CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

FACULTY OF AGROBIOLOGY, FOOD AND NATURAL RESOURCES

DEPARTMENT OF SOIL SCIENCE AND SOIL PROTECTION

Exploitation of spectroscopy in proximal and remote sensing for the spatial prediction of soil properties with an emphasis on soil organic carbon content

Doctoral dissertation

Ph.D. student: Ing. James Kobina Mensah Biney
Supervisor: Prof. Dr. Ing. Luboš Borůvka
Co-supervisor: Asa Gholizadeh, Ph.D.

Prague 2022

DECLARATION

I, James Kobina Mensah Biney, hereby declare that the dissertation thesis titled “

Exploitation of spectroscopy in proximal and remote sensing for the spatial prediction of soil properties with an emphasis on soil organic carbon content” is the result of my original investigation and has not been submitted elsewhere for any other degree or professional qualification.

Prague, February 2022

Ing. James Kobina Mensah Biney

ACKNOWLEDGEMENTS

First and foremost, the Almighty God deserves the most praise for his grace and protection throughout this study.

Second, I would like to express my heartfelt gratitude to my supervisor, Prof. Dr. Ing. Luboš Borůvka for his unwavering support during my Ph.D. studies and research, as well as his patience, motivation, enthusiasm, and vast knowledge. His advice was invaluable throughout the research and writing of this thesis. I could not have asked for a better supervisor and mentor for my Ph.D. studies.

My next thanks go to the Department of Soil Science and Soil Protection at the Faculty of Agrobiology, Food, and Natural Resources, CZU, for offering me a place in this doctoral program. Additionally, to Aleš Klement, Karel Němeček and Miroslav Fér, thank you very much. Your contributions during the field sampling and spectral measurements are highly appreciated.

My sincere thanks also go to Mohammadmehdi Saberioon and Radim Vašát for their support and assistance throughout the Ph.D. study, for which I shall always remain grateful. To my co-author's, I say thank you very much for your wonderful contributions.

Finally, I wish to thank my lovely wife, Dorcas, and the Biney family for their support, encouragement, and prayers throughout my academic studies.

PREFACE

The publications and manuscripts presented in this thesis constitute research activities from 2017 to 2022, comprising several sections that are linked to the main aim of this work. That is to improve the estimates of soil properties, especially SOC, with the data set obtained from three well-known platforms [spectroscopy (in-situ), satellite (Sentinel-2) and unmanned aircraft systems (UAS)], using several modelling techniques and pre-treatment algorithms. Knowledge of soil organic matter (SOM)/SOC quality in the soil environment is an essential factor for the assessment of the environmental balance of organic carbon stocks. Additionally, the emersion of proximal and remote sensing techniques has identified mapping and monitoring of soil as essential applications. The entire thesis was carried out under the supervision of the Department of Soil Science and Soil Protection at the Czech University of Life Sciences (CZU), Prague. Grant providers and co-authors are acknowledged in the respective publications.

Moreover, the thesis investigates the possibility of combining in-situ, UAS, and Sentinel-2 (S2) into a single dataset using data fusion approach to estimate SOC in two agricultural fields in the Czech Republic that are low in organic content. To the best of our knowledge, no other study has explored this, possibly because of its complexity. The manuscript dealing with this research is currently under review. In the context of this Ph.D. thesis, the above-mentioned approach did improve the estimate of SOC in the study field of Nová Ves nad Popelkou. This work was a follow-up to a previous work (Biney et al., 20201) in which the three platforms were in-depth compared using their individual data sets. The current work can help encourage other researchers to also focus on fields with low organic carbon content with the sole goal of improving SOC estimates because of its multiple benefits to the environment. The thesis is composed from the following papers:

Biney, J. K. M., Borůvka, L., Chapman Agyeman, P., Němeček, K., and Klement, A. (2020). Comparison of field and laboratory wet soil spectra in the Vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements. Remote Sensing, 12(18), 3082.

Biney, J. K. M., Blöcher, J. R., Borůvka, L., and Vašát, R. (2021). Does the limited use of orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic carbon need attention? Geoderma, 388, 114945.

Biney, J. K. M., Saberioon, M., Borůvka, L., Houška, J., Vašát, R., Chapman Agyeman, P., Coblinski, J. A., and Klement, A. (2021). Exploring the suitability of UAS-based multispectral images for estimating soil organic carbon: Comparison with proximal soil sensing and spaceborne imagery. Remote Sensing, 13(2), 308.

Biney, J. K. M., Vašát, R., Blöcher, J. R., Borůvka, L., and Němeček, K. (2021). Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. Science of the Total Environment, 151805.

Biney, J. K. M., Vašát, R., Bell, S, M., Kebonye, N. M., Aleš, K., John, K., and Borůvka. L. Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations. Soil and Tillage Research (Under revision)

Biney, J. K. M., Borůvka, L., Klement, A., Houška, J., Červenka, J., Kebonye, N. M., and Salazar, D. U. Verifying the impact of fusion high-resolution simulated in-situ spectroscopy, Sentinel-2, and unmanned aircraft systems data into a single dataset to estimate soil organic carbon content in low-carbon agricultural fields. Catena (Under review).

Additionally, other papers were added with no detailed description or discussion. These studies were either linked with some of the study aims or were done to improve the estimation of SOC in the study field.

Biney, J. K. M., 2022. Verifying the predictive performance for soil organic carbon when employing field Vis-NIR spectroscopy and satellite imagery obtained using two different sampling methods. Computers and Electronics in Agriculture, 194, 106796.

Demattê, J.A.M., Paiva, A.F.d.S., Poppiel, R.R., Rosin, N.A., Ruiz, L.F.C., Mello, F.A.d.O., Minasny, B., Grunwald, S., Ge, Y., Ben Dor, E., Gholizadeh, A., Gomez, C., Chabrillat, S., Francos, N., Ayoubi, S., Fiantis, D., **Biney, J.K.M.,** Wang, C., Belal, A., Naimi, S., Hafshejani, N.A., Bellinaso, H., Moura-Bueno, J.M., and Silvero, N.E.Q. (2022). The Brazilian Soil Spectral Service (BraSpecS): A user-friendly system for global soil spectra communication. Remote Sensing, 14, 740.

TABLE OF CONTENTS

1. LITERATURE REVIEW

1.1 Soil organic carbon

The soil is a complex matrix comprising organic and inorganic (mineral matter) materials, as well as water and air. In the soils, organic matter ranges from decomposed and stable humus to fresh particulate residues of different origins. Skjemstad et al. (1997) found that the distribution of these distinct organic pools in soils influences biological behaviour, nutrient availability and dynamics, soil structure and aggregation, and water retention capacity. Inorganic soil carbon is a product of both carbonic acid and the weathering of rocks in the soil, precipitating as carbonate minerals (Lal, 2009). The inorganic mineral fraction is defined in various classification systems by its particle size distribution (proportions of sand, silt, and clay) and by additional subclasses (Hillel and Hillel, 1998). Usually, the coarse sand particles are typically made up of resistant minerals like quartz and feldspar, whereas the fine particles are made up of various clay minerals that have weathered to varying degrees. According to Jenny (1980), this material fraction can be determined by the parent material, soil age, climate, relief, and landscape position (Jenny, 1980). Soil organic carbon (SOC) is the carbon that exists in soil organic matter (SOM) and on average constitutes almost 58% of SOM (Corsi et al., 2012). SOC is strongly influenced by human actions and environmental circumstances such as topography, geology, climate, and time (Walcott et al., 2009). Research has established over the years that the preservation of SOC concentrations is strongly associated with biological activity and agricultural productivity (Stockmann et al., 2013). Maintaining SOC content above critical limits for specific ecological and climatic zones will help to protect soil resources and maintain crop yields, thus contributing to global food security (Bouma and McBratney, 2013). It has been suggested that where SOC concentrations are reduced below some critical limit, soil nutrients and water holding capacity are hindered, and physical degradation is likely to occur through soil aggregate depletion and increased susceptibility to soil surface crusting and erosion (Amundson et al., 2015). SOC stock measurements allow the identification of its spatial and temporal variations as a cause and are an important resource for policy decisions to protect and conserve soils (Saby et al., 2008).

1.1.1. Role of SOC in agriculture

SOC storage is currently one of the utmost topical research fields because of greenhouse gas increases, food demand enlargement and the severity of human-induced soil degradation (Brevik et al., 2015). The rise in SOC has a positive effect on agricultural productivity, as many species (insects, spiders, snails, mites, nematodes, and some mammals) and microorganisms (bacteria, fungi and protozoa) use SOM as food (Walcott et al., 2009). Researchers have identified SOC as a 'universal keystone variable' in soil quality management (Loveland and Webb, 2003), making it the most important indicator of soil fertility management. SOM, on the other hand, is a major parameter for the storage, exchange and reservoir of water and nutrients. It also improves permeability, aeration, infiltration, aggregate stability and structure (Walcott et al., 2009). Despite the importance of SOM, its loss is a universal concern because it does not result only in a loss of soil quality, but further contributes to greenhouse gases (GHG) in the atmosphere, which leads to climate change (Zimmermann et al., 2006). With regard to the control of erosion, SOC contributes to the stabilization of other parts of the soil and to the formation of aggregates that make the soil more resistant to erosion. SOC also participates in the absorption of many pesticides and other xenobiotics and buffers the soil against pH changes (Walcott et al., 2009). Concerning its interaction with soil water, SOC increases the infiltration rate as well as the water-holding capacity of the soil (Walcott et al., 2009).

1.1.2. SOC and climate change

The Intergovernmental Panel on Climate Change reports (IPCC, 2014) stated that the CO_2 level in the atmosphere increased from 280 parts per million (ppm) to 349 ppm for CO_2 between the preindustrial period and 2005. As a result, the global average temperature (from 13.6 °C to 14.4 °C) and sea level (from 15.2 cm to 22.9 cm) increased throughout the twentieth century. The average Arctic cover of sea ice has decreased at a rate of 2.7% per decade. For instance, between 1850 and 2000, fossil fuel combustion was the major source of CO_2 in the atmosphere, but early scientists proved that from the 1940s to 2009, a large amount of CO_2 was released by terrestrial sources rather than from fossil fuels (Lal, 2009). The soil contains more than 1500 Gt of carbon and is known as the greatest terrestrial carbon pool (Smith, 2008). A small release of CO_2 and CH_4 from the decomposition of SOC in the atmosphere will have adverse impacts on the carbon cycle.

For example, in Europe, SOC storage in farms can be approximately 20% of the global reduction needed during the first commitment period of the Kyoto Protocol (8% of reduction

between 2008 and 2012 from a 1990 base) (EU Soil Thematic Strategy, 2004). This role makes it a good proxy for land degradation assessment. Currently, postKyoto agreements are endeavouring to consider SOC in carbon trade (United Nations, 2015). However, much work still must be done, amid which the ability to monitor, report, and authenticate the levels of SOC is the most critical concern (Walcott et al., 2009). The fact that SOC may be considered in carbon trading is an excellent opportunity for developing countries to be involved in carbon trade by selling carbon credits from sustainable soil management through precision agriculture.

1.2 Characteristics of soil spectroscopy

Soil spectral reflectance is mainly affected by chemical determinants such as SOM, soil moisture, soil mineralogy, and physical structure such as particle size and surface roughness (Lobell and Asner, 2002; Shepherd and Walsh, 2002). Infrared spectroscopy is governed by the principle of radiation absorbance at molecular vibration frequencies (Soriano-Disla et al., 2014). Soil spectral signatures are explained by the reflectance of the electromagnetic spectrum as a function of wavelength (Ben-Dor et al., 1997). The vibrational stretching and bending structures of atoms and their electronic transitions define the spectral absorption features.

1.2.1. Remote and proximal sensing of SOC

Many remote sensing approaches have been used to estimate SOC in the last few decades. They are mostly based on remote spectroscopy (multispectral and hyperspectral), satellite or airborne imagery, and field spectroscopy. Field spectroscopy measurements are mostly used to quantify SOC content within a field (on a small scale) and offer many advantages for applications such as precision agriculture (Barnes et al., 2003). Field spectroscopy with a long sampling interval is also used for the assessment of SOC temporal change over a short period of time. More information about the application of field spectroscopy is provided by Milton et al. (2009). In general, analytical spectral devices (ASDs), such as AgriSpec and Fieldspec, are mostly used as measuring instruments (Rossel et al., 2010). In precision agriculture, for example, ASD is mostly mounted on tractors (Bricklemyer and Brown, 2010) to measure soil properties.

Field spectroscopy measurements are generally less accurate than laboratory measurements because of the surface roughness and moisture content (Christy, 2008; Morgan et al., 2009). However, Stevens et al. (2008) demonstrated that field spectroscopy measurements could provide comparable and accurate result as the laboratory measurements. All these results are specific to the different characteristics of the study area. Stevens et al. (2008) compared the

efficiency of laboratory, field, and airborne spectroscopy to predict SOC using PLSR. They concluded that the RMSE of the field spectroscopy was similar to that of the Walkley and Black method and that airborne spectroscopy was inaccurate.

Satellites, as well as airborne sensors, can be viewed as an excellent opportunity to monitor SOC due to the satellite's temporal repetitiveness and broad field of view ability. Nevertheless, few studies have addressed the contribution of satellite images for the assessment of SOC. In most cases, empirical models incorporating phenomena (such as land management, clay, topography, and moisture) that influence the spatiotemporal dynamics of SOC as covariates are used (Croft et al., 2012), especially on a large scale. According to Vasques et al. (2008), satellite images are not always an effective tool for modelling SOC. These authors obtained an R^2 of 0.51 after the assessment of the efficiency of Hyperion sensor to predict SOC. In addition, the study suggests that an investigation of the EnMAP hyperspectral satellite's (a German sensor) capability to maintain an excellent signal to noise ratio (SNR) is needed. However, Mulder et al. (2011) demonstrated that optical remote sensors cover majority of the information required for soil application. In contrast to satellite sensors, airborne platforms have demonstrated good performance, with an R^2 values between 0.62 and 0.97. On the other hand, Selige et al. (2006) developed a multivariate statistical regression to model SOC concentration using the EnMAP sensor and found a result of $R^2 = 0.89$.

Proximal soil sensing (PSS) refers to the use of field sensors to receive soil signals when the sensor is in contact with or near the soil (within 2 m) (Viscarra Rossel and McBratney, 1998; Viscarra Rossel et al., 1998; 2011). The sensors provide data on physical measures related to the soil and its properties. However, it is widely acknowledged that many proximal soil sensors are developed in the laboratory and that some (e.g., visible–near-infrared sensors) use calibrations derived from laboratory measurements that have been widely used to predict SOC content (McCarty et al., 2002). This is because they are less expensive and faster than the traditional methods used for the estimation of SOC. However, steps such as the collection, grinding, sieving, and drying of the soil, which are crucial during this process, make laboratory spectroscopy slightly slower than field spectroscopy measurements (Stevens et al., 2008). Nevertheless, the laboratory spectroscopy approach is not only the most widely used, but also the most accurate thanks to its high analytical precision. These measurements (lab) are recognized as an alternative to the traditional approach to estimating SOC content. For instance, the rationale for using proximal soil sensors is that although their results may not be as accurate as for conventional laboratory analysis per individual measurement, they facilitate the

collection of soil data using cheaper, simpler, and less labour-intensive techniques, which as an ensemble are very informative (Viscarra Rossel et al., 2011). Additionally, using field spectroscopy, measurements are made under field conditions; data are taken from the surface or within the soil profile, and information is produced almost instantly. Therefore, PSS offers advantages that cannot be achieved by remote sensing or laboratory analysis.

A proximal soil sensor is said to be invasive if, during measurement, there is sensor-to-soil contact; otherwise, it is noninvasive. If the measurements are invasive, the sensors may be further described as in situ (i.e., measurements are carried out inside the soil) or ex situ (i.e., measurements are carried out on excavated soil, e.g., measurements on soil cores). Proximal soil sensors may be described as being mobile, in which case they measure soil properties while moving or 'on-the-go' (Adamchuk et al., 2004a), or they may be stationary, whereby measurements are made in a fixed position and possibly at different depths. A proximal soil sensor that produces its own energy from an artificial source for its measurements is said to be active. It is passive when it uses the sun's or earth's radiation, which occurs naturally. When a physical method is used to calculate the target soil property, then the proximal soil sensor is said to be direct, but if the measurement is a proxy and the inference has a pedotransfer function, then the proximal soil sensor is indirect (Viscarra Rossel et al., 2011). When measurements are made indirectly, the target property must be predicted from sensor measurements by calibrating them (using sensor measurements as well as soil samples obtained and analysed in the laboratory). In this case, it will require a calibration sampling design that optimizes property (or feature) space coverage. Ideally, the sampling would also cover geographical space such that the calibrations involve landscape position and other location-induced phenomena. De Gruijter et al. (2010) describe geographic and property space sampling for fine-resolution soil mapping using proximal soil sensors, and Adamchuk et al. (2011a) compare designs for mobile PSS that consider geographic and property space, field boundaries, and other transition zones.

1.2.2. Visible and near-infrared spectroscopy

Researchers' interest in the use of visible-near-infrared (vis–NIR) diffuse reflectance spectroscopy in soil science has grown over the past decade (Stenberg et al., 2010) since the use of this technique has many advantages. It is non-destructive, requires a minimum number of samples to be prepared, and involves no hazardous chemicals. Measurements take only a few seconds, and several soil properties can be estimated from a single scan. In addition, the

technique allows flexible measurement configurations and in situ as well as laboratory-based measurements (Viscarra Rossel et al., 2006).

Reflectance spectra in the visible (400–800 nm) and near-infrared (800–2500 nm) regions are the result of interactions between the radiating energy and the bonds (molecules) of soil constituents. In the visible region, the high energy of the radiation causes transitions of electrons between molecular orbits with different energy levels (Miller, 2001). With lower radiation energy, corresponding to longer wavelengths, the absorption of energy occurs due to vibrations in molecular bonds. Absorption in the NIR region is due to overtones and combinations of fundamental vibrations in the mid-infrared region. The absorbed energy quanta are bond specific but are also influenced by the chemical matrix and environmental factors such as functional group size, adjacent molecules, and hydrogen bonds (Miller 2001). It allows recognition of a number of molecules that can contain the same bond form. With decreasing intensity and increasing overtone order, the same molecule can give rise to several overtone and combination bands over the NIR field. Because of this, small or large, overlapping absorption features can characterize the NIR region. Soil diffuse reflectance is also affected by soil physical properties associated with particle size and surface structure, as well as water films on the soil surface.

The Vis-NIR region provides valuable information about organic and inorganic soil material, and both clay minerals and soil organic matter, which are essential soil constituents, have well-recognized absorption characteristics in this field. Water has a strong influence on the spectra, with some dominant specific bands of absorption near 1400 and 1900 nm, along with weaker bands in other parts of the spectra (Liu et al., 2002). However, the scattering is more forwards-directed with a water film on soil particles, and moist soils appear darker than dry ones (Sherman and Waite, 1985). The mineral part of the soil generally accounts for half the volume of the soil (Schulze, 2002), while the pores with water and air account for the remaining half. It has significant characteristics in the Vis-NIR spectrum, both in terms of surface properties affecting the degree of scattering and absorption. Absorption in the visible region is mainly related to iron-containing minerals such as hematite and goethite, which exhibit strong absorption bands between 400 and 660 nm (Sherman and Waite, 1985). All minerals also exhibit absorption bands close to 900 nm (880 and 930 nm, respectively, for hematite and goethite) but have almost no absorption features at longer wavelengths (Clark et al., 1990).

1.2.3. Past and current role of Vis-NIR in soil science

Over the last few decades, a significant number of approaches have been proposed to predict soil properties with Vis-NIR spectroscopy. Total organic carbon estimation is potentially the most common, followed by clay content and soil N, because of their possible predictive success (Viscarra Rossel et al., 2006). Some other frequently reported properties also include SOM, minerals, soil texture (clay, silt, and sand content), nutrients, water, pH, extractable P, K, Fe, Ca, Na, Mg, and CEC (Stenberg et al., 2010). Some literature has also shown moisture content to be one of the most accurately measured properties with excellent accuracy in the NIR region (Chang et al., 2001; Mouazen et al., 2006a). This can be attributed to the presence of a clear water absorption band at 1450 nm in the second overtone region, resulting in a wide correlation of approximately 1450 nm.

1.3 Data fusion

According to Durrant-Whyte (2001), data fusion is a method that incorporates knowledge from several different sources to provide a coherent and detailed summary of the system or process of interest. Data fusion is of particular importance in any application where large amounts of data need to be combined, fused and distilled to obtain information of appropriate quality and integrity on which to make decisions. Data fusion is used in many military systems, civilian surveillance and monitoring tasks, process control, and information systems. Data fusion methods are particularly important in all of these applications as they drive towards autonomous systems.

Several approaches are used to perform data fusion, which are normally categorized as levels, including a simple combination of the original data (Viscarra Rossel et al., 2006; Ji et al., 2019) (level 1), a simple combination of selected spectral features (Xu et al., 2019b) (level 2), a combination of the measurement results (O'Rourke et al., 2016) (level 3) and many other levels. The combination of the measurement results, also referred to as model averaging (Horta et al., 2015), involves the combination of the different model outcomes to obtain a better outcome. This improves the estimation accuracy and reduces the possibility of aberrant measurements (O'Rourke et al., 2016; Chen et al., 2019). However, its accuracy could decline in fields with low values of soil properties under consideration. In theory, automated data fusion processes allow for the combination of critical measurements and information to provide knowledge of sufficient wealth and integrity to formulate and execute decisions autonomously.

2. HYPOTHESES AND AIMS

Paper 1: Comparison of field and laboratory wet soil spectra in the vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements

Hypothesis: Field collection of spectra or laboratory spectra measurement on naturally wet soil samples for SOC prediction can provide a fast alternative to laboratory dry soil sample measurement with comparable accuracy.

Aim: This study aims to compare field and naturally acquired lab-wet spectral datasets for the prediction of SOC using their raw and pre-treatment states. The study also determined which of these datasets could be more suitable in the absence of a lab-dry measurement or when a quicker analysis is required to estimate SOC.

Paper 2: Does the limited use of orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic carbon need attention?

Hypothesis: Applying the wrong type or applying a pre-processing method that is too severe to vis-NIR spectroscopy data can result in the removal of valuable information or even the introduction of unwanted variation, thereby affecting prediction accuracy.

Aim: This research aims to verify orthogonal signal correction (OSC) effectiveness in terms of the predictive accuracy of SOC against nine most commonly used pre-treatment methods for both VIS and vis-NIR spectra.

Paper 3: Exploring the suitability of UAS-based multispectral images for estimating soil organic carbon: Comparison with proximal soil sensing and spaceborne imagery

Hypothesis: Although spectroscopy under proximal sensing remains one of the best approaches to accurately predict SOC, spectroscopy's limitation to estimating SOC on a larger spatial scale remains a concern. It is believed that for the efficient quantification of SOC content on a larger scale, the use of remote sensing approaches with spectral indices is a viable option.

Aim: This study aims to evaluate and compare the capabilities of unmanned aircraft systems (UASs) for the monitoring and estimation of SOC with those obtained from spaceborne (Sentinel-2) and proximal soil sensing (field spectroscopy measurements) on an agricultural field with a low SOC content and to verify the effect of soil and vegetation indices. The spatial SOC distribution map is also computed for the various sensors used in reference to the laboratory SOC measured values.

Paper 4: Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil

Hypothesis: The ensemble model will perform at least better than each of the separate techniques in terms of appropriately utilizing all the available data.

Aim: The study aims to compare the ensemble model (PLSR, SVM, Cubist, and random forest) to each of the calibration techniques in terms of prediction accuracy of As content using pXRF and field spectroscopy data in an agricultural field with no history of contaminants. Other components [e.g., soil organic carbon (SOC), Mn, S, soil pH, Fe] that are known to influence As levels in the soil were also retrieved to assess their correlation with soil As.

Manuscript 5 (under review): Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations.

Hypothesis: Acquiring spectral data normally under different measurement conditions could introduce artefacts that reduce SOC prediction accuracy. Ensemble approach combining the differently measured data can eliminate these artefacts and improve the prediction accuracy. The results of several comparative studies based on these predictive calibration techniques used alone were inconsistent so far.

Aim: The main aim is to predict the SOC across the three agricultural fields with both proximal and remote sensing data using the ensemble model. The effectiveness of the developed model on regional scale dataset is also explored.

Manuscript 6 (under review): Verifying the impact of fusion high-resolution simulated in situ spectroscopy, Sentinel-2, and unmanned aircraft systems data into a single dataset to estimate soil organic carbon content in low-carbon agricultural fields.

Hypothesis: The fusion of high-resolution spectroscopy data with both UAS and S2 datasets into a single dataset could help improve the predictive performance of SOC.

Aim: This study aims to systematically compare the individual and combined abilities of in situ, UAS, and S2 sensors for estimating the content of SOC in two different agricultural fields with different soil types. Specifically, middle-level fusion techniques (fusion of in situ, UAS, and S2 data among each other after variable importance selection using the Boruta algorithm)

were used. The spatial distribution map for both individual and fused approaches are also computed.

3. METHODOLOGY

3.1. General study areas

Three agricultural sites within the Czech Republic were used for the study project. These include Brumovice (90 ha, 48°96′ N; 1688′ E), a village in the Břeclav District of the South Moravian Region, located at an elevation of 210 meters on a gentle slope of the Chřiby hills approximately 6 kilometers southeast of Klobouky u Brna. Údrnice, 52 ha (50°21′ N; 15°15′ E) in Jičín district with a mean altitude of 269 m, and Nová Ves nad Popelkou, 22 ha (50°31′ N; 15°24′ E) in central Bohemian region with a mean altitude of 185 m a.s.l. These areas were representative of soilscapes, which were homogenous and comparable in terms of terrain characteristics, land management, and climatic conditions (Schmidt et al., 2010). According to the World Reference Base (WRB) for soil resources (IUSS Working Group WRB, 2014), the original soil type in Brumovice was Haplic Chernozem on loess, which due to erosion changed into Regosol (steep part) and colluvial soil (foot slope and the tributary valley), Chernozems and Luvisols on loess for Údrnice, while for Nová Ves nad Popelkou, the soil is characterized mainly as Cambisols on sedimentary rocks. The three study areas are shown in figure 1 below.

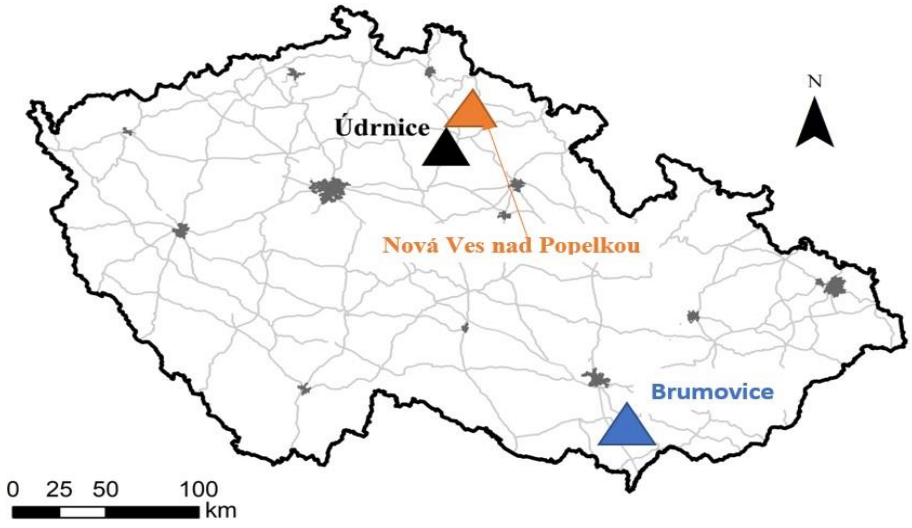


Figure 1. The geographical position of the study areas

3.2. Soil sampling

Although over 350 soil samples were initially proposed to be collected from the three study areas, only 241 of these samples were collected. This includes 111 samples from Údrnice (June of 2019) and 130 samples from Nová Ves nad Popelkou (May of 2019) (figure 2) from the topsoil (0–20 cm) within a regular grid covering the study area. The 241 sampling points were created using a GeoXM (Trimble Inc., Sunnyvale, California, USA) receiver with an accuracy of 1 m before the field visit. For the Brumovice site, 98 samples collected for a previous study were used. Regarding the field size and the chosen sampling algorithm (Pimstein et al., 2011; Ramirez et al., 2014), the selected sample sizes have sufficient coverage of the predictor space, which is a suitable indicator of the population to which the models will be applied. Although the emphasis was on measuring and analysing field and laboratory dry spectra, wet spectra were also considered for comparison between the field and dry measurement approaches using only Nová Ves and Údrnice study sites.

Figure 2. Location of sampling points at Nová Ves nad Popelkou (A), Údrnice (B) and Brumovice (C)

3.3. Spectroscopy data sets

For the study sites located in Nová Ves and Popelkou and Údrnice, the study analysed three spectroscopy data sets, and the best of the three were used for further analysis. These datasets include field, laboratory dry and laboratory wet spectra. For the lab-wet, for instance, its measurement is influenced mainly only by moisture content because most of the other conditions that affect spectral measurement are manipulated in the laboratory. Nevertheless, field data under Vis-NIR measurements are susceptible to external environmental factors, such as temperature, soil moisture and soil structural factors, transient changes in weather conditions during measurement, noise, vegetation cover, illumination sources and variations in illumination due to clouds and wind. The laboratory dry measurement under laboratory-controlled conditions is noted for its reliable prediction of soil organic carbon (SOC) compared to field and wet spectra. This is because a standardized protocol is used; see, e.g., Romero et al. (2018) and Ben Dor et al. (2015). Nevertheless, other issues, such as spectrometer instability, illumination source, detector output, and sample preparation, may persist in the laboratory environment.

3.4. Measurement of field and laboratory spectra and SOC analysis.

The spectral reflectance of soil samples for the different Vis-NIR approaches was measured using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050–2500 nm. The spectrometer was standardized using a Spectralon® panel (Lab-sphere, North Sutton, New Hampshire, USA) with 99% reflectance preceding the first scan and after every six measurements (Shi et al., 2016). For each sample, we collected four spectral measurements, and the average of these measurements was used. The field spectra were measured in the field from four different positions around each of the sampling points. For soil heterogeneity, samples for laboratory analysis were also collected from each of these positions while the field measurement was underway. It was placed into a well-labelled bag and transported to the laboratory for further analysis. Immediately upon reaching the laboratory, spectra readings were taken, and the average of four replicates was used as the wet spectra. After air drying for two weeks, the

samples were gently crushed and sieved (2 mm) before being analysed (ISO 11464:2006) and the laboratory dry spectra were measured. All collected soil samples were also chemically and physically analysed using standard laboratory procedures under a constant laboratory temperature of 20 °C. SOC was measured as total oxidized carbon using a wet oxidation approach (ISO1998.14235). This process utilized the dichromate redox titration approach and was accomplished in two different substeps (Skjemstad et al., 2008). That is, the samples were first oxidized with $K_2Cr_2O_7$ and the solution was then potentiometrically titrated with ferrous ammonium sulphate. However, for the Brumovice site, the spectra and SOC had already been measured, and the data for only the laboratory dry spectra were used.

3.5. Remote sensing imagery

3.5.1. Unmanned aircraft system (UAS) multispectral imaging (airborne)

Multispectral data were acquired using a Trinity F90 fixed-wing drone with a MicaSense Altum dual sensor mounted onboard with two cameras (RGB and Multispectral). The MicaSense Altum dual sensor captures images in six independent spectral bands (multispectral), with the last band being a thermal infrared sensor (Blue 475 nm (B4), Green 560 nm (B5), Red 668 nm (B6), Red edge 717 nm (B7), Near-infrared 840 nm (B8), and Thermal 11 μm (B9)). The RGB sensor also captures images in three bands (red–green-blue) (400–700 nm). This is a high-resolution digital camera where the lens are separated from the multispectral sensor. This implies that the total number of bands captured by Trinity F90 was nine. The location of the on-board Global Navigation Satellite System (GNSS) and Inertial Navigation Unit has been saved in the metadata files using the Exchangeable Image File Format (EXIF). The camera is equipped with a sun sensor that gathers information about the light conditions and saves the radiant flux data produced in EXIF format. The image was acquired on November 25th, 2019, at Nová Ves nad Popelkou and Údrnice under clear sky conditions. The flight plan was prepared using a QBase 3D mobile app (mission planning software). This served as the primary interface between the user and the UAS device. QBase 3D offers real-time information such as altitude, distance, battery life, about the UAS and mission telemetry data that provides the operator with updated information about the flight at all times. The flight height was 190 m, and the spatial resolution was 8.8 cm, covering an area of 31 ha in Nová Ves nad Popelkou. For Údrnice, the altitude was 160 m, covering an area of 52 ha with a spatial resolution of 7.6 cm. We also ensured that we had sufficient batteries for the total flight duration for the two study field . The images were captured automatically, and the calculated position was consistent with 85%

front and 75% side overlap. The images were accurately oriented, a 3D model was extracted, the digital elevation model (DEM) was calculated based on the generated cloud point (during the flying period), and orthorectified images were calculated and then exported as one mosaic in a GeoTIFF file in EPGS 4326-Geographical coordinates on the WGS-84 ellipsoid. Before generating this orthophoto, calibration was performed. The obtained image (before calibration) is already in the reflectance format. However, the actual reflectance values are obtained by dividing each band by 32768 to obtain the values normalized in the interval between 0 and 1. The band center value is 32768, which represents 100% reflectance. For geometrical correction, the ground-based points and the differential global positioning system (DGPS) were used, while for both radiation correction and reflectance transformation, the grayscale correction method was employed. AgiSoft Metashape Professional 1.5.0 (AgiSoft LLC, St. Petersburg, Russia) photogrammetric processing was used. The software's consistent performance in photogrammetric processing has been demonstrated in previous studies (Verhoeven, 2011). To differentiate bare soil areas, the normalized difference vegetation index (NDVI = (infrared–red)/(infrared+red)) was employed to mask a threshold of 0.2. R software (R Development Core Team, Vienna, Austria) was used for all other data processing.

3.5.2. Satellite data acquisition (Sentinel-2 imagery)

The extracted cloud-free Sentinel-2B imagery used for each study field was carried out at the European Space Agency's Copernicus Open Access Hub, which depended on its closest date to field samplings. The Sentinel-2 mission consists of two similar satellites: Sentinel-2A and Sentinel-2B. Each satellite has a Multispectral Instrument (MSI) that generates images of the Earth. The Sentinel-2 images are processed to Level 1C, which implies that they have been ortho-corrected, map-projected images containing top-of-air reflectance data. This image will need further pre-processing by the user, but the level 2A Sentinel-2 imagery can be used instantly because its dataset has been processed by the suppliers using the Sen2Cor processor, which is integrated into the Sentinel Application Platform (SNAP) tool (Muller-Wilm, 2017; Shoko and Mutanga, 2017). These processes include geometric, radiometric, and atmospheric corrections. For the two study fields, level 2A Sentinel-2 imagery was acquired. The Sentinel-2 image consists of 13 spectral bands. These spectral bands range from visible and near infrared (Vis-NIR) to shortwave infrared (SWIR). They include four bands at 10 m resolution ((B2, 490 nm), (B3, 560 nm), (B4, 665 nm), (B8, 842 nm)); six bands at 20 m resolution ((B5, 705 nm), (B6, 740 nm), (B7, 775 nm) and (B8A, 865 nm); 2 SWIR large bands, (B11, 1610 nm) and (B12, 2190 nm)). Finally, three bands were observed at 60 m resolution ((B1, 443 nm), (B9,

940 nm) and (B10, 1380 nm)). Before the download, all 13 bands were resampled to 10 m using SNAP software (by pixel resolution). With the exception of B1, B9, and B10, which were omitted, all the remaining bands were used for analysis for each study field. The Sentinel-2 user handbook describes the whole protocol European Space Agency (ESA), 2016.

3.6. Data pre-processing and model assessment

3.6.1. Spectral pre-treatment

The first step in the development of calibration models is the pre-treatment of the spectral data. Soil spectra are first reduced to eliminate noise on both sides of the spectra. Until modelling, all the datasets were pre-processed. Murray and Williams (1987) stated that removing outliers improves prediction accuracy. Therefore, the outliers from these datasets are removed using a local outlier factor (LOF) algorithm procedure proposed by Breunig et al. (2000). The LOF is a measure that looks at a certain point's neighbours to determine its density and then compares it with the density of other points and uses its local approach to better detect outliers in their neighbours. Additionally, the ensemble sparse partial least squares (enpls) was also explored in some instances.

Before using the data calibration models, the noisy portions between 350 and 399 nm were also eliminated. The datasets were then subjected to the following set of pre-treatment techniques: sg (Savitzky–Golay) from signal R package (Signal developers, 2013), dwt (discrete wavelet transformation) calculated with dwt function from wavelets R package (Aldrich, 2013), d1 (first-order derivative) (Duckworth, 2004), sg_d1, dwt_d1, d2 (second-order derivative), sg_d2, dwt_d2, msc (multiplicative scatter correction) which was calculated using pls R package (Mevik and Wehrens, 2007), sg_msc, dwt_msc, snv (standard normal variate) which was obtained by subtracting each reflectance value from the spectrum's mean reflectance value, and then it was divided by standard deviation, sg_snv, dwt_snv, snv_msc, sg_snv_msc, dwt_snv_msc, log (logarithmic transformation (log(1/R))), sg_log, dwt_log, log_msc, sg_log_msc, dwt_log_msc, log_snv, sg_log_snv, dwt_log_snv, cr (continuum removal) which was obtained from tripack R package (Renka, 1996), sg_cr, dwt_cr, cmr (correction by the maximum reflectance) (Vašát et al., 2017), sg_cmr, dwt_cmr in order to optimize the fitting of target values against spectra. In addition to the other treatment algorithms, orthogonal signal correction (OSC) was explored, using the Unscrambler Program, version X11, CAMO, Norway, for its application. These algorithms also seek to remove or minimize undesirable side

effects (i.e., artefacts) in the spectra while also improving the relevant details about the soil property being estimated.

3.6.2. Modelling development and performance

To ensure that the results were not dependent on the multivariate model, four different multivariate techniques were evaluated separately, namely, Cubist, support vector machine regression (SVMR), partial least squares regression (PLSR) and random forest (RF). The Cubist method was used to calibrate the regression tree models using the train function of the caret package in R. Cubist uses linear regression models at each node instead of the average. To avoid overfitting (Kuhn and Johnson, 2013), the default number of committees (1, 10 and 20) and neighbours (0, 5, and 9) from the train function were utilized. The root mean square error (RMSE) was used to select the best models. Comparably, the SVMR is tuned to different cost parameters with the built-in tuning function using the grid search (precisely 0.001, 0.01, 0.1 and 1) with a linear kernel function, while the epsilon parameter is left to its default value (0.1). The Package e1071 library in R was used. Based on the RMSE, the best cost parameter is selected from bootstrap results based on 10-fold cross-validation. For the PLSR algorithm, a set of new predictor variables identified as latent variables is developed as a linear combination of the initial predictor variables. The model runs and tests itself for each number of components, i.e., from 1 to 10 (the maximum number of model components was set to 10). The optimum number of components is selected based on the lowest RMSE. With the optimal number of components obtained, the model is re-calibrated and validated, and the coefficient of determination (R^2) and the RMSE are computed.

Finally, the RF algorithm is formulated to reduce experimental noise and improve prediction accuracy (Liaw and Wiener, 2002). The dataset under consideration is randomly divided into numerous training sets, and decision trees are developed using bootstrap re-sampling capabilities. The average of the individual tree outputs is then utilized to calculate the final prediction. The Random Forest R package was used, which includes homonyms (Liaw and Wiener, 2002). This is founded on the principles of Leo Breiman and Adele Cutler's Fortran code. A total of 500 trees were grown, with 35 variables randomly selected as candidates at each split. The R programming language (R Development Core Team, 2015, Vienna, Austria) was used for spectra pre-processing and modelling techniques.

The model's output was assessed by five-fold cross-validation for each regression procedure of the calibration (75%) and validation set (25%) of the samples using Cubist, SVMR, PLSR and

RF modelling techniques. The accuracy of the prediction was assessed based on the coefficient of determination for cross-validation (R^2_{CV}), the ratio of performance to interquartile range (RPIQ), the ratio of performance to deviation (RPD), which is the ratio of a parameter's standard deviation to the standard error of that parameter's prediction by a specific model, the root mean square error of prediction (RMSEPcv) (measures the overall model prediction accuracy) of the 5-fold cross-validation and the bias. The bias represents the error of means and is independent. The R^2_{CV} ranges from 0 to 1, where $R^2_{CV} = 1$ is the optimal value. For the RPD, Chang and Laird's (2002) categorization was applied: RPD > 2 indicates good models, RPD between 1.4 and 2 indicates moderate predictive ability, and RPD lower than 1.4 indicates weak models. The five-fold cross-validation was repeated 100 times to ensure model stability and reliability.

3.7. Methodology summary for each paper

3.7.1. Methodology 1: Comparison of field and laboratory wet soil spectra in the Vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements

The pre-treatment use can be found in the attached article as paper 1. The study focuses on each of the separate spectral regions VIS; 400–800, NIR; 800–2500, and the whole Vis-NIR; 400–2500. In all, there were 24 different output models to be tested for each of the two datasets. Almost all the signal transformations were plotted to visualize differences between different pre-processing methods. However, the reflectance and absorbance plots were separated for visual assessment of variation in the spectra and their similarities. For modelling assessment, PLSR, principle component regression (PCR) and SVM were used. The LOF was used for outlier detection and elimination. In all, a total of seven outliers were removed from each dataset. Additionally, for a detailed comparison of the obtained spectra (lab-wet and field), that is, to determine the stable part of the spectra (the part not affected by moisture), the part that differs, and the part with no meaningful information, median filter smoothing (MFS) with a segment size of 7, spectroscopic transformation-absorbance (STA) and gap-segment second derivative (GSD) with a gap size of 6, and a segment size of 25 (Unscramble Software, Version X11, CAMO, Oslo, Norway) was used. The order was MFS–STA–GSD. The study area used is located at Nová Ves nad Popelkou.

3.7.2. Methodology 2: Does the limited use of the orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic carbon need attention?

A total of 259 soil samples from three study fields were used for this study, including Brumovice and Nová Ves. The samples include lab-dry and field spectra. Nine pre-treatments

were used as can be found in the attached paper (paper 2). Outliers within the datasets were eliminated using a local outlier factor (LOF) algorithm proposed by Breunig et al. (2000). In total, nine outliers were removed (field 1:1; field 2:1; field 3:7). All pre-treatment methods were calculated three separate times from raw spectra, sg spectra and dwt spectra, and the best results were selected and reported.

3.7.3. Methodology 3: Exploring the suitability of UAS-based multispectral images for estimating soil organic carbon: Comparison with proximal soil sensing and spaceborne imagery

For this study, nine calculated spectral indices were applied to both the Sentinel-2 and UAS datasets as independent variables, which were anticipated to enhance the prediction capability of these datasets. These spectral indices include the *colour index (CI), normalized difference vegetation index (NDVI), infrared percentage vegetation index (IPVI), normalized difference red edge (NDRE), soil adjusted vegetation index (SAVI), vegetation (V), green–red vegetation index (GNDVI), difference vegetation index (DVI), and brightness index (BI).*

- *Data Pre-processing Approaches*

The field spectra and the other two datasets (UAS and Sentinel-2) were subjected to the following set of pre-processing techniques: dwt, snv, (log(1/R), as well as the combination of dwt with snv (dwt + snv) and with log(1/R) (dwt + log(1/R)).

- *Modelling and Prediction Assessment*

The spectra obtained from Sentinel-2 and UAS sensors, including the determined spectral indices, were each linked to the SOC determined in the laboratory using collected soil samples from the field. Two separate multivariate models were evaluated for all spectral data, namely, RF and SVMR. For the modelling assessment, see the attached full article as paper 3.

Prior to evaluating the predictive models, the normality of the distribution of the SOC contents was examined (skewness <1).

A correlation matrix was computed to visualize the relationships between the three datasets and their parameters (indices) with SOC (to examine which dataset is more correlated or significantly correlated). For the remote sensing data sets (UAS and Sentinel-2), this was done between SOC and their bands and indices. However, for the field spectra, the correlation was made with SOC using only selected wavelengths (based on UAS and Sentinel-2 wavelengths) due to the enormous amount of spectral data available (350–2500 nm). For a visual comparison of the SOC spatial distribution predicted by models based on different data and laboratory

measurements, SOC maps were created using the inverse distance weighting (IDW) interpolation method.

3.7.4. Methodology 4: Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil.

Soil organic matter (SOM) can have a strong influence on redox transformations of toxic elements and of soil minerals (Borch et al., 2010). Among the several elements explored, the prediction of As was high. Therefore, to better understand the role of soil organic matter (SOM) on the change in As and to verify whether these changes affected the prediction accuracy of SOC within our study field, this study was conducted. This is because the prediction accuracy of SOC in this field was poor. A robust model was needed because our study field had no history of pollutants. An ensemble model, which comprises four individual modelling techniques, was used

- *pXRF measurements*

A portable X-ray fluorescence spectrometer Delta Premium (Olympus) was used to analyse As, Mn, S, and Fe. The main component, however, was As. The other elements were included (for correlation with As) because they have been shown to influence As (directly or indirectly, particularly Fe). Fe oxides, for instance, are well known as spectrally active soil properties that can strongly adsorb PTEs or have a high affinity for certain PTEs (Axe et al., 2000; Ben-Dor and Banin, 1995). For the elemental measurement using the pXRFs, see the attached article as paper 4.

- *Soil chemical analyses*

Other auxiliary soil properties that can influence As due to their high adsorption with PTEs (soil organic carbon (SOC), soil pH) were also determined (correlation test). More information on the soil chemical analysis and the pre-treatment used can be found in the attached article as paper 4. Four outliers were removed to improve predictive performance.

- *Correlation and spatial distribution maps*

A correlation matrix was created to observe the relationships between As and the selected auxiliary components (soil pH, SOC, Mn, S, Fe). Spatial variability of soil As contents was mapped using the inverse distance weighting (IDW) interpolation method with the R package gstat (Pebesma, 2004; Gräler et al., 2016). IDW uses a linear combination of values to estimate

the values of the unknown area within the sampling space and allocates weights using its inverse function. Due to its ability to assign weights before prediction, IDW can have a lower error margin than other interpolation methods, which makes it more suitable for creating spatial distribution maps more accurately (Liao et al., 2018; Xie et al., 2011).

- *Multivariate modelling and models*

For this study, four separate multivariate techniques evaluated individually, namely, Cubist, SVM, PLSR, and RF, were used for the study as individual techniques and then combined, forming the ensemble model. The ensemble model details, as well as the model and spatial distribution map validation assessment, can also be found in the attachment article as paper 4.

3.7.5. Methodology 5 (currently under revision): Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations.

For this manuscript, three different agricultural fields were used. Three different proximal sensing data (in-situ. Lab-wet and lab-dry) and remote sensing data from one of the study field were used. Additionally, a merged data of in-situ+lab-wet+lab-dry and lab-dry+lab-dry were also explored. Figure 3 shows the flow diagram for paper five as a manuscript under revision. For the introduction, methods, results, and conclusion, please see the attached manuscript.

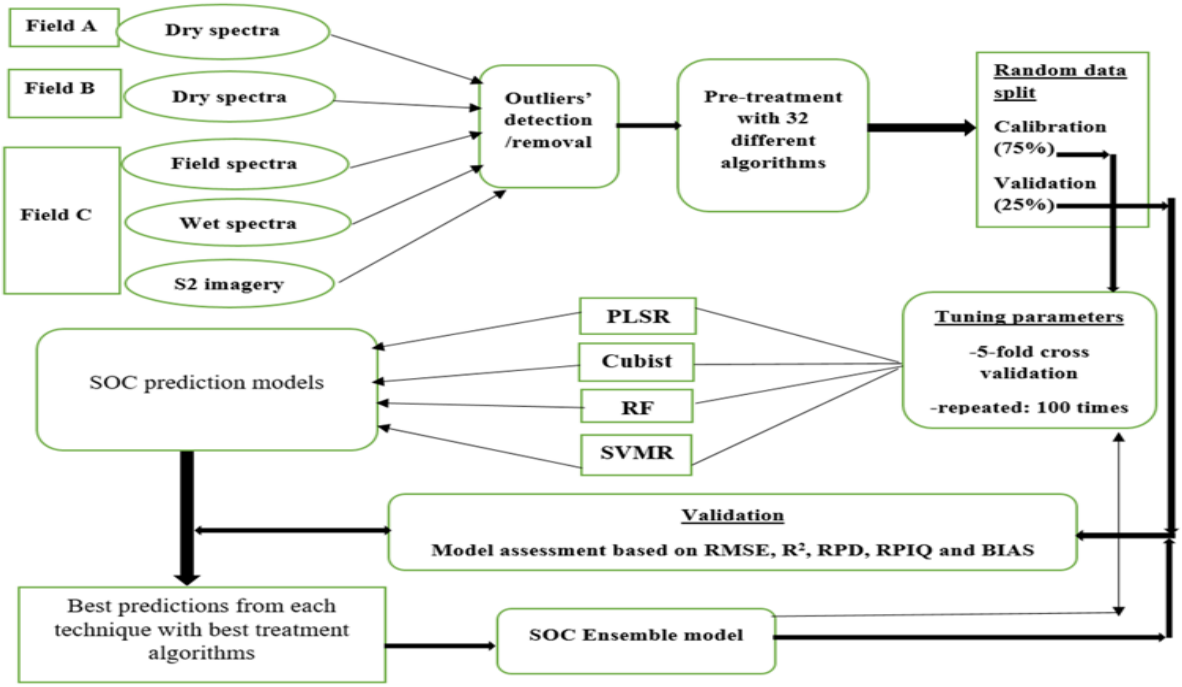


Figure 3: Schematic diagram illustrating the experimental design

3.7.6. Methodology 6 (currently under review): Verifying the impact of fusion high-resolution simulated in situ spectroscopy, Sentinel-2, and unmanned aircraft systems data into a single dataset to estimate soil organic carbon content in low-carbon agricultural fields.

For this manuscript, two fields low in organic carbon content were used for the data fusion approach using three different platforms [in-situ, Unmanned Aircraft Systems (UAS), and Sentinel-2 (S2)], which were combined into a single dataset to predict SOC. Before the data fusion approach, in-situ spectral data was simulated into 12 bands. Figure 4 shows the flow diagram for paper six as a manuscript under review. For the introduction, methods, results, and conclusion, please see the attached manuscript.

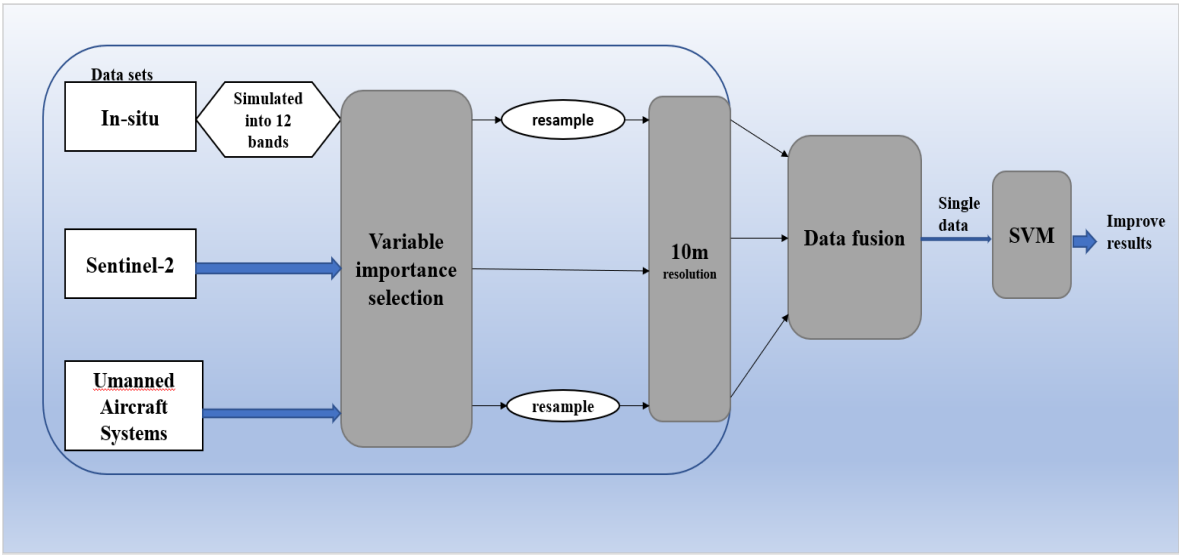


Figure 4: Schematic diagram illustrating the experimental design

4. SUMMARY AND CONCLUDING REMARKS

4.1. Summary and key findings

In general, the study explores the suitability of using data sets obtained from proximal and remote sensing applications in their individual and combined forms to develop a model to improve the estimation of soil properties, especially SOC. The data sets collected include spectroscopy data (in the visible-near infrared (Vis-NIR) range), unmanned aircraft systems (UAS), and Sentinel-2 (S2) data. For the Vis-NIR spectroscopy approach, three data sets were collected under different spectral measurement conditions, namely, lab-dry, lab-wet, and in situ/field spectral conditions. Although the lab-dry method is normally the most accurate of the three measurement conditions, the study decided to use the lab-dry method as reference data and instead focused on the lab-wet and field data in the first study. Moreover, these data can be used instantly, unlike the lab-dry data, which need further analysis. Additionally, the whole idea was to use spectroscopy data that were measured under the same environmental conditions as the remote sensing data (UAS, S2) for fair comparison.

Paper 1

- The focus of this paper was on the comparison of field and naturally acquired lab-wet spectra in the Vis and NIR ranges using several pre-treatment approaches, including orthogonal signal correction (OSC), to identify the most suitable data. The comparison was concluded with the development of validation models for SOC prediction based on PLSR, SVMR, and principal component regression (PCR). The study concludes that in the absence of the lab-dry, both the field and the lab-wet could be used to estimate SOC. However, one significant concern associated with the lab-wet measurement data had to do with finding an appropriate method of transportation to the laboratory for analysis. This is because during transportation, a certain amount of trapped heat can cause variability in moisture content (increasing the moisture content), which can negatively influence the accuracy of predicting SOC accurately. OSC_PLSR was the best model.

Paper 2

- The robustness of the OSC_PLSR model was further explored on larger soil samples in this paper because OSC_PLSR was rarely used in soil science, especially on spectra to estimate SOC. The idea was to use this model on both proximal and remote sensing data in the third study. The second study verified the effectiveness of OSC against nine commonly used pre-treatment methods across three different agricultural fields using

lab-dry and in situ spectral data. In this paper, OSC improves SOC predictive performance compared to the nine commonly used pre-treatment algorithms. One of OSC's strong points was that it helps eliminate multiple artefacts at the same time (e.g., a baseline slope and scatter effect) while ensuring that prediction accuracy will be enhanced during the process compared to some of the other treatment algorithms that remove vital information in the process of improving the dataset under consideration. Now convinced of OSC's robustness, it was initially decided to verify its impact on remote sensing data to predict and map SOC more accurately in the third study as stated already.

Paper 3

- This paper was about a holistic comparison between field spectra, UAS, and Sentinel-2 for the monitoring and mapping of SOC. Additionally, calculated spectral indices were also added to the remote sensing data to enhance the predictive capability of SOC. Several pre-treatment combinations as well as the OSC were also explored (these data sets were measured under disturbing environmental conditions). However, the OSC approach was eliminated because it did not improve the predictive performance of the remote sensing dataset. The results of the study show that, in terms of prediction accuracy, the field spectra was better than UAS and S2 because its dataset was correlated more with SOC than the remote sensing data. Additionally, the field spectra's high spectral resolution gives it an advantage over S2 and UAS because of its ability to capture the target object in more detail. S2 imagery, on the other hand, provided the worst result because all of its bands and calculated spectral indices show no correlation with SOC. Furthermore, UAS's higher spatial resolution than S2 gives it an advantage. This research also shows that UAS imagery can be exploited more efficiently using spectral indices.

Paper 4

- The effectiveness of combining four models in an ensemble approach (PLSR + SVMR + RF + Cubist) against the results obtained by each individual algorithm was assessed in this paper. The goal was to decide whether to use the individual algorithms or the ensemble approach in the next study. The paper used both portable X-ray fluorescence and visible near-infrared spectroscopy datasets. The study showed that the ensemble model was better than each of the individual models. Additionally, the ensemble model

was also tested in another study (which is under revision at Soil and Tillage Research-journal), titled "Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations".

Manuscript 5 (under revision)

- This study confirmed that when different prediction techniques are combined to form an ensemble model using different calibration techniques, prediction and signal pre-treatment algorithms, the prediction accuracy was superior to any of the modelling techniques used individually. The ensemble model built for this study accurately captured the trend of all study fields as well as the various datasets gathered with a minor error and improved the prediction accuracy of SOC. Furthermore, it provides a more robust and reliable approach than each of the individual model estimates do alone. The findings demonstrated that the ensemble model could be an effective tool for reducing overall error in SOC modelling. It was also successful on almost all the data obtained under different spectral measurement conditions with an order of dry > wet > field. The only exception was the accuracy of SOC prediction using Sentinel-2 data, which was poor for the study field employed, likely due to numerous factors (e.g., cloud cover, vegetation) and constraints that affect the acquired Sentinel-2 imagery.

Manuscript 6 (currently under review)

- This paper verifies whether the fusion of high-resolution simulated in situ spectroscopy, Sentinel-2, and unmanned aircraft systems data into a single dataset can be used to improve the prediction of SOC at low organic carbon sites. The study began by ensuring that all of the data were transformed to the same spatial resolution, with an S2 resolution of 10 m serving as a reference. Two study fields were used. Different combinations of fusion of the sample data were investigated [UAS+S2, in situ + UAS, in situ+S2, in situ +UAS+S2, in situ+UAS+S2] and compared with the individual platform. A medium-level fusion approach was used. For the results, the in situ spatial distribution map closely resembles the measured lab map than any other data combination used. The fusion of the three-platform data provided the best estimate of SOC.

Additional Papers

- Paper Biney, 2022 (Computers and Electronics in Agriculture, 194, 106796) aims to compare the differences in SOC prediction when using field spectra (FS) and Sentinel-

2 (S2) data collected separately through simple random (SR) and grid design (GD) on the same agricultural field. Additionally, the impact of spectral indices on S2 data in a merged data approach under the two-sampling strategies is also tested. The results show how crucial the sampling design is for obtaining good prediction accuracy.

- For the paper Demattê et al. 2022 (Remote Sensing, 14, 740), my task was to prepare spectral data measured on soil samples from the Czech Republic. This was an international initiative comprising 26 different authors. It shows how global is this research and exploitation of soil spectroscopy for soil property, particularly SOC, prediction.

4.2. Areas that need further research

- The reason why the OSC_PLSR and the ensemble model were better on proximal sensing data but failed to improve remote sensing data needs further study.
- Obtaining S2 imagery captured on bare soil is another area worth investigating; normally, the field measurement is taken prior to the search for suitable satellite imagery, but due to cloud cover effects, the obtained satellite imagery is sometimes far from the date when the field sampling was taken. Perhaps downloading more images to investigate the optimum date and adjusting the field measurements could be a solution to the issue of obtaining Sentinel-2 imagery captured on bare soil. Other ways to explore this issue need further study. Because getting these agricultural fields at one's "optimum time" is highly dependent on the landowners' and land manager's decisions

4.3. Concluding remarks

The study shows the following conclusions;

Both lab-wet and in-situ spectral data can be used in the absence of lab-dry for the prediction of SOC and especially for quicker analysis.

The use of OSC_PLSR and the ensemble models to improve SOC estimates with spectroscopy data sets (in-situ and lab-dry) is highly recommended.

The predictive performance of UAS data to predict SOC was slightly better than S2 data because UAS provides a better spatial resolution than S2.

Although fusion of proximal and remote sensing data is possible and can serve as an option for SOC estimate improvement, the use of the same fusion algorithm for two or more different studies needs further study.

Using data fusion approach to create a SOC spatial distribution map needs to be explored further.

The fusion of simulated in-situ, S2, and UAS data sets into a single dataset can be used to improve the prediction of SOC in fields poor in organic carbon content.

The study also showed that the simulated in-situ data was not affected, because in terms of prediction accuracy, it was better than both S2 and UAS data.

Overall, it can be concluded that, though there are still many things to explore and study, soil spectroscopy, used in proximal soil sensing, remote sensing or their combination, provides very useful and promising tool for SOC content prediction and monitoring

REFERENCES

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., and Upadhyaya, S. K., 2004a. On-the-go soil sensors for precision agriculture. Computers and Electronics in Agriculture, 44, 71–91.

Adamchuk, V. I., Lund, E. D., Sethuramasamyraja, B., Morgan, M. T., Dobermann, A., and Marx, D. B., 2005. Direct measurement of soil chemical properties on-the-go using ion-selective electrodes. Computer and Electronics in Agriculture, 48, 272– 294.

Adamchuk, V. I., Viscarra Rossel, R. A., Marx, D. B., and Samal, A. K., 2011a. Using targeted sampling to process multivariate soil sensing data, Geoderma, 163, 63–73.

Aldrich, E. 2013. A Package of Functions for Computing Wavelet Filters, Wavelet Transforms and Multiresolution Analyses. 2013. Available online: http://cran.rproject.org/web/packages/wavelets/wavelets.pdf (accessed on 21 September 2012)

Amundson, R., Berhe, A.A., Jan, W., Hopmans, J.W., Olson, C., A., and Sztein, E., 2005. Soil and human security in the 21st century. Science, 348 (6235), 1261071.

Barnes, E.M., Sudduth, K.A., Hummel, J.W., Lesch, S.M., Corwin, D.L., Yang, C., Daughtry, C.S.T., and Bausch, W.C. 2003. Remote and ground-based sensor techniques to map soil properties. Photogrammetric Engineering and Remote Sensing, 69 (6), 619–630.

Ben Dor, E., Ong, C., and Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245-246, 112–124

Ben-Dor, E. and Levin, N., 2000. Determination of surface reflectance from raw hyperspectral data without simultaneous ground data measurements: a case study of the GER 63-channel sensor data acquired over Naan, Israel. International Journal of Remote Sensing, 21(10), 2053-2074.

Ben-Dor, E., Inbar, Y., and Chen, Y., 1997. The reflectance spectra of the visible nearinfrared and shortwave infrared region (400-2500 nm) during a controlled decomposition process. Remote Sensing of Environment, 61, 1-15.

Biney, J. K. M., Borůvka, L., Agyeman, P. C., Němeček, K., and Klement, A., 2020. Comparison of Field and Laboratory Wet Soil Spectra in the Vis-NIR Range for Soil Organic Carbon Prediction in the Absence of Laboratory Dry Measurements. Remote Sensing, 12(18), 3082.

Borch, T., Kretzschmar, R., Kappler, A., Cappellen, P.V., Ginder-Vogel, M., Voegelin, A. and Campbell, K., 2010. Biogeochemical redox processes and their impact on contaminant dynamics. Environmental science & technology, 44(1), pp.15-23.

Bouma, J., and McBratney, A.B., 2013. Framing soils as an actor when dealing with wicked environmental problems. Geoderma, 200–201, 130-139.

Breunig, M. M., Kriegel, H. P., N.G., R. T., and Sander, J., 2000. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (93-104).

Brevik, E.C., Cerdà, A., Matrix-Solera, J., Pereg, L., Quinton, J.N., Six, J. and Van Oost, K., 2015. The interdisciplinary nature of the soil. Soil, 1, 117-129.

Bricklemyer, R. S., and Brown, D.J., 2010. On-the-go Vis NIR: potential and limitations for mapping soil clay and organic carbon. Computers and Electronics in Agriculture, 70 (1), 209–216.

Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., and Casa, R., 2016. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. Remote Sensing of Environment. 179, 54-65.

Chang, C. W., and Laird, D. A., 2002. Near-infrared reflectance spectroscopic analysis of soil C and N. Soil Science, 167(2), 110-116.

Chang, C. W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Science Society America Journal. 65, 480–490.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D. and Shi, Z., 2019. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. Science of the Total Environment, 655, 273-283.

Choe, E., Van Der Meer, F., Van Ruitenbeek, F., Van Der Werff, H., De Smith, B., and Kim, K.W., 2008. Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the Rodalquilar mining area, SE Spain. Remote Sensing of Environment. 112(7), 3222-3233

Christy, C. D., 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. Computers and Electronics in Agriculture, 61 (1), 10–19.

Clark, R. N., King, T. V. V., Klejwa, M., Swayze, G. A., and Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. Journal of Geophysical Research. 95, 12653–12680.

Corsi, S., Friedrich, T., Kassam, A., Pisante, M., and Sà, J. D. M., 2012. Soil organic carbon accumulation and greenhouse gas emission reductions from conservation agriculture: a literature review. Food and Agriculture Organization of the United Nations (FAO). Rome, 89 pages.

Croft, H., Kuhn, N.J., and Anderson, K., 2012. On the use of remote sensing techniques for monitoring spatiotemporal soil organic carbon dynamics in agricultural systems. Catena, 94, 64–7.

De Gruijter, J. J., McBratney, A. B., and Taylor, J., 2010. Sampling for highresolution soil mapping, R. A. Viscarra Rossel, A. B. McBratney and B. Minasny (eds.) Proximal Soil Sensing, New York: Springer-Verlag, 3–14.

Duckworth, J., 2004. Mathematical data preprocessing. Near Infrared Spectroscopy. Agric, 44, 113–132

Durrant-Whyte, H., 2001. Multi-Sensor Data Fusion. Australian Centre for Field Robotics, University of Sydney. Version 1.2.

EU Soil Thematic Strategy. 2004. TWG work package 2: status and distribution of soil in Europe. Final report version 1.0, 31 March 2004.

European Space Agency. 2016. Sen2Cor 2.2.1 software Release Note. Available online at: http://step.esa.int/thirdparties/sen2cor/2.2.1/[L2A-SRN]ESA-EOPG-CSCGS-TN-0014[2.2.1].pdf

European Space Agency, 2015. Sentinel-2 user handbook. ESA Standard Document.

Gholizadeh, A., Amin, M.S.M., Borůvka, L., and Saberioon, M.M., 2014. Models for estimating the physical properties of paddy soil using visible and near-infrared reflectance spectroscopy. Journal of Applied Spectroscopy. 81(3), 534-540.

Gholizadeh, A., Borůvka, L., Saberioon, M., and Vašát, R., 2013. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. Applied Spectroscopy, 67(12), 1349-1362.

Gräler, B., Pebesma, E. J., and Heuvelink, G. B., 2016. Spatio-temporal interpolation using gstat. R J., 8(1), 204.

Hillel, D., and Hillel, D., 1998. Environmental Soil Physics. Academic Press Ltd., London.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R. and Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: a prospective review. Geoderma, 241, 180-209.

IPCC. 2014: Climate Change 2014: Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland.

ISO. Soil Quality–Determination of Organic Carbon by Sulfochromic Oxidation; International Standard 1998.14235; Croatian Standards Institute: Zagreb, Croatia, 1998.

Jenny, H., 1980. The Soil Resource. Springer, Berlin German Federal Republic.

Jensen, J.R., 2007. Remote Sensing of the Environment: An Earth resource perspective. Prentice-Hall, Upper Saddle River: New Jersey, USA, 544.

Ji, W., Adamchuk, V.I., Chen, S., Su, A.S.M., Ismail, A., Gan, Q., Shi, Z. and Biswas, A., 2019. Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. Geoderma, 341, 111-128.

Ji, W., Viscarra Rossel, R.A., and Shi, Z., 2015. Improved estimates of organic carbon using proximally sensed vis–NIR spectra corrected by piecewise direct standardization. European Journal of Soil Science. 66 (4), 670-678.

Kuang, B., Mahmood, S. H., Quraishi, Z. M., Hoogmoed, B. W., Mouazen, A. M., and Eldert J., Van Henten., 2012. Sensing Soil Properties in the Laboratory, In Situ, and On-Line: A Review. In Donald Sparks, editor: Advances in Agronomy, 114, 155223.

Kuhn, M., and Johnson, K.,2013. Applied predictive modeling New York: Springer, (Vol. 26, p. 13). Signal Developers, (2013). Signal: signal processing URL: http://r-forge.r-project.org/projects/signal (2013)

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

Lal, R., 2009. Challenges and opportunities in soil research. European. Journal of Soil Science, 60 (2), 158–169.

Lal, R., Kimble, J., Follett, R., and Stewart, B.A., 1997. Management of carbon sequestration in soil (Advances in Soil Science). CRC, Boca Raton, Fl, USA.

Liao, Y., Li, D., and Zhang, N.,2018. Comparison of interpolation models for estimating heavy metals in soils under various spatial characteristics and sampling methods. Transactions in GIS, 22(2), 409-434.

Liaw, A.; Wiener, M. Classification and regression by random forest. R News 2002, 2, 18–22

Liu, W. D., Baret, F., Gu, X. F., Tong, Q. X., Zheng, L. F., and Zhang, B., 2002. Relating soil surface moisture to reflectance. Remote Sensing Environment. 81, 238–246.

Lobell, Db and Asner, G.P. 2002. Moisture effects on soil reflectance. Soil Science Society of America Journal, 66 (3), 722–727.

Loveland, P. and Webb, J., 2003. Is there a critical level in the agricultural soils of temperate regions? A review. Soil & Tillage Research, 70, 1-18.

Mahmood, H. S., Hoogmoed, W. B., and Van Henten, E. J., 2009. Combined sensor system for mapping soil properties. In E. J. Van Henten, D. Goense, and J. F. M. Huijsmans (Eds.), Precision agriculture 2009: Proceedings of the 7th European conference on precision agriculture, The Netherlands (423–430). The Netherlands: Wageningen Academic Publishers.

McCarty, G.W., Reeves, J.B., Follet, R.F and 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.

Metz, H., Davidson, O. R., Bosch, P, R., Dave, R., and Meyer. L.A (Eds)., 2007. The contribution of working group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Miller, C.E., 2001. Chemical principles of near-infrared technology. P. Williams, K. Norris (Eds.), Near-Infrared Technology in the Agricultural and Food Industries, The American Association of Cereal Chemists Inc., St. Paul, MN, pp. 19-37

Milton, E.J., Schaepman, M.E., Anderson, K., Kneubühler, M. and Fox, N. 2009. Progress in field spectroscopy. Remote Sensing of Environment, 113, 92–109.

Minasny, B. McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling the 639 presences of ancillary information. Computer. Geoscience, 32, 1378-1388.

Morgan, C.L.S., Waiser, T.H., Brown, D. J., and Hallmark, C. T., 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. Geoderma, 151 (3–4), 249–256.

Mouazen, A. M., Karoui, R., De Baerdemaeker, J., and Ramon, H., 2006a. Characterization of soil water content using measured visible and near-infrared spectra. Soil Science Society America Journal 70, 1295–1302.

Mouazen, A. M., Maleki, M. R., De Baerdemaeker, J., and Ramon, H., 2007. On-line measurement of some selected soil properties using a VIS-NIR sensor. Soil &Tillage Research. 93, 13–27.

Mulder, C., Boit, A., Bonkowski. M., D. Ruiter, P.C., Mancinelli, G., Van Der Heijden, Mga, Van Wijnen, H.J., Vonk, Ja, and Rutgers, M., 2011.A below-ground perspective on Dutch agroecosystem. How soil organisms interact to support ecosystem services. Advance Ecology Research 44, 277-358.

Muller-Wilm, U., 2016. Sentinel-2 MSI-Level-2A prototype processor installation and user manual. Telespazio VEGA Deutschland GmbH.

Murray, I., 1988. Aspects of interpretation of NIR spectra, in: Creaser, C.S., Davies, A.M.C. (Eds.), Analytical Application of Spectroscopy. Royal Society of Chemistry: London, UK, 9-21.

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

O'rourke, S. M., Stockmann, U., Holden, N. M., McBratney, A. B., and Minasny, B., 2016. An assessment of model averaging to improve predictive power of portable Vis-NIR and XRF for the determination of agronomic soil properties. Geoderma, 279, 31-44.

Park, H., Kreunen Ss., Curtiss A.J., Dellapenna.D., and Pogson, B.J., 2002. Identification of the carotenoid isomerase provides insight into the carotenoid biosynthesis, prolamellar body formation, photomorphogenesis. Plant and Cell, 14, 321-332.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30(7), 683-691.

Pimstein, A., Ben-Dor, E., and Notesko, G., 2011., Performance of three identical spectrometers in retrieving soil reflectance under laboratory conditions. Soil Science Society America Journal. 75, 110–174.

Post-W.M., Izaurralde R. C, Mann L.K., and Bliss N., 2001. Monitoring and verifying changes of organic carbon in soil. Climatic Change, 51, 73-99.

Ramirez-Lopez, L., Schmidt, K., Behrens, T., Van Wesemael, B., Dematte, J., and Scholten, T., 2014. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma, 226–227, 140–150.

Reeves, J.B., 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma, 158 (1–2), 3–14.

Ren, H.Y., Zhuang, D. F., Singh, A. N., Pan, J.J., Qid, D.S., and Shi, R.H., 2009. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: A case study. Pedosphere. 19, 719-726

Renka, R.J. Algorithm 751: TRIPACK: A constrained two-dimensional Delaunay triangulation package. ACM Trans. Math. Software. 1996, 22, 1–8.

Romero, D.J., Ben-Dor, E., Demattê, J.A.M., Souza, A.B.E., Vicente, L.E., Tavares, T.R., Martello, M., Strabeli, T.F., Da Silva Barros, P.P., Fiorio, P.R., Gallo, B.C., Sato, M.V., and Eitelwein, M.T., 2018. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma 312, 95–103

Roy, D.P., Li, J., Zhang, H.K., and Yan, L., 2016. Best practices for the reprojection and resampling of Sentinel-2 Multi-Spectral Instrument Level 1C data. Remote Sensing Letters, 7, (11), 1023-1032.

Ryan M. G., Law, B. E., 2005. Interpreting, measuring, and modelling soil respiration. Biogeochemistry, 73 (1), 3-27.

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

Saffih-Hdadi, K., and Mary, B., 2008. Modelling consequences of straw residues export on the soil. Soil Biology and Biochemistry, 40 (3), 594-607.

Schlapfer, D., Richter, R., and Feingersh, T., 2014. Operational BRDF effects correction for wide-699 field-of-view optical scanners (BREFCOR). IEEE Transactions on Geoscience for Remote Sensing. 53(4),1855-7001864.

Schmidt, K., Behrens, T., Friedrich, K., and Scholten, T., 2010. A method to generate soilscapes from soil maps. Journal of Plant Nutrition and Soil Science, 173(2), 163172.

Schulze, D. G., 2002. An introduction to soil mineralogy. In ‘‘Soil Mineralogy with Environmental Applications’’ (J. B. Dixon and D. G. Schulze, Eds.), 1–35. Soil Science Society of America Inc., Madison, WI.

Seige, T., Böhner, J., and Schmidhalter, U., 2006. High-resolution topsoil mapping using hyperspectral image and field data in multivariate regression modelling procedures. Geoderma, 136 (1), 235-244.

ESA, European Spatial Agency, 2016. Sentinel-2 user handbook. ESA Standard Document. 64.

Shepherd, K. D., and Walsh, M. G., 2007. Infrared spectroscopy—Enabling an evidence-based diagnostic surveillance approach to agricultural and environmental Management in developing countries. Journal of Near Infrared Spectroscopy. 15, 1–20.

Shepherd, K.D., and Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal, 66 (3), 988– 998.

Sherman, D. M., and Waite, T. D., 1985. Electronic spectra of Fe3+ oxides and oxyhydroxides in the near infrared to ultraviolet. America Mineralogy. 70, 1262–1269.

Shi, T., Wang, J., Chen, Y. and Wu, G., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. International Journal of Applied Earth Observation and Geoinformation, 52, 95-103.

Shoko, C., and Mutanga, O., 2017. Examining the strength of the newly launched Sentinel 2 MSI sensor in detecting and discriminating subtle differences between C3 and C4 grass species. ISPRS Journal of Photogrammetry and Remote Sensing, 129, 32-40.

Skjemstad, J. O., Clarke, P., Golchin, A., and Oades, J. M., 1997. Characterization of soil organic matter by solid-state 13C NMR spectroscopy. In ‘‘Driven by nature: Plant litter quality and decomposition’’ (G. Gadish and K. E. Giller, Eds.), 253– 271. CAB International, Wellington, UK.

Skjemstad, J.O., and Baldock, J.A., 2008. Total and organic carbon. In Soil Sampling and Methods of Analysis, 2nd ed.; Carter, M.R., Gregorich, E.G., Eds.; CRC Press: Boca Raton, FL, USA, 225–238.

Smith, P. 2008. Land use change and soil organic carbon dynamics. Nutrient Cycling in Agroecosystems, 81 (2), 169–178.

Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L. and Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Applied Clay Science, 64, pp.75-83.

Soriano-Disla, J.M., Janik, J., Rossel, R.A., Macdonald, Lm and McLaughlin, M.J., 2014. The Performance of Visible, Near and Mid Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. Applied Spectroscopy Reviews, 49, 139-186.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M. and Wetterlind, J., 2010. Visible and near-infrared spectroscopy in soil science. Advances in Agronomy, 107, 163– 215.

Stevens, A., Van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., and Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma, 144(1-2), 395-404.

United Nations/Framework Convention On Climate Change, 2015. Adoption of the Paris Agreement, 21st Conference of the Parties, Paris: United Nations.

Van Der Werff, H., and Van Der Meer, F., 2016. Sentinel-2A MSI and Landsat 8 OLI provide data continuity for geological remote sensing. Remote Sensing, 8(11), 883.

Van-Camp, L., Bujarrabal, B., Gentile, A. R., Jones, R. J. A., Montanarella, L., Olazabal, C. and Selvaradjou, S. K., 2004. Volume III– and biodiversity. Reports of the technical working groups established under the thematic strategy for soil protection. Office for Official Publications of the European Communities, Luxembourg. 872.

Vašát, R., Kodešová, R., Klement, A., and Borůvka, L., 2017. Simple but efficient signal preprocessing in soil organic carbon spectroscopic estimation. Geoderma, 298, 46-53.

Vaudour, E., Gilliot, J. M., Bel, L., Bréchet, L., Hamiache, J. O. N. A. S., Hadjar, D., and Lemonnier, Y., 2014. Uncertainty of soil reflectance retrieval from SPOT and RapidEye multispectral satellite images using a per-pixel bootstrapped empirical line atmospheric correction over an agricultural region. International Journal of Applied Earth Observation and Geoinformation, 26, 217-234.

Verhoeven, G., 2011. Taking computer vision aloft–archaeological three-dimensional reconstructions from aerial photographs with photoscan. Archaeological Prospection, 18(1), 67-73.

Viscarra Rossel R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., and Lobsey, C., 2011. Proximal soil sensing: updating the pedologist's toolkit. Advances in Agronomy, 113, 237–282.

Viscarra Rossel, R. A., and Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158 (1–2), 46-54.

Viscarra Rossel, R. A., and McBratney, A. B., 1998. Soil chemical analytical accuracy and costs: implications for precision agriculture. Australian Journal of Experimental Agriculture, 38, 765–775.

Viscarra Rossel, R. A., and McBratney, A. B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field VIS-NIR spectroscopy: an Australian case study. Geoderma, 146 (3–4), 403–411.

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., and 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. and Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158, 46–54.

Viscarra Vasques, G. M., Grunwald, S. and Sickman, J. O., 2008. Comparison of multivariate methods for inferential modelling of soil carbon using visible/nearinfrared spectra. Geoderma, 146 (1– 2), 14-25.

Walcott, J., Bruce, S. and Sims, J., 2009. Soil carbon for carbon sequestration and trading: a review of issues for agriculture and forestry. Bureau of Rural Sciences, Department of Agriculture, Fisheries and Forestry, Canberra.

Wehrens, R., and Mevik, B. H., 2007. The pls package: principal component and partial least squares regression in R J. Stat. Softw., 18 (2) (2007), pp. 1-24.

Williams, K. N. (Eds.), Near-Infrared Technology in the Agricultural and Food Industries, American Association of Cereal Chemists, St. Paul, MN, pp. 17-34

Williams, P. 2003. Near-infrared Technology Getting the Best out of Light. PDK Projects, Nanaimo, Canada.

Williams, P. C., and Norris, K., 2001. Variable affecting the near-infrared spectroscopic analysis. In: Near-Infrared Technology in Agricultural and Food Industries (Williams P C; Norris K, eds), 2nd Ed, p 295. American Association of Cereal Chemists, Inc. St. Paul Minnesota, USA.

Xie, Y., Chen, T. B., Lei, M., Yang, J., Guo, Q. J., Song, B., and Zhou, X. Y., 2011. Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. Chemosphere, 82(3), 468-476.

Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., and Shi, Z.,2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. European Journal of Soil Science, 70(1), 162-173.

Yitagesu, F. A., Van Der Meer, F. D., and Van Der Werff, H., 2009. Quantifying engineering parameters of expansive soil from their reflectance spectra: Engineering Geology, 105, 151–160.

Zimmermann, M., Leifeld, J., Schmidt, M.W.I., Smith, P. and Fuhrer, J., 2007. Measured soil organic matter fractions can be related to pools in the RothC model. European Journal of Soil Science, 58(3), 658-667.

Žížala, D., Zádorová, T., and Kapička, J., 2017. Assessment of soil degradation by erosion based on analysis of soil properties using aerial hyperspectral images and ancillary data, Czech Republic. Remote Sensing. 9, 28.

remote sensing

MDPI

Article

Comparison of Field and Laboratory Wet Soil Spectra in the Vis-NIR Range for Soil Organic Carbon Prediction in the Absence of Laboratory Dry Measurements

James Kobina Mensah Biney *, Luboš Borůvka, Prince Chapman Agyeman, Karel Němeček and Aleš Klement

Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic; boruvka@af.czu.cz (L.B.); agyeman@af.czu.cz (P.C.A.); nemecekk@af.czu.cz (K.N.); klement@af.czu.cz (A.K.)

* Correspondence: biney@af.czu.cz

Received: 10 August 2020; Accepted: 18 September 2020; Published: 20 September 2020

Abstract: Spectroscopy has demonstrated the ability to predict specific soil properties. Consequently, it is a promising avenue to complement the traditional methods that are costly and time-consuming. In the visible-near infrared (Vis-NIR) region, spectroscopy has been widely used for the rapid determination of organic components, especially soil organic carbon (SOC) using laboratory dry (lab-dry) measurement. However, steps such as collecting, grinding, sieving and soil drying at ambient (room) temperature and humidity for several days, which is a vital process, make the lab-dry preparation a bit slow compared to the field or laboratory wet (lab-wet) measurement. The use of soil spectra measured directly in the field or on a wet sample remains challenging due to uncontrolled soil moisture variations and other environmental conditions. However, for direct and timely prediction and mapping of soil properties, especially SOC, the field or lab-wet measurement could be an option in place of the lab-dry measurement. This study focuses on comparison of field and naturally acquired laboratory measurement of wet samples in Visible (VIS), Near-Infrared (NIR) and Vis-NIR range using several pretreatment approaches including orthogonal signal correction (OSC). The comparison was concluded with the development of validation models for SOC prediction based on partial least squares regression (PLSR) and support vector machine (SVMR). Nonetheless, for the OSC implementation, we use principal component regression (PCR) together with PLSR as SVMR is not appropriate under OSC. For SOC prediction, the field measurement was better in the VIS range with R^2_{CV} = 0.47 and RMSEPcv = 0.24, while in Vis-NIR range the lab-wet measurement was better with R^2_{CV} = 0.44 and RMSEPcv = 0.25, both using the SVMR algorithm. However, the prediction accuracy improves with the introduction of OSC on both samples. The highest prediction was obtained with the lab-wet dataset (using PLSR) in the NIR and Vis-NIR range with R^2_{CV} = 0.54/0.55 and RMSEPcv = 0.24. This result indicates that the field and, in particular, lab-wet measurements, which are not commonly used, can also be useful for SOC prediction, just as the lab-dry method, with some adjustments.

Keywords: vis-NIR spectroscopy; soil organic carbon; proximal sensing; machine-learning; pretreatment methods; spectral datasets (field-wet)

1. Introduction

Soils are significant natural resources for the survival of humanity. Substantially more carbon is stockpiled in the world's soils than is present in global vegetation and atmosphere combined [1]. Studies have shown over the years that the conservation of soil organic carbon (SOC) concentrations

is strongly linked to biological activity and agricultural productivity [2]. Maintaining SOC contents above critical limits for specific ecological and climatic zones will help to protect soil resources and maintain crop yields, thus contributing to global food security [3]. Prediction of SOC in the soil is, therefore, essential because there is always an improvement in soil health as well as the alleviation of climate change whenever SOC content increases [4].

Soil spectroscopy under proximal soil sensing, developed some decades ago, has been used as a useful tool by more researchers in recent years to complement traditional soil analysis [5,6]. Spectroscopy, being the analysis of the interaction of visible-infrared wavelengths with soil properties, also provides information on soil particle size and thus information on the soil matrix. Another attractive feature of spectroscopy is that spectra can be recorded, at points or by imaging, from different platforms; by proximal sensing in the field, in the laboratory using sampled material, or from remote sensing platforms with multi- and hyperspectral capabilities. Compared with analytical laboratory approaches, its measurement is more cost-effective because it is quicker and can use a single spectral measurement to infer multiple soil properties [5–7]. In laboratory and field environments, the spectroscopy technique is increasingly used to predict numerous soil constituents based on their diagnostic spectral features and approaches to statistical regression [8].

Prediction of soil organic carbon using visible near-infrared (Vis-NIR) spectroscopy under laboratory-controlled conditions has produced the most accurate results (high analytical precision) in comparison to field and remote sensing platforms [9–11]. Under laboratory conditions, external factors such as moisture and environmental conditions that could influence the spectrum are manipulated and are subject to greater control, e.g., spectral noise and atmospheric attenuation. However, steps such as collection, grinding, sieving, and drying of soil, which are vital during this process, make it slower in comparison to field measurements [12,13].

