was SIC content, the higher was this peak (Figure 3). Furthermore, the coefficient of determination between observed SIC contents and $Peak_{2510}$ values was around 0.91 for the 2718 samples of *DB_RMQS* and 0.84 for the 96 soil samples of *DB_Tunisia*.

4.2. SIC prediction models applied to DB_Valid_Tunisia

One spectral outlier was identified within *DB_Calib_Tunisia* so 95 Tunisian soil samples were kept to build the SIC prediction models (*Peak-LR* and *Full-PLSR*). The *Full-PLSR* model for SIC prediction was built using *DB_Calib_Tunisia* and an optimal number of 8 latent variables, and then validated on *DB_Valid_Tunisia*. The performance of this prediction model was very accurate, with an R^2_{cal} of 0.99 and $RMSE_{Cal}$ of 2.4 g kg^{-1} in the calibration step (Figure 5A, Table 2) and an R^2_{val} of 0.99, $RMSE_{val}$ of 3.0 g kg^{-1}, RPD_{val} of 8.7 and $RPIQ_{val}$ of 13.9 when applied to *DB_Valid_Tunisia* (Figure 5B, Table 2). One hundred and forty-two spectral bands were considered as important according to their b-coefficients and VIP values (Figure 5D). These bands included the expected peaks centred at 2510 cm^{-1} and 1800 cm^{-1} corresponding to carbonate peaks (Nguyen,et al., 1991; Legodi et al., 2001; Du and Zhou, 2009). These bands also included ranges from 3600 to 3700 cm^{-1}, which might correspond to clay mineralogy, 1850 to 2020 cm^{-1}, which might correspond to quartz and clay mineralogy, and from 500 to 1100 cm^{-1}, which might correspond to soil minerals such as iron oxides (Le Guillou et al., 2015).

[Figure 5]

[Table 2]

The *Peak-LR* model for SIC prediction, based on $Peak_{2510}$, was built using *DB_Calib_Tunisia* and then validated on *DB_Valid_Tunisia*. The performance of *Peak-LR* prediction model was modest, with an R^2_{cal} of 0.83 and $RMSE_{cal}$ of 10.4 g kg^{-1} in the calibration step (Figure 6A, Table 2) and an R^2_{val} of 0.86, $RMSE_{val}$ of 9.7 g kg^{-1}, RPD_{val} of 2.7 and $RPIQ_{val}$ of 4.2 when applyed to *DB_Valid_Tunisia* (Figure 6B, Table 2). So *Full-PLSR* model provided better performances than *Peak-LR* on Validation dataset (Table 2).

[Figure 6]

4.3. SIC prediction models tested on BD_RMQS database

The performance of *Full-PLSR* model applied to *DB_RMQS* provided moderate accuracy, with an R^2_{test} of 0.71, $RMSE_{test}$ of 13.7 g kg^{-1} and RPD_{test} of 1.2 (Figure 5C, Table 2). These SIC predictions were strongly biased ($bias_{test}$ = -10.6 g kg^{-1}, Figure 5C, Table 2). As expected, the $RPIQ_{test}$ was very low (0.1, Table 2), due to non-normal SIC distribution in *DB_RMQS* (Figure 1A). When considering SIC sub-ranges, $RMSE_{test}$ obtained with *Full-PLSR* model varied strongly, from 4.4 g kg^{-1} (for samples with observed SIC values between 45 and 50 g kg^{-1}, Table 3) to 14.8 g kg^{-1} (for samples with observed SIC values under 5 g kg^{-1}, Table 3).

The performance of *Peak-LR* model applied to *DB_RMQS* provided good accuracy with an R^2_{test} of 0.91, $RMSE_{test}$ of 4.9 g kg^{-1} and RPD_{test} of 3.3 (Figure 6C, Table 2). As expected and as also obtained with *Full-PLSR*, the $RPIQ_{test}$ was very low (0.2, Table 2), due to non-normal SIC distribution in *DB_RMQS* (Figure 1A). Thus *Peak-LR* model provided better performances than *Full-PLSR* model on *DB_RMQS* (Table 2). The $RMSE_{test}$ obtained with *Peak-LR* model varied strongly between observed SIC sub-ranges, from 1.5 g kg^{-1} (for samples with SIC values under 5 g kg^{-1}, Table 3) to 14 g kg^{-1} (for samples with SIC values between 40 and 45 g kg^{-1}, Table 3).

The predicted SIC values were more accurate with *Peak-LR* than with *Full-PLSR* model for SIC-poor samples (< 15 g kg^{-1}). Conversely, they were more accurate with *Full-PLSR* than with *Peak-LR* model for carbonated samples (SIC > 15 g kg^{-1}) (Table 3).

[Table 3]

4.4. Full-PLSR and Peak-LR coupling depending on $Peak_{2510}$

The *Peak-LR* model was applied to the $N1_M$ test samples having a $Peak_{2510}$ value equal to or lower than *M*, while the *Full-PLSR* model was applied on the $N2_M$ test samples having a $Peak_{2510}$ value higher than *M* (Figure 2e), where M was equal to -0.05, -0.025, 0, 0.025 or 0.05. The higher the M value, the higher the prediction performances of the coupling of *Peak-LR* and *Full-PLSR* models,

up to a threshold at M = 0 above which there was no improvement (Figure 7A and B). The higher the M value, the higher the number of test samples $N2_M$ predicted by the Full-PLSR model (Figure 7C). As expected and whatever the value M, the $RPIQ_{test}$ was very low (around 0.3, Figure 7B), due to non-normal SIC distribution in *DB_RMQS* (Figure 1A).

The coupling of *Peak-LR* and *Full-PLSR* models provided the best performances for a *M* value of 0, with $R^2_{test_0}$ of 0.95, $RMSE_{test_0}$ of 3.7 g kg^{-1}, RPD_{test_0} of 4.4. and $RPIQ_{test_0}$ of 0.3 (Figure 7A and B; Figure 8). Depending on SIC sub-range, $RMSE_{test_0}$ obtained with this coupling of *Peak-LR* and *Full-PLSR* models and M=0 varied from 2.1 g kg^{-1} (for samples with observed SIC values < 5 g kg^{-1}) to 11.4 g kg^{-1} (for samples with observed SIC values between 40 and 45 g kg^{-1}, Table 3).

Using this optimal *M* value of 0, 452 test samples were predicted by the *Full-PLSR* model while 1726 test samples were predicted by the *Peak-LR* model (Figure 7C). The 452 test samples predicted by the *Full-PLSR* model corresponded mostly to SIC-rich test samples: their SIC distribution was characterised by a minimum of 0 g kg^{-1}, a maximum of 103.92 g kg^{-1}, a mean of 29.6 g kg^{-1}, a standard deviation of 23.2 g kg^{-1} and a skewness of 0.9 g kg^{-1}. The 1726 test samples predicted by the *Peak-LR* model corresponded mostly to SIC-poor test samples, characterised by a minimum of 0 g kg^{-1}, a maximum of 44.6 g kg^{-1}, a mean of 0.3 g kg^{-1}, a standard deviation of 1.6 g kg^{-1} and a skewness of 17.7 g kg^{-1}. Among these 1726 test samples having a $Peak_{2510}$ value equal to or lower than 0, 98.9 % had an observed SIC content lower than 5 g kg^{-1}. Among the 452 test samples having a $Peak_{2510}$ value higher than 0, 89.9 % had an observed SIC content higher than 5 g kg^{-1}. So the $Peak_{2510}$ allowed separating SIC-poor and SIC-rich test samples before applying suitable prediction model.

