Technical Note: A validated correction method to quantify organic and inorganic carbon in soils using Rock-Eval® thermal analysis

Marija Stojanova[1], Pierre Arbelet[2], François Baudin[3], Nicolas Bouton[4], Giovanni Caria[5, 6],
Lorenza Pacini[1,2], Nicolas Proix[5], Edouard Quibel[2], Achille Thin[2], and Pierre Barré[1]

[1]École Normale Supérieure de Paris, Laboratoire de Géologie, France
[2]Greenback (commercial name: Genesis), Paris, France
[3]UMR ISTeP 7193, Sorbonne Université, CNRS, Paris, France
[4]Vinci Technologies, 27B rue du Port F-92000 Nanterre, France
[5]INRAE, Laboratoire d'Analyses Des Sols, Saint-Laurent Blangy, France
[6]Univ. Lille, CNRS, UMR 8516, LASIRE, Equipe Physico-Chimie De l'Environnement, Lille, France

Correspondence: Marija Stojanova (stojanova@ens.fr) and Pierre Barré (barre@ens.fr)

Abstract. Soils contain large amounts of carbon stored as organic carbon and carbonates. These carbon pools can contribute to climate regulation, and are of primary importance in ensuring proper soil functioning. However, their accurate quantification remains a complex task. Rock-Eval® thermal analysis has emerged as an alternative to classic dry combustion and wet methods, due to its ability to simultaneously provide organic and inorganic carbon measurements on the same subsample. However, it

5   has been observed that Rock-Eval® systematically underestimates the soil organic carbon (SOC), while overestimating the soil inorganic carbon (SIC). In this technical note, we propose a validated correction of both SOC and SIC based on a machine-learning model and using a diverse dataset of 240 soil samples. We show that the proposed correction significantly increases the accuracy of the Rock-Eval® method on the initial dataset, and that it can be successfully applied to data originating from different Rock-Eval® machines, without changing the routine analytical protocol. The transferability of the model allows for

10   its future implementation in the Geoworks software so that Rock-Eval® machines can routinely provide accurate SIC and SOC measurements.

1 Introduction

Soils contain large amounts of carbon stored as organic carbon and carbonates (Batjes, 1996; Zamanian et al., 2018). These carbon pools are dynamic and can contribute to climate regulation (Chenu et al., 2019; Zamanian et al., 2021). Moreover,

15   organic carbon is also of primary importance in ensuring proper soil functioning (Hoffland et al., 2020). As a result, the accurate quantification of soil organic and inorganic carbon (respectively SOC and SIC) content is recognized as an increasingly important, yet not straightforward task, especially in carbonated soils. There exists a wide range of methods for quantifying SOC and SIC, the two main approaches are dry and wet combustion methods, and they both come with their advantages and drawbacks.

20   Dry combustion methods involving CHN (Carbon, Hydrogen, Nitrogen) elemental analyzers are increasingly used to measure total carbon. If the soil sample is not carbonated, the total carbon therefore represents only SOC. In carbonated soils

however, the amount of SIC is quantified using a separate analysis on a second subsample and the quantity of SOC is then determined as the difference between total carbon and SIC. SIC is generally measured using the amount of CO_2 produced by a known quantity of soil after acidification with HCl (Allison and Moodie, 1965). Alternatively, SOC can be directly quantified using CHN if carbonates have been previously removed using acid fumigation (Harris et al., 2001). The standard ISO 10694 method (ISO, 1995, 1999) describes how organic, inorganic, and total carbon can be measured in soil samples using dry methods. The main disadvantages of dry methods are that (1) SIC and SOC quantities are not measured on the same subsamples, and (2) both SOC and SIC measurements can be inaccurate if the decarbonatation is incomplete.

Wet methods are also largely used, especially when CHN analyzers are not available. The most commonly used wet oxidation method is the Walkley-Black method that allows measuring organic carbon directly in carbonated and non-carbonated soils (Walkley and Black, 1934). However, the Walkley-Black method has an average SOC yield of 77 % and therefore the results have to be corrected to account for not recovered SOC (Food and Agriculture Organization of the United Nations, 2020). The correction factor can be significantly different from one soil to another (Neal and Younglove, 1993), representing a limitation for the accuracy of wet oxidation methods (ISO Soil Quality, 1998).