The laboratory domain has become well acknowledged ahead of using the field and other applications. However, direct and timely prediction and mapping of soil properties, especially SOC, can preferably be accomplished by field spectroscopy measurement [14]. Some researchers have even shown the field measurement producing better results than under the laboratory-controlled approach. For example, working with an exceedingly disturbed savannah-like environment, Nocita et al. [15] detected good field predictions of SOC comparative to the same soil samples verified under laboratory conditions. Also, Stevens et al. [16] demonstrated that field measurements could be as accurate as laboratory measurements by comparing the efficiency of the laboratory, field, and airborne spectroscopy to predict SOC using PLSR. They concluded that the RMSE of the field spectral prediction was similar to that of the Walkley and Black method, and that airborne spectroscopy was inaccurate.

Similarly, understanding the impact of different soil components under field conditions is not extensively known. Nevertheless, for the assessment of the applicability of laboratory studies to the natural system, field conditions remain a fundamental requirement [17]. However, natural variability issues must be taken into consideration when sampling materials in the field, as the field environment can display subtle and complex variability. Soil samples obtained from the field can also undergo chemical reactions if special safeguard measures are not applied. For example, an increase in soil pH as a result of CO_2 degassing is because the atmospheric condition and that of the soil atmosphere differ. These changes in atmospheric condition also affect redox-sensitive elements such as Fe, Cr, Hg, Cu, etc., upon their exposure, especially within a depleted oxygen environment [17].

Another form of dataset using the spectroscopy approach (aside field and laboratory dry data) is laboratory wet data. Ideally, for this form of dataset, the sample is collected from the field and then the spectral measurement is taken immediately upon reaching the laboratory. For the wet sample to remain in its natural state, proper and orderly transportation from the field to the laboratory must be ensured. The laboratory wet spectral measurements have also received some considerable attention over the years by researchers, but rather in an artificial form, through a process called rewetting [18–20]. However, in its natural acquired state, it has received relatively less attention. According to Barnard et al. [21] and Bailey et al. [22], when rewetting dry soil, a large pulse of CO_2 is emanated from the soil instantly,

known as the 'Birch Effect', named after "H.F. Birch" who experienced a high mineralization effect in East African soils after the rewetting process [23]. According to Bailey et al. [22], these outcomes can cause a considerable decline in soil carbon stabilization and may even affect its predictability outcome. According to Sparks [17], introducing water to the soil can also create a solution–solid or solution–gas reaction that may result in an unstable solution–soil equilibrium if the reaction time is either too short or too long. This shows that the addition of water to dry soil under laboratory conditions may put the soil under undesirable conditions even before prediction. Artificially generated wet samples (mostly used for experiments) may differ somewhat from the natural collected wet samples due to the rewetting approach.

The first stage of Vis-NIR spectra-based multivariate calibration is often data preprocessing. The intention for this is that Vis-NIR spectra often constitute a subset of the features including noise, scattering of light and variances in spectroscopic path length, which are unrelated to the responses. The variation in the predictor that is unrelated to response can disrupt the multivariate modeling, leading to an inaccurate prediction. Some of these pretreatment methods end up removing relevant information from the predictor, especially multiple signal correction (MSC) and standard normal variate (SNV) [24]. This could either cause an enhancement or have a weakening effect [25]. Orthogonal signal correction (OSC) was firstly introduced by Wold et al. [24] for NIR spectra correction and later on as an improvement to its performance; numerous algorithms have since been published. The key concept of OSC technique is based on eliminating the variation that is not related to the parameter for estimation. This method is achieved through the removal of nonrelevant information of the response in the matrix. Therefore, only information orthogonal to the response is omitted. This is made by ensuring that the removed portion is mathematically orthogonal to the response, or as near as possible to being orthogonal. In some cases, the OSC method can also remove nonlinear relationships between the response and the predicted variables [24]. Though the method often converges fast, it still needs 5–10 repetitions [26].

The aim of this work is to compare field and naturally acquired lab-wet spectral datasets, in their raw and pretreatment state, and also to verify the impact on the prediction accuracy by the introduction of OSC. We will determine which of these datasets could be more suitable in the absence of a lab-dry measurement or when a quicker analysis is required. This will be accomplished by the use of Vis-NIR spectra and their ranges.

2. Materials and Methods

2.1. Study Area

Field spectral data (FSD) were measured in May 2019 on a (not recently ploughed) 22 ha agricultural field located at Nová Ves nad Popelkou (50°31′ N; 15°24′ E), central Bohemian region, with a mean altitude of 185 m a.s.l (Figure 1). The areas are primarily rural and devoted to winter and spring cereals and characterized by dissected relief with side valleys and toe-slopes. The total number of measurement and sampling points over the field was 130. The area chosen was representative of the soils capes that were homogenous and comparable in terms of terrain characteristics, land management, and the climatic conditions [27]. According to the World Reference Base (WRB) for soil resources (IUSS Working Group WRB, 2014), soils of this regions are characterized mainly as Cambisols on sedimentary rocks.

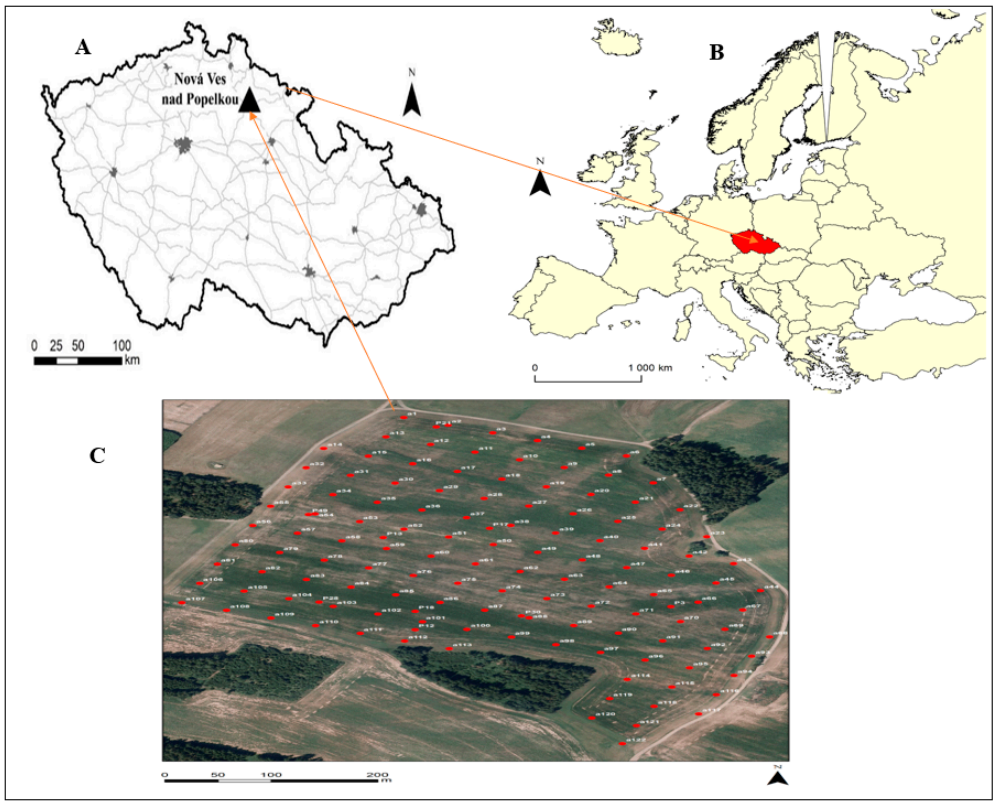


Figure 1. Location of sampling area in the Czech Republic (**A**), location of Czech Republic within Europe (**B**) and location of sampling points at Nová Ves nad Popelkou (**C**).

2.2. Soil Sampling and Spectral Measurement

The field spectral measurement was taken instantly in the field using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, CO, USA) across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050–2500 nm. Measurements from four different positions around each of the 130 sampling points were taken, and the average value was used for further analysis. The measurement and sampling points (130) were created before the field visit (Figure 1) and were located in the field using a GeoXM (Trimble Inc., Sunnyvale, CA, USA) receiver with an accuracy of 1 m. The spectrometer was standardized using the approach of Shi et al. [28]. Samples for laboratory analysis were collected from each of those positions (depth 0–20cm) while the field measurement was underway. Composite samples (approximately 150 to 200 g of soil) were placed into a well-labeled bag and transported to the laboratory for further analysis. Immediately upon reaching the laboratory, spectra readings were taken using the same spectrometer used for the field measurement again in four replicates and the average value used as the lab-wet dataset. The samples were then air-dried, gently crushed, and sieved (≤2 mm) before analyzing for SOC (ISO 11464:2006).

2.3. Spectra Pretreatment and Prediction Model Development

Before modeling, lab-wet and field data were preprocessed. The original spectral range is 350–2500 nm; however, the noisy portions between 350–399 nm were eliminated, leaving the range of 400–2500 nm before spectra pretreatments. Murray [29] stated that removing outliers improves prediction accuracy. Therefore, the outliers from both datasets were removed using a local outlier factor (LOF) algorithm procedure proposed by Breunig et al. [30]. The LOF is a measure that looks at a certain point's neighbours to figure out its density and then compares it with the density of other points and uses its local approach to better detect outliers within the neighborhoods. The field data

set was used as the reference data for the removal of outliers, meaning that the removed outliers from the field dataset were the same outliers as removed from the lab-wet dataset. In all, a total of seven outliers were removed from each dataset. With the exception of the orthogonal signal correction (OSC) (using the Unscramble Software, Version X11, CAMO, Oslo, Norway), all other pretreatment methods used were calculated with R software (R Development Core Team, Vienna, Austria, 2015). This pretreatment includes Savitzky-Golay (SG) filtering, discrete wavelet transformation (DWT), multiplicative scatter correction (MSC), standard normal variate (SNV), correction by the maximum reflectance (CMR), continuum removal (CR), first and second-order derivative (D1 and D2 respectively), as well as logarithmic transformation (Log(1/R)). We used the sgolayfilt algorithm from the signal R package for the SG filtering (adjusted for second-order polynomial fit with 30 smoothing points). For more detail about the pretreatment, the packages used can be found in [31–34].

The PLSR and SVMR predictive models built using five fold leave-group-out cross validation (which was repeated 100× to give more reliable results) were fitted separately, using either raw unsmoothed or smoothed spectra. The models were then adjusted using nine other signal transforms (SG, D1, D2, SNV, log(1/R), DWT, MSC, CR and CMR) with the exception of OSC. All transformations (except SG) were applied in two ways, i.e., the input data were either raw reflectance spectra or smoothed SG spectra and DWT. This was done in the visible (VIS; 400–800), near-infrared (NIR; 800–2500), and the whole Vis-NIR (400–2500) spectral region. In all, there were 24 different output models to be tested for each of the two datasets. Due to insignificant changes and identical performance, only transforms computed from raw spectra are shown, since they were better than using SG in more instances. Almost all the signal transformations were plotted in Figure 2 to visualize differences between different preprocessing methods. However, the reflectance and absorbance plot were separated for visual assessment of variation in the spectra and also their similarities (Figure 3). For the OSC, which is sensitive to the nonlinear algorithm, its assessment was done using PLSR and principal component regression (PCR), not SVMR because SVMR is a nonlinear algorithm. OSC was also done in three spectral regions, just as the nine other signal transformations.

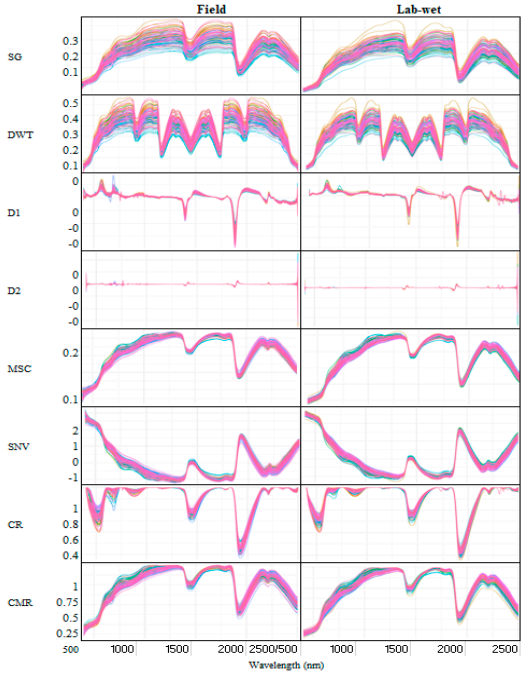


Figure 2. Spectra transforms for both field and lab-wet dataset based on eight different pretreatment methods used.

For a detailed comparison of obtained spectra (lab-wet and field), that is, to determine the stable part of the spectra (the part not affected by moisture), the part that differs, and the part with no

meaningful information, many options were explored without any significant success. Finally, we used three different combinations to analyze the datasets: median filter smoothing (MFS) with segment size of 7, spectroscopic transformation-absorbance (STA) and gap-segment second derivative (GSD) having a gap size of 6, and a segment size of 25 (Unscramble Software, Version X11, CAMO, Oslo, Norway). The order was MFS–STA–GSD.

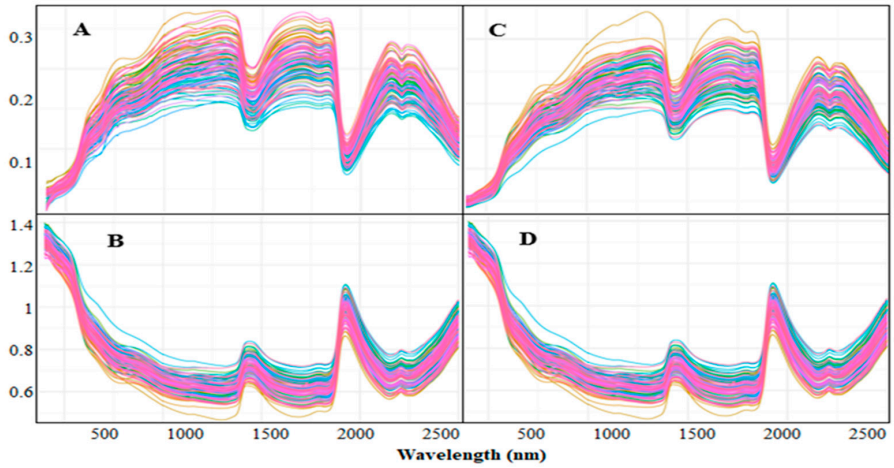


Figure 3. Spectra transforms showing field (left: **A,B**) and lab-wet (right: **C,D**) reflectance (top: **A,C**) and absorbance (bottom: **B,D**) features.

3. Results

3.1. SOC Descriptive Statistics

Table 1 is a summary statistic for SOC characteristic of soil sample in the study area, consisting of standard deviation (SD), coefficient of variation (CV), minimum, maximum, mean value, skewness and range. The statistical distributions of SOC at the study area were positively skewed with a mean value of 1.44 and a CV of 23%. These values usually indicate that the area has a medium to semi-high SOC content.

Table 1. Descriptive statistics of the soil's soil organic carbon (SOC) contents in the study area.

Property	Mean	Median	SD	SV	Kurtosis	Skewness	Range	Min	Max	CV(%)
SOC content (%) (n = 130)	1.44	1.44	0.33	0.11	2.41	0.57	2.33	0.60	2.93	23.00

SD: standard deviation, CV: coefficient of variation, n: number of samples, SV: sample variance

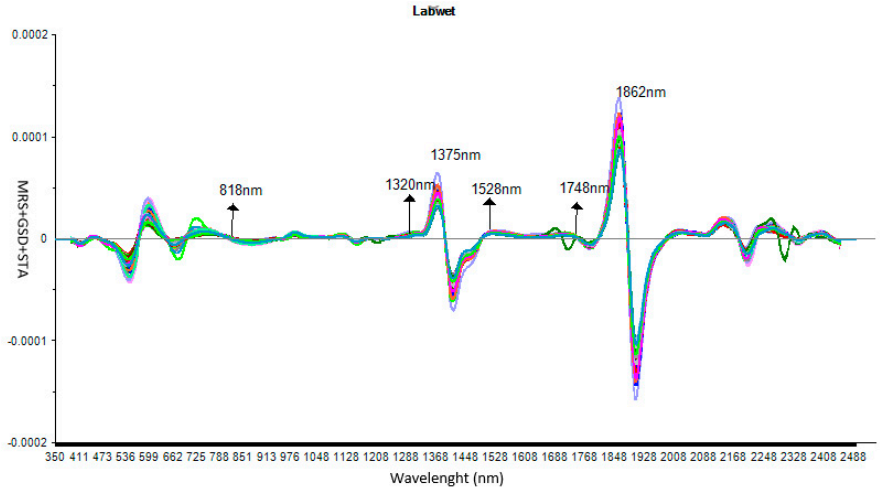
3.2. Basic Comparison of Field And Lab-Wet Spectra

Figure 3 shows the reflectance and absorbance plot from the raw data for each dataset, which was done to explore the patterns and structure of the generated spectra. The key spectral characteristics of a range of soil samples can be perceived from its mean score spectrum, which indicates the average reflectance as well as absorbance in each spectral band for the entire sample sets and the band-specific spectral variance crosswise the total spectral region.

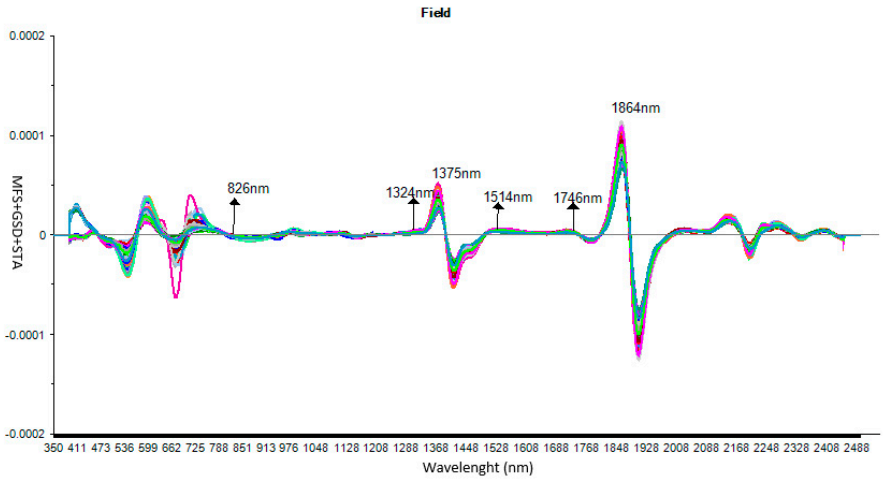
3.3. Detailed Comparison of Field and Lab-Wet Transformed Spectra

As shown by this work (Figure 4), the stable range for lab-wet spectra is from 818 nm to 1320 nm and from 1528 to 1748 nm. For field, it is located between 826 nm and 1324 nm and between 1514 and 1746 nm. This section is categorized as a region that is not influenced by moisture. The concave shape between 450 and 850 nm suggests the presence of crystalline iron [35]. This is also in agreement with Dematte et al. [36] as they stated that soil minerals containing iron, such as hematite and goethite,

result in concave shapes in the visible region of the spectrum. Nevertheless, spectra regions below 820 nm do not show any significant information for either dataset due to noise. However, this is not a justification that this range will not be suitable for prediction, but rather should be interpreted on a case-by-case basis. For example, Islam et al. [37] and Fystro [38] achieved a significantly better result for both Australian and Norwegian soils by using the visible region (350–700 nm) for prediction of SOC. This study also shows that both lab-wet and field spectra between the range of 2000 and 2400 nm display more irregular and unstable patterns, which could be attributed to the relatively low level of incoming radiation for the acquired spectra in the field resulting from the high noise rate. For the lab-wet spectra, this could just be noise or maybe other factors which, for this work, will be very difficult to explain. Poor absorption at 2265 nm for both lab-wet and field suggests the presence of gibbsite [39]. According to research by Howari et al. [40], the absorption characteristics at 990 nm are due to the presence of NaCl, while $NaHCO_3$ shows the absorption characteristics at 1470 nm, 1990 nm and 2170 nm. Absorption at 1400 nm is typically due to vibrations of water molecules and OH groups.



(**A**)



(**B**)

Figure 4. Spectral response after median filter smoothing (MFS) + spectroscopic transformation-absorbance (STA) + gap-segment second derivative (GSD) transformation for lab-wet (**A**) and field (**B**) spectra datasets (highlighting their various spectra features).

The spectrum shown in Figure 4 also illustrates one significant disparity between the field and lab-wet datasets. While the lab-wet dataset displays its peak absorbance value at 1862 nm, the field dataset shows its peak at a shifted wavelength of 1864 nm. Peak shifts are expected due to the effect of temperature change that a sample can sometimes undergo. This could say something about both the physics and chemistry of the determined samples. It may be a risky attempt to remove/mask it because one does not know whether the procedure will end up with the removal of a real and existing signal. Another peak between 1320 and 1528 nm is at the same wavelength of 1375 nm in both field and lab-wet datasets.

However, a concern about the application of field and lab-wet spectra remains because their reflectance may be heavily influenced by moisture content, though Vis-NIR spectroscopy can effectively measure samples with moisture content. Therefore, using any of them as a replacement to the dry spectra may be seen as a wrong decision, because predictive ability and accuracy of Vis-NIR measurement is negatively affected by moisture [41–43]. Despite this, some studies have shown that the field spectra can be more effective than lab-dry measurement for SOC prediction, and Reeves et al. [44] even stated that in the absence of lab-dry measurement, the field spectra should be considered as the most appropriate spectral measurement.

3.4. Comparing Field and Lab-Wet Spectra Predictive Capabilities without OSC

PLSR and SVMR, together with several pretreatment methods, were initially used to compare the prediction accuracy for both field and lab-wet spectral datasets. Leave-group-out cross validation was considered more appropriate because of its design to give more reliable results by means of five fold cross validation (which is repeated 100×). The results show (Tables 2 and 3) that for field data, PLSR gave a better prediction in almost all spectral ranges, particularly in the VIS region with R^2_{CV} = 0.42 and $RMSEP_{CV}$ = 0.26. But this output was made possible with only three out of several pretreatment methods used, namely MSC, SNV and log(1/R). However, PLSR was outperformed by SVMR also in the VIS region with R^2_{CV} = 0.47 and $RMSEP_{CV}$ = 0.24. For the lab-wet dataset, its best prediction accuracy was achieved with SVMR employing log (1/R) transformation shown by R^2_{CV} = 0.44 and $RMSEP_{CV}$ = 0.25 in the Vis-NIR region. Nonetheless, MSC and SNV also provide some improved outcomes relative to other pretreatment procedures used. This shows that the prediction from field spectra was better in the visible range while that from the lab-wet spectra was better in the Vis-NIR range. This implies that R^2_{CV} decreased from field-based (VIS) to lab-wet measurements (vis-NIR) (Table 2), while R^2_{CV} increased from lab-wet-based (VIS) to field-based (Vis-NIR) (Table 3).

Table 2. Statistics of the five fold leave-group-out cross validation for field spectra using both partial least squares regression (PLSR) and support vector machine (SVMR) on different preprocessing methods. The results were calculated as mean values from one hundred independent leave-group-out cross-validation runs.

Pre-	**Field_PLSR**					
Treatment	**VIS**		**NIR**		**VIS-NIR**	
Methods	**R^2_{cv}**	**RMSEPcv**	**R^2_{cv}**	**RMSEPcv**	**R^2_{cv}**	**RMSEPcv**
Raw	0.36	0.27	0.33	0.28	0.36	0.27
SG	0.35	0.27	0.33	0.28	0.35	0.27
DWT	0.36	0.27	0.33	0.28	0.35	0.27
D1	0.36	0.27	0.3	0.28	0.3	0.28
D2	0.21	0.3	0.17	0.31	0.19	0.3
MSC	0.42	0.26	0.27	0.29	0.3	0.28
SNV	0.42	0.26	0.26	0.29	0.28	0.29
LOG	0.42	0.26	0.36	0.27	0.4	0.26
CR	0.28	0.29	0.2	0.3	0.22	0.3
CMR	0.4	0.26	0.26	0.29	0.29	0.29

Table 2. *Cont.*

Pre-Treatment	Field_SVMR					
	VIS		NIR		VIS-NIR	
Methods	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv
Raw	0.35	0.27	0.27	0.29	0.3	0.29
SG	0.33	0.27	0.27	0.29	0.31	0.28
DWT	0.37	0.27	0.28	0.29	0.32	0.28
D1	0.33	0.29	0.24	0.32	0.26	0.31
D2	0.07	0.45	0.22	0.31	0.25	0.3
MSC	0.42	0.26	0.19	0.32	0.19	0.32
SNV	0.41	0.26	0.13	0.35	0.18	0.33
LOG	0.47	0.24	0.31	0.28	0.37	0.27
CR	0.24	0.29	0.23	0.29	0.26	0.29
CMR	0.36	0.27	0.17	0.32	0.19	0.32

Table 3. Statistics of the five fold leave-group-out cross validation for lab wet spectra using both PLSR and SVMR on different preprocessing methods. The results were calculated as mean values from one hundred independent leave-group-out cross-validation runs.

Pre-Treatment	Lab-wet_PLSR					
	VIS		NIR		VIS-NIR	
Methods	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv
Raw	0.32	0.28	0.24	0.29	0.26	0.29
SG	0.33	0.28	0.24	0.29	0.27	0.29
DWT	0.33	0.28	0.23	0.29	0.26	0.29
D1	0.29	0.28	0.21	0.31	0.26	0.29
D2	0.08	0.33	0.20	0.30	0.22	0.30
MSC	0.41	0.26	0.23	0.30	0.27	0.29
SNV	0.41	0.26	0.22	0.30	0.26	0.29
LOG	0.39	0.26	0.26	0.29	0.34	0.27
CR	0.34	0.27	0.17	0.31	0.29	0.29
CMR	0.37	0.27	0.21	0.30	0.27	0.29
Pre-Treatment	**Lab-wet_SVMR**					
	VIS		NIR		VIS-NIR	
Methods	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv	R^2_{cv}	RMSEPcv
Raw	0.33	0.28	0.31	0.28	0.39	0.26
SG	0.32	0.28	0.30	0.28	0.39	0.27
DWT	0.33	0.27	0.31	0.28	0.39	0.26
D1	0.29	0.28	0.14	0.44	0.29	0.32
D2	0.07	0.33	0.09	0.50	0.11	0.44
MSC	0.41	0.26	0.30	0.30	0.40	0.27
SNV	0.41	0.26	0.30	0.29	0.42	0.26
LOG	0.39	0.26	0.32	0.28	0.44	0.25
CR	0.34	0.27	0.19	0.30	0.34	0.27
CMR	0.38	0.27	0.27	0.31	0.40	0.27

3.5. Comparing Field and Lab-Wet Spectra Predictive Capabilities with OSC Approach

Regarding orthogonal signal correction (OSC), as compared to the other pretreatment algorithms, PLSR or PCR modeling after the OSC correction yield improved results (Table 4). For instance, the prediction accuracy for field spectra increased (for both PLSR and PCA), especially in the Vis-NIR range using PLSR with R^2_{CV} = 0.52 and $RMSEP_{CV}$ = 0.25. However, it fell short of the lab-wet dataset in the NIR and vis-NIR region (using PLSR) with R^2_{CV} = 0.54/0.55 and $RMSEP_{CV}$ = 0.24/0.24, which was the overall best prediction for the entire study. PCR and PLSR are related techniques,

and their prediction errors are comparable in most situations. However, PLSR is desired by analysts because it relates response and predictor variables so that the model describes more of the response variance with fewer parameters; also, it could become more interpretable, and the algorithm becomes computationally faster. Each of these approaches can cope with data containing a large number of strongly collinear predictor variables [45].

Table 4. Statistics for SOC prediction from field and lab-wet spectra using both PLSR and PCR based on orthogonal signal correction (OSC).

Dataset	Modelling Method	VIS		NIR		VIS-NIR	
		R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$
Field	PLSR	0.42	0.27	0.51	0.25	0.52	0.25
	PCR	0.45	0.27	0.49	0.25	0.49	0.25
Lab-wet	PLSR	0.45	0.26	0.54	0.24	0.55	0.24
	PCR	0.45	0.26	0.42	0.27	0.43	0.27

4. Discussion

4.1. Comparison of Field and Lab-Wet Spectra

The spectra measured in the field slightly differ from those measured in the laboratory wet conditions, which may be caused by differences in environmental conditions, mainly soil water content, as anticipated, such as soil moisture generally increasing spectral absorption (or decreasing reflectance) of soil compared to dry samples [46]; water replacing the air within soil voids, causing an increase in the forward scattering of light and increasing the absorption of soil at each wavelength [47,48]. The spectra (Figure 3) display similar shapes except for differences in amplitude across the entire range. For example, considering the wavelengths close to 1400 nm and 1900 nm, two obvious features occur because there are either free water or water absorbance bands. The absorption bands can differ slightly and be sharp or wide depending on the dynamics and minerals involved [49]. The absorbance order (Figure 3B,D) assigned to the presence of moisture content was: lab-wet > field, which according to Bishop [50] is attributed to the fundamental widening and bending vibrations of water and hydroxyl bonds. For instance, in overtone regions, water will absorb energy, which can be attributed to water retention forces changing from capillary forces to adsorptive ones. Knadel et al. [51] reported comparable results, too. For the reflectance (Figure 3A,C), it was contrary to that of the absorbance since the order was lab-wet < field, with the internal reflections of reflected radiation being in a water layer covering the soil. However, it was challenging to understand why the reflectance for the lab-wet was lower than for the field since both datasets were expected to have the same moisture content. According to Haubrock et al. [52], the upper surface and the lower parts vary from each other, so that spectrum analysis from the soil surface does not provide details on the properties of lower soil layers.

In this regard, and based on Figures 3 and 4, it could also indicate that our lab-wet samples have been somewhat affected with respect to transportation to the laboratory, because there may have been a certain amount of trapped heat causing variability in moisture content that we might have failed to notice. Variation in moisture content is one of the most significant effects confronting both field and naturally acquired lab-wet samples for NIR spectral prediction [42]. The lab-wet sample is influenced mainly only by moisture content because most of the other conditions that affect spectral measurement are manipulated in the laboratory. Nevertheless, field NIR reflectance measurements are susceptible to external environmental factors, such as temperature, soil moisture and soil structural factors, transient changes in weather conditions during measurement, noise, vegetation cover, illumination sources and variations in illumination due to clouds and wind. One significant concern associated with the lab-wet measurement has been the appropriate method of transportation to the laboratory. How long before they approach the laboratory and for measurement of the sample to commence, is an area of concern.

Sometimes, when the soil is being taken to the laboratory, the samples in the bags appear to 'sweat' as water condensation occurs, and the sample surface may be 'artificially' weathered. This could have also influenced the lab-wet prediction accuracy within this analysis, since for an effective lab-wet dataset, the samples should be in their natural state. Further study is needed; however, for this time around, ensuring an effective means of transportation should be paramount so that variations in the moisture content are taken care of entirely.

In certain circumstances, the spectral response sequence associated mostly with a given parameter may overlap with the response pattern of another factor and thus hinder the estimation effect of that given factor. Therefore, its necessary to understand the physical activity component as well as the environmental conditions of the soil [53]. Some of these components may have direct/indirect bearing on soil spectra, especially within the Vis-NIR region of the soil, in a particular way [54]. For example, according to Adar et al. [55], some absorption features may overlap in such a way that the absorption spectra related to one soil component can be masked, twisted or moved to another position where other soil components may differ. One instance is spectral variation resulting from changes in iron oxide content that can nullify differences in absorption due to organic matter [56]. The NIR spectra contain a combination of diffuse and specular reflectance. Depending on the chemical nature of the sample itself, different wavelengths of the incident light also experience different absorption of the sample. In most cases, this signal may represent our area of interest, so it could be critical to measure it. In some cases, the particle size of the component along the path length may cause a diversion of light at different angles, depending on wavelength, leading to scattering effects, which is a major cause of variation in the Vis-NIR region. Scattering effects can be both additive and multiplicative, which can produce a baseline effect, displacement of the spectrum along the vertical axis, and also modify the local slope of the spectrum [57,58].

4.2. Spectra Pretreatment and Prediction Models

Aside from the log(1/R) transformation, MSC and SNV also show some improved results, especially in the visible range for both field and lab-wet data. This is an indication that the light scatter effect, and the baseline displacement of the spectrum, was one of the main factors affecting the spectroradiometer signal in the visible region [59]. For example, based on Tables 2 and 3, the reason why the prediction accuracy for both field and lab-wet data was better using MSC and SNV than other pretreatment methods (except for log(1/R)) could be attributed to the above-mentioned effect, which was minimized by the use of these pretreatment methods on both datasets. In Vis-NIR region, for instance, the prediction accuracy (using MSC and SNV) reduces especially for the lab-wet (SVMR) data. This is an indication that the above-mentioned effect was not dominant in that region or that it was masked by other components, making its minimization challenging (notably for the lab-wet dataset). Martens et al. [60] proposed that excluding certain parts of the spectral axis that do not represent any necessary information (baseline) would go a long way towards improving the accuracy of the prediction. This makes good spectroscopic sense, however, detecting these parts, particularly for the Vis-NIR signal, is difficult. That is why, typically, the pretreatment is applied across the entire spectra [59].

The reason why the prediction accuracy for the field was better than that of the lab-wet in visible range but less accurate than the lab-wet in the Vis-NIR region using the log transformation, could be attributed to the dominance of nonlinearity responses for both datasets, or more nonlinearity appearing in the visible region than in the Vis-NIR, or less in the visible than Vis-NIR region. According to Minasny et al. [41], the presence of soil moisture does have a substantial, complex and nonlinear impact on reflectance spectra. Therefore, the transformation of reflectance to absorbance using log(1/R) helps to highlight the edges of the absorption characteristics and helps to attain linearization between the spectra and the SOC content [61]. This implies that most of the factors in the absorbance spectra that could have an influence on the spectral measurement were minimized to some extent to improve

prediction. This makes linearization a crucial step for regression models, as many linear modeling responses are easier with nonlinear responses [62].

The use of nine preprocessing methods, i.e., SG, DWT, D1, D2, MSC, SVN, log(1/R), CR and CMR, resulted in a mixed output (better or worse) compared to raw spectra, while the use of SG and DWT, in combination with the above-mentioned pretreatment, did not show any significant improvement compared to using those pretreatments on the raw data alone (Tables 2 and 3); their results are therefore not included in this paper. With this in mind, it is no surprise that log transformation is one of the most common transformations in SOC spectroscopic estimation. This reinforces the need for at least eight or more components to achieve a reasonable estimate, as also reported, for example, by Moron and Cozzolino [63] and Mouazen et al. [64]. The spectral range along the path from 350 to 2500 nm can differ due to several factors causing disparity, and the more the disparity, the less accurate the results. Reducing or eliminating some of the most dominant disparity could improve the accuracy of predictions, as shown in this work. According to the findings shown in this analysis, the OSC of NIR spectra seems to be a successful strategy to boost multivariate calibration models. The findings suggest that the OSC approach also eliminates details from the Vis-NIR data that are not required between the response and predicted variable, and ends up with improved prediction accuracy. This implies that though some pretreatment often removes unrelated attributes from the dataset, that process may end up with the removal of important information. Therefore, in certain instances, the prediction is also positively or negatively affected, as shown by this study (both in lab-wet and field datasets) using several pretreatment methods (Tables 2 and 3). This is also in agreement with Wold et al. [24]. Without OSC, the highest prediction accuracy for lab-wet and field data was R^2_{CV} = 0.44 and $RMSEP_{CV}$ = 0.25 and R^2_{CV} = 0.47 and $RMSEP_{CV}$ = 0.24, respectively, and with OSC, lab-wet was R^2_{CV} = 0.55 and $RMSEP_{CV}$ = 0.24 and field was R^2_{CV} = 0.52 and $RMSEP_{CV}$ = 0.25. In order to use OSC for filtering the signal matrix, a response vector is necessarily required. Similarly, spectra used for the characterization of soil properties such as SOC may appear noisy, and filtering would be warranted. Though OSC has been useful for signal correction for NIR in other analyses, it is rarely used for spectral analyses involving SOC. Despite the improvement brought to the prediction accuracy for both field and lab-wet data by its introduction, further investigation is still needed, such as using it on a larger amount of data, a different type of soil, location, soil variability and many more. This study also shares an opposite view to that of Reeve et al. [44] suggesting that the field spectra should be the most suitable spectral measurement in the absence of laboratory-dry measurement. This is because the lab-wet data with OSC give a slightly better result than the field data. Nevertheless, this should be a case-by-case evaluation (between lab-wet and field spectrum measurement).

Quantifying uncertainty is important for a number of reasons. Measuring uncertainty is needed for the testing of scientific hypotheses [65]. This can improve accuracy by allowing logical combinations of several information sources, such as repeated measurements, other sensors or background knowledge. For example, changes in external environmental conditions during field spectra measurement and ensuring that wet samples do not absorb additional moisture during transport are areas of concern. This particular field is really challenging for SOC prediction, because very poor results have been reported over the years, particularly with laboratory dry spectra measurement [66]. It is important to verify the source of uncertainty from which the sample is collected. Although the R^2 value was not so high for this analysis, it was considered one of the best, based on the history of related research in this area. We believe a detailed analysis of uncertainty about low predictive accuracy is required for this field. This research has now produced results on field and wet spectra measurement in relation to the already existing lab-dry measurement.

5. Conclusions

In this study, the performance of lab-wet and field spectra measurement was evaluated and compared to determine the most appropriate approach without lab-dry measurement. Soil spectra measurement in the field or in wet conditions may carry exclusive and imperative information

about several soil properties still in their natural state. The lab-dry measurement remains the most appropriate for prediction of SOC and some other soil properties. However, field and especially lab-wet measurements can be useful for SOC prediction with the help of pretreatment approaches. Nevertheless, moisture content remains the most challenging effect confronting both lab-wet and field measurements. Obtaining a procedure that would enable predicting soil properties using measurements taken under field conditions or on wet sample could save valuable time needed otherwise for soil sample collection and drying.

The OSC-PLSR method was proven during this study to be the best spectra pretreatment and modeling approach for SOC content estimation via Vis–NIR when dealing with both field and especially lab-wet spectral datasets. OSC-PLSR provided the most accurate result using the lab-wet dataset compared to the nine other tested spectra preprocessing methods, i.e., SG smoothing, DWT, D1, D2, MSC, SNV, CR, Log(1/R) and CMR. Without OSC, log(1/R), MSC, and SNV methods (using SVMR) were better in prediction accuracy based on the field spectra prediction accuracy in the visible region, and concurrently MSC and SNV in the visible region and log(1/R) in the Vis-NIR region on the lab-wet spectra data.

This research reveals many similarities between field and lab-wet spectra measurements with a few variations. The prediction accuracy for lab-wet data was better than for the field spectra, especially with the introduction of OSC (both in NIR and Vis-NIR regions), unlike the use of the other pretreatment approaches.

Due to unknown interactions between soil chromophores, it is difficult to determine the most important wavelengths to describe the composition of the soil. Nonetheless, for quantitative analysis of soil spectra, the optimal bandwidth and number of channels can be very dependent on the soil heterogeneity and the properties to be studied. In addition, further data treatment for lab-wet spectroscopy would be required in order to compete with lab-dry methods, in particular by reducing or removing the effect of moisture. Although the lab-wet data was marginally better than field spectra (Vis-NIR, OSC), and obtained the highest predictive accuracy based on this analysis, this paper proposes that, in the absence of a lab-dry measurement, both datasets may be appropriate, because field spectral measurement was also better in the visible region for all pretreatments, including the OSC. Further study is still needed, especially using a lab-wet data with a proper transportation system to the laboratory.

Author Contributions: Conceptualization, J.K.M.B. and L.B.; methodology, J.K.M.B.; software, J.K.M.B.; validation, J.K.M.B.; formal analysis, J.K.M.B.; investigation, J.K.M.B.; resources, J.K.M.B., K.N., A.K.; data curation, J.K.M.B., A.K.; writing—original draft preparation, J.K.M.B.; writing—review and editing, J.K.M.B., P.C.A.; visualization, J.K.M.B.; supervision, L.B.; project administration, L.B.; funding acquisition, J.K.M.B., L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Czech University of Life Sciences Prague, grant number 21130/1312/3131, and by the Czech Science Foundation, grant number 17-27726S.

Acknowledgments: We will also like to acknowledge the NutRisk grant (European Regional Development Fund, project Center for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products), number CZ.02.1.01/0.0/0.0/16_019/0000845.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Scharlemann, J.P.; Tanner, E.V.; Hiederer, R.; Kapos, V. Global soil carbon: Understanding and managing the largest terrestrial carbon pool. *Carbon Manag.* **2014**, *5*, 81–91. [CrossRef]
2. Stockmann, U.; Adams, M.A.; Crawford, J.W.; Field, D.J.; Henakaarchchi, N.; Jenkins, M.; Minasny, B.; McBratney, A.B.; De Courcelles, V.D.R.; Singh, K.; et al. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agric. Ecosyst. Environ.* **2013**, *164*, 80–99. [CrossRef]
3. Bouma, J.; McBratney, A. Framing soils as an actor when dealing with wicked environmental problems. *Geoderma* **2013**, *200*, 130–139. [CrossRef]

4. Vargas-Rojas, R.; Cuevas-Corona, R.; Yigini, Y.; Tong, Y.; Bazza, Z.; Wiese, L. Unlocking the potential of soil organic carbon: A feasible way forward. In *International Yearbook of Soil Law and Policy*; Springer: Cham, Switzerland, 2018; pp. 373–395.
5. Hutengs, C.; Ludwig, B.; Jung, A.; Eisele, A.; Vohland, M. Comparison of portable and bench-top spectrometers for mid-infrared diffuse reflectance measurements of soils. *Sensors* **2018**, *18*, 993. [CrossRef] [PubMed]
6. Nocita, M.; Stevens, A.; Van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Ben Dor, E.; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil Spectroscopy: An alternative to wet chemistry for soil monitoring. *Adv. Agron.* **2015**, *132*, 139–159.
7. Stevens, A.; Nocita, M.; Toth, G.L.; Montanarella, L.; Van Wesemael, B. Prediction of soil organic carbon at the European Scale by sisible and near InfraRed reflectance spectroscopy. *PLoS ONE* **2013**, *8*, e66409. [CrossRef]
8. Vohland, M.; Besold, J.; Hill, J.; Fründ, H.C. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* **2011**, *166*, 198–205. [CrossRef]
9. Xie, H.; Yang, X.M.; Drury, C.F.; Yang, J.Y.; Zhang, X.D. Predicting soil organic carbon and total nitrogen using mid and near-infrared spectra for Brookston clay loam soil in Southwestern Ontario, Canada. *Can. J. Soil Sci.* **2011**, *91*, 53–63. [CrossRef]
10. Ji, W.; Li, S.; Chen, S.; Shi, Z.; Viscarra-Rossel, R.; Mouazen, A.M. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil Tillage Res.* **2016**, *155*, 492–500. [CrossRef]
11. Udelhoven, T.; Emmerling, C.; Jarmer, T. Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. *Plant Soil* **2003**, *251*, 319–329. [CrossRef]
12. Stevens, A.; Van Wesemael, B.; Bartholomeus, H.; Rosillon, D.; Tychon, B.; Ben-Dor, E. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma* **2008**, *144*, 395–404. [CrossRef]
13. Viscarra-Rossel, R.; Behrens, T.; Ben-Dor, E.; Brown, D.; Demattê, J.; Shepherd, K.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; et al. A global spectral library to characterize the world's soil. *Earth Sci. Rev.* **2016**, *155*, 198–230. [CrossRef]
14. Christy, C. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Comput. Electron. Agric.* **2008**, *61*, 10–19. [CrossRef]
15. Nocita, M.; Kooistra, L.; Bachmann, M.; Müller, A.; Powell, M.; Weel, S. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma* **2011**, *167*, 295–302. [CrossRef]
16. Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Lioy, R.; Hoffmann, L.; Van Wesemael, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45. [CrossRef]
17. Sparks, D.L. *Soil Physical Chemistry*; CRC Press: Boca Raton, FL, USA, 1998 7 July.
18. Nocita, M.; Stevens, A.; Noon, C.; Van Wesemael, B. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. *Geoderma* **2013**, *199*, 37–42. [CrossRef]
19. Wijewardane, N.K.; Ge, Y.; Morgan, C.L.S. Prediction of soil organic and inorganic carbon at different moisture contents with dry ground VNIR: A comparative study of different approaches. *Eur. J. Soil Sci.* **2016**, *67*, 605–615. [CrossRef]
20. Rienzi, E.A.; Mijatovic, B.; Mueller, T.G.; Matocha, C.J.; Sikora, F.J.; Castrignanò, A. Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2014**, *78*, 958–967. [CrossRef]
21. Barnard, R.L.; Blazewicz, S.J.; Firestone, M.K. Rewetting of soil: Revisiting the origin of soil CO_2 emissions. *Soil Biol. Biochem.* **2020**, *147*, 107819. [CrossRef]
22. Bailey, V.; Pries, C.E.H.; Lajtha, K. What do we know about soil carbon destabilization? *Environ. Res. Lett.* **2019**, *14*, 083004. [CrossRef]
23. Birch, H.F. The effect of soil drying on humus decomposition and nitrogen availability. *Plant Soil* **1958**, *10*, 9–31. [CrossRef]
24. Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175–185. [CrossRef]

25. Liu, S.; Shen, H.; Chen, S.; Zhao, X.; Biswas, A.; Jia, X.; Shi, Z.; Fang, J. Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment. *Geoderma* **2019**, *348*, 37–44. [CrossRef]
26. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128. [CrossRef]
27. Schmidt, K.; Behrens, T.; Friedrich, K.; Scholten, T. A method to generate soilscapes from soil maps. *J. Plant Nutr. Soil Sci.* **2009**, *173*, 163–172. [CrossRef]
28. Shi, T.; Wang, J.; Chen, Y.; Wu, G. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 95–103. [CrossRef]
29. Murray, I. Aspects of interpretation of NIR spectra. In *Analytical Application of Spectroscopy*; Creaser, C.S., Davies, A.M.C., Eds.; Royal Society of Chemistry: London, UK, 1988; pp. 9–21.
30. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16 May 2000; pp. 93–104.
31. Wehrens, R.; Mevik, B.H. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* **2007**, *18*. [CrossRef]
32. Renka, R.J. Algorithm 751: TRIPACK: A constrained two-dimensional Delaunay triangulation package. *ACM Trans. Math. Softw.* **1996**, *22*, 1–8. [CrossRef]
33. Aldrich, E. A package of functions for computing wavelet filters, wavelet transforms and multi-resolution Analyses. 2013. Available online: http://cran.rproject.org/web/packages/wavelets/wavelets.pdf (accessed on 21 September 2012).
34. Duckworth, J. Mathematical data pre-processing. *Near Infrared Spectrosc. Agric.* **2004**, *44*, 113–132.
35. Vitorello, I.; Galvão, L.S. Spectral properties of geologic materials in the 400-to 2500 nm range: Review for applications to mineral exploration and lithologic mapping. *Photo Interprétat.* **1996**, *34*, 77–99.
36. Demattê, J.A.; Campos, R.C.; Alves, M.C.; Fiorio, P.R.; Nanni, M.R. Visible–NIR reflectance: A new approach on soil evaluation. *Geoderma* **2004**, *121*, 95–112. [CrossRef]
37. Islam, K.; Singh, B.; McBratney, A. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Res.* **2003**, *41*, 1101–1114. [CrossRef]
38. Fystro, G. The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis–NIR spectroscopy and comparative methods. *Plant Soil* **2002**, *246*, 139–149. [CrossRef]
39. Madeira Netto, J.S. Spectral reflectance properties of soils. *Photo Interprétat.* **1996**, *34*, 59–76.
40. Howari, F.M.; Goodell, P.C.; Miyamoto, S. Spectral properties of salt crusts formed on saline soils. *J. Environ. Qual.* **2002**, *31*, 1453–1461. [CrossRef] [PubMed]
41. Minasny, B.; McBratney, A.; Bellon-Maurel, V.; Roger, J.-M.; Gobrecht, A.; Ferrand, L.; Joalland, S. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma* **2011**, *167*, 118–124. [CrossRef]
42. Bogrekci, I.; Lee, W.S. Effects of soil moisture content on absorbance spectra of sandy soils in sensing phosphorus concentrations using UV-VIS-NIR spectroscopy. *Trans. ASABE* **2006**, *49*, 1175–1180. [CrossRef]
43. Mouazen, A.; Karoui, R.; De Baerdemaeker, J.; Ramon, H. Characterization of soil water content using measured visible and near infrared spectra. *Soil Sci. Soc. Am. J.* **2006**, *70*, 1295–1302. [CrossRef]
44. Reeves, J.; Mccarty, G.; Mimmo, T. The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. *Environ. Pollut.* **2002**, *116*, S277–S284. [CrossRef]
45. Wentzell, P.D.; Montoto, L.V. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 257–279. [CrossRef]
46. Baumgardner, M.F.; Silva, L.F.; Biehl, L.L.; Stoner, E.R. Reflectance properties of soils. *Adv. Agron.* **1986**, *38*, 1–44. [CrossRef]
47. Viscarra-Rossel, R.; McBratney, A. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. *Geoderma* **1998**, *85*, 19–39. [CrossRef]
48. Mouazen, A.; De Baerdemaeker, J.; Ramon, H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res.* **2005**, *80*, 171–183. [CrossRef]

49. Clark, C.C.; Clark, L.; Clark, L. "Anting" behavior by common grackles and European starlings. *Wilson Bull.* **1990**, *102*, 167–169.
50. Bishop, C.W. *Expansion of Moisture Monitoring Network at the Subsurface Disposal Area of the Radioactive Waste Management Complex*; INEL-94/0144; Lockheed Idaho Technologies Company: Idaho Falls, ID, USA, 1994.
51. Knadel, M.; Deng, F.; Alinejadian, A.; De Jonge, L.W.; Moldrup, P.; Greve, M. The effects of moisture conditions-from wet to hyper dry-on visible near-infrared spectra of Danish reference soils. *Soil Sci. Soc. Am. J.* **2014**, *78*, 422–433. [CrossRef]
52. Haubrock, S.; Chabrillat, S.; Lemmnitz, C.; Kaufmann, H. Surface soil moisture quantification models from reflectance data under field conditions. *Int. J. Remote. Sens.* **2008**, *29*, 3–29. [CrossRef]
53. Dwivedi, D.; Riley, W.; Torn, M.; Spycher, N.; Maggi, F.; Tang, J. Mineral properties, microbes, transport, and plant-input profiles control vertical distribution and age of soil carbon stocks. *Soil Biol. Biochem.* **2017**, *107*, 244–259. [CrossRef]
54. Price, J.C. How unique are spectral signatures? *Remote. Sens. Environ.* **1994**, *49*, 181–186. [CrossRef]
55. Adar, S.; Shkolnisky, Y.; Ben-Dor, E. Change detection of soils under small-scale laboratory conditions using imaging spectroscopy sensors. *Geoderma* **2014**, *216*, 19–29. [CrossRef]
56. Poulin, B.A.; Ryan, J.N.; Aiken, G.R. Effects of iron on optical properties of dissolved organic matter. *Environ. Sci. Technol.* **2014**, *48*, 10098–10106. [CrossRef]
57. Maleki, M.R.; Mouazen, A.; Ramon, H.; De Baerdemaeker, J. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosyst. Eng.* **2007**, *96*, 427–433. [CrossRef]
58. Pelliccia, D. Instruments & data tools, two scatter correction techniques for NIR spectroscopy. 2019. Available online: https://www.idtools.com.au/two-scatter-correction-techniques-nir-spectroscopy-python/ (accessed on 21 July 2018).
59. Rinnan, Å.; Van Den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
60. Martens, H.; Jensen, S.A.; Geladi, P. Multivariate linearity transformation for near-infrared reflectance spectrometry. In Proceedings of the Nordic Symposium on Applied Statistics, Stavanger, Norway, 12–14 June 1983; Stokkand Forlag Publishers: Stavanger, Norway, 1983; pp. 205–234.
61. West, J.B.; Bowen, G.J.; Dawson, T.E.; Tu, K.P. *Isoscapes: Understanding Movement, Pattern, and Process on Earth Through Isotope Mapping*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
62. Shi, T.; Chen, Y.; Liu, Y.; Wu, G. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **2014**, *265*, 166–176. [CrossRef] [PubMed]
63. Morón, A.; Cozzolino, D. Application of near infrared reflectance spectroscopy for the analysis of organic C, total N and pH in soils of Uruguay. *J. Near Infrared Spectrosc.* **2002**, *10*, 215–221. [CrossRef]
64. Mouazen, A.; Maleki, M.; De Baerdemaeker, J.; Ramon, H. On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil Tillage Res.* **2007**, *93*, 13–27. [CrossRef]
65. Hobbs, J.; Braverman, A.; Cressie, N.; Granat, R.; Gunson, M. Simulation-based uncertainty quantification for estimating atmospheric CO_2 from satellite data. *SIAM/ASA J. Uncertain. Quantif.* **2017**, *5*, 956–985. [CrossRef]
66. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Boruvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and sentinel-2 spectral imaging. *Remote Sens. Environ.* **2018**, *218*, 89–103. [CrossRef]

Geoderma 388 (2021) 114945

Contents lists available at ScienceDirect

Geoderma

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

Does the limited use of orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic carbon need attention?

James Kobina Mensah Biney [a,*], Johanna Ruth Blöcher [b], Luboš Borůvka [a], Radim Vašát [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic
[b] Department of Water Resources and Environmental Modeling, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic

ARTICLE INFO

Handling Editor: Ingrid Kögel-Knabner

Keywords:
Agricultural soil
Pre-treatment algorithms (OSC)
Machine learning algorithms
SOC

ABSTRACT

Visible-Near-infrared (Vis-NIR) spectroscopy is a relatively modern method that can be used to predict soil properties such as soil organic carbon (SOC). Predictions of soil properties with Vis-NIR requires pre-processing algorithms. Applying a wrong type or applying too extreme pre-processing algorithms may result in the removal of valuable information or may even introduce unwanted variations. This can negatively affect the prediction accuracy of the property being studied. Orthogonal signal correction (OSC) pre-processing method has been used in other fields for visible and near infrared spectra improvement. However, the application of OSC application in soil science remains limited. The main idea behind OSC is the removal of only unwanted variation from the spectrum unlike some other pre-treatment methods which are believed to remove valuable information in the process of removing undesirable variation. This study verifies the effectiveness of the OSC against nine commonly used pre-treatment methods across three different agricultural fields for both lab-dry and in-field spectra. For the prediction, partial least square regression (PLSR) and support vector machine regression (SVMR) algorithms were used. In this study, the OSC method overall improved prediction accuracy the most (e.g. with OSC the best result was $R^2_{CV} = 0.79$, without OSC the best result was $R^2_{CV} = 0.62$) and is therefore a promising tool that should be included in further studies on different soils and other soil properties.

1. Main text

Visible-Near-infrared (Vis-NIR) spectroscopy contains information on different soil properties along its wavelengths between 350 and 2500 nm. However, the measurements can vary under different surface and environmental conditions, e.g. differences in the weather conditions, environment, humidity, temperature, human factors, spectral noise, and atmospheric attenuation (Rinnan et al., 2009). The use of spectroscopy in the range of Vis-NIR under laboratory-controlled conditions has been noted for its reliable prediction of soil organic carbon (SOC) compared to field and remote sensing platforms (Xie et al., 2011). This may be because a standardized protocol is used, see e.g. Romero et al. (2018) and Ben Dor et al. (2015). This can remove lots of unwanted artifacts and stabilize the measurements. Also, external environmental conditions that may have an influence on the spectrum are kept under observation and manipulate under laboratory conditions (Hulley et al., 2010). Nevertheless, other issues such as spectrometer instability, illumination source, detector output, and sample preparation may persist in the laboratory environment. As a result of the above-mentioned disturbing factors, data pre-processing has become a vital tool to achieve reliable results. The primary function of all pre-processing methods is to reduce the unmodeled variability in the data and to enhance the feature sought in the spectra, which is often a linear (simple) relation. However, choosing the most robust pre-processing technique can be challenging because applying a wrong type or applying a pre-processing method that is too severe can result in the removal of valuable information or even the introduction of unwanted variation (Rinnan et al., 2009; Engel et al., 2013). This, according to Wold et al. (2008), could have a negative effect on the model's prediction accuracy. Multiple scatter correction (MSC), standard normal variate (SNV), Savitzky-Golay (SG) filtering,

* Corresponding author.
E-mail address: biney@af.czu.cz (J.K.M. Biney).

https://doi.org/10.1016/j.geoderma.2021.114945
Received 15 September 2020; Received in revised form 11 January 2021; Accepted 12 January 2021
Available online 29 January 2021

derivatives and logarithmic (log(1/R)) transformation are some of the commonly used pre-processing methods (Vašát et al., 2017). However, some of these pre-processing algorithms, particularly MSC and SNV, may end up removing useful information in the attempt to remove undesirable signals (Wold et al., 2008).

The orthogonal signal correction (OSC) for NIR spectral correction was initially introduced by Wold et al. (1998). It had been reasonably utilized on spectra in other fields, but its application in soil science remains limited. The core principle of OSC technique is to exclude variations that are not related to the predicted parameter. It does this without eliminating essential information as some other processing method might do (Wold et al., 1998). With OSC, the assumption is that any variation not related to the response variable is an artifact and should be filtered without removing any vital information (Wold et al., 1998; Engel et al., 2013). This is achieved by ensuring that the omitted component is either mathematically orthogonal to the response, or as close to orthogonal as possible. This research aims to verify OSC's effectiveness in terms of predictive accuracy of SOC against nine of the most commonly used pre-processing methods for both VIS and VIS-NIR spectra.

We used 259 soil samples taken from a depth 0–25 cm from three separate agricultural fields (field 1: 53, year 2012; field 2: 76, year 2013; field 3: 130, year 2019) within the Czech Republic using the sampling grid approach. According to the World Reference Base (WRB) for soil resources (IUSS Working Group, WRB, 2014), the soils are characterized as Rendzic Leptosol and Colluvial soil (field 1), Haplic Luvisol, Regosol and Colluvial soil (field 2) and Cambisol on sedimentary rocks (field 3). For samples from field 1 and 2, the samples were air-dried, gently crushed, and sieved ($\leq$2 mm) before analysis as lab-dry measurement (ISO 11464:2006). SOC was measured as the total oxidizable carbon using the wet oxidation approach (ISO 14235:1998). The spectral measurements were carried out under laboratory conditions using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050–2500 nm. After every 10 samples, the sensor was re-calibrated using regular white reference Spectralon® (Labsphere, North Sutton, NH, USA). For samples from field 3, the spectra measurement was taken in-situ. Samples for laboratory analysis were also collected. In summary, the final three datasets consisted of lab-dry (LD) spectra (field 1& 2) as well as in-field spectra (FS; field 3). Modeling was performed for each field separately. Outliers within the datasets were eliminated using a local outlier factor (LOF) algorithm proposed by Breunig et al. (2000). In total nine outliers were removed (field 1:1; field 2: 1; field 3: 7). Except for the orthogonal signal correction (OSC) (using the Unscrambler Program, version X11, CAMO, Norway), all other pre-treatment methods used were determined utilizing R software (R Development Core Team, 2015). The pre-treatment methods include SG filtering (with a second-order polynomial fit and 21 smoothing points), discrete wavelet transformation (DWT), standard normal variate (SNV), continuum removal (CR), maximum reflectance correction (CMR), multiplicative scatter correction (MSC), first and second-order derivative (D1 and D2) (obtained by using the *locpoly* function from the KernSmooth package), and logarithmic transformation (log(1/R)). The SNV was determined by subtracting each reflectance value from the mean reflectance value of the specific spectrum and dividing it by the standard deviation of the entire spectrum. All pre-treatment methods were calculated three separate times, from raw spectra, SG spectra and DWT spectra and the best results were selected and reported. (Only transforms on the raw spectrum are shown in order to minimize space). For more information on the pre-treatment algorithms, we refer to Vašát et al. (2017) and Biney et al. (2020). The data was processed using partial least square regression (PLSR) and support vector machine regression (SVMR) models in the spectral region of visible (VIS) only and visible and near-infrared (Vis-NIR) built using five-fold leave-group-out cross-validation. PLSR model was tuned in the way that the maximum number of model components was set to 10. Then the model runs and tests itself for each number of components, i.e. from 1 to 10. Based on cross validation (which is leave-group-out with 5 segments) the optimal number of components is chosen based on the lowest RMSE. Finally, the model is re-calibrated with the optimal number of components and validated, R^2 and RMSE calculated. Similarly, the SVMR is tuned with different cost values (specifically 0.001, 0.01, 0.1 and 1), using a linear kernel and an epsilon of 0.1. Based on the RMSE, the best cost parameter is selected from a 10-fold cross-validation. The overall predictive accuracy of the models was

Table 1
Statistics of the five-fold leave-group-out cross-validation for field 1, 2 and 3 spectra using both PLSR and SVMR on different pre-processing methods: Raw (initial spectrum), Savitzky–Golay (SG), discrete wavelet transformation (DWT), first derivative (D1), second derivative (D2), multiplicative scatter correction (MSC), standard normal variate (SNV), log transformed (LOG), continuum removal (CR), maximum reflectance correction (CMR) and orthogonal signal correction (OSC). The pre-processing was applied to raw initial spectra.

Pre-treatment		Raw	SG	DWT	D1	D2	MSC	SNV	LOG	CR	CMR	OSC
Field 1	R^2_{cv}	0.51	0.47	0.43	0.40	0.01	0.47	0.38	0.44	0.11	0.45	0.64
VIS PLSR	RMSEPcv	0.13	0.14	0.15	0.15	0.20	0.14	0.16	0.15	0.19	0.14	0.13
VIS-NIR PLSR	R^2_{cv}	0.45	0.46	0.56	0.27	0.02	0.54	0.57	0.56	0.33	0.53	0.79
	RMSEPcv	0.15	0.15	0.13	0.17	0.20	0.13	0.13	0.13	0.16	0.13	0.10
VIS SVMR	R^2_{cv}	0.54	0.50	0.51	0.36	0.10	0.45	0.37	0.47	0.07	0.21	0.51
	RMSEPcv	0.13	0.14	0.13	0.17	0.22	0.15	0.16	0.14	0.18	0.17	0.14
VIS-NIR SVMR	R^2_{cv}	0.52	0.55	0.59	0.44	0.31	0.59	0.57	0.61	0.12	0.52	0.63
	RMSEPcv	0.13	0.13	0.12	0.14	0.16	0.13	0.13	0.12	0.18	0.14	0.12
Field 2	R^2_{cv}	0.45	0.45	0.45	0.36	0.22	0.48	0.54	0.49	0.11	0.50	0.73
VIS PLSR	RMSEPcv	0.14	0.14	0.14	0.15	0.17	0.14	0.13	0.14	0.19	0.14	0.11
VIS-NIR PLSR	R^2_{cv}	0.44	0.41	0.46	0.37	0.18	0.48	0.48	0.54	0.52	0.50	0.66
	RMSEPcv	0.15	0.15	0.15	0.15	0.18	0.14	0.14	0.13	0.13	0.14	0.13
VIS SVMR	R^2_{cv}	0.53	0.50	0.49	0.44	0.29	0.49	0.48	0.49	0.15	0.40	0.63
	RMSEPcv	0.13	0.13	0.13	0.15	0.19	0.14	0.14	0.14	0.17	0.15	0.12
VIS-NIR SVMR	R^2_{cv}	0.60	0.48	0.52	0.37	0.26	0.62	0.60	0.51	0.52	0.55	0.65
	RMSEPcv	0.12	0.14	0.13	0.16	0.16	0.12	0.12	0.13	0.13	0.13	0.12
Field 3	R^2_{cv}	0.36	0.35	0.36	0.36	0.21	0.42	0.42	0.42	0.28	0.40	0.42
VIS PLSR	RMSEPcv	0.27	0.27	0.27	0.27	0.30	0.26	0.26	0.26	0.29	0.26	0.27
VIS-NIR PLSR	R^2_{cv}	0.36	0.35	0.35	0.30	0.19	0.30	0.28	0.40	0.22	0.29	0.52
	RMSEPcv	0.27	0.27	0.27	0.28	0.30	0.28	0.29	0.26	0.30	0.29	0.25
VIS SVMR	R^2_{cv}	0.35	0.33	0.37	0.33	0.07	0.42	0.41	0.47	0.24	0.36	0.45
	RMSEPcv	0.27	0.27	0.27	0.29	0.45	0.26	0.26	0.24	0.29	0.27	0.26
VIS-NIR SVMR	R^2_{cv}	0.30	0.31	0.32	0.26	0.25	0.19	0.18	0.37	0.26	0.19	0.28
	RMSEPcv	0.29	0.28	0.28	0.31	0.30	0.32	0.33	0.27	0.29	0.32	0.29

evaluated in terms of the index of determination (R^2cv) and root mean square error of prediction (RMSEPcv).

The result shows some inconsistency with the introduction of the other pre-processing methods, where some pre-processing methods worsened the prediction accuracy (Table 1). With the implementation of the OSC, the prediction accuracy for the three data sets generally improved (Table 1). Exceptions include field 3 (VIS range) where Log(1/R) performed better. Furthermore, OSC obtained a lower error (RMSEPcv) compared to the nine pre-treatment methods used. Although the Vis-NIR range for this study was the dominant region in term of predictive accuracy, however, this range was outperformed by the visible range for field 2 (with OSC) and field 3 (without OSC).

Comparing the nine-pre-treatment methods with OSC reveals that only three (for e.g. MSC, SNV and log(1/R)) out of these pre-treatment algorithms had a better and consistent prediction accuracy for SOC (across the three field); however, in terms of the most accurate result, they all fell short to that of OSC. For MSC, one of the possible reasons could be defining an appropriate reference spectrum which according to Rinnan et al. (2009) is one of the challenges associated with MSC. Gallagher et al. (2005) suggested that adding a natural variance to the MSC (using a weighting scheme in the pre-processing phase) could help with this problem. Unfortunately, this impartially straightforward approach does not always work well with NIR data, as a more scattering effect is observed on the dataset due to the spread in the higher wavelength range, which needs correction rather than a reduction in the applied weight (Rinnan et al., 2009). Another strategy noted by Windig et al (2008) is to find the average spectrum from the MSC corrected dataset; however, according to Rinnan et al. (2009), the reference spectrum should be updated by repeating the MSC several times which could limit its accuracy. Although Dhanoa et al. (1994) demonstrate similarities between the SNV and MSC, SNV may be susceptible to noisy inputs in the spectrum because its parameter does not involve a least square fitting (Rinnan et al., 2009). The key ideal for all forms of pre-treatment methods is to only eliminate artifacts that are present in the data, without introducing unnecessary artifacts or variations to the data. According to Engel et al. (2013), one benefit of OSC is the elimination of multiple artifacts at the same time (e.g., a baseline slope and scatter effect) while ensuring prediction accuracy will be enhanced during the process. This study corroborates these findings with OSC improving the prediction accuracy for the three datasets (Table 1). However, OSC is also prone to some issues, especially the presence of a response variable before its application (both measured and predicted is needed). Additionally, OSC method often converges fast, but still requires 5–10 iteration (Trygg and Wold, 2002). The study used 5 iteration. Based on our findings for both laboratory spectra and field data, to improve SOC prediction accuracy OSC appears very promising, and should be included in studies testing Vis-NIR on more soils and other soil properties.

Declaration of Competing Interest

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

Acknowledgement

We gratefully acknowledge Radka Kodešová for field samples in 1 & 2 datasets and Aleš Klement, Karel Němeček and Miroslav Fér for their assistance in soil sampling and spectral measurement.

This study was supported by an internal grant of the Czech University of Life Sciences Prague no: 21130/1312/3131 and project No. 17-27726S of the Czech Science Foundation.

References

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245-246, 112–124. https://doi.org/10.1016/j.geoderma.2015.01.002.

Biney, J. K. M., Borůvka, L., Chapman Agyeman, P., Němeček, K., & Klement, A. (2020). Comparison of field and laboratory wet soil spectra in the Vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements. Remote Sensing, 12(18), 3082.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, vol. 29, (pp. 93-104).

Dhanoa, M.S., Lister, S.J., Sanderson, R., Barnes, R.J., 1994. The link between multiplicative scatter correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra. J. Near Infrared Spectrosc. 2 (1), 43–47. https://doi.org/10.1255/jnirs.30.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L. M.C., 2013. Breaking with trends in pre-processing? TrAC Trends Anal. Chem. 50, 96–106. https://doi.org/10.1016/j.trac.2013.04.015.

Gallagher, N.B., Blake, T.A., Gassman, P.L., 2005. Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil. J. Chemometrics 19 (5-7), 271–281. https://doi.org/10.1002/cem.929.

Hulley, G.C., Hook, S.J., Baldridge, A.M., 2010. Investigating the effects of soil moisture on thermal infrared land surface temperature and emissivity using satellite retrievals and laboratory measurements. Remote Sens. Environ. 114 (7), 1480–1493. https://doi.org/10.1016/j.rse.2010.02.002.

Rinnan, Å., Berg, F.V.D., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28 (10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

Romero, D.J., Ben-Dor, E., Demattê, J.A.M., Souza, A.B.e., Vicente, L.E., Tavares, T.R., Martello, M., Strabeli, T.F., da Silva Barros, P.P., Fiorio, P.R., Gallo, B.C., Sato, M.V., Eitelwein, M.T., 2018. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma 312, 95–103. https://doi.org/10.1016/j.geoderma.2017.09.014.

Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). J. Chemometrics 16 (3), 119–128. https://doi.org/10.1002/cem.695.

Vašát, R., Kodešová, R., Klement, A., Borůvka, L., 2017. Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. Geoderma 298, 46–53. https://doi.org/10.1016/j.geoderma.2017.03.012.

Windig, W., Shaver, J., Bro, R., 2008. Loopy MSC: A simple way to improve multiplicative scatter correction. Appl Spectrosc 62 (10), 1153–1159. https://doi.org/10.1366/000370208786049097.

Wold, S., Antti, H., Lindgren, F., Öhman, J., 1998. Orthogonal signal correction of near-infrared spectra. Chemometr. Intell. Lab. Syst. 44 (1-2), 175–185. https://doi.org/10.1016/S0169-7439(98)00109-9.

Wold, M.B., Connell, L.D., Choi, S.K., 2008. The role of spatial variability in coal seam parameters on gas outburst behaviour during coal mining. International Journal of Coal Geology 75 (1), 1–14. https://doi.org/10.1016/j.coal.2008.01.006.

Xie, H.T., Yang, X.M., Drury, C.F., Yang, J.Y., Zhang, X.D., 2011. Predicting soil organic carbon and total nitrogen using mid- and near-infrared spectra for Brookston clay loam soil in Southwestern Ontario, Canada. Can. J. Soil. Sci. 91 (1), 53–63. https://doi.org/10.4141/cjss10029.

remote sensing

MDPI

Article

Exploring the Suitability of UAS-Based Multispectral Images for Estimating Soil Organic Carbon: Comparison with Proximal Soil Sensing and Spaceborne Imagery

James Kobina Mensah Biney [1,*], Mohammadmehdi Saberioon [2], Luboš Borůvka [1], Jakub Houška [1,3], Radim Vašát [1], Prince Chapman Agyeman [1], João Augusto Coblinski [1,4] and Aleš Klement [1]

1 Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic; boruvka@af.czu.cz (L.B.); jakub.houska@vukoz.cz (J.H.); vasat@af.czu.cz (R.V.); agyeman@af.czu.cz (P.C.A.); coblinskijoao@gmail.com (J.A.C.); klement@af.czu.cz (A.K.)

2 Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Section 1.4 Remote Sensing and Geoinformatics, Telegrafenberg, 14473 Potsdam, Germany; saberioon@gfz-potsdam.de

3 The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Department of Landscape Ecology, 60200 Brno, Czech Republic

4 Department of Soil Science, Faculty of Agronomy, Federal University of Rio Grande do Sul, Bento Gonçalves Avenue, Porto Alegre 91540-000, Brazil

* Correspondence: biney@af.czu.cz

Abstract: Soil organic carbon (SOC) is a variable of vital environmental significance in terms of soil quality and function, global food security, and climate change mitigation. Estimation of its content and prediction accuracy on a broader scale remain crucial. Although, spectroscopy under proximal sensing remains one of the best approaches to accurately predict SOC, however, spectroscopy limitation to estimate SOC on a larger spatial scale remains a concern. Therefore, for an efficient quantification of SOC content, faster and less costly techniques are needed, recent studies have suggested the use of remote sensing approaches. The primary aim of this research was to evaluate and compare the capabilities of small Unmanned Aircraft Systems (UAS) for monitoring and estimation of SOC with those obtained from spaceborne (Sentinel-2) and proximal soil sensing (field spectroscopy measurements) on an agricultural field low in SOC content. Nine calculated spectral indices were added to the remote sensing approaches (UAS and Sentinel-2) to enhance their predictive accuracy. Modeling was carried out using various bands/wavelength (UAS (6), Sentinel-2 (9)) and the calculated spectral indices were used as independent variables to generate soil prediction models using five-fold cross-validation built using random forest (RF) and support vector machine regression (SVMR). The correlation regarding SOC and the selected indices and bands/wavelengths was determined prior to the prediction. Our results revealed that the selected spectral indices slightly influenced the output of UAS compared to Sentinel-2 dataset as the latter had only one index correlated with SOC. For prediction, the models built on UAS data had a better accuracy with RF than the two other data used. However, using SVMR, the field spectral prediction models achieved a better overall result for the entire study (log(1/R), RPD = 1.40; R^2_{CV} = 0.48; RPIQ = 1.65; RMSEPCV = 0.24), followed by UAS and then Sentinel-2, respectively. This study has shown that UAS imagery can be exploited efficiently using spectral indices.

Keywords: soil organic carbon; proximal soil sensing; remote sensing multispectral sensors; agricultural soil; spectral indices

Citation: Biney, J.K.M.; Saberioon, M.; Borůvka, L.; Houška, J.; Vašát, R.; Chapman Agyeman, P.; Coblinski, J.A.; Klement, A. Exploring the Suitability of UAS-Based Multispectral Images for Estimating Soil Organic Carbon: Comparison with Proximal Soil Sensing and Spaceborne Imagery. *Remote Sens.* **2021**, *13*, 308. https://doi.org/10.3390/rs13020308

Received: 19 November 2020
Accepted: 12 January 2021
Published: 17 January 2021

1. Introduction

Soil organic carbon (SOC) content is one of the leading indicators for soil state assessment. Therefore, a thorough and timely observation of SOC content with effective techniques is needed to better understand the function of soil within the carbon cycle

universally [1,2]. However, numerous drawbacks, including complex and unpredictable environmental conditions, and numerous soil-forming conditions, limit the efficiency and performance of their estimation. Due to these unfavorable factors, the mapping of SOC and its attributes requires time and money [2,3]. Therefore, there is a global surge toward the need for fast and less costly techniques for efficient quantification of SOC content.

In response to these challenges, the emersion of proximal soil sensing (PSS) and remote sensing (RS) approaches is described as a useful detection tool for evaluating and analyzing several soil parameters including SOC [4–6]. For proximal sensing, a physical contact is needed to obtain signal from the target using the spectrometer sensor (within 2m apart) [7]. Whereas, for remote sensing (RS), electromagnetic radiation is used to obtain, without physical contact data or occurrence [8]. Spectroscopy (visible near-infrared (vis-NIR)) under PSS is also classified as a useful tool for the accurate quantification, in laboratory [2,9,10], and in the field [11–13] of SOC content with limited resources. Its approach for soil assessment started in the years 1960 to 1980 [14] and intensified between 1990–2000s [15]. Research into vis-NIR spectroscopy approaches within soil science has also increased rapidly in the last couple of decades [16,17]. For example, with infrared spectroscopy, a sole spectrum can allow the identification of contrasting soil constituents concurrently [5]. Nevertheless, when using spectroscopy, one of the suggestions is that the accumulation of an established soil component is linear to a mixture of absorption properties within the spectral range, also the issue where organic and inorganic molecules can absorb at wavelengths beyond 2000 nm cannot be ignored [18,19]. According to Mulder et al. [20], qualitative and quantitative information on soil variables and soil classification can be collected in a cost-effective approach using RS. For example, it is difficult to disregard the short revisit duration of the Sentinel-2 imagery and the large quantity of the data set generated that is available and can also be freely downloaded [21]. In addition, the spectral composition of the soil can be calculated affordably and conveniently, thus providing a trade-off between cost and precision [22]. Nevertheless, in terms of detailed large-scale site monitoring, enhanced results classification, and data reduction, remote sensing has an advantage over PSS [23]. For measurement, RS methods can be categorized into two main types, namely spaceborne (e.g., use of satellites) and airborne (either aircraft or drone). However, aerial surveillance, employing imagery collected by satellites, manned aircraft and unmanned aerial vehicles (UAVs)(actual aircraft (Drone) itself), is one of the most commonly used RS techniques [24]. Airborne imaging can provide a more precise mapping of the variability found in agricultural fields. Even from a single flight mission, the information produced can cover wide areas because the aircraft has adequate flight duration [25]. Moreover, airborne sensors can also provide site segmentation data based on soil heterogeneity, while expanding existing soil property datasets to support digital soil mapping [20]. Spaceborne remotely sensed imagery, on the other hand, has an enormous potential as an enabling instrument for generating soil profile maps, due to the relation that can be created between the soil's complex chemical bonds and electromagnetic radiation. For example, with the introduction of the first satellites in the 1980s, optical satellite (multispectral) imagery was widely utilized for a comprehensive SOC assessment [26]. However, the traditional airborne and satellite remote sensing frameworks where most sensors (e.g., multispectral, hyperspectral, etc.,) are mounted, have not always satisfied the researchers' and environmental demands [27]. In case of environmental applications, some of these platforms are prone to several issues like high cost and especially poor spatial and temporal resolution. Satellite data can be very appealing because of its broad spatial coverage including inaccessible areas that were historically too remote or too harmful to reach while using traditional aerial photography [28,29]. Nevertheless, issues such as low resolution and excessive noise while using Hyperion satellites [11] and the 16-day Landsat-8 revisit period suggest that the available options for time series research and bare soil observation may be minimal [30,31]. According to Crucil et al. [32], some of the above-mentioned issues with spaceborne still remain unresolved even with the emersion of

new satellites, especially Sentinel-2. Moreover, Sentinel-2 data may even have to undergo several pre-processing steps which could affect its prediction capability of soil properties.

Although remote sensing imagery (e.g., hyperspectral images) offers detailed bare soil spectral data, the effects posed by some RS factors, especially the soil water content and dissolved organic material within the soil, cannot be ignored [33,34]. Consequently, one or more appropriate spectral indices could be necessary to help limiting the effect posed by the above listed factors on RS imagery [33,35,36]. For examples, Jin et al. [37], reported improved results by using several indices to predict soil organic matter.

Over the past decade, the development of UAS, also known as unmanned aircraft system, has made it possible to obtain valuable data that have been beneficial to determine spatial variability within soil properties [27], especially SOC [38] that would have been difficult to identify utilizing conventional frameworks for RS. UAS can be categorized depending on the nature of its wings either as non-movable (non-mobile) or movable (mobile), with the non-mobile wings typically having higher speed and greater duration, while the mobile wings can offer greater maneuverability. UAS appear as versatile platforms with the potential to augment RS survey collected from spaceborne or manned aircraft [27]. Although UAS cannot compete based on the spatial coverage with satellite imagery, they offer unparalleled spatial and temporal resolutions unrivalled by satellite alternatives [39]. Because of its spatial-temporal advantage, the UAS-based approach can provide greater return time by providing high performance rates for many flights throughout the day and monitoring processes at a very high spatial and temporal resolution [40]. UAS technology is now mainstream and cost-effective and is being utilized for a broad variety of environmental applications, for example, estimating evapotranspiration, or assessing water stress for sustainable agriculture and precision agriculture [41–45]. For monitoring/quantifying of SOC in agricultural or arable lands, UAS borne imagery has received attention from some researchers [32,38,46], but most of these studies were focused on fields with rather high SOC concentration. This implies that a field low in SOC has not yet been explored. Although UAS has numerous advantages, it is also prone to problems such as restricted payload, short flight endurance, and difficulties in maintaining flight speed and stability during heavy winds and turbulence [47]. However, in terms of new technological advancement, most of the technical problems of the UAS could be solved by collaboration involving environmental experts and UAS engineers [48].

Nevertheless, it is worth mentioning that the Association for Unmanned Vehicle Systems International (AUVSI) has estimated that in the coming years, about 80% of UASs will be used for agricultural activities [43]. UAS sales in Germany, for example, approached 400,000 units in 2017 and were projected to grow to over a million by the end of 2020. Moreover, UAS sales doubled in the US in the same year, with an increase of 117 percent compared to the previous year [49]. Finally, as reported by Kriehn [45], in 2019, there were 900,000 registered UAS drones in the United States, with about 17 percent being used for agriculture.

Clearly, the use of UAS is increasing rapidly, which calls for further studies to assess and enhance its prediction capability for soil properties, especially SOC. Although there have been some studies on both Sentinel-2 and UAS imagery for exploring SOC content, the focus has mainly been on fields with high SOC content. This study aims to focus on a field that is poor in SOC and, importantly, to verify the effect of spectral indices from UAS data (which is rarely used by researchers), as remote sensing data are vulnerable to many disturbing external environmental parameters. To the best of our knowledge, no studies have evaluated the capability of UAS for the agriculture fields with a low amount of SOC when coupled with spectral indices. Therefore, this study's primary objective is to evaluate and compare UAS monitoring capabilities and estimation of SOC with those obtained from spaceborne (Sentinel-2) and proximal soil sensing (field spectroscopy measurements) on an agricultural field low in SOC content as well as verifying the effect of soil and vegetation indices. The spatial SOC distribution map will also be computed for the various sensors used in reference to the laboratory SOC measured values.

2. Materials and Methods

2.1. Study Area

The site used for this study is a 22 ha agricultural land situated at Nová Ves nad Popelkou (50.310° N, 15.240° E), in the Liberec Region (Figure 1), in the Czech Republic. The region has consistent mean windspeeds of 6 km/h, a humidity of 74% with an average altitude of 185 m a.s.l. The region is predominantly agricultural and is dedicated to winter and spring cereals and is dominated by dissected relief with side valleys and toe-slopes. Local prevalent soils types are mainly Cambisols and Stagnosols on crystalline and sedimentary rocks according to the World Reference Base for Soil Resources (IUSS Working Group WRB, 2014).

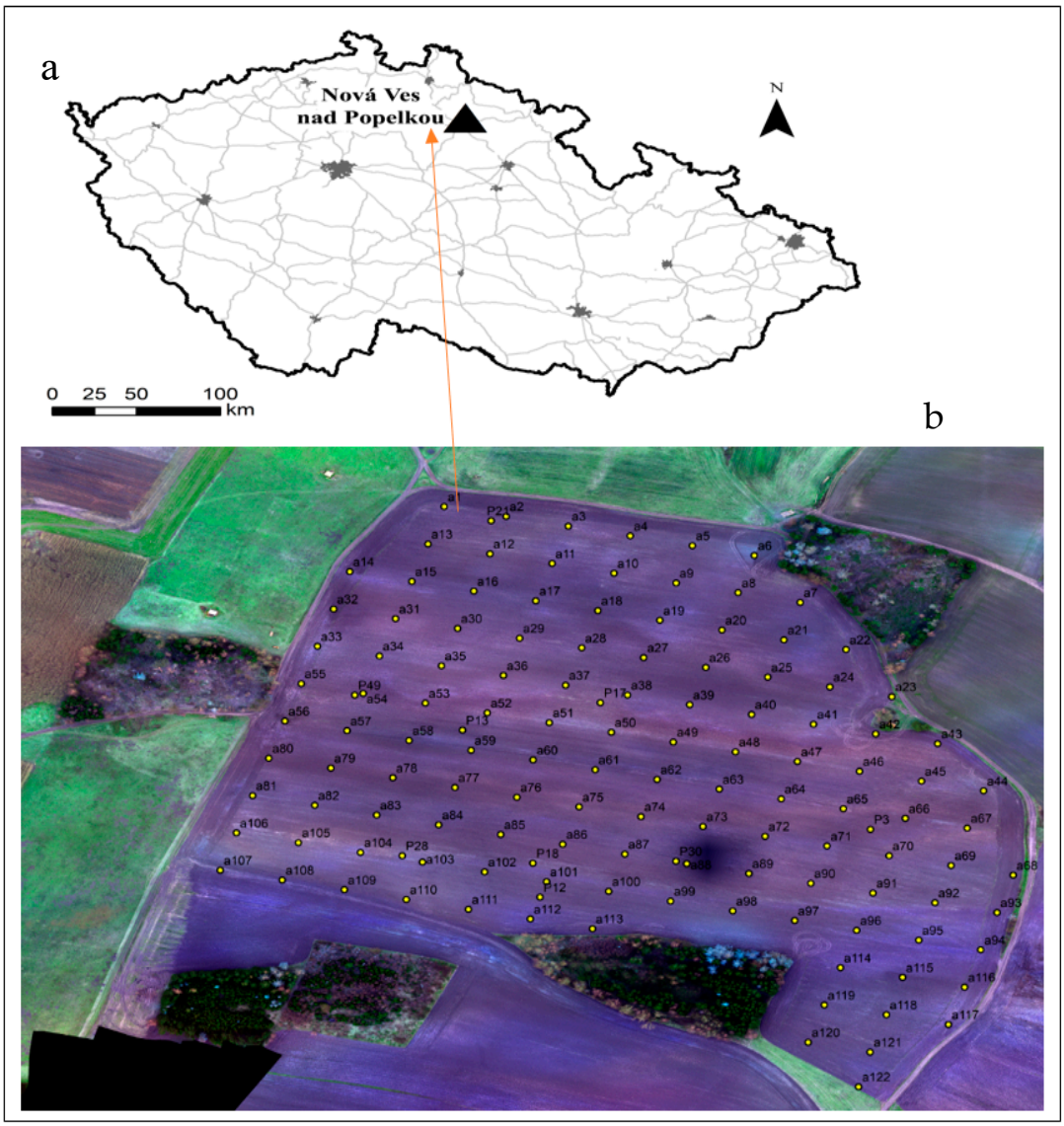


Figure 1. Position of the sampling area in the Czech Republic (**a**) and position of the Nová Ves nad Popelkou sampling points (**b**).