[Figure 7]

[Figure 8]

5. Discussion

5.1. Peak-LR and Full-PLSR performances on the Tunisian Validation set

Before being applied to the French RMQS database, the *Full-PLSR* and *Peak-LR* models were calibrated and validated on the Tunisian set. Such PLSR regression models calibrated from a region *A* and then validated on samples from the same region *A,* where soil and climate conditions could be considered similar, have been slightly developed and analyzed in the literature. The results obtained from the *Full-PLSR* model are in accordance with the literature (Table 2, Figures 5A and B). Barthès et al. (2016) obtained similar performances for SIC prediction (R^2 of 0.98 and *RPD* of 7.8 in six-group cross-validation) using the same Tunisian MIRS database. McCarty et al. (2002) obtained similar performances (R^2_{val} = 0.98) using a Central United State MIRS database characterised by SIC range from 0 to 65.4 g kg^{-1}.

The validation performance of the *Peak-LR* model was lower than the one obtained with the *Full-PLSR* model (R^2_{val} of 0.85 and 0.99 for the *Peak-LR* and *Full-PLSR* model, respectively, Table 2, Figures 5B and 6B). This is in accordance with the result previously obtained by Gomez et al. (2008), who also compared a SIC *Full-PLSR* model based on the entire visible, near- infrared and shortwave infrared spectra (VNIR-SWIR, 400–2500 nm) and a SIC *Peak-LR* model based on spectral absorption band centred at 2341 nm. Higher performance obtained by *Full-PLSR* model might be explained by the higher content of spectral information used. While the *Peak-LR* model benefits only from the absorption peak centred at 2510 cm^{-1}, the *Full-PLSR* model used this same absorption peak with additional spectral ranges, such as bands centred at 1800 cm^{-1} (Figure 5D) corresponding to another carbonate peak in the MIR range (e.g., Nguyen,et al., 1991; Legodi et al., 2001). Moreover, the *Full-PLSR* model used also spectral ranges from 3600 to 3700 cm^{-1}, 1850 to 2020 cm^{-1} and 500 to 1100 cm^{-1}, which might correspond to clay mineralogy, quartz and clay mineralogy, and soil mineralogy such as iron oxides, respectively (Le Guillou et al., 2015).

5.2. Peak-LR and Full-PLSR performances on French test set

Once calibrated and validated on the Tunisian set, both *Full-PLSR* and *Peak-LR* models were tested on the French RMQS dataset. Such MIR spectra-based regression models calibrated from a region *A* and then tested on samples from a region *B,* where *A* and *B* have no common area so have potential differences in soil and climate, have been little developed and analyzed in the literature. The test performance obtained with the *Full-PLSR* model calibrated from the Tunisian subset and applied to the French RMQS dataset (R^2_{test} and $RMSE_{test}$ of 0.71 and 13.7 g kg^{-1}, Table 2, Figure 5C) was lower than the one obtained with a PLSR model calibrated from the French RMQS dataset and applied to the Tunisian dataset (same methodology but inverted datasets, Gomez et al., 2020). So a Full-PLSR model provided better performance when the test set was well-represented by the calibration set (as in Gomez et al., 2020) than when the test set was poorly-represented by the calibration set.

The test performance of the *Peak-LR* model was higher than the one obtained with the *Full-PLSR* model (R^2_{test} of 0.91 and 0.71 for the *Peak-LR* and *Full-PLSR* model, respectively, Table 2, Figures 5C and 6C). So a *Peak-LR* model provided better performance than a *Full-PLSR* model on a test set poorly represented by the calibration set. In particular, the prediction performances for SIC-poor test soil samples (SIC < 15 g kg^{-1}), poorly represented by the calibration set, were lower based on the *Full-PLSR* model than based on the *Peak-LR* model. Conversely the prediction performances for SIC-rich test soil samples (SIC > 15 g kg^{-1}) well represented by the calibration set were higher based on the *Full-PLSR* model than based on the *Peak-LR* model (Table 3).

The *Full-PLSR* model used spectral bands corresponding to carbonate peaks (centred at 2510 cm^{-1} and 1800 cm^{-1}) and additional spectral bands corresponding to soil mineralogy specific to the pedological Tunisian context (Figure 5D). While these additional spectral bands brought information and improved *Full-PLSR* model compared to the *Peak-LR* model when applied to the

Tunisian Validation dataset, they seemed to affect the SIC predictions negatively when applied to SIC-poor French soil samples. Most test soil samples were SIC-poor, and these SIC-poor samples had a very diverse mineralogy (Arrouays et al., 2002). In contrast, the calibration dataset included few SIC-poor samples, which therefore could not represent the wide diversity of SIC-poor test samples (Figure 4). So a regression procedure based on full spectra of all calibration samples (i.e. *Full_PLSR*) could hardly be applied successfully to SIC-poor test samples that might have very different mineralogies and thus very different spectra (cf. samples with SIC = 0 g kg^{-1}, Figure 3A vs. 3B), because much spectral information from most SIC-poor test samples could hardly be managed using a calibration set that lacked such information. For SIC prediction on SIC-poor test samples poorly represented by the calibration set, basic regression procedure using one carbonate peak height (i.e. *Peak_LR*) was by far more appropriate as the carbonate peak height was present in all spectra.

In contrast, SIC-rich samples from calibration and test sets had comparable spectra (cf. spectra plotted in red corresponding to samples with SIC = 90-100 g kg^{-1}, Figure 3A vs. 3B); so such test samples were much better represented by the calibration set (cf. RMQS samples with SIC ≥ 45 g kg^{-1} vs. Tunisian samples in Figure 4). In addition, the calibration set included many carbonated samples. This may explain the benefit of using the full spectrum (i.e. full information) instead of one peak only (basic information) for SIC prediction on SIC-rich test samples, most of which were correctly represented by the calibration set.

5.3. Carbonate absorbance peak as a driver to discriminate SIC-poor and SIC-rich samples

The SIC contents and $Peak_{2510}$ values were highly correlated for both *DB_RMQS* and *DB_Tunisia* dataset (R^2 of 0.91 and 0.84, respectively), as suggested by Legodi et al. (2001), who also reported high correlation (0.993) between MIR carbonate peak centred at 2510 cm^{-1} and SIC contents (description of their carbonate peak calculation not found in Legodi et al., 2001). Based on this observation, the use of this MIR carbonate peak allowed to separate the SIC-poor and SIC-rich test samples, as 98.9 % of samples with $Peak_{2510} \leq 0$ had an observed SIC content lower than 5 g kg^{-1} and 89.9 % of samples with $Peak_{2510} > 0$ had an observed SIC content higher than 5 g kg^{-1}. So the $Peak_{2510}$ value could be used as a simple proxy of SIC absence/presence in soil samples, like the acid test consisting in placing a drop of dilute (5% to 10%) hydrochloric acid on a soil sample and watching for bubbles of carbon dioxide gas to be released (Soil Survey Staff, 1993). As the absence of bubbles of carbon dioxide evidences low SIC content, absence of a carbonate peak at 2510 cm^{-1} (which corresponds to $Peak_{2510} \leq 0$) indicated a SIC content lower than 5 g kg^{-1}.

Considering the development of large soil laboratory VNIR-SWIR spectral libraries (at national, Knadel et al., 2012; continental, Toth et al., 2013, Stevens et al., 2013 and even global scale, Viscarra Rossel et al., 2016), similar condition of dissimilarity between calibration and test datasets might be met and a pre-analysis of test soil VNIR-SWIR spectra might be considered to

separate SIC-poor and SIC-rich samples, using the carbonate absorption band centred at 2341 nm (Gaffey, 1986; Gomez et al., 2008).