More recently, thermal analysis methods have been proposed as an interesting alternative to quantifying both SOC and SIC on a single subsample (Disnar et al., 2003; Vuong et al., 2016; Apesteguia et al., 2018). Among thermal methods, Rock-Eval® (RE) thermal analysis has received a particular attention in the recent years (Koorneef et al., 2023; Hazera et al., 2023), suggesting that if suitable corrections are applied, the Rock-Eval® method could provide accurate quantification of both SIC and SOC. Moreover, a recent study has shown that the results provided by the Rock-Eval method are highly repeatable and reproducible, with relative errors of the order of 2 to 4 % for the measurements of organic and inorganic C contents (Pacini et al., 2023). The aim of this technical note is to provide appropriately validated corrections to enable quantification of both SIC and SOC by Rock-Eval® and to test the transferability of corrections between Rock-Eval® machines.

2 Materials and methods

2.1 Soil samples

The study mobilized 240 soil samples taken from the surface layer (0-30 cm) of French agricultural soils. A stratification procedure was used to select the samples and ensure that each combination of texture class, SIC quantity, and SOC quantity was represented with at least 7 samples for low-carbonated soils, and 14 samples for carbonated soils. As a result, the selected samples cover a wide range of particle-size distribution, and SIC and SOC content, as shown in Figure 1. SIC and SOC quantification was conducted at the Soil Analyses Laboratory (LAS) of Saint-Laurent Blangy, accredited by Cofrac (French accreditation committee). LAS is a public laboratory operated by INRAE (French National Research Institute for Agriculture, Food, and Environment) that analyzes samples for the French and European research community. SIC and SOC measured at the LAS can be considered as highly accurate and are therefore suitable to constitute reference measurements.
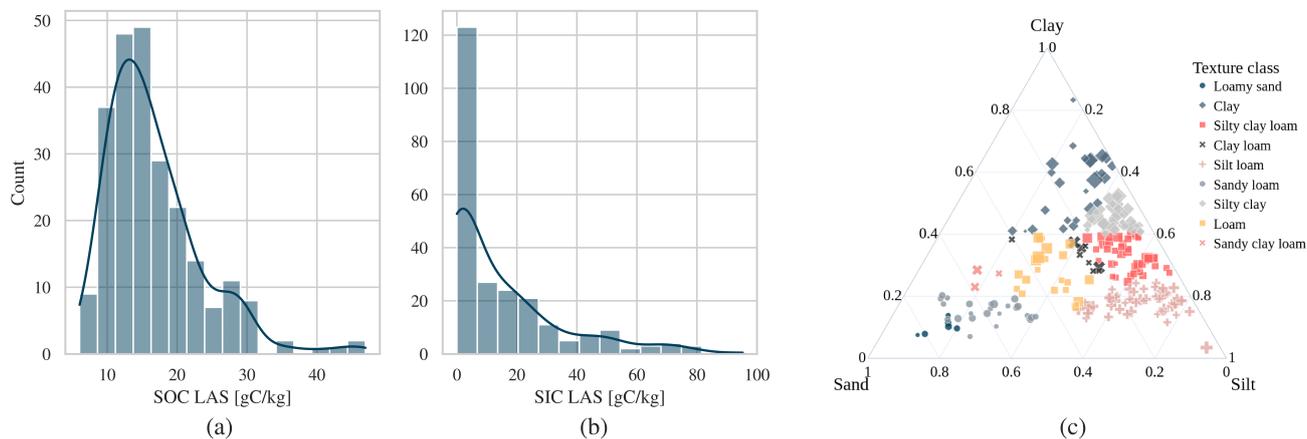
Figure 1. Distributions of SOC (a), SIC (b), and particle-size as a proportion of sand, silt, and clay (c), in the dataset of 240 samples, as measured by the LAS Saint-Laurent Blangy.

2.1.1 Rock-Eval® measurements

Each of the 240 soil samples were ground using a Fritch pulverisette 6 tungsten carbide planetary mill and divided in three subsamples. One subsample was sent to LAS Saint-Laurent Blangy for SIC and SOC determination. The two other subsamples were analyzed using Rock-Eval® at ISTeP (Earth Science Institute, CNRS, Sorbonne University, Paris, France), and at Vinci Technologies (Nanterre, France). At ISTeP, samples were analyzed using a Rock-Eval® 6 Turbo (RE6Turbo), whereas at Vinci Technologies samples were analyzed using an RE6 Standard instrument and a Rock-Eval® 7 (RE7) instrument. The RE7 is the new version of Rock-Eval® machines and we therefore considered it useful to check whether a correction designed on RE6 can be transferred on the new RE7 version.