2.2. Soil Sampling and Field Spectral Measurement

A sampling grid method comprising 130 sampling points spread across the whole field was used, as shown in Figure 1. Prior to the actual site survey, those sampling points (130) were generated and identified in the field employing GeoXMM. (Trimble Inc., Sunnyvale,

CA, USA) receiver with an accuracy of 1 m. The field spectra were measured instantly in the field on 6 May 2019 using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, CO, USA) across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050–2500 nm. Measurements from four different positions around each of the 130 sampling points were taken, and the average value was used as the field spectral dataset. The spectroradiometer was calibrated before the first scan and after every six measurements, using a white Spectralon TM (Lab-sphere, North Sutton, NH, USA) [50]. Soil samples were also collected from each of those positions (depth 0–20 cm) while the field measurement was underway. Composite samples (approximately 150 to 200 g of soil) were placed into well-labelled bags and transported to the laboratory for further analysis. These samples were then air-dried, gently crushed, and sieved ($\leq$2 mm) and SOC was measured as total oxidized carbon using wet oxidation approach [51]. This process utilized the dichromate redox titration approach and was accomplished in two different sub-steps [52]. That is, the samples were first oxidized with $K_2Cr_2O_7$ and the solution was then potentiometrically titrated with ferrous ammonium sulphate.

2.3. Remote Sensing Imagery

The remote sensing data used were the Sentinel-2 and UAS imagery at different resolution. Table 1 provides an overview of their individual missions' characteristics.

Table 1. Key radiometric features of multi-spectral sensors shown in this analysis.

Features of the Sensor	**Sentinel-2 [53]**	**Trinity F90 Fixed-Wing Drone**
Mission	Spaceborne	UAS
Sensor type	Push-broom	MicaSense Altum dual sensor
Spectral bands	13	9
Used spectral bands	10	6
Spectral range	9 vis-NIR 3 SWIR	9VNIR
FWHM (nm)	20–200	
SNR (typical)	129 (444) nm 154 (497) nm 168 (560) nm 142 (664) nm 117(704) nm 89 (740) nm 10 (783) nm 174 (843) nm 72 (865) nm 114 (943) nm 50 (1377) nm 100 (1613) nm 100 (2200) nm	32 (475) nm 14 (531) nm 27 (560) nm 16 (650) nm 14 (668) nm 10 (705) nm 12 (717) nm 57 (842) nm thermal infrared 8–14 um
GSD (spatial resolution)	10/20/60 m	Variable (8.8 cm)
Positional accuracy	12 m	3 m
Acquisition date	10 June 2019	25 November 2019

UAS: unmanned aircraft system; vis-NIR: visible and near-infrared; FWHM: full width at half maximum; SNR: signal-to-noise ratio (Wavelength mentioned); SWIR: short-wave infrared; GSD: ground sampling distance.

2.3.1. UAS Multispectral Imagery

Multispectral data were acquired using a Trinity F90 fixed-wing drone with a MicaSense Altum dual sensor mounted onboard with two cameras (RGB + Multispectral). The MicaSense Altum dual sensor captures images in six independent spectral bands

(multispectral) with the last band being a thermal infrared sensor (Blue 475 nm (B4), Green 560 nm (B5), Red 668 nm (B6), Red edge 717 nm (B7), Near-infrared 840 nm (B8), and Thermal 11 µm (B9)). The RGB sensor also captures images in three bands (red-green-blue) (400–700 nm). This is a high-resolution digital camera that is separated from the multispectral sensor. This implies that the total bands captured by the Trinity F90 were nine. The location of the on-board Global Navigation Satellite System (GNSS) and Inertial Navigation Unit has been saved in the metadata files using the Exchangeable Image File Format (EXIF). The camera is equipped with a sun sensor that gathers information about the light conditions and saves the radiant flux data produced in the EXIF format. The image was acquired on 25 November 2019 at Nová Ves nad Popelkou in a clear sky condition. The flight plan was prepared using a QBase 3D mobile app (mission planning software), this served as the primary interface between the user and the UAS device. QBase 3D offers real-time information, such as altitude, distance, battery life about the UAS, and mission telemetry data that provide the operator with updated information about the flight at all times. The flight height was 190 m and the spatial resolution was 8.8 cm, covering an area of 31 ha. We also ensured that we had sufficient batteries for the total flight duration over the entire study field. The images were captured automatically, and the calculated position was consistent with 85% front and 75% side overlap. The images were accurately oriented, 3D model was extracted, the digital elevation model (DEM) was calculated based on the generated cloud point (during the flying period), and orthorectified images were calculated and then exported as one mosaic in GeoTIFF file in EPGS 4326—Geographic coordinates on WGS-84 ellipsoid. Before generating this orthophoto, calibration is performed. The obtained image (before calibration) is already in the reflectance format, however, the actual reflectance values are obtained by dividing each band by 32,768 to get the values normalized in the interval between 0 and 1. The 32,768 is the band center value which represents 100 percent of reflectance. For geometrical correction, the ground-based points and the Differential Global Positioning System (DGPS) were used while for both radiation correction and transformation of reflectance, the Gray Scale Correction method was utilized. AgiSoft Metashape Professional 1.5.0 (AgiSoftLLC, St. Petersburg, Russia), photogrammetric processing was used. The software's consistent performance in photogrammetric processing has been demonstrated in previous studies [54]. In order to differentiate bare soil areas, the Normalized Difference Vegetation Index (NDVI) was employed to mask a threshold of 0.2. The R software (R Development Core Team, Vienna, Austria) was used for all other data processing. For this study, it was only the multispectral section (Trinity F90) with six bands that was used for further analysis.

2.3.2. Sentinel-2 Imagery

The extracted cloud-free Sentinel-2B imagery used for this study was carried out at the European Space Agency's Copernicus Open Access Hub on 10 June 2019. The Sentinel-2 mission consists of two similar satellites: Sentinel-2A, and Sentinel-2B, respectively. Each satellite has a Multi-Spectral Instrument (MSI) that generates images of the earth. The Sentinel-2 images are processed to Level-1C, which implies that they have been ortho-corrected, map-projected images containing top-of-the-air reflectance data. This image will need further pre-processing by the user, but the level 2A Sentinel-2 imagery can be used instantly because its dataset has been processed by the suppliers using Sen2Cor processor. These processes include geometric, radiometric, and atmospheric corrections. For this study the level 2A Sentinel-2 imagery was used. The Sentinel-2 image consists of 13 spectral bands. These spectral bands range from the visible and near infrared (vis-NIR) to the short-wave infrared (SWIR). They include four bands at 10 m resolution ((B2, 490 nm), (B3, 560 nm), (B4, 665 nm), (B8, 842 nm)); six bands at 20 m resolution ((B5, 705 nm), (B6, 740 nm), (B7, 775 nm), and (B8A, 865 nm); 2 SWIR large bands, (B11, 1610 nm) and (B12, 2190 nm). Finally, three bands at 60 m resolution ((B1, 443 nm), (B9, 940 nm), and (B10, 1380 nm)). Before downloading, all the 13 band were resampled to 10 nm using the SNAP software (by pixel resolution). With the exception of B1, B9, and B10 that were omitted, all

the remaining bands were used for this study. The Sentinel-2 user handbook [55] describes the whole protocol.

Soil optical properties can be influenced by certain factors such as soil water content, mineral composition, and organic matter content [37]. Therefore, nine calculated spectral indices have been applied to both Sentinel-2 and UAS datasets as independent variables, anticipated to enhance the prediction capability of the datasets. The added spectral indices were Colour Index (CI), Normalized Differences Vegetation Index (NDVI), Infrared Percentage Vegetation Index (IPVI), Normalized Difference Red Edge (NDRE), Soil Adjusted Vegetation Index (SAVI), Vegetation (V), Normalized Difference Vegetation Index (GNDVI), Difference Vegetation Index (DVI), and Brightness Index (BI). The equations used to determine these indices are shown in Table 2. SNAP was used to obtain bare soil pixel values at sampling locations.

Table 2. Indices derived from Sentinel-2 and UAS spectra.

Index	Definition Based on Sentinel-2	Definition Based on UAS	References
CI	$\frac{B4-B3}{B4+B3}$	$\frac{B6-B5}{B6+B5}$	Pouget et al. [56]
NDVI	$\frac{B8-B4}{B8+B4}$	$\frac{B8-B6}{B8+B6}$	Rouse et al. [57]
IPVI	$\frac{1}{2}(NDVI+1)$	$\frac{1}{2}(NDVI+1)$	Crippen [58]
NDRE	$\frac{B8-B5}{B8+B5}$	$\frac{B8-B7}{B8+B7}$	Barnes et al. [59]
SAVI	$\frac{(B8-B4)*(1+L)}{B8-B4+L}$ L = 0.5	$\frac{(B8-B6)*(1+L)}{B8-B6+L}$ L = 0.5	Huete [60]
GNDVI	$\frac{B8-B3}{B8+B3}$	$\frac{B8-B5}{B8+B5}$	Gitelson et al. [61]
DVI	B8-B4	B8-B6	Richardson and Wiegand [62]
BI	$\frac{\sqrt{(B4*B4)+(B3*B3)}}{2}$	$\frac{\sqrt{(B6*B6)+(B5*B5)}}{2}$	Escadafal [63]
V	$\frac{B8}{B4}$	$\frac{B8}{B6}$	Jordan [64]

Sentinel-2 (B3: Green, B4: Red, B5: Red Edge, B8: NIR); UAS (B5: Green, B6: Red, B7: Red Edge, B8: NIR).

2.4. Data Pre-Processing Approaches

The initial field spectral range was 350–2500 nm; however, the noisy segments of 350–399 nm were removed prior to spectra treatment, retaining only 400–2500 nm range. The field spectra and the other two dataset (UAS and Sentinel-2) were then subjected to the following set of pre-processing techniques: discrete wavelet transformation (DWT), standard normal variate (SNV), logarithmic transformation (log(1/R)), as well as the combination of DWT with SNV (DWT + SNV) and with log(1/R) (DWT + log(1/R)). The DWT is a known technique for signal smoothing and/or noise reduction. This function was determined using the wavelet package in the R software [65]. Also, all other pre-treatment algorithms have been computed using the R software.

2.5. Modelling and Prediction Assessment

The spectra obtained from Sentinel-2 and UAS sensors, including the determined spectral indices, were each linked to the SOC determined in the laboratory using collected soil samples from the field. For the field data, the spectral measurement in the field was used. The above-mentioned datasets were used to build SOC predictive models. The spatial resolution for the UAS remains the same (8.8 cm). Two separate multivariate models were evaluated for all spectral data, namely random forest (RF) and support vector machine regression (SVMR). SVMR is a nonlinear algorithm used for regression and classification processes with a set of related supervised learning algorithms, which has an excellent ability to be a universal predictor of any multivariate function to any defined degree of accuracy. Even if the discriminant feature gathered is based on minimal data, the independent test set's prediction error can still be small. RF is also a technique for classification and regression. RF belongs to the ensemble machine learning algorithm family that predicts a soil parameter response from a set of predictors that could be a training data matrix. This

is done by creating and aggregating multiple decision trees. RF also adjusts splitting by picking the best split from a randomly chosen subset of predictors [66]. In multivariate regression models, spectral reflectance data were used as predictor variables and selected soil parameters data as output responses. The model output was evaluated for each regression procedure by five-fold cross-validation of the training set (75%) and the testing set of 25% of the samples using SVMR and RF modelling techniques. The prediction accuracy was evaluated by index of determination (R^2_{CV}), the ratio of performance to interquartile range (RPIQ), the root mean square error of prediction (RMSEPcv) of the 5-folds cross-validation, and the ratio of performance to deviation (RPD). The RPD was determined as an auxiliary indicator of model reliability as the ratio of the RMSEPcv to the standard data deviation. The larger the RPD, the better the model for prediction. Prior to evaluating the predictive models, the normality of the distribution of the SOC contents was examined (skewness <1).

A correlation matrix was also calculated to visualize the relationships between the three datasets and their parameters (indices) with SOC (examine which dataset is more correlated or significantly correlated). For the remote sensing data set (UAS and Sentinel-2), this was done between SOC and their bands and indices. However, for the field spectra, the correlation was made with SOC using only selected wavelengths (based on UAS and Sentinel-2 wavelengths) due to the enormous amount of spectral data available (350–2500 nm).

For a visual comparison of SOC spatial distribution predicted by models based on different data and laboratory measurement, SOC maps were created using the inverse distance weighting (IDW) interpolation method.

3. Results

3.1. Soil Organic Carbon (SOC) Frequency Histogram and Descriptive Statistics

Figure 2 is a frequency histogram and a statistics summary of SOC characteristics in soil samples within the study area comprising standard deviation (SD), coefficient of variation (CV), minimum, maximum, mean value, skewness, and standard error (SE). The statistical distribution of the SOC within the sample site was positively skewed. A visual inspection of the SOC histogram showed that the value distribution (tail region) has shifted to the left side. Generally, the overall result signifies a low to medium SOC content of the area.

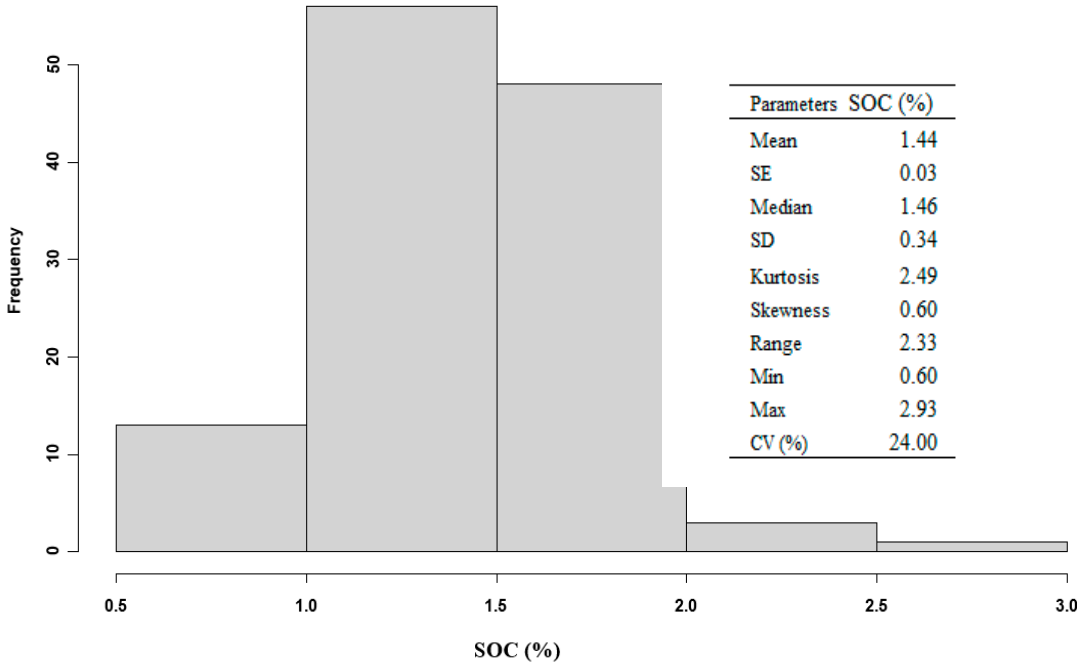


Figure 2. Frequency histogram and statistics summary of soil organic carbon (SOC) content.

3.2. Correlation of SOC with Reflectance Bands and Spectral Indices for Sentinel-2 and UAS Datasets

To visualize the differences between UAS, Sentinel-2 imagery datasets and the calculated indices (for each dataset) within the study area, correlation matrices between their parameters (bands and indices) and SOC were built (Figure 3). The correlation matrices helped to determine among the datasets strong or significant correlations (in both positive and negative directions) so as to identify which spectral bands or indices are the key determinants in the prediction of SOC. For the UAS dataset, the most significant correlations were found between SOC and CI, band 7 and band 6, followed by NDVI, NDRE, IPVI, and BI. For the Sentinel-2 spectral bands, it was only CI that provided significant correlation with SOC. Although, neither dataset was strongly correlated with SOC, there were strong correlations between some of the bands and indices.

3.3. Correlation between SOC and Selected Wavelength of Field Spectra

Figure 4 displays the correlation matrix of SOC with selected wavelengths of field spectra. These wavelengths were selected using the wavelength values that were similar or closer to that of UAS and Sentinel-2 bands. Considering all selected wavelengths, the strongest significant correlations between SOC and field spectra were obtained from 443, 665, 668, 705, and 717 nm while the remaining wavelengths showed good correlations. Although strong correlation was seen between all selected wavelengths (among each other), there were no strong correlation witnessed between SOC and the selected wavelengths.

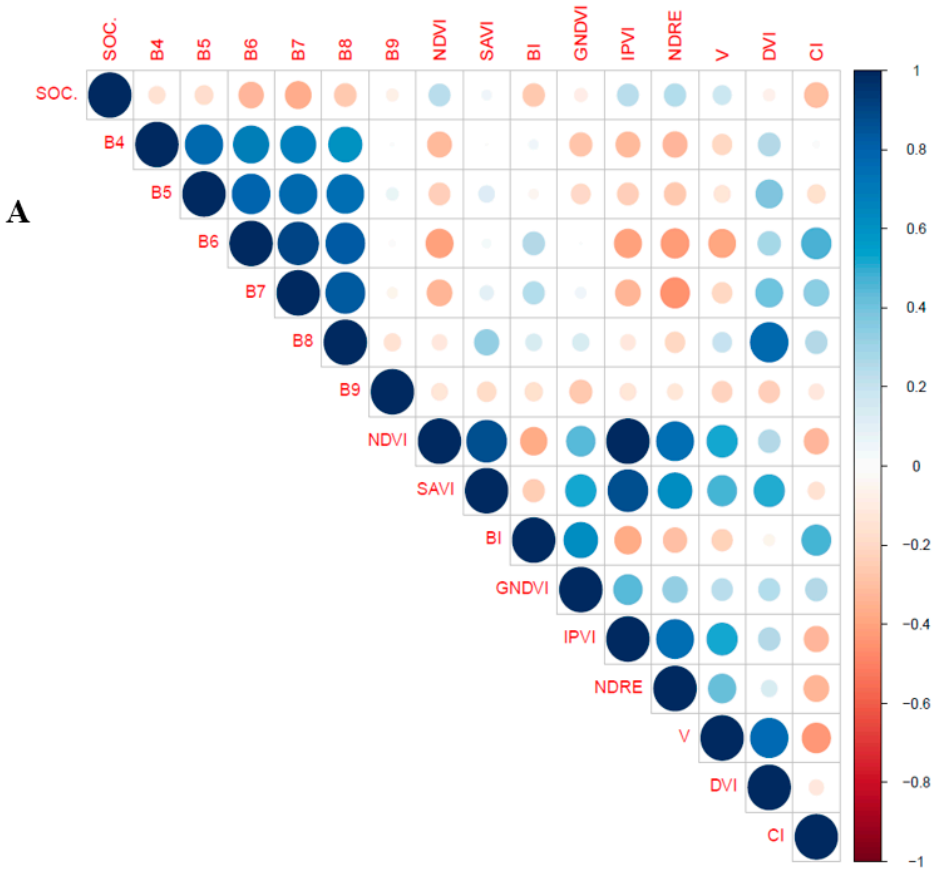


Figure 3. *Cont.*

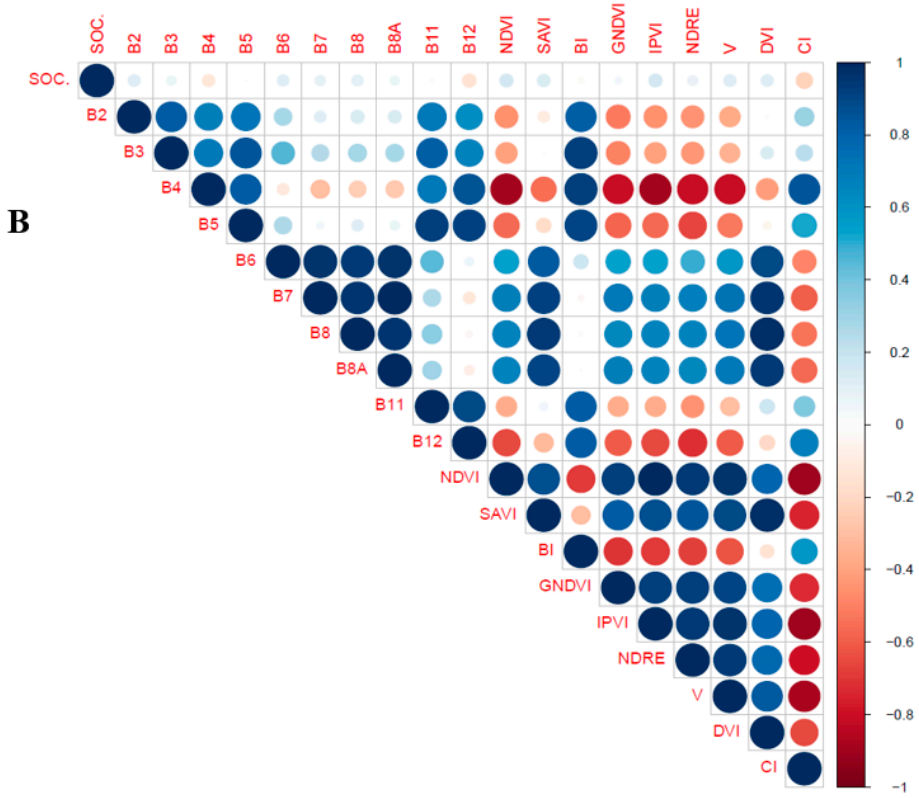


Figure 3. Correlation matrices of SOC with UAS reflectance bands and calculated spectral indices (**A**), and Sentinel-2 reflectance bands and some calculated spectral indices (**B**) at the study locations. Positive correlations are shown in blue while red color represents negative correlations. Color intensity and the size of the circle are proportional to the correlation coefficient values. On the right side of the correlation matrices, the legend color indicates the correlation coefficients and the corresponding colors.

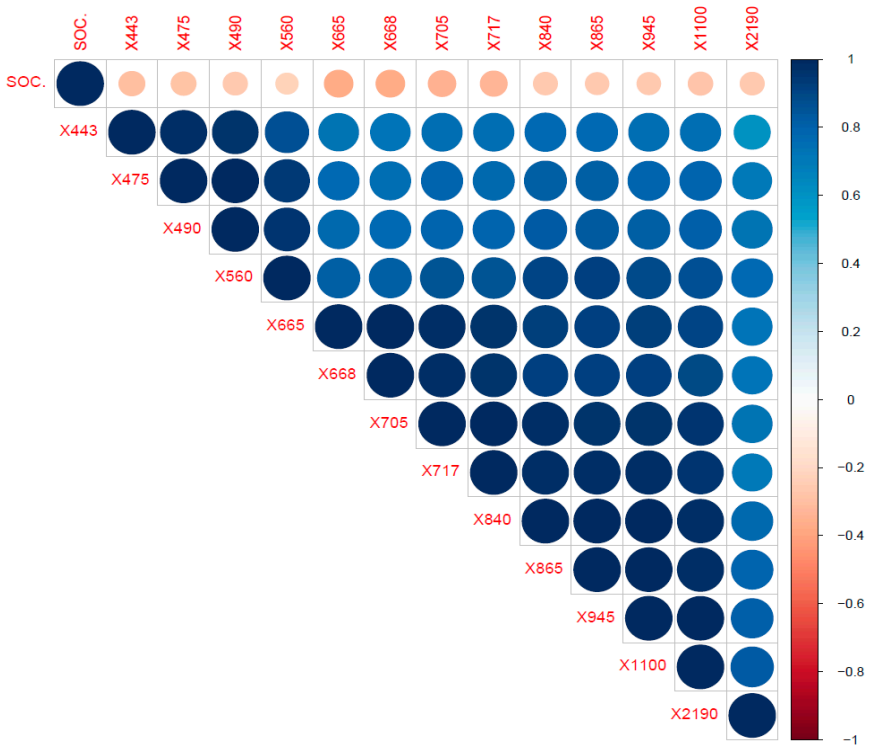


Figure 4. Correlation matrix between SOC and some selected wavelength length (based on both UAS and Sentinel-2 bands) of field spectral measurements.

3.4. Prediction of SOC Using UAS, Sentinel-2 and Field Spectra Data Sets

The prediction results (Table 3) showed that the highest prediction accuracy of SOC (RPD = 1.4; R^2_{CV} = 0.48; RPIQ = 1.65; $RMSEP_{CV}$ = 0.24, for log(1/R)) were obtained with field spectral data using SVMR algorithm. This was followed by UAS (DWT, RF) with RPD = 1.13, R^2_{CV} = 0.27, RPIQ = 1.36, and $RMSEP_{CV}$ = 0.30, and finally Sentinel-2 (SNV), SVMR) with RPD = 1.08, R^2_{CV} = 0.24, RPIQ = 1.31, and $RMSEP_{CV}$ = 0.31, respectively. Moreover, other improved result was also obtained using SNV + DWT, log + DWT, and SNV (SVMR) methods with the field spectra and log(1/R) (RF) with UAS.

Table 3. Statistics of the fivefold leave-group-out cross-validation for field spectra, UAS and Sentinel-2 using random forest (RF) and support vector machine regression (SVMR) with different pre-processing methods.

	UAS				**Sentinel-2**				**Field Spectra**			
					RF							
Treatment	**RPD**	**R^2_{cv}**	**RPIQ**	**RMSEPcv**	**RPD**	**R^2_{cv}**	**RPIQ**	**RMSEPcv**	**RPD**	**R^2_{cv}**	**RPIQ**	**RMSEPcv**
Raw	1.05	0.11	1.24	0.32	1.04	0.04	1.22	0.32	0.99	0.05	1.14	0.34
DWT	1.13	0.27	1.36	0.29	1.02	0.02	1.2	0.33	1.02	0.08	1.18	0.33
SNV	1.00	0.14	1.19	0.34	0.96	0.15	1.27	0.35	1.12	0.22	1.31	0.3
SNV + DWT	1.01	0.03	1.22	0.33	1.02	0.01	1.23	0.33	1.08	0.18	1.28	0.31
Log(1/R)	1.04	0.22	1.31	0.32	1.02	0.05	1.29	0.33	1.01	0.06	1.14	0.33
Log + DWT	1.10	0.17	1.27	0.3	1.01	0.01	1.18	0.33	1.02	0.05	1.14	0.33
					SVMR							
Raw	1.01	0.11	1.18	0.33	1.07	0.15	1.27	0.31	1.21	0.36	1.44	0.28
DWT	1.05	0.14	1.16	0.32	1.03	0.11	1.13	0.33	1.23	0.35	1.44	0.27
SNV	1.06	0.22	1.25	0.31	1.08	0.24	1.31	0.33	1.31	0.44	1.53	0.26
SNV + DWT	1.04	0.12	1.12	0.32	1.05	0.11	1.14	0.32	1.35	0.45	1.59	0.25
Log(1/R)	1.09	0.19	1.29	0.31	1.08	0.16	1.28	0.31	1.4	0.48	1.65	0.24
Log + DWT	1.07	0.11	1.15	0.31	1.06	0.12	1.12	0.32	1.33	0.45	1.56	0.25

4. Discussion

Spectroscopy under proximal soil sensing has now become a common way of estimating SOC and other soil parameters because of its high accuracy level compared to the other forms of measurement stated above [7]. In comparison with the other two data sets, namely UAS and Sentinel-2, the field spectra under proximal soil sensing show the best prediction output as expected (RPD = 1.4; R^2_{CV} = 0.48; RPIQ = 1.65 and $RSMEP_{CV}$ = 0.24, for log(1/R), SVMR). Although the RPD and R^2_{CV} value for this field is not so high, it is comparable to other research findings [67]. Nonetheless, Stevens et al. [25] demonstrated in one of their studies the efficiency of field measurements in comparison to airborne spectroscopy to predict SOC. However, under field measurement, spectroscopy is prone to external environmental conditions, primarily soil moisture, while under laboratory conditions its final output can be influenced by issues such as spectrometer instability, illumination source, detector output, and sample preparation.

Considering the field spectra correlation with SOC using the selected wavelengths based on both UAS and Sentinel-2 bands, it reveals that most of the wavelengths were significantly correlated with SOC compared to the other two datasets. Likewise, almost all selected wavelengths of the field spectra were strongly correlated with each other. This might have accounted for the improved performance of field measurements using vis-NIR spectroscopy approach. However, field spectroscopy inability to cover large spatial areas is one of its major disadvantages. This is because the costs and work and time demands associated with field and laboratory evaluation makes it difficult to undertake soil properties assessment on a vast scale area [68].

The vast frequent data streams generated by satellite sensors can also ensure that soil monitoring and mapping techniques for larger areas can be accurately, rapidly, and effectively established [29,69]. In this study, the accuracy of SOC predictions using Sentinel-

2 imagery was the lowest compared with the other two datasets, although the differences were rather small compared to that of the UAS (Table 2). One of the probable reasons for its worst performance could be the low correlation of all Sentinel-2 bands and almost all the calculated indices used with SOC. It could also be presumed that our quest to acquire a cloud-free image which shifted the Sentinel-2 image collection date from early May to June, a likely period for some vegetation to emerge on the field, could have affected the accuracy of Sentinel-2 imagery prediction. For instance, Castaldi et al. [21] attributed the weak output of Sentinel-2 imagery data in their study to the probable influence of maize seedlings on some of the Sentinel-2 bands (wavelengths), highly sensitive to vegetation. This is so because their Sentinel-2 image was collected in May when maize seedlings may have been emerging. According to Bartholomew et al. [70], spectral reflectance form can sometimes be affected because of the apparent existence of fresh or dry vegetation (less than 20%) and therefore the predictive accuracy of soil properties could be affected. Satellite data can be desirable due to its wide spatial coverage, fast revisit time, and the ability to acquire data unaffected by local air traffic restrictions, however, as a result of cloudiness or when parched and bald soil conditions are needed, these predetermined revisit times may not be suitable for adequate temporal coverage [32]. Other challenges for satellite applications are the relatively low image resolution, restricted availability of high-quality temporal and spatial images, primarily as a result of adverse atmospheric conditions and sensor requirements [71]. For example, in Brazil, Friedel et al. [72] utilized spectroscopy techniques and spaceborne (Hyperion satellite) imagery to quantify soil obtained from the tropics. They indicated that because of the presence of shadow within the study area, satellite image efficiency was hampered. In addition, Steinberg et al. [73] evaluated the potential of both airborne and spaceborne (simulated EnMap) imaging spectroscopy for SOC and clay prediction. Their finding was that the airborne imagery revealed a small improvement with regard to the accuracy of prediction compared to the spaceborne domain.

UAS may be a cheaper and more realistic replacement to satellites, general aviation aircraft and even ground spectroscopy (thanks to large spatial coverage). UAS light-bearing sensors are now being used effectively to track vegetation in precision agriculture [74,75]. An advantage of UAS consists in the small distance between the UAS sensor and the outermost layer of soil, compared to airborne or satellite sensors, which can lead to a comprehensive retrieval of soil spectra. A further advantage of UAS over both airborne and satellite is its ability to yield accurate surface reflectance, especially in case of the need for high-resolution remote sensing-based data, because of the possibility of UAS to be fitted with an incoming sunlight sensor, whereas both satellite and airborne data may require an atmospheric correction model for reflectance measurement [32,76]. Although the spatial resolution of airborne sensors (using aircraft) is higher and could be an alternative to that of satellite data rather than UAS, the acquisition of multitemporal data in an optimum state is hampered by high operating costs especially in case of a change in environmental conditions during measurement [77]. For instance, Stevens et al. [25] used an aircraft-mounted CASI + SASI sensor (444–2500 nm) to detect the shift in carbon stock on a larger scale survey, one of the main problem encounters being the spectral model calibration which they attributed to the several troubling factors including soil water content and enormous aircraft noise that influenced the final carbon stock estimate. For this study, the prediction accuracy for UAS (Table 3) was slightly better than Sentinel-2 satellite (with RPD = 1.13; R^2_{CV} = 0.27; RPIQ = 1.36 and $RMSEP_{CV}$ = 0.30, against Sentinel-2 with RPD = 1.08; R^2_{CV} = 0.24; RPIQ = 1.31 and $RMSEP_{CV}$ = 0.31). One possible reason could be that some of the UAS bands and indices (CI, Band 7, and Band 6 followed by NDVI, IPVI, NDRE, and BI) showed some level of significant correlation with SOC unlike the Sentinel-2 data (only CI). Furthermore, the UAS image was acquired during a favorable weather condition, which is one of UAS strongest advantages over spaceborne and airborne (using aircraft). According to Gomez et al. [33], some of the reasons that could affect the difference in prediction accuracy of SOC between airborne and spaceborne are the sensor spectral and spatial information quality, the distance between sensors and target, and atmospheric

conditions [33]. The use of UAS has become almost ubiquitous in the last five years owing to a reasonable price of its aircraft and the payload camera (from vis-NIR to thermal and 3D) [78]. Although in this study, UAS acquired imagery have shown some positives over satellite images, to uncover its maximum potential for soil properties estimation especially SOC, some outstanding issues would need to be tackled and probably resolved. However, some researchers have suggested solutions to some of the UAS limitations. The UAS lightweight system, for example, could signify an unstable camera positioning resulting from a discrete spatial resolution on the same flight route between two or more captured images [79]. Nevertheless, according to Hardin et al. [80] and Vericat et al. [81], cases of manual geometric correction can be successfully used to solve the above-mentioned problem. According to Aldana-Jague et al. [38], the amount of data produced by the UAS is huge and this demands a significant portion of processing time. Nevertheless, recent developments in UAS GPS systems, coupled with lofty performance inertial measurement units (IMUs), have helped to minimize these large amounts of processing time by using direct-georeferencing approaches [82]. Before UAS images are merged, geometric correction and ortho-rectification are necessary, which is due to the small swath area and platform instability. However, according to Xiang and Tian [71], this issue could be addressed by already developed methods such as the manual use of georeferencing tools (ground-based GCPs), image matching, as well as the use of automated georeferencing (data navigation along with camera lens distortion model). Moreover, it is proposed that the accuracy of the images gathered from a UAS platform should first be assessed for practical purposes to select the most suitable pre-processing technique [83].

Another common limitation of the UAS is the issue of the vignetting effect, that normally induces a shade along the extreme parts of the acquired image, resulting in a blackening of the boundaries compared to the center of the image taken. Nonetheless, Aldana-Jague et al. [38], minimize this issue by taking several images of a "canvas white" beneath consistent daylight conditions and averaging these images for the number of bands used. Lelong et al. [79] also suggested additional method to help resolve the above-mentioned concern, as well as the issue of the bi-directional reflectance distribution function (BRDF) effects faced by UAS, however, according to Hardin and Jensen [48], to better solve these concerns further experiments are still needed. For Lebourgeois et al. [84], though the problem of vignetting is less likely to have an impact on multispectral and hyperspectral imaging sensors that are custom designed, there will still be some degree of vignetting present irrespective of the action taken to rectify its effect. Though, BRDF could influence UAS images as stated above, according to Aber et al. [85,86], the use of the UAS platform is one of the easiest ways to evaluate BRDF models on other remote sensing systems. The use of UAS in agriculture for aerial imaging is still fairly new and may need some bit of patience as well as modest expertise. However, this system continues to gain considerable popularity among environmental scientists, despite all UAS drawbacks as outlined by Hardin and Hardin [47], such as instability, short flight times, distortion within captured imagery, and payload limitation [87]. In a study by Moran et al. [88], they stated that issues related to UAS imaging are also similar to conventional aerial and satellite image applications, e.g., instrument calibration, atmospheric correction, vignetting correction, band-to-band registration, and frame mosaicing. However, with UAS (as stated above) most of these issues can be corrected or adjusted compared to the other ways of measurement. Notwithstanding, as noted already, using UAS comes with numerous benefits including simple to utilize, rapid and accurate set-up at low costs, versatile while flying, and the ability to capture images with very fine resolution. The future for UAS looks promising and it could be used to replace the spaceborne or supplement that of proximal soil sensing if the suggestions by Hardin and Jensen [48] are fully carried out, that is, most of the technical challenges faced by UAS could be overcome by a broad cooperation among environmental experts and UAS engineers during the development of new UAS devices.

One area that is worth mentioning is the issue related to data transfer between spectroscopy and RS especially for large scale site estimation/monitoring of soil properties,

especially SOC. As a result of limited studies [89,90] to date in such a significant area, more work is still needed in order to help limit the uncertainty of atmospheric correction, which according to Castaldi et al. [90] does affect spectral responses from remote sensing data. UAS could have been the best option in terms of all remote sensing approaches due to its small distance between its sensor and the soil. However, it was difficult to locate studies that have tested the feasibility of that approach. A study into such areas in the foreseeable future is highly needed, especially data transfer between spectroscopy and UAS for large scale estimation of soil properties.

SOC predictive performance can be highly variable when data are collected using different procedures, sampling techniques, sample preparation prior to its analyses, instrument requirements, and analytical approaches and algorithms. This is because spectroscopic models can be seriously affected by the properties described above [91,92]. Moreover, this is not exceptional to SOC measurement with spaceborne, UAS and field spectroscopy where the method of data collection does differ as was the case in this study. Choosing the most comprehensive pre-processing strategy may assist in achieving a more accurate prediction models [31], however, linking this with the most appropriate modelling method can positively or negatively affect prediction accuracy. For the modelling and pre-treatment algorithms used in this study, it can be noted that though the field spectra performance increases using SVMR with log transformation (Table 3), for RF, its prediction accuracy experiences a decrease even compared to the UAS dataset. This confirms the need for the use of more techniques for better comparison and to achieve a fair estimation of different form of datasets as noted by Moron and Cozzolino [93] and Mouazen [94] and have been also confirmed by some other studies [10,13]. Finally, the SOC maps derived from the predictive modeling based on the data from the different sensors used are shown in the Figure 5. This was done to view the distribution of SOC within the study field in reference to the laboratory measured values. This demonstrates that all the sensors imagery could predict both low and high SOC values. The field spectra yield a map more similar to the reference map compared to the models using the other two datasets. Sentinel-2 imagery showed a better similarity than UAS imagery to the reference map possibly due to SWIR bands in Sentinel-2, however, map based on UAV imagery on the other hand was similar to the reference map where SOC is lower.

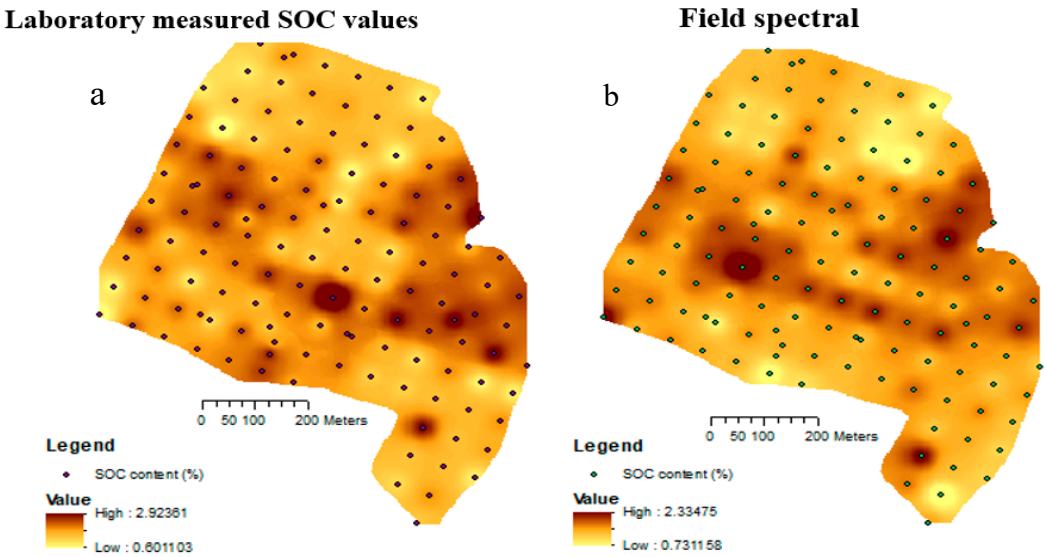


Figure 5. *Cont.*

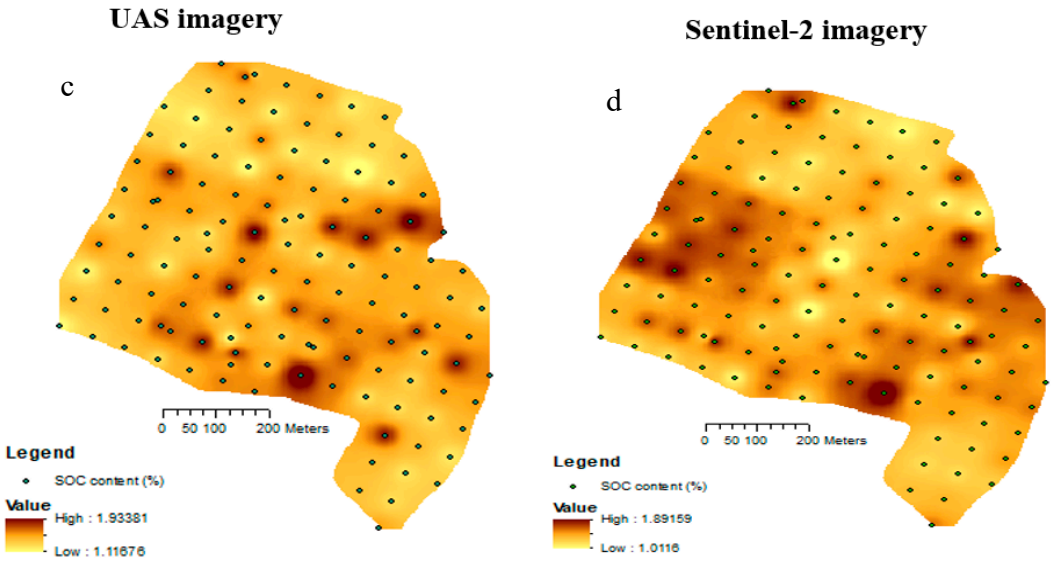


Figure 5. Spatial SOC distribution maps based on prediction using various sensors at Nová Ves nad Popelkou study site as compared to the map based on laboratory SOC measurement: (**a**) reference laboratory conditions, (**b**) field spectral, (**c**) Sentinel-2 multispectral imagery, (**d**) UAS multispectral imagery.

5. Conclusions

This study compared and explored the ability to predict SOC in a field with low SOC content using UAS imagery with spectral indices to that of field spectroscopy and Sentinel-2 datasets. Although for prediction accuracy of SOC, the field spectroscopy was better, the low SOC content within the field makes it difficult to compare the actual performance between UAS and Sentinel-2. However, although the difference was small, the UAS imagery was slightly better than the Sentinel-2 output. This was attributed to the correlation of the spectral indices and bands with SOC. Unlike UAS that had CI, Band 7, and Band 6, followed by NDVI, IPVI, NDRE, and BI that were significantly correlated with SOC, it was only CI for Sentinel-2. It is worth mentioning that all the three datasets show no strong correlation with SOC. However, the spatial distribution map shows that these sensors can detect both high and low SOC values. For comparison especially between UAS and Sentinel-2, the study shows both forms of measurement have their positive features, that is for Sentinel-2 larger spatial coverage and for UAS the reduce distance between the sensor and the soil surface can contribute to a more comprehensive retrieval of soil spectra. Also, they are prone to several limitations especially for Sentinel-2, such as cloud cover and a lot of pre-processing steps, and for UAS they include instability, short flight times, and payload limitation. However, for UAS, most of these issues can be corrected or adjusted compared to other ways of measurement. In conclusion, UAS and Sentinel-2 sensors exploitation for SOC estimation in fields with low SOC need further study, such as using different spectral indices, different machine learning algorithms, and the use of both high and low SOC content fields to determine their actual differences. UAS-based imagery will not substitute the use of manned aircraft or satellite imagery for larger scale assessments but will greatly contribute to local management at small to medium scales. The application of UAS for aerial imagery in agriculture is still relatively new and requires patience and moderate experience. This research has shown that UAS imagery can be exploited more efficiently using spectral indices.

Author Contributions: Conceptualization, J.K.M.B. and M.S.; methodology, J.K.M.B.; software, R.V. and J.K.M.B.; validation, J.K.M.B.; formal analysis, J.K.M.B. and M.S.; investigation, J.K.M.B.; resources, J.K.M.B., J.H., P.C.A. and A.K.; data curation, J.K.M.B., J.H., A.K. and J.A.C.; writing—original draft preparation, J.K.M.B.; writing—review and editing, J.K.M.B., L.B. and M.S.; visualization,

J.K.M.B.; supervision, L.B.; project administration, L.B.; funding acquisition, J.K.M.B. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Czech University of Life Sciences Prague, grant number 21130/1312/3131, and by the Czech Science Foundation, grant number 17-27726S.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the need to respect the rights of the land owners and land managers.

Acknowledgments: The authors are grateful to Miroslav Fér and Karel Němeček for their supporting effort during the field sample collection and to Cervenka Jakub for the UAS imagery processing. We will also like to acknowledge the NutRisk grant (European Regional Development Fund, project Center for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products), number CZ.02.1.01/0.0/0.0/16_019/0000845.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sanchez, P.A.; Ahamed, S.; Carré, F.; Hartemink, A.E.; Hempel, J.; Huising, J.; Lagacherie, P.; McBratney, A.B.; McKenzie, N.J.; de Lourdes Mendonça-Santos, M.; et al. Digital soil map of the world. *Science* **2009**, *325*, 680–681. [CrossRef] [PubMed]
2. Guo, Z.; Han, J.; Li, J.; Xu, Y.; Wang, X. Effects of long-term fertilization on soil organic carbon mineralization and microbial community structure. *PLoS ONE* **2019**, *14*, e0211163. [CrossRef]
3. Vasques, G.M.; Grunwald, S.; Harris, W.G. Spectroscopic models of soil organic carbon in Florida, USA. *J. Environ. Qual.* **2020**, *39*, 923–934. [CrossRef] [PubMed]
4. Ben-Dor, E. Quantitative remote sensing of soil properties. *Adv. Agron.* **2002**, *75*, 173–243.
5. Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [CrossRef]
6. Viscarra Rossel, R.A.; Taylor, H.J.; McBratney, A.B. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. *Eur. J. Soil Sci.* **2007**, *58*, 343–353. [CrossRef]
7. Viscarra Rossel, R.A.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Proximal soil sensing: An effective approach for soil measurements in space and time. *Adv. Agron.* **2011**, *113*, 243–291. [CrossRef]
8. Elachi, C.; Van Zyl, J.J. *Introduction to the Physics and Techniques of Remote Sensing*; John Wiley & Sons: Hoboken, NJ, USA, 2006; Volume 28.
9. Reeves, J.B.; McCarty, G.W.; Meisinger, J.J. Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. *J. Near Infrared Spectrosc.* **2000**, *8*, 161–170. [CrossRef]
10. Vašát, R.; Kodešová, R.; Klement, A.; Borůvka, L. Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. *Geoderma* **2017**, *298*, 46–53. [CrossRef]
11. Gomez, C.; Rossel, R.A.V.; McBratney, A.B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma* **2008**, *146*, 403–411. [CrossRef]
12. Kühnel, A.; Bogner, C. In-situ prediction of soil organic carbon by vis–NIR spectroscopy: An efficient use of limited field data. *Eur. J. Soil Sci.* **2017**, *68*, 689–702. [CrossRef]
13. Biney, J.K.M.; Borůvka, L.; Chapman Agyeman, P.; Němeček, K.; Klement, A. Comparison of Field and Laboratory Wet Soil Spectra in the Vis-NIR Range for Soil Organic Carbon Prediction in the Absence of Laboratory Dry Measurements. *Remote Sens.* **2020**, *12*, 3082. [CrossRef]
14. Bowers, S.A.; Hanks, A.J. Reflection of radiant energy from soil. *Soil Sci.* **1965**, *100*, 130–138. [CrossRef]
15. Shepherd, K.D.; Walsh, M.G. Development of reflectance spectral libraries for characterization of soil properties. *Soil Sci. Soc. Am. J.* **2002**, *66*, 988–998. [CrossRef]
16. Brown, D.J.; Shepherd, K.D.; Walsh, M.G.; Mays, M.D.; Reinsch, T.G. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290. [CrossRef]
17. Wetterlind, J.; Stenberg, B.; Söderström, M. The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale. *Precis. Agric.* **2008**, *9*, 57–69. [CrossRef]
18. Clark, R.N.; King, T.V.; Klejwa, M.; Swayze, G.A.; Vergo, N. High spectral resolution reflectance spectroscopy of minerals. *J. Geophys. Res. Solid Earth* **1990**, *95*, 12653–12680. [CrossRef]
19. Bellon-Maurel, V.; McBratney, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils–Critical review and research perspectives. *Soil Biol. Biochem.* **2011**, *43*, 1398–1410. [CrossRef]

20. Mulder, V.L.; De Bruin, S.; Schaepman, M.E.; Mayr, T.R. The use of remote sensing in soil and terrain mapping—A review. *Geoderma* **2011**, *162*, 1–19. [CrossRef]
21. Castaldi, F.; Hueni, A.; Chabrillat, S.; Ward, K.; Buttafuoco, G.; Bomans, B.; Vreys, K.; Brell, M.; van Wesemael, B. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 267–282. [CrossRef]
22. O'rourke, S.M.; Holden, N.M. Optical sensing and chemometric analysis of soil organic carbon–a cost effective alternative to conventional laboratory methods? *Soil Use Manag.* **2011**, *27*, 143–155. [CrossRef]
23. Yokoya, N.; Chan, J.C.W.; Segl, K. Potential of resolution-enhanced hyperspectral data for mineral mapping using simulated EnMAP and Sentinel-2 images. *Remote Sens.* **2016**, *8*, 172. [CrossRef]
24. Matese, A.; Toscano, P.; Di Gennaro, S.F.; Genesio, L.; Vaccari, F.P.; Primicerio, J.; Belli, C.; Zaldei, A.; Bianconi, R.; Gioli, B. Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture. *Remote Sens.* **2015**, *7*, 2971–2990. [CrossRef]
25. Stevens, A.; van Wesemael, B.; Bartholomeus, H.; Rosillon, D.; Tychon, B.; Ben-Dor, E. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma* **2008**, *144*, 395–404. [CrossRef]
26. Frazier, B.E.; Cheng, Y. Remote sensing of soils in the eastern Palouse region with Landsat Thematic Mapper. *Remote Sens. Environ.* **1989**, *28*, 317–325. [CrossRef]
27. Whitehead, K.; Hugenholtz, C.H. Remote sensing of the environment with small unmanned aircraft systems (UASs), part 1: A review of progress and challenges. *J. Unmanned Veh. Syst.* **2014**, *2*, 69–85. [CrossRef]
28. Held, A.; Ticehurst, C.; Lymburner, L.; Williams, N. High resolution mapping of tropical mangrove ecosystems using hyperspectral and radar remote sensing. *Int. J. Remote Sens.* **2003**, *24*, 2739–2759. [CrossRef]
29. Berger, M.; Moreno, J.; Johannessen, J.A.; Levelt, P.F.; Hanssen, R.F. ESA's sentinel missions in support of Earth system science. *Remote Sens. Environ.* **2012**, *120*, 84–90. [CrossRef]
30. Immitzer, M.; Vuolo, F.; Atzberger, C. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
31. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Borůvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sens. Environ.* **2018**, *218*, 89–103. [CrossRef]
32. Crucil, G.; Castaldi, F.; Aldana-Jague, E.; van Wesemael, B.; Macdonald, A.; Van Oost, K. Assessing the performance of UAS-compatible multispectral and hyperspectral sensors for soil organic carbon prediction. *Sustainability* **2019**, *11*, 1889. [CrossRef]
33. Gomez, C.; Adeline, K.; Bacha, S.; Driessen, B.; Gorretta, N.; Lagacherie, P.; Roger, J.M.; Briottet, X. Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. *Remote Sens. Environ.* **2018**, *204*, 18–30. [CrossRef]
34. Jin, X.; Du, J.; Liu, H.; Wang, Z.; Song, K. Remote estimation of soil organic matter content in the Sanjiang Plain, Northest China: The optimal band algorithm versus the GRA-ANN model. *Agric. For. Meteorol.* **2016**, *218*, 250–260. [CrossRef]
35. Guo, L.; An, N.; Wang, K. Reconciling the discrepancy in ground-and satellite-observed trends in the spring phenology of winter wheat in China from 1993 to 2008. *J. Geophys. Res. Atmos.* **2016**, *121*, 1027–1042. [CrossRef]
36. Liu, S.; An, N.; Yang, J.; Dong, S.; Wang, C.; Yin, Y. Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China. *Catena* **2015**, *133*, 137–144. [CrossRef]
37. Jin, X.; Song, K.; Du, J.; Liu, H.; Wen, Z. Comparison of different satellite bands and vegetation indices for estimation of soil organic matter based on simulated spectral configuration. *Agric. For. Meteorol.* **2017**, *244*, 57–71. [CrossRef]
38. Aldana-Jague, E.; Heckrath, G.; Macdonald, A.; van Wesemael, B.; Van Oost, K. UAS-based soil carbon mapping using VIS-NIR (480–1000 nm) multi-spectral imaging: Potential and limitations. *Geoderma* **2016**, *275*, 55–66. [CrossRef]
39. Manfreda, S.; McCabe, M.F.; Miller, P.E.; Lucas, R.; Pajuelo Madrigal, V.; Mallinis, G.; Ben Dor, E.; Helman, D.; Estes, L.; Ciraolo, G.; et al. On the use of unmanned aerial systems for environmental monitoring. *Remote Sens.* **2018**, *10*, 641. [CrossRef]
40. Hunt, E.R., Jr.; Daughtry, C.S. What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture? *Int. J. Remote Sens.* **2018**, *39*, 5345–5376. [CrossRef]
41. Gago, J.; Douthe, C.; Coopman, R.; Gallego, P.; Ribas-Carbo, M.; Flexas, J.; Escalona, J.; Medrano, H. UAVs challenge to assess water stress for sustainable agriculture. *Agric. Water Manag.* **2015**, *153*, 9–19. [CrossRef]
42. Hoffmann, H.; Nieto, H.; Jensen, R.; Guzinski, R.; Zarco-Tejada, P.; Friborg, T. Estimating evaporation with thermal UAV data and two-source energy balance models. *Hydrol. Earth Syst. Sci.* **2016**, *20*, 697–713. [CrossRef]
43. Radoglou-Grammatikis, P.; Sarigiannidis, P.; Lagkas, T.; Moscholios, I. A compilation of UAV applications for precision agriculture. *Comput. Netw.* **2020**, *172*, 107148. [CrossRef]
44. Niu, H.; Wang, D.; Chen, Y. Estimating actual crop evapotranspiration using deep stochastic configuration networks model and UAV-based crop coefficients in a pomegranate orchard. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping V*; International Society for Optics and Photonics: Washington, DC, USA, 2020; Volume 11414, p. 114140C. [CrossRef]
45. Kriehn, D. Current and Future Applications of Unmanned Aircraft Systems in Precision Agriculture. 2020. Available online: http://www.wcngg.com/2020/10/21/current-and-future-applications-of-unmanned-aircraft-systems-in-precision-agriculture/ (accessed on 21 October 2020).

46. Žížala, D.; Minařík, R.; Zádorová, T. Soil organic carbon mapping using multispectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions. *Remote Sens.* **2019**, *11*, 2947. [CrossRef]
47. Hardin, P.J.; Hardin, T.J. Small-scale remotely piloted vehicles in environmental research. *Geogr. Compass* **2010**, *4*, 1297–1311. [CrossRef]
48. Hardin, P.J.; Jensen, R.R. Small-scale unmanned aerial vehicles in environmental remote sensing: Challenges and opportunities. *Gisci. Remote Sens.* **2011**, *48*, 99–111. [CrossRef]
49. Sylvester, G. (Ed.) *E-agriculture in Action: Drones for Agriculture*; Food and Agriculture Organization of the United Nations and International Telecommunication Union: Bangkok, Thailand, 2018.
50. Shi, T.; Wang, J.; Chen, Y.; Wu, G. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 95–103. [CrossRef]
51. ISO. *Soil Quality–Determination of Organic Carbon by Sulfochromic Oxidation*; International Standard 1998.14235; Croatian Standards Institute: Zagreb, Croatia, 1998.
52. Skjemstad, J.O.; Baldock, J.A. Total and organic carbon. In *Soil Sampling and Methods of Analysis*, 2nd ed.; Carter, M.R., Gregorich, E.G., Eds.; CRC Press: Boca Raton, FL, USA, 2008; pp. 225–238.
53. ESA Radiometric—Resolutions—Sentinel-2 MSI—User Guides—Sentinel Online. Available online: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric (accessed on 10 November 2019).
54. Verhoeven, G. Taking computer vision aloft–archaeological three-dimensional reconstructions from aerial photographs with photoscan. *Archaeol. Prospect.* **2011**, *18*, 67–73. [CrossRef]
55. ESA. *Sentinel-2 User Handbook*; Revision 2; ESA Standard Document: Paris, France, 2015; Volume 64.
56. Pouget, M.; Madeira, J.; Le Floch, E.; Kamal, S. *Caracteristiques Spectrales des Surfaces Sableuses de la Region Cotiere Nord-Ouest de l'Egypte: Application aux Donnees Satellitaires SPOT. 4–6/12/1990*; ORSTOM, Collection Colloques et Seminaires: Paris, France, 1990.
57. Rouse, J.W.; Haas, J.R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systemsin the Great Plains witherts. In Proceedings of the 3rd ERTS Symposium, Washington, DC, USA, 1 May 1974.
58. Crippen, R.E. Calculating the vegetation index faster. *Remote Sens. Environ.* **1990**, *34*, 71–73. [CrossRef]
59. Barnes, E.M.; Clarke, T.R.; Richards, S.E.; Colaizzi, P.D.; Haberland, J.; Kostrzewski, M.; Waller, P.; Choi, C.; Riley, E.; Thompson, T.; et al. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA, 16–19 July 2000; Volume 1619.
60. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
61. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
62. Richardson, A.J.; Wiegand, C.L. Distinguishing vegetation from soil background information. *Photogramm. Eng. Remote Sens.* **1977**, *43*, 1541–1552.
63. Escadafal, R. Remote sensing of arid soil surface colour with Landsat thematic mapper. *Adv. Space Res.* **1989**, *9*, 159–163. [CrossRef]
64. Jordan, C.F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]
65. Aldrich, E. A Package of Functions for Computing Wavelet Filters, Wavelet Transforms and Multiresolution Analyses. 2013. Available online: http://cran.rproject.org/web/packages/wavelets/wavelets.pdf (accessed on 21 September 2012).
66. Liaw, A.; Wiener, M. Classification and regression by random forest. *R News* **2002**, *2*, 18–22.
67. Vaudour, E.; Gilliot, J.M.; Bel, L.; Lefevre, J.; Chehdi, K. Regional prediction of soil organic carbon content over temperate croplands using visible near-infrared airborne hyperspectral imagery and synchronous field spectra. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *49*, 24–38. [CrossRef]
68. Conant, R.T.; Ogle, S.M.; Paul, E.A.; Paustian, K. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. *Front. Ecol. Environ.* **2011**, *9*, 169–173. [CrossRef]
69. Malenovský, Z.; Rott, H.; Cihlar, J.; Schaepman, M.E.; García-Santos, G.; Fernandes, R.; Berger, M. Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sens. Environ.* **2012**, *120*, 91–101. [CrossRef]
70. Bartholomeus, H.; Kooistra, L.; Stevens, A.; van Leeuwen, M.; van Wesemael, B.; Ben-Dor, E.; Tychon, B. Soil organic carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 81–88. [CrossRef]
71. Xiang, H.; Tian, L. Development of a low-cost agricultural remote sensing system based on an autonomous unmanned aerial vehicle (UAV). *Biosyst. Eng.* **2011**, *108*, 174–190. [CrossRef]
72. Friedel, M.J.; Buscema, M.; Vicente, L.E.; Iwashita, F.; Koga-Vicente, A. Mapping fractional landscape soils and vegetation components from Hyperion satellite imagery using an unsupervised machine-learning workflow. *Int. J. Digit. Earth* **2017**, *11*, 670–690. [CrossRef]
73. Steinberg, A.; Chabrillat, S.; Stevens, A.; Segl, K.; Foerster, S. Prediction of common surface soil properties based on Vis-NIR airborne and simulated EnMAP imaging spectroscopy data: Prediction accuracy and influence of spatial resolution. *Remote Sens.* **2016**, *8*, 613. [CrossRef]
74. Baluja, J.; Diago, M.P.; Balda, P.; Zorer, R.; Meggio, F.; Morales, F.; Tardaguila, J. Assessment of vineyard water status variability by thermal and multispectral imagery using an unmanned aerial vehicle (UAV). *Irrig. Sci.* **2012**, *30*, 511–522. [CrossRef]

75. Primicerio, J.; Di Gennaro, S.F.; Fiorillo, E.; Genesio, L.; Lugato, E.; Matese, A.; Vaccari, F.P. A flexible unmanned aerial vehicle for precision agriculture. *Precis. Agric.* **2012**, *13*, 517–523. [CrossRef]
76. Soriano-Disla, J.M.; Janik, L.J.; Allen, D.J.; McLaughlin, M.J. Evaluation of the performance of portable visible-infrared instruments for the prediction of soil properties. *Biosyst. Eng.* **2017**, *161*, 24–36. [CrossRef]
77. Itten, K.I.; Dell'Endice, F.; Hueni, A.; Kneubühler, M.; Schläpfer, D.; Odermatt, D.; Seidel, F.; Huber, S.; Schopfer, J.; Kellenberger, T.; et al. APEX-the hyperspectral ESA airborne prism experiment. *Sensors* **2008**, *8*, 6235–6259. [CrossRef]
78. Xue, J.; Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *J. Sens.* **2017**, *2017*, 1353691. [CrossRef]
79. Lelong, C.C.; Burger, P.; Jubelin, G.; Roux, B.; Labbé; S; Baret, F. Assessment of unmanned aerial vehicles imagery for quantitative monitoring of wheat crop in small plots. *Sensors* **2008**, *8*, 3557–3585. [CrossRef]
80. Hardin, P.J.; Jackson, M.W.; Anderson, V.J.; Johnson, R. Detecting squarrose knapweed (Centaurea virgata Lam. Ss squarrosa Gugl.) using a remotely piloted vehicle: A Utah case study. *Gisci. Remote Sens.* **2007**, *44*, 203–219. [CrossRef]
81. Vericat, D.; Brasington, J.; Wheaton, J.; Cowie, M. Accuracy assessment of aerial photographs acquired using lighter-than-air blimps: Low-cost tools for mapping river corridors. *River Res. Appl.* **2009**, *25*, 985–1000. [CrossRef]
82. Turner, D.; Lucieer, A.; Wallace, L. Direct georeferencing of ultrahigh-resolution UAV imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 2738–2745. [CrossRef]
83. Berni, J.A.J.; Zarco-Tejada, P.J.; Suárez, L.; González-Dugo, V.; Fereres, E. Remote sensing of vegetation from UAV platforms using lightweight multispectral and thermal imaging sensors. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2009**, *38*, 6.
84. Lebourgeois, V.; Bégué, A.; Labbé, S.; Mallavan, B.; Prévot, L.; Roux, B. Can commercial digital cameras be used as multispectral sensors? A crop monitoring test. *Sensors* **2008**, *8*, 7300–7322. [CrossRef] [PubMed]
85. Aber, J.S.; Aber, S.W.; Buster, L.; Jensen, W.E.; Sleezer, R.O. Challenge of infrared kite aerial photography: A digital update. *Kans. Acad. Sci. Trans.* **2009**, *112*, 31–39.
86. Aber, J.S.; Marzolff, I.; Ries, J.B. *Small-format Aerial Photography: Principles, Techniques and Geoscience Applications*; Elsevier: Amsterdam, The Netherlands, 2010; pp. 72–73.
87. Hardin, P.J.; Lulla, V.; Jensen, R.R.; Jensen, J.R. Small Unmanned Aerial Systems (sUAS) for environmental remote sensing: Challenges and opportunities revisited. *Gisci. Remote Sens.* **2019**, *56*, 309–322. [CrossRef]
88. Moran, M.S.; Inoue, Y.; Barnes, E.M. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sens. Environ.* **1997**, *61*, 319–346. [CrossRef]
89. Nouri, M.; Gomez, C.; Gorretta, N.; Roger, J.M. Clay content mapping from airborne hyperspectral Vis-NIR data by transferring a laboratory regression model. *Geoderma* **2017**, *298*, 54–66. [CrossRef]
90. Castaldi, F.; Chabrillat, S.; Jones, A.; Vreys, K.; Bomans, B.; Van Wesemael, B. Soil organic carbon estimation in croplands by hyperspectral remote APEX data using the LUCAS topsoil database. *Remote Sens.* **2018**, *10*, 153. [CrossRef]
91. Ben-Dor, E.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245*, 112–124. [CrossRef]
92. Gholizadeh, A.; Carmon, N.; Klement, A.; Ben-Dor, E.; Borůvka, L. Agricultural soil spectral response and properties assessment: Effects of measurement protocol and data mining technique. *Remote Sens.* **2017**, *9*, 1078. [CrossRef]
93. Moron, A.; Cozzolino, D. Application of near infrared reflectance spectroscopy for the analysis of organic C, total N and pH in soils of Uruguay. *J. Near Infrared Spectrosc.* **2002**, *10*, 215–221. [CrossRef]
94. Mouazen, A.M.; Maleki, M.R.; De Baerdemaeker, J.; Ramon, H. On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil Tillage Res.* **2007**, *93*, 13–27. [CrossRef]

Science of the Total Environment xxx (xxxx) xxx

Contents lists available at ScienceDirect

Science of the Total Environment

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

Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil

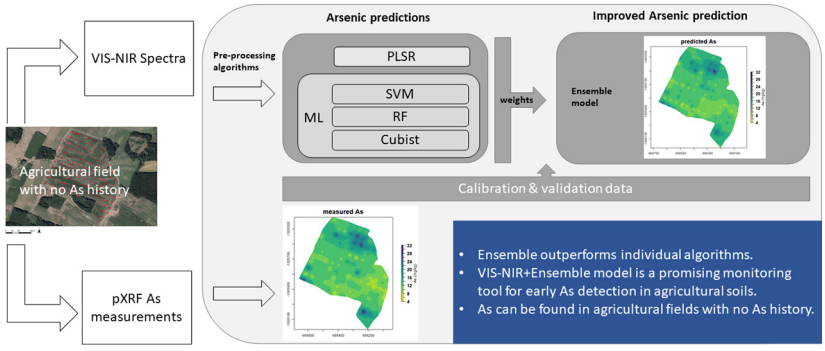
James Kobina Mensah Biney [a,b,*], Radim Vašát [a], Johanna Ruth Blöcher [c], Luboš Borůvka [a], Karel Němeček [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic
[b] The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Department of Landscape Ecology, Lidická 25/27, Brno, 602 00, Czech Republic
[c] Department of Water Resources and Environmental Modeling, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic

HIGHLIGHTS

- Soil arsenic was found to correlated with Fe.
- Ensemble model was better than the individual modelling techniques for *As* estimation.
- Identifying the best pre-treatment algorithms helps improve prediction accuracy.
- Multiple pre-treatments combined are better than using a single treatment algorithm.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 September 2021
Received in revised form 7 November 2021
Accepted 15 November 2021
Available online xxxx

Editor: Filip M.G. Tack

Keywords:
Agricultural soil
Machine learning
Ensemble model
Arsenic
Portable X-ray fluorescence (pXRF)
Field spectroscopy

ABSTRACT

Increasing concentrations of potentially toxic elements (PTE) in agricultural soils remain a major source of public concern. Monitoring PTEs in an agricultural field with no history of contaminants necessitate adequate analysis utilizing a robust model to accurately uncover hidden PTEs. Detecting and mapping the distribution of soil properties using portable X-ray fluorescence (pXRF) and proximal sensing techniques is not only rapid, but also relatively inexpensive. In this study, an ensemble model, consisting of partial least square regression (PLSR), support vector machine (SVM), random forest (RF) and cubist, was used for the prediction and mapping of soil *As* content in an agricultural field with no history of pollution. The datasets were collected using pXRF and field spectroscopy techniques. The main goal was to compare the ensemble model to each of the calibration techniques in terms of prediction accuracy of *As* content in such a field. Other components [e.g., soil organic carbon (SOC), *Mn*, *S*, *soil pH*, *Fe*] that are known to influence *As* levels in the soil were also retrieved to assess their correlation with soil *As*. The models were evaluated using the root mean squared error ($RMSE_{CV}$), the coefficient of determination (R^2_{CV}) and the ratio of performance to interquartile range (RPIQ). In terms of prediction accuracy, the ensemble model outperformed each of the individual techniques (R^2_{CV} = 0.80/0.75) and obtained the least error margin ($RMSE_{CV}$ = 1.91/2.16). Overall, all the predictive techniques were able to detect both low and high estimated values of soil *As* within the study field, but with the ensemble model resembling the measurements better. The ensemble model, a promising tool as demonstrated by the current study, is highly recommended to be included in future studies for more accurate estimation of *As* and other PTEs in other agricultural fields.

* Corresponding author at: Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic.
E-mail address: biney@af.czu.cz (J.K.M. Biney).

https://doi.org/10.1016/j.scitotenv.2021.151805
0048-9697/

1. Introduction

Soil is considered a diverse and complex ecosystem that supports both human food production systems and a wide range of biodiversity. It is also one of the most important natural resources on the planet. However, increasing accumulation of potentially toxic elements (PTEs) in the soil (particularly, agricultural soils) has been a major source of public concern, as plants grown in such soils can take up a large amount of these PTEs (e.g., *Cu*, *Pb*, *Cd*, *Zn*, *As*), which threatens food quality and security and increases potential risk to human health as well as other organisms through the food chain (Järup, 2003; Wu et al., 2015; Yin et al., 2021). For example, *Zn* lowers immunological function and high-density lipoprotein levels, excess *Cu* has been linked to liver damage, and *Pb* and *Cd* are hazardous even at low concentrations. Whiles Pb induces renal tumors, increases blood pressure, *Cd* on the other hand, can induce kidney dysfunctions (FDA, 2001; Ikem and Egiebor, 2005; Llobet et al., 2003).

PTEs can penetrate and remain in agricultural soils naturally, based primarily on the composition of the geological parent materials, as well as other means, but at a level that is considered not toxic to the human body (FDA, 2001; Lu et al., 2012; Singh and Garg, 2006). However, deficiency and toxic effects are observed beyond this normal threshold due to other sources such as wastewater irrigation, solid waste disposal, livestock manure, inorganic fertilizers, agrochemicals, sludge applications, vehicular exhaust, industrial activities, as well as atmospheric deposition and pesticides (Hu and Cheng, 2013; Liu et al., 2020; Panchenko et al., 2018; Plyatsuk et al., 2019; Wu et al., 2020). Therefore, accurate prediction and monitoring of soil PTEs are imperative for effective contingency planning.

Soil arsenic is one of the most harmful contaminants, according to the United States Toxic Commission (Jiang et al., 2015). This is due to the potential health risks it poses to humans, particularly when exposed to a higher concentration of its content in a short period of time. These includes skin cancer, abnormal cell metabolism, diseases of the bladder, liver, lungs, kidneys, and even death (Bennett, 1996). Arsenic can form organic and inorganic compounds when it reacts with other elements, however, the latter are thought to be more toxic than the former. Furthermore, the concentration of *As* in the soil can be influenced by other soil components. Some studies (Cao and Ma, 2004; Horta et al., 2015; Mandal and Suzuki, 2002; Warren et al., 2003) have even suggested that assessing the relationship between these components (e.g., soil organic carbon (SOC), Mn, soil pH, total Fe, temperature, S, clay) and *As* could help explain the concentration of *As* in the soil. The main factors that contribute to soil *As* pollution have been recognized as rock weathering and atmospheric deposition as well as anthropogenic activities (Li et al., 2021; Patel et al., 2005). In agricultural soil, the enrichment of *As* poses a threat to food security as a result of its bioaccumulation in plant (Cui et al., 2018). Additionally, soil *As* has been a serious environmental threat in a number of countries throughout the world affecting millions of people (Smedley and Kinniburgh, 2002). So, highlighting the role of soil safety to human health and sustainable agriculture, identifying areas polluted by toxic metals is critical (Brevik et al., 2018). Therefore, as a result of the growing prevalence of *As* and its potential toxicity, there is an urgent need for a rapid assessment in agricultural soils so that an efficient mitigation framework can be established to minimize or eliminate the potential health risk of *As* and other PTEs.

PTEs in agricultural soil can be identified using simplified, quick, and low-cost technologies such as X-ray fluorescence (XRF) [a recent advancement being the portable XRF (pXRF)], and visible near-infrared (Vis-NIR) as well as several other traditional approaches that are costly and also non- environmentally friendly [e.g., inductively coupled plasma (ICP) or atomic absorption spectrometry (AAS)]. In most cases, pXRF measurements of soil PTEs could be read directly from the analyser, however, the reading may fail due to the analyser's detection limit, especially if the elements concentration is low (O'Rourke et al., 2016). According to Sacristán et al. (2016), the main advantage of Vis-NIR spectroscopy over XRF spectroscopy is that information on multiple soil features can be obtained with a single measurement from the same spectrum. Nevertheless, the absorption bands in the soil Vis-NIR spectra are feeble, comprehensive, and could also overlap. This may cause the information of certain critical spectra needed to improve the predictive performance of soil features to be hidden. Therefore, to accurately identify soil characteristics, a proxy approach must be developed to retrieve this hidden information (Gholizadeh et al., 2016; Rossel and Behrens, 2010). Furthermore, Vis-NIR spectral information may be affected by the conditions under which the soil is scanned, reducing the accuracy of soil property estimation due to unwanted variance. Normally, multiple pre-treatments are utilized, because a single treatment algorithm may not be able to handle all the variations that may exist (artefacts). Also, Vis-NIR spectra combined with chemometrics have been used by several researchers (Chakraborty et al., 2015; Ren et al., 2009; Shi et al., 2016; Wei et al., 2019; Wu et al., 2005, 2007) to predict the presence of *As* and other heavy metals in the soil.

As the need for rapid, accurate, and environmentally friendly methods to estimate PTEs in agricultural soil grows, the results of some comparative studies (e.g., Gholizadeh et al., 2020; Fang et al., 2021; Kebonye et al., 2021; Xie et al., 2021; Zhou et al., 2021) based on the use of multiple machine learning (ML) methods to predict PTE in soil (within a single study) were not consistent; these ML algorithms were evaluated individually. But, different MLs can indicate different sets and important predictors variables when estimating soil properties. The solution to this problem according to Dietterichl (2002) may lie in combining multiple individual prediction techniques into a single model, and thus employing the ensemble learning theory. The hypothesis is that the new model (ensemble model) will perform at least better than each of the separate techniques in terms of appropriately utilizing all the available data (Diks and Vrugt, 2010). Even though the ensemble theory is not new, it has yet to be thoroughly investigated on Vis-NIR spectra to estimate PTEs (specifically *As*) using ML and statistical approaches.

For this study, the focus is primarily on the possibility of improving the predictive performance of spectroscopic models for *As* content because it is a major PTE with related health implications to living organisms in the ecosystem. Four well-known and tested individual techniques were used to develop the ensemble model, namely three ML algorithms [random forest (RF), support vector machine regression (SVM) and cubist] and one statistical method [partial least squares regression (PLSR)]. Although, comprehensive research studies on the prediction and monitoring of PTEs in agricultural soil have already been conducted by a number of researchers, the majority of these studies (Agyeman et al., 2021; Chakraborty et al., 2015, 2017; Ettler et al., 2005; Gholizadeh et al., 2015; Han et al., 2019; Kebonye et al., 2021; Kebonye and Eze, 2019; Kotková et al., 2019; Tremlová et al., 2017; Vaněk et al., 2008; Xie et al., 2021) has focused on agricultural sites or areas close to landfills (where solid waste were dumped) or areas in proximity to industrial plants or industries (e.g., coal, mining) that generate toxic waste. Nevertheless, even normal agricultural practices (e.g., application of pesticides and fertilizers; mineral fertilizers, organic fertilizers; parent material composition; vehicular exhaust; agrochemicals) can generally cause the enrichment of PTEs in agricultural soil (Chen et al., 2008; Nicholson et al., 2003).

Additionally, *As* is also found in the continental crust, primarily in inorganic form (arsenic compounds) associated with igneous and sedimentary rocks. For example, Holub (1997), discovered a high concentration of arsenic in Central Bohemian (Czech Republic) Pluton parent rocks ranging from 20 to 50 mg $\cdot$ kg^{-1}. Although there is no historical record of PTEs in the selected study field, the soil type in the area is Cambisols on sedimentary rocks. This highlights the possibility of PTEs, particularly As, because, according to Li et al. (2021) and Patel et al. (2005), as stated above, weathering of rock is a likely source of soil *As* pollution.

Therefore, the objective of this study was to investigate the feasibility of using an ensemble model with multiple pre-treatment algorithms and coupled with pXRF and field Vis-NIR spectra data for the prediction and mapping of soil *As* content. Specifically, to apply this methodology in an agricultural soil with no historical background of pollution or close to any toxics production sites or industries. The main goal was to compare the results of the individual calibration algorithms (PLSR, SVM, RF, and cubist) with the ensemble model, and to verify through comparison which approach could predict and map soil *As* content more accurately. Additionally, spatial distribution maps of *As* within the study site were assessed for each model using the pXRF-As values and the predicted values using the Inverse Distance Weighting (IDW) Interpolation method.

2. Material and methods

2.1. Study site, soil sampling, and field spectra measurement

The study site is an agricultural field located in Nová Ves nad Popelkou, 22 ha (50°31′ N; 15° 24′ E), Czech Republic with a mean altitude of 185 m above sea level. This area is known for some agricultural and irrigation activities. The chosen area was representative of the soil caps, which were homogeneous and comparable in terms of terrain characteristics, land management, and climatic conditions (Schmidt et al., 2010). The most dominant crops are winter and spring cereals. According to the IUSS Working Group WRB (2014), the soils in this region are predominantly Cambisols on sedimentary rocks.

During soil surveys, using the grid sampling approach (with 40 m spacing) spread across the entire field (Fig. 1), 130 spectra observations were collected in the field using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, CO, USA) in the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050-2500 nm. Additionally, to prevent meteorological conditions constraints, a contact probe device with a 2-cm-diameter circle viewing area and its own light source was used (Waiser et al., 2007). The spectrometer was standardized using a Spectralon® panel (Lab-sphere, North Sutton, New Hampshire, USA) with 99% reflectance preceding the first scan and after every ten measurements (Shi et al., 2016). During the field spectra measurement, soil samples from each of the 130-

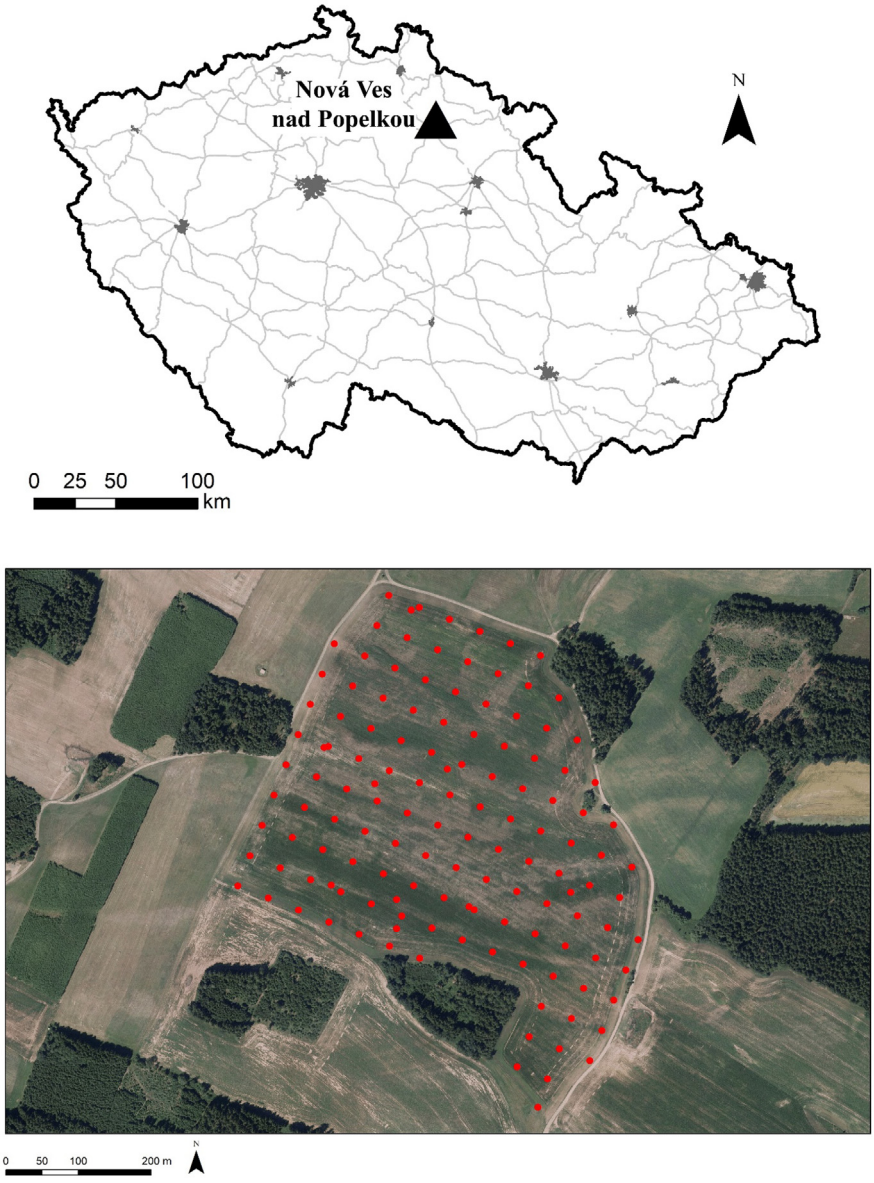


Fig. 1. Location of sampling site at Nová Ves nad Popelkou in Czech Republic.

sampling points were also collected (depth of 15 cm) in well-labelled bags (composite sample of approximately 150 to 200 g) and transported to the laboratory for further analysis. The scanning of each sampling point and the collection of samples was conducted in three replicates with an average of three measurements. In the laboratory, these soil samples were oven dried, ground up and sieved through a <2 mm stainless steel sieve before analysis.

2.2. Laboratory analysis

2.2.1. pXRF measurements

A portable X-ray fluorescence spectrometer was used to analyse *As*, *Mn*, *S*, and *Fe*. Several researchers have also emphasized the effectiveness of using p-XRFs to rapidly obtain elemental measurements of soil (Eze et al., 2016; Paulette et al., 2015; Ravansari et al., 2020). The main component, however, was *As*. The other elements were included (for correlation with *As*) because, they have been shown to influence *As* (directly or indirectly, particularly *Fe*). *Fe* oxides for instance, are well-known as spectrally active soil properties that can strongly adsorb PTEs or have a high affinity for certain PTEs (Axe et al., 2000; Ben-Dor and Banin, 1995). For the elemental measurement, the pXRFs was operated in accordance with the manufacturer's instructions (e.g., calibration, setting the time spent during analysis), to helps reduce inconsistencies and inadequacies in the resulting output. The sample was lightly tapped within the cup for uniformity before scanning. This was done to enhance the sample surface area and the layer through which the X-rays may penetrate adequately. After that, each sample was scanned for 60 s using a pXRF spectrometer (Delta Premium XPD 600, Olympus Innov-X, USA) connected to a computer loaded with the pXRF software (Weindorf et al., 2013, 2016). The “Soils Mode” was used for measurement. Each sample was scanned three times (amounting to 180 s total time), and the average of the three measurements was calculated and used as the final element. For quality control and assurance, the analysis used the quality assurance and control procedure, as well as the standard reference material for a portable device (i.e., XRF 2711a NIST, the National Institute of Standards and Technology). The detection limits for the main elements were *As* <5 mg/kg and <10 mg/kg (Fe).

2.2.2. Soil chemical analyses

Other auxiliary soil properties that can influence *As* due to their high adsorption with PTEs (soil organic carbon (SOC), soil pH) were also determined (correlation test). Measurement of SOC was carried out in two steps using the dichromate redox titration method (Skjemstad and Baldlock, 2008). The samples were initially oxidized with K2Cr2O7 and afterwards, the solution was potentiometrically titrated with ferrous ammonium sulphate (FeH8N2O8S2). The soil pH was measured using a 1:5 (*w*/*v*) ratio of soil and water (pH-H2O) and 1 M potassium chloride (pH-KCl) solution (ISO 10390:1994) using an inoLab Level 1 pH meter (WTW GmbH & Co. KG, Weilheim. Germany).

2.3. Spectra data pre-processing approaches

Spectral data are influenced by a variety of undesirable variations (Ben Dor et al., 2015), and as a result, treatment algorithms are commonly used as a remedy. One of the major benefits of using data preprocessing, particularly for field spectra data, is that it can help minimize or eliminate majority of these unwanted variations (e.g., baseline changes, peak shifts, scattering, noises, missing values, e.t.c) that can occur during samples collection and preparation. These variations may at times obscure the “true” chemically relevant underlying structure, thereby lowering the prediction accuracy of the parameter under consideration (Engel et al., 2013; Rinnan et al., 2009).

The original spectral range was 350–2500 nm; however, the noisy portions between 350 and 399 nm was eliminated, leaving the range of 400–2500 nm before spectral pre-treatment. The raw spectra were transformed into reflectance and were then subjected to the following set of pre-processing techniques. Savitzky-Golay filter (sg) (adjusted for second-order polynomial fit with 31 smoothing points) from signal R package (Signal Developers, 2013). Discrete wavelet transformation (dwt), first-order derivative (d1), second-order derivative (d2) and multiplicative scatter correction (msc) were calculated using pls R package (Mevik and Wehrens, 2007). Standard normal variate (snv), logarithmic transformation (log) as (log(1/R)) and continuum removal (cr) (Vašát et al., 2017) and correction by the maximum reflectance (cmr) from the tripack R package (Renka, 1996). In addition following combinations were also used sg_d1, dwt_d1, sg_d2, dwt_d2, sg_msc, dwt_msc, sg_snv, dwt_snv, snv_msc, sg_snv_msc, dwt_snv_msc, sg_log, dwt_log, log_msc, sg_log_msc, dwt_log_msc, log_snv, sg_log_snv, dwt_log_snv, sg_cr, dwt_cr, sg_cmr, dwt_cmr. This was done to select the most accurate result for each individual modelling techniques. Before the prediction of As, four outliers were removed to improve predictive performance. It is worth mentioning that only the three best pre-treatment results and the raw dataset (no pre-treatment applied) for each calibration technique are shown. For the spectra pre-processing, the R software's (R Core Team and Team, 2015) was used.