Finally, future works could study several parameters to optimize discrimination between SIC-rich and SIC-poor samples. In addition to our $Peak_{2510}$, some other ways to characterize this peak could be explored, such as the area of the peak or a peak height considering the right side of the peak around 2450 cm^{-1}.

6. Conclusion

This work highlighted that when the test set was poorly represented by the calibration set, a simple linear regression based on the height of a carbonate peak provided better SIC predictions than PLSR using the full spectrum. However, when the test set was well represented by the calibration set, PLSR provided better SIC predictions than that simple linear regression. This work also highlighted the interest of pre-analyzing the test set to separate it in two classes (*a-priori* SIC-poor vs. SIC-rich test samples) based on the height of the carbonate absorbance peak at 2510 cm^{-1}. Then the most suitable prediction model might be selected for each test sample, depending on the class and so on how it was represented by the calibration dataset: PLSR using the full spectrum was selected for test samples with peak height > 0, well represented by the calibration set; and simple linear regression based on carbonate peak height was selected for test samples with peak height < 0, poorly represented by the calibration set. This study confirmed the very high applicability of MIRS for SIC determination, even when the calibration and test sets come from different pedo-climatic contexts, so when the former do not fully represent the latter spectrally.

Acknowledgments

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 government agency for environmental protection and energy management (ADEME), the Institut de recherche pour le développement (IRD, which is a French public research organization dedicated to southern countries) and the Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE, which is a French public research organization 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 through its LE STUDIUM Research Consortium Program.

References

Arrouays, D., Jolivet, C., Boulonne, L., Bodineau, G., Saby, N., Grolleau, E., 2002. A new initiative in France: a multi-institutional soil quality monitoring network. Compt. Rend. Acad. Agric. France 88, 93–105.

Apesteguia, M., Plante, A. F., Virto, I. 2018. Methods assessment for organic and inorganic carbon quantification in calcareous soils of the Mediterranean region. *Geoderma Regional*, 12, 39-48.

Barthès, B.G., Kouakoua, E., Moulin, P., Hmaidi, K., Gallali, T., Clairotte, M., Bernoux, M., Bourdon, E., Toucet, J., Chevallier, T., 2016. Studying the physical protection of soil carbon with quantitative infrared spectroscopy. *J. Near Infrared Spectrosc*. 24, 199–214.

Barthès B.G., Kouakoua E., Coll P., Clairotte M., Moulin P., Saby N.P.A., Le Cadre E., Etayo A., Chevallier T., 2020. Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration – The case of soil inorganic carbon prediction by mid-infrared spectroscopy, *Geoderma*, 369, 114272.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc*. 43, 772–777.

Batjes, N. H., 1996. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* 47, 151-163.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric indicators commonly used for assessing the quality of the prediction, *Trends in Analytical Chemistry* (TRAC) 29(9), 1073-1081.

Bernoux, M., Chevallier, T. 2014. Carbon in drylands. Multiple essential functions. Les dossier thématiques du CSFD. N°10. CSFD/Agropolis International, Montpellier, France.

Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst*. 78, 103–112.

Comstock, J.P., Sherpa, S.R., Ferguson, R., Bailey, S., Beem-Miller, J.P., Lin, F., Lehmann, J., Wolfea, 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 Syst. 3, 11.

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

Du, C. and Zhou, J. 2009. Evaluation of soil fertility using infrared spectroscopy: a review. *Environmental Chemistry Letters*, 7, 97–113.

Gaffey, S.J. 1986. Spectral reflectance of carbonate minerals in the visible and near infrared (0.35-2.55 microns); calcite, aragonite, and dolomite. *American Mineralogist*, 71 (1-2), 151–162.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics Intell. Lab. Syst*. 110 (1), 168e176.

Gomez, C., Lagacherie, P., Coulouma, G., 2008 Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma,* 148, 141–148.

Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., Barthès, B.G. 2020. Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. *Geoderma*, 375, 114469.

Grinand, C., Barthès, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *European Journal of Soil Science*, 63 (2), 141-151.

Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, 158, 66–77.

Hannam, K. D., Kehila, D., Millard, P., Midwood, A. J., Neilson, D., Neilson, G. H., Forge, T. A., Nichol, C., and Jones, M. D. 2015. Bicarbonates in irrigation water contribute to carbonate formation and CO2 production in orchard soils under drip irrigation, *Geoderma*, 266, 120-126.

Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N. 2006. Le Réseau de mesures de la qualité des sols de France (RMQS). Etat d'avancement et premiers résultats. *Etude et Gestion des Sols*, 13, 149–164.

ISO (International Organization for Standardisation), 1995a. ISO 10694:1995 - Soil Quality - Determination of Organic and Total Carbon after Dry Combustion (Elementary Analysis). ISO, Geneva.

ISO (International Organization for Standardisation), 1995b. ISO 10693:1995 – Determination of Carbonate Content – Volumetric Method. ISO, Geneva.

IUSS Working Group WRB, 2014. International Union of Soil Sciences, Working Group World Reference Base for Soil Resources. World Reference Base for Soil Resources 2014. International Soil Classification System for Naming Soils and Creating Legends of Soil Maps. FAO, Rome.

Knadel, M., Deng, F., Thomsen, A. and Greve, M.H. 2012. Development of a Danish national Vis-NIR soil spectral library for soil organic carbon determination, Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping (2012), Sydney, Australia. 403.

Lal, R. 2003. Soil erosion and the global carbon budget. *Environment International*, 29(4), 437-450.

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

Legodi, M.A., de Waal, D., Potgieter, J.H., Potgieter, S.S. 2001. Rapid determination of $CaCO_3$ in mixtures utilising FT-IR spectroscopy. *Minerals Engeering*. 14(9), 1107-1111.

Le Guillou, F., 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 Research* 53, 913-921.

Liu, Y., Shi, Z., Zhang, G., Chen, Y., Li, S., Hong, Y., Shi, T., Wang, J., Liu, Y. (2018) Application of Spectrally Derived Soil Type as Ancillary Data to Improve the Estimation of Soil Organic Carbon by Using the Chinese Soil Vis-NIR Spectral Library. *Remote Sens*. 10, 1747.

Mark, H.L., Tunnell, D., 1985. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Analytical Chemistry* 57, 1449–1456.

McCarty, G.W., Reeves III, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. *Soil Science Society of America Journal*. 66: 640–646.

Mevik, B.-H., Wehrens, R., 2007. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software* 18, 1-24.

Mc Crea, J.M. 1950. On the isotopic chemistry of carbonates and a paleotemperature scale, *J. Chem. Phys*., 18, 849–857.

Mi, N., Wang, S.Q., Liu, J.Y., Yu, G.R., Zhang, W.J., Jobbágy, E. 2008. Soil inorganic carbon storage pattern in China. *Glob. Chang. Biol.* 14, 2380–2387.

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., and Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.

Nguyen, T., Janik, L., Raupach, M. 1991. Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. *Aust J Soil Res.* 29: 49.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology & Biochemistry*, 68, 337-347.

Pearson, R.K. 2002. Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology* 10 (1), 55–63.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R foundation for Statistical Computing, Vienna. http://www.R-project.org/.

Reeves, J.B., Smith, D.B. 2009. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. *Appl Geochemistry* 24, 1472–1481.

Romanyà, J., Rovira, P. 2011. An appraisal of soil organic C content in Mediterranean agricultural soils. *Soil Use And Management* 27, 321-332.