The Rock-Eval® thermal analysis consists of a pyrolysis step, followed by an oxidation step. During pyrolysis, the sample is gradually heated in an inert environment (N_2) and evolved hydrocarbons, CO_2, and CO effluents are continuously monitored with time. Then, the sample is cooled down before the start of the oxidation stage. During oxidation, the sample is gradually heated in purified laboratory air and the quantity of evolved CO_2 and CO is continuously monitored. The used analytical setup was the "SOIL" routine. The pyrolysis step started with a 3 min isotherm at 200° C followed by a heating ramp of 30° C min^{-1} from 200° C to 650° C (Disnar et al., 2003). The heating routine for the oxidation step started with a 1 min isotherm at 300 ° C followed by a heating ramp of 20° C min^{-1} until 850° C and terminated with a 5 min isotherm at 850° C (Baudin et al. (2015), adapted from Behar et al. (2001)). Pacini et al. (2023) details the RE6 thermal analysis, as well as the usage of the "SOIL" routine in Geoworks.

All samples (ca. 60 mg) on all machines were analyzed using this routine. The thermograms were integrated to provide RE parameters and indices using the Geoworks v1.7 software. Notably, TOCre6 and MinC (corresponding to Rock-Eval® SOC and SIC measurements) were provided by Geoworks.

2.2 Design of the correction methods

It has been long known that TOCre6 and MinC tend to slightly (ca. 10%) underestimate SOC and overestimate SIC, respec-
75 tively (Disnar et al., 2003). The data we collected from the LAS Saint-Laurent Blangy and the Rock-Eval® machines (ISTeP
and Vinci Technologies) also respects this general tendency, as shown in Figure 2. The goal of this study is to design and
validate a correction method for soils that would align the TOCre6 and MinC values measured by Rock-Eval® as closely as
possible with reference SOC and SIC measurements provided by the LAS. To do so, we considered several correction meth-
ods, each one implementing a different Machine Learning (ML) model. We began by learning a correction method using the
80 data obtained with the RE6 Turbo from ISTeP and pairing it with the soil organic and inorganic carbon data from LAS Saint-
Laurent Blangy. We then tested the generalizability of the correction by applying it to the data obtained using other machines,
i.e., the RE6 and RE7 data from Vinci Technologies. In the interest of brevity and legibility, certain modeling choices and
techniques are omitted from the main text and are instead included in our Supplementary materials in the form of a Jupyter
notebook (Kluyver et al., 2016) using open-source code Python 3 code (Van Rossum and Drake, 2009).
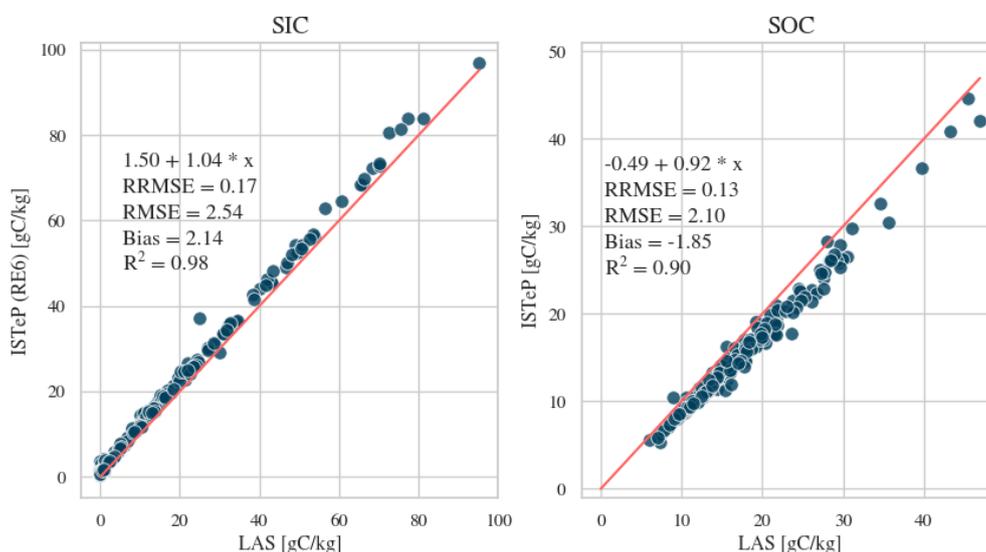
Figure 2. Comparison of the SIC and SOC measurements before correction from the RE6 analysis by ISTeP and reference measurements by
LAS Saint-Laurent Blangy. The red line represents the 1:1 diagonal.