2.4. Correlation and spatial distribution maps

A correlation matrix was created to observe the relationships between *As* and the selected auxiliary components (*soil pH*, *SOC*, *Mn*, *S*, *Fe*). This was done to determine which of these components were closely related to *As* content and could potentially influence its estimation. Lastly, the spatial variability of soil *As* contents was mapped using the inverse distance weighting (IDW) interpolation method with the R package gstat (Pebesma, 2004; Gräler et al., 2016). Spatial interpolation techniques are noted to estimate values at an unmeasured location by using point measurements within a given sample space (Qiao et al., 2018). However, selecting an appropriate interpolation strategy for different occurrences and datasets presents numerous challenges (Liao et al., 2018; Qiao et al., 2018). For example, field intricacy, and a large amount of inefficiently collected geographical data can be problematic. The Inverse Distance Weighting (IDW) algorithm is among the most frequently used spatial interpolation techniques in soil science because of its ease of application. IDW uses a linear combination of values to estimate the values of the unknown area within the sampling space and allocates weights using its inverse function. Due to its ability to assign weights before prediction, IDW can have a lower error margin than other interpolation methods which make it more suitable for creating spatial distribution maps more accurately (Liao et al., 2018; Xie et al., 2011).

2.5. Multivariate modelling and models

To make sure the results were not dependent on the multivariate model, four separate multivariate techniques were evaluated individually, namely cubist, SVM, PLSR, RF. The R software (R Core Team and Team, 2015) was used for both modelling and prediction for each of the technique detailed below.

2.5.1. Cubist

The cubist method was used to calibrate the regression tree models using the train function of the caret package in R. The cubist model is a kind of tree structure modelling based on an M5 algorithm (Quinlan, 1992) To avoid overfitting, cubist uses linear regression models at each node instead of the average (Kuhn and Johnson, 2013), For this study, the default number of committees (1, 10 and 20) and neighbours (0, 5, and 9) from the train function were used, giving us a total of nine models. The root means square error (RMSE) was used to select the best models.

2.5.2. Support vector machine (SVM)

Similarly, for this study, the SVM was tuned to different cost parameters using the grid search's built-in turning function (specifically, 0.001, 0.01, 0.1, and 1) in R's package e1071 library, while the epsilon parameter was left at its default value (0.1). Based on the RMSE, the best cost parameter was selected from bootstraps based on a 10-fold cross validation.

2.5.3. Partial least square regression (PLSR)

The PLSR algorithm offers the benefit of removing the issue of multiple collinearities between the independent variables (Næsset et al., 2005). For the current study, the model was made to run and test itself for each number of components, i.e. from 1 to 10 (the maximum number of model components was set to 10). The optimal number of components was selected based on the lowest RMSE. The model was then re-calibrated and validated, the R^2 and the RMSE were computed.

2.5.4. Random forest (RF)

The RF algorithm was specially formulated to reduce experimental noise and improve prediction accuracy (Liaw and Wiener, 2002). The dataset under evaluation was divided randomly into several training sets, and decision trees created to utilize the bootstrap re-sampling capability. The final prediction was calculated using the average of the individual tree outputs. The random forest R package, which includes homonymous (Liaw and Wiener, 2002), was used. A total of 500 trees were used, with 35 variables randomly selected as candidates at each split.

2.6. Construction of the ensemble model

Each of the four best performing predictions obtained from the four techniques (SVM, PLSR, RF, and cubist) was assigned a specific weight using an automated approach that involved testing all possible combinations of predefined weights (routine sequence values from 0.05 to 0.95). The only requirement is that they must add up to one. The weights for each base method were determined by testing multiple possible cases with a precision of two hundredths. In practice, we proceeded by creating all possible permutations with repetition about the size of four elements (the number of predictive algorithms used) from a set of possible numbers n as weights, which was a sequence from 0 to 1 by 0.02. Subsequently only the foursomes that summed up to 1 were used as weights. All those sets of weights were then used one by one for the weighted average of the predictive algorithms and the particular validation statistics was recorded. The set of weights that yielded the lowest RMSE was taken as the most suitable one. In more detail, since an n-fold cross-validation was employed, all four predictive algorithms were run at each data split and weighted averaged. As an input to the ensemble model, these four individual predictions (which had already been cross-validated) were adjusted by the weights of the initial permutation and then added. In addition, for each approach, the ensemble model employed the best treatment algorithm (each technique is selected alongside its best treatment algorithm). The cross-validation statistics were computed by comparing the combined (ensemble) prediction and observed values, using100-repetition runs. The validation statistics were then calculated from these number of runs as a mean value. The same approach had been used to process all permissible permutations, yielding a set of ensemble models with cross-validation statistics. As a consequence, a large number of ensembles learning runs were created based on the number of admissible permutations. The result with the lowest RMSEcv was selected as the most suitable.

2.7. Model and spatial distribution map validation assessment

The model's output was assessed by a five-fold cross-validation for each regression procedure of the calibration (75%) and validation set (25%) of the samples using cubist, SVM, PLSR and RF modelling techniques. The accuracy of the prediction was assessed based on the coefficient of determination (R^2_{CV}), the ratio of performance to interquartile range (RPIQ), and the root mean square error of prediction (RMSEcv) (measures the model overall prediction accuracy) of the 5-folds cross-validation. The bias represents the amount that a model's prediction differs from the target value. The R^2_{CV} ranges from 0 to 1, where $R^2_{CV} = 1$ is the optimal value. The five-fold cross-validation was repeated 100 times to ensure model stability and reliability.

3. Results and discussion

3.1. Measured data

3.1.1. Descriptive statistics of as, and five other soil attributes (SOC, Fe, S, Mn and soil pH)

According to the summary statistics results (Table 1), arsenic (*As*) had the most positively skewed distribution, with a skewness of 1.68, a standard deviation of 4.3 and a significant variability ranging from 6.85 to 29.35 [all in mg/kg]. Additionally, *As* content was found to vary more than any of the parameters measured, presenting a substantial coefficient of variation (CV) value of 32%, which implies that *As* has a high chance of being influenced by external causes such as human activities (Chen et al., 2008). Total *Fe* provided the highest mean of 18,314.60 mg/kg, as well as the second-best skewness value of 1.16 mg/kg. SOC had the lowest mean of 1.44% because its predicted content within the study area is low according to other studies in the area (Biney et al., 2021; Biney et al., 2020; Gholizadeh et al., 2018). The soil pH ranged from moderately acidic (5.56) to slightly alkaline (7.76), with a mean value of 6.39. The CV for pH is not reported because, the CV is restricted to variables measured on scales with an absolute zero: pH has an arbitrary zero. The arbitrary choice of zero affects its CV. Sulphur was the least varied parameter, with a CV of 18 mg/kg and also the lowest skewness value. Finally, Mn had the second highest mean, trailing only Fe, as well as the second most varied element, with a CV of 23%.

Table 1
Statistical summary of soil attributes.

	n[a]	Mean	Median	SD[b]	Skewness	Min[c]	Max[d]	CV[e]%
As (mg/kg)	126.00	13.42	12.75	4.30	1.68	6.85	29.35	32
S	130.00	249.97	249.00	44.31	0.14	148.00	394.50	18
Mn	130.00	1153.82	1139.00	260.68	0.71	622.00	2101.50	23
Fe	130.00	18,314.60	16,997.75	4000.30	1.16	11,653.50	30,086.50	22
SOC (%)	130.00	1.44	1.45	0.33	0.57	0.60	2.93	23
PH	130.00	6.39	6.39	0.28	0.99	5.56	7.76	

[a] n: number of samples.
[b] SD: standard deviation.
[c] Min: minimum.
[d] Max: maximum.
[e] CV: Coefficient of variation.

3.1.2. Influence of other soil attributes on as

Several studies conducted over the last two decades discovered that *As* and other PTEs availability, mobility and distribution in the soil could be influenced by their adsorption relationship to many factors and parameters. These includes, but not limited to the following organic matter concentration, *pH*, *Fe*, *Mn*, temperature, SOC, clay, soil particle size, phosphate and Fe-oxide content in the soil (Cao and Ma, 2004; Hale et al., 1997; Horta et al., 2015; Mandal and Suzuki, 2002; Warren et al., 2003; Xie et al., 2012). In order to test and confirms this hypothesis, the correlation coefficients between *As* and some of these soil properties and metals (*S*, *Mn*, *Fe*, *SOC*, *soil PH*) were calculated (Fig. 2). This was done to help ascertain the primary contributing factors of high *As* prediction found in an agricultural site that is not nearby any industries or landfills site that generate waste substances which may contain some amount of toxic elements.

For the result, total *Fe* ($R = 0.53$) was the most correlated component to *As*. SOC correlated negatively with *As* ($R = -0.34$), the remaining components were only weakly correlated with *As*. According to Fitz and Wenzel (2002), *Fe-oxides* and *Fe-hydroxides* have been identified as some of the primary active elements that influenced soil-*As* retention ability. Furthermore, PTEs have been found to bond and interact with primary spectrally active soil constituents, including SOC and other forms of Fe content in the soil (Song et al., 2012; Wu et al., 2005). In the current study, these two soil properties were the only components with some degree of correlation with soil *As*. Khosravi et al. (2017) showed that Fe can affect the prediction of *As* and some other PTEs. Although, SOC may also play a role in *As* prediction. One possible reason for this is that, if the inorganic constituent level in soils is higher than the organic constituent content, spectral measurements can be significantly influenced. Moreover, the bonding effects of *Fe* and *Al* compounds promote *As* adsorption by soil organic matter (SOM) (Lin et al., 2004). For example, Hartley et al. (2004) used *Fe-oxide* to mitigate *As* concentration in an *As*-contaminated soil because *Fe* is known to adsorb *As*. Even though, *Fe* had a good relationship with *As* for the current study, however, presumably, its content was not enough or the total *Fe* were non-reactive to cause the adsorption of *As* for this study field.

Table 2
Prediction performance showing Statistics of the five-fold leave-group-out cross-validation for soil *As* content using PLSR (partial least square regression), RF (random forest), cubist, SVMR (support vector machine) and the ensemble model (combination of all four models) with several pre-treatment algorithms combination.

Treatment	R^2cv	RMSEcv	RPIQ	BIAS
SVM				
Raw	0.64	2.57	1.41	−0.0674
sg	0.71	2.35	1.83	−0.0317
dwt	0.75	2.16	1.99	0.0613
msc	0.71	2.33	1.84	−0.0422
RF				
Raw	0.30	3.59	0.99	−0.1059
sg_d1	0.57	2.84	1.25	−0.0799
sg_d2	0.64	2.65	1.34	−0.0605
dwt_d2	0.63	2.64	1.34	−0.0861
PLSR				
Raw	0.64	2.58	1.37	0.0061
d1	0.71	2.29	1.55	−0.0468
dwt_d1	0.72	2.28	1.56	−0.0093
dwt_d2	0.71	2.30	1.54	−0.0506
Cubist				
Raw	0.56	2.92	1.21	−0.0203
Log	0.65	2.55	1.39	−0.0115
snv_msc	0.64	2.59	1.37	−0.0115
msc	0.61	2.70	1.32	−0.0337
Ensemble				
Combined treatment dwt+ sg_d2+ dwt_d1+ Log	0.80	1.91	2.11	−0.0111

Ensemble structure (with weights): $As = 0.20As_{PLSR;\ dwt\_d1} + 0.54As_{SVM;\ Dwt} + 0.24As_{RF;\ sg\_d2} + 0.02As_{cubist;\ log}$.

According to Misra and Tiwari (1963), several studies have found out that, soils with high levels of reactive iron parameters absorb more *As* than soils with a comparable texture but low levels of iron.

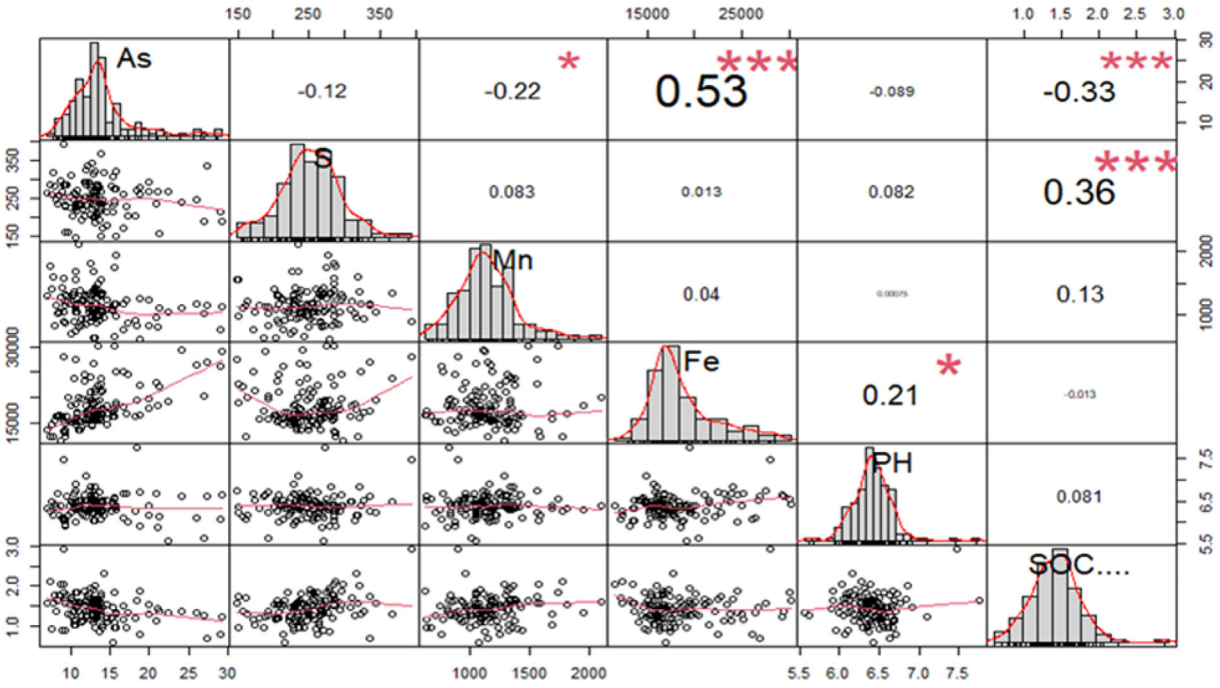


Fig. 2. Correlation matrix showing scatter plots, histograms, and Pearson's correlation coefficients between *As* and the other soil attribute values. *, ** and *** represent the significant correlation at *p*-values of 0.05, 0.01 and 0.001, respectively.

3.2. Evaluation and comparison of models, as well as the impact of pre-treatment algorithms

3.2.1. Predictive performance of soil As content using the stand-alone modelling algorithms

Field spectra in the Vis-NIR range were used to predict soil *As* throughout the study area using an *As* data set measured in the laboratory with the pXRF for each of the 130 samples collected. Though, all the separate calibration techniques demonstrated acceptable prediction of soil *As* content [ranging from $R^2_{CV} = 0.64$ to 0.75, RPIQ = 1.34 to 1.99, and RMSEcv = 2.16 to 2.59], however, in general, SVM_dwt model provided the most appropriate result (Table 2) with the highest R^2_{CV} and lowest error values compared to the other techniques in an order of PLSR_dwt_d1 > cubist_log > RF_sg_d2. According to Stevens et al. (2010), SVM is known for its ability to estimate and improve nonlinear structures in multidimensional domains. Several other studies have also demonstrated the pre-eminence of SVM over other multivariate techniques (Lucà et al., 2017; Xu et al., 2020).

The predictions from the separate multivariate techniques (Table 2) were obtained with distinct pre-treatment algorithms (dwt, sg_d2, dwt_d1, and log). For example, the dwt algorithm is commonly used to filter and normalize spectra data, thus accounting for discrepancies during sample preparation (Viscarra Rossel et al., 2016), smoothing with Sg is often used to eliminate artificial noises within the working spectral range, log is used for the attainment of linearization between the predictors and response variables, while the first derivatives were used in the study by Gholizadeh et al. (2018) to eliminate baseline offset. This is an indication that, the field spectra data were most likely affected by multiple variations (artefacts) and therefore, a single treatment strategy was insufficient to provide a comprehensive rectification of these variations in order to improve the prediction accuracy of *As*. For example, without the introduction of the treatment algorithms, the results obtained for each calibration technique were poor to average [ranging from $R^2_{CV} = 0.3$ to 0.56, with error margins of 2.92 to 3.59], except for SVM and PLSR, which provided good results [$R^2_{CV} = 0.64$, (Table 2, raw data)].

The Scatterplots (Fig. 3) show the results of predicted versus observed for As predictive accuracy using the four individual predictive techniques. Though, some extreme points are seen for each technique particularly for RF_sg_d2 and cubist log_log, these points may not necessary be classified as outliers or if outliers, then they could be positive outliers containing vital information about the data set. For instance, before prediction of *As*, four (4) outliers were removed from the data set (result not shown) using a local outlier factor (LOF) algorithm procedure proposed by Breuniq et al. (2000). This is because, the prediction accuracy for the techniques improved after the outlier's removal except for cubist (which remained the same). The LOF algorithms examine the density of a specific point's neighbours to determine its density, then compare it to the density of other points, and employ a local approach to detect outliers within the neighbourhoods. This is in agreement with Murray (1988), who stated that, outlier removal improves prediction accuracy. However, according to Frost (2019), some of these outliers may also contain vital information and their removal could affect the accuracy of prediction. In terms of prediction accuracy on the test set, it was clear that SVM and PLSR look similar, whereas RF and cubist provided nearly identical outputs in terms of coefficient of determination but differed in terms of RPIQ values. This

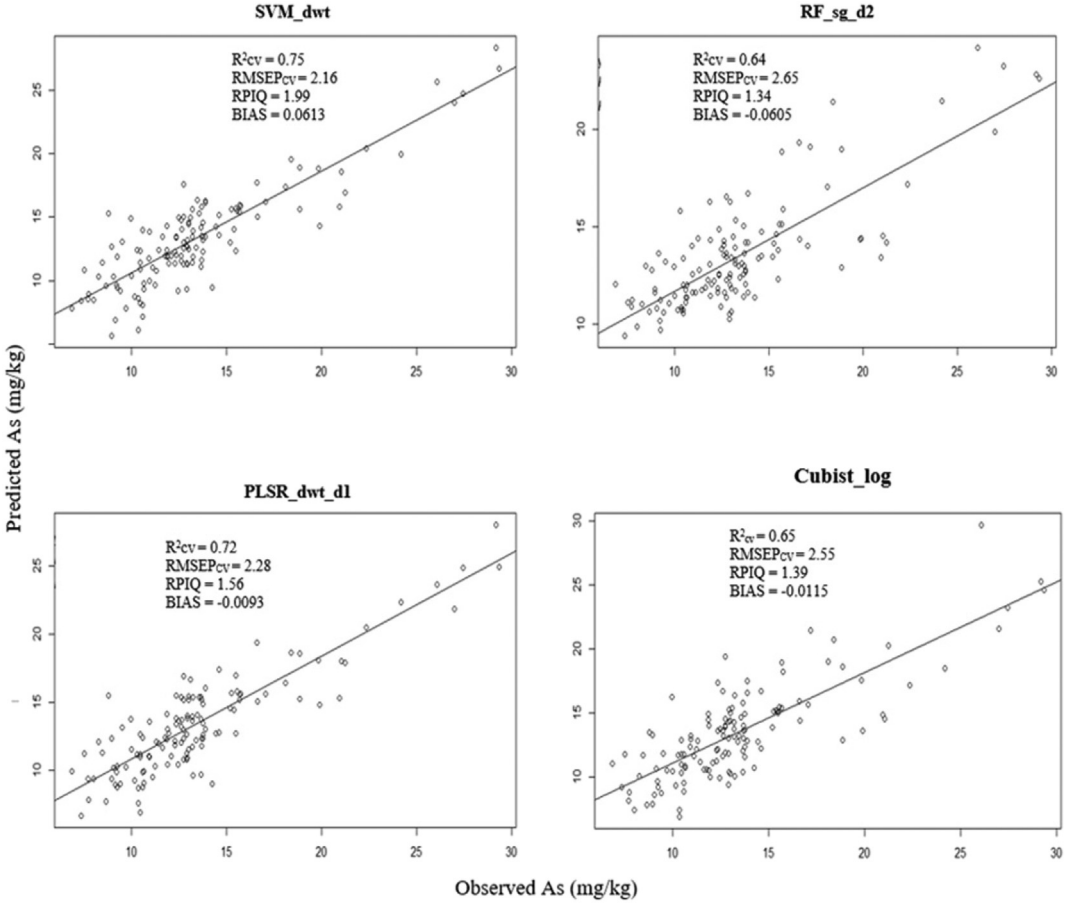


Fig. 3. Observed vs. predicted *As* content (%) values for each of the four calibrations techniques (PLSR, SVM, MLR, and RF) with the best treatment algorithms employed.

was the generated output which was fed into the ensemble order to improve the prediction accuracy of *As* by combining the four modelling techniques predictions.

3.2.2. Assessment of the ensemble model for soil as prediction

Consequently, the ensemble Vis-NIR spectra model (which is combining of several individual predictions made with different calibration techniques into a single final one), outperformed these algorithms in terms of soil *As* prediction accuracy. It provided the lowest RMSEcv (1.91 mg/kg) as well as the highest prediction accuracy [$R^2_{CV} = 0.80$, RPIQ = 2.11, BIAS = − 0.111] compared to the best result obtained by the separate techniques [$R^2_{CV} = 0.75$, RMSEcv =2.16, RPIQ = 1.99, BIAS = 0.0613, SVM]. Moreover, the ensemble model, in particular, obtained RMSEcv improvements of 25–74%, highlighting the potential of the ensemble model for soil *As* prediction. This result (ensemble) was achieved by minimizing the limitations and drawbacks of each technique and maximizing the responses of the combined models (Arabameri et al., 2019; Kalantar et al., 2018; Martre et al., 2015). Furthermore, diverse signal pre-processing strategies were explored, yielding a large number of individual predictions as input for the ensemble model. Notably, the best pre-processing algorithms for each technique that makes up the ensemble model were used, allowing for a complete correction of the variety of artefacts present (Mishra et al., 2020). Additionally, the variability of *As* (Table 1) within the study site could have probably be a contributing factor for the ensemble model's improved performance. According to Liu et al. (2015), the ensemble model may be more effective if the individual predictive techniques vary in terms of the nature of the predictive algorithm and the structure of the provided parameters.

It is also worth noting that, choosing the best pre-treatment for a specific spectra data set can be difficult, especially when using multivariate techniques separately. According to Oliveri et al. (2019), this is due to a lack of clear guidelines for determining when to use a specific pre-processing approach, i.e. whether to use a single or a combination of treatment algorithms. Normally, the user is required to explore all possible alternatives in order to determine which algorithm may be best for the data set under consideration (Mishra et al., 2020). According to the findings in this current study; using an ensemble model as a solution to the aforementioned problem could be a viable option.

3.3. Comparison of soil spatial distribution maps produced by individual techniques and the ensemble model

Both pXRF-laboratory measured and Vis-NIR-predicted *As* content (using the individual techniques and the ensemble model) were used as an example to validate the feasibility of detecting the spatial distribution of soil *As* content. This was done in order to detect areas with higher *As* content as well as the spatial prediction of *As* for each algorithm and the ensemble model. As *Fe* was most strongly correlated, a spatial distribution map of *Fe* was also created using *Fe-pXRF*-measurements, to compare to *As-pXRF*. The results show that, while the hotspots for *Fe* were spread across the entire study field, for *As*, it was concentrated in a few selected sections (specifically, the northern section), but was broad in scope than that of *Fe* (Fig. 4). Finally, the spatial distribution maps of the measured and predicted *As* presented in Fig. 5 show that overall all the predictive techniques were able to detect both low and high estimated values of soil *As* within the study field with the ensemble model showing greater detail. While high concentration levels of *As* were predicted around the northern section, lower levels were predicted around the east-southern section of the study field (Fig. 5). Fig. 6 shows the difference between the interpolated predicted and measured values. Although areas of overprediction and underprediction are similar among the algorithms, the prediction error appears less extreme for the ensemble model. This shows that especially with the ensemble model the spatial variability of soil *As* and its level of concentration could be reasonably identified by interpolating predicted *As* values. This implies that, Vis-NIR data in combination with IDW has the potential to roughly and rapidly reveal spatial patterns of soil *As*.

In summary, we identified and could predict levels of Arsenic in an agricultural field with no history of pollution or close to any landfills or industries that generate toxic waste element. Therefore, this study agrees with Chen et al. (2008) and Nicholson et al. (2003) that, common agricultural practices can influence the cause of PTEs in agricultural soil. Disentangling the origin of As is a question this current study could not answer. However, this study demonstrates that the focus for estimating PTEs should not be limited to agricultural soil or areas near industries

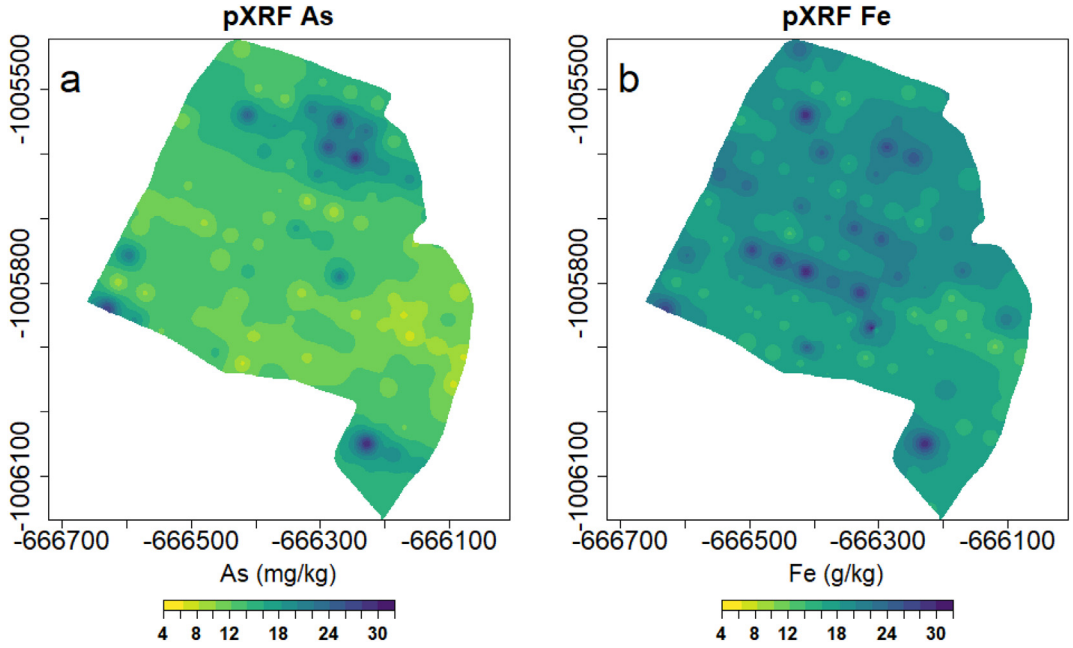


Fig. 4. Spatial *As* and *Fe* distribution maps based on laboratory measurement pXRF values.

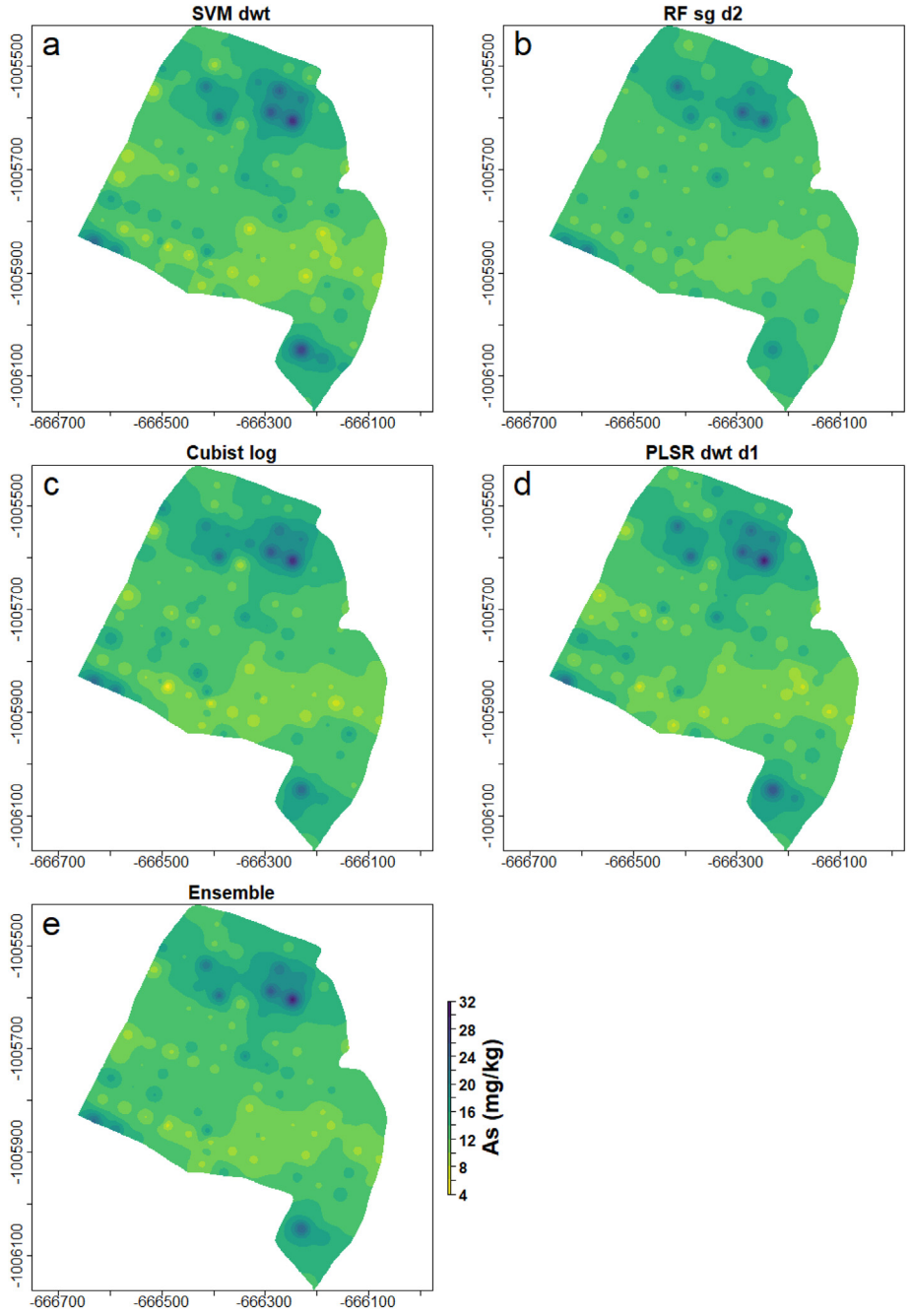


Fig. 5. Spatial *As* distribution maps based on the best prediction outcome from field spectroscopy and pXRF data for individual techniques and the ensemble model [SVM(a), RF (b), PLSR (c), cubist (d), ensemble model (e)].

with toxic waste, but also to agricultural fields with no such records to help discover the early level of PTEs. The developed ensemble models for this study are robust and effective at not only predicting PTEs in agricultural soils but also reduces the error margin.

4. Conclusion

The use of pXRF (lab) and field spectroscopy data coupled with an ensemble model [PLSR, SVM, random forest, and cubist] predicted soil *As* content more accurately than each of calibration techniques used separately. Among the other auxiliary components used (soil PH, Mn, S, Fe, and SOC), known to influence *As* content in the soil, *Fe* was correlated most strongly to *As* than any of the other components. The study also shows that, agricultural field with no historical background of PTEs pollution or in close proximity to any industries that produce and release harmful substances, need more attention from researchers to unearth hidden PTEs at the early stages. This is due to the tendency that, unknowing, the health of human and other organisms could be negatively impacted. Although there were similarities in terms of the spatial distribution map between the individual approaches, the ensemble model better resembled the measured data. Based on our findings, the ensemble model appears very promising because it also provided the lowest error margin [$RMSE_{CV} = 1.91$ as against 2.16], an error improvement of between 24 and 74% among the individual techniques. Therefore, the study recommends its inclusion in studies testing Vis-NIR spectroscopy and pXRF data to estimate and mapped soil *As* more accurately, as well as other PTEs, though further studies are still required, especially using larger data sets.

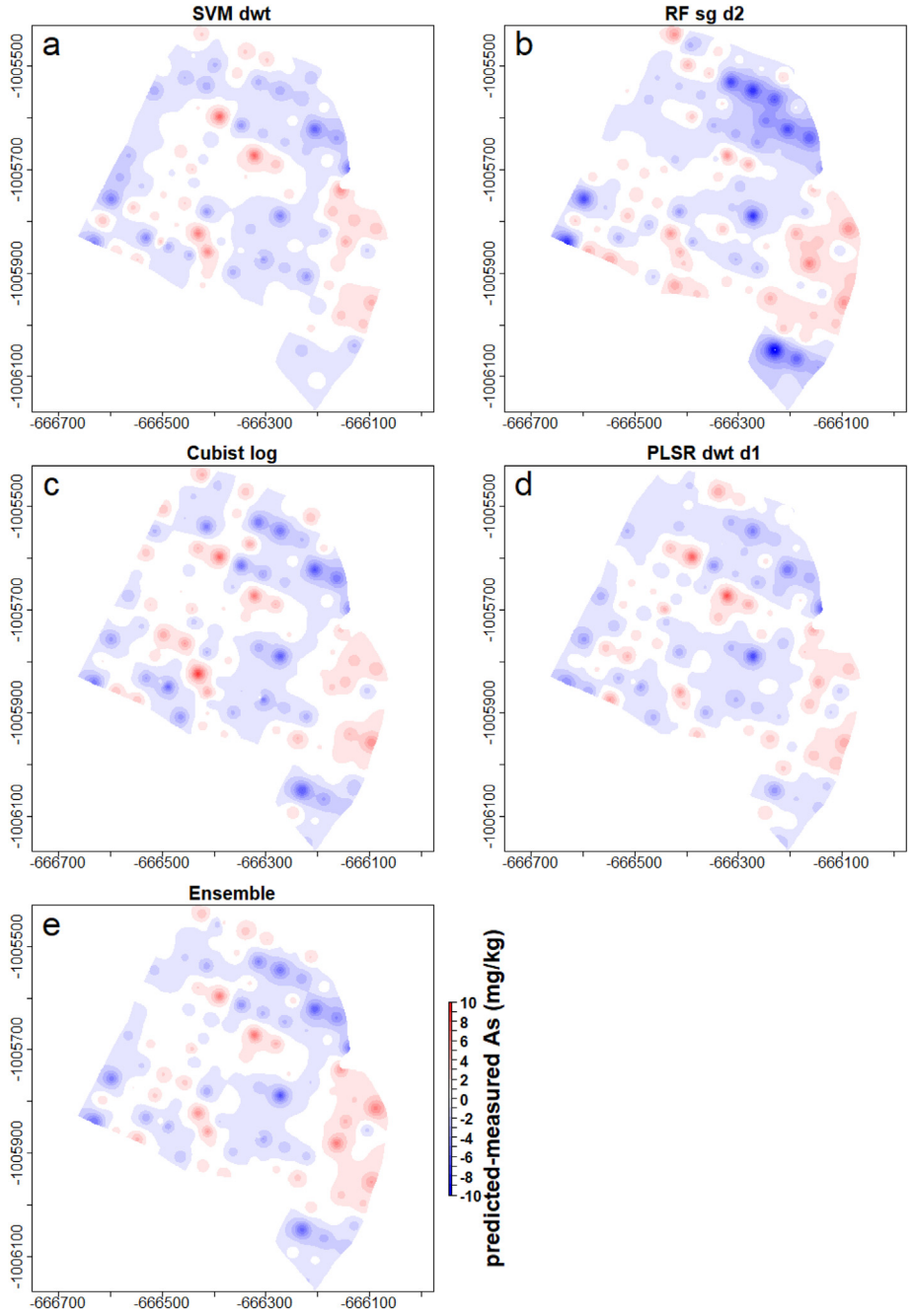


Fig. 6. Differences in spatial *As* distribution (predicted-measured) for individual techniques and the ensemble model [SVM(a), RF (b), PLSR (c), cubist (d), ensemble model (e)].

Funding

This study was supported by an internal grant of the Czech University of Life Sciences Prague, project no. SV20-5-21130.

CRediT authorship contribution statement

James Biney: Conceptualization, Methodology, Writing- Original draft preparation, Analysis, Data curation and Investigation, Visualization, Editing. **Radim Vašát**: software, Data curation. **Johanna Ruth Blöcher**: Software, Editing, Visualization. **Luboš Borůvka**: Supervision, Visualization. **Karel Němeček**: Data Curation and Visualization.

Declaration of competing interest

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

Acknowledgement

The authors also acknowledge the support of the European Regional Development Fund Project Centre for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products (no. CZ.02.1.01/0.0/0.0/16_019/0000845).

We also gratefully acknowledge Aleš Klement and Miroslav Fér for their assistance in soil sampling and spectral measurement.

References

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., Kocarek, M., 2021. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environ. Geochem. Health 43 (1). https://doi.org/10.1007/s10653-020-00743-8.

Arabameri, A., Pradhan, B., Rezaei, K., Sohrabi, M., Kalantari, Z., 2019. GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms. J. Mt. Sci. 16 (3). https://doi.org/10.1007/s11629-018-5168-y.

Axe, L., Tyson, T., Trivedi, P., Morrison, T., 2000. Local structure analysis of strontium sorption to hydrous manganese oxide. J. Colloid Interface Sci. 224 (2). https://doi.org/10.1006/jcis.2000.6737.

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: standards and protocols. Geoderma, 245–246. https://doi.org/10.1016/j.geoderma.2015.01.002.

Ben-Dor, E., Banin, A., 1995. Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. Soil Sci. 159 (4). https://doi.org/10.1097/00010694-199504000-00005.

Bennett, G.F., 1996. Arsenic in the environment — part I: cycling and characterization. J. Hazard. Mater. 48 (1–3). https://doi.org/10.1016/s0304-3894(96)90005-9.

Biney, J.K.M., Boruvka, L., Agyeman, P.C., Němeček, K., Klement, A., 2020. Comparison of field and laboratory wet soil spectra in the vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements. Remote Sens. 12 (18). https://doi.org/10.3390/RS12183082.

Biney, J.K.M., Saberioon, M., Borůvka, L., Houška, J., Vašát, R., Agyeman, P.C., Coblinski, J.A., Klement, A., 2021. Exploring the suitability of uas-based multispectral images for estimating soil organic carbon: comparison with proximal soil sensing and spaceborne imagery. Remote Sens. 13 (2). https://doi.org/10.3390/rs13020308.

Breuniq, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. SIGMOD Record (ACM Special Interest Group on Management of Data). 29 (2). https://doi.org/10.1145/335191.335388.

Brevik, E.C., Pereg, L., Steffan, J.J., Burgess, L.C., 2018. Soil ecosystem services and human health. Current Opinion in Environmental Science and Health. vol. 5. https://doi.org/10.1016/j.coesh.2018.07.003.

Cao, X., Ma, L.Q., 2004. Effects of compost and phosphate on plant arsenic accumulation from soils near pressure-treated wood. Environ. Pollut. 132 (3). https://doi.org/10.1016/j.envpol.2004.05.019.

Chakraborty, S., Man, T., Paulette, L., Deb, S., Li, B., Weindorf, D.C., Frazier, M., 2017. Rapid assessment of smelter/mining soil contamination via portable X-ray fluorescence spectrometry and indicator kriging. Geoderma 306. https://doi.org/10.1016/j.geoderma.2017.07.003.

Chakraborty, S., Weindorf, D.C., Paul, S., Ghosh, B., Li, B., Ali, M.N., Ghosh, R.K., Ray, D.P., Majumdar, K., 2015. Diffuse reflectance spectroscopy for monitoring lead in landfill agricultural soils of India. Geoderma Reg. 5. https://doi.org/10.1016/j.geodrs.2015.04.004.

Chen, M., Li, X.ming, Yang, Q., Zeng, G. ming, Zhang, Y., Liao, D.xiang, Liu, J.jin, Hu, J.mei, Guo, L., 2008. Total concentrations and speciation of heavy metals in municipal sludge from Changsha, Zhuzhou and Xiangtan in middle-south region of China. J. Hazard. Mater. 160 (2–3). https://doi.org/10.1016/j.jhazmat.2008.03.036.

Cui, J.li, Zhao, Y.ping, Li, J.shan, Beiyuan, J.zi, Tsang, D.C.W., Poon, C.sun, Chan, T.shan, Wang, W.xiong, Li, X.dong, 2018. Speciation, mobilization, and bioaccessibility of arsenic in geogenic soil profile from Hong Kong. Environ. Pollut. 232. https://doi.org/10.1016/j.envpol.2017.09.040.

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk A 24 (6). https://doi.org/10.1007/s00477-010-0378-z.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L.M.C., 2013. Breaking with trends in pre-processing? TrAC -Trends Anal. Chem. vol. 50. https://doi.org/10.1016/j.trac.2013.04.015

Ettler, V., Vaněk, A., Mihaljevič, M., Bezdička, P., 2005. Contrasting lead speciation in forest and tilled soils heavily polluted by lead metallurgy. Chemosphere 58 (10). https://doi.org/10.1016/j.chemosphere.2004.09.084.

Eze, P.N., Mosokomani, V.S., Udeigwe, T.K., Oyedele, O.F., 2016. Quantitative geospatial dataset on the near-surface heavy metal concentrations in semi-arid soils from Maibele Airstrip North,Central Botswana. Data Brief 8. https://doi.org/10.1016/j.dib.2016.08.026.

Fang, H., Jamali, B., Deletic, A., Zhang, K., 2021. Machine learning approaches for predicting the performance of stormwater biofilters in heavy metal removal and risk mitigation. Water Res. 200. https://doi.org/10.1016/j.watres.2021.117273.

FDA, 2001. Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc. Report of the Panel on Micronutrients. National Academy Press, Washington, DC, Food and Drug Administration. Dietary supplements. Center for Food Safety and Applied Nutrition.

Fitz, W.J., Wenzel, W.W., 2002. Arsenic transformations in the soil-rhizosphere-plant system: fundamentals and potential application to phytoremediation. J. Biotechnol. 99 (3). https://doi.org/10.1016/S0168-1656(02)00218-3.

Frost, J., 2019. Guidelines for removing and handling outliers in data. https://statisticsbyjim.com/basics/remove-outliers.

Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., 2016. A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra. Remote Sens. 8 (4). https://doi.org/10.3390/rs8040341.

Gholizadeh, A., Borůvka, L., Vašát, R., Saberioon, M., Klement, A., Kratina, J., Tejnecký, V., Drábek, O., 2015. Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: a case study. PLoS ONE 10 (2). https://doi.org/10.1371/journal.pone.0117457.

Gholizadeh, A., Saberioon, M., Ben-Dor, E., Rossel, R.A.V., Borůvka, L., 2020. Modelling potentially toxic elements in forest soils with vis–NIR spectra and learning algorithms. Environ. Pollut. 267, 115574. https://doi.org/10.1016/j.envpol.2020.115574.

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sens. Environ. 218. https://doi.org/10.1016/j.rse.2018.09.015.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R J. 8 (1). https://doi.org/10.32614/rj-2016-014.

Hale, J.R., Foos, A., Zubrow, J.S., Cook, J., 1997. Better characterization of arsenic and chromium in soils: a field-scale example. Soil Sediment Contam. 6 (4). https://doi.org/10.1080/15320389709383573.

Han, Y., Tang, Z., Sun, J., Xing, X., Zhang, M., Cheng, J., 2019. Heavy metals in soil contaminated through e-waste processing activities in a recycling area: implications for risk management. Process Saf. Environ. Prot. 125. https://doi.org/10.1016/j.psep.2019.03.020.

Hartley, W., Edwards, R., Lepp, N.W., 2004. Arsenic and heavy metal mobility in iron oxide-amended contaminated soils as evaluated by short- and long-term leaching tests. Environ. Pollut. 131 (3). https://doi.org/10.1016/j.envpol.2004.02.017.

Holub, M, 1997. Produkce arzenu při těžbě drahých a barevných kovů [Outturn of arsenic at exploitation of precious and non-ferrous metals]. (in Czech). Vesmír 76, 389–392.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R., Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: a prospective review. Geoderma. vols. 241–242. https://doi.org/10.1016/j.geoderma.2014.11.024.

Hu, Y., Cheng, H., 2013. Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a large-scale region. Environ. Sci. Technol. 47 (8). https://doi.org/10.1021/es304310k.

Ikem, A., Egiebor, N.O., 2005. Assessment of trace elements in canned fishes (mackerel, tuna, salmon, sardines and herrings) marketed in Georgia and Alabama (United States of America). J. Food Compos. Anal. 18 (8). https://doi.org/10.1016/j.jfca.2004.11.002.

IUSS Working Group WRB, 2014. World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106 https://doi.org/10.1017/S0014479706394902.

Järup, L., 2003. Hazards of heavy metal contamination | British Medical Bulletin | Oxford Academic. British Medical Bulletin. vol. 68, Issue 1.

Jiang, Y., Zeng, X., Fan, X., Chao, S., Zhu, M., Cao, H., 2015. Levels of arsenic pollution in daily foodstuffs and soils and its associated human health risk in a town in Jiangsu ProvinceChina. Ecotoxicol. Environ. Saf. 122. https://doi.org/10.1016/j.ecoenv.2015.07.018.

Kalantar, B., Pradhan, B., Amir Naghibi, S., Motevalli, A., Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics Nat. Hazards Risk 9 (1). https://doi.org/10.1080/19475705.2017.1407368.

Kebonye, N.M., Eze, P.N., 2019. Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. Environ. Monit. Assess. 191 (11). https://doi.org/10.1007/s10661-019-7812-6.

Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. Geoderma 384. https://doi.org/10.1016/j.geoderma.2020.114792.

Khosravi, V., Doulati Ardejani, F., Yousefi, S., 2017. Spectroscopic-based assessment of the content and geochemical behaviour of arsenic in a highly heterogeneous sulphide-rich mine waste dump. Environ. Earth Sci. 76 (13). https://doi.org/10.1007/s12665-017-6793-4.

Kotková, K., Nováková, T., Tůmová, Kiss, T., Popelka, J., Faměra, M., 2019. Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. Geomorphology 329. https://doi.org/10.1016/j.geomorph.2018.12.010.

Kuhn, M., Johnson, K., 2013. Over-fitting and model tuning. Applied Predictive Modeling https://doi.org/10.1007/978-1-4614-6849-3_4.

Liao, Y., Li, D., Zhang, N., 2018. Comparison of interpolation models for estimating heavy metals in soils under various spatial characteristics and sampling methods. Trans. GIS 22 (2). https://doi.org/10.1111/tgis.12319.

Liaw, A., Wiener, M., 2002. Classification and regression by random Forest. R News 2 (3).

Li, C., Wang, J., Yan, B., Miao, A.J., Zhong, H., Zhang, W., Ma, L.Q., 2021. Progresses and emerging trends of arsenic research in the past 120 years. Crit. Rev. Environ. Sci. Technol. 51 (13). https://doi.org/10.1080/10643389.2020.1752611.

Lin, H.T., Wang, M.C., Li, G.C., 2004. Complexation of arsenate with humic substance in water extract of compost. Chemosphere 56 (11). https://doi.org/10.1016/j.chemosphere.2004.05.018.

Liu, W., Du, P., Wang, D., 2015. Ensemble learning for spatial interpolation of soil potassium content based on environmental information. PLoS ONE 10 (4). https://doi.org/10.1371/journal.pone.0124383.

Liu, Y.M., Liu, D.Y., Zhang, W., Chen, X.X., Zhao, Q.Y., Chen, X.P., Zou, C.Q., 2020. Health risk assessment of heavy metals (Zn, Cu, Cd, Pb, As and Cr) in wheat grain receiving repeated Zn fertilizers. Environ. Pollut. 257. https://doi.org/10.1016/j.envpol.2019.113581.

Llobet, J.M., Falcó, G., Casas, C., Teixidó, A., Domingo, J.L., 2003. Concentrations of arsenic, cadmium, mercury, and lead in common foods and estimated daily intake by

children, adolescents, adults, and seniors of Catalonia, Spain. J. Agric. Food Chem. 51 (3). https://doi.org/10.1021/jf020734q.
Lu, A., Wang, J., Qin, X., Wang, K., Han, P., Zhang, S., 2012. Multivariate and geostatistical analyses of the spatial distribution and origin of heavy metals in the agricultural soils in Shunyi, Beijing,China. Sci. Total Environ. 425. https://doi.org/10.1016/j.scitotenv.2012.03.003.
Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., Buttafuoco, G., 2017. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma 288. https://doi.org/10.1016/j.geoderma.2016.11.015.
Mandal, B.K., Suzuki, K.T., 2002. Arsenic round the world: a review. Talanta. vol. 58, Issue 1. https://doi.org/10.1016/S0039-9140(02)00268-0.
Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rötter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob. Chang. Biol. 21 (2). https://doi.org/10.1111/gcb.12768.
Mevik, B.H., Wehrens, R., 2007. The pls package: principal component and partial least squares regression in R. J. Stat. Softw. 18 (2). https://doi.org/10.18637/jss.v018.i02.
Mishra, U., Gautam, S., Riley, W.J., Hoffman, F.M., 2020. Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. Front. Big Data 3. https://doi.org/10.3389/fdata.2020.528441.
Misra, S.G., Tiwari, R.C., 1963. Studies on arsenite-arsenate system adsorption of arsenate. Soil Sci.Plant Nutr. 9 (6). https://doi.org/10.1080/00380768.1963.10431056.
Murray, I., 1988. Aspects of the interpretation of NIR spectra. In: Creaser, C.S., Davies, A.M.C. (Eds.), Analytical Applications of Spectroscopy. 1988. Royal Society of Chemistry, London.
Næsset, E., Bollandsås, O.M., Gobakken, T., 2005. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. Remote Sens. Environ. 94 (4). https://doi.org/10.1016/j.rse.2004.11.010.
Nicholson, F.A., Smith, S.R., Alloway, B.J., Carlton-Smith, C., Chambers, B.J., 2003. An inventory of heavy metals inputs to agricultural soils in England and Wales. Sci. Total Environ. 311 (1–3). https://doi.org/10.1016/S0048-9697(03)00139-6.
Oliveri, P., Malegori, C., Simonetti, R., Casale, M., 2019. The impact of signal pre-processing on the final interpretation of analytical outcomes – a tutorial. Analytica Chimica Acta. vol. 1058. https://doi.org/10.1016/j.aca.2018.10.055.
O'Rourke, S.M., Stockmann, U., Holden, N.M., McBratney, A.B., Minasny, B., 2016. An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. Geoderma 279. https://doi.org/10.1016/j.geoderma.2016.05.005.
Panchenko, T., Pradivlyana, A., Ursul, O., 2018. Estimation of the influence of thermal energy objects on the environment. Materialy XLVII Naukovo-Tekhnichnoyi Konferentsiyi Pidrozdiliv VNTU. Vinnytsya . https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2018/paper/view/4574.
Patel, K.S., Shrivas, K., Brandt, R., Jakubowski, N., Corns, W., Hoffmann, P., 2005. Arsenic contamination in water, soil, sediment and rice of central India. Environ. Geochem. Health 27 (2). https://doi.org/10.1007/s10653-005-0120-9.
Paulette, L., Man, T., Weindorf, D.C., Person, T., 2015. Rapid assessment of soil and contaminant variability via portable x-ray fluorescence spectroscopy: Copşa Mică,Romania. Geoderma 243–244. https://doi.org/10.1016/j.geoderma.2014.12.025.
Pebesma, E., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.
Plyatsuk, L., Balintova, M., Chernysh, Y., Demcak, S., Holub, M., Yakhnenko, E., 2019. Influence of phosphogypsum dump on the soil ecosystem in the Sumy region (Ukraine). Appl. Sci. 9 (24). https://doi.org/10.3390/app9245559.
Qiao, P., Lei, M., Yang, S., Yang, J., Guo, G., Zhou, X., 2018. Comparing ordinary kriging and inverse distance weighting for soil as pollution in Beijing. Environ. Sci. Pollut. Res. 25 (16). https://doi.org/10.1007/s11356-018-1552-y.
Quinlan, J., 1992. Learning with continuous classes. In: Adam (Ed.), Proceedings of AI'92, second World Scientific.
Ravansari, R., Wilson, S.C., Tighe, M., 2020. Portable X-ray fluorescence for environmental assessment of soils: not just a point and shoot method. Environment International. vol. 134. https://doi.org/10.1016/j.envint.2019.105250.
R Core Team R., core Team, Team, R.D.C., 2015. R: A Language and Environment for Statistical Computing.
Ren, H.Y., Zhuang, D.F., Singh, A.N., Pan, J.J., Qiu, D.S., Shi, R.H., 2009. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: a case study. Pedosphere 19 (6). https://doi.org/10.1016/S1002-0160(09)60167-3.
Renka, R.J., 1996. Algorithm 751: TRIPACK: a constrained two-dimensional delaunay triangulation package. ACM Trans. Math. Softw. 22 (1). https://doi.org/10.1145/225545.225546.
Rinnan, Å., van den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC -Trends Anal. Chem. vol. 28, Issue 10. https://doi.org/10.1016/j.trac.2009.07.007.
Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2). https://doi.org/10.1016/j.geoderma.2009.12.025.
Sacristán, D., Viscarra Rossel, R.A., Recatalá, L., 2016. Proximal sensing of Cu in soil and lettuce using portable X-ray fluorescence spectrometry. Geoderma 265. https://doi.org/10.1016/j.geoderma.2015.11.008.
Schmidt, K., Behrens, T., Friedrich, K., Scholten, T., 2010. A method to generate soilscapes from soil maps. J. Plant Nutr. Soil Sci. 173 (2). https://doi.org/10.1002/jpln.200800208.
Shi, T., Wang, J., Chen, Y., Wu, G., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. Int. J. Appl. Earth Obs. Geoinf. 52. https://doi.org/10.1016/j.jag.2016.06.002.
Signal Developers, 2013. Signal: signal processing. URL: http://r-forge.r-project.org/projects/signal.
Singh, V., Garg, A.N., 2006. Availability of essential trace elements in indian cereals, vegetables and spices using INAA and the contribution of spices to daily dietary intake. Food Chem. 94 (1). https://doi.org/10.1016/j.foodchem.2004.10.053.
Skjemstad, J.O., Baldlock, J.A., 2008. Total and organic carbon. In: Carter, M., Gregorich, E. (Eds.), Soil sampling and methods of analysis, 2nd edn CRC Press, Boca Raton.
Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behaviour and distribution of arsenic in natural waters. Applied Geochemistry. vol. 17, Issue 5. https://doi.org/10.1016/S0883-2927(02)00018-5.
Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta,China. Applied Clay Science 64. https://doi.org/10.1016/j.clay.2011.09.010.
Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., van Wesemael, B., 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158 (1–2). https://doi.org/10.1016/j.geoderma.2009.11.032.
Dietterichl, Thomas G., 2002. The Handbook of Brain Theory and Neural Networks - Ensemble Learning. 40(2). MIT Press.
Tremlová, J., Sehnal, M., Száková, J., Goessler, W., Steiner, O., Najmanová, J., Horáková, T., Tlustoš, P., 2017. A profile of arsenic species in different vegetables growing in arsenic-contaminated soils. Arch. Agron. Soil Sci. 63 (7). https://doi.org/10.1080/03650340.2016.1242721.
Vaněk, A., Ettler, V., Grygar, T., Borůvka, L., Šebek, O., Drábek, O., 2008. Combined chemical and mineralogical evidence for heavy metal binding in mining- and smelting-affected alluvial soils. Pedosphere 18 (4). https://doi.org/10.1016/s1002-0160(08)60037-5.
Vašát, R., Kodešová, R., Klement, A., Borůvka, L., 2017. Simple but efficient signal preprocessing in soil organic carbon spectroscopic estimation. Geoderma 298. https://doi.org/10.1016/j.geoderma.2017.03.012.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Ji, W., 2016. A global spectral library to characterize the world's soil. In. Earth Sci. Rev. 155. https://doi.org/10.1016/j.earscirev.2016.01.012.
Waiser, T.H., Morgan, C.L., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Sci. Soc. Am. J. 71 (2), 389–396. https://doi.org/10.2136/sssaj2006.0211.
Warren, G.P., Alloway, B.J., Lepp, N.W., Singh, B., Bochereau, F.J.M., Penny, C., 2003. Field trials to assess the uptake of arsenic by vegetables from contaminated soils and soil remediation with iron oxides. Sci. Total Environ. 311 (1–3). https://doi.org/10.1016/S0048-9697(03)00096-2.
Wei, L., Yuan, Z., Yu, M., Huang, C., Cao, L., 2019. Estimation of arsenic content in soil based on laboratory and field reflectance spectroscopy. Sensors (Switzerland) 19 (18). https://doi.org/10.3390/s19183904.
Weindorf, D.C., Chakraborty, S., Herrero, J., Li, B., Castañeda, C., Choudhury, A., 2016. Simultaneous assessment of key properties of arid soil by combined PXRF and Vis-NIR data. Eur. J. Soil Sci. 67 (2). https://doi.org/10.1111/ejss.12320.
Weindorf, D.C., Paulette, L., Man, T., 2013. In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna,Romania. Environ. Pollut. 182. https://doi.org/10.1016/j.envpol.2013.07.008.
Wu, H., Yang, F., Li, H., Li, Q., Zhang, F., Ba, Y., Cui, L., Sun, L., Lv, T., Wang, N., Zhu, J., 2020. Heavy metal pollution and health risk assessment of agricultural soil near a smelter in an industrial city in China. Int. J. Environ. Health Res. 30 (2). https://doi.org/10.1080/09603123.2019.1584666.
Wu, Q., Leung, J.Y.S., Geng, X., Chen, S., Huang, X., Li, H., Huang, Z., Zhu, L., Chen, J., Lu, Y., 2015. Heavy metal contamination of soil and water in the vicinity of an abandoned e-waste recycling site: implications for dissemination of heavy metals. Sci. Total Environ. 506–507. https://doi.org/10.1016/j.scitotenv.2014.10.121.
Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71 (3). https://doi.org/10.2136/sssaj2006.0285.
Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J., Qin, Z., 2005. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Appl. Geochem. 20 (6). https://doi.org/10.1016/j.apgeochem.2005.01.009.
Xie, M., Li, H., Zhu, Y., Xue, J., You, Q., Jin, B., Shi, Z., 2021. Predicting bioaccumulation of potentially toxic element in soil–rice systems using multi-source data and machine learning methods: a case study of an industrial city in Southeast China. Land 10 (6). https://doi.org/10.3390/land10060558.
Xie, X.L., Pan, X.Z., Sun, B., 2012. Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a copper smelter. Pedosphere 22 (3). https://doi.org/10.1016/S1002-0160(12)60022-8.
Xie, Y., Lei, M., Yang, J., Guo, Q.J., Song, B., Zhou, X.Y., Chen, T.bin, 2011. Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: accuracy and uncertainty analysis. Chemosphere 82 (3). https://doi.org/10.1016/j.chemosphere.2010.09.053.
Xu, X., Ren, M., Cao, J., Wu, Q., Liu, P., Lv, J., 2020. Spectroscopic diagnosis of zinc contaminated soils based on competitive adaptive reweighted sampling algorithm and an improved support vector machine. Spectrosc. Lett. 53 (2). https://doi.org/10.1080/00387010.2019.1696828.
Yin, F., Wu, M., Liu, L., Zhu, Y., Feng, J., Yin, D., Yin, C., Yin, C., 2021. Predicting the abundance of copper in soil using reflectance spectroscopy and GF5 hyperspectral imagery. Int. J. Appl. Earth Obs. Geoinf. 102. https://doi.org/10.1016/j.jag.2021.102420.
Zhou, W., Yang, H., Xie, L., Li, H., Huang, L., Zhao, Y., Yue, T., 2021. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. Catena 202. https://doi.org/10.1016/j.catena.2021.105222.

Manuscript Under Revision (Soil and Tillage Research)

Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations.

James Kobina Mensah Biney[1,2], RadimVašát[1], Stephen Mackenzie Bell[3], Ndiye Michael Kebonye[1], Aleš Klement[1], Kingsley John[1], Luboš Borůvka[1]

[1]Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic

[2]The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Department of Landscape Ecology, Lidická 25/27, Brno, 602 00, Czech Republic

[3]Institute of Environmental Science and Technology (ICTA-UAB), Universitat Autònoma de Barcelona Carrer de les Columnes s/n, Campus UAB, Edifici ICTA-ICP, 08193 Cerdanyola del Vallès, Barcelona Spain

Correspondence E-mail: no2james@yahoo.com (J.K.M. Biney)

Abstract

Estimating soil organic carbon (SOC) using visible near infrared (Vis-NIR) spectroscopy has proven to be a rapid and reliable approach. However, when working across large geographical scales, remote sensing may be more suitable. Acquiring these spectra data normally under different measurement conditions could introduce artefacts that reduce SOC prediction accuracy. A common procedure has been using calibration or multivariate techniques in conjunction with one or more pre-treatment algorithms. The results of several comparative studies based on these predictive calibration techniques used alone were inconsistent. Moreover, protocols to select the most appropriate pre-treatment algorithms rarely exist. This study combines predictions from different techniques into a single model based on an ensemble learning approach. The main objective is to improve the accuracy of SOC prediction by assessing the effectiveness of using different calibration techniques individually against an ensemble model consisting of one statistical method, which includes partial least squares regression

(PLSR), and three machine learning (ML) algorithms, including random forest (RF), support vector machine regression (SVMR), and Cubist. Several pre-treatment algorithms were also employed to improve the spectral data before prediction. Spectra data were collected from three different agricultural fields (with different soil types), under different spectral measurement conditions (field, wet and dry). Additionally, Sentinel-2 (S2) data was collected from one of these fields. Furthermore, to ascertain the effectiveness of the developed model on regional scale dataset, two options were employed: (1) merged data from all fields, and (2) merged data from fields measured under the same spectral measurement conditions. The models were evaluated using root mean square error of prediction ($RMSEP_{CV}$), the coefficient of determination (R^2_{CV}), the ratio of performance to interquartile range (RPIQ), the ratio of performance to deviation (RPD) and BIAS. The results show that, across the three agricultural fields, the ensemble model predicted SOC more accurately than each of the individual calibration techniques (R^2_{CV} = 0.92, $RMSEP_{CV}$ = 0.10, RPD = 3.06, RPIQ = 3.74, BIAS = 0.0067). The models derived from merged data (regional dataset) show that the ensemble approach predicted SOC more accurately with option 2 than option 1. Finally, while the ensemble model improves SOC accuracy with S2 data, the final output was poor. Further research to determine the underlying problem is strongly recommended. Nonetheless, these results indicate that the ensemble model is advantageous because it improved the prediction accuracy of SOC and reduced the error margin.

Keywords: Ensemble predictive model; Soil organic carbon; Agricultural soil; Spectroscopy (field-wet-dry); Sentinel-2; Pre-treatment

1. Introduction

Essential ecological resources, such as food and fibre production, are under threat because of the pressures on soils and their carbon stocks stemming from rapid urban growth, land degradation, and intensive agriculture. There is a growing need for accurate monitoring of soil organic carbon (SOC), as carbon losses can have negative impacts on agricultural soil productivity and ecosystem health while contributing to increasing atmospheric greenhouse gas concentrations (Lausch et al., 2019; Paustian et al., 2016; Zádorová et al., 2015; Sanchez et al., 2009; Van Oost et al., 2007; Lal, 2004)

Visible near-infrared (Vis-NIR) spectroscopy has proven to be a rapid, accurate, and trustworthy approach to estimate SOC. However, acquiring these spectra data under different measurement conditions could introduce artefacts or other factors that increase the chances for inaccuracies. These can result, for example, from the number of field observations, the density of soil samples, the target variables, possible soil properties (i.e., chemical, physical, biological, etc.) and perhaps other soil constituents (see e.g., Stenberg et al., 2010). According to Ben-Dor et al. (2015), a key Vis-NIR limitation is that soil spectral information is susceptible to the conditions in which the soil is scanned. This can have a significant impact on the measured reflectance spectra and obtained calibration models, reducing the accuracy of SOC prediction (Ge et al., 2011).

In recent years, more modern, dynamic, and comprehensive multivariate calibration approaches using linear and non-linear methods and machine learning (ML) algorithms (e.g., Lamichhane et al., 2019; Heung et al., 2016; Stevens et al., 2010; Viscarra Rossel and Behrens, 2010) have been used to retrieve vital information from proximal and remote sensing datasets (e.g., Carmon and Ben-Dor, 2017; Viscarra Rossel et al., 2016) to estimate SOC as well as other soil properties. Additionally, through the use of digital soil mapping (DSM), these ML algorithms can create a possible relationship between soil and environmental predictor variables to predict soil properties such as SOC for areas that have not been physically sampled (Padarian et al., 2020; Taghizadeh-Mehrjardi et al., 2020). Unfortunately, the results of several comparative studies based on ML algorithms were not consistent (Wang et al., 2018; Jeong et al., 2017; Were et al., 2015). This is because the predictive algorithms' output may be influenced by a range of factors or artefacts. Furthermore, different ML algorithms can identify different sets of important predictor variables when estimating SOC. According to Diettrich. (2002), the solution to this issue may lie in combining multiple individual prediction models made with different calibration techniques into a single model, using ensemble learning theory and selecting appropriate pre-treatment options to minimise or eliminate artefacts.

Ensemble models are defined as machine learning models in which several learners are trained to address the same problem (Dietterich, 2002). As a result, the ensemble model will be presented with a broader range of individual predictions as data input. The eventual prediction is estimated based on

either an average or weighted average of these individual algorithms. The assumption is that the new model would be at least as good as all individual models effectively using all available information (Diks and Vrugt, 2010). However, to further improve the predictive accuracy of the calibration, model pre-treatment methods may be needed; but, there is no hard-and-fast rule for determining a specific pre-treatment and, more importantly, the specific multivariate algorithm to match it.

The ensemble model constructed for this study consists of a combination of four individual modelling techniques: one statistical method, which includes partial least squares regression (PLSR), and three ML algorithms, including random forest (RF), support vector machine regression (SVMR), and Cubist. Additionally, 32 pre-treatment algorithms with several combinations for removing noise and other irrelevant information from the spectra before prediction were also employed to assess the likelihood of enhancing the predictive efficiency of spectroscopic models to predict SOC accurately. The model will first return the results for each technique and their best pre-treatment algorithms (after exploring all the several pre-treatment algorithms available). The best output for each technique, together with its optimal pre-treatment, will then be fed into the ensemble model in a combined form (single unit) as an input to generate the ensemble model prediction outcome using a weighted average approach.

Although the ensemble theory is not new, it has yet to be fully explored on all the numerous ML and statistical approaches currently available. Furthermore, its application in soil science is limited in contrast to other fields of study (Li et al., 2021; Li et al., 2021; Huang et al., 2020; Kalantar et al., 2020; Ma et al., 2019; Althuwaynee et al., 2014; Engler et al., 2013; Chi et al., 2009, etc.). In soil carbon research, it's been used for the assessment of SOC biogeochemical processes (Farina et al., 2021), for identifying trends in SOC stock using time series data (Riggers et al., 2019), for modeling SOC using topographic attributes and vegetation indices (Tajik et al., 2020), and for analysing laboratory dry data to estimate SOC content (Vaát et al., 2017). It has, however, not yet been widely researched for SOC prediction using wet and field spectra data and employing both Sentinel-1 and 2 imagery.

In this study, Vis-NIR spectroscopy datasets were collected in three different agricultural fields in the Czech Republic (with different soil types) under different spectral measurement conditions (dry, field and wet). The goal of this study is twofold: (i) to use Vis-NIR spectroscopy datasets to predict SOC

more accurately using ensemble models while employing the most appropriate signal pre-treatment strategy, and (ii) to verify the suitability of the developed ensemble model on two other datasets (i.e., a remote sensing dataset [Sentinel-2 imagery (S2)] acquired from only one of the study fields and a regional scale dataset obtained by merging different spectra data).

2. Materials and methods

2.1. Study areas

For this study, three different agricultural sites of varying soil types and in different parts of the Czech Republic were selected: Vidim, Bromovice, and Nová Ves nad Popelkou (Field A, Field B, and Field C, respectively) (Figure 1). These fields are examples of intensively farmed arable land.

2.1.1. Vidim (Field A)

Vidim is an agricultural site spanning 8 ha (50°28'4.262 "N, 14°31'32.968"E, altitude 315–323 m above sea level (a.s.l.), average annual temperature 7–8 °C, average annual precipitation 550–650 mm) situated in the northern part of the Czech Republic. This area features traditionally cultivated farmlands that are severely affected by water erosion due to significant slope and intensive ploughing. The soil texture is mainly silt loam (Antonín et al., 2021a). From the top to bottom, the soil units were recognised as Haplic Luvisol on loess and loess clays, Regosol, and Colluvic Regosol (Zádorová et al., 2014). According to the World Reference Base (WRB) for soil resources (IUSS Working Group, WRB, 2014), the soils are characterised as Haplic Luvisol, Regosol and Colluvial soil

.2.1.2. Brumovice (Field B)

The Bromovice study area (100 ha, 48°57'38.864"N, 16°53'46.153"E, altitude 187–227 m a.s.l., average annual temperature 9–10 °C, average annual precipitation 550–650 mm) developed on loess substrate and is located in the southeast of the Czech Republic. The area's soil is among the most fertile in the Czech Republic. Wheat and sweet corn are the two most dominant crops grown in this area. The soil texture is mainly Silt loam, Loam (1^{st}) (Antonín et al., 2021b). However, as a result of significant slope, the land is affected by water erosion, with distinct erosion furrows occurring mostly in the steepest parts

of the area. The area was once covered evenly with a distinct soil unit known as Haplic Chernozem. Nevertheless, three different soil types; Regosols (degraded Chernozem), colluvial Chernozem, and colluvial, later evolved due to intensive machinery cultivation and subsequent intense water erosion depending on the topographical conditions of the area. As per the World Reference Base for Soil Resources (IUSS Working Group WRB, 2014), all the soil units were finally classified as Haplic Chernozem, Regosol, Colluvial Chernozem, and Colluvial soil. See, for example, Zádorová et al. (2011a) or Jakšík et al., (2015) for more extensive characterisation of the area.

2.1.3. Nová Ves nad Popelkou (Field C)

The study field at Nová Ves nad Popelkou (22 ha, 50°31′ N; 15°24′ E) is located in the central Bohemian region with a mean altitude of 185 m a.s.l. The soils of the representative area chosen are homogenous and comparable in terms of terrain characteristics, land management, and climatic conditions (Schmidt et al., 2010). Soil texture in area can be classified as sandy-loam and loam (Gholizadeh et al., 2018). The most dominant crop is winter and spring cereals. According to the World Reference Base (WRB) for soil resources (2014), soils of this region are predominantly Cambisols on sedimentary rocks.

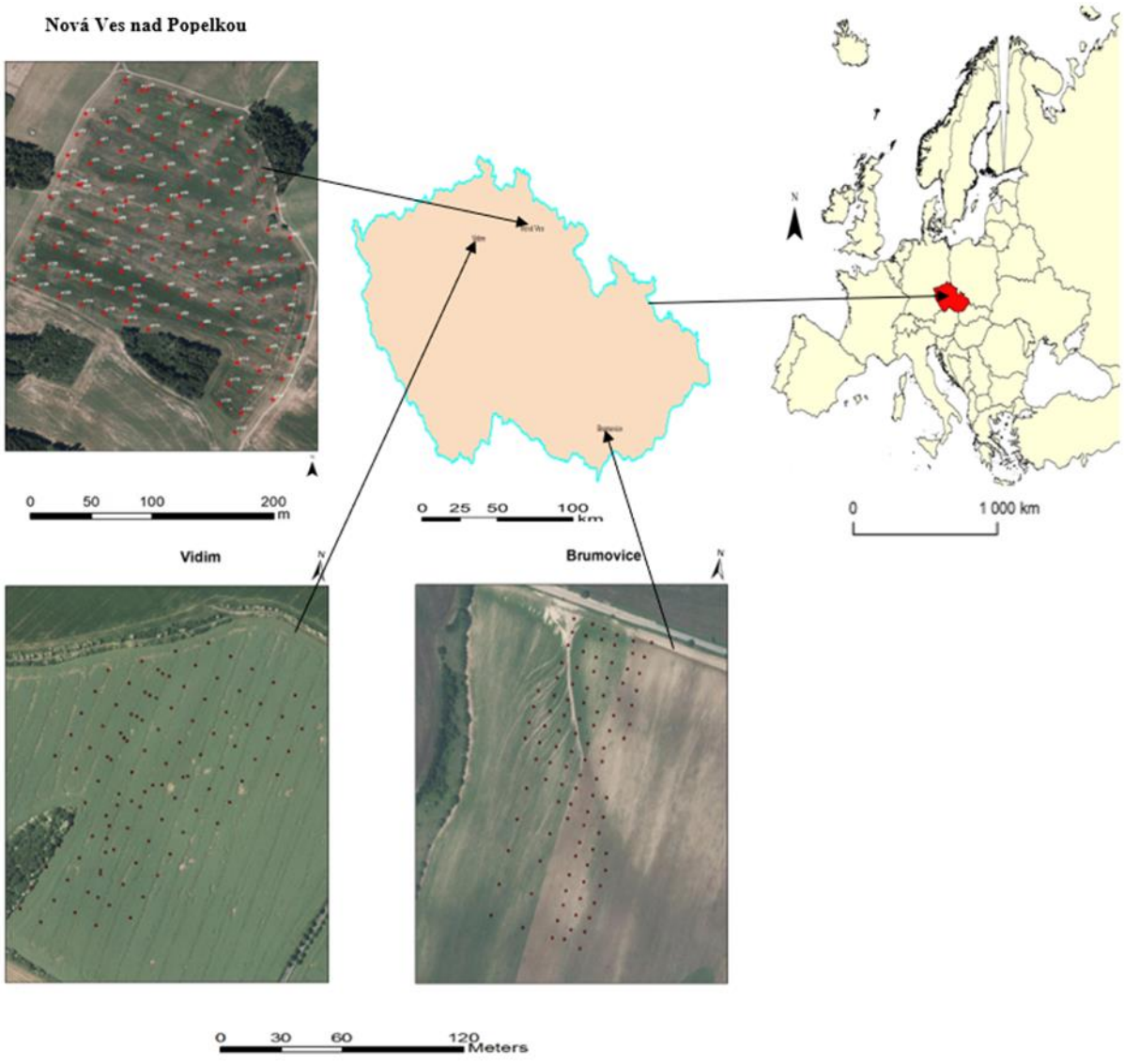


Figure 1: Locations of each study site in the Czech Republic

2.2. Soil sampling and soil organic carbon laboratory determination

A total of 303 soil samples [Field A (76), Field B (97) and Field (130)] were collected from the respective agricultural fields via a rectangular grid design from the topsoil layer (0 to 25 cm depth) in 2013 (Field A), 2010 (Field B) and 2019 (Field C) and placed in clearly labelled bags (approximately 150 to 200 g). The size of the sample provided adequate coverage of the fields and was representative of the area and samples to which the models were developed. Each field's sampling points were selected separately before the field visit and were located on the field using a GeoXM (Trimble Inc., 2007) receiver at an accuracy of 1 m. The samples were then air-dried, gently crushed, and sieved to particle

fraction (to 2 mm) before analysis. SOC measurement was carried out in two steps using the dichromate redox titration method (Skjemstad and Baldock, 2008). The specimens (samples) were initially oxidised with K2Cr2O7, and afterwards, the solution was potentiometrically titrated with ferrous ammonium sulphate (FeH8N2O8S2).

2.3. Laboratory and field spectra measurement

The spectral reflectance of the soil samples for the different Vis-NIR approaches was measured using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) with a high-intensity contact probe across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the range of 350–1050 nm and 10 nm for the range of 1050–2500 nm. The device also has its light source (100 W halogen reflectors lamp), which was used to acquire the soil spectra in the laboratory (Weiser et al., 2007). For Field C, two different spectral measurements were taken, namely field and wet spectra. The field spectra were measured in the field. Three spectral measurements for each sample were collected. The average of these measurements was used as field spectra for Field C. The sensor was re-calibrated preceding the first scan and after every ten runs by scanning a Spectralon® (Labsphere, North Sutton, NH, USA) standard white reference panel with 99% reflectance, reflective area diameter (inches):1.25, housing material: delrin, diameter (inches): 1.5, thickness (inches): 0.55, and operating relative humidity: 5% - 95%. During the field sampling and spectral acquisition for Field C, as stated earlier, samples for laboratory analysis were also collected and immediately upon arrival at the laboratory, spectra readings were taken (three repetitions) and the average value was used as the wet spectra dataset for Field C. The samples for Fields A and B were measured in the laboratory using the same procedure as for Field C. However, before the spectral measurement in the lab, these soil samples were placed in 9 cm diameter Petri dishes to form 2 cm layers of soil. The samples were levelled with a stainless-steel blade, and three replicate measurements was taken in the centre of the sample, and the average of each sample was used as datasets for Fields A and B. The spectrometer was standardised using a Spectralon® panel (Lab-sphere, North Sutton, NH, USA) (Shi et al., 2016).

2.4. Sentinel-2 imagery acquisition and analysis

The Multi-Spectral Sentinel-2B imagery used was a cloud-free image-level 2A, which means it is ready to be used right away because it has already been processed by the providers using Sen2Corprocessor. These processes include geometric, radiometric, and atmospheric corrections. The imagery was obtained from the Copernicus Open Access Hub of the European Space Agency on June 26, 2019. The Sentinel-2 image (S2) is made up of 13 spectral bands. These spectral bands range from visible and near-infrared (vis-NIR) to short-wave infrared (SWIR). There are four bands with a 10 m resolution [(B2, 490 nm), (B3, 560 nm), (B4, 665 nm), (B8, 842 nm)]; six bands at 20 m resolution [(B5, 705 nm), (B6,740 nm), (B7, 775 nm), and (B8A, 865 nm); 2 SWIR large bands, (B11, 1610 nm) and (B12,2190 nm)], and finally, three bands at 60 m resolution [(B1, 443 nm), (B9, 940 nm), and (B10,1380 nm)]. Before extracting these bands, a pixel resolution re-sampling was performed using 10 m as the reference to ensure that all the bands were at the same resolution. This was done using the SNAP software. For further analysis, three bands (B1, B9, B10) were excluded. According to Elhag and Bahrawi (2017), the selected 10 bands are usually used to assess soil properties. Technical details of the S2 bands used in this study can be found in the European Space Agency work book, (2010).

2.5. Comparison of spectra data measured under different conditions and the detection and removal of outliers

When scanning soil materials with spectroscopy in the Vis-NIR range, certain soil properties (e.g., minerals, soil moisture content, etc.) and other factors of a wide variety, including the changing working conditions of the spectrometer device (e.g., temperature, re-calibration intervals, etc.), may influence the spectra during the measurement phase. Since the magnitude of such occurrences or influences cannot be seen with the naked eye, the different spectra data (field, wet, and dry samples) were transformed with log transformation and continuum removal to visualise and compare the various spectral forms. The raw spectra without modification were used as references.

The initial spectral range was 350-2500 nm. Before data transformation with the several pre-treatment algorithms, the noisy region between 350–399 nm was removed, leaving a range of 400-2500 nm. Furthermore, the presence of outliers was explored using ensemble sparse partial least squares (enpls),

because according to Balakrishnan (1994) and Frost. (2019), outliers could influence prediction accuracy. The number of outliers removed was Field A (0), Field B (2), Field C (wet (4), field (5)).

2.6. Dataset pre-processing

The raw spectra (after outlier removal) were transformed into reflectance. The datasets were then subjected to the following set of pre-treatment techniques: sg (Savitzky-Golay) from signal R package (Signal developers, 2013), dwt (discrete wavelet transformation) calculated with dwt function from wavelets R package (Aldrich, 2013), d1 (first-order derivative) (Duckworth, 2004), sg_d1, dwt_d1, d2 (second-order derivative), sg_d2, dwt_d2, msc (multiplicative scatter correction) which was calculated using pls R package (Mevik and Wehrens, 2007), sg_msc, dwt_msc, snv (standard normal variate) which was obtained by subtracting each reflectance value from the spectrum's mean reflectance value, and then it was divided by standard deviation, sg_snv, dwt_snv, snv_msc, sg_snv_msc, dwt_snv_msc, log (logarithmic transformation (log(1/R))), sg_log, dwt_log, log_msc, sg_log_msc, dwt_log_msc, log_snv, sg_log_snv, dwt_log_snv, cr (continuum removal) which was obtained from tripack R package (Renka, 1996), sg_cr, dwt_cr, cmr (correction by the maximum reflectance) (Vašát el al, 2017), sg_cmr, dwt_cmr in order to optimise the fitting of target values against spectra. These algorithms also seek to remove or minimise undesirable side effects (i.e., artefacts) in the spectra while also improving the relevant details about the soil property being estimated. It is worth mentioning that only the best pre-treatment results for each calibration technique are shown.

The following spectroscopic datasets were obtained for further analysis: Field A (dry spectra), Field B (dry spectra), and Field C (wet spectra and field spectra). Additionally, the robustness and parsimoniousness of the models to predict SOC with different spectral datasets in a combined form (e.g., regional dataset) were also examined using two options. For option 1, spectra from Field A, Field B, and Field C (field spectra) were merged to form a single dataset: three field merged data [Field (A+B+C (field spectra)]. For option 2, dry spectra datasets were also merged together to form a single dataset: two field merged data (Field A+ Field B).

2.7. Modelling development and performance

To ensure the results were not dependent on the multivariate model, four different multivariate techniques were evaluated separately, namely Cubist, support vector machine regression (SVMR), partial least squares regression (PLSR) and random forest (RF). The Cubist method was used to calibrate the regression tree models using the train function of the caret package in R. Cubist uses linear regression models at each node instead of the average. To avoid overfitting (Kuhn and Johnson, 2013), the default number of committees (1, 10 and 20) and neighbours (0, 5, and 9) from the train function were utilised. The root mean square error (RMSE) was used to select the best models. Comparably, the SVMR is tuned to different cost parameters with the built-in tuning function using the grid search (precisely 0.001, 0.01, 0.1 and 1) with a linear kernel function while the epsilon parameter is left to its default value (0.1). The Package e1071 library in R was used. Based on the RMSE, the best cost parameter is selected from bootstrap results based on 10-fold cross-validation. For the PLSR algorithm, a set of new predictor variables identified as latent variables is developed as a linear combination of the initial predictor variables. The model runs and tests itself for each number of components, i.e., from 1 to 10 (the maximum number of model components was set to 10). The optimum number of components is selected based on the lowest RMSE. With the optimal number of components obtained, the model is re-calibrated and validated, and the coefficient of determination (R^2) and the RMSE are computed.

Finally, the RF algorithm is formulated to reduce experimental noise and improve prediction accuracy (Liaw and Wiener, 2002). The dataset under consideration is randomly divided into numerous training sets, and decision trees are developed using bootstrap re-sampling capabilities. The average of the individual tree outputs is then utilised to calculate the final prediction. The Random Forest R package was used, which includes homonymous (Liaw and Wiener, 2002). This is founded on the principles of Leo Breiman and Adele Cutler's Fortran code. A total of 500 trees were grown, with 35 variables randomly selected as candidates at each split. The R programming language (R Development Core Team, 2015, Vienna, Austria) was used for spectra pre-processing and modelling techniques.

The model's output was assessed by five-fold cross-validation for each regression procedure of the calibration (75%) and validation set (25%) of the samples using Cubist, SVMR, PLSR and RF modelling techniques. The accuracy of the prediction was assessed based on the coefficient of determination

(R^2_{CV}), the ratio of performance to interquartile range (RPIQ), the ratio of performance to deviation (RPD), which is the ratio of a parameter's standard deviation to the standard error of that parameter's prediction by a specific model, the root mean square error of prediction (RMSEPcv) (measures the overall model prediction accuracy) of the 5-folds cross-validation and the bias. The bias represents the error of means and is independent. The R^2_{CV} ranges from 0 to 1, where $R^2_{CV} = 1$ is the optimal value. For the RPD, Chang and Laird's (2002) categorisation was applied: RPD > 2 indicates good models, RPD between 1.4 and 2 indicates moderate predictive ability, and RPD = 1.4 indicates weak models. The five-fold cross-validation was repeated 100 times to ensure model stability and reliability.

2.8. The ensemble predictive model's construction

The four tested calibration techniques were combined together in order to build the ensemble model, which in turn is actually a weighted average of four individual predictions (one for each calibration technique). To ensure that the best possible result was achieved, all permissible combinations of weights for individual predictions were tested. To do so, an automated procedure was utilised that proceeds in a way that a regular sequence (ranging from 0 to 1 by 0.05) of desired values is first created, for which subsequently all possible permutations of the length of four are computed. To ensure that the assigned weights sum up to one, only permutations whose sum is equal to one were taken for further calculations. Next, the four individual cross-validation results were repeatedly weighted averaged using the valid permutations one by one (there were as many runs as there were the number of valid permutations) and the respective validation statistics (weighted average vs. observed values) were calculated and recorded. Note that the individual cross-validations are composed of an average of one hundred individual runs. Finally, the set of weights that correspond to the best validation statistics (the lowest RMSE) form the ensemble model.

Furthermore, only those predictions that were achieved using the best pre-treatment method were used to build the ensemble model. To test the predictive performance of the ensemble model, the validation statistics were compared to those achieved with the four individual techniques for all datasets used (field spectra, combined dataset approach, and the Sentinel-2 data). Figure 2 schematically displays the experimental design.