Socrates, G., 2001. Infrared and Raman Characteristic Group Frequencies: Tables and Charts. John Wiley & Sons, Chichester, UK.

Soil Survey Division Staff. 1993. Soil survey manual. United States Department of Agriculture Handbook 18.

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).

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. 254 pp.

Toth, G., Jones, A., Montanarella, L., 2013. LUCAS Topsoil Survey: Methodology, Data, and Results. 10. Publications Office of the European Union, Luxembourg, p. 141 (2788/97922).

Yang Y., Fang J., Ji C., Ma W., Mohammat A., Wang S., Wang S., Datta A., Robinson D., Smith P. 2012. Widespread decreases in topsoil inorganic carbon stocks across China's grasslands during 1980s–2000s, *Glob. Change Biol.*, 18, 3672-3680.

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., 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 Review*. 155, 198-230.

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.

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 Sci. Soc. Am. J.* 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, 241-246.

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

Wold, S., Sjöström, M., Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58, 109-130.

Zamanian, K., Zhou, J. Kuzyakov, Y. 2021. Soil carbonates: The unaccounted, irrecoverable carbon source. *Geoderma*, 384, 114817.

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}	Standard deviation g kg^{-1}	Skewness
DB_Tunisia	96	0.0	92.9	43.3	48.5	25.6	-0.2
DB_RMQS	2178	0.0	103.9	6.4	0.0	16.1	3.1

Table 2. Figures of merit obtained with *Peak-LR* and *Full-PLSR* models calculated on the entire SIC ranges over calibration, validation and test databases.

	***Peak-LR* model**	***Full-PLSR* model**
Calibration		
R^2_{cal}	0.83	0.99
$RMSE_{cal}$ (g kg^{-1})	10.4	2.4
RPD_{cal}	2.5	10.5
$RPIQ_{cal}$	4.0	17.1
Validation		
R^2_{val}	0.86	0.99
$RMSE_{val}$ (g kg^{-1})	9.7	3.0
RPD_{val}	2.7	8.7
$RPIQ_{val}$	4.2	13.9
$Bias_{val}$ (g kg^{-1})	0.6	0.3
Test		
R^2_{test}	0.91	0.71
$RMSE_{test}$ (g kg^{-1})	4.9	13.7
RPD_{test}	3.3	1.2
$RPIQ_{test}$	0.2	0.1
$Bias_{test}$ (g kg^{-1})	1.0	-10.6

Table 3. *$RMSE_{test}$* calculated for eleven observed SIC sub-ranges from *Peak-LR*. *Full-PLSR*. and *Peak-LR* and *Full-PLSR* coupling models (*Peak-LR* was applied if *$Peak_{2510}$* value ≤ *M*. otherwise *Full-PLSR* was applied. with *M=0*).

Observed SIC sub-ranges	**Nbr of Test samples**	***$RMSE_{test}$ (g kg^{-1})***		***$RMSE_{test_0}$ (g kg^{-1})***
		Peak-LR model	**Full-PLSR model**	**Peak-LR and Full-PLSR coupling**
< 5 g kg^{-1}	1756	**1.5**	14.8	2.1
[5-10] g kg^{-1}	75	**5.4**	7.9	6.4
[10-15] g kg^{-1}	38	**7.1**	7.3	7.4
[15-20] g kg^{-1}	48	9	**7.8**	7.8
[20-25] g kg^{-1}	37	8.4	**6.4**	6.8
[25-30] g kg^{-1}	42	8.4	**7.3**	7.3
[30-35] g kg^{-1}	33	10.2	**7.6**	8.1
[35-40] g kg^{-1}	24	6.8	**5.9**	5.9
[40-45] g kg^{-1}	20	14	**10.1**	11.4
[45-50] g kg^{-1}	27	10.7	**4.4**	4.4
> 50 g kg^{-1}	78	17	**6.9**	6.9

Figure 1. A) Frequency of SIC values for the Tunisian samples (*DB_Tunisia*, in green) and the RMQS samples (*DB_RMQS*, in pink). B) Zoom to highlight the distribution of *DB_Tunisia*.

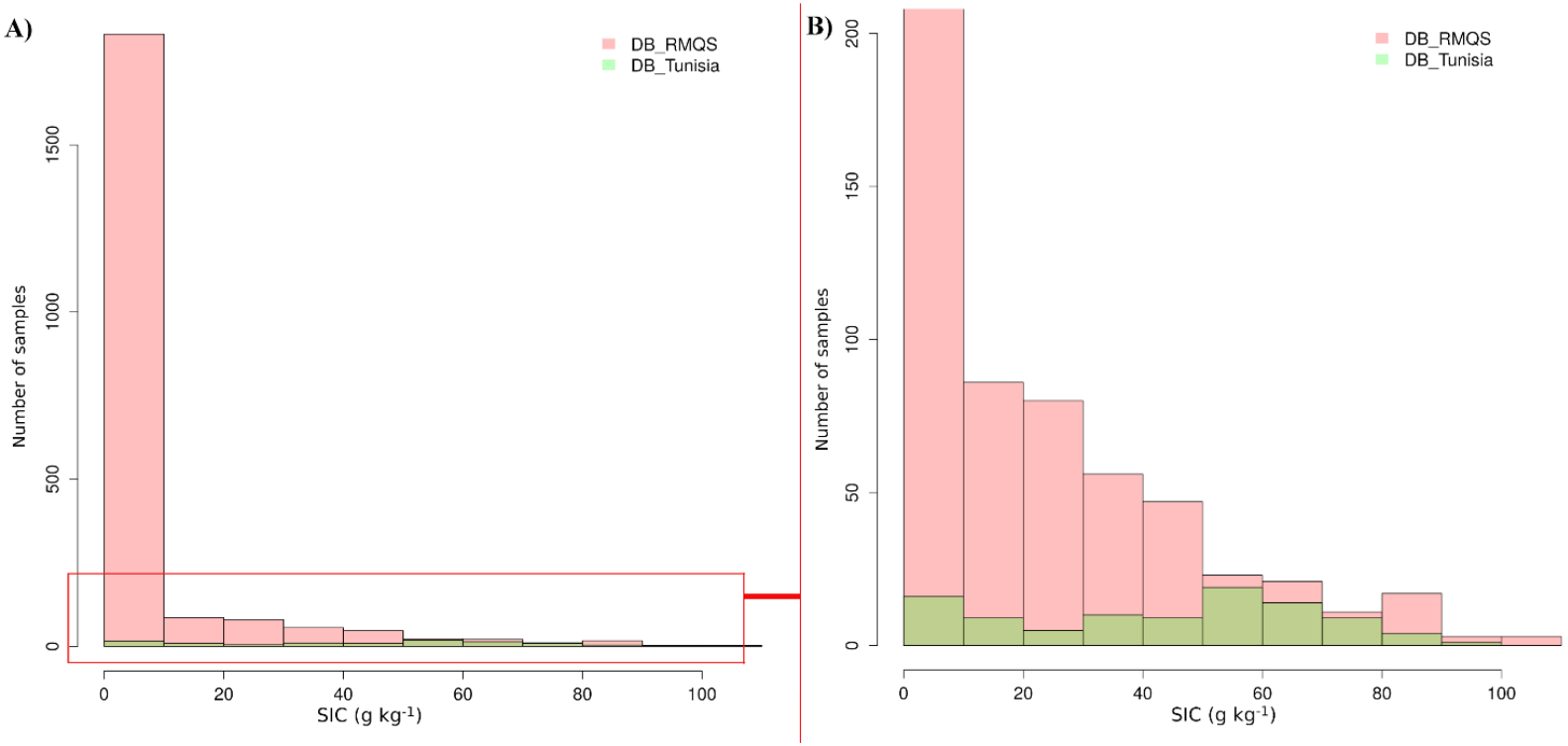

Figure 2. Workflow. *Peak*$_{2510}$ is the height of the absorbance peak at 2510 and 2650 cm^{-1} as defined in equation (1). *M* is a threshold of *Peak*$_{2510}$ used for separating SIC-rich from SIC-poor Test samples and that could take the values -0.05, -0.025, 0, 0.025, 0.05. $N1_M$ and $N2_M$ are the numbers of Test samples with *Peak*$_{2510}$ lower and higher than this value *M*, respectively. *Peak-LR* is a simple linear regression built to model the relationship between SIC content and *Peak*$_{2510}$. *Full-PLSR* is a multivariate regression built to model the relationship between SIC content and the full MIR spectra.