85 After having tested several different approaches, the corrective model we propose here is a Support Vector Machine (SVM)
(Cortes and Vapnik, 1995) regression as implemented in the scikit-learn Python library (Pedregosa et al., 2011). SVMs are a
collection of supervised ML models that are well-suited to our usage as they adapt to both linear and non-linear data. SVMs
using the RBF (Radial Basis Function) kernel offer two different regularization hyper-parameters, which allows us to avoid
overfitting the training data. Other models we have tested are a regular least-squares Linear Regression (LR), a constrained
90 Ridge Regression (Ridge), and a Random Forest (RF) model. The LR model is an ordinary least squares linear regression that

minimizes the residual sum of the differences between the observed and the predicted data. The Ridge regression is a special case of LR where regularization is added to the coefficients using the l2-norm, so as to avoid overfitting the training data. Lastly, the RF is an ensemble modeling method that uses averaging of a collection of randomized decision trees, i.e., random subset of features and of samples, to predict the target variable. Like SVMs, RFs are capable of capturing linear and non-linear

95 relationships in the data. After correcting the SIC and SOC, it can happen that the corrected value is a small negative number close to zero, especially for low-carbonated soil samples. In this case, we simply set these negative SIC and SOC values to zero.

The dataset contains all the RE parameters (see Pacini et al. (2023)) for the 240 soil samples including the TOCre6 (namely SOC_{RE6}) and MinC (namely SIC_{RE6}), as well as SOC and SIC provided by the LAS. We develop two separate models, for

100 predicting the needed correction of the SOC and the SIC. Our target variable is the difference between the SOC (respectively SIC) measured by the LAS and the RE6 analysis:

$$y_{SOC} = SOC_{CHN} - SOC_{RE6} \; ; \tag{1}$$

$$y_{SIC} = SIC_{CHN} - SIC_{RE6} \; . \tag{2}$$

105 All the models are crossed-validated during the training phase using two-thirds of the available data, and then tested on the remaining one-third of samples. After a correlation analysis between the target variables and the RE6 parameters, as well as running several multivariable models, we concluded that the Rock-Eval® SOC (respectively SIC) alone is sufficient to accurately estimate the needed correction. Therefore, in the remainder of this note, all the proposed corrections are single-feature models. Finally, the corrected SIC and SOC of RE can be calculated as:

110 $$SOC_{corrected} = SOC_{RE6} + \widehat{y}_{SOC} \; , \tag{3}$$

and

$$SIC_{corrected} = SIC_{RE6} + \widehat{y}_{SIC} \; , \tag{4}$$

where $\widehat{y}_{SOC}$ and $\widehat{y}_{SIC}$ are the predicted corrections of the ML model.

For the sake of comparison, we also implemented the corrective approach described by Hazera et al. (2023), where two

115 separate models are proposed for carbonated and non-carbonated samples. The correction for carbonated soils enriched in poorly degraded organic compounds is:

$$SOC_{corrected} = 1.17 * SOC_{RE6} \; , \tag{5}$$

and

$$SIC_{corrected} = SIC_{RE6} - 0.091 * SOC_{RE6} \; . \tag{6}$$

120 While for non-carbonated samples they propose the correction:

$$SOC_{corrected} = (SOC_{RE6} + SIC_{RE6}) * 1.068 \,,$$ (7)

and

$$SIC_{corrected} = 0 \,.$$ (8)

As the Hazera et al. (2023) model divides the soil samples into carbonated and non-carbonated, we also divide our data
125 using a similar threshold of 2 gC/kg of SIC. This approach results in three different datasets: the non-carbonated (80 samples),
carbonated (160 samples), and all data (240 samples).

2.3 Metrics to evaluate the correction methods

We use several different metrics to quantify the quality of the provided correction. The R2 coefficient of determination is the
proportion of the variable in the dependent variable that can be explained by the independent variable. Its upper bound and best
130 score is 1, while its lower bound is minus infinite. An R2 score of 0 means that none of the dependent variable's variability is
explained and the model only predicts the mean value. It is calculated as:

$$R^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y})^2} \,,$$ (9)

where $\widehat{y}_i$ is the predicted value of the i-th sample and $\widehat{y}$ is the average value over all samples. The Root Mean Squared Error
(RMSE) and Relative RMSE (RRMSE) are calculated as follows:

135 $$RMSE = \sqrt{\frac{(y_i - \widehat{y}_i)^2}{N}} \,;$$ (10)

$$RRMSE = \frac{RMSE}{\overline{y}} \,,$$ (11)

where N is the number of samples. We also calculate the bias as the difference between the mean of the measured values and
the mean of the predicted values. The bias allows us to know if the model tends to systematically over- or under-estimate the
140 target variable.