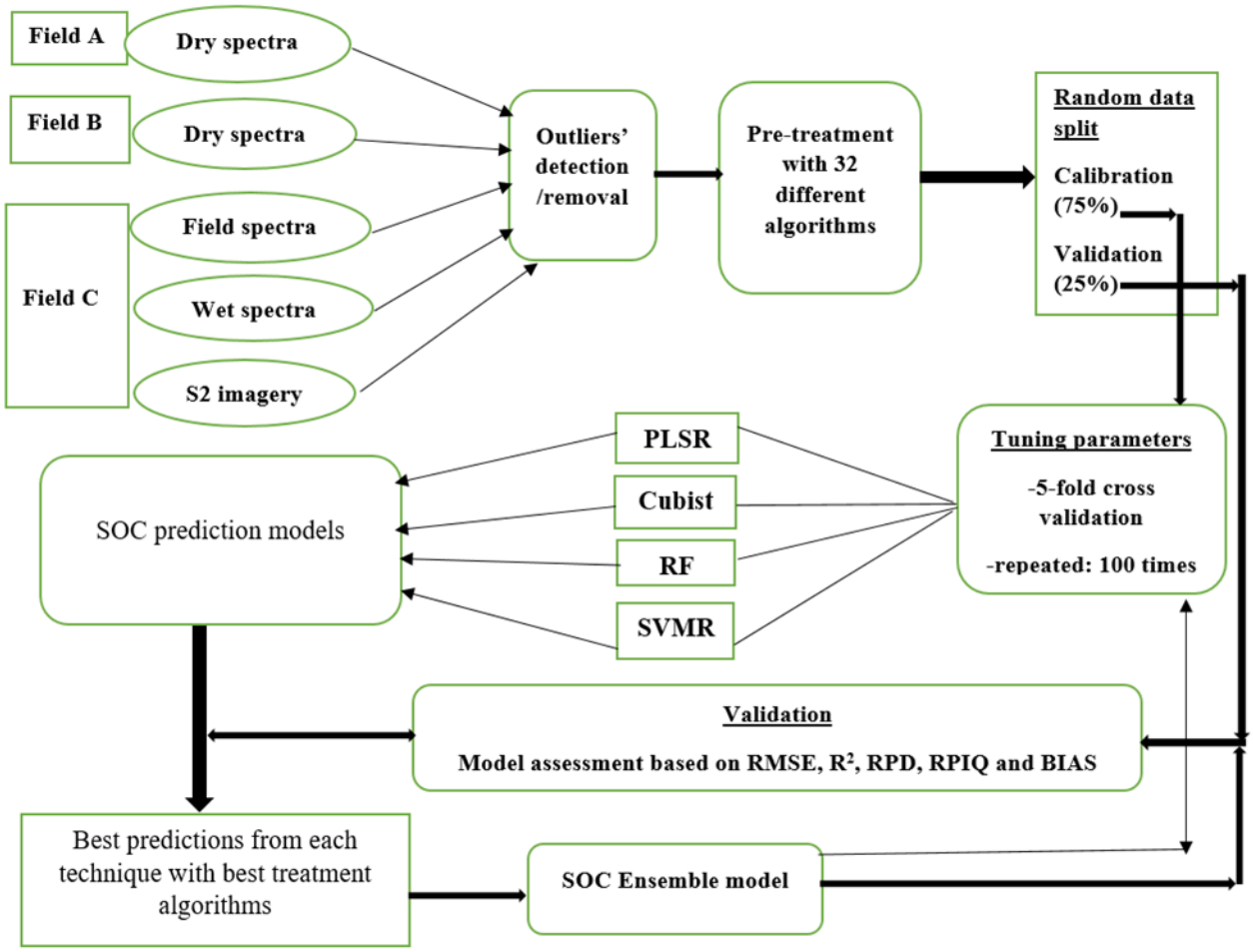


Figure 2. Schematic diagram illustrating the experimental design. [PLSR (partial least squares regression), RF (random forest), SVMR (support vector machine regression), S2 (Sentinel-2), Field A (Vidim), Field B (Bromovice), Field C (Nová Ves nad Popelkou)]

3. Results

3.1 Soil organic carbon (SOC) descriptive statistics

The descriptive statistics (Table 1) show the results of SOC analysis within the three study fields for the whole calibration and validation data sets, comprising the mean, median, minimum (Min), maximum (Max), standard deviation (SD), coefficient of variation (CV), and skewness. For Field A, the calibration data CV was the same as the whole data, whereas for Field B, both the whole and validation data were also the same. Generally, the study fields (using the whole data set) were significantly different in SOC content. For instance, the lowest mean content of 1.02% was observed in Field A, while Field C obtained

300 the highest mean content of 1.44%. Furthermore, skewness was used to evaluate the normality of SOC
301 content. It was discovered that all of the study fields' SOC content was presumed to be normally
302 distributed, with skewness values close to 0.8, particularly for Field B. The SOC content for these fields
303 could be described as medium to semi-high.

304 Table 1: Descriptive statistics of soil organic carbon (SOC%) contents at the three study fields

Fields (Samples)	n[a]	Mean	Median	SD[b]	Skewness	Range	Min[c]	Max[d]	CV%[e]
Field A (dry) (WD[f])	76	1.02	0.98	0.19	0.59	0.94	0.66	1.6	19.00
Calibration data	57.00	1.04	1.04	0.20	0.38	0.94	0.66	1.60	19.00
Validation data	19.00	0.93	0.90	0.11	0.54	0.38	0.77	1.15	12.00
Field B (dry) (WD)	97	1.06	0.99	0.32	0.79	1.48	0.5	1.98	30.00
Calibration data	73.00	0.98	0.95	0.25	0.57	1.08	0.50	1.58	26.00
Validation data	24.00	1.30	1.37	0.39	0.07	1.29	0.69	1.98	30.00
Field C (field & wet) (WD)	130	1.44	1.44	0.33	0.57	2.33	0.6	2.93	23.00
Calibration data	97.00	1.45	1.47	0.35	0.62	2.33	0.60	2.93	24.00
Validation data	33.00	1.42	1.43	0.29	0.15	1.39	0.71	2.10	20.00

305

306 [a]n: number of samples

307 [b]SD: standard deviation

308 [c]Min: minimum

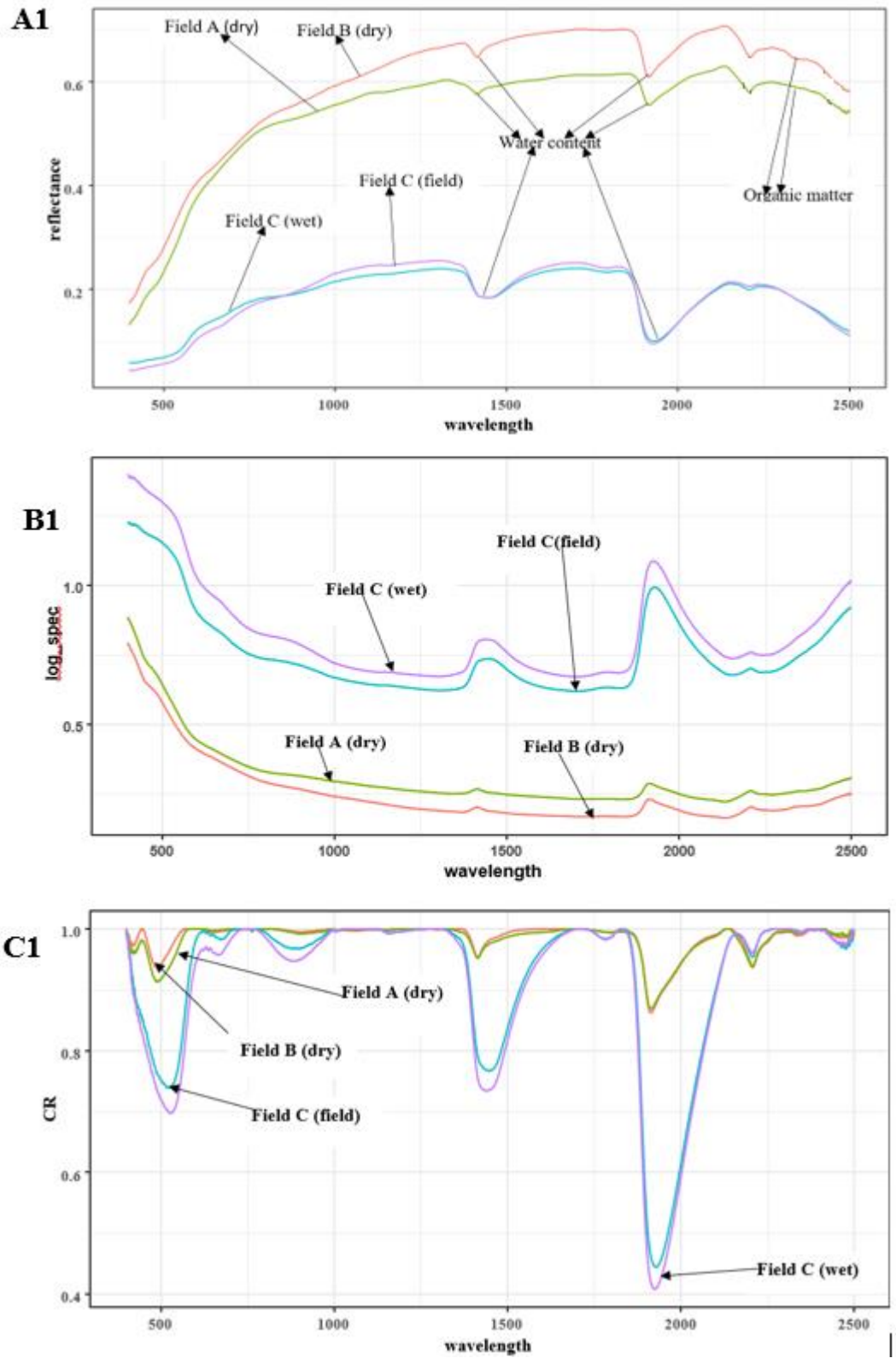
309 [d]Max: maximum

310 [e]CV: coefficient of variation

311 [f]WD: whole data set

312

Figure 3 (B1 & C1) displays the effect of using logarithmic transformation and CR algorithms on the raw spectra data under different measurement conditions (field, wet, and dry), whereas Figure 3 (A1) shows only the raw spectra without transformation. The CR algorithm is commonly used to compare different spectra measurements because of its ability to scale spectra to unity. The CR line in Figure 3 (C1) shows the specific field absorption along the spectra wavelengths can be attributed to the influence of H2O/OH and carbonate. According to Howari et al. (2002), absorption at 990 and 1400 nm is typically caused by NaCl and water molecular vibrations and OH groups. Additionally, absorption and reflection at bands usually centred at 1400, 1900, 2200, and 2365 nm are due to water and mineral influences (Soriano-Disla et al., 2014). This shows that Field C (wet) had the highest water content and that the spectra formed for Field B were slightly better than Field A, particularly in the 500 nm range. The log(1/R) algorithm transforms the reflectance into absorbance (Minasny et al., 2011), which helps detect the absorption characteristics' edges. The sequence of absorbance in terms of soil moisture is displayed in Figure 3 (B1) as Field C (wet) > Field C (field) > Field A > Field B. The reflectance, as shown in A1, was the exact opposite of B1. Correspondingly, on display was organic matter with wavelengths ranging from 2000 to 2500 nm (Ben-Dor et al., 1999).



328

329 Figures 3: Spectra showing raw data (A1), absorbance features (B1), and continuum removal plot (C1)

330 for the three fields under difference spectra measurement conditions.

331

3.3. Evaluating the predictive performance of calibration methods for all datasets

3.3.1 Individual modelling algorithm.

The assessment of SOC with individual calibration techniques and the pre-treatment algorithms (different combinations) shows the model prediction accuracy differs significantly from one field to the other, while Sentinel-2 data (only from Field C) provided the worst results (Tables 1&2). Considering Field A (dry spectra) (Table 1), both PLSR and SVMR yielded almost the same results with R^2_{CV} = 0.64/0.63 and RMSEP = 0.12/0.13 when compared to the other two individual techniques (Cubist and RF), which provided poor results. However, there was a slight difference between the RPD (1.66/1.59) and RPIQ values (2.44/2.35) favouring PLSR. Furthermore, for Field B [(dry spectra) (Table 2)], the order of individual prediction accuracy was SVMR (sg_msc) > PLSR (cmr) > RF (dwt_cr) > cubist (sg_log), with R^2_{CV} values ranging from 0.84 to 0.89. For Field C (Table 2), SVMR (R^2_{CV}= 0.49, $RMSEP_{CV}$ = 0.23, RPD = 1.37, RPIQ = 1.51) provided the most appropriate result, followed by PLSR for both the wet and field spectra dataset. Additionally, based on the obtained RPD values, the result demonstrates a poor model in which only high and low values are observable using Chang and Laird's (2002) categorisation.

3.3.2 Combined data set and Sentinel-2 imagery

For the combined data using the individual calibration techniques and treatment algorithms, the results (Table 2) indicated that a more accurate prediction was achieved with option 2 [Field A (dry spectra) + Field B (dry spectra) (R^2_{CV} = 0.79, $RMSEP_{CV}$ = 0.13, RPD = 2.16, RPIQ = 2.79)] compared to option 1 [Field A (dry spectra) + Field B (dry spectra) + Field C (field spectra)] (R2CV= 0.71, $RMSEP_{CV}$= 0.20, RPD = 1.78, RPIQ = 2.69). Nonetheless, the best result for both options was realised with the same calibration technique (SVMR) and treatment algorithm (log).

3.3.3 Evaluating the predictive performance of the Ensemble Model

The predictive performance for SOC with the ensemble model using all the datasets shows some improvement compared to the individual calibration techniques (Tables 1&2). However, the margin of increase in prediction accuracy varied among each dataset. The Field B dataset for instance, provided

the overall best results (R^2_{CV} = 0.92, RMSEPcv = 0.10, RPD = 3.06, RPIQ = 3.74, BIAS = 0.0067), while the Sentinel-2 dataset (Field C) obtained the worst results (R^2_{CV} = 0.27, RMSEPcv = 0.29, RPD = 1.2, RPIQ = 1.31, BIAS = -0.004). For the Sentinel-2 dataset, this was a slight improvement compared to the result achieved by the individual techniques. The prediction accuracy of SOC obtained from the combined dataset (option 2) using the ensemble was better than the individual techniques and was more appropriate than option 1.

Table 2: Prediction performance showing statistics of the five-fold leave-group-out cross-validation for four spectra measurement at three distinct areas: Field A (dry spectra in the lab), Field B (dry spectra in the lab), Field C (wet spectra in the lab), and Field C (field spectra on the field) using PLSR (partial least squares regression), RF (random forest), Cubist, SVMR (support vector machine regression) and the ensemble model (combination of all four models) with several pre-treatment algorithms combination: Raw (initial spectrum), Savitzky–Golay (sg), discrete wavelet transformation (dwt), first derivative (d1), second derivative (d2), multiplicative scatter correction (msc), standard normal variate (snv), log transformed (log), continuum removal (cr), maximum reflectance correction (cmr)

Models	Best pre-treatment	R^2_{CV}	RMSEPcv	RPD	RPIQ	BIAS
	Field A (dry soil sample)					
PLSR	log_msc	0.64	0.12	1.66	2.44	-0.0008
Cubist	sg_log_msc	0.48	0.16	1.42	2.08	-0.0080
RF	sg_log_snv	0.48	0.15	1.40	2.06	-0.0083
SVMR	sg_snv	0.63	0.13	1.59	2.35	-0.0069
Ensemble	log_msc+sg_log_msc+	0.70	0.12	1.88	2.66	0.0034

	sg_log_snv+sg_snv +					
	with all four models					
	Field B (dry soil sample)					
PLSR	sg_msc	0.86	0.13	2.66	3.24	0.0052
Cubist	cmr	0.84	0.14	2.50	3.04	0.0098
RF	dwt_cr	0.85	0.13	2.54	3.09	-0.0033
SVMR	sg_log	0.89	0.12	2.97	3.65	0.0102
Ensemble	sg_msc+cmr+dwt_cr+sg_log+	0.92	0.10	3.06	3.74	0.0067
	with all four models					
	Field C (field soil sample)					
PLSR	dwt_log	0.41	0.26	1.30	1.49	0.0009
Cubist	log	0.34	0.27	1.24	1.42	0.0115
RF	sg_d1	0.31	0.28	1.21	1.38	0.0054
SVMR	log	0.47	0.25	1.31	1.50	-0.0047
Ensemble	dwt_log+log+sg_d1+log+	0.50	0.25	1.39	1.59	-0.0132
	with all four models					

	Field C (wet soil sample)					
PLSR	dwt_log_msc	0.37	0.27	1.25	1.43	-0.0022
Cubist	sg_log_msc	0.31	0.28	1.20	1.37	-0.0106
RF	dwt_cr	0.30	0.29	1.18	1.35	0.0081
SVMR	log_msc	0.49	0.25	1.37	1.51	-0.0238
Ensemble	dwt_log_msc+sg_log_msc+ dwt_cr+log_msc+ with all four models	0.54	0.24	1.42	1.63	-0.0103

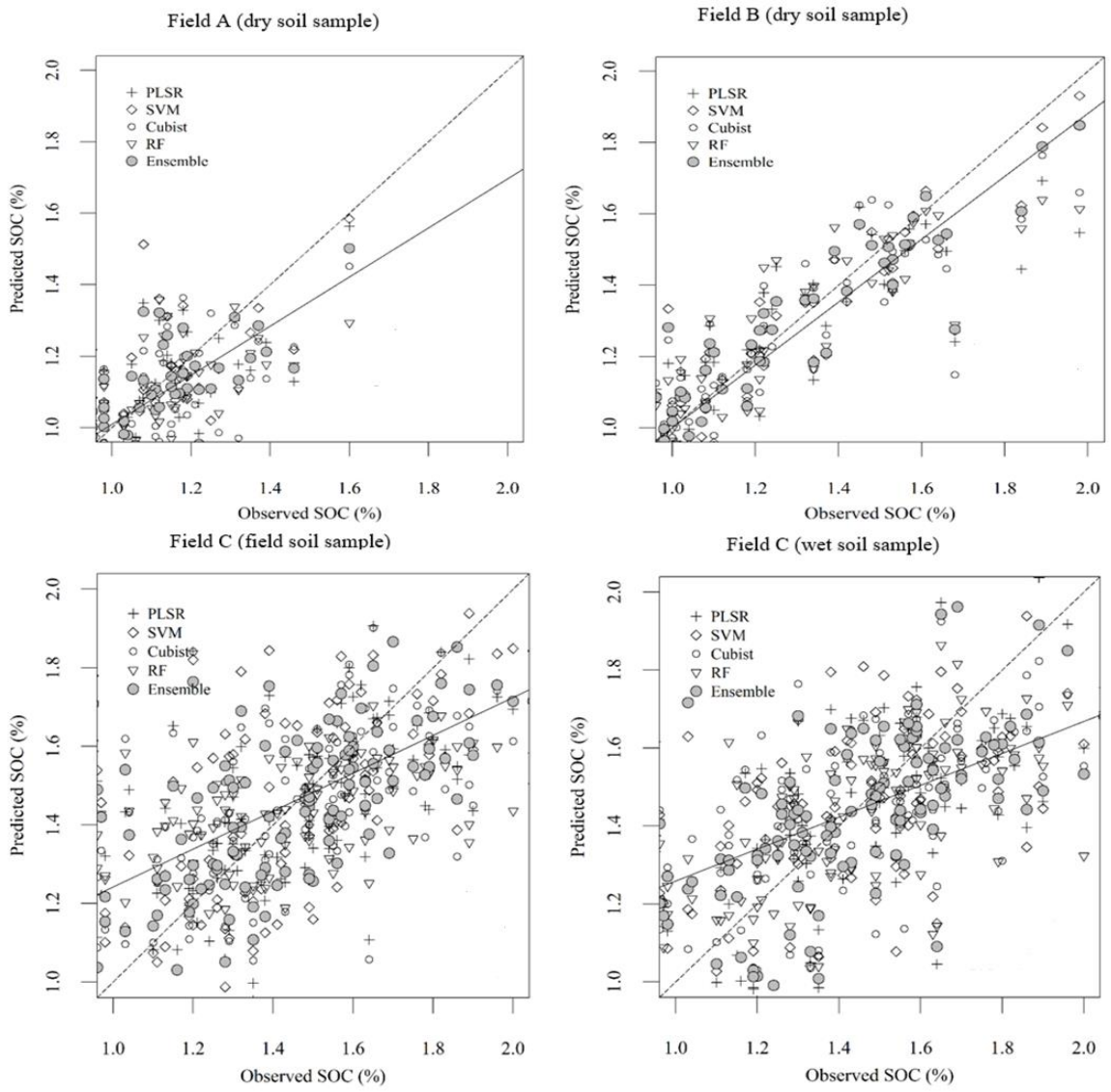
372

373 Table 3: Prediction performance showing statistics of the five-fold leave-group-out cross-validation for
374 three (A+B+C) and two (A+B) field combined using PLSR, RF, Cubist, SVMR (support vector machine
375 regression) and the ensemble model with several pre-treatment algorithms combination: Raw (initial
376 spectrum), Savitzky–Golay (sg), discrete wavelet transformation (dwt), first derivative (d1), second
377 derivative (d2), multiplicative scatter correction (msc), standard normal variate (snv), log transformed
378 (log), continuum removal (cr), maximum reflectance correction (cmr).

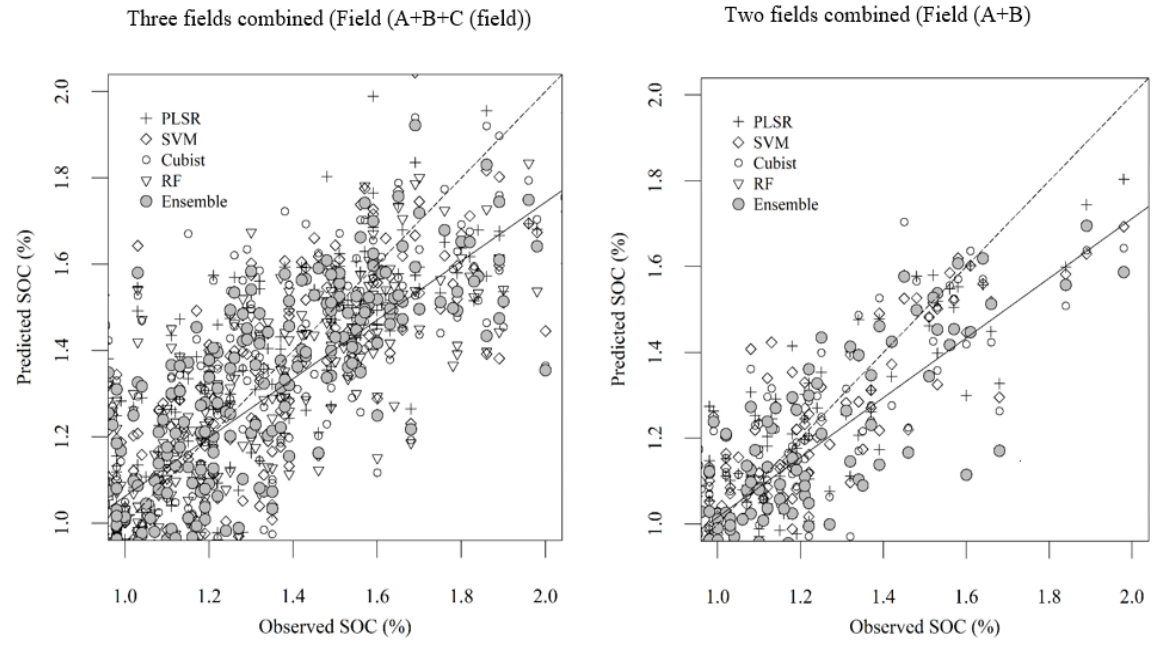
Models	Best pre-treatment	R^2cv	RMSEPcv	RPD	RPIQ	BIAS
	Three fields combined (A+B+C)					
PLSR	sg_log	0.66	0.21	1.71	2.59	0.0009
Cubist	log	0.68	0.21	1.73	2.61	0.0036
RF	cr	0.66	0.21	1.71	2.59	-0.0011
SVMR	log	0.71	0.20	1.78	2.69	0.0101
Ensemble	sg_log+log+cr+log with all four models	0.75	0.17	1.91	2.76	0.0045

	Two fields combined (A+B)					
PLSR	sg_d1	0.76	0.13	2.04	2.48	0.0041
Cubist	dwt_msc	0.78	0.13	2.13	2.59	-0.0024
RF	sg_d1	0.75	0.14	1.98	2.40	-0.0031
SVMR	log	0.79	0.13	2.16	2.79	0.0098
Ensemble	sg_di+dwt_msc+sg_di+log with all four models	0.83	0.12	2.30	2.96	0.0065

The scatterplots (Figure 4) show the results of predicted versus observed contents of SOC predictive accuracy using the different spectra measurement conditions dataset (field, wet and dry), the merged data options and the Sentinel-2 imagery. In this study, the discrepancies in predicting SOC across different datasets and measurement conditions vary on all R^2_{CV}, RMSEPcv, RPD, and RPIQ model performance indicators. None of the SOC predictions resembled a 1:1 fit, especially for Field C (both wet and field), the combined dataset option 1 (field dry) and the Sentinel-2 data.


Field A (dry soil sample)
Field B (dry soil sample)
Field C (field soil sample)
Field C (wet soil sample)
PLSR
SVM
Cubist
RF
Ensemble
Predicted SOC (%)
Observed SOC (%)
1.0
1.2
1.4
1.6
1.8
2.0

386


Three fields combined (Field (A+B+C (field))
Two fields combined (Field (A+B)
PLSR
SVM
Cubist
RF
Ensemble
Predicted SOC (%)
Observed SOC (%)
1.0
1.2
1.4
1.6
1.8
2.0

387

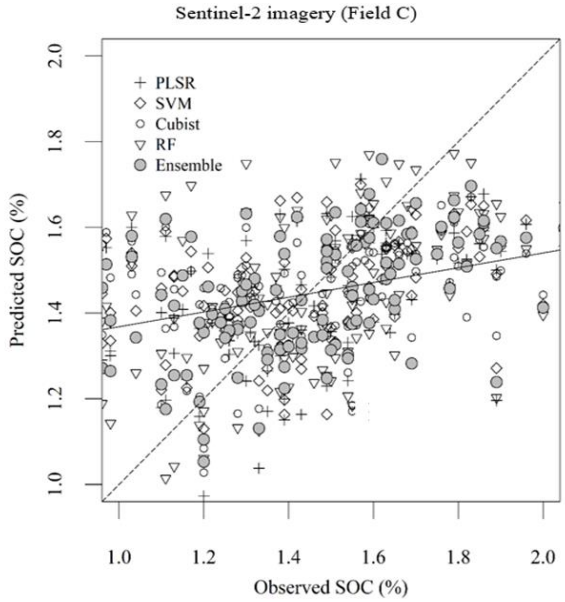


Figure 4: Observed vs predicted SOC content (%) values for each of the four best calibrations model (PLSR, SVM, MLR, and RF) as well as the final ensemble calibration model for (A2): three different fields (four measured spectra), (B2): for combined data option 1 (A+B+C (field)) and combined data option 2 (A+B), with least-squares fit (solid) and hypothetical, optimal fit (dashed).

4. Discussion

4.1. Comparison of the individual modelling techniques for all datasets

The inconsistency in the predictive performance of the four individual techniques (PLSR, RF, Cubist and SVMR) (Table 2) could be attributable to the fact that accuracy assessment of SOC models can vary significantly from one field to another depending on the spectral measurement conditions and the signal pre-treatment procedures used (Mishra et al., 2020). Furthermore, the validation assessment reveals SVMR as the most reliable method for this study. In comparison to the other individual techniques, it provides the best prediction result of SOC [e.g., Field B (R^2_{CV} = 0.89)] on almost all datasets except Field A, where the PLSR result is slightly better (R^2_{CV} = 0.64); however, the difference between PLSR and SVMR (R^2_{CV} = 0.63) for the said field could be considered negligible (difference of only 0.01). It is worth mentioning that, in some instances (Table 2), the prediction accuracy of SVMR and the other individual techniques (PLSR, RF, and Cubist) were nearly comparable, particularly for the Field B

406 sample. However, in other instances, the results obtained for all the individual techniques were not
407 favourable. A specific example is datasets from Field C [(field or wet spectra) (R^2_{CV} = 0.49) and
408 Sentinel-2 imagery (R^2_{CV} = 0.21)], where all individual techniques fail to explain at least 0.5% of the
409 variance in SOC. This implies that the model's inputs failed to explain nearly half of the observed
410 variation.

411 Large RMSE values suggest that the predicted and true responses vary substantially, whereas a small
412 RMSE suggests that the predicted and true responses are very close. According to Ben-Dor et al. (2015),
413 large or small RMSE values could result from differences in measurement protocols and the use of
414 sample techniques under different conditions. For the current study, the RMSEPcv, as well as the other
415 evaluation criteria (RPD, RPIQ and BIAS) results using the modeling techniques separately (Table 1)
416 show that SVMR obtained the lowest error margin and better assessment metrics (either better or slightly
417 better) than the other techniques, except for Field A where PLSR provided slightly improved results.

418 Despite this, this study highlights SVMR's robustness in identifying the existence of correlated and
419 outdated variables. For example, in a study comparing SVMR with other multivariate techniques (e.g.,
420 PLSR) to predict SOC contents, Viscarra Rossel and Behrens (2010) found that SVMR yielded the
421 lowest RMSEs in the prediction of SOC contents at all Vis–NIR wavelengths. Additionally, according
422 to Thissen et al. (2004), SVMR models are less susceptible to noise and outliers. Although SVMR is
423 noted to approximate and enhance a non-linear structure among multidimensional spaces (Stevens et al.,
424 2010), its positive results can also presumably be attributable to the fact that there are both linear and
425 non-linear correlations between the many spatial variables that SVMR could effectively evaluate. The
426 superiority of SVMR has also been established in many other studies (e.g., Xu et al., 2018; Lucà et al.,
427 2017; Kuang et al., 2015). Among the individual techniques, RF was considered the least well-
428 performing algorithm. This further affirms its poor predictive ability, as previously reported, e.g., by
429 Viscarra Rossel and Behrens (2010).

430 *4.2. The impact of spectra pre-processing algorithms on data sets*

The accuracy of SOC prediction may also be influenced by numerous factors of uncertainty under different conditions during data collection and analysis resulting from different data collection mediums that exit [particularly proximal (spectroscopy) and remote (airborne, satellite, and space-borne) sensing approaches]. For example, temperature increases may result in a non-linear carbon deficit (Ciais et al., 2005; Reichstein et al., 2006). Sampling and laboratory biases also need to be considered, while samples from various surveys may differ significantly due to a range of factors (Lal et al., 2001; Neff et al., 2002; Ogle et al., 2006). Under certain conditions, the spectral characteristics pattern related to a given parameter during spectral measurement may overlap with the response pattern (e.g., SOC) associated with some other factor, causing the prediction of that given factor to be negatively impacted or other significant information masked out, both leading to a decrease in prediction accuracy. For this study, all three different spectroscopy measurement conditions were used (laboratory (wet and dry) and field) as well as space-borne imagery acquisition (Sentinel-2). Therefore, almost all the various disturbing factors that could influence the spectral measurement were anticipated (i.e., soil moisture, texture, noise, transient changes in weather conditions during measurement, illumination sources, etc.). Some of these factors can be visually seen using the reflectance plot (Figure 3 (A1)). Likewise, some of these effects on the spectra wavelength can also be seen in Figure 3 (B1 & C1) using the absorbance and the CR plots.

Similarly, the spatial, spectral, and temporal resolutions of remote sensing sensors vary, which could also affect the accuracy of SOC estimation. The use of the several pre-treatment algorithms employed before prediction helped to improve the applicable spectral features. According to Bowers and Hanks (1965), one of the major factors influencing spectral measurement is the effect of an increased moisture content. This can initiate the phenomenon of stretching and bending vibrations of water and hydroxyl bonds, which can negatively affect soil property predictive performance (Minasny et al. 2009; Bricklemyer and Brown 2010). Moreover, the baseline height, as well as other spectral attributes across the entire spectral range, may also be affected (Muller and Decamps, 2000). Additionally, the effect of soil moisture content and other constraints on spectral measurements for SOC prediction has also been

457 reported in several other studies (e.g., Minasny et al., 2011; Tekin et al., 2012; Nocita et al., 2013;
458 Wijewardane et al. 2016; de Santana et al. 2019).

459 The 32 (including different combinations) spectral treatments provided mixed results (better or worse,
460 the majority of the worse results are not shown). Because these spectral datasets were obtained under
461 different conditions, as already stated, they may contain artefacts indicating the presence of unwanted
462 variation (Mishra et al., 2020). These artefacts could result from but are not limited to the following:
463 instrumental drifts, measuring modality, state of sample and environmental influences (Roger et al.,
464 2020). According to Engel et al. (2013), no single method is universally adequate for all datasets, and
465 signal pre-treatment techniques can influence the entire analysis. For Field A, the best result using the
466 individual calibration techniques was achieved with log_msc (R^2_{CV} = 0.64), indicating that linearisation
467 attainment between the spectra and SOC content and light scattering effects were some of the dominant
468 artefacts affecting this dataset. Furthermore, the second-best result was obtained with the sg_snv
469 algorithm (R^2_{CV} = 0.63), suggesting that the light scattering effect was also one of the spectral defects.
470 However, because the prediction accuracy was modest, the possibility that a critical section was masked
471 by other components or removed by the treatment algorithm application cannot be ruled out. This is
472 because failing to select the most appropriate pre-treatment algorithm may result in the removal of
473 relevant information related to the property of interest (Engel et al., 2013; Oliveri et al., 2019).
474 According to Wold et al. (1998), both SNV and MSC have the potential to remove vital information
475 from the spectra if misused. On the other hand, for Field B, the most relevant result (also with the
476 individual techniques) was achieved with the sg_log algorithm (R^2_{CV} = 0.89), signifying that the
477 attainment of linearisation between the spectra and SOC content was presumably effective. Moreover,
478 the log approach has also been shown to remove baseline effects to enhance spectral characteristics,
479 resulting in enhanced prediction accuracy (Ben-Dor et al., 1997; Schlerf et al., 2010). The results
480 obtained with the other calibration techniques (Table 2) presumably show the rectification of other
481 abnormalities, e.g., the second-best result was achieved with the sg_msc algorithm (R^2_{CV} = 0.86) (light
482 scatter effect), and the third-best result was achieved with the dwt_cr algorithm (R^2_{CV} = 0.85) (resolving
483 overlapping bands and unwanted background noise as well as vertical offset and/or slope effect). This

484 study, therefore, agrees with the study of Mishra et al. (2020), which stated that pre-treatment could
485 improve spectral quality by removing undesired effects from a dataset. However, due to the constraints
486 of different techniques and the intricacy of these undesirable effects, a single pre-treatment procedure
487 may be unable to entirely eliminate all of these deficiencies. As a result, using a combination of diverse
488 pre-treatment methods is one of the most effective options. Although the log_msc algorithm provided
489 the best results for Field C, the prediction accuracy was, conversely, not encouraging (R^2_{CV} = 0.49). This
490 indicates that the effect of soil moisture content (Figure 3) was likely dominant, as the spectra used for
491 this field were spectra measured under field and wet conditions: the soil sample had not been dried to
492 reduce the amount of water in the soil. This highlights the challenge for pre-treatments to eliminate the
493 influence of water content on Vis-NIR spectra measured in the field or under wet conditions. Although,
494 the external parameter orthogonalization (EPO) algorithm was utilised by Minasny et al. (2011) to
495 remove the effect of soil water content from Vis-NIR spectra for SOC content calibration, however, in
496 order to develop EPO pre-processed dataset, prior knowledge of soil water content information is
497 needed.

498 *4.3. Ensemble model predictive performance of SOC*

499 *4.3.1. Using the single data set*

500 The ensemble model (a merger of SVMR, PLSR, Cubist and RF) predicted SOC more accurately than
501 any of the individual models as mentioned earlier. The model significantly improves prediction accuracy
502 certainty by minimising the limitations and drawbacks of each model while maximising the combined
503 models' responses. (Kalantar et al., 2018; Arabameri et al., 2019; Martre et al., 2015). The ensemble
504 model's good performance was achieved by selecting the best pre-treatment algorithms for each
505 technique to aid in the correction of a variety of artefacts that the individual techniques might have failed
506 to address (Mishra et al., 2020). The ensemble model not only improved the prediction accuracy of SOC
507 (from R^2_{CV} = 0.89 to 0.92), but also reduced the error margin (from RMSEPcv = 0.12 to 0.10). Other
508 assessment parameters of the ensemble model (RPD, RPIQ, and BIAS) were also improved when
509 compared to the individual techniques. Furthermore, for Field C, the reason why the prediction accuracy
510 for its two spectra was not the same and lower in comparison to the other two study fields could be

attributed to the mode of spectral measurement. For instance, while the field spectral measurement was taken under uncontrolled environmental conditions prone to several environmental factors as stated above, the wet spectra were measured under a controlled laboratory environment, meaning its spectra data was influenced mainly by soil moisture content, as shown in Figure 4 because the measurement was taken on the soil in its wet state. The other fields performed better, most likely due to lower moisture content (measured using a laboratory regulated dry method) and other effects mitigated by the pre-treatment combination techniques employed. Under laboratory measurement conditions, standardised protocols are followed, which aid in removing unwanted factors (Romero et al., 2018; Ben Dor et al., 2015). Despite these effects, the ensemble model improved the accuracy of prediction of SOC for Field C (using the wet spectral) from R^2_{CV} = 0.49 (best for the individual techniques) to R^2_{CV} = 0.54 (ensemble), while other assessment parameters also improved accordingly in addition to the reduction in error (Table 2).

4.3.2. Combined data sets

For the merged dataset (option 1), the ensemble model shows its superiority over each of the individual calibration techniques by improving the prediction accuracy of SOC for each option (Table 3). However, the differences in accuracy between these options could be attributable to the influence of moisture content and probably other environmental and stochastic factors. The addition of the field spectra (from Field C) to the two dry spectra (Fields A and B) might have introduced some new defects to the other variations that already existed in the field dataset during measurement (e.g., moisture, etc). This is because the accuracy of the SOC predictive performance increased with option 2 (without the Field C dataset). This is in agreement with Waiser et al. (2007) who found that fields with high surface moisture content could influence prediction accuracy. Furthermore, as per Minasny et al. (2011), the presence of water within the soil would have a significant, intricate, and non-linear impact on the reflective spectra, potentially affecting prediction accuracy. This implies that our ensemble model could be used for the prediction of a merged dataset measured under different spectra conditions (dry, wet and field) or different soil types (e.g., regional dataset). However, for improved accuracy, the merger or combination

of these soil spectra data should be done with spectra obtained under the same spectra measurement conditions (dry spectra samples only, wet spectra samples only or field spectra samples only).

4.3.3. Using the satellite dataset

Finally, even though the ensemble model outperformed all individual calibration techniques using the Sentinel-2 dataset (an improvement of 0.06) (Table 3), it was nevertheless deemed unsuccessful due to its poor final predictive performance (R^2_{CV} = 0.27, RMSEPcv = 0.29, RPD = 1.20, RPIQ = 1.31, BIAS = -0.0154). The influence of other factors or occurrences cannot be ruled out. All the models utilised, including the individual techniques and signal pre-treatment approaches, failed to yield satisfactory results on this dataset (Sentinel-2). Vegetation cover on the surface of the soil is one of the major disrupting environmental factors that normally negatively influence the predictive performance of Sentinel-2 imagery for estimating soil properties, including SOC. Although our sampling period was in May, to obtain a cloud-free image, the downloaded Sentinel-2 imagery was an image dated June 20, a possible time during which some vegetation in the field is anticipated to emerge. The presence of this vegetation [e.g., green or dry vegetation (> 20%)] can significantly alter the form of spectral reflection, potentially affecting the prediction accuracy of soil properties (Bartholomew et al., 2011; Castaldi et al., 2019). Perhaps downloading more images to investigate the optimum date and adjusting the field measurements could be a solution to the issue of obtaining Sentinel-2 imagery captured on bare soil. However, getting these agricultural fields at one's "optimum time" is highly dependent on the landowners' decisions. The above-mentioned issues of vegetation cover and some other occurrences could probably be some of the reasons for the Sentinel-2 imagery dataset's poor performance with the ensemble as well as the individual techniques. More research in this specific area could help ascertain the true predictive performance of Sentinel-2 data using the ensemble model, which has received limited attention from researchers.

4.4. Scatter plots comparison and outliers assessment

The scatterplots in Figure 3 highlight the disparities pattern of the ensemble model and the modeling techniques for the spectra datasets and the Sentinel-2 imagery data. The order of disparity between the

measured and estimated SOC values under Field A (dry spectra), Field B (dry spectra) and Field C (field and wet spectra) was: Field C > Field A > Field B. This could be due to the sensor's spectral information content measured under different conditions and other parameters (Figure 3). According to Gomez et al. (2018), some of these disparities could be ascribed to light sources, instrumental noise, and environmental conditions (i.e., laboratory and field). For instance, Fields A and B may be affected differently by these defects, although both datasets were measured under laboratory-controlled conditions. For Field C, the field and wet plot were nearly identical, with minor changes that were not statistically significant. For the other two merged data, option 2 was better than option 1 because of the addition of the field spectra to option 1. Lastly, the measured and observed plot of Sentinel-2 data may be due to the lower predictive performance of satellite imagery which is often related to environmental conditions, spatial resolution, and the condition of the soil (Zhang and Zhou 2016; Steinberg et al., 2016).

It is also worth noting that the disparities between the measured and observed values (particularly Field C and the Sentinel-2 imagery) cannot be entirely attributed to the influence of outliers. Sometimes, keeping or removing outliers, particularly in regression models before prediction, could result in either a positive or negative outcome. Some of these outliers contain essential information about the data and could aid in prediction, so removing them may adversely affect prediction (Frost, 2019; Blatná, 2006). Figure 5 shows the outlier plot using enpls (Aggarwal, 2013) for spectroscopy and the Sentinel-2 dataset. Although there were several extreme points considered outliers in Field A, these extreme points were considered positive outliers containing vital information about the dataset because removing any of those points before prediction negatively affected the predictive performance of SOC for Field A, so they were not removed. The result improves by removing some of the extreme values in both Field B [points (p) 93 and 97] and Field C [wet (p 71,31,54,28)], field (p 71,23,35,48,105). However, due to the interesting response of this process for Sentinel-2 data, no outliers were removed to achieve a fair comparison. This is because while removing some of these extreme values improves the prediction accuracy for the individual techniques from R^2_{CV} = 0.21 to 0.25, it diminishes the ensemble model's accuracy (from R^2_{CV} = 0.27 to 0.24).

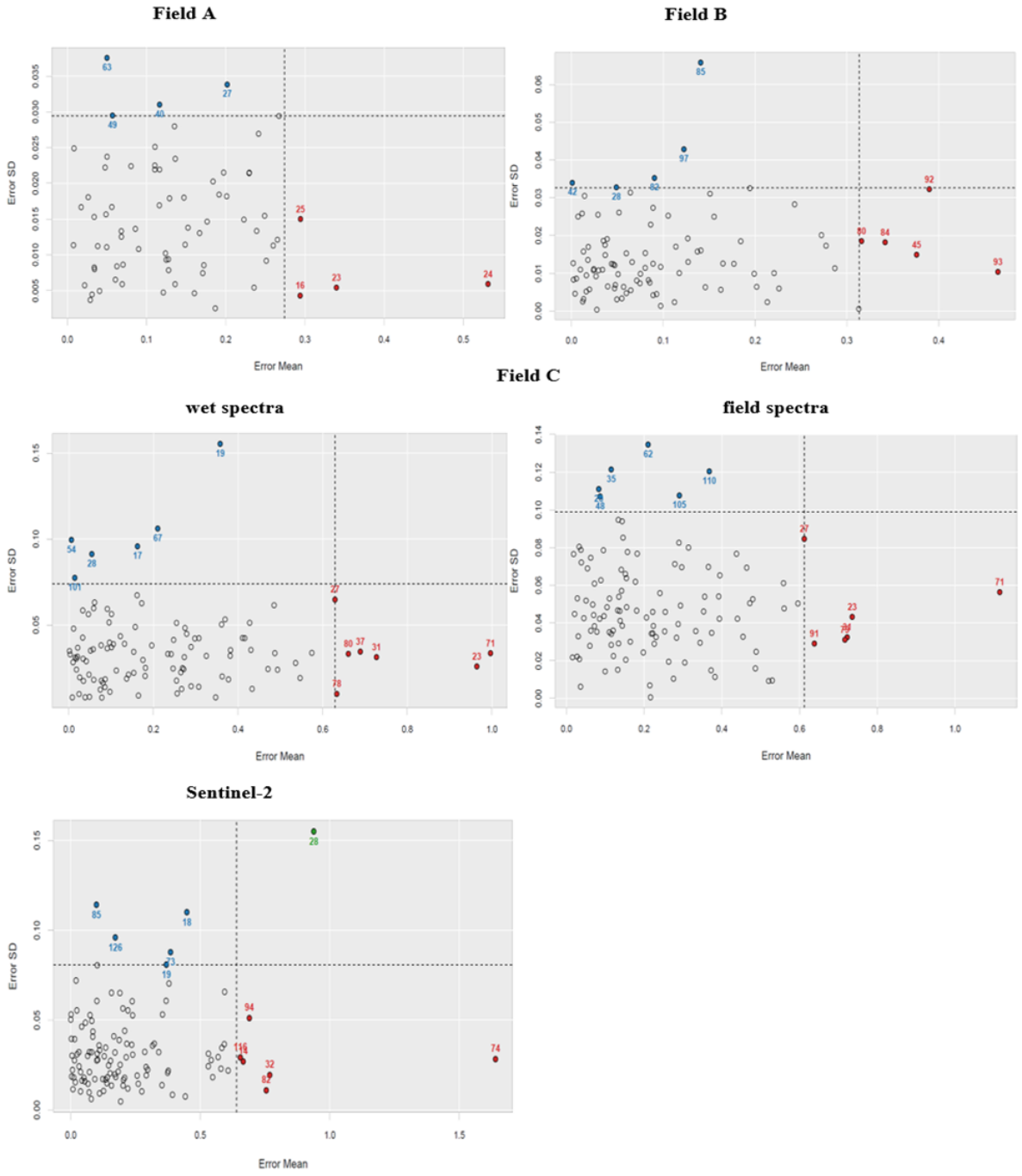


Figure 5: Outliers plot showing the mean and the standard deviation (SD) of the prediction error distribution for spectroscopy and Sentinel-2 datasets [Among the 4 regions, the lower left region, which occupies most off the data, is the normal samples that have a small mean value and SD. The upper left area is the x outliers, which have a small mean value but a large SD. Conversely, the lower right region is the y outliers, which have a large mean value but a small SD. The upper right region contains some of the extreme outliers or abnormal samples.

In summary, the ensemble model built for this study proved to be more robust and reliable than the individual modeling techniques. This has also been confirmed by other studies (e.g., Riggers et al., 2019; Tajik et al., 2020; Mishra et al., 2020). Previously, other studies focused on just one type of spectroscopy measurement condition for the estimation of SOC using the ensemble model approach. The resulting models were sometimes less applicable to other forms of spectroscopy measurement conditions as well as space-borne imagery data. For this study, the collected samples were from three separate sites with various soil types, and they were taken under the primary spectroscopy measurement conditions (wet, dry and field).

Additionally, satellite data and regional datasets (spectra data merged) were also tested. The focus was to enhance the precision of SOC prediction. A slight change in the SOC pool could considerably impact the global carbon cycle and, ultimately, climate change in general (Powlson et al., 2011). Therefore, it is necessary to develop models as shown in this study to estimate SOC accurately under different measurement conditions. Finally, there are many more ML algorithms and data mining tools that have yet to be explored in an ensemble model. Exploring these options is highly recommended. According to Martre et al. (2015), one of the benefits of using the ensemble model is its ability to compensate for errors across the models and provide better synchronisation of the model procedures.

5. Conclusion

The goal of the study was to compare the effectiveness of using an ensemble model against individual techniques forming the ensemble model to predict SOC using spectral data (field, wet, and dry), merged spectral data (regional), and Sentinel-2 data. This study confirmed that when different prediction techniques are combined to form an ensemble model using different calibration techniques, prediction and signal pre-treatment algorithms, the prediction accuracy was superior to any of the modeling techniques used individually. The ensemble model built for this study accurately captured the trend of all study fields as well as the various datasets gathered with a minor error and improved the prediction accuracy of SOC. Furthermore, it provides a more robust and reliable approach than each of the individual model estimates do alone. The findings demonstrated that the ensemble model could be an effective tool for reducing overall error in SOC modelling. It was also successful on almost all the data

obtained under different spectral measurement conditions with an order of dry > wet > field. The only exception was the accuracy of SOC prediction using Sentinel-2 data, which was low for the study field employed, likely due to numerous factors (e.g., cloud cover, vegetation) and constraints that affect the acquired Sentinel-2 imagery. This is because all the individual techniques also produced similarly poor results with the Sentinel-2 data. Nonetheless, future studies to verify the effectiveness of an ensemble model on remote sensing data are highly recommended, especially using remote sensing data with fewer defects.

Using the ensemble model on a regional dataset is highly feasible. However, to obtain a more accurate results of SOC prediction, the selected regional dataset should be made up of soil samples measured under the same spectroscopy conditions. Selection of the most appropriate treatment combination could be one option to eliminate or minimise several artefacts at the same time. The ensemble model demonstrates the ability to choose the best treatment algorithms for each dataset.

Funding.

This study was supported by an internal grant from the Czech University of Life Sciences Prague, project No. SV20-5-21130.

Acknowledgement:

The authors also acknowledge the support of the European Regional Development Fund Project Center for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products (No. CZ.02.1.01 / 0.0 / 0.0 / 16_019 / 0000845)

Compliance with Ethical Standards:

The authors declare no conflicts of interest.

References

649 Aggarwal, C. C. (2013). Outlier ensembles: position paper. ACM SIGKDD Explorations Newsletter,
650 14(2), 49-58.
651 Aldrich, E. (2013). Wavelets: A package of functions for computing wavelet filters, wavelet transforms
652 and multiresolution analyses. R package version 0.3-0. URL http://CRAN.R-
653 project.org/package=wavelets.
654 Althuwaynee, O. F., Pradhan, B., Park, H. J., & Lee, J. H. (2014). A novel ensemble decision tree-based
655 CHi-squared Automatic Interaction Detection (CHAID) and multivariate logistic regression models in
656 landslide susceptibility mapping. Landslides, 11(6), 1063-1078.
657 Arabameri, A., Pradhan, B., Rezaei, K., Sohrabi, M., & Kalantari, Z. (2019). GIS-based landslide
658 susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear
659 multivariate regression and boosted regression tree algorithms. Journal of Mountain Science, 16(3), 595-
660 618.
661 Balakrishnan, N. (1994). Order statistics from non-identical exponential random variables and some
662 applications. Computational statistics & data analysis, 18(2), 203-238.
663 Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., &
664 Tychon, B. (2011). Soil organic carbon mapping of partially vegetated agricultural fields with imaging
665 spectroscopy. International Journal of Applied Earth Observation and Geoinformation, 13(1), 81-88.
666 Ben Dor, E., Ong, C., & Lau, I. C. (2015). Reflectance measurements of soils in the laboratory:
667 Standards and protocols. Geoderma, 245, 112-124.
668 Ben-Dor, E., Inbar, Y., & Chen, Y. (1997). The reflectance spectra of organic matter in the visible near-
669 infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process.
670 Remote Sensing of Environment, 61(1), 1-15.
671 Ben-Dor, E., Irons, J. R., & Epema, G. F. (1999). Soil reflectance. Remote sensing for the earth sciences:
672 Manual of remote sensing, 3, 111-188.
673 Blatná, D. (2006). Outliers in regression. Trutnov, 30, 1-6.
674 Bowers, S.A., R.J. Hanks. (1965). Reflection of radiant energy from soils. Soil Sci., 100, 130-138
675 Bricklemyer, R. S., & Brown, D. J. (2010). On-the-go VisNIR: Potential and limitations for mapping
676 soil clay and organic carbon. Computers and Electronics in Agriculture, 70(1), 209-216.

677 Carmon, N., & Ben-Dor, E. (2017). An advanced analytical approach for spectral-based modelling of
678 soil properties. Int. J. Emerg. Technol. Adv. Eng, 7, 90-97.
679 Castaldi, Fabio, Andreas Hueni, Sabine Chabrillat, Kathrin Ward, Gabriele Buttafuoco, Bart Bomans,
680 Kristin Vreys, Maximilian Brell, and Bas van Wesemael. (2019). "Evaluating the capability of the
681 Sentinel 2 data for soil organic carbon prediction in croplands." ISPRS Journal of Photogrammetry and
682 Remote Sensing 147: 267-282
683 Chakraborty, P., Das, B. S., & Singh, R. (2017). An ensemble modeling approach for estimating
684 diffusive tortuosity for saturated soils from porosity. Soil Science, 182(2), 45-51.
685 Chang, C. W., & Laird, D. A. (2002). Near-infrared reflectance spectroscopic analysis of soil C and N.
686 Soil Science, 167(2), 110-116.
687 Chi, M., Kun, Q., Benediktsson, J. A., & Feng, R. (2009). Ensemble classification algorithm for
688 hyperspectral remote sensing data. IEEE Geoscience and Remote Sensing Letters, 6(4), 762-766.
689 Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., ... & Valentini, R. (2005). Europe-
690 wide reduction in primary productivity caused by the heat and drought in 2003. Nature, 437(7058), 529-
691 533.
692 de Santana, F. B., de Giuseppe, L. O., de Souza, A. M., & Poppi, R. J. (2019). Removing the moisture
693 effect in soil organic matter determination using NIR spectroscopy and PLSR with external parameter
694 orthogonalisation. Microchemical Journal, 145, 1094-1101.
695 Dietterich, T. G. (2002). Ensemble learning. The handbook of brain theory and neural networks, 2(1),
696 110-125.
697 Diks, C. G., & Vrugt, J. A. (2010). Comparison of point forecast accuracy of model averaging methods
698 in hydrologic applications. Stochastic Environmental Research and Risk Assessment, 24(6), 809-820.
699 Duckworth, J. (2004). Mathematical data preprocessing. In: Roberts, C.A., Workman Jr.J.,
700 Reeves IIIJ.B. (Eds.), Near-infrared Spectroscopy in Agriculture. ASA-CSSA-SSSA, Madison, WI.,
701 USA, pp. 115–132.
702 Elhag, M., & Bahrawi, J. A. (2017). Soil salinity mapping and hydrological drought indices assessment
703 in arid environments based on remote sensing techniques. Geoscientific Instrumentation, Methods and
704 Data Systems, 6(1), 149-158.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. (2013). Breaking with trends in pre-processing?. TrAC Trends in Analytical Chemistry, 50, 96-106.

Engler, R., Waser, L. T., Zimmermann, N. E., Schaub, M., Berdos, S., Ginzler, C., & Psomas, A. (2013). Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecology and Management, 310, 64-73.

European Space Agency. (2016). Sen2Cor 2.2.1-Software Release Note.

Farina, R., Sándor, R., Abdalla, M., Álvaro-Fuentes, J., Bechini, L., Bolinder, M.A., Brilli, L., Chenu, C., Clivot, H., Migliorati, M.D.A. & Di Bene, C. (2021). Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. Global Change Biology, 27(4), 904-928.

Frost. J. (2019). Guidelines for Removing and Handling Outliers in Data. https://statisticsbyjim.com/basics/remove-outliers/ (accessed on 23 October 2019).

Ge, Y., Morgan, C. L., Grunwald, S., Brown, D. J., & Sarkhot, D. V. (2011). Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. Geoderma, 161(3-4), 202-211.

Gholizadeh, A., Žižala, D., Saberioon, M., & Borůvka, L. (2018). Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sensing of Environment, 218, 89-103.

Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P., ... & Briottet, X. (2018). Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. Remote Sensing of Environment, 204, 18-30.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77

IUSS Working Group WRB (2014) World reference base for soil resources 2014. Edited by Schad P, van Huyssteen C, Micheli E. World Soil Resources Reports No. 106. FAO, Rome. 189 p. ISBN 978-92-5-108369-7

731 Jakšík, O., Kodešová, R., Kubiš, A., Stehlíková, I., Drábek, O., & Kapička, A. (2015). Soil aggregate
732 stability within morphologically diverse areas. Catena, 127, 287-299.

733 Jeong, G., Oeverdieck, H., Park, S. J., Huwe, B., & Ließ, M. (2017). Spatial soil nutrients prediction
734 using three supervised learning methods for assessment of land potentials in complex terrain. Catena,
735 154, 73-84.

736 Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., & Mansor, S. (2018). Assessment of the effects
737 of training data selection on the landslide susceptibility mapping: a comparison between support vector
738 machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Natural
739 Hazards and Risk, 9(1), 49-69.

740 Kalantar, B., Ueda, N., Saeidi, V., & Ahmadi, P. (2020). Application of machine learning algorithms
741 and their ensemble for landslide susceptibility mapping. In Workshop on World Landslide Forum,
742 Springer, Cham, 233-239.

743 Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and
744 partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic
745 carbon, pH and clay content. Soil and Tillage Research, 146, 243-252.

746 Kuhn, M., & Johnson, K. (2013). Applied predictive modeling New York: Springer, (Vol. 26, p. 13).

747 Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. science,
748 304(5677), 1623-1627.

749 Lal, R. A. T. T. A. N. (2001). Soil degradation by erosion. Land degradation & development, 12(6),
750 519-539.

751 Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for
752 soil organic carbon mapping and their implications: A review. Geoderma, 352, 395-413.

753 Lausch, A., Baade, J., Bannehr, L., Borg, E., Bumberger, J., Chabrilliat, S., Dietrich, P., Gerighausen,
754 H., Glässer, C., Hacker, J.M. and Haase, D. (2019). Linking remote sensing and geodiversity and their
755 traits relevant to biodiversity—part I: soil characteristics. Remote sensing, 11(20), 2356.

Li, G., Zheng, Y., Liu, J., Zhou, Z., Xu, C., Fang, X., & Yao, Q. (2021). An improved stacking ensemble learning-based sensor fault detection method for building energy systems using fault-discrimination information. Journal of Building Engineering, 102812.

Li, K., Tu, L., & Chai, L. (2020). Ensemble-model-based link prediction of complex networks. Computer Networks, 166, 106978.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma, 288, 175-183.

Ma, S., & Chu, F. (2019). Ensemble deep learning-based fault diagnosis of rotor bearing systems. Computers in industry, 105, 143-152.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., ... & Wolf, J. (2015). Multimodel ensembles of wheat growth: many models are better than one. Global change biology, 21(2), 911-925.

Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J. M., Gobrecht, A., Ferrand, L., & Joalland, S. (2011). Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma, 167, 118-124.

Minasny, B., McBratney, A. B., Pichon, L., Sun, W., & Short, M. G. (2009). Evaluating near infrared spectroscopy for field prediction of soil properties. Soil Research, 47(7), 664-673.

Mishra, U., Gautam, S., Riley, W., & Hoffman, F. M. (2020). Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. Frontiers in big Data, 3, 40.

Muller, E., & Decamps, H. (2001). Modeling soil moisture–reflectance. Remote sensing of Environment, 76(2), 173-180.

779 Neff, J. C., Townsend, A. R., Gleixner, G., Lehman, S. J., Turnbull, J., & Bowman, W. D. (2002).
780 Variable effects of nitrogen additions on the stability and turnover of soil carbon. Nature, 419(6910),
781 915-917.

782 Nikodem, A., Kodešová, R., Fér, M., & Klement, A. (2021a). Using scaling factors for characterizing
783 spatial and temporal variability of soil hydraulic properties of topsoils in areas heavily affected by soil
784 erosion. Journal of Hydrology, 593, 125897.

785 Nikodem, A., Kodešová, R., Fér, M., & Klement, A. (2021b). Variability of topsoil hydraulic
786 conductivity along the hillslope transects delineated in four areas strongly affected by soil erosion.
787 Journal of Hydrology and Hydromechanics, 69(2), 220-231.

788 Nocita, M., Stevens, A., Noon, C., & van Wesemael, B. (2013). Prediction of soil organic carbon for
789 different levels of soil moisture using Vis-NIR spectroscopy. Geoderma, 199, 37-42.

790 Ogle, S. M., Breidt, F. J., & Paustian, K. (2006). Bias and variance in model results associated with
791 spatial scaling of measurements for parameterisation in regional assessments. Global Change Biology,
792 12(3), 516-523.

793 Oliveri, P., Malegori, C., Simonetti, R., & Casale, M. (2019). The impact of signal pre-processing on
794 the final interpretation of analytical outcomes–A tutorial. Analytica chimica acta, 1058, 9-17.

795 Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review
796 aided by machine learning tools. Soil, 6(1), 35-52.

797 Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G. P., & Smith, P. (2016). Climate-smart soils.
798 Nature, 532(7597), 49-57.

799 Powlson, D. S., Whitmore, A. P., & Goulding, K. W. (2011). Soil carbon sequestration to mitigate
800 climate change: a critical re-examination to identify the true and the false. European Journal of Soil
801 Science, 62(1), 42-55.

802 R Development Core Team (2014). R: A Language and Environment for Statistical Computing. Vienna,
803 Austria: R Foundation for Statistical Computing.

804 Reichstein, M., Ciais, P., Papale, D., Valentini, R., Running, S., Viovy, N., ... & Zhao, M. (2007).
805 Reduction of ecosystem productivity and respiration during the European summer 2003 climate
806 anomaly: a joint flux tower, remote sensing and modelling analysis. Global Change Biology, 13(3), 634-
807 651.

808 Renka, R. J. (1996). Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation
809 package. ACM Transactions on Mathematical Software (TOMS), 22(1), 1-8.

810 Riggers, C., Poeplau, C., Don, A., Bamminger, C., Höper, H., & Dechow, R. (2019). Multi-model
811 ensemble improved the prediction of trends in soil organic carbon stocks in German croplands.
812 Geoderma, 345, 17-30.

813 Roger, J. M., Boulet, J. C., Zeaiter, M., & Rutledge, D. N. (2020). Pre-processing Methods. In
814 Comprehensive Chemometrics, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.:
815 Amsterdam, The Netherlands, Volume 3, 1–75

816 Romero, D. J., Ben-Dor, E., Demattê, J. A., e Souza, A. B., Vicente, L. E., Tavares, T. R., ... & Eitelwein,
817 M. T. (2018). Internal soil standard method for the Brazilian soil spectral library: Performance and
818 proximate analysis. Geoderma, 312, 95-103

819 Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., & Schüler, G. (2010). Retrieval of
820 chlorophyll and nitrogen in Norway spruce (Picea abies L. Karst.) using imaging spectroscopy.
821 International Journal of Applied Earth Observation and Geoinformation, 12(1), 17-26.

822 Schmidt, K., Behrens, T., Friedrich, K., & Scholten, T. (2010). A method to generate soilscapes from
823 soil maps. Journal of Plant Nutrition and Soil Science, 173(2), 163-172.

824 Shi, T., Wang, J., Chen, Y., & Wu, G. (2016). Improving the prediction of arsenic contents in agricultural
825 soils by combining the reflectance spectroscopy of soils and rice plants. International journal of applied
826 earth observation and geoinformation, 52, 95-103.

827 Signal Developers, (2013). Signal: signal processing URL: http://r-forge.r-project.org/projects/signal
828 (2013)

Skjemstad, J. O., Baldock, J. A., Carter, M. R., & Gregorich, E. G. (2008). Soil sampling and methods of analysis. Total and organic carbon'. 2nd edn.(Eds MR Carter, EG Gregorich), 225-237.

Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., & Foerster, S. (2016). Prediction of common surface soil properties based on Vis-NIR airborne and simulated EnMAP imaging spectroscopy data: Prediction accuracy and influence of spatial resolution. Remote Sensing, 8(7), 613.

Stenberg, B., Rossel, R. A. V., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. Advances in agronomy, 107, 163-215.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & Van Wesemael, B. (2010). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma, 158(1-2), 32-45.

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., & Schmidt, K. (2020). Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma, 376, 114552.

Tajik, S., Ayoubi, S., & Zeraatpisheh, M. (2020). Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Regional, 20, e00256.

Tekin, Y., Tumsavas, Z., & Mouazen, A. M. (2012). Effect of moisture content on prediction of organic carbon and pH using visible and near-infrared spectroscopy. Soil science society of America journal, 76(1), 188-198.

Thissen, U., Üstün, B., Melssen, W. J., & Buydens, L. M. (2004). Multivariate calibration with least-squares support vector machines. Analytical Chemistry, 76(11), 3099-3105.

Van Oost, K., Quine, T.A., Govers, G., De Gryze, S., Six, J., Harden, J.W., Ritchie, J.C., McCarty, G.W., Heckrath, G., Kosmas, C. and Giraldez, J.V. (2007). The impact of agricultural soil erosion on the global carbon cycle. Science, 318(5850), 626-629.

Vašát, R., Kodešová, R., Klement, A., & Borůvka, L. (2017). Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. Geoderma, 298, 46-53.

Viscarra Rossel, R., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1-2), 46-54.

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., ... & Ji, W. (2016). A global spectral library to characterise the world's soil. Earth-Science Reviews, 155, 198-230.

Waiser, T. H., Morgan, C. L., Brown, D. J., & Hallmark, C. T. (2007). In situ characterisation of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal, 71(2), 389-396.

Wang, G., Ye, J. C., Mueller, K., & Fessler, J. A. (2018). Image reconstruction is a new frontier of machine learning. IEEE transactions on medical imaging, 37(6), 1289-1296.

Wehrens and Mevik . (2007). The pls package: principal component and partial least squares regression in R J. Stat. Softw., 18 (2) (2007), pp. 1-24.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52, 394-403.

Wijewardane, N. K., Ge, Y., & Morgan, C. L. (2016). Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalisation. Geoderma, 267, 92-101.

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent laboratory systems, 44(1-2), 175-185.

Xu, D., Ma, W., Chen, S., Jiang, Q., He, K., & Shi, Z. (2018). Assessment of important soil properties related to Chinese Soil Taxonomy based on vis–NIR reflectance spectroscopy. Computers and electronics in agriculture, 144, 1-8.

Zádorová, T., Penížek, V., Šefrna, L., Rohošková, M., & Borůvka, L. (2011a). Spatial delineation of organic carbon-rich Colluvial soils in Chernozem regions by Terrain analysis and fuzzy classification. Catena, 85(1), 22-33.

879 Zádorová, T., Penížek, V., Vašát, R., Žížala, D., Chuman, T., & Vaněk, A. (2015). Colluvial soils as a
880 soil organic carbon pool in different soil regions. Geoderma, 253, 122-134.

881 Zadorova, T., Žížala, D., Penížek, V., & Čejková, Š. (2014). Relating extent of colluvial soils to
882 topographic derivatives and soil variables in a Luvisol sub-catchment, Central Bohemia, Czech
883 Republic. Soil and Water Research, 9(2), 47-57.

884 Zhang, D., & Zhou, G. (2016). Estimation of soil moisture from optical and thermal remote sensing: A
885 review. Sensors, 16(8), 1308.

886

887

888

889

890

891

892

893

Manuscript under review (Catena)

Verifying the impact of fusion high-resolution simulated in-situ spectroscopy, Sentinel-2, and Unmanned Aircraft Systems data into a single dataset to improve soil organic carbon content estimation in low-carbon agricultural fields.

James Kobina Mensah Biney*[1,2,3], Luboš Borůvka[1], Aleš Klement[1], Jakub Houška [2], Jakub Cervenka [2], Ndiye Michael Kebonye[1], Diego Urbina Salazar[3]

[1]Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic

[2]The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Department of Landscape Ecology, Lidická 25/27, Brno, 602 00, Czech Republic

[3]UMR EcoSys, AgroParisTech, INRAE, Université Paris-Saclay, 78850 Thiverval-Grignon, France

Correspondence E-mail: no2james@yahoo.com (J.K.M. Biney)

Abstract

Accurate estimation of soil organic carbon (SOC) content remains key because of its numerous benefits to the environment, including but not limited to contributing to food security and mitigating the greenhouse gas effect. The results of SOC estimates provided by researchers using proximal (spectroscopy) and remote sensing (airborne or spaceborne) data varied (from bad to excellent) across several fields. However, most of these studies focus on areas high in organic carbon, whereas areas low in SOC have received limited attention. It is believed that merging high-resolution spectroscopy with remote sensing datasets could improve the estimation of SOC. Currently, no single sensor or technique can estimate all soil properties accurately, including SOC. However, integrating data from these sensors remains a challenge due to differences in spatial/spectral resolution for each sensor, making the combination problematic. Therefore, the present study aims to explore improving the prediction and mapping of SOC content in two agricultural fields low in organic carbon where detailed information needs to be captured. This was done by merging data from in-situ spectroscopy, Unmanned Aircraft

Systems (UAS), and Sentinel-2 (S2) into a single dataset through data fusion. Before the fusion process, each platform's variable importance (VI) data was selected as the final data fused into a single dataset. The study confirmed that fusion of these datasets using VI provides better SOC results than separately using individual datasets. Concerning the SOC spatial distribution map, the in-situ map resembled the referenced measured map compared to the other platforms. Although the data fusion approach used in this study is a promising tool, further research is highly recommended, mainly using remote sensing data with fewer defects and other modeling techniques. The results obtained from the two study fields varied, ranging from an improvement in one field compared to no improvement in the other field.

Keywords: Soil organic carbon; Data fusion; In-situ spectroscopy; Remote sensing; Variable importance; Agricultural soil

1. Introduction

The Spatio-temporal variability of soil organic carbon (SOC) content, regulated by both natural and anthropogenic influences, is the interaction of numerous processes and factors that vary from one locality to another. This is because no two soils are the same. Moreover, the soil is considered a complex environmental system with temporal and spatial variation at multiple scales (Webster & Oliver, 2007; Post & Kwon, 2000; Guenet et al., 2021; Hugue et al., 2016; Guo and Gifford, 2002). However, as a result of the soil's complex features and the variable spectral response of organic matter, resulting in a lack of clear and narrow spectral features, estimating the behaviours of soil attributes, such as SOC, has provided mixed results (bad and good) despite the use of several machine learning and regression models on datasets obtained from both proximal and remote sensing techniques. Therefore, accurate estimation of SOC content remains vital because of its central role and benefits in various soil functions in the environment. This includes contributing to food security, greatly influencing soil structure, fertility, water-holding capacity, and mitigating greenhouse gas levels in the atmosphere. As a result, failure to accurately estimate its content may lead to wrong decisions by farmers and policymakers, affecting not only a community but possibly a country as a whole. This is because minor changes in SOC content or stock may have a negative impact on atmospheric CO_2 concentrations (Davidson & Janssens, 2006). To aid farmers and decision-makers in accurate decision-making and natural resource planners for optimal

planning, reliable estimates of SOC content must be examined across various spatial scales (Vaudour et al., 2019).

In the last few decades, a wide variety of studies have explored the use of several existing, updated, and newly developed machine learning (ML) and regression models, as well as other approaches in the quest to improve the estimation of SOC (e.g., Lamichhane et al., 2019; Heung et al., 2016; Heuvelink et al., 2021; Taghizadeh-Mehrjardi et al., 2016). Nevertheless, some of these models have yielded average to excellent results, especially in fields high in organic content. In contrast, for fields low in organic carbon content, the performance of these models ranges from average to abysmal results. Even though progress in developing these algorithms (ML and regression models) outweighs proximal and remote sensing techniques, there is still no best global predictive algorithm for SOC estimation. Because specific criteria and factors must be considered to predict SOC, for example, accurately, the geospatial characteristics of the study site, sample size and the specified covariate (Yimer et al., 2006; Wang et al., 2019; Yao et al., 2019). This implies that the varying results for SOC obtained using these algorithms can also be attributed to how data sets are collected and processed using these sensing techniques [proximal (e.g., spectroscopy) and remote (e.g., spaceborne and airborne).

The use of spectroscopy as a high-resolution technique (Munnaf et al., 2020), specifically in the visible near-infrared (Vis-NIR) range, has facilitated the rapid determination of organic components, particularly SOC. Spectroscopy, as a cost-effective method, has been shown to produce accurate measurements of soil properties, including SOC, compared to remote sensing methods. This is a result of its ability to retrieve soil information more effectively due to the proximity of its sensors to the target attributes during spectral measurement in the field or the laboratory (Viscarra Rossel et al., 2006; Kuang et al., 2012; Angelopoulou et al., 2019; Grunwald et al., 2015).

Although spectroscopy is more often used in the laboratory to predict a variety of soil constituents using diagnostic spectral features and statistical regression approaches (Bayer et al., 2016), in-situ applications are increasingly being utilised (Ben-Dor et al., 2009; Kweon and Maxton, 2013). This technique can also be mounted on platforms ranging from handhelds to fixed installations or tractor-embedded sensors. However, on a larger scale or in areas difficult to access, the use of spectroscopy that provides

information only at selected points is limited (Schwartz et al., 2011; Stenberg et al., 2010; Viscarra Rossel et al., 2006; McCarty et al., 2002). Additionally, over larger areas, its prediction levels of accuracy tend to decline, owing primarily to nonlinear correlations between soil characteristics and spectra, resulting in more significant prediction errors (Stevens et al., 2013; Stenberg et al., 2010).

Remote sensing (RS), on the other hand, is commonly associated with the use of satellite (e.g., Sentinel-2 (S2)) and airborne (Unmanned System Aircraft (UAS)) platforms that employ multi or hyperspectral imagery. Regular updates of RS data provide great potential for estimating and mapping SOC over a large domain, including inaccessible areas (needs physical contact). As an alternative to costly and time-consuming conventional field sampling and analysis, this sensing technique is rapid and provides a spectral reference base for soil attributes (Castaldi et al., 2019a). However, its low spatial resolution is one of its significant limitations, including several disturbing environmental factors, as stated in the literature.

As a result, identifying a universally accepted single sensor to function under such circumstances (soil's complex nature) to measure all soil parameters reliably remain key due to these sensors' limitations. This is because each of these sensors does provides distinct spatial and temporal perspectives of soil at different spatial resolutions (Grunwald et al., 2015) [e.g., spectroscopy (ranges in nanometers (nm), satellites (S2 (meters (m)), airborne (UAS (centimetres (cm)]. Finding a strategy for integrating data from each of these sensors and fusing it into a single data set, especially in fields with low organic carbon content, could thus improve SOC estimation accuracy.

For instance, some studies have estimated and mapped SOC and other soil properties within the last few decades using proximal and remote sensing data. However, most of these studies (Biney et al., 2021; Gholizadeh et al., 2018; Žížala et al., 2019; Gomez et al., 2008; Crucil et al., 2019) use these data separately in a single study. Although in some literature, some of these datasets were merged to estimate soil properties (Sabetizade et al., 2021; Vohland et al., 2014; Peng et al., 2015; Bousbih et al., 2019; Wang et al., 2020; Vohland et al., 2022). Moreover, this was mostly done with only airborne, spaceborne, or spectroscopy datasets and spaceborne and airborne data combined. Additionally, several procedures were used in the said studies to achieve their set objectives. This ranges from upscaling and

downscaling (Li et al., 2017; Wang et al., 2017; Peng et al., 2015), data integration, sensor fusion or data fusion (Wang et al., 2017; Ji et al., 2019; Khaleghi et al., 2013; Simone et al., 2002). According to Grunwald et al. (2015), sensor/data fusion could be a viable solution for integrating soil attributes at multiple scales of variation (both horizontally and vertically) to improve soil attribute estimation, including SOC (i.e., merging proximal and remote sensing data).

Data fusion is the process of combining data and information from multiple sensors into a single dataset, thereby improving the interpretation performance of the obtained data to achieve greater accuracy than each of the source data alone (Hall and Llinas, 1997). Data fusion incorporates concepts from various disciplines, including signal processing, information theory, statistical estimation and inference, and artificial intelligence (Khaleghi et al., 2013). Consequently, researchers have attempted to use multiple sensors to obtain more accurate results (Horta et al., 2015; O'Rourke et al., 2016; Xu et al., 2019). Several approaches are used to perform data fusion, including a simple combination of the original data (Viscarra Rossel et al., 2006; Ji et al., 2019), a simple combination of selected spectral features (Xu et al., 2019b), and a combination of the measurement results (O'Rourke et al., 2016). The combination of measurement results, also called model averaging (Horta et al., 2015), involves various model outcomes to obtain a better result. This improves the estimation accuracy and reduces the possibility of aberrant measurements (O'Rourke et al., 2016; Chen et al., 2019). However, the accuracy could decline in fields with low soil properties under consideration. Although various categories of fusion approaches have been developed for different purposes (e.g.), some of these methods have performed better than others in the broad area of fusion technology (Grunwald et al., 2015).

As global climate change remains a major environmental concern (Enriquez-de-Salamanca et al., 2017), SOC will continue to play an essential role in the carbon cycle globally because, according to Lal (2004), it remains the most significant carbon pool on land. As a result, several studies are ongoing using modified existing and new techniques and approaches to estimate and map SOC content and improve upon existing and new data for SOC estimation. Also, previous work on the current fields (Gholizadeh et al., 2018; Biney et al., 2021) has shown medium-to-poor SOC estimate results using proximal and remote sensing datasets individually.

The objective of the present study was to compare the individual and merged abilities of in situ, UAS, and S2 sensors for estimating the content of SOC in two different agricultural fields with varying types of soil. In the case of the merged approach, different combinations of the above data sets were explored. To the best of our knowledge, no studies have tested the impact of merging these three datasets into a single dataset through data fusion to estimate SOC. As a result, this study's tasks included (i) estimating the contents of SOC using the in-situ, UAS, and S2 data separately in their original resolution, (ii) performing the estimations with middle-level fusion techniques (simulated in situ + UAS + S2) after variable importance selection using the Boruta algorithm, (iii) verifying the performances of support vector machine (SVM) on the fused data to estimate SOC, and (iv) creating spatial distribution maps of SOC content in the study areas using the inverse distance weighting (IDW) interpolation method.

2. Materials and methods

2.1. Site description

The study area selected from which data were collected consists of two different agricultural fields located at Nová Ves nad Popelkou and Udrnice. The Nová Ves nad Popelkou study area (Field A) is 22 ha (50°31′ N; 15°24′ E) in central Bohemia, with a mean altitude of 185 m asl, while Udrnice (Field B) is 52 ha (50°21′ N; 15°15′ E) in the district of Jicin, with a mean altitude of 269 m asl (Fig. 1). These fields are representations of arable land that has been extensively farmed. The areas are primarily rural, with the most dominant crops being winter and spring barley, spring cereals, and maize. Additionally, the selected sites were representative of soil capes, which were homogenous and comparable in terms of terrain characteristics, land management, and climatic conditions (Schmidt et al., 2010). These areas were also characterised by dissected relief with side valleys, toe-slopes, and back-slopes. According to the World Reference Base (WRB) for soil resources (World Reference Base for Soil Resources 2014), the soil types in Udrince consist mainly of Chernozems and Luvisols on loess, while those in the Nová Ves nad Popelkou region are predominantly Cambisols on sedimentary rocks. Table 1 shows the study sites and data collection details.

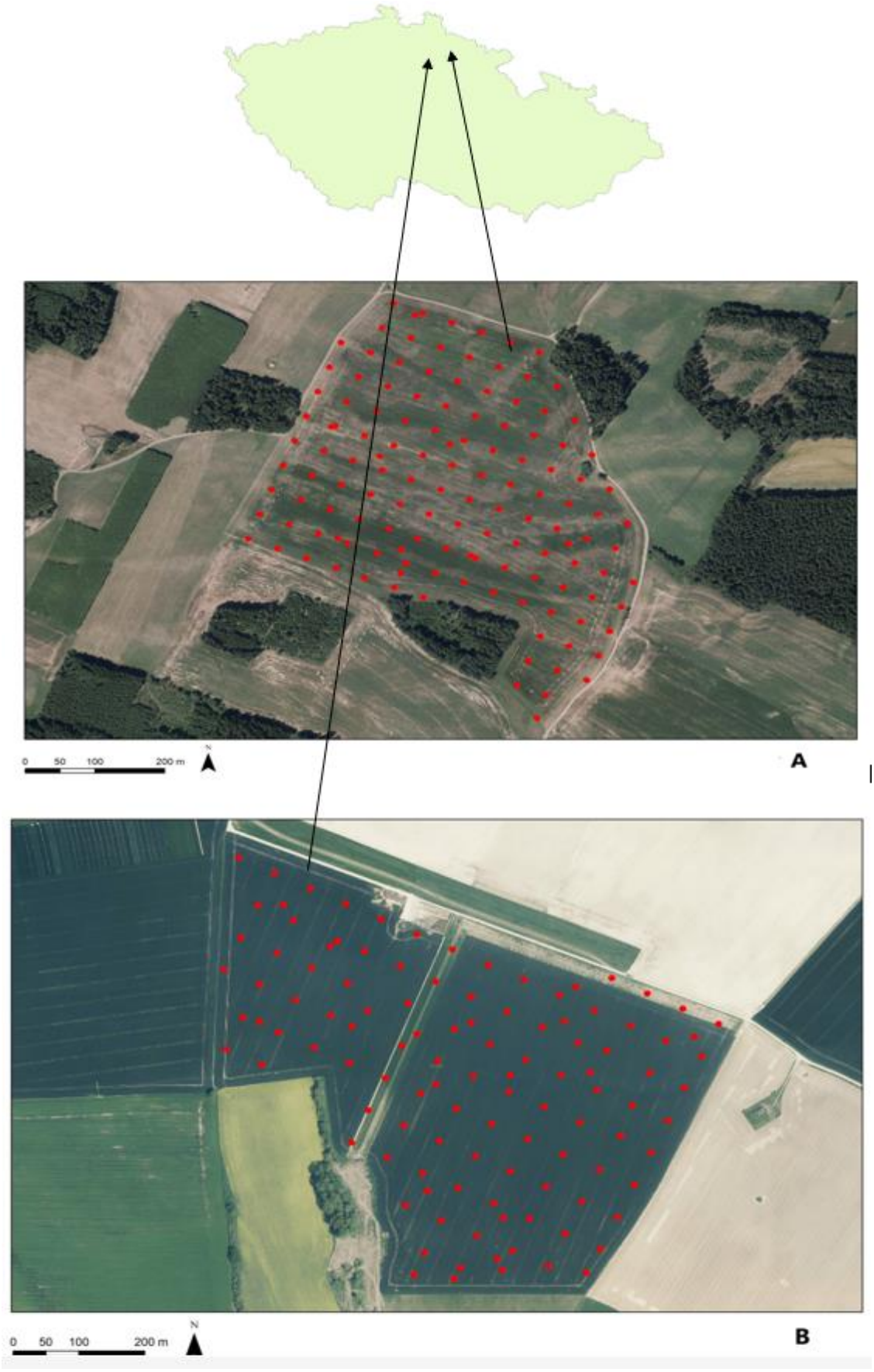


Fig. 1: Location of sampling points in the Czech Republic [Nová Ves nad Popelkou (A), Udrnice (B)]

161 Table 1: Study area and data collection details.

Location	Area size (ha)	Dominant soil unit	Samples (no.)	Soil sampling	UAS	S2
Udrnice	22	Chernozems and Luvisols on loess	130	10.06.2019	24.11.2019	25.07.2019
Nová Ves nad Popelkou	52	Cambisols on sedimentary rocks.	111	24.05.2019	24.11.2019	10. 07.2019

162

163 *2.2. Soil sampling and in situ spectra measurement*

164 A total of 241 soil samples were comparably taken from the corresponding agricultural fields [Field A
165 (130), Field B (111)]. A rectangular grid sampling strategy with a space interval of 40 m (Field A) and
166 60 m (Field B) was adopted. To be transported to the laboratory for chemical analysis, soil samples from
167 the topsoil layer (0 to 20 cm depth) were collected and placed in clearly labelled bags (approximately
168 150 to 200 g). The sample size provided adequate coverage of the fields and was symbolic of the area
169 and samples for which the models were implemented. Each field's sampling points were located in the
170 field using a GeoXM (Trimble Inc., 2007) receiver at an accuracy of 1 m. Both fields had not been
171 recently ploughed (soil not disturbed). These samples were air-dried for two weeks in the lab, and large
172 clods of the dried soil samples were gently crushed in a porcelain bowl and then passed through a 2 mm
173 sieve. Soil organic carbon content (SOC, %) was measured in two steps using the dichromate redox
174 titration method (Skjemstad and Baldock, 2008). The specimens (samples) were initially oxidised with
175 $K_2Cr_2O_7$, and afterwards, the solution was potentiometrically titrated with ferrous ammonium sulfate
176 ($FeH_8N_2O_8S_2$). Concurrently, in situ spectral measurements were performed on the soil using an ASD
177 Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) with a high-intensity
178 contact probe across the 350–2500 nm wavelength range during the respective field sampling. Before
179 that, the removal of some undesired materials from the soil surface (excluding plant material, crop
180 residues, or stones) was carried out. The spectroradiometer spectral resolution was 3 nm for the region
181 of 350–1050 nm and 10 nm for the region of 1050–2500 nm. Three spectral measurements for each
182 sample were collected, and the average values were used for further analysis. The sensor was

recalibrated every ten runs using a Spectralon® (Labsphere, North Sutton, NH, USA) standard white reference panel (Shi et al., 2016).

2.3. Remote sensing data acquisition

2.3.1. Sentinel-2 imagery (S2)

The Multispectral Sentinel-2B imagery used was a cloud-free image-level 2A product ready to be used. This is because processes including geometric, radiometric, and atmospheric corrections have already been performed by the manufacturers using Sen2Corprocessor. The imagery, made up of 13 spectral bands (10 m, 20 m, or 60) (Table 2), was obtained through the Copernicus Open Access Hub of the European Space Agency on July 10, 2019 (Field A) and July 25, 2019 (Field B). To ensure that all the bands had the exact resolution, Snap software was used to resample from the original Bottom of Atmospheric (BoA) 20 or 60 m spatial resolution to 10 m using the nearest neighbor resampling approach because it is computationally efficient and conserves the input image pixel values (Roy et al., 2016). Three bands were excluded for further analysis (B1, B9, and B10). The remaining 10 bands [(VNIR-SWIR, B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12] were used in this study. According to Elhag and Bahrawi (2017), the selected bands are usually used to assess soil properties. Technical information about the S2 bands used can be found in the European Space Agency's handbook (2010).