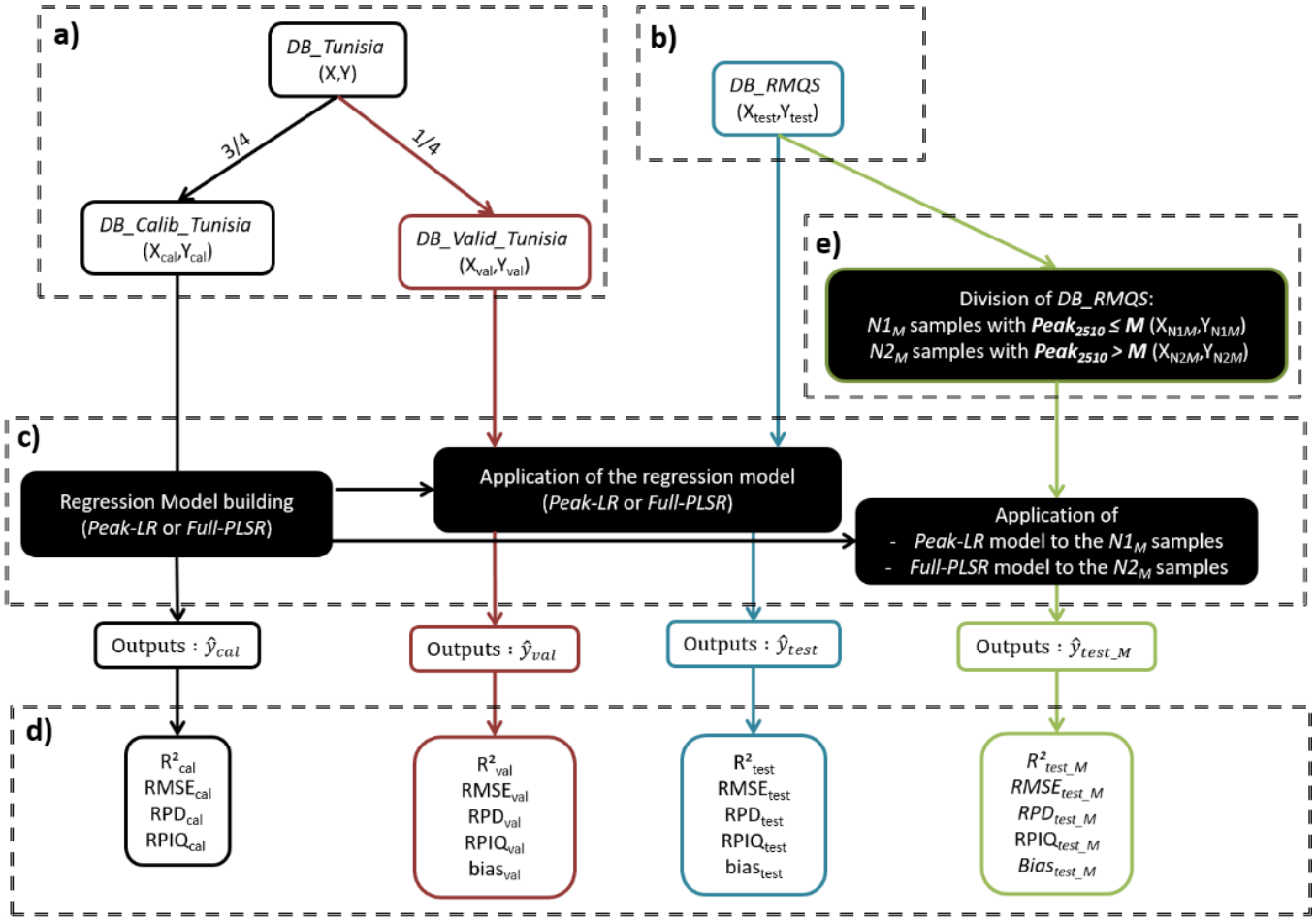

Figure 3. Examples of four SNV-corrected MIR absorbance spectra of soil samples from A) *DB_Tunisia* and B) *DB_RMQS* (their SIC content is specified in the top-left corner). Red vertical plain and dotted lines indicate the centre of the carbonate peak at 2510 cm^{-1} and the boundary at 2650 cm^{-1}, respectively.