3 Results and discussion

3.1 Compared performances of the corrective methods

Having introduced several models for both the SIC and the SOC, we proceed with assessing their corrective performance on
the test dataset. The metric we are mainly interested in is the RMSE, and to a lesser extent the Bias, the R2 coefficient of

145   determination, and the ordinary least squares linear equation describing the relationship between the reference CHN values
and the (corrected) RE values.

Figure 3 shows the RMSE for the original data and of the corrected data after using one of the four different tested ML models
and the correction factors provided by Hazera et al. (2023). Concerning the SIC correction, the overall best performance is
achieved by the SVM model, as it is the only one to significantly reduce the error in all three datasets (carbonated, non-
150   carbonated, and all samples). With the exception of the RF model for non-carbonated data, all the models we propose offer a
considerable decrease in RMSE. The RF model, due to its intrinsic complexity and bootstrapping, is probably more impacted
than other ML models by the smaller dataset size, as only the 80 non-carbonated samples are considered, and one-third of those
are set aside for testing. Regarding the SOC correction, the models we propose strongly increase the adequation between LAS
and Rock-Eval® ISTeP SOC values. They have virtually equal performance, resulting in a two- to three-fold decrease in RMSE.
155   A more detailed comparison of the behavior of the models as a function of the input data is available in the Supplementary
material.
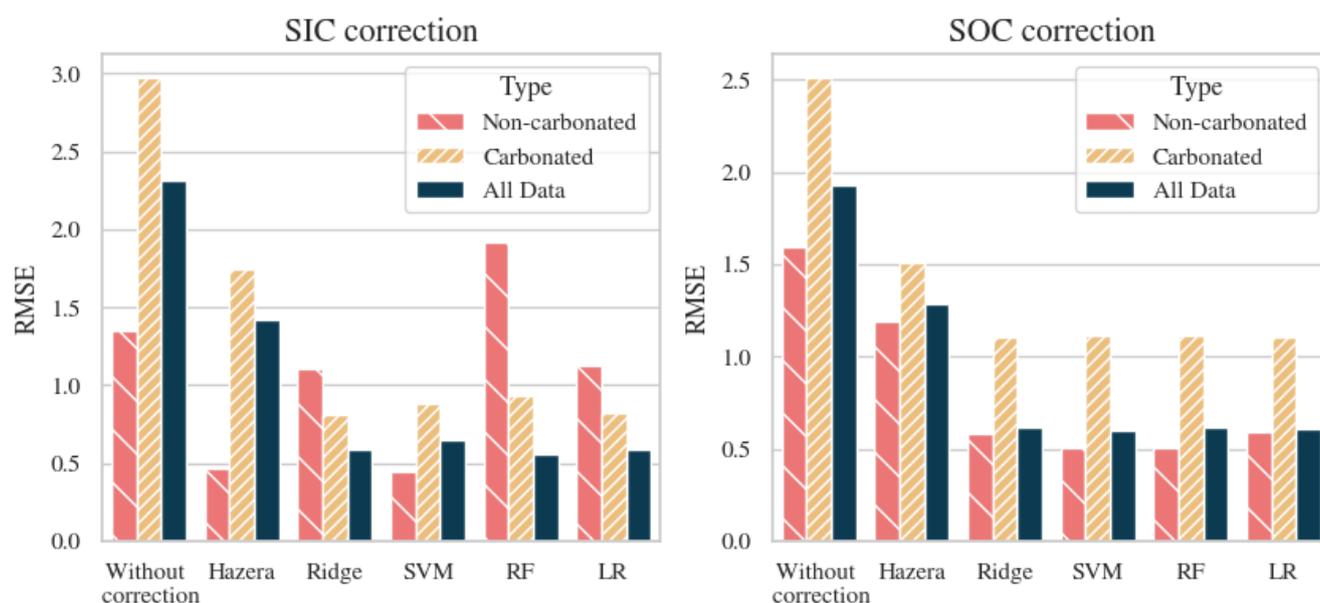
Figure 3. Correction accuracy comparison of the four tested ML models, the Hazera et al. (2023) model, and the original data before
correction.