2.3.2. Unmanned Aircraft System Multispectral Imagery

Multitemporal spectral images were collected on November 25, 2019, using a fixed-wing Unmanned Aircraft Systems (UAS) (Trinity F90 fixed-wing) mounted with two cameras: a sony camera (RGB, 3 bands) and a MicaSense Altum camera (Multispectral, 6 bands), totalling nine bands (Table 3). The sensor resolution was 2064 x 1544 pixels with a long wave infrared band of 160 x 120 pixels. Additionally, the field of view was 47° x 37° (multi-spectral) and 57° x 44° (thermal), respectively. The camera has a sun sensor that gathers information about the light conditions and saves the radiant flux data in EXIF format. A QBase 3D smartphone app (mission planning software) was used to create the flight plan, which acted as the primary interface between the user and the UAS device. The QBase 3D provides real-time information about the UAS, such as altitude, distance, battery life, and mission

telemetry data, so the operator always has the most up-to-date information on the flight. Batteries were added to last the flight duration for the two study areas. The images were taken between 9:30 and 13:30 under cloudless conditions to guarantee high-quality imagery. The flight height was 190 m (Field A) and 145 m (Field B), with a spatial resolution of 8.8 cm (Field A) and 7.7 cm (Field B) (MicaSense Altum camera), covering an area of 31 ha for Field A and 52 ha for Field B. All collected images during the flying of the UAS were processed in one go. Then, an exterior orientation based on tens of thousands of identical points was created. Afterwards, point clouds and DEMs (digital elevation models) were generated to collect ground control points. When the residuals on each GCP are satisfied, an orthophoto mosaic is created by recalculating all the models again. The orthorectified image is exported as one mosaic in GeoTIFF file format on the WGS-84 ellipsoid. Calibration was performed prior to generating this orthophoto. AgiSoft Metashape Professional 1.5.0 (AgiSoft LLC, St. Petersburg, Russia), photogrammetric processing was used. The normalised difference vegetation index (NDVI) was used to mask a 0.2 threshold to differentiate bare soil areas for both study fields.

Table 2: Specifications of Sentinel-2B Multispectral Instrument sensor

Spectral band	Spectral-domain	Central wavelength (nm)	Spatial resolution (m)
B1	visible (vis)	433	60
B2	vis	490	10
B3	vis	560	10
B4	vis	665	10
B5	Red-edge	705	20
B6	Red-edge	740	20
B7	Red-edge	783	20
B8	Near-infrared (NIR)	842	10
B8A	NIR	865	20
B9	NIR	945	60
B10	Short wave infrared (SWIR)	1380	60
B11	SWIR	1610	20
B12	SWIR	2190	20

224 Table 3: Specifications of UAS Multispectral Instrument sensor

Spectral band	Spectral-domain	Central wavelength (nm)
RGB		
B1	Blue	475
B2	Green	560
B3	Red	650
Multispectral		
B4	Red	668
B5	Red-edge	705
B6	Red-edge	717
B7	Red-edge	740
B8	Near-infrared	842
B9	Thermal (long wave infrared)	945

225

226 *2.4. Pearson correlation matrix and alignment of in-situ spectroscopy data to multi-spectral imaging*

227 The degree of a linear relationship between two variables is measured by correlation (predictors and
228 responses). Before the correlation matrix was applied to the S2 and UAS, as well as the in-situ spectral
229 data with SOC for each study field, some modifications were made to the in-situ data. As already stated,
230 the in-situ spectra range from 350–2500 nm, but these are point-based measurements. Therefore, the
231 reflectance data from the in-situ spectroradiometer were simulated by S2 and UAS multi-spectral
232 imaging sensors to obtain simulated broadband reflectance data at ground level, resulting in 12 bands
233 (S2 format) and 8 bands (UAS format). Due to noise in the 350–399 nm region, only the reflectance
234 values between 400 nm and 2500 nm were used.

235 A "read.asd" tool was developed in Visual Basic for Applications (VBA) under Excel® software
236 (Vaudour et al., 2014) to process and automate the in-situ spectra data. The tool was also designed using
237 the Relative Spectral Response Function (RSRF) to weight the average in-situ reflectance values
238 automatically. The relative spectral response (RSR) for S2 and the specific UAS imagery used in this
239 study were obtained and utilised. The RSRF integrated the in-situ spectrum reflectance into the UAS

and S2 band formats. Even though it is possible to acquire a synthesised value by computing a simple mean from the nominal FWHM bandwidth, it is preferable to use the available RSR to account for the sensitivity of each sensor band (satellite and drone) (Feilhauer et al., 2013). The S2 data, on the other hand, were extracted into reflectance values using Snap software, while the UAS values for each band were extracted into digital numbers format. However, it was converted into reflectance values using the band centre value, which was 32768 and represented 100% of the reflectance. Each band was divided by 32768 to obtain normalised values ranging from 0 to 1. Finally, the Pearson correlation matrix was performed using the UAS and S2 data and the simulated in-situ spectral data (S2 and UAS) with SOC for comparison. This was done to help explain the impact of each dataset on SOC prediction. Additionally, the simulated in-situ datasets (UAS or S2 format) were compared, and the data showing the best correlation with SOC were selected as in-situ spectral final data.

2.5. Pretreatment and outlier removal.

During in-situ spectral measurement, the obtained spectra are prone to many defects, ranging from noise and other undesirable side effects (i.e., artefacts). These artefacts can negatively influence the obtained spectra, thereby affecting the predictive performance of the attribute under consideration (SOC). Therefore, to optimise the fitting of SOC against the spectral data while also improving the SOC relevant details, pretreatment algorithms were applied to the in-situ spectral data to help reduce the unmodeled variability in the data and enhance the features sought in the in-situ spectral data. The pretreatment methods used include SG filtering (with a second-order polynomial fit and 21 smoothing points) [sgolayfilt function from the signal R package (Signal developers, 2013) was used], logarithmic (log (1/R)) transformation (to linearise the relation between spectral values and the concentration of absorbing soil constituents) and the standard normal variate (SNV) (normalisation of the spectra). The SNV was calculated by subtracting every reflectance value from the mean reflectance value of the particular spectrum and dividing this value by the standard deviation of the whole spectrum.

According to Murray (1988), outliers can be classified as undesirable data that can influence model output and must be checked and removed to enhance the model's predictive performance. Therefore, before the spectral treatment, the presence of outliers was examined using ensemble sparse partial least

squares (enpls). In total, 6 outliers (Field A (4), Field B (2)) were removed from the in-situ data. No outliers were removed from either the S2 or UAS datasets.

2.6. Variable importance selection

The selection of featured bands (variable importance) is an efficient way of improving the model's accuracy and robustness by reducing the bias caused by uninformative variables (Xu et al., 2019). This is because not all variables are significant for estimating the desired outcome. As a result, removing irrelevant variables may aid in improving the model's predictive performance. The Boruta algorithm, which is an ensemble learning technique, and a random forest classification wrapper were used for the future selection approach. This algorithm can select all feature sets related to the dependent variables and can also detect the influencing factors of the dependent variables to achieve effective and superior feature selection. Boruta employs a random forest classifier with two main adapters: permutation importance (raw permutation adapter and normalised permutation adapter) and Gini importance. The Gini importance of each feature is calculated as the sum of the number of splits and includes the feature. It also considers the fluctuations in the mean accuracy loss among the generated forest trees.

2.7. Data set preparation and development before data fusion

2.7.1. Transforming datasets to the same spatial resolution

As previously stated, S2 and UAS data were collected with varying spatial resolutions [(Field A (S2 (10 m), UAS (8.8 cm), Field B (S2 (10 m), UAS (7.7 cm)], whereas for the in-situ data, it has no spatial resolution but was collected with a spectral resolution of between 2 nm and 10 nm for both fields. Therefore, these datasets were transformed to a 10 m resolution format (S2 data as reference). This was done separately for each study field using only the future selected variables for each data. For the in-situ data, a kriging map was created using empirical Bayesian kriging (EBK). This kriging method (EBK) is a geostatistical interpolation technique that consists of two geostatistical models: intrinsic random function kriging (e.g., Chilès and Delfiner, 1999) and a linear mixed model (Diggle and Ribeiro, 2007). This method is unique and differs from other kriging techniques. EBK uses the process of subsetting and simulation to produce optimal results. Furthermore, parameter determination is performed

automatically. In the process, EBK assumes an estimated semivariogram for the interpolation region as well as a linear prediction that includes variable spatial damping, thereby producing less error. In contrast to ordinary kriging, which utilises weighted least squares to estimate semivariogram parameters, this method uses restricted maximum likelihood estimation.

Using the GPS coordinates of the study area (used to locate the sampling points in the field), including one of the selected bands, a map was generated with the band values overlaid on the image (for each sampling point 126). The map (referred to as the "estimated spectral map") was transformed to a raster with a resolution of 10 m. Afterwards, the spectral band values were extracted from all sampling points for each study field separately [Field A (126) and Field B (109)]. The procedure was repeated for each selected band based on the future selection approach. The final data (in-situ spectral) at 10 m resolution were used for further analysis. For the S2 data (already at 10 m), the future selected bands were retained as the final data set for each field. However, for the UAS, the imagery was resampled to a resolution of 10 m, and the future selection bands were also retained as the final data. The kriging and resampling or upscaling processes were performed in the ArcGIS map (ArcGIS 10.8.1) and ENVI 5.6.1

2.7.2. Data fusion

However, middle-level fusion was considered for this study by using the combination of UAS, S2 and in situ spectral feature bands selected by the Boruta algorithm. However, the low-level fusion was also tested, but the result is not shown because it was less accurate than the middle-level approach. In this study, the variables UAS, S2 and in-situ were denoted by i, j, and k, respectively, and the number of soil samples was denoted by n. The combinations of the spectra (UAS $\otimes$ S2 $\otimes$ in-situ) produced n outer product matrices with the multiplied intensities of the original three domains. The $i \times j \times k$ matrix was then unfolded to an $i \times j \times k$ vector, resulting in a new matrix with n rows and $i \times j \times k$ columns for the chemometric analysis.

$$Y = Wo + (C_{UAS} \times X_{UAS}) + (C_{S2} \times X_{S2}) + (C_{in\text{-}situ} \times X_{in\text{-}situ}) \quad (1)$$

317 Where Y is the vector with the soil property of interest (measured element contents), XUAS, XS2 and
318 Xin-situ are the independent variables (prediction outcomes of UAS, S2 and in-situ), Wo is the intercept,
319 and CUAS, CS2 and Cin-situ are the coefficients of UAS, S2 and in-situ outcomes.

320 *2.8. Chemometric analysis and model assessment*

321 *2.8.1. Support vector machine (SVM)*

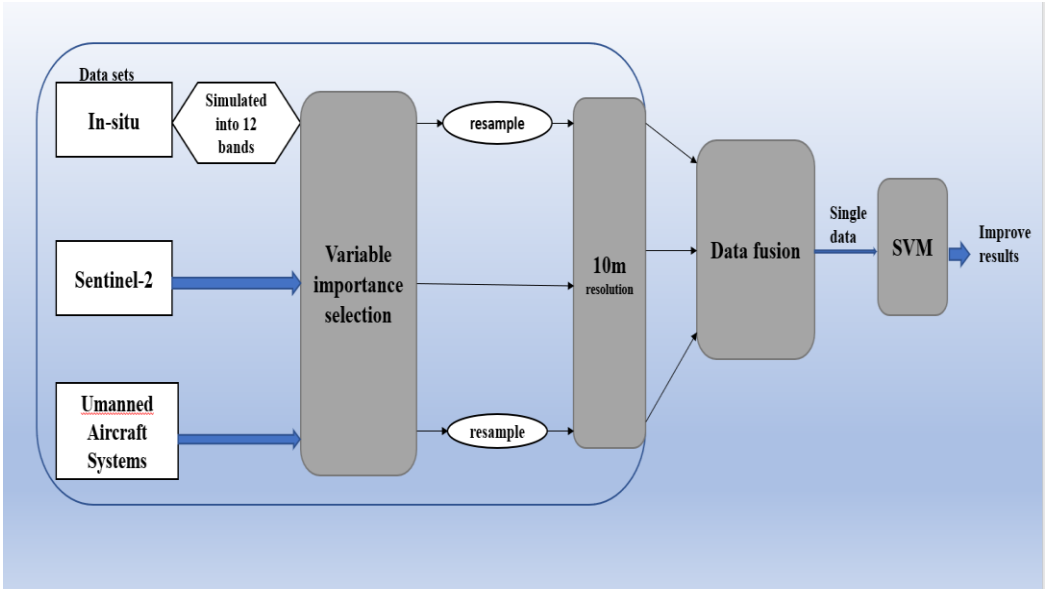
322 The support vector machine (SVM) algorithm was proposed by Guyon et al. (2002). This algorithm is
323 noted for updating the ranking criterion at each step of a backwards strategy: the criterion must be
324 reviewed at each stage, and the variable that minimises this measure must be excluded. SVM use kernel
325 functions such as degreed polynomial, radial basis, or hyperbolic tangent to project the data onto a new
326 hyperspace where complex nonlinear patterns can be simply represented (Gunn, 1998; Cortes and
327 Vapnik, 1995). For this study, SVM was tuned with different cost parameters with the built-in tuning
328 function of the grid search (specifically 0.001, 0.01, 0.1, and 1) using a linear kernel while the epsilon
329 parameter was left at its default value (0.1). The best cost parameter is determined from a 10-fold cross-
330 validation based on the RMSE. The package e1071 library in R was used.

331 *2.8.2. Data partitioning, model accuracy assessments and spatial distribution maps*

332 The model's output was assessed by fivefold cross-validation for each regression procedure, where the
333 whole dataset was randomly divided into calibration (75%) and validation sets (25%) of the samples.
334 The accuracy of the prediction was assessed based on the coefficient of determination (R^2_{cv}) [the R^2_{cv}
335 ranges from 0 to 1, where $R^2_{cv} = 1$ is the optimal value], the root mean square error of prediction
336 (RMSEcv) (measures the overall model prediction accuracy) and the ratio of performance to
337 interquartile range (RPIQ), which is defined as the interquartile range of the observed values divided by
338 RMSEcv. The RPIQ considers both prediction errors and variation in observed values, resulting in a
339 model validity metric that is more easily comparable across model validation studies. The RPIQ does
340 not make any assumptions about the distribution of the observed values. The greater the RPIQ is, the
341 better the model's predictive capacity [very poor model (RPIQ < 1.4), fair (1.4 ≤ RPIQ < 1.7), good
342 models (1.7 ≤ RPIQ < 2.0), very good models (2.0 ≤ RPIQ ≤ 2.5), and excellent models (RPIQ > 2.5)].

The RPD is the ratio of a parameter's standard deviation to a specific model's standard error of that parameter's prediction. For the RPD, Chang and Laird's (2002) categorisation was applied: RPD > 2 indicates good models, RPD between 1.4 and 2 indicates moderate predictive ability, and RPD = 1.4 indicates weak models. The fivefold cross-validation was repeated 100 times to ensure model stability and reliability.

Finally, the spatial variability of SOC contents was mapped using the inverse distance weighting (IDW) interpolation method using the predicted values for each platform data as well as the merged data and different fusion data sets. According to Qiao et al. (2018), this interpolation method can use point measurements at a given location to estimate other values at an unknown location. Additionally, due to its ease of application, the IDW method is classified among the most frequently used interpolation techniques in soil science. One of its major advantages is its ability to assign weights before prediction, thereby providing a lower error margin and creating a more accurate distribution map than other techniques (Liao et al., 2018). Fig. 2 schematically displays the experimental design.



3. Results

3.1. Descriptive statistics of soil organic carbon (SOC, %) content and simulated in situ spectral data

Table 4 shows the statistical results of SOC for two agricultural fields (A and B), indicating the standard deviation (SD), mean, skewness, minimum, maximum, and coefficient of variation.

361 Table 4: Descriptive statistics of soil organic carbon (SOC, %) content at two study fields

Samples	Mean	Median	SD	Skewness	Minimum	Maximum	CV(%)
Field A	1.44	1.44	0.33	0.57	0.6	2.93	23
Field B	1.09	1.03	0.35	2.20	0.53	2.88	31

CV: coefficient of variation, SD: standard deviation

362

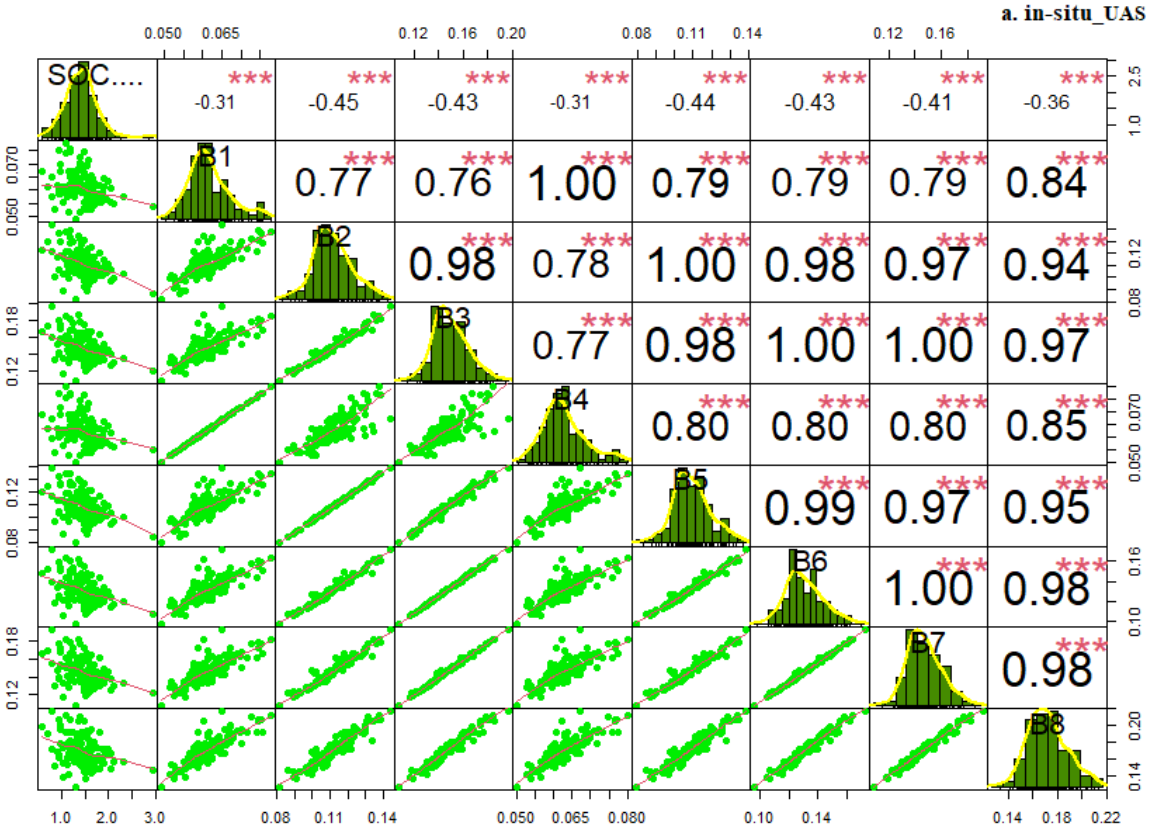
363 SOC values were approximately normally distributed across both study fields, with skewness values of
364 0.57 in Field A and 2.20 in Field B. The study fields differed significantly in terms of SOC content.
365 Field A had the highest mean content of 1.44%, while Field B, with a mean of 1.09%, was the lowest.
366 The results also show that the SOC distribution is more homogeneous in Field A than in Field B, with
367 the former having CV values of 23% and the latter having CV values of 31%. These SOC values could
368 indicate medium to semi-high SOC content for Field A and medium to poor SOC content for Field B.

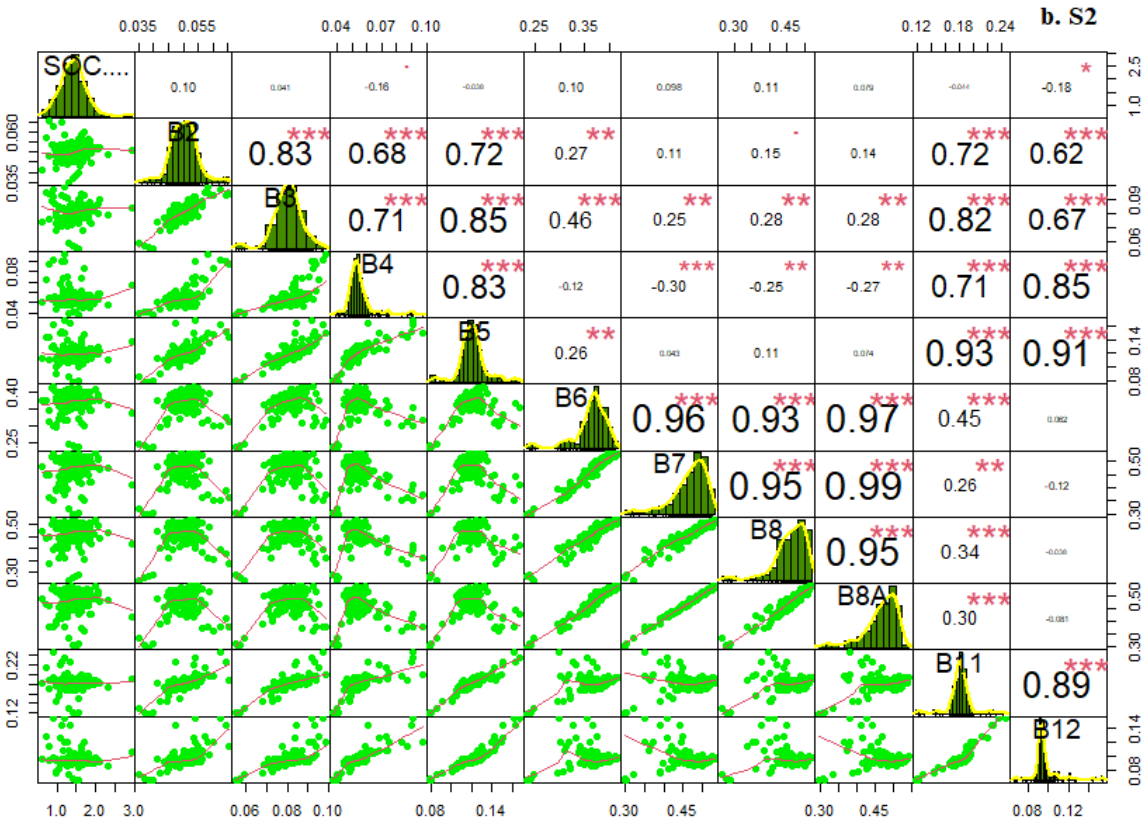
369 The simulated in-situ spectra (UAS format) were selected over the in-situ spectra (S2 format) as the
370 final in-situ dataset for each study field. The approach used, however, differs for each study field. For
371 example, for Field A, the majority of its simulated in-situ spectral bands (UAS) (Fig. 3a) significantly
372 correlate with SOC compared to the other in-situ spectral data (S2 format) (see supplementary file). In
373 the case of Field B, none of the bands in its simulated in-situ spectral data (UAS and S2 form) was
374 correlated with SOC. Nonetheless, the in-situ spectra (UAS format) show a strong correlation among
375 their individual bands compared to the other in-situ spectral data (S2 format) (see supplementary file).
376 Hence, only these in-situ data (UAS format) were considered for subsequent investigations in this study.

377 *3.2. Correlation matrix between SOC and bands for each platform data*

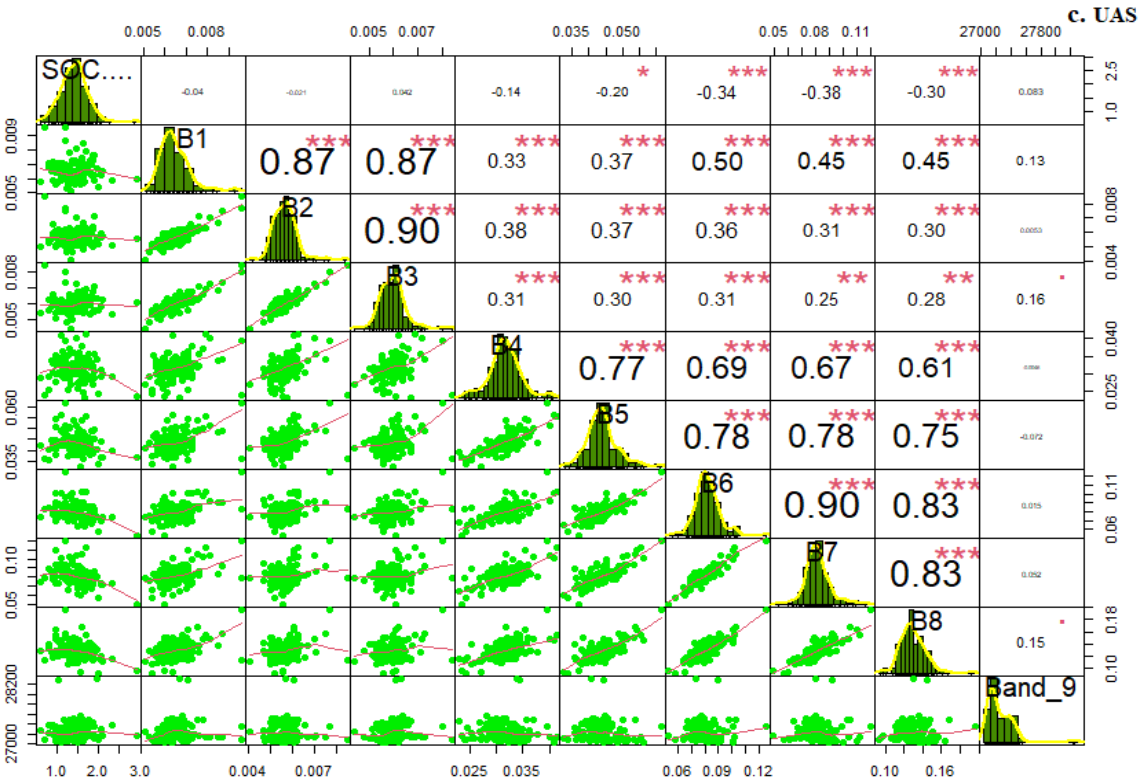
378 For Field A, the Pearson correlation analysis (Fig. 3b) showed no correlation between SOC and all S2
379 bands. In contrast, for the in-situ spectral data (simulated data), a moderately significant correlation was
380 found between SOC and five of its bands [B2, B3, B5, B6, B7, r =-0.45, -0.43, -0.41, -0.41, -0.41], while
381 three bands of the UAS data (Figure 3c) [B6, B7, B8, r =-0.34, -0.38, -0.30] showed weak correlation

with SOC. In the case of Field B, almost all the bands for the three datasets [in-situ spectra (simulated data), S2 and UAS] (Fig. 4 a, b, c) showed a vague or no relationship with SOC, except for Band 3 (r = -031) of S2 data (Fig. 4b), which had a weak correlation with SOC. Notwithstanding, the in-situ data were slightly better than the remaining two datasets (Field B) (order: in situ > S2 > UAS) in terms of better interrelationships among the individual bands, highlighting their strong mutual dependence.




b. S2
SOC....
B2
B3
B4
B5
B6
B7
B8
B8A
B11
B12

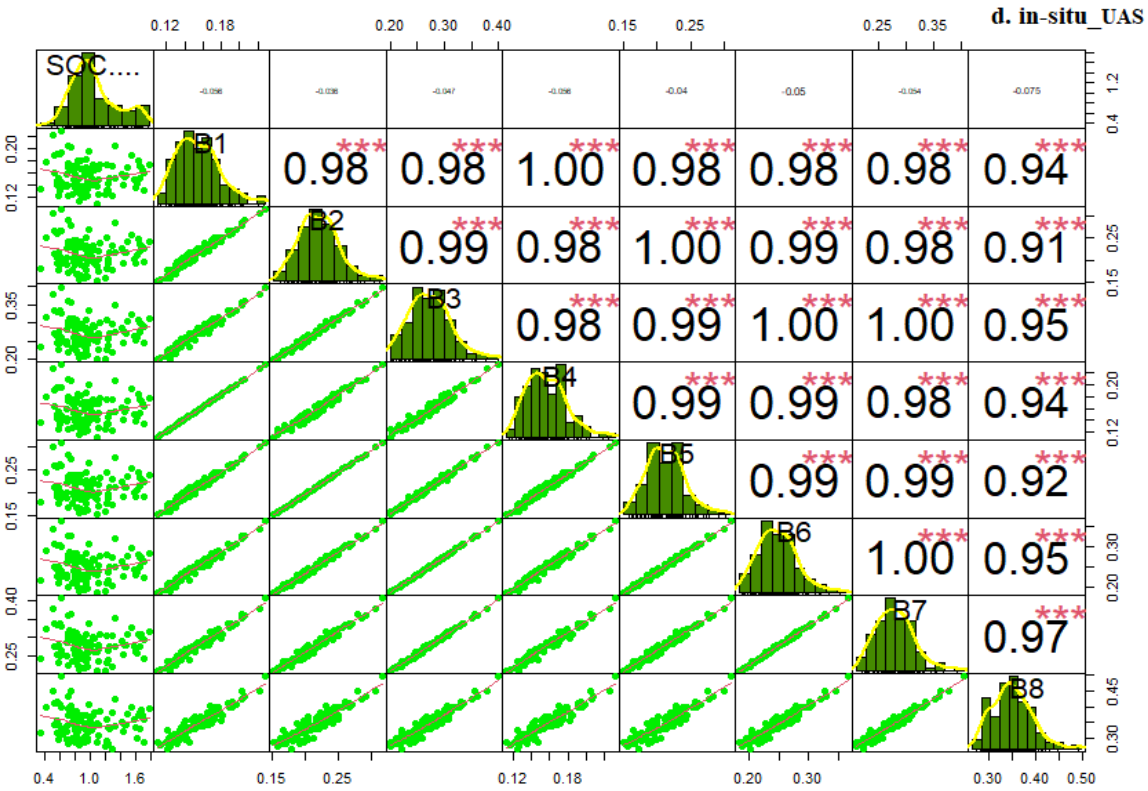
390


c. UAS
SOC....
B1
B2
B3
B4
B5
B6
B7
B8
Band_9

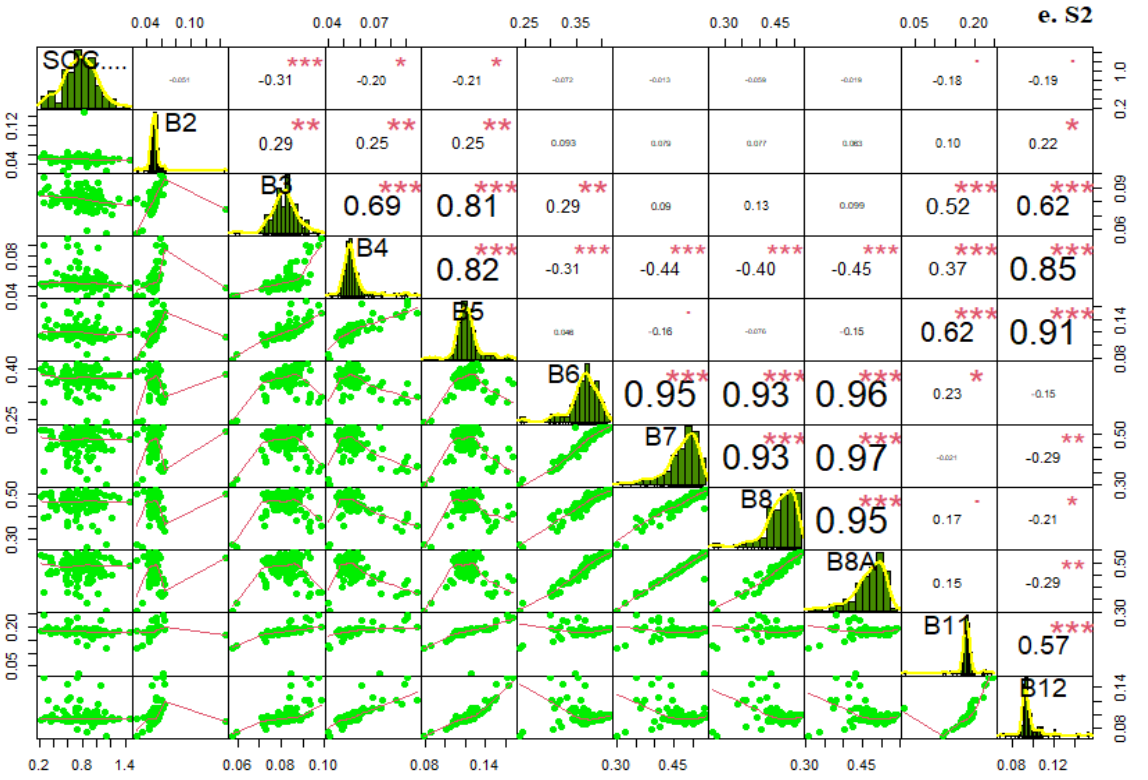
391

392 Fig. 3: Correlation matrices of SOC with the best simulated in-situ spectra (UAS) (a), S2 reflectance

393 bands (b) and UAS reflectance bands (c) for Field B



394



395

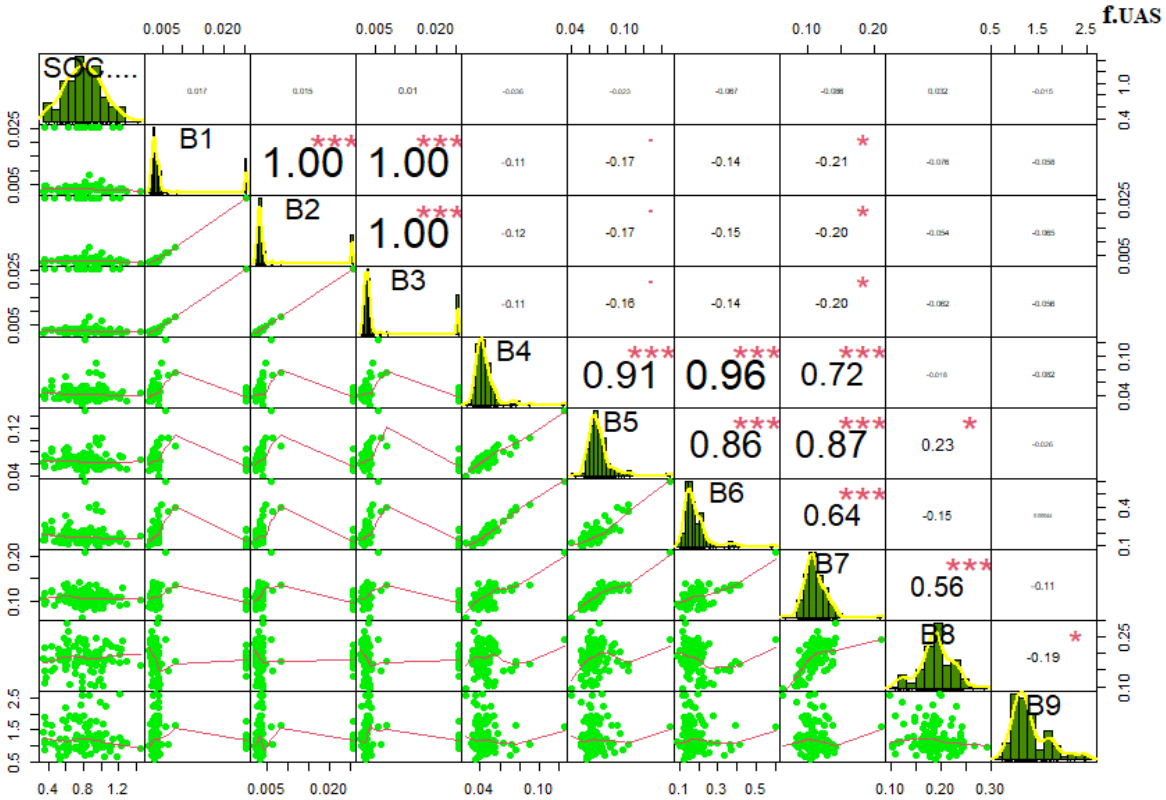
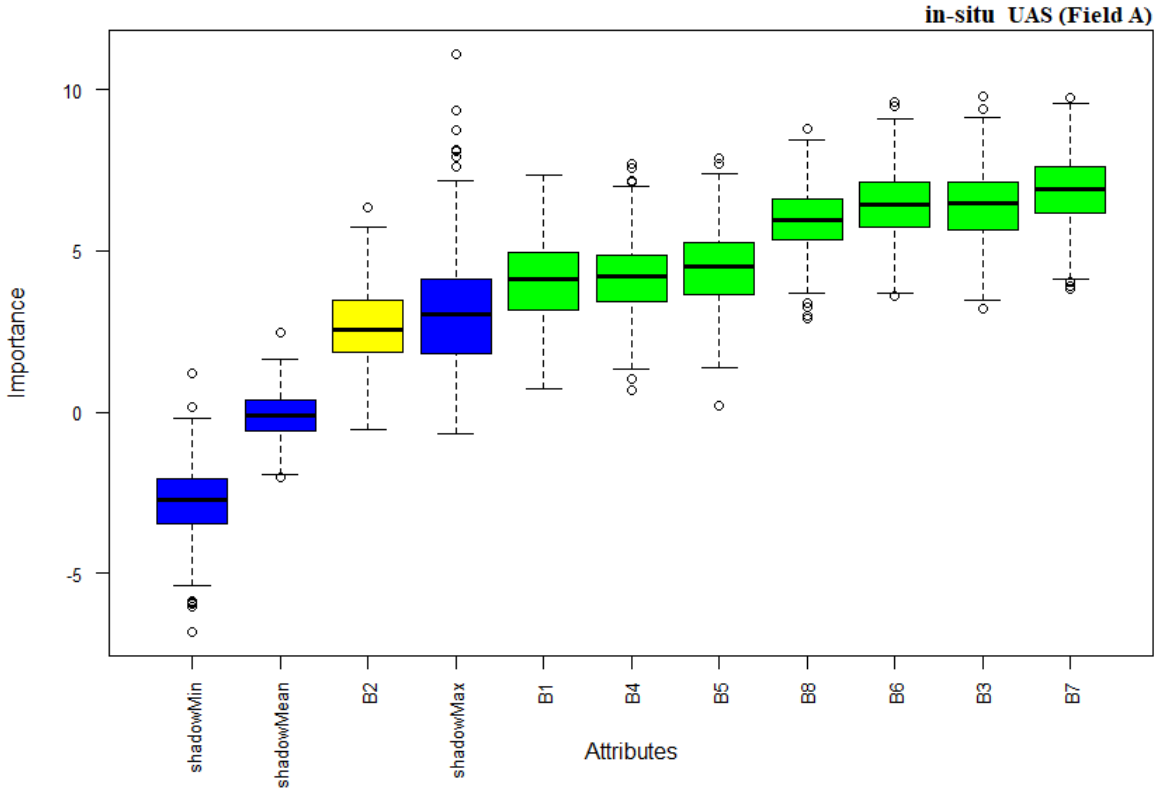


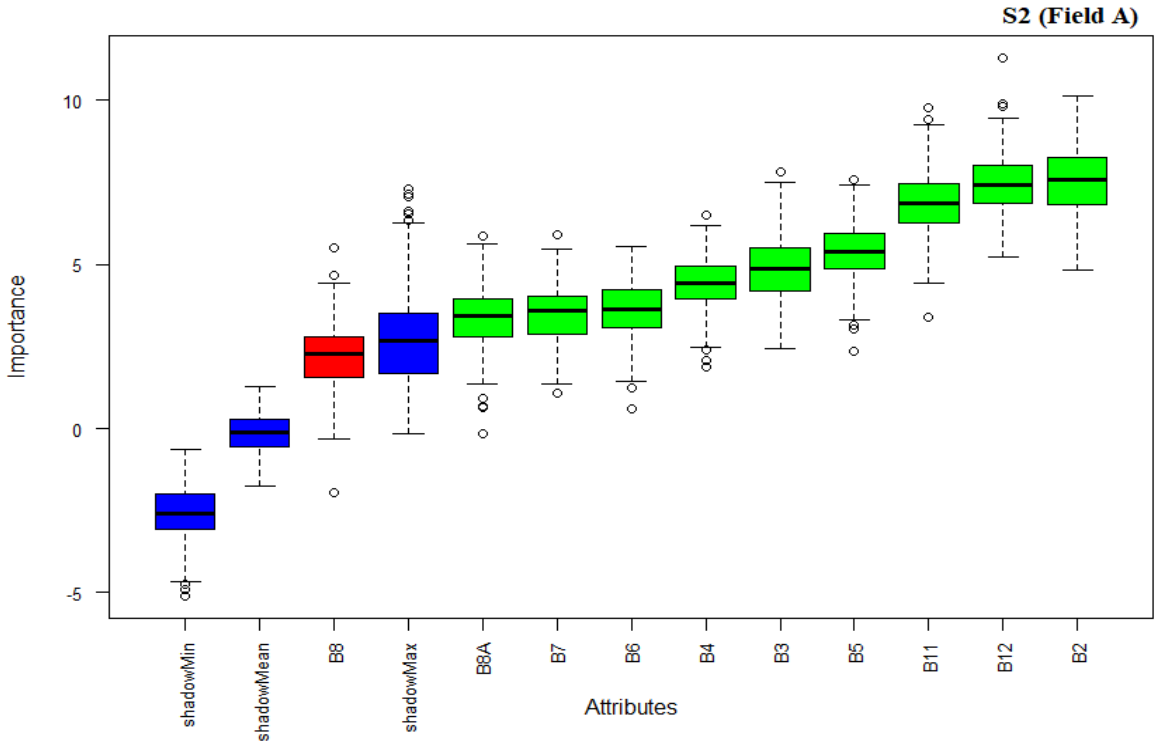
Fig. 4: Correlation matrices of SOC with the best simulated in-situ spectra (UAS) (d), S2 reflectance bands (e) and UAS reflectance bands (f) for Field B

3.3. Future selection with the Boruta algorithm

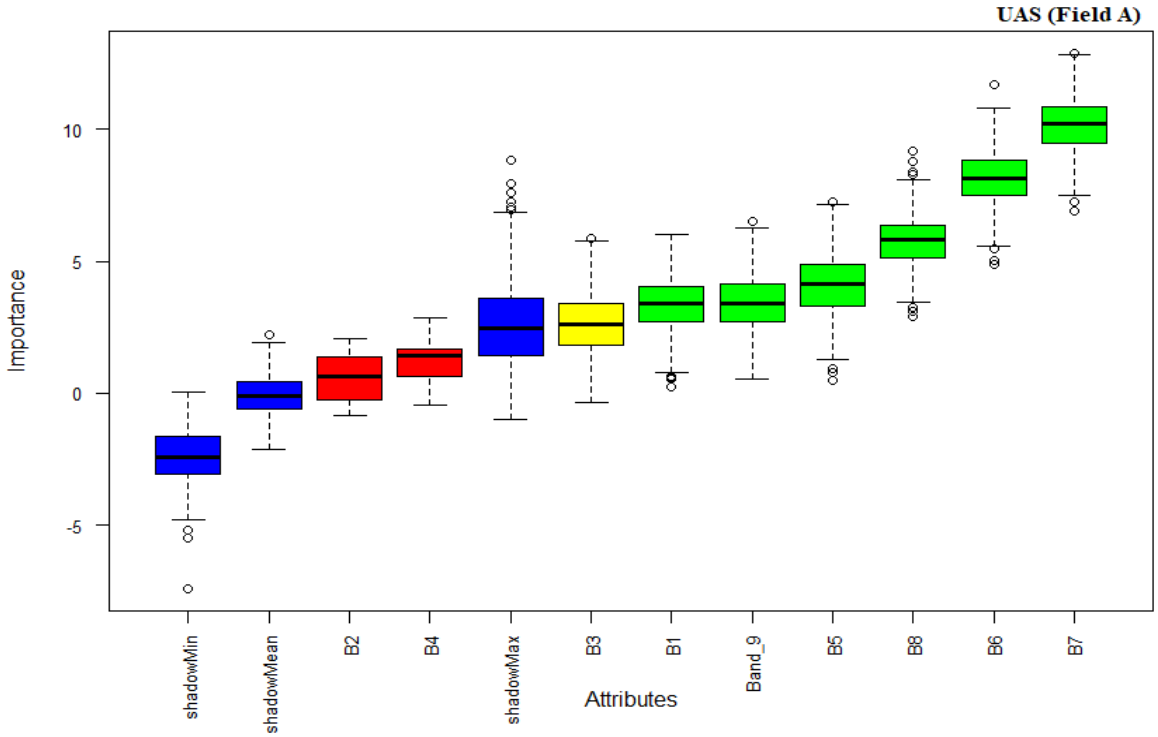
Following the application of the Boruta algorithm to each of the three datasets (in situ, UAS, and S2) for both Field A and Field B, some of the bands constituting these datasets (bands before the shadow max index) (Fig. 5 a and b) were deemed irrelevant and were removed. This includes the following bands: [in-situ (B2), S2 (B8), UAS (B2, B4) (Field A)], and [in-situ (B1, B4), S2 (B2, B6, B7, B8, B8A), UAS (B4, B5, B6, B7, B8, B9 (Field B)]. As a result, only the most appropriate bands (bands after the shadow max index) were considered for further investigation in this study, specifically for the data fusion approach (Fig 5 a and b).


in-situ UAS (Field A)
Importance
10
5
0
-5
shadowMin
shadowMean
B2
shadowMax
B1
B4
B5
B8
B6
B3
B7
Attributes

407

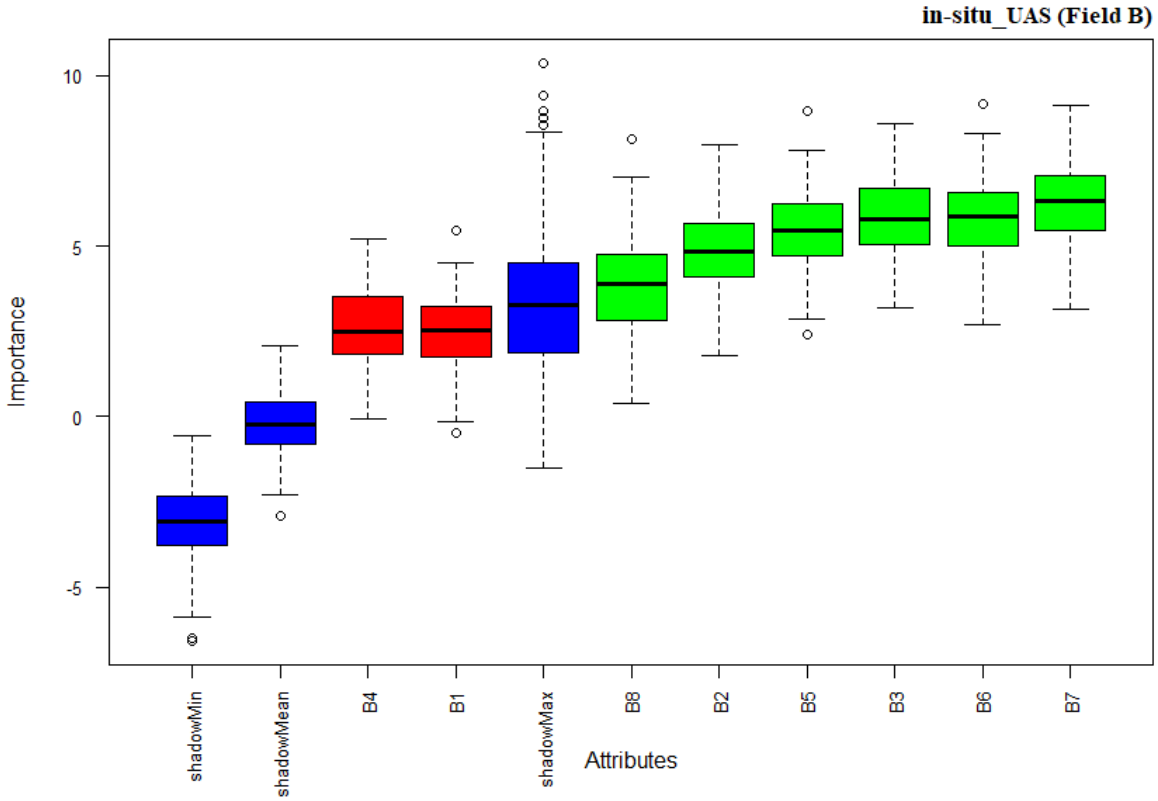

S2 (Field A)
Importance
10
5
0
-5
shadowMin
shadowMean
B8
shadowMax
B8A
B7
B6
B4
B3
B5
B11
B12
B2
Attributes

408

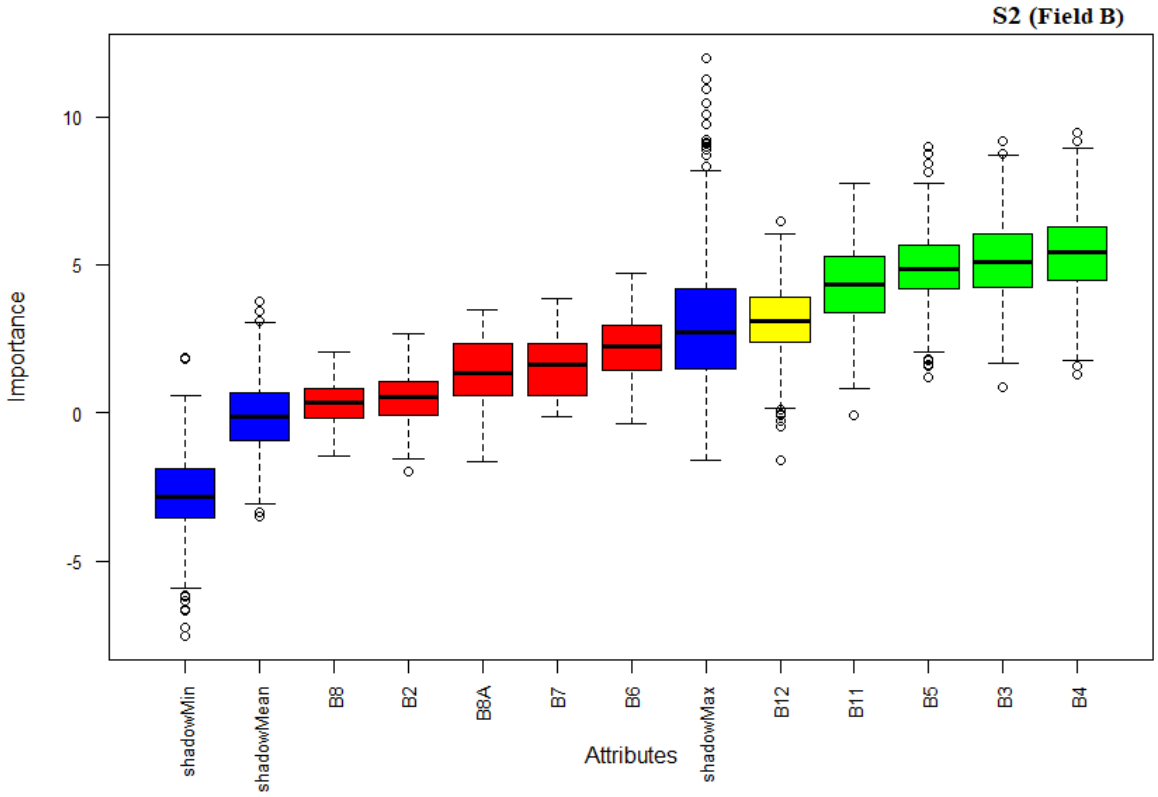


409

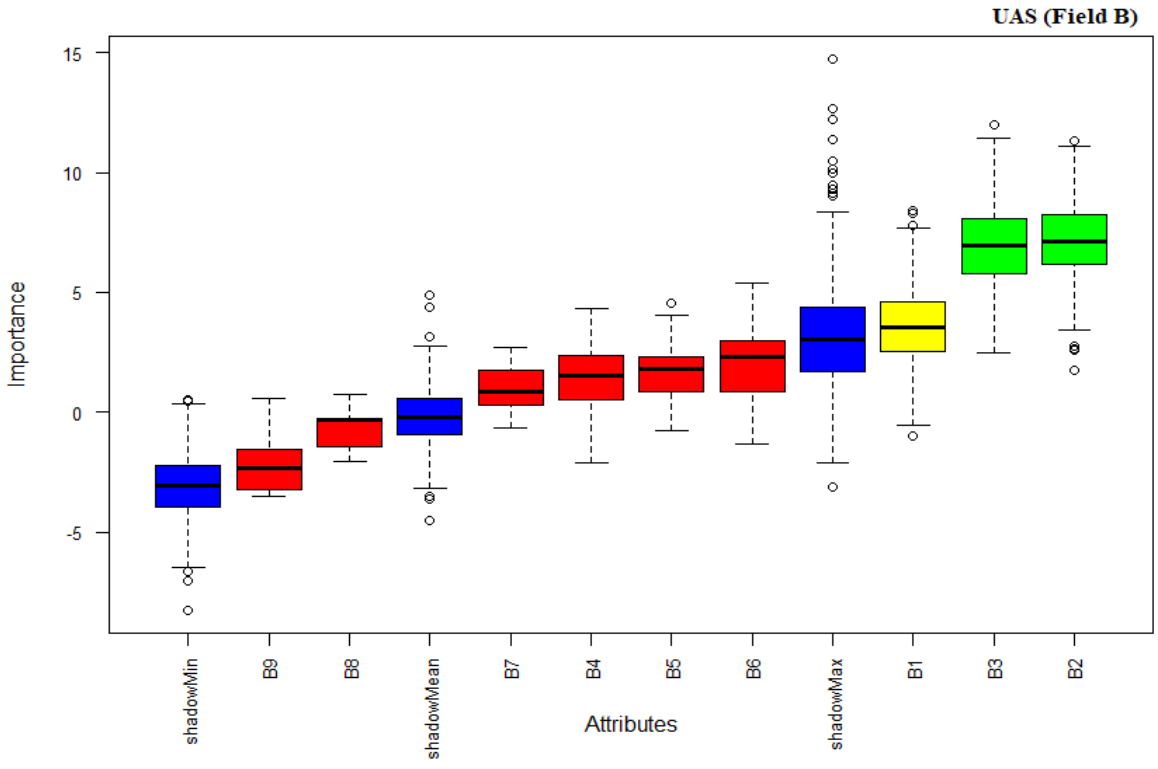
410 Fig. 5a: Variable importance of Field A for the three-platform data.



411



412



413

414 Fig. 5b: Variable importance of Field B for the three-platform data.

3.4. SOC prediction performance for the three platforms using individual and fused data

The results of the estimation models using SVM for the three datasets (in-situ, S2 and UAS) for each study field (A and B), including the prediction for the individual data, merged data (without considering variable importance), and fused data (considering only variable importance), are presented in Table 5. The obtained results varied for each category of the dataset. The estimations of SOC using the in-situ spectra under the individual data category provided the best results compared to both S2 and UAS data for Field A (R^2_{cv} = 0.51, RMSE = 0.23, RPIQ = 1.74, RPD = 1.56) and Field B (R^2_{cv} = 0.34, RMSE = 0.29, RPIQ = 1.36, RPD = 1.34). Whereas the result for Field A could be described as fair because the amount of variance explained by the model was 0.51%, the result for Field B was less satisfying, as the explained variance was only 0.35%. Table 5 also shows that for Field A, although the merged data (without considering VI) were better than UAS and S2 (Field A), these data, however, were outperformed by the in-situ data in terms of the predictive performance of SOC. In contrast, the merged data (Field B) were almost comparable to the in situ data but slightly better than the other individual and fused data sets. However, the error margin provided by the in-situ data was slightly better than that of the merged data (Field B) (Table 5).

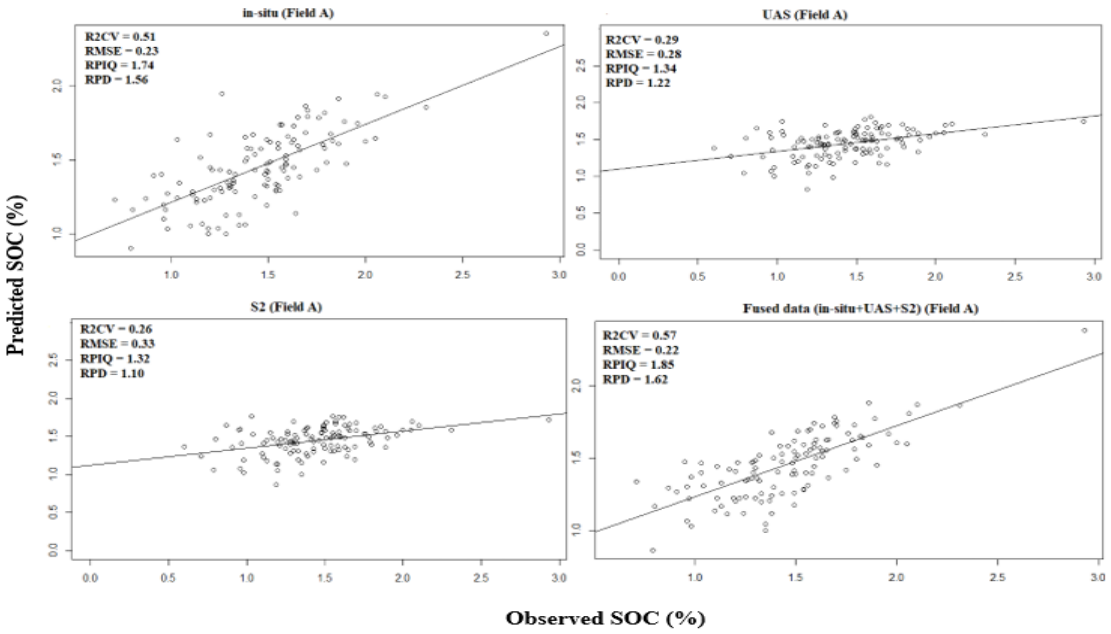
According to the results (Table 5), using the variable importance datasets to obtain improved SOC estimates from each study field differed substantially. In Field A, the fused approach using the three-platform data (in-situ, + UAS + S2) provided the best overall result; in Field B, the merged and in situ datasets provided slightly better results than the fused datasets. Additionally, for Field A, two of the fusing combinations (in-situ + S2 and UAS + S2) showed less predictive performance of SOC compared to the in situ dataset.

Table 5: Prediction performance showing fivefold leave-group-out cross-validation statistics for three platforms (in-situ, UAS, and S2) as well as different combinations among the dataset (fusion and merged format) at two different sites (Fields A and B) using SVM (support vector machine).

Datasets	R^2_{CV}	RMSE	RPIQ	RPD
Field A				
in-situ	0.51	0.23	1.74	1.56
UAS	0.29	0.28	1.34	1.22
S2	0.26	0.33	1.32	1.1
Merged data (without considering VI)				
In situ + UAS + S2	0.49	0.28	1.68	1.5
Fused data (with only VI)				
in-situ + UAS	0.52	0.23	1.74	1.54
in-situ + S2	0.47	0.26	1.59	1.42
UAS + S2	0.31	0.31	1.24	1.11
in-situ + UAS + S2	0.57	0.22	1.85	1.62
Field B				
in-situ	0.34	0.29	1.36	1.34
UAS	0.21	0.38	1.27	1.25
S2	0.18	0.41	1.18	0.91
Merged data (without considering VI)				
In-situ+UAS+S2	0.35	0.37	1.26	1.34
Fused data (with only VI)				
in-situ + UAS	0.31	0.30	1.33	1.30
in-situ + S2	0.24	0.33	1.22	1.22
UAS + S2	0.18	0.34	1.09	0.96
In-situ + UAS + S2	0.30	0.27	1.34	1.32

The scatterplots of predicted SOC using SVM against observed SOC are shown in Fig. 6a and 6b for three platform datasets and combinations from two different locations (Fields A and B). Overall, the in-situ and fused datasets (in situ+UAS+S2) look similar and differ in extreme values, with the former having more extreme values than the latter (Field A). Additionally, UAS and S2 also look identical,

448 but although more points were located closer to the 1:1 line, the arrangement of points looks different
449 (concentrated together) compared to the in-situ data. Moreover, the 1:1 line for both the UAS and S2
450 was shifted because of the arrangement of the points based on these data, indicating an underestimated
451 SOC content. The difference in predicting SOC among the various platforms was not the same, with
452 the in-situ data showing better results than the UAS and S2 data. However, the results improved with
453 the use of the fused datasets.



454

455 Fig. 6a. Observed vs predicted SOC content (%) for the individual platforms and the overall fused data
456 set in Field A using SVM

457 Visually, the scatterplots for the four datasets in Field B look similar but differ in terms of the extreme
458 value distribution. UAS and S2 had more extremely distributed values than the in-situ data. For
459 prediction, the in situ data were better. However, all the different platforms and combinations failed to

account for at least 0.5% of the variance in SOC for the area.

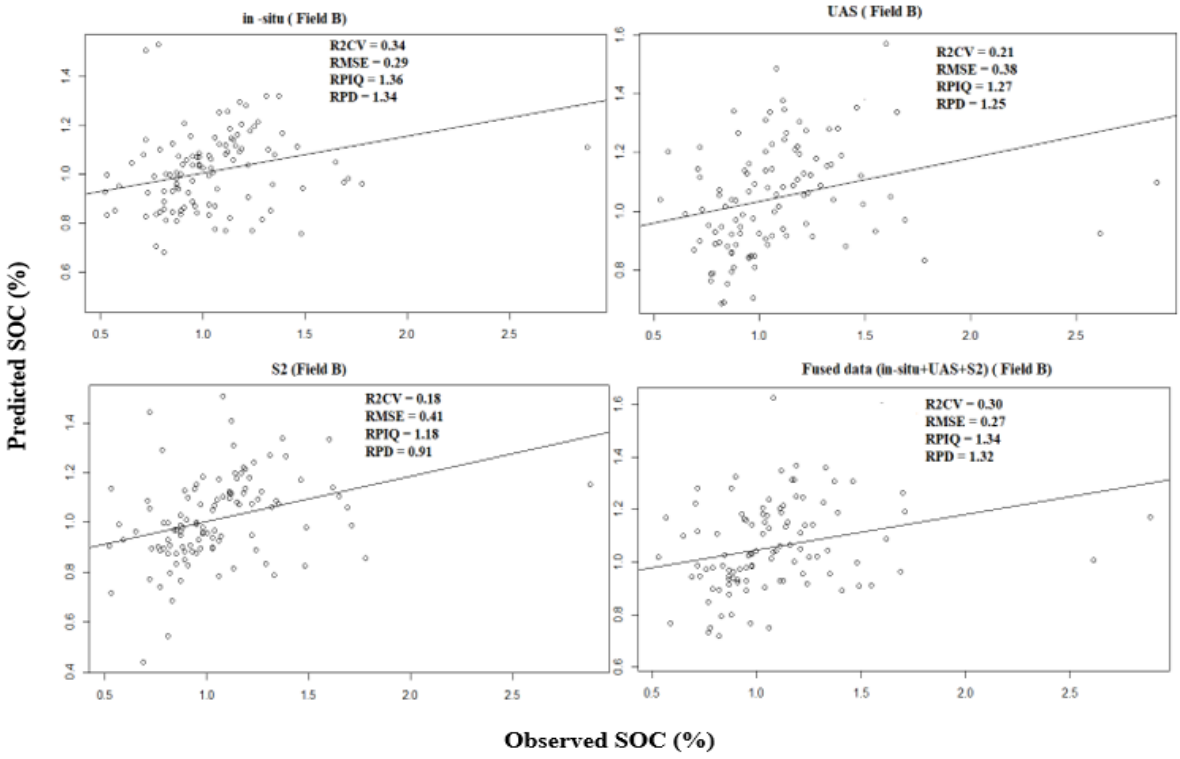


Fig. 6b. Observed vs predicted SOC content (%) for the individual platforms and the overall fused data set in Field B using SVM

Spatial distribution of SOC using in situ, UAS and S2 with different data combinations

The spatial distribution maps of SOC obtained using SVM predictive modeling from the in-situ spectral, UAS, and S2 data as well as different fusion combinations among these datasets, including fused data [(in-situ + UAS + S2)], (in-situ + S2), (in-situ + UAS), (S2+UAS)] and merged data (in-situ + UAS + S2), are illustrated in Figs. 7 and 8.

Fig. 7 shows that the in situ and fused data (in-situ+S2) spatial distributions of the SOC maps look similar to the lab-measured SOC map. In the case of UAS and S2, their spatial distribution maps also look similar, especially in both the upper and lower sections, but differ in the middle section, where S2 shows lower SOC content than UAS. Additionally, comparing both the S2 and UAS maps to the lab-measured map, the S2 SOC distribution map looks better than the map displayed by the UAS imagery, and both the UAS and the fused data (UAS+S2) maps look similar in the upper, middle, and lower halves of the study area. Finally, the fused data (in-situ + UAS) and (in-situ + UAS+S2) produced

identical SOC spatial distribution maps. However, the former detection of high SOC values is similar to the lab measurement compared to the latter.

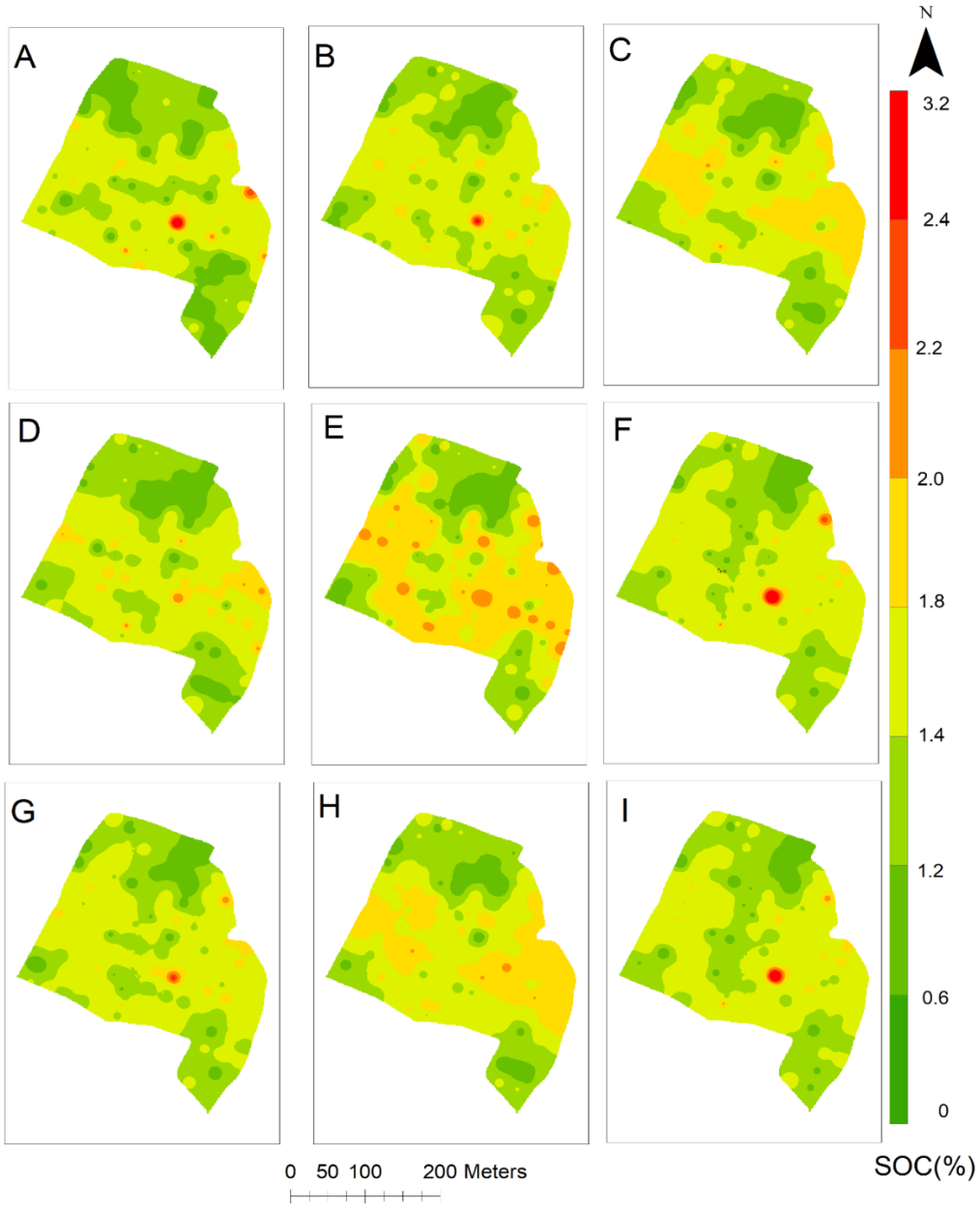


Fig. 7. Spatial SOC distribution maps in Field A based on the best prediction outcome from the individual platforms merged and fused dataset with SVM [lab measured (A), in-situ (B), UAS (C), S2 (D), merged data (in-situ+UAS+S2) (E), fused data (in-situ +UAS) (F), fused data (in-situ+S2) (G), fused data (UAS+S2) (H), fused data (in-situ+UAS+S2)]

Fig. 8 shows that none of the different platforms and the additional data combinations have any similarity with the lab-measured spatial distribution map of SOC, particularly regarding detecting high

SOC values. However, the fused data sets [in-situ +UAS, in-situ +S2, and in-situ +UAS] as well as the merged data exhibit similar characteristics of SOC distribution in the study area, with low SOC values in the middle to lower section. Additionally, UAS and S2 were almost identical except for the detection of high SOC values, where S2 was better than UAS.

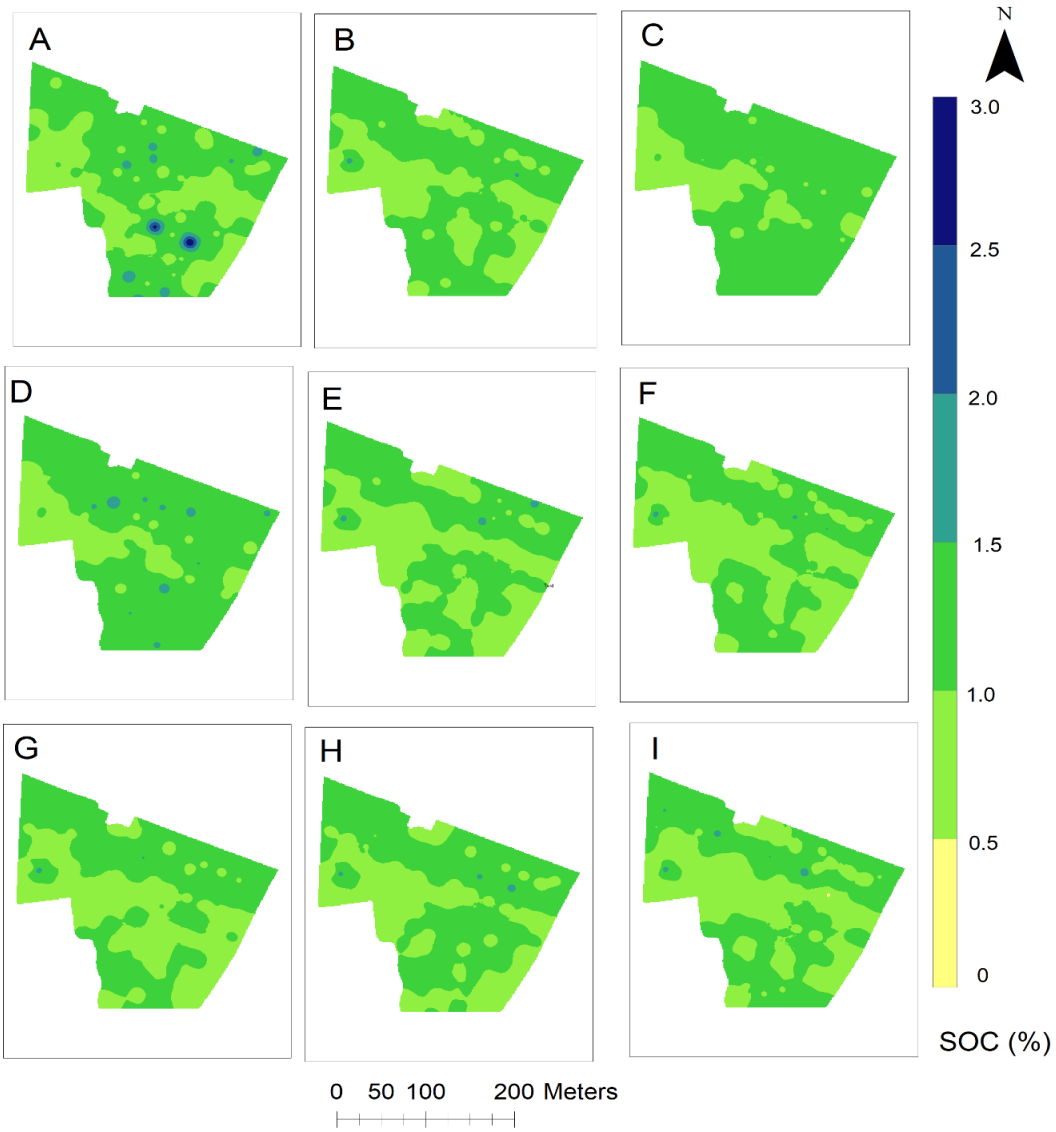


Fig. 8. Spatial SOC distribution maps in Field B based on the best prediction outcome from the individual platforms merged and fused dataset with SVM [lab measured (A), in situ (B), UAS (C), S2 (D), merged data (in situ+UAS+S2) (E), fused data (in-situ +UAS) (F), fused data (in-situ+S2) (G), fused data (UAS+S2) (H), fused data (in-situ+UAS+S2)].

4. Discussion

The differences in model output among the three platforms individually could be attributed to the conditions and approaches under which each dataset was obtained to estimate SOC. Normally, high spatial and spectral resolution remote sensing data (closer to the target object during measurement) are needed to help estimate SOC more accurately (Peng et al., 2015), especially in an area with low organic carbon content detailed information is required. The performance of the in-situ platform over other platforms is not surprising, as several studies have shown the superiority of spectroscopy (especially lab spectra) to remote sensing data (Hrelja et al., 2021; Angelopoulou et al., 2019; Gomez et al., 2008; Lagacherie et al., 2008, Stevens et al., 2008). The use of the in-situ spectral data ensured that all three platforms were prone to the same disturbed environmental conditions for a fair comparison. This environmental condition includes weather conditions, soil roughness, different measurement conditions, acquisition heights, noise interferences, spatial resolution, atmospheric conditions, pixel purity, and crop residues (Gomez et al., 2018; Lagacherie et al., 2008; Zhang and Zhou, 2016). However, the results show that these disturbing factors influence remote sensing data (S2, UAS) more than the in-situ data (Table 4). The study also showed that the simulated in-situ spectral data (UAS format as the final in situ-spectra) were not affected. It performed better than S2 and UAS data in estimating SOC (in Field A).

Similarly, the projection of the in-situ simulated bands into 10 m resolution for the data fusion approach was unaffected, as shown in Table 5. However, for Field B, although the simulated in-situ was better, all three platforms' predictive performances of SOC were poor because all the data sets had no relationship with SOC (Fig. 4). Moreover, as shown in Table 4, the statistical distribution of SOC in this field was poor, with a low mean value of 1.09%.

This study also evaluated the importance of predictors used to explain the variability of SOC. The Boruta algorithm technique, with a unique ability to reveal the significance of predictors to any machine learning model, was adopted (with an order of importance in ascending order) (section 3.3) before the data fusion approach to use only relevant variables to improve the accuracy of SOC. For Field A, for instance, the total number of bands for the three platforms using the data fusion approach was 27. Four

bands were discarded [in-situ (B1, B4, B5, B8, B6, B3, B7), UAS (B3, B1, B9, B5, B8, B6, B7), and S2 (B8A, B7, B6, B4, B3, B5, B11, B12, B2)], with B8 being the worst band (Figure 5a). Many studies have demonstrated the sensitivity of B8 to SOC (Mondal et al., 2017). The significance of this band in estimating SOC content within a study area in South Africa was emphasised by Odebiri et al. (2020). They reasoned that B8 could provide valuable information on the physiological state of vegetation concerning SOC, where chlorophyll content attributes are present. This could have influenced the S2 data used to estimate SOC in the current study (Table 5).

One of our primary goals was to assess the impact of data fusion across these platforms using various combinations after variable importance (VI) selection using the Boruta algorithm. As expected, the fusion approach (Field A) resulted in a better overall SOC estimation accuracy than the individual platforms, except for the in-situ platform, where SOC estimation was better than the following fusion approaches: in-situ+S2 and UAS+S2. One possible explanation is that the S2 data had a negative influence on both the in-situ and UAS data because, as shown in Fig. 3b, none of the bands of the S2 imagery was correlated with SOC. Moreover, the influence of B8 on S2, as stated above, cannot be ruled out. Regardless, the results (Table 5) show that the data fusion model with the combined three platform techniques using only variable importance datasets (in-situ+UAS+S2) can be used to map and improve the prediction of topsoil SOC with cross-validation R^2_{cv} values of 0.57, RMSE = 0.22, RPD = 1.62, and RPIQ = 1.85 in a study field (Field A) low in organic carbon content (with SOC values ranging between 0.6 and 2.93%). This improvement was most likely due to the minimal effect of the S2 data on the combined data of the in-situ and UAS platforms before prediction, as these two datasets fused provided the second-best result (R^2_{cv} = 0.52, RMSE = 0.23, RPIQ = 1.74, RPD = 1.54).

Furthermore, the removal of interference caused by each sensor technique may have resulted in comprehensive information for predicting the target parameter (Terra et al., 2019; Viscarra Rossel et al., 2006; Xu et al., 2019). This is because, after variable importance selection, the redundant and undesirable bands were discarded. For example, considering the merged data without variable importance selection, the model fails to account for at least 0.5% of the variance in SOC (R^2_{cv} = 0.49). This suggests that the three-platform merged data approach included redundant bands, influencing the

549 model output. This improvement in SOC predictive performance is also consistent with findings from
550 other studies (e.g., Knox et al., 2015; Johnson et al., 2019; Terra et al., 2019), although, in some
551 instances, the individual approaches were slightly better or better off than some of the fusion
552 combinations. It is worth noting that other researchers also obtained contradictory results (Clairotte et
553 al., 2016; Viscarra Rossel et al., 2006) using a data fusion approach to estimate SOC, where no
554 improvement was reported. The study by Clairotte et al. (2016) attributed this phenomenon to several
555 disturbing factors (e.g., noise, undesirable information, or artefacts) when more than one spectral range
556 or band is merged.

557 Moreover, the baseline height and other spectral attributes throughout the entire spectral range or bands
558 may negatively influence the prediction accuracy of the soil properties under consideration (Muller and
559 Decamps, 2000). In the case of Field B, generally, the obtained result was very poor, irrespective of
560 whether the datasets were used individually, in merged form, or even in different combination formats
561 before the data fusion application. This could result from the fact that all three platform bands (Fig. 4d,
562 e, and f) exhibited either no or poor correlation with SOC, thereby negatively affecting the predictive
563 capability of SOC. Moreover, many redundant bands were detected and removed after the application
564 of the Boruta algorithms (Fig. 5b) compared to Field A, which did not even help because the obtained
565 results after the data fusion application were slightly poorer than using the in-situ data individually or
566 the merged data approach. One possible issue that could have warranted this problem is using the same
567 data fusion techniques for different study fields with varying soil types. This might not work out, as was
568 the case for the current study. As a result, caution should be exercised when selecting data fusion
569 approaches for different study fields to avoid introducing unexpected sources of error that can reduce
570 prediction accuracy (Javadi et al., 2021). Therefore, further studies on the impact of using the same data
571 fusion techniques on different study fields under different soil types and conditions could help verify
572 the possibility of obtaining a universal data fusion approach to improve SOC estimates. Additionally,
573 the poor SOC distribution statistics (Table 4) for this study location, with a low mean score of 1.09%,
574 as stated earlier, could probably be a contributing factor. Overall, data fusion can generate better, well-

resolved, and apparent results than the use of a single data set, especially in poor to medium organic carbon fields.

The observed vs predicted results for SOC predictive accuracy for study locations A and B using the in-situ, UAS, and S2 data individually and under fusion conditions are shown in the scatterplot (Fig. 6 a and b). The extreme values, especially in Field B, can be attributed to the spectral information content of the sensors under different measurement conditions as well as other soil types and the ability of each sensor to detect both low and high SOC at the study location. Moreover, each sensor platform was prone to disturbing external environmental conditions that generally affect proximal and remote sensing measurements in the field, as stated above. It is worth mentioning that extreme points cannot be classified as negative outliers. During the data analysis process, outliers within the datasets were checked based on their impact on the prediction accuracy of SOC using the enpls technique. As stated earlier, all undesirable data were removed. According to Balakrishnan (1994) and Frost (2019), outliers can affect the prediction accuracy, but some outliers are still significant, and removing them may affect the model output.

Overall, from the output maps (Field A), SOC is moderately prevalent in the study area, especially in the middle part (Fig. 7). Compared to the in-situ platform, both UAS and S2 depict a slightly different scenario, especially UAS showing a field with semi-high SOC in the middle part, which can be attributed to the poor prediction accuracy of SOC using these data sets. In addition, UAS and S2 also underestimated the highest SOC values in the area due to the various wavelength variables used in the development of each prediction model. According to Wulder et al. (2015), selecting better S2 imagery under ideal conditions with fewer interferences could help obtain a reliable prediction accuracy, thereby providing improved mapping of a given study area under consideration. Additionally, differences in acquisition date, the presence of clouds, masking techniques, or better masking effects (Immitzer et al., 2016; Steinberg et al., 2016) could also affect the accuracy of remote sensing data sets in mapping soil attributes. Although the fused datasets were reasonably comparable to the lab-generated map, the highest SOC values were likely overestimated due to fusing their platform datasets. This also shows that the IDW spatial distribution map generated by the fusion of different sensor data needs further research. For

Field B (Fig. 8), although all the individual and fused data underestimated the poor distribution of SOC in the study area in reference to lab-measured SOC values, none of these data sets was able to detect the highest SOC value, which could result from the poor prediction of SOC by each data set using the SVM model.

5. Conclusion

The study verified the impact of estimating and mapping SOC in study fields low in organic carbon content by merging high-resolution simulated in-situ data with Sentinel-2 (S2) and Unmanned Aircraft Systems (UAS) data through a data fusion approach. Prior to SOC estimation, all the data were converted to the same spatial resolution, and a variable importance approach was also applied.

This study confirmed that data fusion could generate better, more intense, and apparent results than the use of a single data set, especially in poor to medium organic carbon fields. The findings also demonstrated that the data fusion approach could effectively reduce the overall error in SOC modeling. Although it was successful in the Nova Ves study field (Field A), the accuracy of SOC did not improve for the Udrince study field, likely because all three platform bands exhibited either no or poor correlation with SOC, thereby negatively affecting the predictive capability of SOC. Additionally, the obtained results for this study field were very poor, irrespective of whether the datasets were used individually, in merged form, or even in different combination formats before the data fusion application. As a result, caution should be exercised when selecting data fusion approaches for other study fields to avoid introducing unexpected sources of error that can reduce prediction accuracy. Although there were similarities in the spatial distribution map between the individual approaches, the in situ platform better resembled the measured data in Field A. In contrast, for Field B, none of the platforms resembled the measured map because no correlation existed between these data sets and SOC. Nonetheless, future studies to verify the effectiveness of the fusion approach on both proximal and remote sensing data are highly recommended, mainly using remote sensing data with fewer defects and other modeling techniques.

628 Funding.

629 This study was supported by an internal grant from the Czech University of Life Sciences Prague, project
630 No. SV20-5-21130, and by the Technology Agency of the Czech Republic project No. SS02030018.

631 Acknowledgements:

632 The authors also acknowledge the support of the European Regional Development Fund Project Centre
633 for the investigation of synthesis and transformation of nutritional substances in the food chain in
634 interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of
635 soil contamination risks for the quality of agricultural products (No.
636 CZ.02.1.01/0.0/0.0/16_019/0000845)

637 Compliance with Ethical Standards:

638 The authors declare no conflicts of interest.

639 References

640 Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., & Bochtis, D. (2019). Remote sensing
641 techniques for soil organic carbon estimation: A review. Remote Sensing, 11(6), 676.

642 Balakrishnan, N. (1994). Order statistics from nonidentical exponential random variables and some
643 applications. Computational statistics & data analysis, 18(2), 203-238.

644 Bayer, A. D., Bachmann, M., Rogge, D., Müller, A., & Kaufmann, H. (2016). Combining field and
645 imaging spectroscopy to map soil organic carbon in a semiarid environment. IEEE Journal of Selected
646 Topics in Applied Earth Observations and Remote Sensing, 9(9), 3997-4010.

647 Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Taylor, G. R., Hill, J., Whiting, M. L., & Sommer, S.
648 (2009). Using imaging spectroscopy to study soil properties. Remote sensing of environment, 113, S38-
649 S55.

650 Biney, J.K.M., Saberioon, M., Borůvka, L., Houška, J., Vašát, R., Chapman Agyeman, P., Coblinski,
651 J.A. and Klement, A. (2021). Exploring the Suitability of UAS-Based Multispectral Images for

652 Estimating Soil Organic Carbon: Comparison with Proximal Soil Sensing and Spaceborne Imagery.
653 Remote Sensing, 13(2), 308.

654 Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N., Ben Aissa, N. &
655 Mougenot, B. (2019). Soil texture estimation using radar and optical data from Sentinel-1 and Sentinel-
656 2. Remote Sensing, 11(13), 1520.

657 Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M. and
658 van Wesemael, B. (2019). Evaluating the capability of the Sentinel 2 data for soil organic carbon
659 prediction in croplands. ISPRS Journal of Photogrammetry and Remote Sensing, 147, 267-282.

660 Chang, C. W., & Laird, D. A. (2002). Near-infrared reflectance spectroscopic analysis of soil C and N.
661 Soil Science, 167(2), 110-116.

662 Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H.... & Shi, Z. (2019). A high-resolution
663 map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and
664 its implications for pollution. Science of the Total Environment, 655, 273-283.

665 Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P., Bernoux, M., & Barthès, B. G.
666 (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance
667 spectroscopy. Geoderma, 276, 41-52.

668 Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

669 Crucil, G., Castaldi, F., Aldana-Jague, E., van Wesemael, B., Macdonald, A., & Van Oost, K. (2019).
670 Assessing the performance of UAS-compatible multi-spectral and hyperspectral sensors for soil organic
671 carbon prediction. Sustainability, 11(7), 1889.

672 Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and
673 feedbacks to climate change. Nature, 440(7081), 165-173.