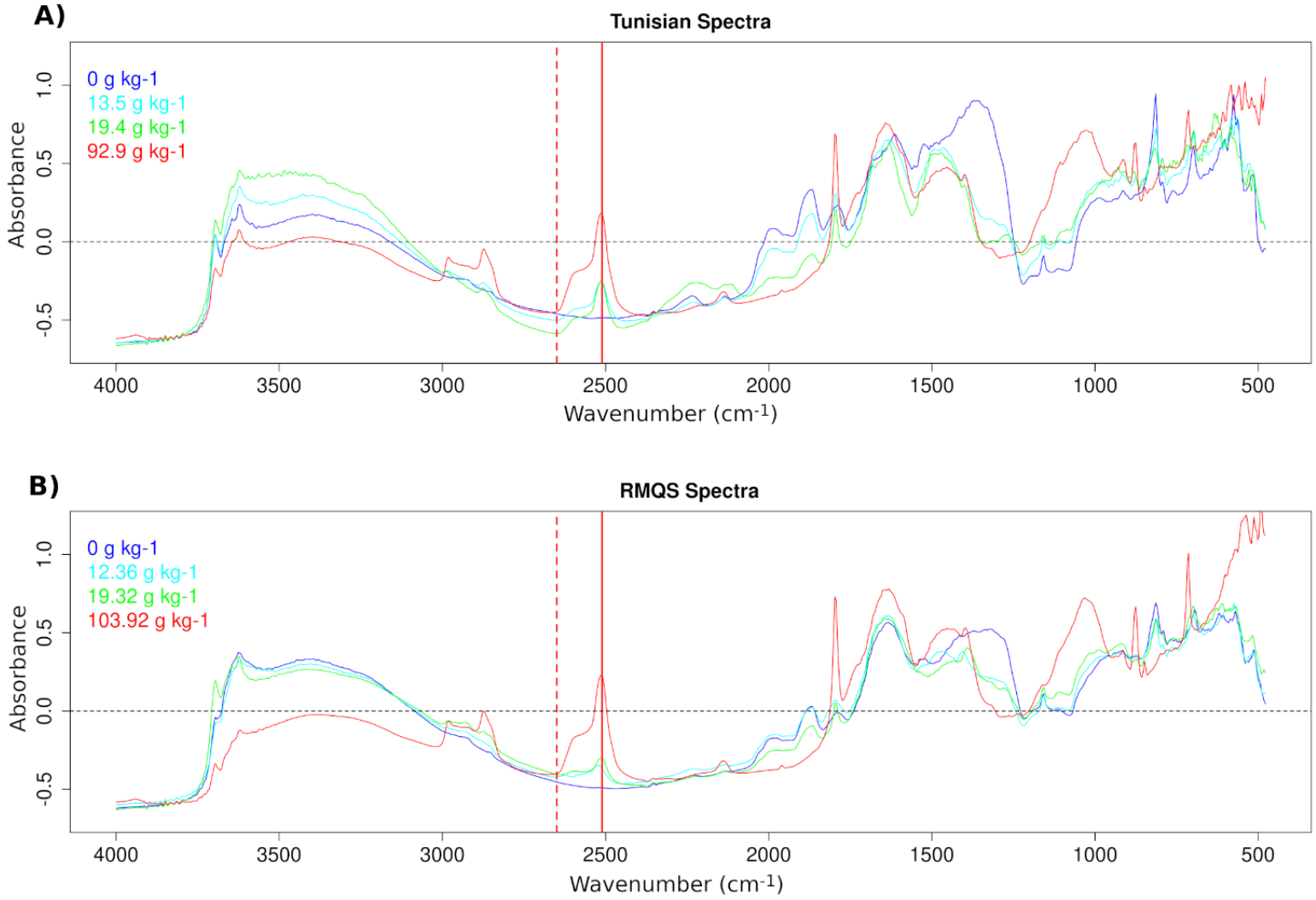

Figure 4. Projection of RMQS spectra (points, coloured depending on their SIC value) onto the first two principal components built with Tunisian spectra (black stars).

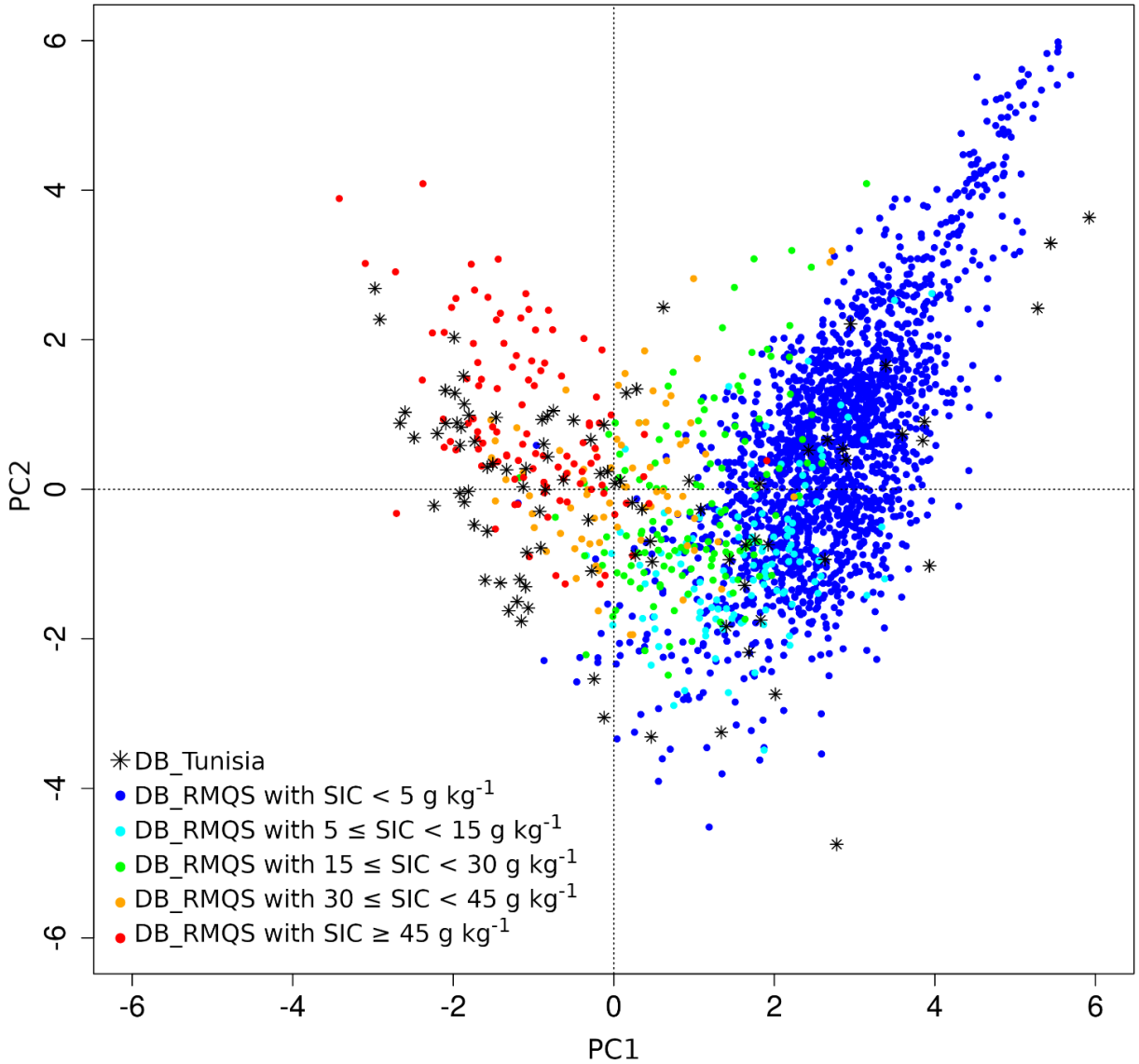

Figure 5. Observed versus predicted SIC values obtained from the *Full-PLSR* prediction model for A) *DB_Calib_Tunisia*, B) *DB_Valid_Tunisia*, C) *DB_RMQS*; and D) significant wavenumbers (represented by vertical red lines) for this model; the black circles represent a Tunisian soil spectrum with 57 g kg^{-1} of SIC.