The correction factors proposed by Hazera et al. (2023) significantly decreased the RMSE for SIC and SOC for all datasets.
However, Figure 3 clearly shows that this correction is not optimal as the RMSE is on average about twice higher compared
to the ML-based corrections. This is not very surprising as, contrary to the other methods, the correction proposed in Hazera
160   et al. (2023) was not validated on test data. Even though the four ML-based corrections render highly similar performance, the
overall best-performing model is the SVM. The results of the SVM correction are shown in Figure 4. The RMSE is divided by

a factor of more than 3 for the SIC and more than 2 for the SOC with respect to the raw data before correction. For both the SIC and the SOC, the equation describing the relationship between corrected RE6 and LAS data is almost identical to the 1:1 ratio, showing that our proposed correction closes the gap between the two methods.
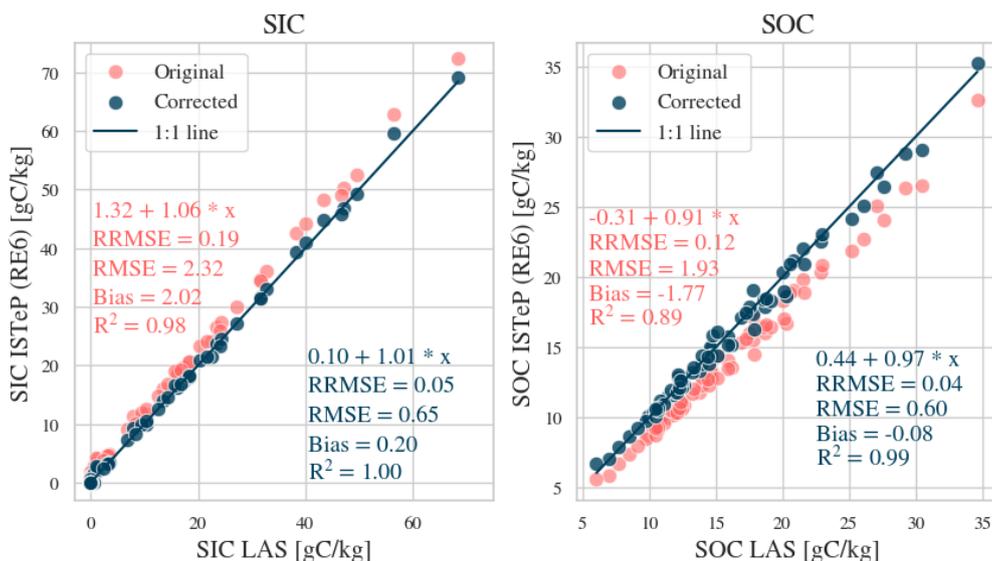
Figure 4. Precision of the proposed SVM correction on the test dataset.

3.2 Transferability of the corrective method to other Rock-Eval® machines

In order to test the transferability of our results to other Rock-Eval® machines, we apply our SVM model learned on ISTeP's RE6 Turbo machine to the data obtained using Vinci Technologies' RE6 and RE7 machines. The reference SIC and SOC data stays the same, i.e., provided by the LAS Saint-Laurent Blangy. Figure 5 shows the comparison for the RE6 (a and b) and RE7 data (c and d), respectively. The correction quality is comparable to that of the ISTeP data with, not surprisingly, only slightly higher RMSEs. The overall bias is not only of the same order but sometimes lower, in part due to the larger sample size as the original correction metrics are calculated on only one third of the dataset used as testing data. The possibility to transfer a correction model for SIC and SOC learned on one Rock-Eval® machine to another was somehow expected, as Pacini et al. (2023) observed that TOCre6 and MinC are very reproducible across different RE6 instruments and for RE6 and RE7 instruments.