674 Diggle, P. J., & Ribeiro, P. J. (2007). Geostatistical design. Model-based geostatistics, 199-212.

675 Elhag, M., & Bahrawi, J. A. (2017). Soil salinity mapping and hydrological drought indices assessment
676 in arid environments based on remote sensing techniques. Geoscientific Instrumentation, Methods and
677 Data Systems, 6(1), 149-158.

678 Enríquez-de-Salamanca, Á., Díaz-Sierra, R., Martín-Aranda, R. M., & Santos, M. J. (2017).
679 Environmental impacts of climate change adaptation. Environmental Impact Assessment Review, 64,
680 87-96.

681 European Space Agency. (2016). Sen2Cor 2.2.1-Software Release Note

682 Feilhauer, H., Thonfeld, F., Faude, U., He, K. S., Rocchini, D., & Schmidtlein, S. (2013). Assessing
683 floristic composition with multi-spectral sensors—A comparison based on monotemporal and
684 multiseasonal field spectra. International Journal of Applied Earth Observation and Geoinformation, 21,
685 218-229.

686 Frost. J. (2019). Guidelines for Removing and Handling Outliers in Data.
687 https://statisticsbyjim.com/basics/remove-outliers/(accessed on 23 October 2019).

688 Gholizadeh, A., Žižala, D., Saberioon, M., & Borůvka, L. (2018). Soil organic carbon and texture
689 retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sensing of
690 Environment, 218, 89-103.

691 Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P.... & Briottet, X. (2018).
692 Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multi-
693 spectral to hyperspectral scenarios. Remote Sensing of Environment, 204, 18-30.

694 Gomez, C., Rossel, R. A. V., & McBratney, A. B. (2008). Soil organic carbon prediction by
695 hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma,
696 146(3-4), 403-411.

697 Grunwald, S., Vasques, G. M., & Rivero, R. G. (2015). Fusion of soil and remote sensing data to model
698 soil properties. Advances in Agronomy, 131, 1-109.

699 Guenet, B., Gabrielle, B., Chenu, C., Arrouays, D., Balesdent, J., Bernoux, M., Bruni, E., Caliman, J.P.,
700 Cardinael, R., Chen, S. and Ciais, P. (2021). Can N2O emissions offset the benefits from soil organic
701 carbon storage?. Global Change Biology, 27(2), 237-256.

702 Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report,
703 14(1), 5-16.

704 Guo, L. B., & Gifford, R. M. (2002). Soil carbon stocks and land use change: a meta analysis. Global
705 change biology, 8(4), 345-360.

706 Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using
707 support vector machines. Machine learning, 46(1), 389-422.

708 Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE,
709 85(1), 6-23.

710 Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview
711 and comparison of machine-learning techniques for classification purposes in digital soil mapping.
712 Geoderma, 265, 62-77

713 Heuvelink, G. B., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R.... & Sanderman,
714 J. (2021). Machine learning in space and time for modelling soil organic carbon change. European
715 Journal of Soil Science, 72(4), 1607-1623.

716 Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R. and
717 Pozza, L. (2015). Potential of integrated field spectroscopy and spatial analysis for enhanced assessment
718 of soil contamination: a prospective review. Geoderma, 241, 180-209.

719 Hrelja, I., Šestak, I., & Bogunović, I. (2021, April). Estimation of soil organic matter using proximal
720 and satellite sensors after a wildfire in Mediterranean Croatia. In EGU General Assembly Conference
721 Abstracts (pp. EGU21-3896).

722 Hugue, F., Lapointe, M., Eaton, B. C., & Lepoutre, A. (2016). Satellite-based remote sensing of running
723 water habitats at large riverscape scales: Tools to analyse habitat heterogeneity for river ecosystem
724 management. Geomorphology, 253, 353-369.

725 Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First experience with Sentinel-2 data for crop and tree
726 species classifications in central Europe. Remote sensing, 8(3), 166.

727 IUSS Working Group, W. R. B. (2006). World reference base for soil resources. World Soil Resources
728 Report, 103.

729 Javadi, S. H., Munnaf, M. A., & Mouazen, A. M. (2021). Fusion of Vis-NIR and XRF spectra for
730 estimation of key soil attributes. Geoderma, 385, 114851.

731 Ji, W., Adamchuk, V. I., Chen, S., Su, A. S. M., Ismail, A., Gan, Q.... & Biswas, A. (2019). Simultaneous
732 measurement of multiple soil properties through proximal sensor data fusion: A case study. Geoderma,
733 341, 111-128.

734 Johnson, J. M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K. D., & Saito, K. (2019). Near-
735 infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice
736 fields in sub-Saharan Africa. Geoderma, 354, 113840.

737 Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of
738 the state-of-the-art. Information fusion, 14(1), 28-44.

739 Knox, N. M., Grunwald, S., McDowell, M. L., Bruland, G. L., Myers, D. B., & Harris, W. G. (2015).
740 Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy.
741 Geoderma, 239, 229-239.

742 Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J.
743 (2012). Sensing soil properties in the laboratory, in-situ, and on-line: a review. Advances in Agronomy,
744 114, 155-223.

745 Kweon, G., Lund, E., & Maxton, C. (2013). Soil organic matter and cation-exchange capacity sensing
746 with on-the-go electrical conductivity and optical sensors. Geoderma, 199, 80-89.

747 Lagacherie, P., Baret, F., Feret, J. B., Netto, J. M., & Robbez-Masson, J. M. (2008). Estimation of soil
748 clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote
749 Sensing of Environment, 112(3), 825-835.

750 Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. science,
751 304(5677), 1623-1627.

752 Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for
753 soil organic carbon mapping and their implications: A review. Geoderma, 352, 395-413.

754 Li, Z., Zhang, H. K., Roy, D. P., Yan, L., Huang, H., & Li, J. (2017). Landsat 15-m panchromatic-
755 assisted downscaling (LPAD) of the 30-m reflective wavelength bands to Sentinel-2 20-m resolution.
756 Remote Sensing, 9(7), 755.

757 Liao, Y., Li, D., & Zhang, N. (2018). Comparison of interpolation models for estimating heavy metals
758 in soils under various spatial characteristics and sampling methods. Transactions in GIS, 22(2), 409-
759 434.

760 McCarty, G. W., Reeves, J. B., Reeves, V. B., Follett, R. F., & Kimble, J. M. (2002). Mid-infrared and
761 near-infrared diffuse reflectance spectroscopy for soil carbon measurement.

762 Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., & Mukhopadhyay, A. (2017). Spatial soil
763 organic carbon (SOC) prediction by regression kriging using remote sensing data. The Egyptian Journal
764 of Remote Sensing and Space Science, 20(1), 61-70.

765 Munnaf, M. A., Haesaert, G., Van Meirvenne, M., & Mouazen, A. M. (2020). Site-specific seeding
766 using multi-sensor and data fusion techniques: A review. Advances in Agronomy, 161, 241-323.

767 Murray, I. (1988). C.S. Creaser, A.M.C. Davies (Eds.), Analytical Applications of Spectroscopy, 1988,
768 Royal Society of Chemistry, London (1988)

769 Odebiri, O., Mutanga, O., Odindi, J., Peerbhay, K., & Dovey, S. (2020). Predicting soil organic carbon
770 stocks under commercial forest plantations in KwaZulu-Natal Province, South Africa using remotely
771 sensed data. GIScience & Remote Sensing, 57(4), 450-463.

O'rourke, S. M., Stockmann, U., Holden, N. M., McBratney, A. B., & Minasny, B. (2016). An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. Geoderma, 279, 31-44.

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., & Greve, M. H. (2015). Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. PloS one, 10(11), e0142295.

Post, W. M., & Kwon, K. C. (2000). Soil carbon sequestration and land-use change: processes and potential. Global change biology, 6(3), 317-327.

Qiao, P., Lei, M., Yang, S., Yang, J., Guo, G., & Zhou, X. (2018). Comparing ordinary kriging and inverse distance weighting for soil as pollution in Beijing. Environmental Science and Pollution Research, 25(16), 15597-15608.

Roy, D. P., Li, J., Zhang, H. K., & Yan, L. (2016). Best practices for the reprojection and resampling of Sentinel-2 Multi Spectral Instrument Level 1C data. Remote Sensing Letters, 7(11), 1023-1032.

Sabetizade, M., Gorji, M., Roudier, P., Zolfaghari, A. A., & Keshavarzi, A. (2021). Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. Catena, 196, 104844

Schmidt, K., Behrens, T., Friedrich, K., & Scholten, T. (2010). A method to generate soilscapes from soil maps. Journal of Plant Nutrition and Soil Science, 173(2), 163-172.

Schwartz, G., Eshel, G., & Ben-Dor, E. (2011). Reflectance spectroscopy as a tool for monitoring contaminated soils. Soil Contam, 6790.

Shi, T., Wang, J., Chen, Y., & Wu, G. (2016). Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. International journal of applied earth observation and geoinformation, 52, 95-103.

Signal Developers, (2013). Signal: signal processing URL: http://r-forge.r-project.org/projects/signal (2013)

797 Simone, G., Farina, A., Morabito, F. C., Serpico, S. B., & Bruzzone, L. (2002). Image fusion techniques
798 for remote sensing applications. Information fusion, 3(1), 3-15.

799 Skjemstad, J. O., Baldock, J. A., Carter, M. R., & Gregorich, E. G. (2008). Soil sampling and methods
800 of analysis. Total and organic carbon'. 2nd edn.(Eds MR Carter, EG Gregorich), 225-237.

801 Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., & Foerster, S. (2016). Prediction of common surface
802 soil properties based on Vis-NIR airborne and simulated EnMAP imaging spectroscopy data: Prediction
803 accuracy and influence of spatial resolution. Remote Sensing, 8(7), 613

804 Stenberg, B., Rossel, R. A. V., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared
805 spectroscopy in soil science. Advances in agronomy, 107, 163-215.

806 Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil
807 organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PloS one,
808 8(6), e66409.

809 Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., & Ben-Dor, E. (2008).
810 Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils.
811 Geoderma, 144(1-2), 395-404.

812 Tabacchi, E., Lambs, L., Guilloy, H., Planty-Tabacchi, A. M., Muller, E., & Decamps, H. (2000).
813 Impacts of riparian vegetation on hydrological processes. Hydrological processes, 14(16-17), 2959-
814 2976.

815 Taghizadeh-Mehrjardi, R., Nabiollahi, K., & Kerry, R. (2016). Digital mapping of soil organic carbon
816 at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma, 266, 98-110.

817 Terra, F. S., Rossel, R. A. V., & Demattê, J. A. (2019). Spectral fusion by Outer Product Analysis (OPA)
818 to improve predictions of soil organic C. Geoderma, 335, 35-46.

819 Vaudour, E., Gilliot, J. M., Bel, L., Bréchet, L., Hamiache, J. O. N. A. S., Hadjar, D., & Lemonnier, Y.
820 (2014). Uncertainty of soil reflectance retrieval from SPOT and RapidEye multi-spectral satellite images

using a per-pixel bootstrapped empirical line atmospheric correction over an agricultural region. International Journal of Applied Earth Observation and Geoinformation, 26, 217-234.

Vaudour, E., Gomez, C., Fouad, Y., & Lagacherie, P. (2019). Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. Remote Sensing of Environment, 223, 21-33.

Viscarra Rossel, R., Walvoort, D. J. J., McBratney, 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(1-2), 59-75.

Vohland, M., Ludwig, B., Seidel, M., & Hutengs, C. (2022). Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in-situ than for laboratory-based modelling approaches. Geoderma, 405, 115426.

Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2014). Determination of soil properties with visible to near-and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma, 223, 88-96.

Wang, Q., Blackburn, G. A., Onojeghuo, A. O., Dash, J., Zhou, L., Zhang, Y., & Atkinson, P. M. (2017). Fusion of Landsat 8 OLI and Sentinel-2 MSI data. IEEE Transactions on Geoscience and Remote Sensing, 55(7), 3885-3899.

Wang, Y., Guan, Z., & Zhao, T. (2019). Sample size determination in geotechnical site investigation considering spatial variation and correlation. Canadian Geotechnical Journal, 56(7), 992-1002.

Wang, X., Zhang, Y., Atkinson, P. M., & Yao, H. (2020). Predicting soil organic carbon content in Spain by combining Landsat TM and ALOS PALSAR images. International Journal of Applied Earth Observation and Geoinformation, 92, 102182.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists. (second edition), John Wiley & Sons, Chichester.

Wulder, M. A., Hilker, T., White, J. C., Coops, N. C., Masek, J. G., Pflugmacher, D., & Crevier, Y. (2015). Virtual constellations for global terrestrial monitoring. Remote Sensing of Environment, 170, 62-76.

Xu, D., Chen, S., Rossel, R. V., Biswas, A., Li, S., Zhou, Y., & Shi, Z. (2019). X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. Geoderma, 352, 61-69.

Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., & Shi, Z. (2019). Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. European Journal of Soil Science, 70(1), 162-173.

Yimer, F., Ledin, S., & Abdelkadir, A. (2006). Soil organic carbon and total nitrogen stocks as affected by topographic aspect and vegetation in the Bale Mountains, Ethiopia. Geoderma, 135, 335-344.

Yao, X., Yu, K., Deng, Y., Zeng, Q., Lai, Z., & Liu, J. (2019). Spatial distribution of soil organic carbon stocks in Masson pine (Pinus massoniana) forests in subtropical China. Catena, 178, 189-198.

Zhang, D., & Zhou, G. (2016). Estimation of soil moisture from optical and thermal remote sensing: A review. Sensors, 16(8), 1308.

Žížala, D., Minařík, R., & Zádorová, T. (2019). Soil organic carbon mapping using multi-spectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions. Remote Sensing, 11(24), 2947.

remote sensing

MDPI

Article

The Brazilian Soil Spectral Service (BraSpecS): A User-Friendly System for Global Soil Spectra Communication

José A. M. Demattê [1,*], Ariane Francine da Silveira Paiva [1], Raul Roberto Poppiel [1], Nícolas Augusto Rosin [1], Luis Fernando Chimelo Ruiz [1], Fellipe Alcantara de Oliveira Mello [1], Budiman Minasny [2], Sabine Grunwald [3], Yufeng Ge [4], Eyal Ben Dor [5], Asa Gholizadeh [6], Cecile Gomez [7], Sabine Chabrillat [8,9], Nicolas Francos [5], Shamsollah Ayoubi [10], Dian Fiantis [11], James Kobina Mensah Biney [6], Changkun Wang [12], Abdelaziz Belal [13], Salman Naimi [10], Najmeh Asgari Hafshejani [10], Henrique Bellinaso [1], Jean Michel Moura-Bueno [14] and Nélida E. Q. Silvero [1]

1 Department of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo, Pádua Dias Avenue, 11, Piracicaba 13418-900, Brazil; ariane.silveira@usp.br (A.F.d.S.P.); raulpoppiel@usp.br (R.R.P.); narosin@usp.br (N.A.R.); luisruiz@usp.br (L.F.C.R.); fellipeamello@usp.br (F.A.d.O.M.); henrique.bellinaso@sp.gov.br (H.B.); neli.silvero@usp.br (N.E.Q.S.)
2 School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia; budiman.minasny@sydney.edu.au
3 Soil and Water Sciences Department, University of Florida, 2181 McCarty Hall, P.O. Box 110290, Gainesville, FL 32611, USA; sabgru@ufl.edu
4 Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583, USA; yge2@unl.edu
5 Department of Geography, Porter School of Environmental and Earth Science, Faculty of Exact Science, Tel Aviv University, P.O. Box 39040, Tel Aviv 6997801, Israel; bendor@post.tau.ac.il (E.B.D.); nicolasf@mail.tau.ac.il (N.F.)
6 Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic; gholizadeh@af.czu.cz (A.G.); biney@af.czu.cz (J.K.M.B.)
7 Laboratoire d'Etude des Interactions entre Sol-Agrosystème-Hydrosystème, University of Montpellier, National Research Institute for Agriculture, Food and Environment, Institute of Research for Development, Montpellier SupAgro, 34060 Montpellier, France; cecile.gomez@ird.fr
8 Helmholtz Center Potsdam GFZ German Research Centre for Geosciences, Remote Sensing and Geoinformatics, Telegrafenberg, 14473 Potsdam, Germany; chabrillat@ifbk.uni-hannover.de
9 Institute of Soil Science, Germany Leibniz University Hannover, Herrenhäuser Straße 2, 30419 Hannover, Germany
10 Department of Soil Science, College of Agriculture, Isfahan University of Technology, Isfahan 84156-83111, Iran; ayoubi@iut.ac.ir (S.A.); s.naimi@ag.iut.ac.ir (S.N.); Najme.Asgari@ag.iut.ic.ir (N.A.H.)
11 Department of Soil Science, Faculty of Agriculture, Universitas Andalas, Padang 25163, Indonesia; dianfiantis@faperta.unand.ac.id
12 State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China; ckwang@issas.ac.cn
13 Agriculture Application, Soil and Marine Division, National Authority for Remote Sensing and Space Sciences (NARSS), Alfmaskan, Cairo P.O. Box 1564, Egypt; abelal@narss.sci.eg
14 Department of Agronomy, University of Cruz Alta, Highway Jacob Della Méa, km 5.6, Cruz Alta 98005-972, Brazil; jbueno@unicruz.edu.br
* Correspondence: jamdemat@usp.br

Citation: Demattê, J.A.M.; Paiva, A.F.d.S.; Poppiel, R.R.; Rosin, N.A.; Ruiz, L.F.C.; Mello, F.A.d.O.; Minasny, B.; Grunwald, S.; Ge, Y.; Ben Dor, E.; et al. The Brazilian Soil Spectral Service (BraSpecS): A User-Friendly System for Global Soil Spectra Communication. *Remote Sens.* **2022**, *14*, 740. https://doi.org/10.3390/rs14030740

Academic Editor: Yue Wu, Kai Qin, Maoguo Gong and Qiguang Miao

Received: 19 January 2022
Accepted: 2 February 2022
Published: 5 February 2022

Abstract: Although many Soil Spectral Libraries (SSLs) have been created globally, these libraries still have not been operationalized for end-users. To address this limitation, this study created an online Brazilian Soil Spectral Service (BraSpecS). The system was based on the Brazilian Soil Spectral Library (BSSL) with samples collected in the Visible–Near–Short-wave infrared (vis–NIR–SWIR) and Mid-infrared (MIR) ranges. The interactive platform allows users to find spectra, act as custodians of the data, and estimate several soil properties and classification. The system was tested by 500 Brazilian and 65 international users. Users accessed the platform (besbbr.com.br), uploaded their spectra, and received soil organic carbon (SOC) and clay content prediction results via email. The BraSpecS prediction provided good results for Brazilian data, but performed variably for other countries.

Prediction for countries outside of Brazil using local spectra (External Country Soil Spectral Libraries, ExCSSL) mostly showed greater performance than BraSpecS. Clay R^2 ranged from 0.5 (BraSpecS) to 0.8 (ExCSSL) in vis–NIR–SWIR, but BraSpecS MIR models were more accurate in most situations. The development of external models based on the fusion of local samples with BSSL formed the Global Soil Spectral Library (GSSL). The GSSL models improved soil properties prediction for different countries. Nevertheless, the proposed system needs to be continually updated with new spectra so they can be applied broadly. Accordingly, the online system is dynamic, users can contribute their data and the models will adapt to local information. Our community-driven web platform allows users to predict soil attributes without learning soil spectral modeling, which will invite end-users to utilize this powerful technique.

Keywords: proximal soil sensing; soil spectral library; spectroscopy; soil analysis; soil quality; precision agriculture; community practice; soil health monitoring

1. Introduction and Contextualization

Soil is an important component of the environment as it offers vital services such as food production, clean water, and carbon sequestration [1]. To achieve sustainable use of these resources, the world's soil community must form partnerships and seek reliable methods for obtaining its information. So far, the traditional soil laboratory has been the most common way to obtain soil data, but it is not environmentally friendly, and it becomes expensive when large amount of samples need to be analyzed [2]. This is especially crucial in developing countries, where farmers either do not conduct soil analysis due to high costs or the absence of locally accessible laboratory services. Despite the disadvantages, traditional laboratory analysis is, and will continue to be, the most suitable way to obtain soil data. However, alternatives such as soil spectroscopy have proved to be a convenient way to optimize soil analysis and a rapid alternative to disseminate the results to all interested parties. Indeed, a recent study [3] proved that wet laboratories' analysis results have more variation between laboratories than between spectral sensors.

Soil spectroscopy is well-documented in the literature with a strong background in science and evidence [4–7]. Understanding the infrared phenomena on soil has provided researchers with confidence in its use to quantify soil properties, with much research conducted post-2000. Soil researchers are encouraged by the power of the infrared technique and seek a global communication tool, such as the so-called Soil Spectral Libraries (SSLs). The first publication on developing an SSL with global soil reflectance data was presented by Stoner and Baumgartner in 1981 [8], followed by others, [9,10]. The latter example had 92 participating countries. In addition, countries have developed their own SSLs, such as the Brazilian Soil Spectral Library (BSSL) [11,12], the Czech Republic [13], France [14], Denmark [15], Mozambique [16], Spain [17], Australia [18], China [19–21], USA [22–24], New Zealand [25], and Tajikistan [26].

Soil spectroscopy is mostly understood by researchers and has gathered hundreds of papers in the last 60 years [6,7]. Despite the substantial research, the technique has not advanced to the end-users. Traditional wet chemistry soil analysis has continued to be performed since early 1800. There is no doubt regarding the importance of conventional wet chemistry lab analysis, but the demand for soil analysis is increasing and the dependency on wet chemistry is not be sustainable [27].

Many researchers have demonstrated the efficiency of soil spectroscopy and robust predictive capabilities for multiple soil properties [28], summarized in [7]. In addition, the MIR spectral range has been proven to provide superior prediction compared to vis–NIR–SWIR spectra [29–31]. The SSLs, thus, are important research initiatives [10] but the data are only available through scientific journal publications. Other initiatives have adopted an 'open spectra' approach. This includes regional programs such as the ICRAF-ISRIC Soil VNIR Spectral Library [32], the LUCAS framework (Land Use/Cover Area Frame Survey;

http://eusoils.jrc.ec.europa.eu/projects/Lucas accessed on 12 January 2022) [33] with data from 23 countries in Europe [34], African Soil Information System (AFSIS) [35] and the GEOCRADLE with samples from 9 countries in the Balkans, Middle East, north and central Africa [36], the Open Soil spectral Library [37] promoted by the Soil Spectroscopy for a Global Good, based on the Rapid Carbon Assessment [38] spectral data from USA and Africa [32,35]. In both cases, the closed and open spectra data still lack operationalization to make them readily available to end-users, such as farmers and land managers.

Spectral data require scientific expertise to infer soil properties using complex processing algorithms not available to the general public. Moreover, there are variations in spectra aroused from different measurement protocols [39]. As an analogy, when satellite imagery was available for free for the first time, it was not widely adopted. Most users lacked computational competencies in pre-processing issues (e.g., atmospheric correction and georeferencing) and the complexity in supervised and unsupervised classification methods [40]. These shortcomings were removed when the images were made available to general users in pre-processed and georeferenced data format. Nowadays, we have a similar situation, where many SSLs and software processing are available [41] but did not make the bridge to making their use easier for the end-users, and thus they (farmers, consulters, others) cannot see the importance. As the first step on a learning curve, why not start delivering the spectral soil products directly to users? Such a win-win approach would boost even more research aimed at providing the best possible spectral-derived soil data and at the same time benefit end-users.

In this study, we present a free online platform called the Brazilian Soil Spectral Service (BraSpecS) for soil properties prediction using visible–Near–Short Wave Infrared (vis–NIR–SWIR) and Mid-Infrared (MIR) spectral ranges. This platform is a pioneering initiative, which aims to demonstrate its application for predicting many soil attributes, but here with a focus on soil organic carbon (SOC) and clay contents with spectral data from Brazil and the world. Furthermore, by establishing its direct application, we hope to foster a new generation of collaboration towards building a global online service for soil analysis.

2. Materials and Methods

2.1. The Brazilian Soil Spectral Service (BraSpecS) Construction

We developed an online service called The BraSpecS (Brazilian Soil Spectral Service) with support of the Geotechnologies in Soil Science Group (GEOCIS, https://esalqgeocis.wixsite.com/english accessed on 30 January 2022) Laboratory at the Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP). The web interface of the platform BraSpecS was created in JavaScript language. JavaScript is a lightweight, interpreted and object-based language, mainly used in building web interfaces [34]. BraSpecS is divided into three complementary modules: data localization, soil data visualization, and soil processing and quantification (Figure 1). The web platform can be accessed at http://besbbr.com.br/ (accessed on 30 January 2022).

In the data locations module, the user visualizes the number of samples by Brazilian states and identifies the authors and partner institutions. The map interaction of the Brazilian states was elaborated using the Leaflet library [42]. The soil data visualization module shows soil spectra in the vis–NIR–SWIR, and MIR bands filtered by classifications, orders, groups, layers, and textures.

All spectra and models are kept inside the system, maintaining data privacy and not publicly disclosed. Scripts for modeling are in the system's backend, so the user only needs to choose the desired properties to quantify. The system delivers the following soil properties which user can choose: soil color (Hue, Value and Chroma), clay, sand, silt, SOC, pH in water, exchangeable/available contents (Ca^{2+}, Mg^{2+}, K^{+}, Al^{3+}, H + Al, and P), sum of bases (SB = Ca^{2+} + Mg^{2+} + K^{+}), cation exchange capacity (CEC = SB + H+Al), base saturation (V% = SB/CEC × 100), aluminum saturation (m% = Al^{3+}/(SB + Al^{3+}) × 100), pseudo-total contents (Fe_2O_3, TiO_2, MnO, SiO_2, Al_2O_3), and Ki weathering index (Ki = SiO_2/Al_2O_3 × 1.7).

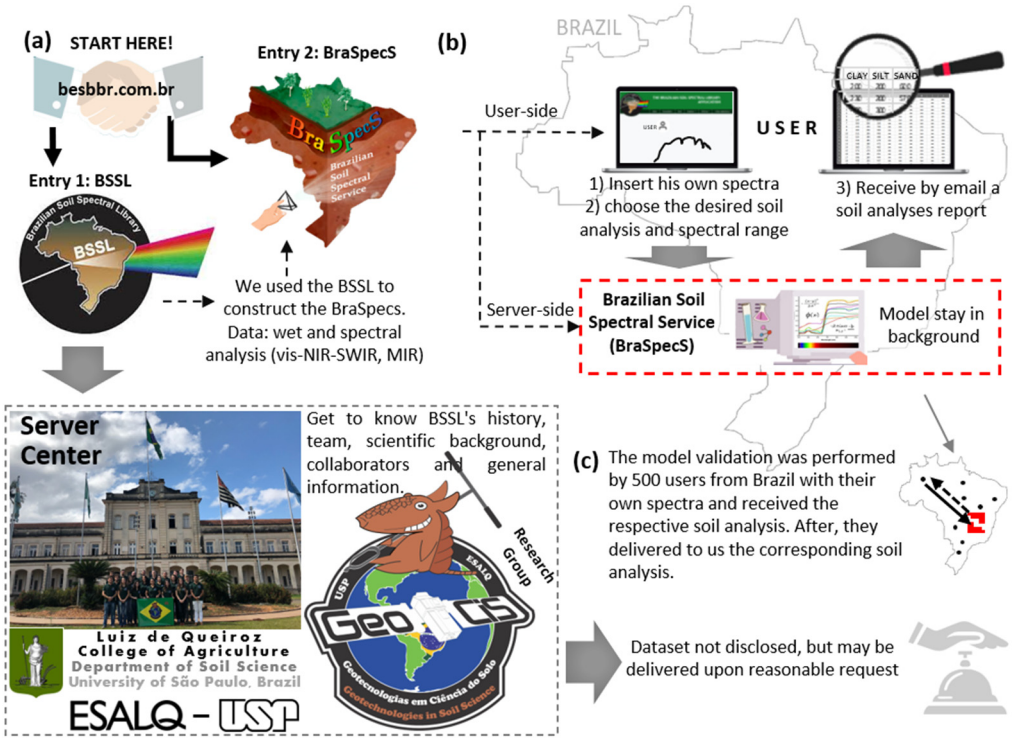


Figure 1. Flowchart of the directions inside the system: (**a**) Cover page, Entry directions, System functional architecture, Group, Server center, dataset, (BSSL, Demattê et al., 2018); (**b**) Dynamics of the system; (**c**) Evaluation–Validation procedure.

This system was prepared only for the BraSpecS as an experimental prototype. The predictive models were prepared in R software [43]. The steps of the soil processing and quantification module are described in the sections: soil dataset construction (Section 2.2) and data processing, models, and validation (Section 2.3). The baseline of the system functional architecture is illustrated in Figure 1.

For the web server, a workstation was acquired with 2 XEON 5120T processor hardware, with 14 cores each, and a video card with 4000 GPU's, which are essential for the application of the predictive models. The web server was created using the Apache software [44] and PHP programming language [45,46]. Apache is an open-source Hypertext Transfer Protocol (HTTP) server project to provide a secure, efficient, and extensible web server on HTTP standards. PHP is a fast and flexible scripting language, mainly used in web development [47]. The R scripts used in the soil processing and quantification module were integrated into the Apache web server through rApache software. The rApache allows the execution of scripts developed in the R programming language on Apache web servers [48].

Soil processing and quantification is a task that requires high computational resources [49], and to meet this requirement, we employed tools for organizing the spectral data submitted by users in queues and distributing them to other computers. The First-In-First-Out (FIFO) [50] model was selected and implemented on the server, allowing for a dynamic queue data structure that allows the removal and insertion of processing on the server. A high-performance processing cluster was created and thus made it possible to distribute the processes with low-cost computers in the R environment.

2.2. Internal Soil Dataset of BraSpecS

As mentioned, the BraSpecS is a service based on the soil dataset from the BSSL [12] where details of soil sampling and spectra can be achieved. In summary, soil samples are from different depths (cm)—A (0–20), B (40–60), C (80–100) and D (100–120)—which were acquired by auger or inside pits. Using these data, we constructed the platform with

vis– NIR–SWIR, resulting in 49,753 soil samples donated by 81 collaborators, representing 69 institutions from all over the country (https://bibliotecaespectral.wixsite.com/english/lista-de-cedentes, accessed on 30 January 2022). The BraSpecS in the MIR range comprises 4951 soil samples.

The BraSpecS contains laboratory analysis for several soil attributes in vis–NIR–SWIR and MIR regions. In this paper, we focused only on clay and SOC. The total content of SOC was determined according to a modification of the Walkley–Black method [51] where SOC is oxidized with potassium dichromate ($K_2Cr_2O_7$) in the presence of sulfuric acid (H_2SO_4), and the heat released in the acid dilution is used to catalyze the redox reaction. After digestion, the remaining unreduced $K_2Cr_2O_7$ is titrated with ferrous ammonium sulfate ($Fe(NH_4)_2(SO_4)_26H_2O$). The methodological procedure for this analysis was followed as described by [52]. For clay, in general the informed method was determined by measurements from [53]. For the vis–NIR–SWIR analysis, the soil samples were dried at 45 °C for 48 h, crushed, sieved with a 2 mm mesh, and homogeneously distributed in petri dishes prior to the measurement of the spectra in the 400–2500 nm range [12]. The spectral data were acquired using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO, USA). The sampling interval was 1 nm, reporting 2151 channels. The light source was provided by two external 50-W halogen lamps, which were positioned at a distance of 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. The sensor is calibrated using a white Spectralon plate (Lab-sphere, North Sutton, NH, USA) representing a 100% reflectance standard (reflectance factor 1.0). The reflectance of each sample was calculated as the radiance ratio between the soil sample and the Spectalon reference.

For spectral analysis in the MIR, the soil samples were ground and passed through 100 mesh. Reflectance spectra were obtained with the Alpha Sample Compartment RT-DLaTGS ZnSe (Bruker Optik GmbH, Ettlingen, Germany) equipped with an accessory for acquiring Diffuse Reflectance Infrared Fourier Transform (DRIFT). The sensor has a HeNe laser positioned inside the equipment and a calibration pattern for each wavelength. It has a KBr beam allowing a high amplitude of the incident radiance to penetrate the sample. Spectra were acquired between 4000 to 600 cm^{-1} (which is about 2500–16,667 nm) with a spectral resolution of 5 cm^{-1} and 32 scans per minute per spectrum. A gold reference plate was used as standard, and the sensor was calibrated every four measurements.

2.3. Data Modeling Provided by BraSpecS

Different pre-processing methods were evaluated for the vis–NIR–SWIR range and those that presented the best results for each soil property was selected. The Standard Normal Variable (SNV) and the Continuous Removal (CR) were used for clay and SOC, respectively, and all calculations were done using the 'prospectr' package in R [34]. In order to minimize the influence of noise at the tail ends of measured spectra, the ranges from 350 to 420 nm and from 2480 to 2500 nm were removed to have the spectra range from 420 to 2480 nm. Finally, we resampled the spectra at a resolution of 10 nm to reduce spectral multicollinearity and processing time and improve the modeling efficiency for this large dataset [54,55]. For the MIR spectral range, the Savitzky–Golay first Derivative (SGD) with a first-order polynomial and a window size of 9 nm and SNV were applied.

The datasets for each soil property were randomly split into a calibration (training; 70%) and a validation (testing; 30%) datasets, independently for each property. The complete BraSpecS dataset was used to calibrate spectroscopy models using the cubist machine learning algorithm [56]. Cubist is a rules-based algorithm that applies the M5 (Model Tree) approach to create categorical decision trees to deal with continuous classes. The algorithm produces 'trees' through rules that use boost training [56]. Reinforcement training is based on converting weak learners into strong learners, in addition to giving stronger learners more weight [57]. Cubist has been successfully applied to model clay and SOC from vis–NIR–SWIR spectra in numerous other studies (e.g., [10,21,58,59]). According to a comprehensive review by [60], Cubist stood out as a method, among other machine

learning methods, to predict SOC reliably from vis–NIR–SWIR spectra with R^2 between 0.76–0.89 and residual prediction deviation (RPD) between 1.99–2.88 in several studies.

The final model is regulated by a set of nodes along the tree and two hyperparameters (committees and neighbors), which improve the model's performance. The model construction and estimation process were performed by the *caret* package in R [61], which has a set of functions that seek to simplify the process of creating predictive models. The first criteria used to select the optimal models was the Coefficient of Determination (R^2). However, the Root Mean Square Error (RMSE) and Ratio of Performance to Interquartile Distance (RPIQ) also were used for interpretation of results.

The online system was tested by users using their own spectra with a total of 500 Brazilian participants recruited along a spectroscopy course (https://esalqgeocis.wixsite.com/english/probase, accessed on 30 January 2022) composed by laboratory technicians, researchers, students, farmers, consultants, and distributed along 20 states of the federation (two spectra per user, total of 1000 spectra) for the vis–NIR–SWIR and 200 samples for MIR using equipment and protocol equal to that used for the construction of BraSpecS. The spectral soil predictions for clay and SOC were made on-the-fly immediately after submission of the spectra. These blind set soil predictions were compared post-hoc after retrieval of participants' soil analytical lab data to evaluate deviations. Users were also invited to provide critiques of the system for further improvement.

2.4. Data Modeling Provided by BraSpecS

We compared clay and SOC models derived using the world, national (BraSpecS), and local vis–NIR–SWIR and MIR datasets. The following approaches were used: (a) We entered the world spectral data to predict clay and SOC using the BraSpecS soil models. The global data entailed 28,598 soil samples with vis–NIR–SWIR scans from 65 countries and 8039 samples from 4 countries with MIR scans (390 from Australia, 170 from Iran, 2728 from the USA, and 4751 from Brazil); (b) we created for each of the 65 countries Local Models (ExCSSL) with their spectral population and predicted clay and SOC locally; (c) finally, we merged the BraSpecS with the spectra from the other 65 countries and generated a GSSL.

The processing was the same as the previously described for the BraSpecS, that is, random data split with 70% for model calibration (training) and 30% for validation (testing) and modeling using the machine learning algorithms. Finally, we compared the results from the BraSpecS tested by other countries and compared them with the developed ExCSSL, BraSpecS and GSSL models were compared with the same 65 countries. This made it possible to evaluate the differences between global, national, and local datasets on the quantification of soil properties. The workflow process is illustrated in Figure 2. The number of samples per country is provided in Supplementary Materials.

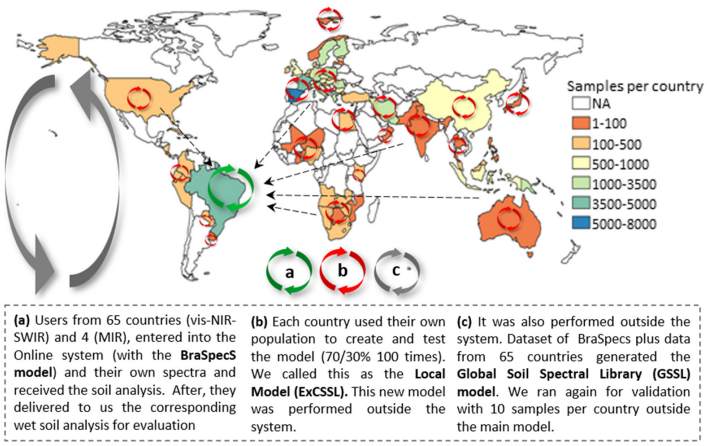


Figure 2. World participants. Exploiting different populations and models to quantify soil properties, (**a**) spectra from 65 countries were tested into the BraSpecS model, (**b**) the Local model, (**c**) the Global Soil Spectral Model.

3. Results

3.1. Online Interaction Experience

The website is available on "besbbr.com.br" (accessed on 12 January 2022) and brings together spectral information in vis–NIR–SWIR and MIR ranges (Figure 3). This is a user-friendly interface intended to provide a favorable experience for users. The web is designed for: (a) end-users, who want the soil analysis; (b) researchers and academic employees, who want to test and evaluate their models; (c) students who are interested to learn; (d) pedologists and soil scientists to test and have new insights and view the soil spectral signatures patterns; (e) startups to create their own market.

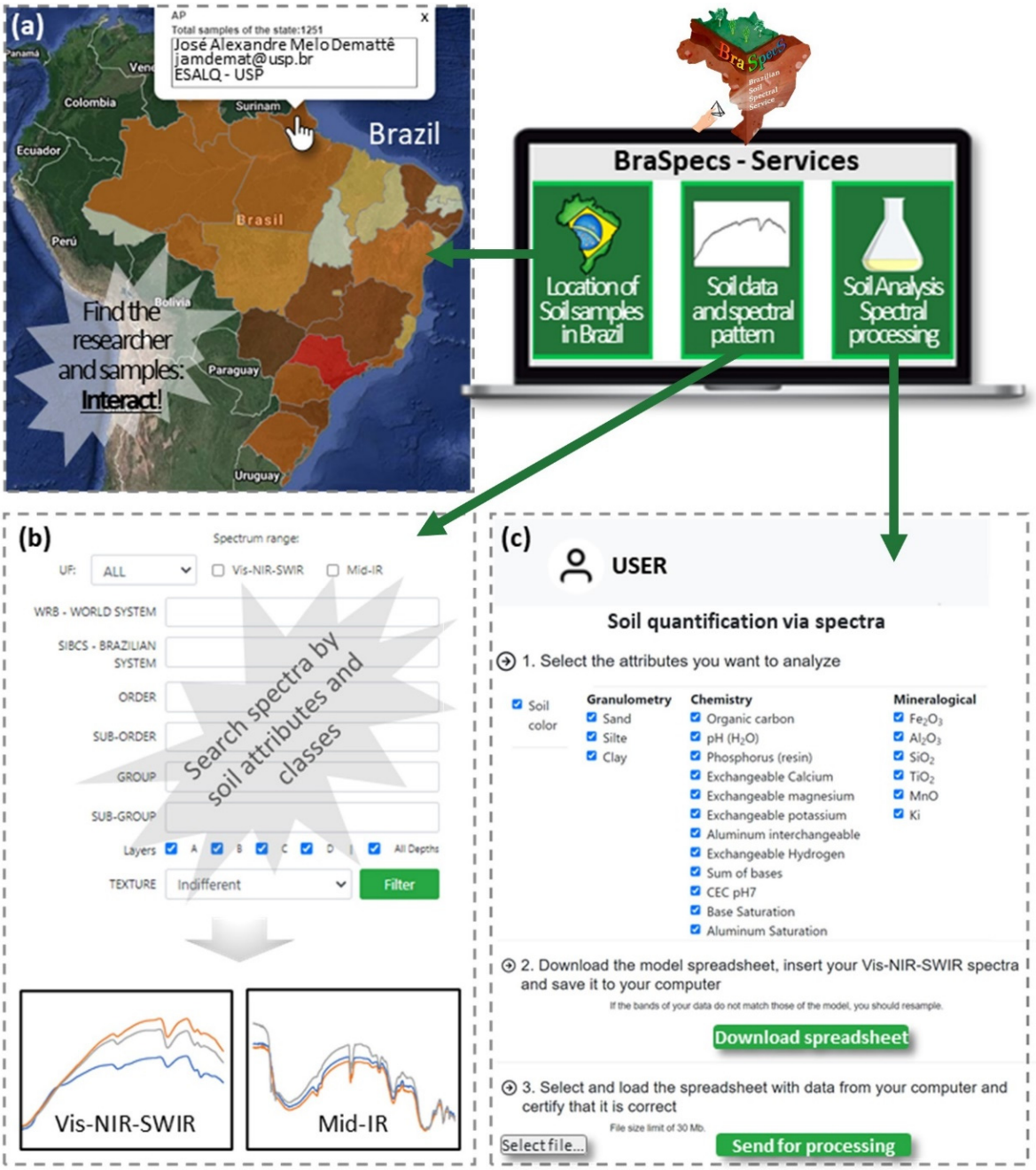


Figure 3. BraSpecS Services description: (**a**) find researchers that have data; (**b**) observe patterns of spectra per soil attribute and per soil classification, filtering per state, per researcher, per soil type and others; (**c**) Insert your spectra, choose the desired soil analysis (Granulometric, chemistry and mineralogical) and receive the report by email with statistics.

The website presents the following sequence (Figure 3). First, a user registers in the system. Afterward, the user can view the general information on how the BraSpecS was developed or go directly to BraSpecS-related services.

The tool offers three services (Figure 3), which are as follows. First, (1) an alignment tool, where the user can search for the owners for spectral data and personal contact information. With this information, users can contact the data provider directly and ask for specific datasets to initiate a collaboration. Furthermore, the user will see where to find potential users on spectroscopy. The idea was to stimulate users to interact and create a new collaboration. The interactive map of contributors allows one to search for specific institutions or researchers and visualize Brazilian State spectral data. The map also allows the interaction between researchers and that leads to spectral data sharing and partnerships. Second, (2) users find several examples of soil types of patterns. One may ask for a specific soil class, for example, a Ferralsol, and the system will filter and show the average of all Ferralsols in the dataset or from a specific state. Users can run searches by specific criteria, such as soil depth, soil type, and specific soil property. For example, the user can ask to retrieve vis–NIR–SWIR samples of sandy soils at surface depth and in a specific state, owner, region, or the whole library. The result of the search is the average of soil spectra regarding the chosen characteristic. Depending on the number of soil spectra falling under specific criteria, the process may take some time. As an example, we selected the São Paulo (SP) state, the vis–NIR–SWIR spectra, the sandy textural class from the first layer (A), with no indications for soil classification. This query took about 3 min. Figure 3 gives detail on spectral patterns that users have access. Users can see the spectra regarding clay content or soil classification and compare them with their spectra.

Finally, (3) we have the Soil Analysis Spectral Service. Here the user has spectral data and wants to make a soil prediction anywhere in the world. To access the prediction module, the user must register on the platform with an email address to receive the results. Afterward, the user must log into his/her account in the platform to: (1) download the template to organize the soil spectra; (2) upload the file (.csv format) with the soil spectra; (3) select the attributes to be predicted; and (4) send the data for processing in the platform. Thus, the user uploads the spectra, and chooses among vis–NIR–SWIR or MIR spectral range and the desired soil properties, and then runs the processing. The system runs all scripts in the background, not displayed on the web but presented in this paper. After about 15 minutes, depending on the filtering and number of samples chosen, the user will receive a report by email. The report entails all soil analyses of the specific spectra, method (cubist), and statistical performance metrics of the backend. The user also has the option to provide feedback online and share the wet soil analytics of user spectra, so the system will be recharged with new data and will increase the dataset.

3.2. The Quantification

The descriptive metrics for spectroscopy clay content and SOC estimated concentrations using different population models, that is the BraSpecS, GSSL, and the ExCSSL are shown in Figure 4. For clay, the predicted content standard of all models w very similar with the observed distribution, with 90% of the population ranging mainly from 150 to 400 $g \cdot kg^{-1}$. For SOC, the predicted value distributions obtained from the GSSL and the ExCSSL were in accordance with the observed values, with 90% of the population ranging mainly from 10 to 180 $g \cdot kg^{-1}$. However, the BraSpecS model underestimated SOC values compared to SOC observations.

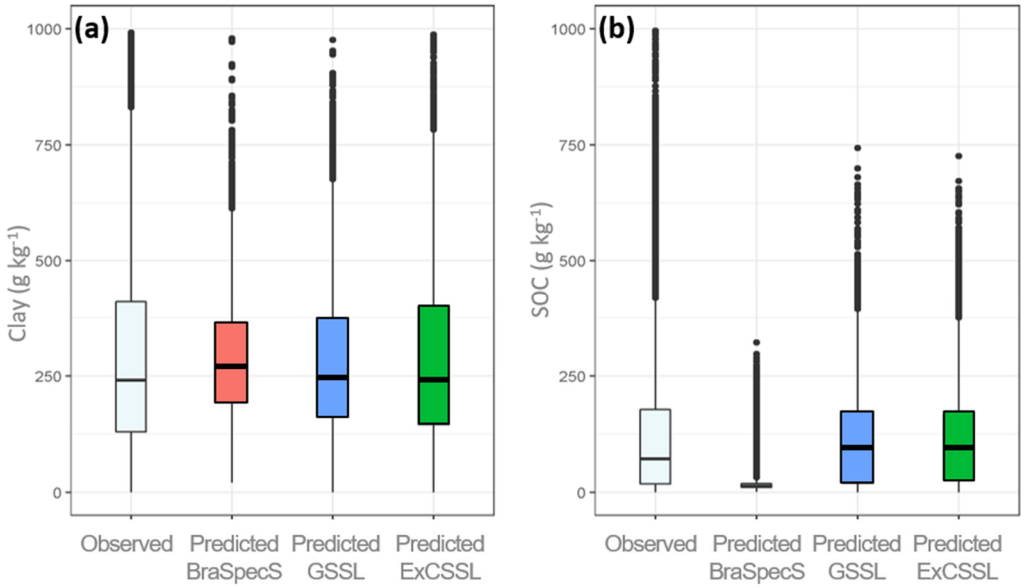


Figure 4. Boxplots of observed and predicted (**a**) clay and (**b**) SOC using different spectral libraries' vis–NIR–SWIR range.

3.3. Prediction Models Based on Different Populations

Figures 5 and 6 present the R^2 values of the vis–NIR–SWIR models results for clay and SOC using the different models (ExCSSL, BraSpecS and GSSL) while the RMSEs and RPIQ of clay and SOC models are shown in Figures 7–10, respectively. ExCSSL presented better results followed by the global and the BraSpecS models. The BraSpecS model was very good for Brazilian samples, but not for other countries.

The BraSpecS model showed a very different value from the ExCSSL in other countries for both clay and SOC in the MIR range (Figures 11 and 12). Prediction values derived from MIR closely matched the observed properties' distributions, except for clay predictions using BraSpecS. Interestingly, the R^2 for clay validation models mirrored the results of vis–NIR–SWIR. However, for SOC, the R^2 derived from MIR spectra were substantially better than those derived from vis–NIR–SWIR using the BraSpecS.

Spectra from a large country such as Brazil has advantages due to its high variability in soils and biomes. Table 1 shows that, using the BraSpecS model, 24 countries achieved an R^2 of greater than 0.5 for clay with vis–NIR–SWIR data. When the BraSpecS model was applied in 16 different African countries, 7 had R^2 score over 0.5, which means that BraSpecS was feasible. In fact, many countries from Africa have similar soils as Brazil. In contrast, the BraSpecS models did not perform well on spectra from Asia. Using ExCSSL, 11 models in Asia achieved $R^2 \geq 0.7$. The GSSL models outperformed the BraSpecS models, but did not perform as well as the ExCSSL. In total, 54 countries achieved an R^2 higher than 0.5 for clay estimation using GSSL compared to only 24 countries using the BraSpecS.

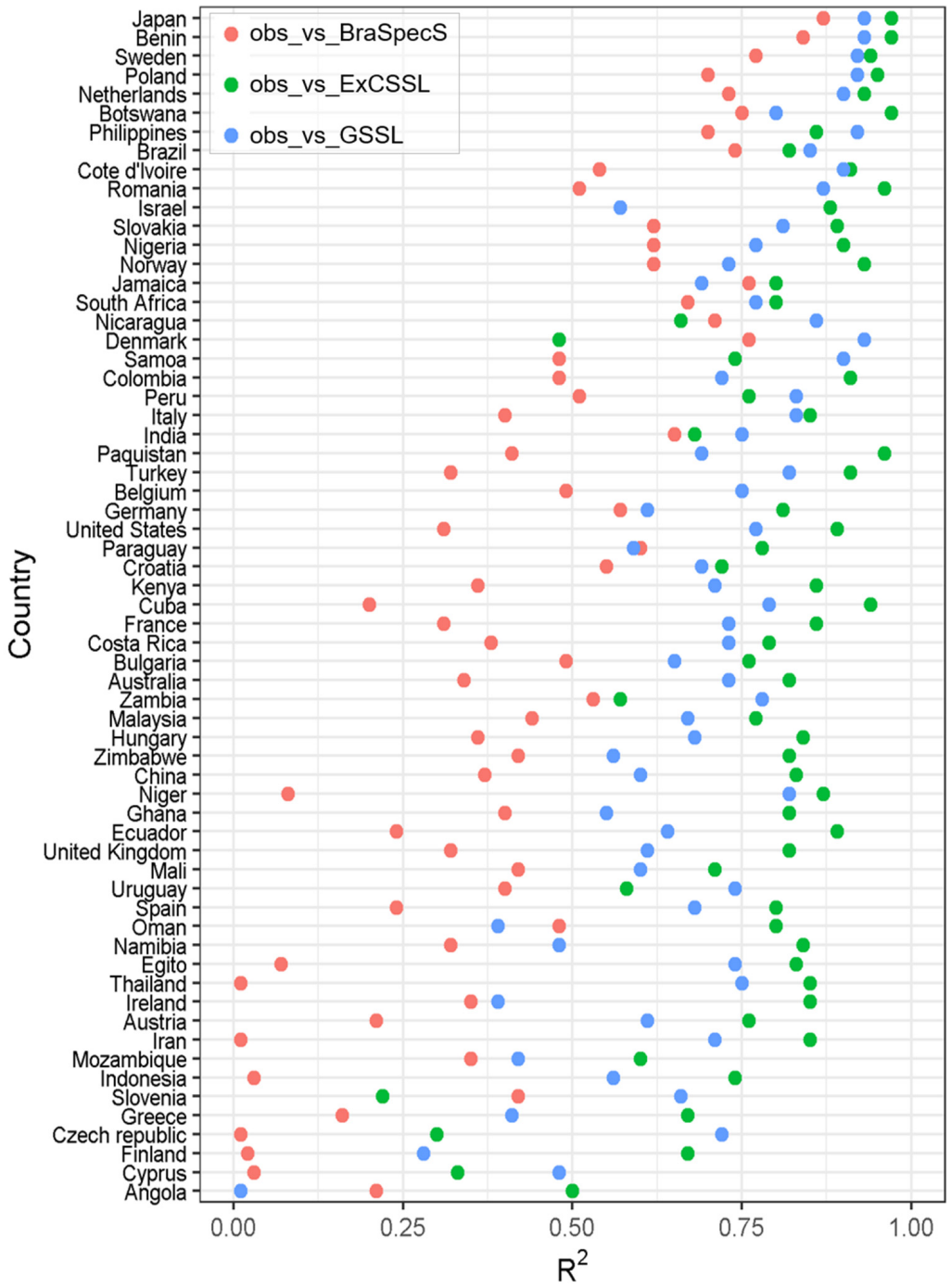


Figure 5. R^2 for clay models per country using vis–NIR—range.

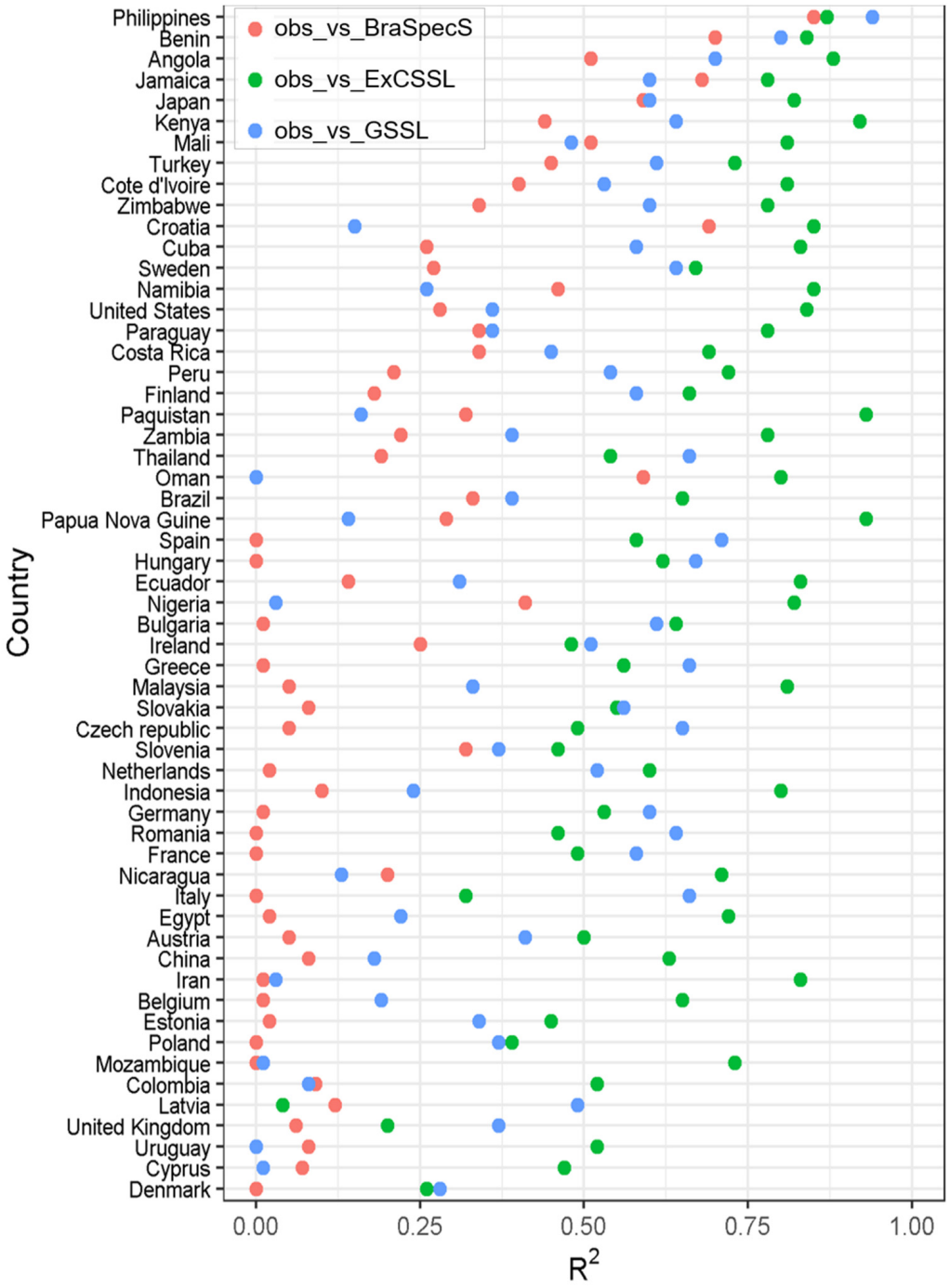


Figure 6. R^2 for SOC models per country using vis–NIR–SWIR range.

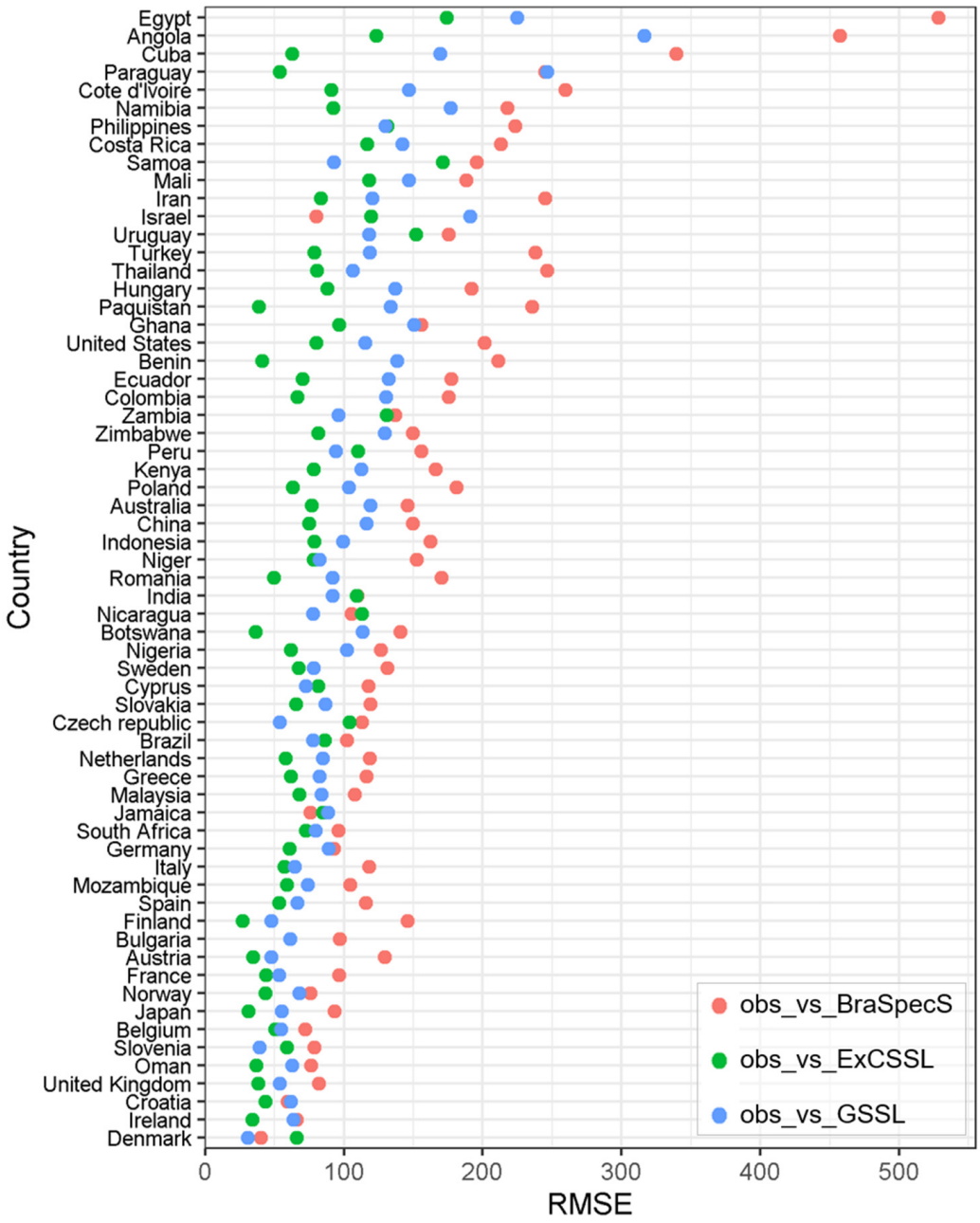


Figure 7. RMSE for clay models per country using vis–NIR–SWIR range.

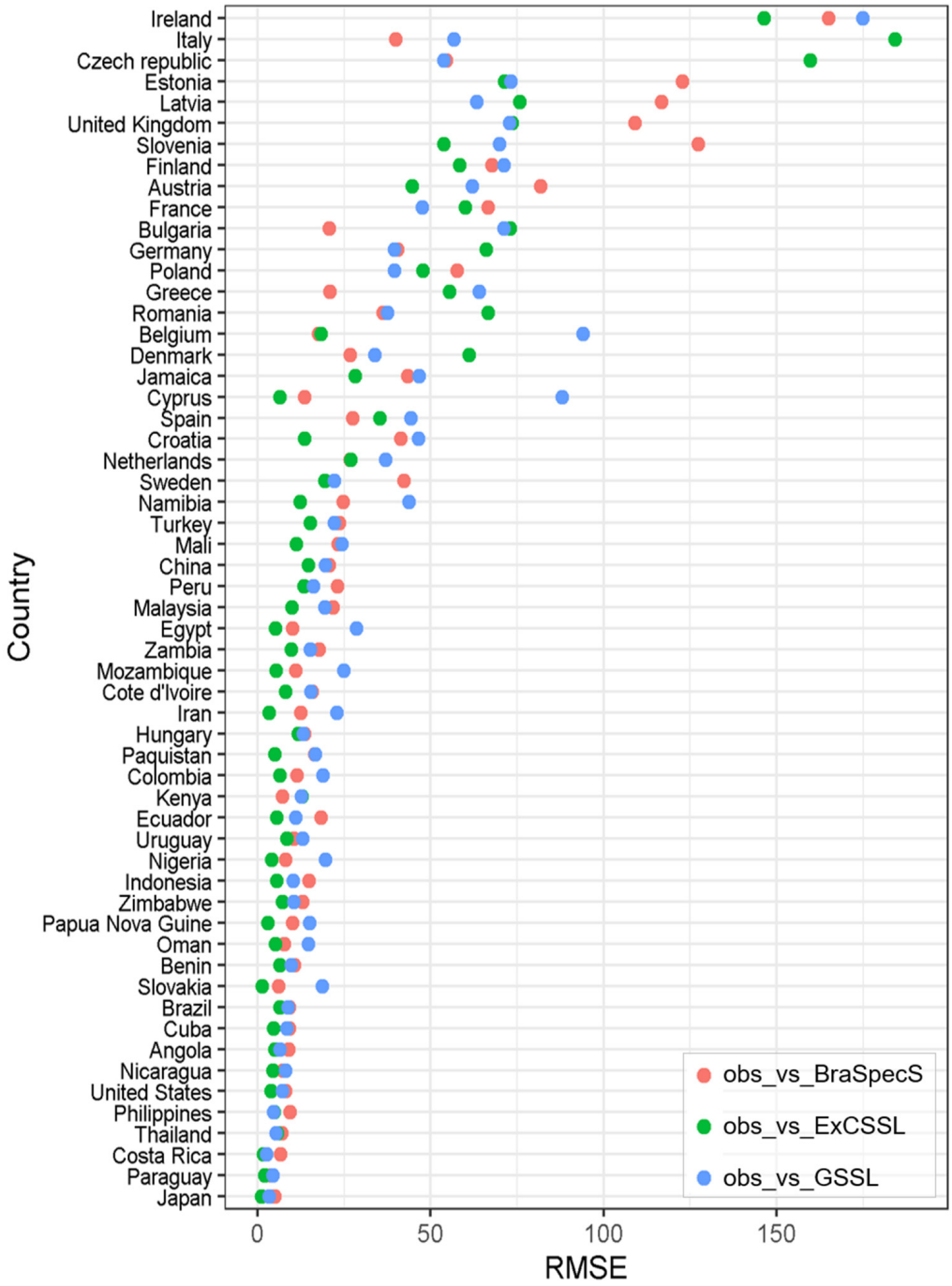


Figure 8. RMSE for SOC models per country using vis–NIR–SWIR range.

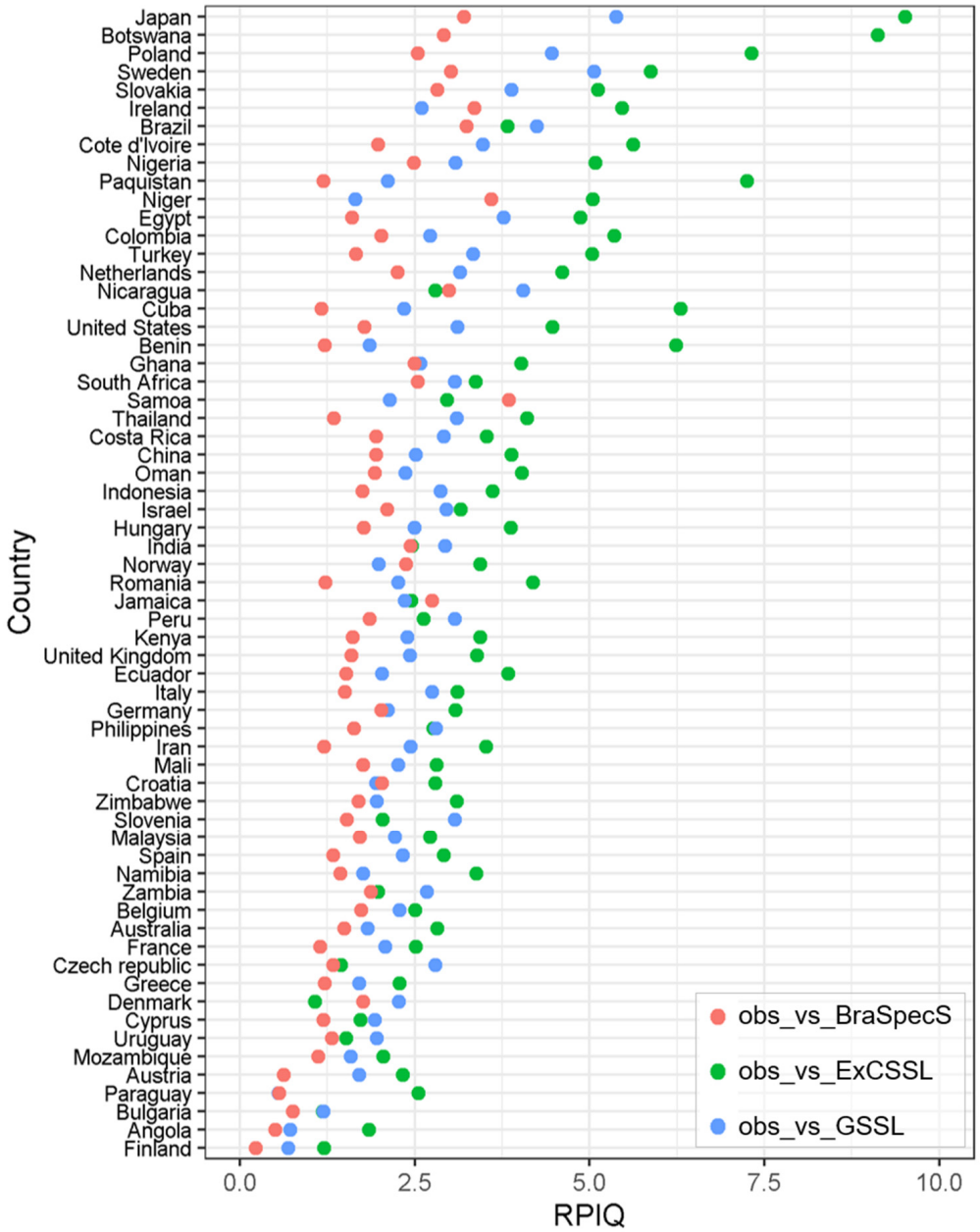


Figure 9. RPIQ for clay models per country using vis–NIR–SWIR data.

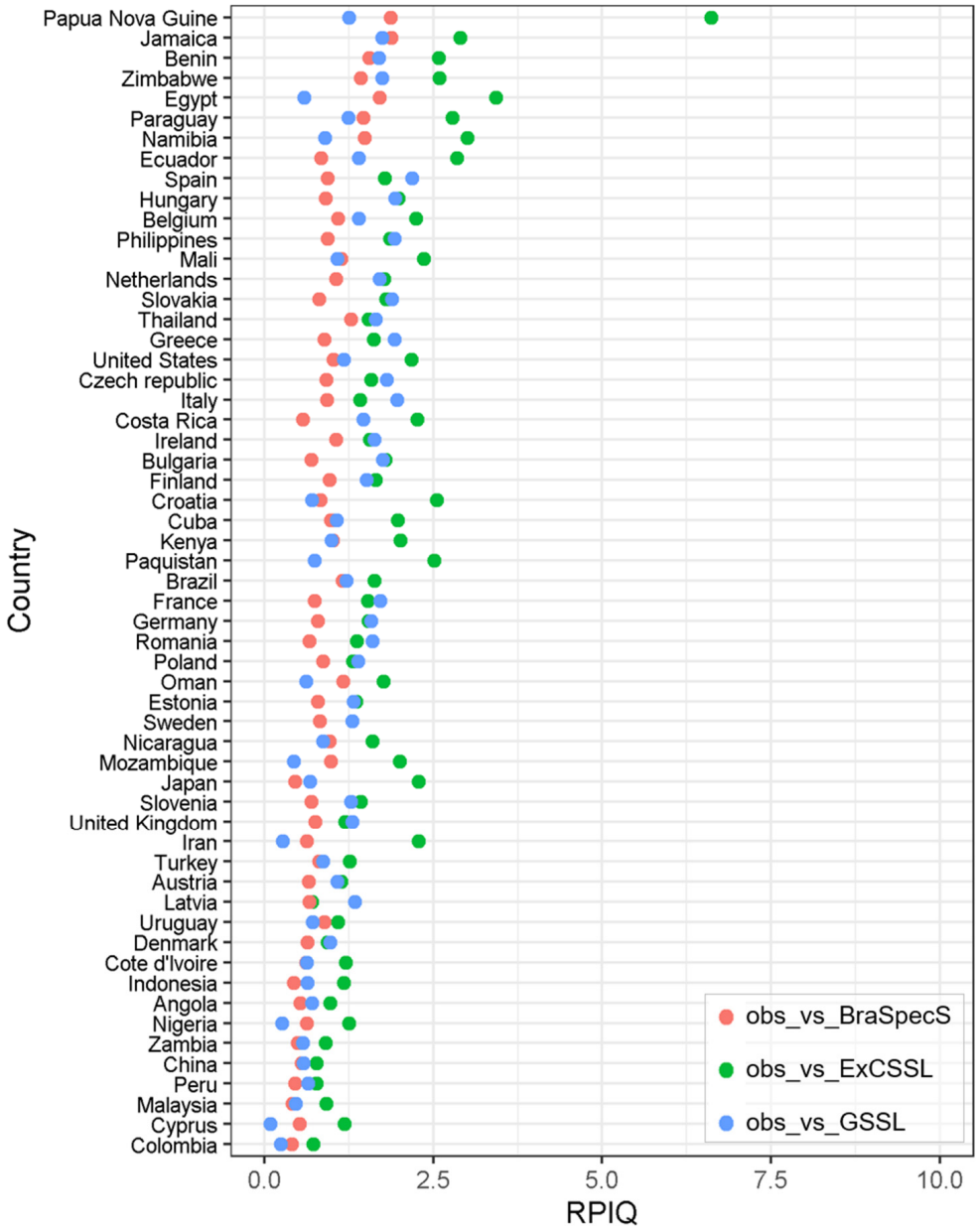


Figure 10. RPIQ for SOC models per country using vis–NIR–SWIR data.

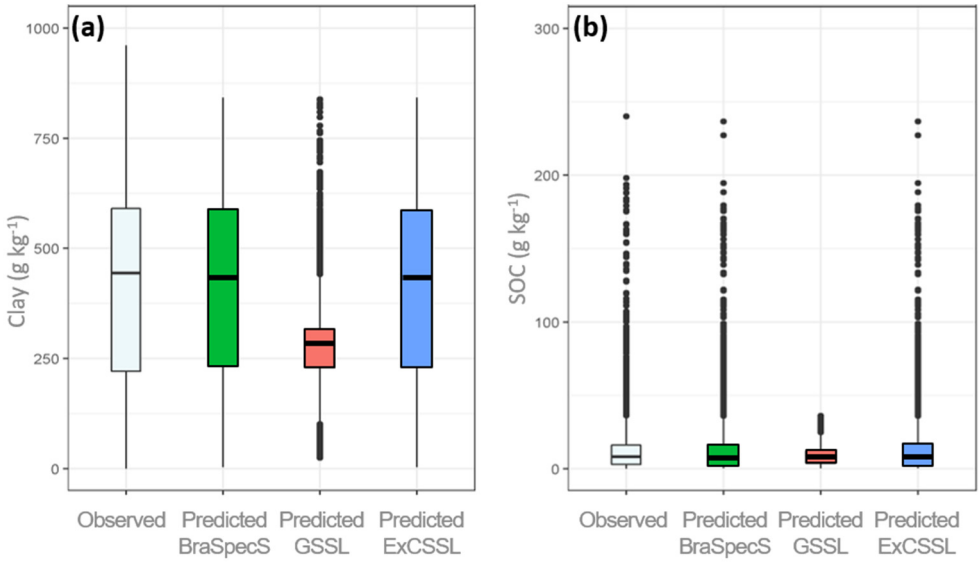


Figure 11. Boxplots of (**a**) clay and (**b**) SOC using MIR dataset per country.

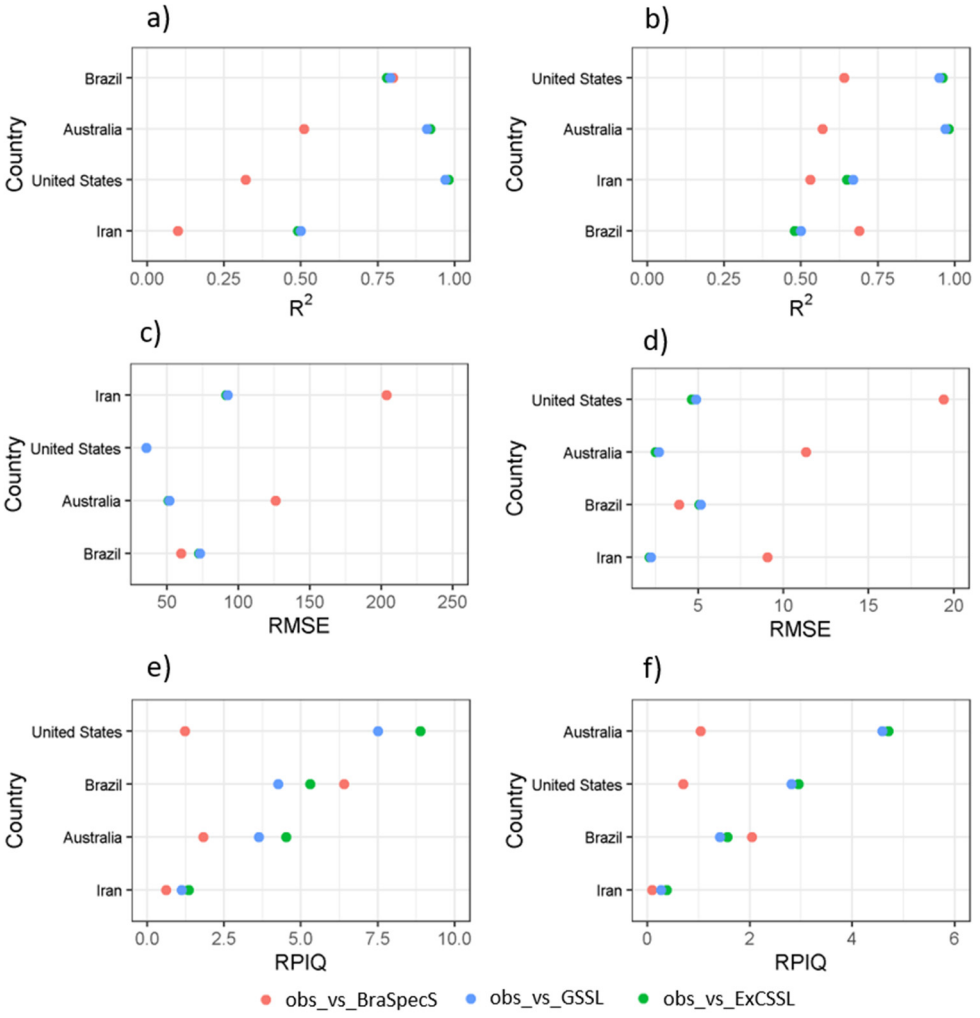


Figure 12. R^2 for (**a**) clay and (**b**) soil organic carbon (SOC) models, RMSE (**c**) clay and (**d**) soil organic carbon (SOC) models and RPIQ (**e**) clay and (**f**) soil organic carbon (SOC) models per country using MIR data.

Table 1. Number of countries with R^2 for clay model using vis–NIR–SWIR data.

Model	R^2	Total	Africa	Asia	Europe	North America	Oceania	South America
ExCSSL Clay	0–0.3	1	0	0	1	0	0	0
	0.3–0.5	3	0	0	3	0	0	0
	0.5–0.7	8	3	1	2	1	0	1
	0.7–0.8	11	1	2	4	1	1	2
	>0.8	40	11	9	13	3	1	3
BraSpecS Clay	0–0.3	14	3	3	6	1	0	1
	0.3–0.5	25	6	5	8	2	2	2
	0.5–0.7	12	4	1	5	0	0	2
	0.7–0.8	9	1	1	4	2	0	1
	>0.8	3	2	1	0	0	0	0
GSSL Clay	0–0.3	3	1	1	1	0	0	0
	0.3–0.5	6	2	1	3	0	0	0
	0.5–0.7	18	3	4	8	1	0	2
	0.7–0.8	18	5	3	4	3	1	2
	>0.8	18	4	3	7	1	1	2

In summary, the best clay model performance was obtained for ExCSSL > GSSL > BraSpecS irrespective of different continents with ExCSSL clay models showing R^2 larger than 0.5 in 59 countries. Interestingly, even in South America, BraSpecS clay models were outperformed by GSSL and ExCSSL models. What also stands out is the high performance of ExCSSL clay models in Europe, with 13 countries out of 23 achieving a $R^2 > 0.8$. Another interesting point is that 4 countries from Europe had the same R^2 (0.7–0.8) with all models (BraSpecS, ExCSSL and GSSL).

For SOC, the results for R^2 in validation mode using vis–NIR–SWIR data were less promising, when using the BraSpecS (Table 2) but maintained the trend of better for the GSSL and the best for the ExCSSLs. The BraSpecS and somewhat GSSL SOC models performed especially poorly in European countries. Indeed, these ones have very different mineralogy and carbon contents.

Table 2. Number of countries with R^2 for Soil Organic Carbon (SOC) models using vis–NIR–SWIR data.

Model	R^2	Total	Africa	Asia	Europe	North America	Oceania	South America
ExCSSL SOC	0–0.3	3	0	0	3	0	0	0
	0.3–0.5	9	0	0	9	0	0	0
	0.5–0.7	17	0	2	11	1	0	3
	0.7–0.8	9	4	1	0	4	0	0
	>0.8	19	7	7	1	2	1	1
BraSpecS SOC	0–0.3	38	3	5	22	3	1	4
	0.3–0.5	11	5	2	1	1	0	2
	0.5–0.7	6	2	2	1	1	0	0
	0.7–0.8	1	0	0	1	0	0	0
	>0.8	1	0	1	0	0	0	0
GSSL SOC	0–0.3	17	4	5	4	1	1	2
	0.3–0.5	14	2	1	6	2	0	3
	0.5–0.7	22	3	3	13	2	0	1
	0.7–0.8	2	1	0	1	0	0	0
	>0.8	2	1	1	0	0	0	0

4. Discussion

4.1. The Web Service Advantages and Limitations

The BraSpecS online platform presented here overcomes some practical limitations and makes soil spectra a service accessible to anyone, democratizing its usage. Although a highly complex, state-of-the-art machine learning method (Cubist) was used in this study, our platform frees the end-user from having to learn spectral modeling. *A user only needs to upload a spectrum and will receive the soil analysis. This may start to bring years of infrared research directly to the public.*

The spectral platform approach enhances equity of making spectral soil models and knowledge readily available to the global community at no cost. The data-driven knowledge engrained in soil models, such as ExCSSL, BraSpecS, and GSSL, developed from vis–NIR–SWIR and MIR spectral data is shared with users who become participants to co-create larger and larger global soil spectral libraries that serve the greater good.

The web services on the spectral platform are user-friendly, fast, and facilitate the formation of an active and engaged community of experts, soil scientists, students, farmers, consultants, and other stakeholders. As a living technology platform, suggestions from the user community can be readily integrated. People who belong to a global soil spectral community can also benefit by retrieving soil analytics from their uploaded spectral data.

While other global and continental soil spectral models were driven by researchers and professional soil databases other ongoing global spectral community efforts (e.g., SoilSpec4GG) are more vertical with researchers subsuming people's spectral data without a data sharing policy that fully acknowledges and credits the user's labor and costs of field data collection. Works such as from [10], the European LUCAS dataset modeled by [62,63], as well as the U.S. soil spectral data modeled by [23,64,65], were important to show the community the importance and potential of the technique. A free repository such as by [66] have great importance to make this grow since users have access to data. Therefore, our initiative demonstrates the importance of providing online results to end-users and this may encourage other working groups to improve similar new systems.

One of the reasons the accuracy varies between SSLs is the difference in measurement protocol used by the SSL owners. This issue needs to be resolved in the near future with an initiative to establish an agreed standard and protocols amongst the global users. This effort is being carried out by the IEEE SA P4005 working group.

In addition to the service, developing a soil-spectral data web platform that is anticipated to grow even further in the future with the submission of new spectra provides a virtual space to build community. Due to data ownership, the system used a non-disclosed dataset idea. If one is interested in the data, the system indicates the data owner, and encourages the user to contact the data custodian, increasing community knowledge.

The pedometric and soil modeler community have become quite specialized in AI, scripting, modeling, and high-end data processing, which has somewhat disconnected them from work with field pedologists and farmers cropping the fields. Thus, our soil-spectral web platform helps bridge the gap between modelers and users of soil data. Our tool offers people to collaborate, form partnerships, get to know others who are interested in soil spectroscopy, and better understand soils in all regions of the globe. Reconnecting soil modelers and soil spectral data collectors offers new avenues to build community. In essence, new connections can be made between "the machine" that models soils and people with interest in soils.

To understand the usefulness of shared SSL, consider the following example: Stakeholders (farmers or researchers) could send their soil samples to a central SSL (e.g., national, or global SSL), where they would be scanned and the spectral data stored, or they could send already acquired soil spectra. Local SSL can be explored for personal interests or to meet other needs (e.g., soil monitoring) and feed the global SSL, growing a global repository. Once a global SSL is trained and evaluated, spectra from a profile of an unknown type can be compared with spectra in the global SSL and a preliminary soil classification or specific soil properties, such as SOC or clay content, can be estimated. Global, regional, and

local SSL can co-exist because they serve different needs and purposes. While local SLs are customized to specific soil regions, they tend to perform better to predict SOC and clay than regional (BraSpecS) and global (GSSL) models, as demonstrated in our study. However, the outlook for SSL is that as more soil-spectral data pairs are added, global prediction performances using more advanced analysis are expected to become better at modelling local soil variations. Growing a global SSL will eventually converge to a saturation level at which regional soil variabilities are well represented; thus, it is expected that global SSL using AI technology will provide as robust and accurate soil prediction models as local SSL in the future.

One requirement for a robust model is that the dataset must be standardized with the same spectral bands and the same soil analysis method. This is because the interaction between wet analytical data and spectra can be different when spectral models are trained using unharmonized spectra and SOC data. This could increase the predictive uncertainty and reduce the interpretability of the model. Data quality is crucial for superior results, although it is difficult to achieve with legacy datasets. Thus, it is imperative to start using agreed-upon standards and protocols on careful and agreed spectral soil measurements with the quality of soil wet chemistry analysis, which are the basis for success in delivering accurate model estimates for soil properties.

Another limitation is that end-users will need to acquire or gain access to spectroradiometers to collect soils' vis–NIR–SWIR and/or MIR data. These limitations are viewed as temporary since soil sensor technology has become more widespread with the advent of precision agriculture and "smart" agricultural management. In addition, regional cooperatives or centers may serve farmers and end-users in more resource-limited settings.

4.2. Brazilian Users of the BraSpecS

For the Brazilian vis–NIR–SWIR dataset, clay content presented good results using the platform, with R^2 = 0.75. In contrast, SOC presented lower values (R^2 of 0.45). These results indicate that the SOC is more dependent on so many factors such as biomes, land use, mineralogy [12,67], and has great dynamics due to climate and microorganisms that mean its quantification can become a challenge. This agrees with past studies, e.g., [6]. In this scenario, the use of SSL and local models can be a strategy for the online service to return more accurate estimates to end-users. Moreover, SOC spectral estimation has been a challenge in Brazilian agricultural areas because of the low soil carbon content. The SOC results using MIR were significantly better than vis–NIR–SWIR, since they reached R^2 of 0.8 and 0.7 for clay and SOC, respectively, in agreement with past studies [68–70]. Thus, the BraSpecS online platform can be used as an important service for soil analysis over Brazil, considering the level of accuracy and the clay and SOC property.

4.3. International Users of the BraSpecS Based on the Internal BraSpecS

International users from several countries submitted spectra via the online platform to identify whether their local samples could be predicted by the BraSpecS service. We observed that for clay, three countries from Africa, two from Asia, four from Europe and two from North America showed R^2 values of over 0.7. Despite that, in Europe there were still five countries with models in the R^2 range of 0.5–0.7 for clay. The results were less satisfactory for SOC, which agrees that this property is more dependent on other factors such as biomes, land use, and others [10]. In the case of clay, results indicated that the BraSpecS model presented good results for some countries. For example, spectra from Thailand, Benin, Denmark, Jamaica, Japan, The Netherlands, Nicaragua, Poland, Philippines, South Africa, and Sweden reached R^2 over 0.7. This indicates that for clay, a model built using spectra from a large, diverse country can quantify spectra from other countries. On the other hand, many countries reached low values. This gives two indications: (1) a country model can assist other countries but not all of them; (2) the user will have the opportunity to choose the SSL depending on soil similarity. For example, if the user lives in a country that does not have an SSL, the user can choose a global one or another region with similar

soil. These limitations can be added to the online spectrum service platform, enabling the user to make the decision to use local or global models based on the accuracy of the estimates required for each application of the results.