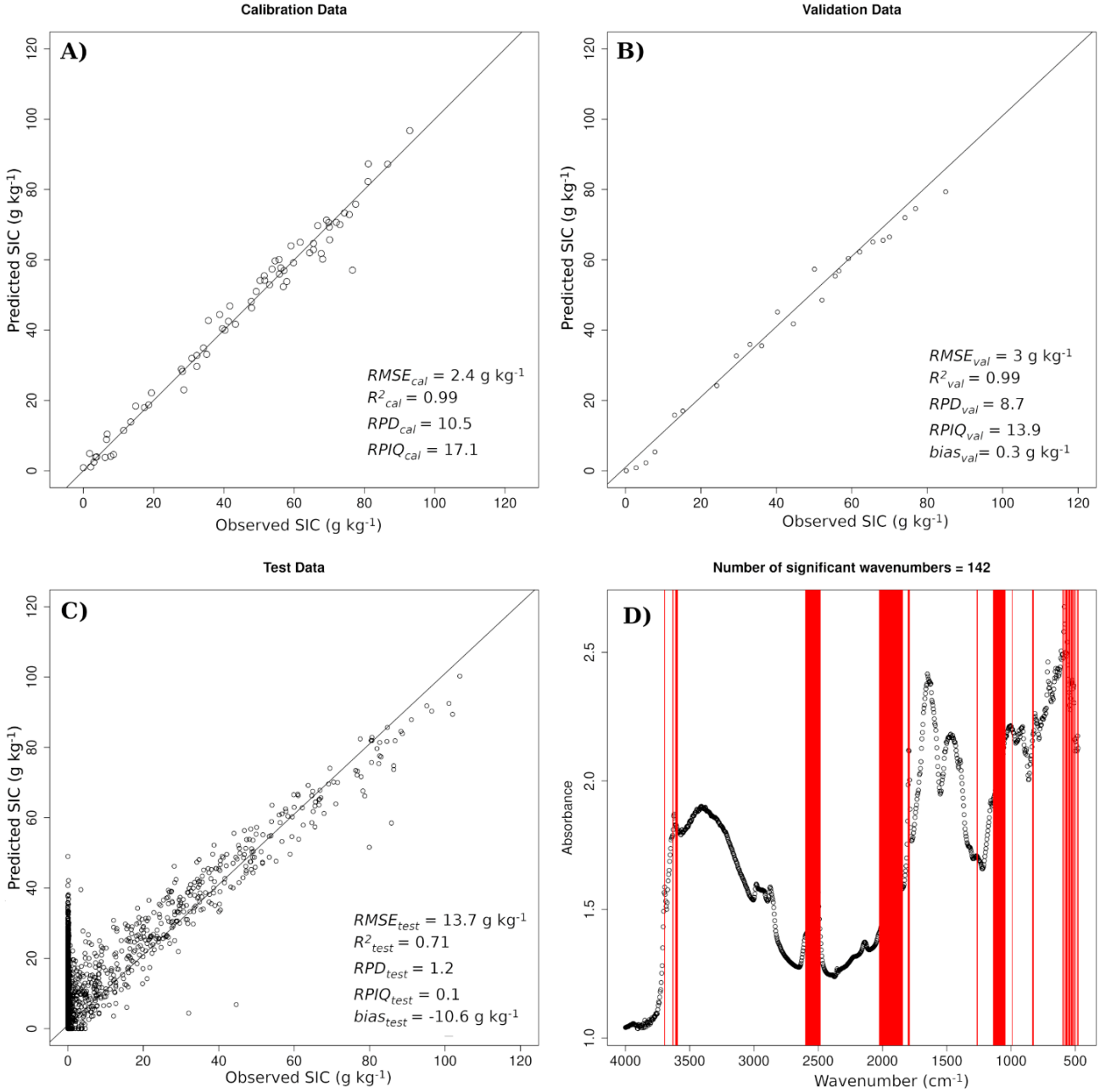

Figure 6. Observed versus predicted SIC values obtained from *Peak-LR model* for A) *DB_Calib_Tunisia*, B) *DB_Valid_Tunisia*, and C) *DB_RMQS*.

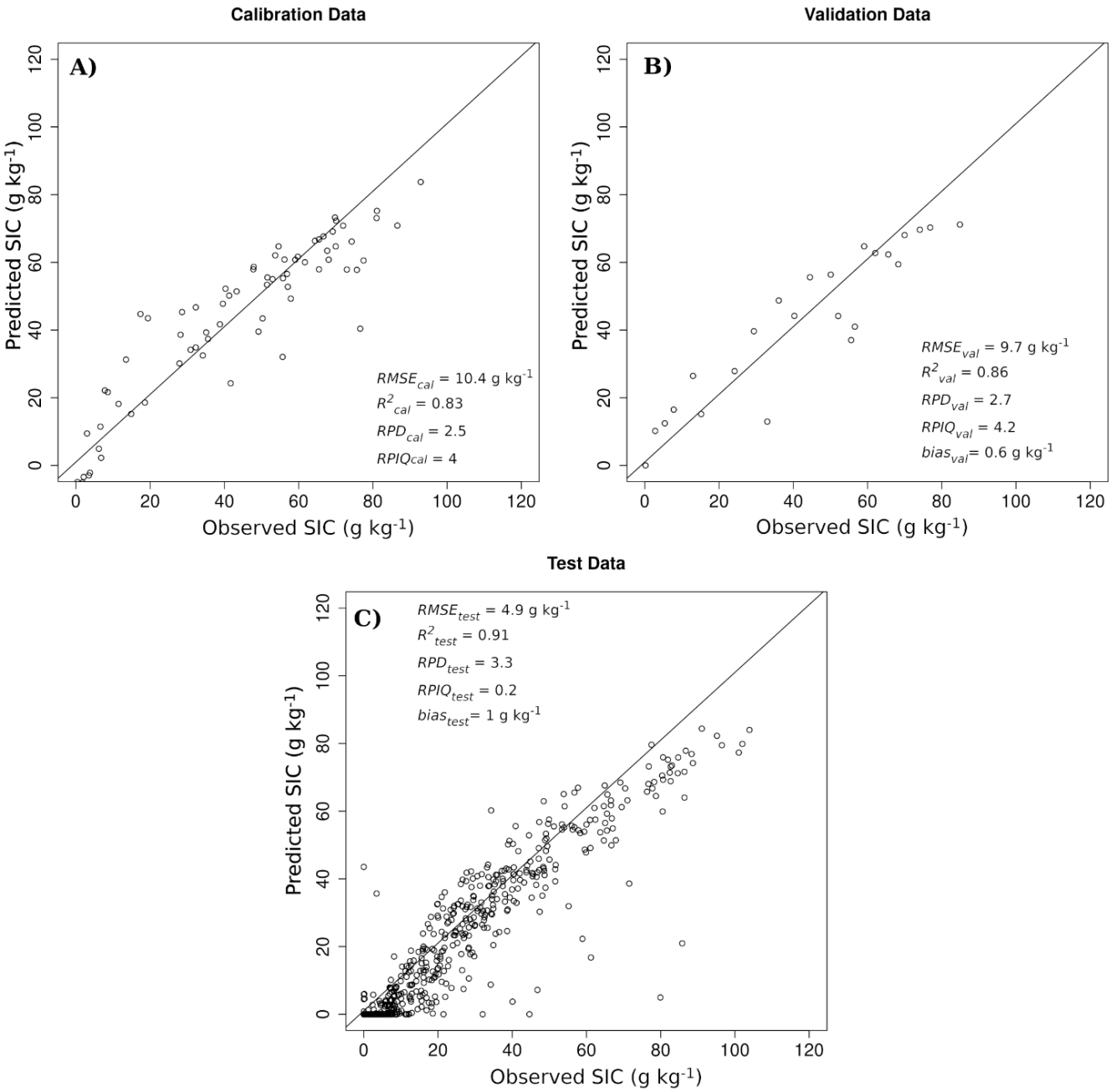

Figure 7. Figures of merit obtained when using *Full-PLSR* and *Peak-LR* depending on the height of $Peak_{2510}$: A) $R^2t_{est_M}$ (black points) and $RMSE_{test_M}$ (red points), B) RPD_{test_M} (black points) and $RPIQ_{test_M}$ (red points), and C) number $N1_M$ of test samples having a $Peak_{2510}$ value equal to or lower than *M* and being predicted by the *Peak-LR* model (green circles), and number $N2_M$ of test samples having a $Peak_{2510}$ value higher than *M* and being predicted by the *Full-PLSR* model (orange stars).