4 Conclusions

Accurately estimating the quantity of soil organic carbon has important implications for the monitoring of soil health, as well as climate regulation techniques and policies. The classic dry combustion and wet methods, though standardized decades
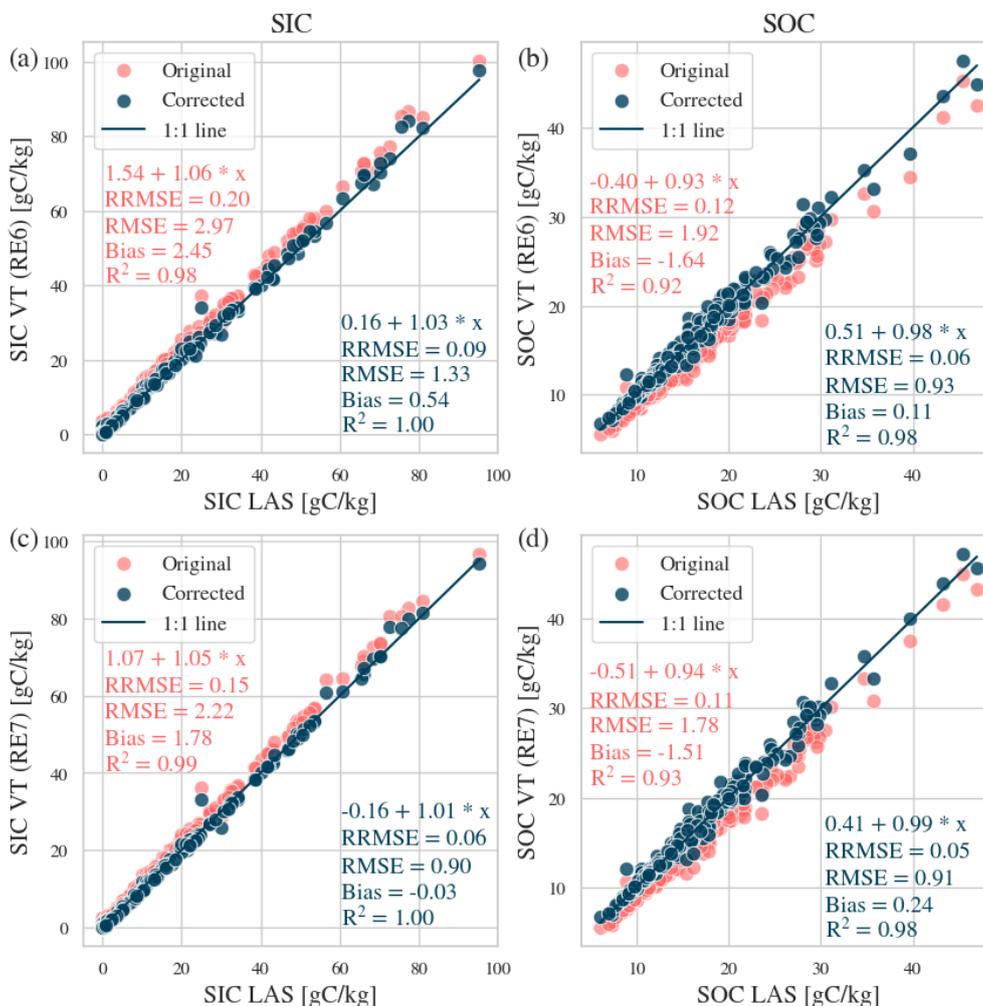
Figure 5. Results of applying the proposed SVM correction to data obtained using Vinci Technologies' RE6 and RE7 machines.

ago, come with a series of drawbacks when considering carbonated soils. Previous studies suggested that, on top of providing information on soil organic matter biogeochemical stability (Barré et al., 2016), thermal analyses could be an accurate means of

180 determining SOC and SIC quantities in carbonated soil with lower experimental uncertainty. Our work confirms this hypothesis and proposes the first validated correction method to accurately determine the quantities of SOC and SIC from a Rock-Eval® thermal analysis. This correction method, based on SVM machine-learning, can be transferred to different RE6 and RE7 instruments and will be implemented in the Geoworks software so that Rock-Eval® machines can routinely provide accurate SIC and SOC measurements, at least in the SOC and SIC value ranges investigated in this study (up to ca. 5% and ca. 8% for

185 SOC and SIC respectively).

Code and data availability. The Jupyter notebook implementing the code and the data accompanying this technical note are both made available (Stojanova et al., 2024).

.

Author contributions. MS, PA, FB, NB, LP, and PB designed the study, MS conducted the data treatment in collaboration with PA, LP, and AT. PA and EQ provided the samples, GC and NP realized SIC and SOC reference measurements, FB and NB provided the Rock-Eval results, MS and PB wrote the first version of the draft which was corrected by all co-authors.

Competing interests. This work was funded by Vinci Technologies, which develops and markets the Rock-Eval machine.

Acknowledgements. The authors thank Florence Savignac (ISTeP) and Theirry Devillaz (Vinci Technologies) who carried out the Rock-Eval® analyses.

References

Allison, L. and Moodie, C.: Carbonate, Methods of Soil Analysis: Part 2 Chemical and Microbiological Properties, 9, 1379–1396, 1965.

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

Barré, P., Plante, A. F., Cécillon, L., Lutfalla, S., Baudin, F., Bernard, S., Christensen, B. T., Eglin, T., Fernandez, J. M., Houot, S., et al.: The energetic and chemical signatures of persistent soil organic matter, Biogeochemistry, 130, 1–12, 2016.

Batjes, N. H.: Total carbon and nitrogen in the soils of the world, European journal of soil science, 47, 151–163, 1996.

Baudin, F., Disnar, J.-R., Aboussou, A., and Savignac, F.: Guidelines for Rock–Eval analysis of recent marine sediments, Organic Geochemistry, 86, 71–80, 2015.

Behar, F., Beaumont, V., and Penteado, H. D. B.: Rock-Eval 6 technology: performances and developments, Oil & Gas Science and Technology, 56, 111–134, 2001.

Chenu, C., Angers, D. A., Barré, P., Derrien, D., Arrouays, D., and Balesdent, J.: Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations, Soil and Tillage Research, 188, 41–52, 2019.

Cortes, C. and Vapnik, V.: Support-vector networks, Machine learning, 20, 273–297, 1995.

Disnar, J.-R., Guillet, B., Kéravis, D., Di-Giovanni, C., and Sebag, D.: Soil organic matter (SOM) characterization by Rock-Eval pyrolysis: scope and limitations, Organic geochemistry, 34, 327–343, 2003.

Food and Agriculture Organization of the United Nations: A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes, 2020.

Harris, D., Horwáth, W. R., and Van Kessel, C.: Acid fumigation of soils to remove carbonates prior to total organic carbon or carbon-13 isotopic analysis, Soil Science Society of America Journal, 65, 1853–1856, 2001.

Hazera, J., Sebag, D., Kowalewski, I., Verrecchia, E., Ravelojaona, H., and Chevallier, T.: Adjustments of the Rock-Eval® thermal analysis for soil organic and inorganic carbon quantification, Biogeosciences Discussions, 2023, 1–18, 2023.

Hoffland, E., Kuyper, T. W., Comans, R. N., and Creamer, R. E.: Eco-functionality of organic matter in soils, Plant and Soil, 455, 1–22, 2020.

ISO: Qualité du sol, dosage du carbone organique et du carbone total après combustion sèche (analyse élémentaire), AFNOR, 1995.

ISO: 10693 (1995) Qualité du sol–Détermination de la teneur en carbonate–Méthode volumétrique, Association Française de Normalisation: Paris, 1999.

ISO Soil Quality: Determination of Organic Carbon by Sulfochromic Oxidation, International Standard, 1998.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C.: Jupyter Notebooks – a publishing format for reproducible computational workflows, in: Positioning and Power in Academic Publishing: Players, Agents and Agendas, edited by Loizides, F. and Schmidt, B., pp. 87 – 90, IOS Press, 2016.

Koorneef, G. J., de Goede, R. G., Pulleman, M. M., van Leeuwen, A. G., Barré, P., Baudin, F., and Comans, R. N.: Quantifying organic carbon in particulate and mineral-associated fractions of calcareous soils–A method comparison, Geoderma, 436, 116 558, 2023.

Neal, R. H. and Younglove, T.: The use of a dry combustion infrared instrumental technique to determine total and organic carbon in California soils, Communications in soil science and plant analysis, 24, 2733–2746, 1993.

Pacini, L., Adatte, T., Barré, P., Boussafir, M., Bouton, N., Cécillon, L., Lamoureux-Var, V., Sebag, D., Verrecchia, E., Wattripont, A., et al.: Reproducibility of Rock-Eval® thermal analysis for soil organic matter characterization, Organic Geochemistry, 186, 104 687, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825–2830, 2011.

Stojanova, M., Arbelet, P., Baudin, F., Bouton, N., Caria, G., Pacini, L., Proix, N., Quibel, E., Thin, A., and Barré, P.: Code and Data for Technical note: A validated correction method to quantify organic and inorganic carbon in soils using Rock- Eval® thermal analysis, https://doi.org/10.5281/zenodo.10706303, 2024.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, ISBN 1441412697, 2009.

Vuong, X., Prietzel, J., and Heitkamp, F.: Measurement of organic and inorganic carbon in dolomite-containing samples, Soil Use and Management, 32, 53–59, 2016.

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

Zamanian, K., Zarebanadkouki, M., and Kuzyakov, Y.: Nitrogen fertilization raises CO2 efflux from inorganic carbon: A global assessment, Global Change Biology, 24, 2810–2817, 2018.

Zamanian, K., Zhou, J., and Kuzyakov, Y.: Soil carbonates: The unaccounted, irrecoverable carbon source, Geoderma, 384, 114 817, 2021.