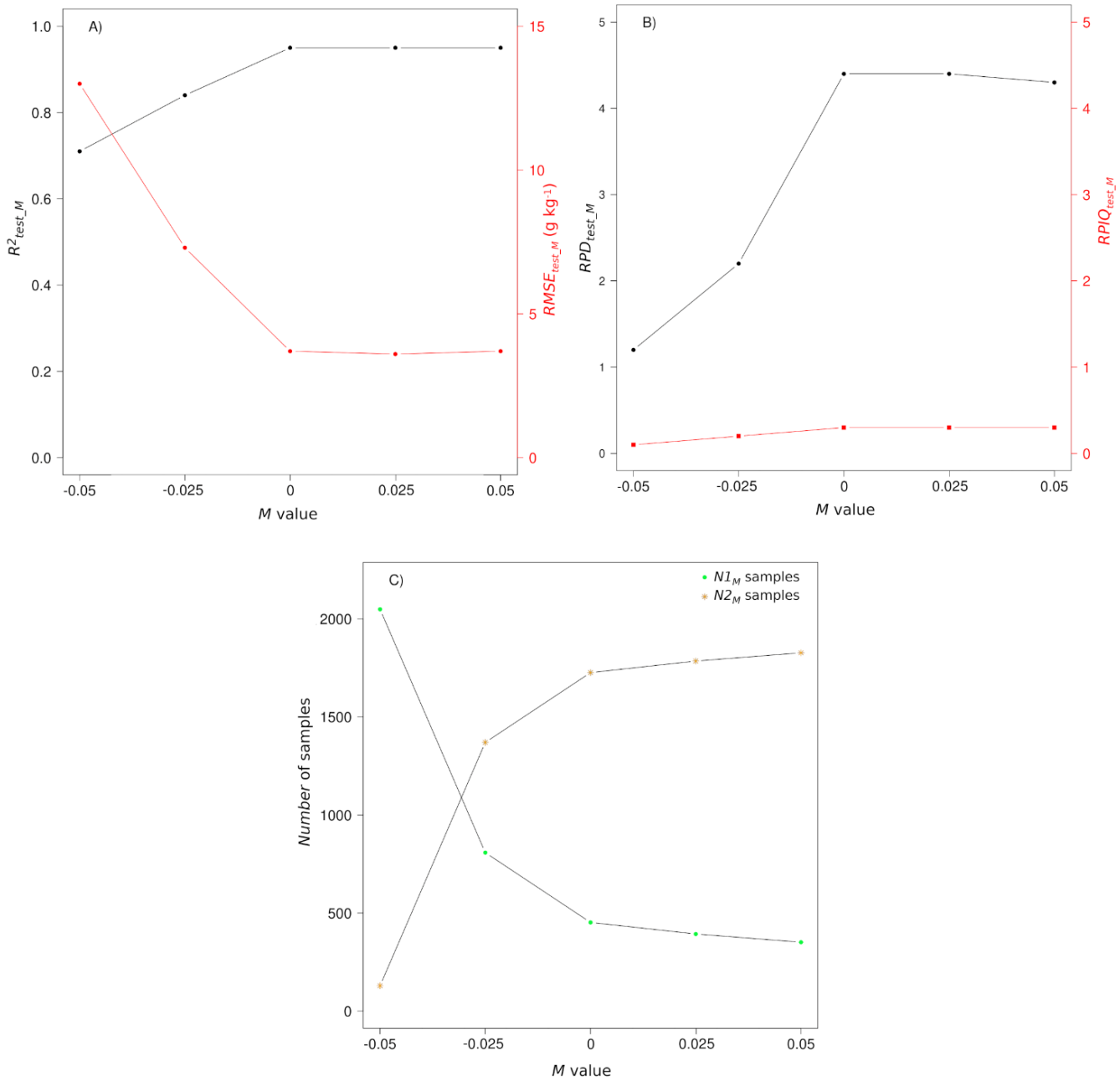
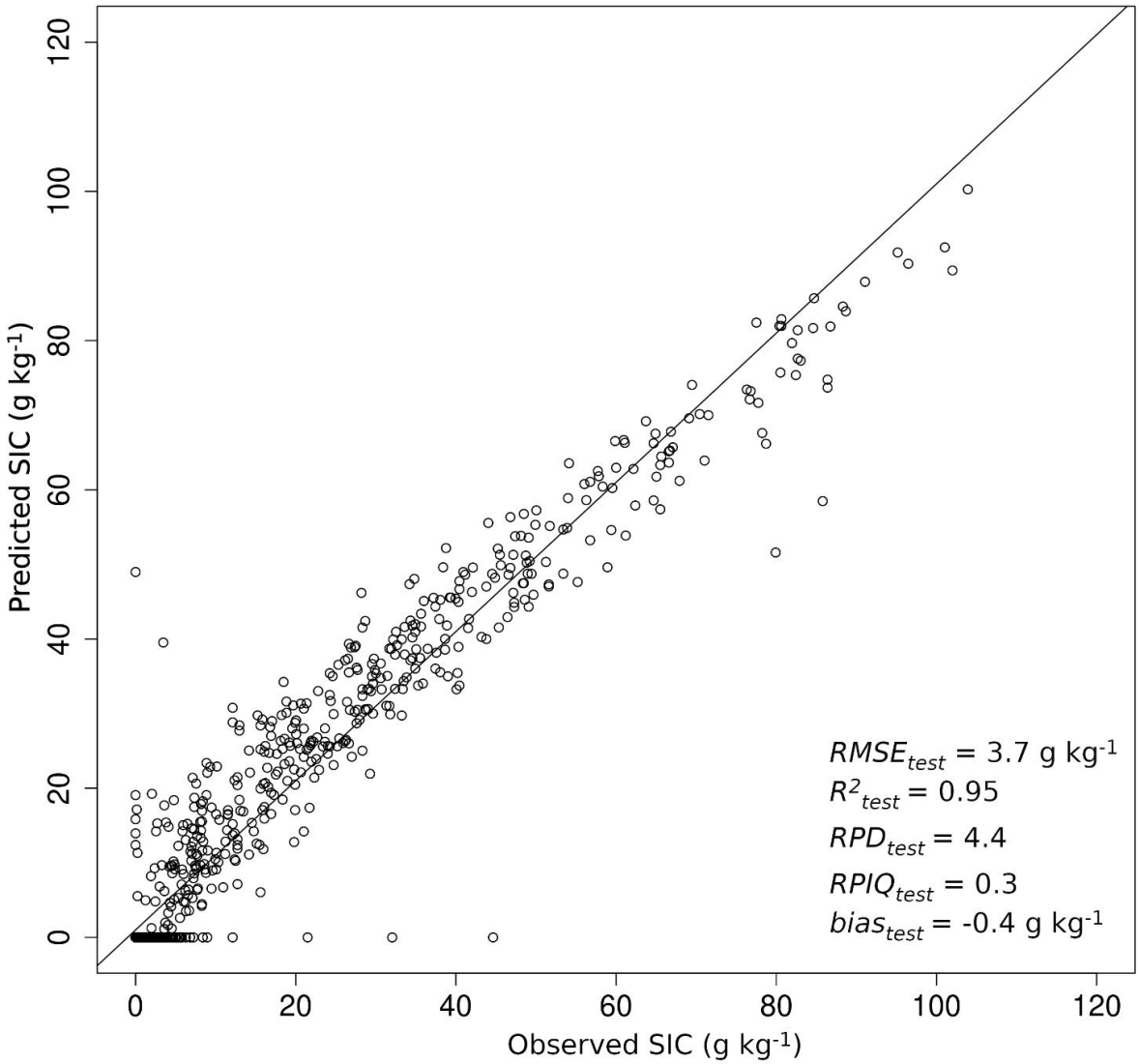
Figure 8. Observed versus predicted SIC values using the *Peak-LR* model when $Peak_{2510}$ was lower than M = 0 and the *Full-PLSR* model when $Peak_{2510}$ was equal to or higher than M = 0.