4.4. International Users of the BraSpecS Based on Local Datasets

SSLs may adopt several approaches and levels: a farm [71], a region [72] a country [12,19], a continent [33], or the world [10]. The present paper presented different approaches to understand soil population specific modeling (ExCSSL, BraSpecS, and GSSL).

We observed that ExCSSL were clearly better at quantifying clay and SOC in almost all cases and continents. The user-specific ExCSSL preserved the main characteristics of their regional soils, parent materials, biomes, and other information which spectra carry. This finding agrees with [73], for whom the transfer of vis–NIR–SWIR models from global to local scale, the latter were the best. In our study, better model performances for both clay and SOC were observed in local models when compared to the GSSL, irrespective of different continents with diverse soils. We observed only a few cases in Europe where model performances were quite low for both SOC and clay ($R^2 < -0.3$).

Our results clearly demonstrate that the GSSL model performed better than the BraSpecS model for both SOC and clay content, while the local country-specific models outperformed both BraSpecS and GSSL models.

The caveat is that local datasets had different sample sizes (unbalanced sampling design) which may have influenced model performances. The issue of unbalanced datasets in testing the transfer of soil spectral models was addressed by [74] who used a standardized balanced sampling design; however, in their study the transferability and scalability of spectral models (local to regional scale, and vice versa) for soil carbon were inconclusive. The study found that the transferability and up- and downscaling of the soil spectral models were limited by the following factors: (a) the spectral data domain; (b) soil attribute domain; (c) methods (e.g., machine learning or deep learning AI methods) that describe the relations between vis-NIR-SWIR and soil carbon; and (d) environmental domain space of attributes that control soil carbon dynamics.

Other soil spectral studies, such as [75,76], showed that spiking libraries improved the performance of soil prediction models. These spiking studies suggest that building larger spectral libraries (continental and world libraries) achieves better results to model soil properties than regional and local libraries. However, [76,77] pointed out that local model calibrations customized and optimally fitted to field/farm/local soilscapes are the best to model soil properties, even with small datasets with as few as 25 samples. Whether local or global soil spectral models perform better may be more a question of homogenization of data to reduce the variability in soil, spectral, and/or associated soil-environmental characteristics. A study [60] found for soils in southern Brazil that the stratification of a large spectral library into more homogeneous sample groups by environmental criteria (physiographic regions and land-use/land-cover) improved the accuracy of SOC predictions compared to pedological (soil texture) and vis–NIR–SWIR spectral (spectral classes) criteria. Subsetting can be considered as an approach to localize and homogenize soil spectral sample populations, but it is not always successful and depends on the soil and environmental conditions [78]. In another study [79], they found that stratification by mineralogical uniform clusters improved predictive performance of clay content, irrespective of the geographic region, using a large tropical soil spectral set.

There are several factors that play a role in building world soil spectral libraries that have contradictory effects on modeling of soil properties. First, adding soil spectral data may introduce noise to the global library, specifically if the data quality is of poor quality due to incorrect measurement, or different protocols. Second, adding redundant soil spectral data that are already present in the world library is unlikely to boost model performance of soil properties. Indeed, studies demonstrated that more soil data in a spectral library does not necessarily mean better soil predictions. According to [80], there was relatively little significant increase in prediction capacity of soil attributes with the

use of an entire data set compared to a smaller subset, which increased the R^2 from 0.63 to 0.72 for SOC and R^2 from 0.71 to 0.73 for clay, respectively. On the other hand, representing the actual soil variability exhaustively of every region of the globe—including cold regions, mountainous or wetland regions that are difficult access, or politically restricted regions—would ensure that data gaps are filled, and all soil types are included in a global dataset. Such efforts are underway, though have been hampered by investment in new soil sample campaigns and standardized analytics and spectral protocols to ensure high data quality.

The issue of legacy datasets that have dominated national and global soil libraries and the lack of a consistent global soil monitoring network are noteworthy; [81,82] stressed that although striking global soil maps have been generated, future soil mapping and modeling efforts will depend on data mining all existing soil data and an increase in soil monitoring efforts. A suggestion to use an internal soil sample (ISS) that is disseminated between all laboratories and align all measurements to this ISS was proposed and validated by [39]. Various users successfully adopted this method using different spectrometers [83] and measurement conditions [84]. This direction may minimize the variation between many new SSLs and help better use our system in the global domain as demonstrated by [85].

The GSSL of clay compared to the ExCSSL, agrees with [86], who presented a methodological framework for using vis–NIR–SWIR spectroscopy at local and global scales by spectral treatment and regression methods. In our study, MIR-based predictions of clay showed the highest R^2 for ExCSSL, followed by GSSL and last BraSpecS. MIR data (and pooled MIR + vis-NIR-SWIR) compared to solely using vis-NIR-SWIR data have shown superior results to predict SOC and/or clay in numerous studies (e.g., [64,78,87,88]). Although MIR spectra have fingerprinting capabilities to trace fundamental spectral elemental bonds, vis-NIR-SWIR is limited to identifying overtones in spectra. However, the former is much more costly and laborious to use. This explains the rapid growth of national and continental vis-NIR-SWIR libraries, while large MIR libraries that cover the variability of soils around the globe are still limited.

Figure 13 shows an example of application. The Israel dataset was inserted in BraSpecS and reached an R^2 of 0.88. When using the local model, the performance was still at 0.88 with a lower error. For SOC in Kenya, BraSpecS reached 0.44, and with the local model, 0.92. As both datasets were inserted in the GSSL, results varied; the Israel prediction became worse but it improved for Kenya. The examples indicate that BraSpecS can be used depending on the country and soil similarity. This may be the track for SSLs of the world, a user can seek SSL that provides the best result for its spectral data. For example, 'if' Israel did not have any SSL, which SSL would they choose to use: BraSpecS or a global one? As we suggest for the future, the user can test both and use the best one according to the user's objective. We need to have global, continental, country, and local SSLs. To alleviate the problem of the current model of central data custodian, SSLs need to have a distributed model where users can contribute towards a global SSL but their data ownership and privacy are preserved [49]. In the future, distributed SSLs linked via a system such as a blockchain would ensure data ownership is respected, yet users can still access the global dataset.

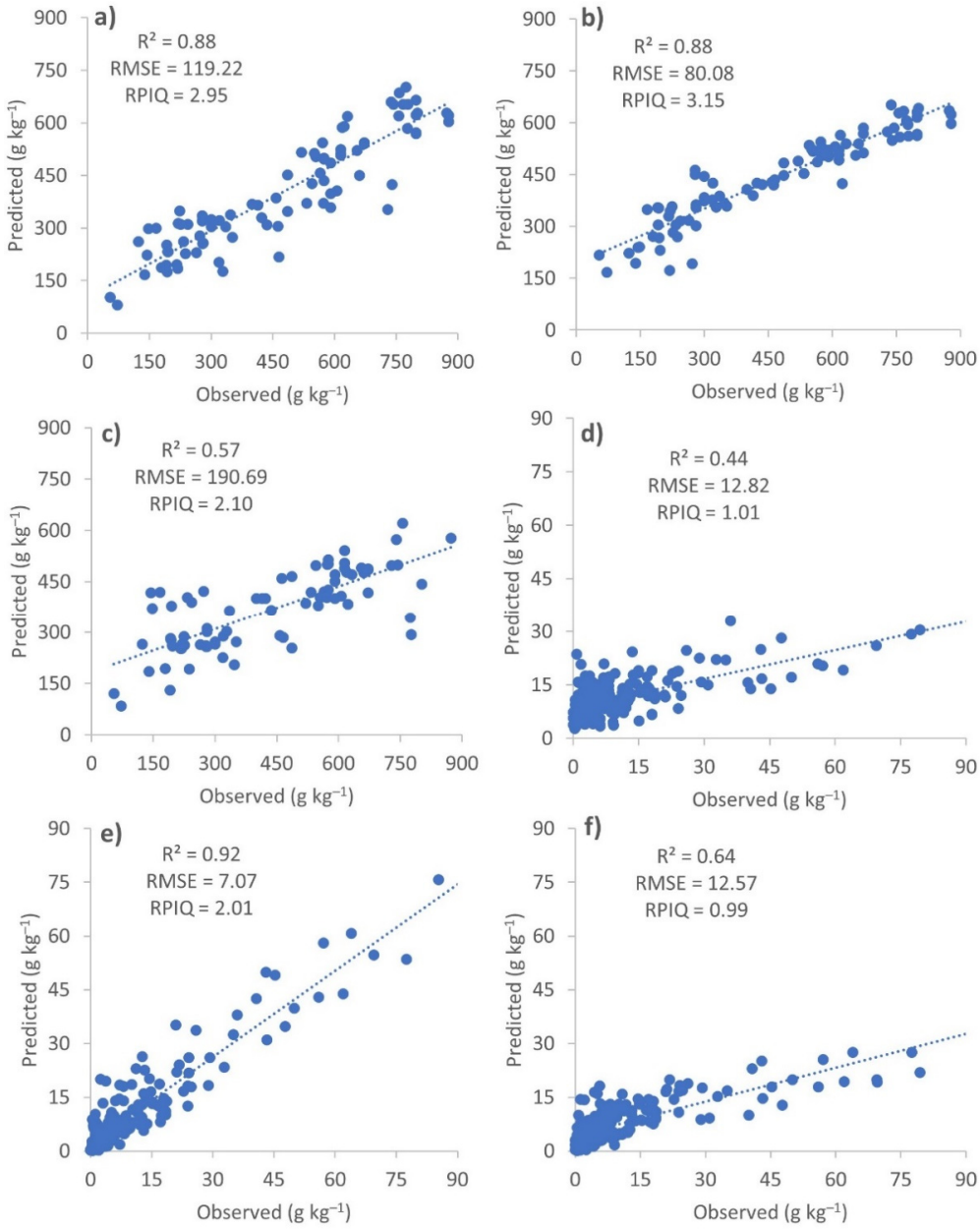


Figure 13. Israel scatter plots for Clay: (**a**) Observed × Predicted by the BraSpecS model, (**b**) Observed × Predicted ExCSSL, and (**c**) Observed × Predicted GSSL. Kenya scatter plots for SOC: (**d**) Observed × Predicted by the BraSpecS model, (**e**) Observed × Predicted ExCSSL, and (**f**) Observed × Predicted GSSL.

5. Conclusions and Final Considerations

It was possible to construct a platform where its importance can cover resource-limited regions which may consider the opportunity to submit spectra and retrieve estimated soil data in an established online service, such as BraSpecS. End-users can already interact with infrared technology.

The BraSpecS system facilitates dynamic communication between worldwide users and delivers important soil information. The presented system can be applied for several purposes, including research, farming, soil analytical laboratories, industries, consulting companies, creation of startups, teaching, pedology research, digital soil mapping, precision agriculture, and more. The system is user-friendly and does not require the user to have competency and literacy in soil spectral modeling. Users simply insert the soil spectra into

the system and receive soil estimates with statistical metrics information. The user also has the ability to find the owners of spectra, request data, get in contact, and build partnerships. The platform also allows users to view spectral patterns of soil classes, soil texture, SOC content, and many other soil properties. Finally, the user receives a report of the soil model results for spectra that were submitted to the web platform. Nevertheless, this system is not without drawbacks and limitations, which will be resolved once the system is in use and feedback is received from worldwide users.

In the case of clay and SOC quantification, vis–NIR–SWIR presented reasonable data (best for clay) and MIR reached the best in both cases. In terms of data model population, statistics increased as follows: Global Model—BraSpecS model—Local model.

The cascading future growth of GSSL holds much promise through the pooling of local and regional soil spectral data to represent the global soil variability. End-user and stakeholder engagement in the BraSpecS will be profoundly important to build a robust and sustainable global soil spectral library that serves the greater good of soils. Our approach of a community-driven GSSL in which people and stakeholders participate in collecting new soil samples and spectra that represent actual soil conditions will overcome some of the limitations of global SOC and clay maps/grids derived via digital soil mapping that were mainly produced from legacy soil data representing historic soil conditions. Monitoring soil change is profoundly important in an age of multi-hazard natural disasters (such as wildfires and flooding), global climate change, and interconnected soil, food, social, economic, and ecological dilemmas. The need for up-to-date SOC, clay, and other soil properties is imminent. BraSpecS has operationalized soil spectroscopy to address these urgent needs for accurate and current soil information as well as assessment of soil change around the globe.

6. Future Works

The quality of model performances is influenced by multiple factors including: (a) quality and consistency of spectral data (absence of an agreed protocol); (b) quality and consistency of soil wet laboratory analysis (method and accuracy); and (c) soil forming factors, mineralogy, biome, and other environmental factors that influence soil genesis. These should be addressed in future studies to achieve the best results to assess soils. We would like to stress that community-driven global soil spectral libraries allow the contributors to construct clay, SOC, and other soil property models of various kinds. This work paves the way to investigate spectra modeling using spectra similarities from a global or regional SSL. Similarities of soil and environmental factors between regions matter when transferring spectral models overseas is certainly the best approach, as indicated in our work.

An important direction for the future should be a filter inside the system, which would achieve the best spectra to create the model (i.e., spectral fitting). Thus, for each spectrum (and for each soil attribute), the user could have a different model, increasing the use worldwide. Looking at the same ideas, the system of soil classification (which is also presented here) could be improved with photo and soil description, to assist pedologists and soil survey. Along these insights, the system could also allow pre and post-processing which would gather other end-users in a larger community such as researchers, consultants, and industries.

Web sites

(1) The Brazilian Soil spectral Service (BraSpecS): besbbr.com.br or http://143.107.213.227/layout/_en/apresenta_temp.php. (Accessed on 2 February 2022)
(2) The Brazilian Soil Spectral Library (BSSL): https://bibliotecaespectral.wixsite.com/english or http://143.107.213.227/layout/. (Accessed on 2 February 2022)
(3) The Group that developed, Geotehnologies on Soil Science Group (GeoCis): https://esalqgeocis.wixsite.com/english. (Accessed on 2 February 2022)
(4) Corresponding author profile: https://jamdemat.wixsite.com/dematte (Accessed on 2 February 2022)

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs14030740/s1, Table S1: Number of samples and contact of researchers of other countries.

Author Contributions: Conceptualization, J.A.M.D.; methodology, A.F.d.S.P.; software, A.F.d.S.P.; writing—original draft preparation, J.A.M.D., A.F.d.S.P., R.R.P., N.A.R., L.F.C.R.; writing—review and editing, J.A.M.D., A.F.d.S.P., R.R.P., N.A.R., L.F.C.R., F.A.d.O.M., B.M., S.G., Y.G., E.B.D., A.G., C.G., S.C., N.F., S.A., D.F., J.K.M.B., C.W., A.B., S.N., N.A.H., H.B., J.M.M.-B., N.E.Q.S.; visualization, R.R.P.; supervision, J.A.M.D.; project administration, J.A.M.D.; funding acquisition, J.A.M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by São Paulo Research Foundation (FAPESP) (grant numbers 2014/22262-0, 2016/26176-6, and 2020/04306-0).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are grateful to the Geotechnologies in Soil Science Group (GeoCiS/ESALQ-USP; http://esalqgeocis.wixsite.com/english accessed on 30 January 2022) for team support. We are grateful to Sérgio Ricardo Scagnolato, Luciano Brandine de Negreiros and Fábio Chaddad for the site technical support. We specially knowledge all participants that delivered spectra and dataset to make this work possible, in special the ones that can be found in the Brazilian Soil Spectral Library (https://bibliotecaespectral.wixsite.com/english/lista-de-cedentes accessed on 30 January 2022; [16]) and the overseas ones whose countries are indicated in the tables and in the site (https://gossats.wixsite.com/home accessed on 30 January 2022). This work was supported by the São Paulo Research Foundation (FAPESP) (grant numbers 2014/22262-0 2016/26176-6, and 2020/04306-0).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lal, R.; Bouma, J.; Brevik, E.; Dawson, L.; Field, D.J.; Glaser, B.; Hatano, R.; Hartemink, A.E.; Kosaki, T.; Lascelles, B.; et al. Soils and sustainable development goals of the United Nations: An international union of soil sciences perspective. *Geoderma Reg.* **2021**, *25*, e00398. [CrossRef]
2. Rossel, R.A.V.; McBratney, A.B. Soil chemical analytical accuracy and costs: Implications from precision agriculture. *Aust. J. Exp. Agric.* **1998**, *38*, 765. [CrossRef]
3. Demattê, J.A.M.; Dotto, A.C.; Bedin, L.G.; Sayão, V.M.; Souza, A.B. E Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma* **2019**, *337*, 111–121. [CrossRef]
4. Viscarra Rossel, R.A.; Cattle, S.R.; Ortega, A.; Fouad, Y. In situ measurements of soil colour, mineral composition and clay content by vis—NIR spectroscopy. *Geoderma* **2009**, *150*, 253–266. [CrossRef]
5. Ben-Dor, E.; Chabrillat, S.; Demattê, J.A.M.; Taylor, G.R.; Hill, J.; Whiting, M.L.; Sommer, S. Using imaging spectroscopy to study soil properties. *Remote Sens. Environ.* **2009**, *113*, S38–S55. [CrossRef]
6. Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; MacDonald, L.M.; McLaughlin, M.J. The performance of visible, near and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. [CrossRef]
7. Nocita, M.; Stevens, A.; van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Ben Dor, E.; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 139–159.
8. Stoner, E.R.; Baumgardner, M.F. Characteristic variations in reflectance of surface soils. *Soil Sci. Soc. Am. J.* **1981**, *45*, 1161–1165. [CrossRef]
9. Brown, D.J.; Shepherd, K.D.; Walsh, M.G.; Dewayne Mays, M.; Reinsch, T.G. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290. [CrossRef]
10. 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.; et al. A global spectral library to characterize the world's soil. *Earth Sci. Rev.* **2016**, *155*, 198–230. [CrossRef]
11. Bellinaso, H.; Demattê, J.A.M.; Romeiro, S.A. Soil spectral library and its use in soil classification. *Rev. Bras. Ciência Solo* **2010**, *34*, 861–870. [CrossRef]
12. Demattê, J.A.M.; Dotto, A.C.; Paiva, A.F.; Sato, M.V.; Dalmolin, R.S.; Maria do Socorro, B.; da Silva, E.B.; Nanni, M.R.; ten Caten, A.; Noronha, N.C.; et al. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* **2019**, *354*, 113793. [CrossRef]

13. Brodský, L.; Klement, A.; Penížek, V.; Kodešová, R.; Borůvka, L. Building soil spectral library of the Czech soils for quantitative digital soil mapping. *Soil Water Res.* **2011**, *6*, 165–172. [CrossRef]
14. Gogé, F.; Joffre, R.; Jolivet, C.; Ross, I.; Ranjard, L. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemom. Intell. Lab. Syst.* **2012**, *110*, 168–176. [CrossRef]
15. Knadel, M.; Deng, F.; Thomsen, A.; Greve, M. Development of a Danish national Vis-NIR soil spectral library for soil organic carbon determination. In *Digital Soil Assessments and Beyond*; CRC Press: Boca Raton, FL, USA, 2012; pp. 403–408.
16. Cambule, A.H.; Rossiter, D.G.; Stoorvogel, J.J.; Smaling, E.M.A. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma* **2012**, *183*, 41–48. [CrossRef]
17. Bas, M.V.; Meléndez-Pastor, I.; Navarro-Pedreño, J.; Gómez, I.; Mataix-Solera, J.; Hernández, E. Saline soils spectral library as a tool for digital soil mapping. In Proceedings of the EGU General Assembly Conference Abstracts, Vienna, Austria, 7–12 April 2013.
18. Viscarra Rossel, R.A. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. *J. Geophys. Res.* **2011**, *116*, F04023. [CrossRef]
19. Shi, Z.; Wang, Q.; Peng, J.; Ji, W.; Liu, H.; Li, X.; Viscarra Rossel, R.A. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Sci. China Earth Sci.* **2014**, *57*, 1671–1680. [CrossRef]
20. Ji, W.; Li, S.; Chen, S.; Shi, Z.; Viscarra Rossel, R.A.; Mouazen, A.M. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil Tillage Res.* **2016**, *155*, 492–500. [CrossRef]
21. Liu, S.; Shen, H.; Chen, S.; Zhao, X.; Biswas, A.; Jia, X.; Shi, Z.; Fang, J. Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment. *Geoderma* **2019**, *348*, 37–44. [CrossRef]
22. Condit, H.R. The spectral reflectance of American soils. *Photogramm. Eng.* **1970**, *36*, 955–966.
23. Wijewardane, N.K.; Ge, Y.; Wills, S.; Loecke, T. Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment Project. *Soil Sci. Soc. Am. J.* **2016**, *80*, 973–982. [CrossRef]
24. Wijewardane, N.K.; Ge, Y.; Wills, S.; Libohova, Z. Predicting physical and chemical properties of US Soils with a mid-infrared reflectance spectral library. *Soil Sci. Soc. Am. J.* **2018**, *82*, 722–731. [CrossRef]
25. Baldock, J.A.; McNally, S.R.; Beare, M.H.; Curtin, D.; Hawke, B. Predicting soil carbon saturation deficit and related properties of New Zealand soils using infrared spectroscopy. *Soil Res.* **2019**, *57*, 835. [CrossRef]
26. Hergarten, C.; Nazarmavloev, F.; Wolfgramm, B. Building a soil spectral library for Tajikistan comparing local and global modeling approaches. In Proceedings of the 3rd Global Workshop on Proximal Soil Sensing, Potsdam, Germany, 26–29 May 2013; pp. 265–269.
27. Tuğrul, K.M. Soil management in sustainable agriculture. In *Soil Management and Plant Nutrition for Sustainable Crop Production*; Hasanuzzaman, M., Teixeira Filho, M.C.M., Nogueira, T.A., Eds.; IntechOpen: London, UK, 2019; pp. 111–126.
28. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
29. Dangal, S.; Sanderman, J.; Wills, S.; Ramirez-Lopez, L. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Syst.* **2019**, *3*, 11. [CrossRef]
30. Baumann, P.; Helfenstein, A.; Gubler, A.; Keller, A.; Meuli, R.G.; Wächter, D.; Lee, J.; Viscarra Rossel, R.; Six, J. Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring. *SOIL* **2021**, *7*, 525–546. [CrossRef]
31. Ng, W.; Minasny, B.; Jeon, S.H.; McBratney, A. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Secur.* **2022**, *6*, 100043. [CrossRef]
32. World Agroforestry (ICRAF). International Soil Reference and Information Centre (ISRIC) ICRAF-ISRIC Soil VNIR Spectral Library. Available online: https://data.worldagroforestry.org/dataset.xhtml?persistentId=doi:10.34725/DVN/MFHA9C (accessed on 11 November 2021).
33. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the largest expandable soil dataset for Europe: A review. *Eur. J. Soil Sci.* **2018**, *69*, 140–153. [CrossRef]
34. Stevens, A.; Nocita, M.; Tóth, G.; Montanarella, L.; van Wesemael, B. Prediction of soil organic carbon at the european scale by visible and near infrared reflectance spectroscopy. *PLoS ONE* **2013**, *8*, e66409. [CrossRef]
35. Vågen, T.-G.; Winowiecki, L.; Tondoh, J.E.; Desta, L.T.; Gumbricht, T. Mid-Infrared Spectra (MIRS) from ICRAF Soil and Plant Spectroscopy Laboratory: Africa Soil Information Service (AfSIS) Phase I 2009–2013. Available online: https://data.worldagroforestry.org/dataset.xhtml?persistentId=doi:10.34725/DVN/QXCWP1 (accessed on 25 January 2022).
36. Tziolas, N.; Tsakiridis, N.; Ben-Dor, E.; Theocharis, J.; Zalidis, G. A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation. *Geoderma* **2019**, *340*, 11–24. [CrossRef]
37. Woodwell Climate Research Center (OSSL). Open Soil Spectroscopy Library. Available online: https://www.woodwellclimate.org/open-soil-spectral-library/ (accessed on 10 December 2021).
38. United States Department of Agriculture (USDA). Rapid Carbon Assessment (RaCA) Project. Available online: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054164 (accessed on 22 September 2021).
39. Ben Dor, E.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245*, 112–124. [CrossRef]
40. Bunting, P. Pre-processing of remotely sensed imagery. In *The Roles of Remote Sensing in Nature Conservation*; Diaz-Delgado, R., Lucas, R., Hurford, C., Eds.; Springer: Manhattan, NY, USA, 2017; pp. 39–63.

41. Dotto, A.C.; Dalmolin, R.S.D.; Caten, A.T.; Gris, D.J.; Ruiz, L.F.C. AlradSpectra: A quantification tool for soil properties using spectroscopic data in R. *Rev. Bras. Ciência Solo* **2019**, *43*, e0180263. [CrossRef]
42. Sang, K.; Piovan, S.; Fontana, G.L. A WebGIS for Visualizing historical activities based on photos: The project of Yunnan–Vietnam railway web map. *Sustainability* **2021**, *13*, 419. [CrossRef]
43. R Development Core Team. The R Project for Statistical Computing. Available online: https://www.r-project.org (accessed on 6 June 2021).
44. Apache Software Foundation Community-Led Development "The Apache Way". Available online: https://www.apache.org/ (accessed on 23 July 2020).
45. Abeysinghe, S. *PHP Team Development*; Packt Publishing: Birmingham, UK, 2009.
46. Tasneem, S.; Ammar, R. Performance study of a distributed web server: An analytical approach. *J. Softw. Eng. Appl.* **2012**, *05*, 855–863. [CrossRef]
47. Mitchell, L.J. *PHP Web Services: APIs for the Modern Web*; O'Reilly Media: Newton, MA, USA, 2016.
48. Horner, J. RApache: Web Application Development with R and Apache. Available online: http://www.rapache.net/ (accessed on 25 June 2021).
49. Padarian, J.; Minasny, B.; McBratney, A.B. Machine learning and soil sciences: A review aided by machine learning tools. *SOIL* **2020**, *6*, 35–52. [CrossRef]
50. Manohar, H.M.; Appaiah, S. Stabilization of FIFO system and inventory management. *Iternational Res. J. Eng. Technol.* **2017**, *4*, 5631–5638.
51. Walkley, A.; 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 Sci.* **1934**, *37*, 29–38. [CrossRef]
52. Van Raij, B.; Andrade, J.C.; Cantarella, H.; Quaggio, J.A. *Análise Química para Avaliação da Fertilidade de Solos Tropicais*; IAC: Campinas, Brazil, 2001; ISBN 9788585564056.
53. Teixeira, P.C.; Donagema, G.K.; Fontana, A.; Teixeira, W.G. *Manual de Métodos de Análise de Solo*, 3rd ed.; EMBRAPA: Brasília, Brazil, 2017.
54. Zhang, X.; Huang, B.; Ji, J.F.; Hu, W.Y.; Sun, W.X.; Zhao, Y.C. Quantitative prediction of soil salinity content with visible-near infrared hyper-spectra in Northeast China. *Spectrosc. Spectr. Anal.* **2012**, *32*, 2075–2079.
55. Zhang, X.; Huang, B. Prediction of soil salinity with soil-reflected spectra: A comparison of two regression methods. *Sci. Rep.* **2019**, *9*, 5067. [CrossRef]
56. Quinlan, J. Learning with continuous classes. In Proceedings of the AI'92, 5th Australian Conference on Artificial Intelligence, Hobart, Tasmania, 16–18 November 1992; Adams, A., Sterling, L., Eds.; World Scientific: Singapore, 1992; pp. 343–348.
57. Khaledian, Y.; Miller, B.A. Selecting appropriate machine learning methods for digital soil mapping. *Appl. Math. Model.* **2020**, *81*, 401–418. [CrossRef]
58. Xiong, X.; Grunwald, S.; Myers, D.B.; Kim, J.; Harris, W.G.; Comerford, N.B. Holistic environmental soil-landscape modeling of soil organic carbon. *Environ. Model. Softw.* **2014**, *57*, 202–215. [CrossRef]
59. Moura-Bueno, J.M.; Dalmolin, R.S.D.; Horst-Heinen, T.Z.; Grunwald, S.; ten Caten, A. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. *Geoderma* **2021**, *393*, 114981. [CrossRef]
60. Angelopoulou, T.; Balafoutis, A.; Zalidis, G.; Bochtis, D. From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—A Review. *Sustainability* **2020**, *12*, 443. [CrossRef]
61. Kuhn, M. Caret: Classification and Regression Training. R Package Version 6.0-90. 2010. Available online: https://CRAN.R-project.org/package=caret (accessed on 20 March 2020).
62. Tsakiridis, N.L.; Keramaris, K.D.; Theocharis, J.B.; Zalidis, G.C. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* **2020**, *367*, 114208. [CrossRef]
63. Yang, J.; Wang, X.; Wang, R.; Wang, H. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis—NIR spectroscopy. *Geoderma* **2020**, *380*, 114616. [CrossRef]
64. Ng, W.; Minasny, B.; Montazerolghaem, M.; Padarian, J.; Ferguson, R.; Bailey, S.; McBratney, A.B. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* **2019**, *352*, 251–267. [CrossRef]
65. Sequeira, C.H.; Wills, S.A.; Grunwald, S.; Ferguson, R.R.; Benham, E.C.; West, L.T. Development and update process of VNIR-based models built to predict soil organic carbon. *Soil Sci. Soc. Am. J.* **2014**, *78*, 903–913. [CrossRef]
66. Summerauer, L.; Baumann, P.; Ramirez-Lopez, L.; Barthel, M.; Bauters, M.; Bukombe, B.; Reichenbach, M.; Boeckx, P.; Kearsley, E.; Van Oost, K.; et al. The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis. *SOIL* **2021**, *7*, 693–715. [CrossRef]
67. Moura-Bueno, J.M.; Dalmolin, R.S.D.; Horst-Heinen, T.Z.; ten Caten, A.; Vasques, G.M.; Dotto, A.C.; Grunwald, S. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? *Sci. Total Environ.* **2020**, *737*, 139895. [CrossRef]
68. Terra, F.S.; Demattê, J.A.M.; Viscarra Rossel, R.A. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geoderma* **2015**, *255–256*, 81–93. [CrossRef]
69. Clairotte, M.; Grinand, C.; Kouakoua, E.; Thébault, A.; Saby, N.P.A.; Bernoux, M.; Barthès, B.G. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma* **2016**, *276*, 41–52. [CrossRef]

70. Dotto, A.C.; Dalmolin, R.S.D.; ten Caten, A.; Grunwald, S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma* **2018**, *314*, 262–274. [CrossRef]
71. Ramirez-Lopez, L.; Wadoux, A.M.J.-C.; Franceschini, M.H.D.; Terra, F.S.; Marques, K.P.P.; Sayão, V.M.; Demattê, J.A.M. Robust soil mapping at the farm scale with vis–NIR spectroscopy. *Eur. J. Soil Sci.* **2019**, *70*, 378–393. [CrossRef]
72. Rizzo, R.; Medeiros, L.G.; Mello, D.C.; de Marques, K.P.P.; de Souza Mendes, W.; Quiñonez Silvero, N.E.; Dotto, A.C.; Bonfatti, B.R.; Demattê, J.A.M. Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in southeastern Brazil. *Geoderma* **2020**, *361*, 114018. [CrossRef]
73. Brown, D.J. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma* **2007**, *140*, 444–453. [CrossRef]
74. Grunwald, S.; Yu, C.; Xiong, X. Transferability and scalability of soil total carbon prediction models in Florida, USA. *Pedosphere* **2018**, *28*, 856–872. [CrossRef]
75. Sankey, J.B.; Brown, D.J.; Bernard, M.L.; Lawrence, R.L. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma* **2008**, *148*, 149–158. [CrossRef]
76. Wetterlind, J.; Stenberg, B. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *Eur. J. Soil Sci.* **2010**, *61*, 823–843. [CrossRef]
77. Ng, W.; Minasny, B.; Jones, E.; McBratney, A. To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. *Geoderma* **2022**, *406*, 115501. [CrossRef]
78. McDowell, M.L.; Bruland, G.L.; Deenik, J.L.; Grunwald, S.; Knox, N.M. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma* **2012**, *189*, 312–320. [CrossRef]
79. Araújo, S.R.; Wetterlind, J.; Demattê, J.A.M.; Stenberg, B. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *Eur. J. Soil Sci.* **2014**, *65*, 718–729. [CrossRef]
80. Debaene, G.; Niedźwiecki, J.; Pecio, A.; Żurek, A. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma* **2014**, *214–215*, 114–125. [CrossRef]
81. Grunwald, S.; Thompson, J.A.; Boettinger, J.L. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Sci. Soc. Am. J.* **2011**, *75*, 1201–1213. [CrossRef]
82. McBratney, A.; de Gruijter, J.; Bryce, A. Pedometrics timeline. *Geoderma* **2019**, *338*, 568–575. [CrossRef]
83. Kopačková, V.; Ben-Dor, E. Normalizing reflectance from different spectrometers and protocols with an internal soil standard. *Int. J. Remote Sens.* **2016**, *37*, 1276–1290. [CrossRef]
84. Chabrillat, S.; Gholizadeh, A.; Neumann, C.; Berger, D.; Milewski, R.; Ogen, Y.; Ben-Dor, E. Preparing a soil spectral library using the internal soil standard (ISS) method: Influence of extreme different humidity laboratory conditions. *Geoderma* **2019**, *355*, 113855. [CrossRef]
85. Gholizadeh, A.; Carmon, N.; Klement, A.; Ben-Dor, E.; Borůvka, L. Agricultural soil spectral response and properties assessment: Effects of measurement protocol and data mining technique. *Remote Sens.* **2017**, *9*, 1078. [CrossRef]
86. Genot, V.; Colinet, G.; Bock, L.; Vanvyve, D.; Reusen, Y. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. *J. Near Infrared Spectrosc.* **2011**, *19*, 117–138. [CrossRef]
87. Knox, N.M.; Grunwald, S.; McDowell, M.L.; Bruland, G.L.; Myers, D.B.; Harris, W.G. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma* **2015**, *239*, 229–239. [CrossRef]
88. Riedel, F.; Denk, M.; Müller, I.; Barth, N.; Gläßer, C. Prediction of soil parameters using the spectral range between 350 and 15,000 nm: A case study based on the permanent soil monitoring program in saxony, Germany. *Geoderma* **2018**, *315*, 188–198. [CrossRef]

Computers and Electronics in Agriculture 194 (2022) 106796

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Verifying the predictive performance for soil organic carbon when employing field Vis-NIR spectroscopy and satellite imagery obtained using two different sampling methods

James Kobina Mensah Biney *

Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic
The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Department of Landscape Ecology, Lidická 25/27, Brno 602 00, Czech Republic

ARTICLE INFO

Keywords:
Sampling design
Soil organic carbon (SOC)
Agricultural soil
Spatial variability
Pretreatment

ABSTRACT

In soil research, the most employed sampling design techniques can be categorized as random sampling (stratified or simple random (SR)) or systematic techniques (transects or grid). Many other sampling approaches have also been developed by researchers based on these sampling principles. The purpose of this study is to compare the differences in SOC prediction when using field spectra (FS) and Sentinel-2 (S2) data collected separately through SR and grid design (GD) on the same agricultural field. Additionally, the impact of spectral indices on S2 data in a merged data approach under the two-sampling strategies will also be tested. The data for each sampling method were obtained based on a previous study in which 130 soil samples were collected from a full grid design (with 40 m spacing) covering the entire area. Although the full GD method was used for this current study, the distance between the samples was increased (80 m apart). The schemes were therefore structured for the collection of 65 samples in the field for each sampling technique. However, 63 samples were collected with the GD because two of the sampling points fell on rocky areas and were eliminated accordingly. For SR sampling, the study field was not stratified, and no requirements were used for minimum sample spacing. Sixty-five samples and spectral data were collected at various locations. To achieve the mentioned objective, this study builds a five-fold cross-validation model based on support vector machines (SVMs). Different pretreatment combinations were also implemented. The results showed that the GD was better than the SR approach using the merged dataset ($R^2_{CV} = 0.45$, $RMSE_{CV} = 0.26$, RPD = 1.41, bias = −0.0073); however, SOC prediction under SR sampling using FS yielded the highest accuracy and lowest error margin ($R^2_{CV} = 0.60$, $RMSE_{CV} = 0.21$, RPD = 1.66, and bias = 0.0045). Despite the above-mentioned disparity between the single and merged data, this study shows that using different sampling design methods on the same field separately is a very promising approach for SOC estimation, particularly in fields with low SOC. Based on these results, the robustness of this approach should be investigated next in future studies using larger sample sizes as well as other modeling techniques. Based on these results, the robustness of this approach should be investigated next in future studies using larger sample sizes as well as other modeling techniques.

1. Introduction

Soil organic carbon is considered one of the essential soil components because of its contribution to soil fertility and crop production (Muñoz and Kravchenko, 2011). However, its impact extends beyond the physical, chemical, and biological properties of the soil to include climatic change mitigation (Karmakar et al., 2016). Therefore, its prediction is crucial because decision-makers and farmers rely on these values to map up an informed strategy for the benefit of a community, region, and a country as a whole. Nonetheless, the efficient prediction of SOC content and its change depends largely on the spatial variability of SOC and the long-term and seasonal temporal variability; monitoring programs that can address this variation are vital (Allen et al., 2010).

Soil sampling is primarily intended to provide a representative

* Address: Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague-Suchdol, Czech Republic
E-mail address: biney@af.czu.cz.

https://doi.org/10.1016/j.compag.2022.106796
Received 24 October 2021; Received in revised form 30 January 2022; Accepted 13 February 2022
Available online 21 February 2022

sample of soil properties in an area, using mainly a statistical approach to achieve that. However, the number of collected samples can vary, depending on the variation in the region (James and Wells, 1990). The first step for obtaining the spatial distribution of soil characteristics is an appropriate sampling design because the predictive performance of the soil property accuracy under study could be compromised if an inappropriate sampling design is chosen (Zhu et al., 2015). According to Chang et al. (2016), the applicability of the sampling design is subject to the soil monitoring objective. However, for Stein and Ettema (2003), the main goal is to ensure that the prediction error for mapping or spatial investigation is minimized.

According to an FAO report titled "Soil carbon monitoring based on repeated measurement'' (FAO, 2012), for designing a scheme, especially for SOC prediction, certain key criteria must be considered. This includes the intensity of the sampling, sampling interval, soil layers and, most importantly, the position of the sample points and the total number that must be collected. SOC typically has high spatial variability because, in some instances, soil-forming features include different spatial scales, thus revealing the spatial variance characteristics of SOC may face some difficulties (Miller et al., 2016). Some researchers have also reported that SOC prediction varies with sampling density, especially in complex terrain areas compared to simple topography regions (e.g., Heim et al. 2009; Tsui et al. 2013).

Therefore, choosing the most suitable sampling design for a particular study area is an enormous task to accomplish, particularly for prediction in a large area or even a new field (Stevens et al., 2008). This is because according to Yang et al. (2019), distinct sampling design techniques produce different sample sets, which will normally have a direct impact on the predictive performance of soil properties. However, according to Carter & Gregorich (2007), the aim of selecting a suitable sampling method for SOC estimation is that the sampling scheme should be a total representation of the entire field at a reasonable cost.

Over the past decades and more, there have been numerous sampling methods used by researchers according to the project types and budgets across several fields (reviewed in Gilbert, 1987; Mulla and McBratney, 2002; Liao et al., 2009). For instance, in soil research, the most commonly employed sampling design methods can be categorized as random or systematic or by convenience (Carter & Gregorich, 2007; Eggleston et al., 2006). The most commonly used sample procedures with respect to systematic sampling are transects or grid techniques, while for random sampling, they are simple and stratified random sampling (Carter & Gregorich, 2007). Many other sampling techniques have also been used or developed that are comparable to or modified versions of the above-mentioned sampling designs. Details can be found in the studies by Brus et al. (2011), Allen et al. (2010), Vasat et al. (2012), Wang et al. (2021), and Minasny and McBratney (2006). With the availability of both satellite and space-borne images, new sampling approaches have been developed not only to cover the whole area but also to consider the spatial representation (Hank et al., 2019). However, the fact remains that ground-based measurements will always serve as the primary or reference value for remote sensing-based analysis. Numerous studies in soil science have used both remote sensing (RS) (satellite imagery) and proximal soil sensing (PSS) (e.g., spectroscopic reflectance) (Ben-Dor and Banin, 1994; Ben-Dor and Banin, 1995) for the analysis of soil samples obtained using several sampling approaches. For example, Stevens et al. (2008) predicted SOC content on a regional scale using laboratory, field and airborne spectroscopy.

The prediction of SOC with PSS techniques, proven to be rapid and cost-effective, has provided more accurate results compared to the traditional approach, which is time-consuming (e.g., Viscarra Rossel and Bouma, 2016). PSS also differs significantly in terms of its sensing platform, data structure, and research objectives. However, one of the primary limitations of spectroscopy (under PSS) is its inability to cover larger-scale sampling areas because it is a point-based approach, and the costs involved, including labour, can be enormous. RS, on the other hand, is known for its ability to cover larger-scale sites, as well as for monitoring and to enhance result classification, but its limited spatial resolution as well as the effect of external environmental factors are some of its key limitations (Lagacherie et al., 2008; Angelopoulou et al., 2019). It is assumed that if the proper sampling design is not selected, the predictions of soil properties, such as SOC may be negatively impacted.

In this context, the main goal of this study is to assess the prediction accuracy for SOC when using two different sampling designs—simple random and grid design—to collect field spectra and Sentinel-2 data separately on the same field for each sampling method. The main goal is to determine which sampling designs are consistent across the two datasets and whether they improved SOC predictive performance. The assessment will be performed using (i) field spectra and Sentinel-2 data individually and (ii) in a combined form (Sentinel-2 + calculated spectral indices). Spectral indices from sentinel imagery will be used for the combined data approach. This is to also verify the effect of these indices on the Sentinel-data under two different sampling strategies.

2. Materials and methods

2.1. Study area and soil sampling

The study site (Fig. 1) is a 22-ha field in Nová Ves nad Popelkou (50°31′ N; 15°24′ E) in the Czech Republic's central Bohemian region, with a mean altitude of 185 m a.s.l. It is an agricultural field that stretches two kilometres southeast of the town of Lomnice nad Popelkou along the Popelka River. The main crops grown in the area are winter wheat and spring barley. The areas are primarily rural, characterized by dissected relief with side valleys and toe-slopes. Additionally, the study field chosen is a representative of soil capes, which are homogenous and comparable in terms of terrain characteristics, land management, and climatic conditions (Schmidt et al., 2010). During the measurement campaign, the soil was bare and undisturbed (it had not recently been plowed). According to the World Reference Base (WRB) for soil resources (IUSS Working Group WRB, 2014), the soils of these regions are characterized mainly as Cambisols on sedimentary rocks.

Two sets of field spectra data were collected separately in the field using two different sampling methods, namely, simple random (SR) and grid-design (GD). The sampling points for each sampling strategy were selected in reference to a previous study in the same study area (Biney et al., 2020). For the said study, 130 soil samples were collected using a full grid design (with 40 m spacing) covering the entire study area. Although the full GD technique was still adopted for the current study, covering the entire area; however, the spacing between each sampling point was modified (80 m apart) compared to the previous study. This was done to also assess the impact of two grid-sampling designs (having different sample spacing) on the estimation of SOC in a study field low in SOC content. The schemes were therefore structured for the collection of 65 samples in the field for each sampling technique. Notwithstanding, 63 samples were collected with the GD because two of the sampling points fell on rocky areas and were eliminated accordingly. Additionally, for the SR design, three different sampling schemes were created using the data management tool in ArcGIS (ESRI, The Redlands, CA, USA) to ensure that the generated result was not by chance. However, due to financial constraints, only one of the ArcGIS SR design schemes was used in the field. Using a descriptive statistics plot, this scheme was compared to the two remaining schemes (where no field samples were collected) (result not shown). The SR scheme used in the field was selected for further analysis because it provided a better mean, coefficient of variation (CV), and standard deviation (SD). Additionally, during the SR sampling design, the study field was not stratified, and no requirements were used for minimum sample spacing. The sampling points (SR and GD) were created separately and fed into a GeoXM (Trimble Inc., Sunnyvale, California, USA) receiver with an accuracy of 1 m before the field visit, and the positions of each sampling point were located in the field using the same instrument.

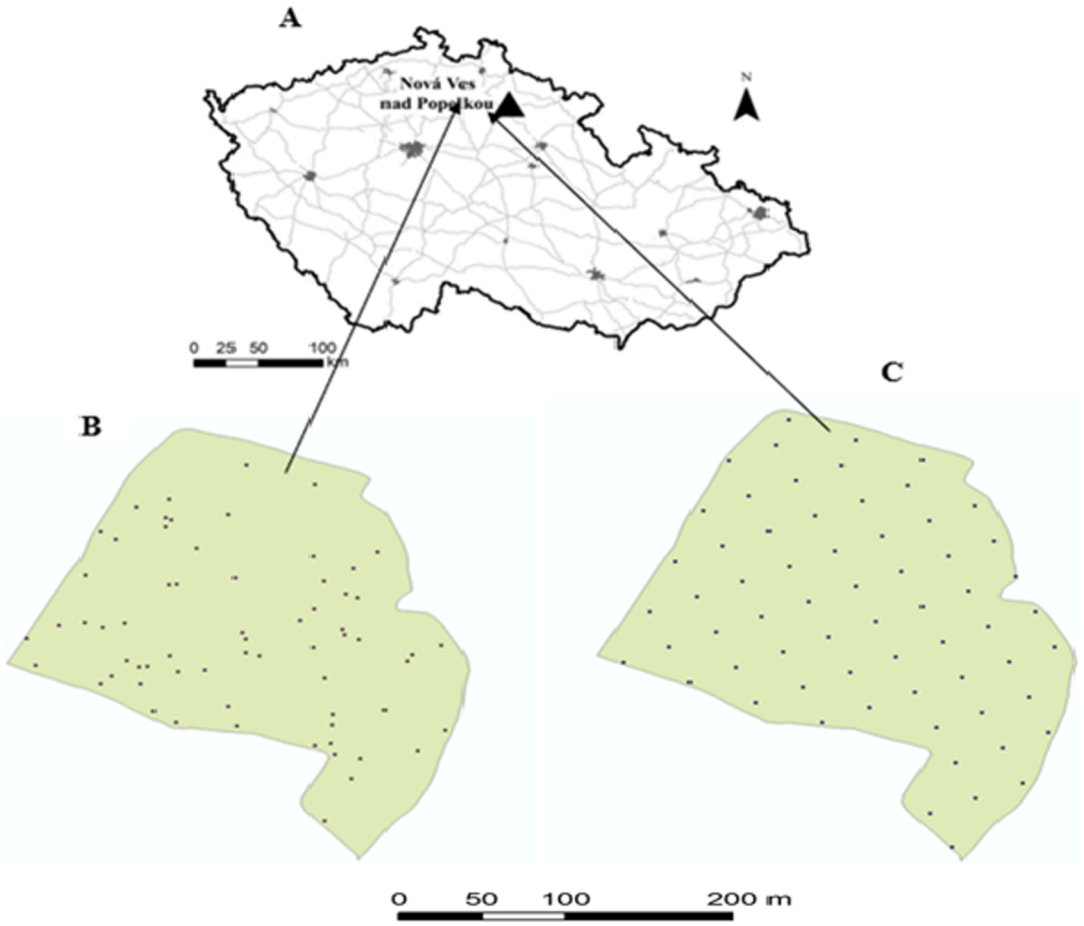


Fig. 1. Sampling area location in the Czech Republic (A), display of simple random (B) and grid design (C) sampling points at NováVes nad Popelkou.

2.2. Field spectral measurement and soil analysis

The GD sampling points were spread evenly across the whole field, as shown in Fig. 1C. The field spectra (FS) were measured on May 24, 2019, using an ASD Field Spec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) across the 350–2500 nm wavelength range. The spectroradiometer spectral resolution was 2 nm for the region of 350–1050 nm and 10 nm for the region of 1050–2500 nm. The spectral measurements were carried out on the soil surface at three individual locations around each sampling point, uniformly distributed and then averaged into one composite spectrum per sampling point for each spectral region. The spectrometer was standardized prior to the first measurement and after every 10 measurements using a white Spectralon™ panel (Labsphere, North Sutton, NH, USA) (Shi et al., 2016). The same procedure was repeated for the SR field spectra sample (Fig. 1B). In addition, soil samples (using the SR and GD) were collected from each sampling point (depth, 0–2 cm) during the FS measurement, placed into a well-labelled bag and conveyed to the laboratory (composite samples, approximately 130 to 170 g of soil) for further analysis. These samples were then air-dried, gently crushed, and sieved (≤2 mm) before being analysed for SOC (ISO 11464:2006).

2.3. Sentinel-2 imagery acquisition and analysis

The Multispectral Sentinel-2B imagery used was a cloud-free image level 2A product, which means it is ready to be used right away because the suppliers using Sen2Corprocessor have already processed it. These processes include geometric, radiometric, and atmospheric corrections. The best sentinel-2 imagery used was imagery dated July 10, 2019, obtained from the Copernicus Open Access Hub of the European Space Agency. Additionally, two other comparable imagery dates, June 15 and 30, were also collected to obtain imagery that was closer to the field sampling date. The imagery (S2) consists of 13 spectral bands. These spectral bands range from visible and near-infrared (vis-NIR) to shortwave infrared (SWIR). There were four bands at 10 m resolution [((B2, 490 nm), (B3, 560 nm), (B4, 665 nm), (B8, 842 nm))] and six bands at 20 m resolution [((B5, 705 nm), (B6, 740 nm), (B7, 775 nm), and (B8A, 865 nm)]. The remaining bands were two SWIR large bands [(B11, 1610 nm) and (B12, 2190 nm)] and three 60 m resolution bands [(B1, 443 nm), (B9, 940 nm), and (B10, 1380 nm)]. Prior to extracting these bands, a resampling by pixel resolution approach (10 m resolution as the reference) was performed to ensure that all the bands were at the resolution. This was done using the SNAP software. For further analysis, three bands (B1, B9, and B10) were excluded. This implies that all the remaining bands used were at the same resolution of 10 m. Technical details of the S2 bands used in this study can be found in work book of the European Space Agency. (2010).

2.4. Merged data set approach

Combining both remote sensing data at different spatial resolutions (10, 20, and 60 m) with soil data collected with field spectra at 2 or 3 nm is problematic or not appropriate. These datasets cannot be combined because they are associated with different support sizes: tens of meters for remote sensing data and point support for field survey data. Due to this difficulty, the study did not merge the S2 and FS data. However, future studies exploring this approach are recommended to help verify the impact of high-resolution data such as the FS merged with S2 on SOC estimation, since, according to Grunwald et al. (2015), there is no single sensor or technique that can accurately estimate all soil properties, such

as SOC.

Spectral indices, usually a single number derived from the spectral reflectance of two or more wavebands (Ji and Peters, 2007), are believed to improve the interpretation of remote sensing data (Grunwald et al., 2015). For instance, according to Peng et al. (2015), spectral indices are preferable to raw spectral bands as indicators for SOC estimation, because when compared to spectral bands, these indices are more specific in distinguishing bare soil, vegetation, and other factors that could influence SOC estimation. Moreover, spectral indices were utilized by Jin et al. (2017) and Liu et al. (2015) for the estimations of various soil properties. To explore this hypothesis on the two different sampling strategies on the same study field, nine calculated spectral indices as covariates were extracted from the S2 imagery to serve as an additional covariate that will be merged with the S2 bands to verify the impact of merged data on SOC estimation. These indices include the normalized difference vegetation index (NDVI) (Rouse et al., 1974), soil adjusted vegetation index (SAVI) (Huete, 1988), green–red vegetation index (GRVI) (Tucker, 1979), and modified soil adjusted vegetation index (MSAVI) (Qi et al., 1994). The remaining indices are the brightness index (BI) (Escadafal, 1989), redness index (RI) (Pouget et al., 1990), infrared percentage vegetation index (IPVI) (Crippen, 1990), normalized difference red edge (NDRE) (Barnes et al., 2000), and finally the differenced vegetation index (DVI) (Richardson and Wiegand, 1977). SNAP was used to obtain the values at the sampling locations.

The formulas to derive these indices are shown in Table 1.

2.5. Dataset preprocessing and predictive modeling performance

The initial spectroscopic measurements were within the range of 350–2500 nm, but before further processing, the extremely noisy part of the spectra (350–399 nm) was removed to improve the accuracy of prediction. Subsequently, the datasets were subjected to the following set of pretreatment techniques, sg (Savitzky–Golay) from the signal R package (Signal developers, 2013), dwt (discrete wavelet transformation) calculated with the dwt function from the wavelet R package (Aldrich, 2013), d1 (first-order derivative) (Duckworth, 2004), sg_d1, msc (multiplicative scatter correction), which was calculated using the pls R package (Mevik and Wehrens, 2007), snv (standard normal variate), log (logarithmic transformation (log(1/R))), dwt_log, dwt_log_snv, sg_msc, sg_log, raw, sg_log_msc, dwt_log_msc, log_msc, log, sg_d1, sg_log_snv, and sg_log_msc, to optimize the fitting of target values against spectra. The four best pretreatment algorithms will be reported for each approach to avoid nonrelevant results. We refer to Vašát et al. (2017) and Biney et al. (2020) for more details on the pretreatment algorithms used for this current study. R software was used to evaluate all the pretreatment techniques (R Development Core Team, 2014). In total, the following datasets were obtained for the two sampling methods for both the individual and merged data approaches: (a) individual approaches [GD (two datasets were used, that is, S2 and FS) and SR (two datasets were also used, that is, S2 and FS)], (b). for the merged data approaches [GD (S2 + spectra indices (SID)) and for the SR (S2 + SID]. In addition, a Pearson correlation was computed between SOC and each of the data that formed the merged dataset (S2 and SID). This was done to aid in determining the level of relationship between SOC and the datasets in question. This was performed for each sampling strategy used in this study (GD and SR).

Table 1
Derived indices.

Index	S2 imagery Classification
NDVI	$\frac{B8-B4}{B8+B4}$
IPVI	$\frac{1}{2}(NDVI+1)$
NDRE	$\frac{B8-B5}{B8+B5}$
SAVI	$\frac{(B8-B4)*(1+L)}{B8-B4+L}$ L = 0.5
GRVI	$\frac{B3-B4}{B3+B4}$
DVI	B8-B4
BI	$\frac{\sqrt{(B4*B4)+(B3*B3)}}{2}$
MSAVI	$\frac{(1+L)(B8-B4)}{(B8+B4+L)}$
RI	$\frac{B4*B4}{B3*B3*B3}$

The predictive models were calibrated using the SVM algorithm, a machine learning method (Vapnik, 2000). SVM has been utilized across different scientific disciplines for classification and regression problems. The e1071 R package (Meyer et al., 2014) is for computations, specifically the linear kernel function. Before fitting the final models, the cost parameter was fine-tuned using the built-in tuning function. The epsilon parameter was left at its default value (0.1). For information on the tuning of hyperparameters and the optimal number of cost values and the type of kernel for SVM, the reader is directed to Biney et al. (2021a, 2021b). The overall model (SVM) output was evaluated by the index of determination (R^2_{CV}), the ratio of performance to deviation (RPD) [which is calculated as the ratio of standard error of the estimate (RMSEcv) to standard deviation of the data], and the root mean square error of prediction (RMSEcv) [measures the model overall prediction accuracy].

The whole dataset for each sampling approach was randomly divided into two subsets for calibration (75%) and validation (25%). Each model was fitted using the calibration data, while the validation evaluated model performance. A 5-fold leave-group-out cross validation (5-fold LGO CV) was applied to the training dataset for each of the models utilized. The R^2_{CV} ranges from 0 to 1, where $R^2_{CV} = 1$ is the optimal value, and for RPD, Chang and Laird's (2002) categorization was applied: RPD > 2 indicates good models, RPD between 1.4 and 2 indicates moderate predictive ability, and RPD = 1.4 indicates weak models.

3. Results and discussion

3.1. SOC descriptive statistics and distribution of GD and SR sampling points in the study field

Table 1 is a statistics summary for SOC characteristics in soil samples for the GB and SR approaches, comprising standard deviation (SD), coefficient of variation (CV), minimum (Min), maximum (Max), mean value, and skewness. Both sampling approaches captured the same Min and Max value and showed an approximate normal distribution (skewness = 1.21/1.24). According to Blanca et al. (2013), acceptable values of skewness should fall between − 3 and + 3. The CV for both sampling methods indicated a moderate variability of SOC ($0.15 < CV < 0.35$), which could be attributed to random factors such as inherent spatial variation in soil (Wilding, 1985). Generally, the overall result (Table 2) signifies a low to medium SOC content for the area.

This study examined the use of two different sampling techniques (on the same field) for designing field sampling schemes to obtain soil samples with field spectroscopy (FS) and Sentinel-2 (S2) imagery to separately evaluate the predictive performance of SOC. The results show that the GD sampling (under the systematic sampling approach) and the SR sampling methods (under the random sampling approach) varied considerably from each other (Table 3) in terms of predictive performance and the level of error margin for SOC estimation. One of the primary goals of selecting an appropriate sampling procedure is to accurately estimate the values of soil attributes in specific locations without difficulties. As stated by Brus et al. (2011), the significance of selecting a sampling design that is simple to implement and leads to easily interpretable statistical estimation procedures cannot be overestimated. Soils are heterogeneous on a range of spatial and temporal scales due to soil property variations (Fitter et al., 2000). For example, if

Table 2
Soil organic carbon (SOC) descriptive statistics for grid design (GD) and simple random (SR).

SOC (%)	Mean	Median	[a]SD	Kurtosis	Skewness	[b]Min	[c]Max	[d]CV (%)
GD	1.48	1.49	0.34	4.60	1.21	0.87	2.93	23.00
SR	1.49	1.50	0.33	4.92	1.24	0.87	2.93	22.00

[a] SD: standard deviation.
[b] Min: minimum.
[c] Max: maximum.
[d] CV: coefficient of variation.

Table 3
Statistics of the five-fold leave-group-out cross-validation for SOC prediction from field spectra (FS) and S2 using the individual datasets collected through grid design (GD) and simple random (SR) sampling methods with support vector machine (SVM) algorithm on different preprocessing combination algorithms.

		GD FS		
Treatment	dwt_log	dwt_log_snv	sg_msc	sg_log
R^2cv	0.42	0.49	0.44	0.47
RMSEcv	0.26	0.24	0.25	0.24
RPD	1.30	1.35	1.32	1.37
Bias	−0.0316	0.0055	−0.0100	−0.0036
		S2		
Treatment	raw	dwt_log	Dwt	d1
R^2cv	0.22	0.38	0.35	0.14
RMSEcv	0.34	0.29	0.30	0.36
RPD	1.10	1.25	1.23	1.04
Bias	0.0220	−0.0067	0.0159	−0.0131
		SR FS		
Treatment	log_snv	sg_log_msc	dwt_log_msc	dwt_log_snv
R^2cv	0.60	0.51	0.48	0.39
RMSEcv	0.21	0.24	0.24	0.26
RPD	1.66	1.38	1.38	1.29
Bias	0.0045	0.0055	0.0121	0.0134
		S2		
Treatment	sg_log	log_msc	Log	sg_d1
R^2cv	0.24	0.24	0.30	0.32
RMSEcv	0.35	0.37	0.33	0.34
RPD	1.06	1.01	1.11	1.14
Bias	0.0025	0.0380	0.0117	−0.0060

the heterogeneity of variance of soil properties is not adequately represented with the selected sampling design method, the prediction accuracy for such a field with all collected sampling points could be affected. Fig. 1B & C reveals that the two-sampling approach's selected sampling points and positions in the same field differed, indicating differences in prediction performance.

3.2. Comparison and evaluation of SOC prediction with field spectra under GD and SR sampling techniques

The individual datasets collected from these sampling approaches (GD and SR) in Table 3 clearly indicate that the data collected through the SR sampling method using the FS provided not only the overall most appropriate result but also the least error margin (R^2_{CV} = 0.60, $RMSE_{CV}$ = 0.21, RPD = 1.66, bias = 0.0045, log_snv) compared to the data collected through the GD sampling approach (R^2_{CV} = 0.49, $RMSE_{CV}$ = 0.24, RPD = 1.35, bias = 0.0055, dwt_log_snv). In comparison to previous studies in this field (Biney et al., 2020; Gholizadeh et al., 2018), the obtained results could be described as an improvement in SOC accuracy as well as a decrease in error margin. Nonetheless, the modified GD sampling point distance (spacing distance) had no significant effect on the final SOC prediction accuracy when compared to the results obtained from a previous study (Biney et al., 2020). For example, the study by Biney et al. (2020) utilized only the GD sampling approach, with a limited distance (40 m apart), as opposed to the current study, which employs the GD and SR approaches separately with an increase in sample spacing for the GD approach (80 m apart), as previously mentioned. Both results (using the GD) were almost the same. Additionally, for the study of Gholizadeh et al. (2018), the conditioned Latin Hypercube Sampling (cLHS) method was utilized. This implies that sometimes the sampling design does not depend on the number of sampling points or distances but rather on the heterogeneity of the soil in the area, which may vary on a case-by-case basis. As a result, failure to select the most appropriate sampling method could have an impact on the ultimate result. Mishra et al. (2009) showed 48 samples having better-represented variance heterogeneity of the soil than 240 samples which did not and for prediction, the latter was better than the former. According to Zhao et al. (2016), for sampling structure establishment, one should ensure that there is a compromise between uniform and irregular distribution of sampling points, and especially the inclusion of previous knowledge in the sampling design should be a key consideration. For Stenberg et al. (2010), the main feature controlling the final prediction accuracy of SOC might be the SOC variability characteristics. However, Heung et al. (2017) stated that the collection of spectral data in the field for an accurate prediction of SOC is highly dependent on a proper sampling design prior to the field visit.

The findings of the current study contradict those of Mallarino and Wittry (2004), who concluded that grid sampling was perhaps the most efficient method for determining organic matter and soil pH. Presumably, the good results of the SR sampling approach were explained by better representativeness and possibly a normal statistical distribution. Moreover, SR sampling techniques have also been acknowledged for soil surveys by some researchers (Roels and Jonker, 1983; Shadish, 2002). For instance, according to Shadish (2002), SR sampling ensures that the findings from the dataset under consideration are close to the outcome that would be obtained if the full study field was used. Although the SR was better than the GD for this study, when compared to some other sampling approaches, especially the conditioned Latin Hypercube Sampling (cLHS) stratified random strategy (Minasny and McBratney, 2006), SR could show some downsides. These include its inability to select a sample if the units or parameters are widely dispersed, it may not achieve good geographical coverage, and finally its poor distributions of variables (Minasny and McBratney 2006; Worsham et al., 2012). One of the key advantages of the cLHS stratified random strategy is that, instead of using geographic space, it makes use of the available feature space (Worsham et al., 2012). Consequently, its main drawback is the possibility of obtaining an ideal Latin hypercube, which tends to decrease as the number of input variables or sample locations increases. This is due to the effect of equal stratification, which can result in selecting more samples that are near the mean of the sample space.

Although GD sampling has been noted to provide a comprehensive view of the research area, it is also described as both expensive and time-consuming (Nanni et al. 2011; Stamper et al. 2014). Nevertheless, one of its major issues is locating where the soil sample should be taken within each grid or the possible size of the grid, since taking the centres of the chosen grid cells is inaccurate, as all other point locations would have zero chances of being selected. Thus, the estimated quality measures will necessarily apply to a finite set of point locations rather than the entire grid. As a result, several researchers (Nanni et al. 2011; Stamper et al. 2014; Wollenhaupt and Wolkowski, 1994) have expressed

dissatisfaction with the approaches used to select suitable grid sizes or cells that will provide a comprehensive understanding of soil property variations. To address this issue, Brus et al. (2011) suggest that one or more point locations be chosen randomly in each grid cell. As per Flowers et al. (2005), the best grid size for sampling a field may not be identified until after the sampling has taken place. For this study field, for instance, both grid and cLHS in previous studies (on the same field) failed to achieve a better result than this current study using FS collected through SR sampling.

3.3. Assessment of SOC prediction using S2 imagery obtained via GD and SR sampling techniques

Table 3 also indicates that both sampling conditions (GD and SR) using S2 data provided poor results. Nevertheless, the results attained through GD sampling ($R^2_{CV} = 0.38$, RMSEcv = 0.29, RPD = 1.25, bias = −0.0067, dwt_log) were slightly better than those obtained through the SR sampling strategy ($R^2_{CV} = 0.32$, RMSEcv = 0.34, RPD = 1.14, bias = −0.0060 sg_dl). Fig. 2 also shows that the Pearson correlation coefficient results for these sampling approaches using the SOC and S2 bands show poor correlation with some bands under GD (e.g., B3, B5, B6, B7, B12), whereas there is a vague or no relationship with all the bands under the SR method (except for B4, B11), which were even poorly correlated with SOC). This might have accounted for the slightly better results obtained with S2 under GD. Notwithstanding, some of the bands were moderately to strongly correlated with each other. This implies that these bands may be important for specific soil types or conditions but may not work for all datasets. This shows that S2 imagery underpredicts SOC content using the two sampling methods by missing out on some useful information. For instance, SOC was poorly correlated or showed no relationship with these bands regardless of which sampling strategy was adopted (Fig. 2).

Furthermore, S2 imagery being influenced by vegetation cover on the surface of the field cannot be ruled out. This is because, due to cloud cover issues, the study could not obtain S2 imagery that corresponded to the very same date when the field sample was collected. As stated earlier, the best S2 imagery used was dated July 10, 2019, because in terms of prediction, it was better than the two other comparable collected imagery dated June 10 and 30, 2019, as stated above (results not shown). Normally, in July, the presence of vegetation at its initial stage in agricultural fields is possible. According to Bartholomeus et al. (2011), the presence of vegetation cover in the field before spectral measurement can sometimes influence the spectral reflectance shape,

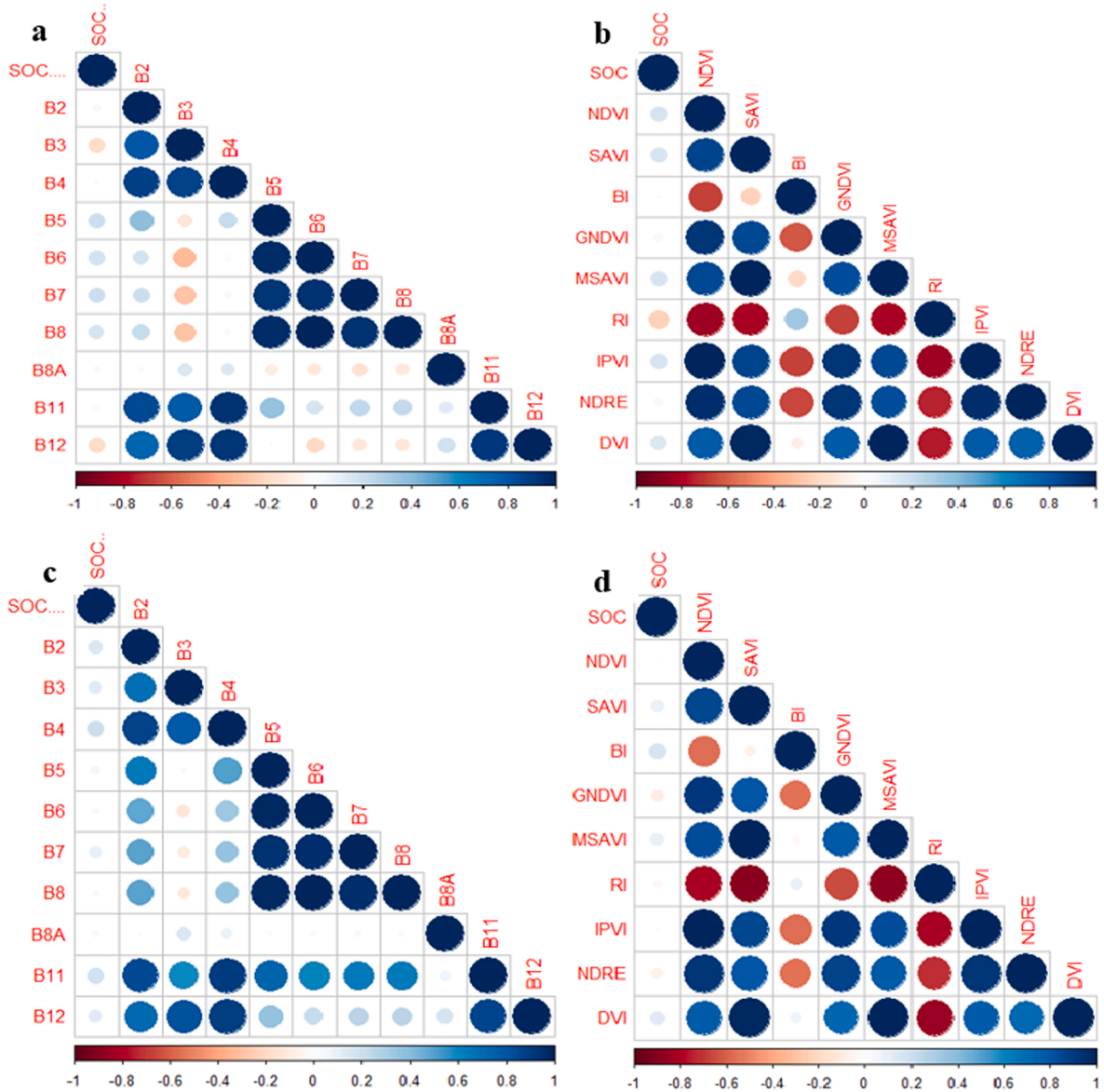


Fig. 2. Pearson correlation of SOC based on S2 bands and calculated indices from different sampling strategies [GD (a, b) and SR (c, d) sampling approaches].

and this could influence the predictive performance of soil properties, such as SOC. For the current study, the field sampling was taken on bare soil. One probable solution to the issue of obtaining Sentinel-2 imagery captured on bare soil is the possibility of downloading several images to investigate the optimum date so that the field measurements can be adjusted accordingly. However, getting these agricultural fields at one's "optimum time" is highly dependent on the landowners' or farmers' decisions, which, in most cases, are final because these fields are utilized for commercial purposes. Despite these occurrences, several studies have confirmed the superiority of spectroscopy over remote sensing approaches (e.g., Gomez et al., 2018; Stevens et al., 2008; Lagacherie et al., 2008) for SOC estimation.

3.4. Combined dataset model and pretreatment performance under different sampling methods for SOC estimation

The best result for the combined data (Table 4) was obtained using GB sampling (R^2_{CV} = 0.45, RMSEcv = 0.26, RPD = 1.41, bias = −0.0073, sg_log_snv). This is because the S2 band as a covariate under GD is marginally better than SR. Additionally, some of the indices obtained under GD show some partial relationship with SOC, particularly the RI (Fig. 2b). This index could have positively influenced the merged data. According to Madeira et al. (1997), the RI index (spectral index) is noted for measuring soil redness variation and can also account for the intensity of absorption features that characterize certain soil properties (e.g., iron oxide, organic matter, etc.). Table 4 also shows that the SR sampling approach was outperformed by the GD sampling method in the combined dataset strategy using the S2 and the indices datasets. One possibility is that S2 imagery performance under the SR approach was poor. Therefore, indices obtained from the same spectral data using the SR sampling approach did not add much important information to the models. Moreover, in terms of correlation, the indices derived under the SR method were poorly correlated with SOC (Fig. 2d). Although, under the right conditions, some of these sampling methods can yield meaningful results, there is no guarantee that, under a different set of circumstances or approaches, the good results will be sustained (Cochran, 1977). Therefore, it is worth mentioning that when merging two or more distinct sets of data acquired using different measurement techniques, employing mathematical and statistical techniques (e.g., data fusion) that will only merge the positive attributes of the datasets and exclude the defects from these datasets will be much more appropriate for increasing prediction accuracy. If not, as stated above, the dataset's vulnerabilities may have an impact on the overall combined dataset. Although the overall best merged data result was less comparable to the overall best result using the individual data (Table 3), the addition of the spectral indices under the GD improved the overall result provided by the S2 data (Table 4).

Table 4
Statistics of the five-fold leave-group-out cross-validation for SOC prediction from the combined dataset approach (S2 + SID) for grid design (GD) and simple random (SR) sampling methods with SVM on different preprocessing combination algorithms.

		GD S2 + SID		
Treatment	sg_log_snv	dwt_log_snv	sg_log_msc	log_snv
R^2cv	0.45	0.38	0.34	0.37
RMSEcv	0.26	0.28	0.27	0.28
RPD	1.41	1.33	1.24	1.22
bias	−0.0053	−0.0068	−0.0042	−0.0119
		SR S2 + SID		
Treatment	sg_log_snv	sg_log_msc	dwt_log_msc	dwt_log_snv
R^2cv	0.37	0.27	0.32	0.24
RMSEcv	0.29	0.32	0.30	0.34
RPD	1.27	1.21	1.19	1.14
bias	0.0035	0.0045	0.0111	0.0151

The scatterplots (Fig. 3) show the results of the predicted versus observed SOC predictive accuracy using the FS and S2 sensing platform datasets obtained through the GD and SR sampling methods. In this study, the disparities in predicting SOC across different platforms vary from one platform to the other as well as from one sampling approach to the other based on all R^2_{CV}, $RMSE_{CV}$ and RPD parameters. There were substantially more scatters in SOC prediction, particularly when using S2 (for individual data) as well as the combined data (FS + SID, both sampling approaches). This confirms that the difference in SOC prediction between the two sampling approaches was not obvious because, as shown in Fig. 1, the sampling points captured by the sampling techniques were distributed across different locations within the same field. Additionally, it was shown that the individual (particularly field spectra) and the combined dataset differed. According to Gomez et al. (2018), light sources, instrumental noise, spatial resolution, and atmospheric conditions can all have a negative impact on sensor accuracy for spectroscopy and remote sensing during measurement.

The use of several pretreatment combinations (Table 3, 4) provided mixed results on the various datasets. Pretreatment approaches are essential in eliminating irrelevant information and may primarily improve the performance of the predictive model. Choosing the right pretreatment, however, is critical for spectroscopy and remote sensing data since no method can be considered universally suitable for any dataset (Engel et al., 2013). The log_snv combination, noted for linearization attainment between the spectra and SOC content as well as the correction of the light scattering effect, was the best pretreatment algorithm (helped to achieve the best overall result). Additionally, the log algorithm also eliminates baseline effects and enhances the spectral features, thereby causing an increase in prediction accuracy (Schlerf et al. 2010). According to Liu et al. (2019), numerous factors could influence the dataset; the selection of only one pretreatment could sometimes fail to cope with all of these factors. The best pretreatment, therefore, depends on the dataset used, and a combination of several pretreatments could be more useful, but further study is needed on that. This implies that various pretreatments on sampling design are not influenced but can be differently efficient when combined with different sampling designs.

4. Conclusion

This study examined the possibility of enhancing the precision of SOC prediction in an agricultural field low in SOC content by employing different sampling techniques [simple random (SR) and grid design (GD)], utilizing field spectra (FS) and Sentinel-2 imagery (S2) based on the SVM algorithm. To predict the SOC content, two approaches based on FS and S2 imagery were developed for each sampling technique: using the individual data only and merging the S2 band with nine spectral indices (SID) extracted from the Sentinel-2 imagery (S2 + SID). The study showed that the GD and SR data obtained with S2 provided poor results compared to the same sampling conditions using the FS. One of the possible reasons was that, in terms of correlation, all S2 selected bands were either poorly correlated or showed no relationship with SOC no matter which sampling strategy was adopted. However, the overall result provided by the S2 data under the GD sampling method improved with the addition of the spectral indices (merged data approach) (R^2_{CV} = 0.45, RMSEcv = 0.26). However, the best result and lowest error margin for the entire study were achieved with the SR sampling method using FS (R^2_{CV} = 0.6, RMSEcv = 0.21). Despite the fact that in the past, different sampling methods have been utilized in the same study field (particularly grid and conditioned Latin Hypercube Sampling (cLHS) methods), the obtained results with SR sampling for the current study are by far the most appropriate. Moreover, this shows that the use of different sampling designs on the same field separately is a promising option to predict SOC accurately, especially in fields with low SOC content. The addition of these approaches in future studies is

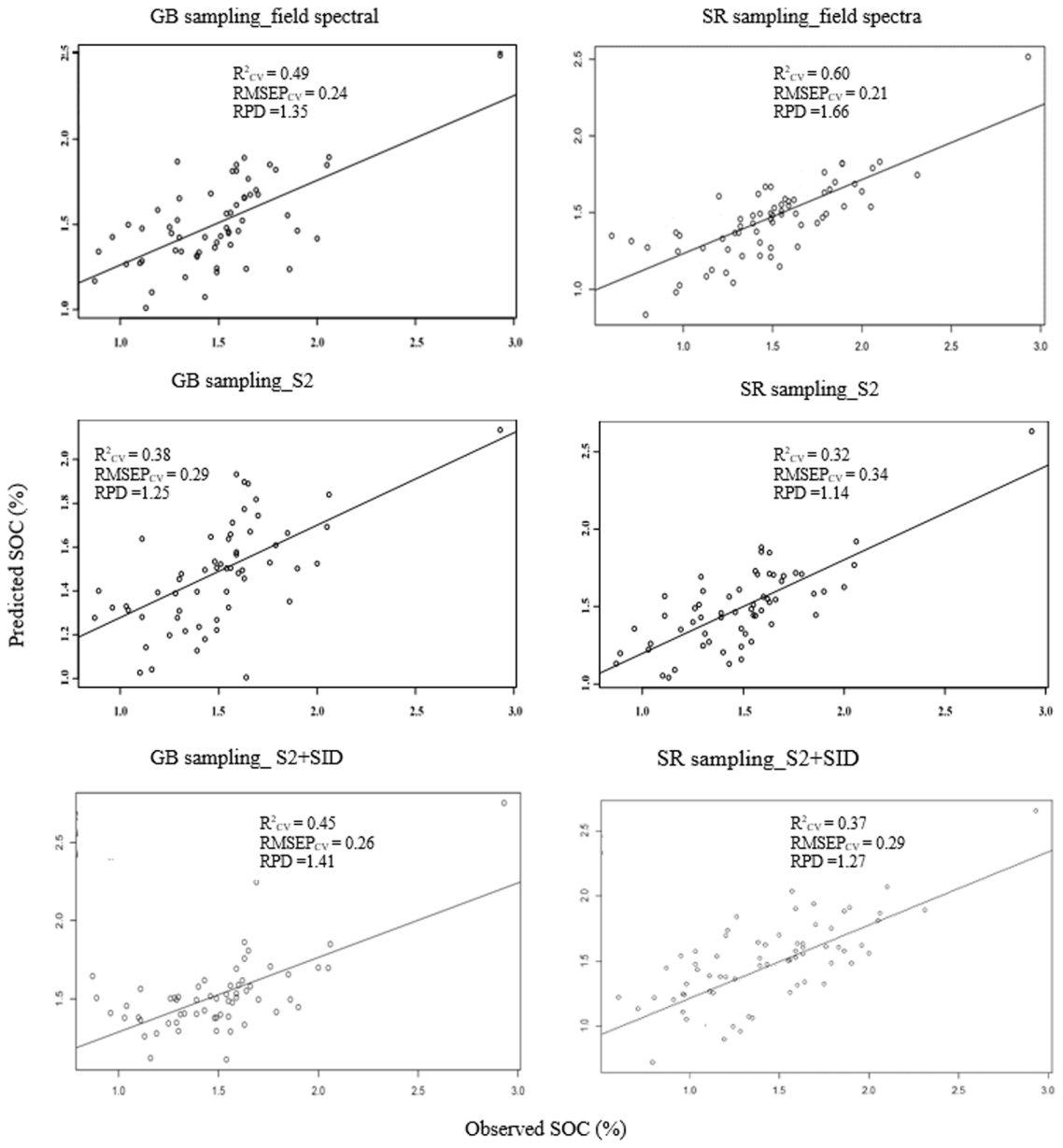


Fig. 3. Predicted versus observed SOC with FS and S2 sensing platforms obtained through different sampling strategies (GD and SR) for both individual and combined datasets.

highly recommended, especially using larger datasets to test their robustness.

Funding

This study was supported by an internal grant from the Czech University of Life Sciences Prague, project No. SV20-5-21130, and by the Technology Agency of the Czech Republic project No. SS02030018.

CRediT authorship contribution statement

James Kobina Mensah Biney: Conceptualization, Methodology, Writing – original draft, Visualization, Data curation, Investigation, Software.

Declaration of Competing Interest

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

Acknowledgements

The author also acknowledges the support of the European Regional Development Fund Project Center for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products (No. CZ.02.1.01/0.0/0.0/16_019/0000845).

I also gratefully acknowledge Aleš Klement, Karel Němeček and Miroslav Fér for their assistance in soil sampling and spectral measurement and Radim Vašát for some of the software used.

References

Aldrich, E., 2013. Wavelets: A package of functions for computing wavelet filters, wavelet transforms and multiresolution analyses. R package version 0.3-0. URL http://CRAN.R-project.org/package=wavelets.

Allen, D.E., Pringle, M.J., Page, K.L., Dalal, R.C., 2010. A review of sampling designs for the measurement of soil organic carbon in Australian grazing lands. Rangeland J. 32 (2), 227–246. https://doi.org/10.1071/RJ09043.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. 11 (6), 676. https://doi.org/10.3390/rs11060676.

Barnes, E.M., Clarke, T.R., Richards, S.E., Colaizzi, P.D., Haberland, J., Kostrzewski, M., Moran, M.S., 2000, July. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In: Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA (Vol. 1619).

Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., Tychon, B., 2011. Soil organic carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 13 (1), 81–88. https://doi.org/10.1016/j.jag.2010.06.009.

Ben-Dor, E., Banin, A., 1994. Visible and near-infrared (0.4–1.1 μm) analysis of arid and semiarid soils. Remote Sens. Environ. 48 (3), 261–274. https://doi.org/10.1016/0034-4257(94)90001-9.

Ben-Dor, E., Banin, A., 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59 (2), 364–372. https://doi.org/10.2136/sssaj1995.03615995005900020014x.

Biney, J.K.M., Blöcher, J.R., Borůvka, L., Vašát, R., 2021a. Does the limited use of orthogonal signal correction pretreatment approach to improve the prediction accuracy of soil organic carbon need attention? Geoderma 388, 114945. https://doi.org/10.1016/j.geoderma.2021.114945.

Biney, J.K.M., Borůvka, L., Chapman Agyeman, P., Němeček, K., Klement, A., 2020. Comparison of Field and Laboratory Wet Soil Spectra in the Vis-NIR Range for Soil Organic Carbon Prediction in the Absence of Laboratory Dry Measurements. Remote Sensing 12 (18), 3082. https://doi.org/10.3390/rs12183082.

Biney, J.K.M., Vašát, R., Blöcher, J.R., Borůvka, L., Němeček, K., 2021b. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. Sci. Total Environ. 151805.

Blanca, M.J., Arnau, J., López-Montiel, D., Bono, R., Bendayan, R., 2013. Skewness and kurtosis in real data samples. Methodology 9 (2), 78–84. https://doi.org/10.1027/1614-2241/a000057.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x.

Carter, M.R., Gregorich, E.G. (Eds.), 2007. Soil sampling and methods of analysis. CRC Press. https://doi.org/10.1201/9781420005271.

Chang, C.-W., Laird, D.A., 2002. Near-infrared reflectance spectroscopic analysis of soil C and N. Soil Sci. 167 (2), 110–116.

Chang, X., Bao, X., Wang, S., Zhu, X., Luo, C., Zhang, Z., Wilkes, A., 2016. Exploring effective sampling design for monitoring soil organic carbon in degraded Tibetan grasslands. J. Environ. Manage. 173, 121–126. https://doi.org/10.1016/j.jenvman.2016.03.010.

Cochran, W.G., 1977. Sampling Techniques, 3rd edition. John Wiley & Sons, New York.

Crippen, R., 1990. Calculating the vegetation index faster. Remote Sens. Environ. 34 (1), 71–73. https://doi.org/10.1016/0034-4257(90)90085-Z.

Duckworth, J., 2004. Mathematical data preprocessing. Near-infrared spectroscopy in agriculture 44, 113–132. https://doi.org/10.2134/agronmonogr44.c6.

Eggleston, S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K., 2006. IPCC guidelines for national greenhouse gas inventories. IGES, Hayama, Japan.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L. M.C., 2013. Breaking with trends in preprocessing? TrAC, Trends Anal. Chem. 50, 96–106. https://doi.org/10.1016/j.trac.2013.04.015.

Escadafal, R., 1989. Remote sensing of arid soil surface color with Landsat thematic mapper. Adv. Space Res. 9 (1), 159–163. https://doi.org/10.1016/0273-1177(89)90481-X.

European Space Agency, 2016. Sen2Cor 2.2.1—Software Release Note. European Space Agency, Paris, France.

FAO, 2012. Soil carbon monitoring based on repeated measurements. FAO Forestry Paper 2012 No.168 pp. http://www.fao.org/3/i2793e/i2793e02.pdf.

Fitter, A., Hodge, A., Robinson, D., 2000. Plant response to patchy soils. In: Hutchings, M.J., John, E.A., Stewart, A.J.A. (Eds.), The Ecological Consequences of Environmental Heterogeneity, pp. 71–90.

Flowers, M., Weisz, R., White, J.G., 2005. Yield-based management zones and grid sampling strategies: Describing soil test and nutrient variability. Agron. J. 97 (3), 968–982. https://doi.org/10.2134/agronj2004.0224.

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sens. Environ. 218, 89–103. https://doi.org/10.1016/j.rse.2018.09.015.

Gilbert, R.O., 1987. Statistical methods for environmental pollution monitoring. John Wiley & Sons.

Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P., Roger, J.M., Briottet, X., 2018. Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. Remote Sens. Environ. 204, 18–30. https://doi.org/10.1016/j.rse.2017.10.047.

Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of soil and remote sensing data to model soil properties. Adv. Agron. 131, 1–109. https://doi.org/10.1016/bs.agron.2014.12.004.

Hank, T.B., Berger, K., Bach, H., Clevers, J.G.P.W., Gitelson, A., Zarco-Tejada, P., Mauser, W., 2019. Space-borne imaging spectroscopy for sustainable agriculture: Contributions and challenges. Surv. Geophys. 40 (3), 515–551. https://doi.org/10.1007/s10712-018-9492-0.

Heim, A., Wehrli, L., Eugster, W., Schmidt, M.W.I., 2009. Effects of sampling design on the probability to detect soil carbon stock changes at the Swiss CarboEurope site Lägeren. Geoderma 149 (3–4), 347–354. https://doi.org/10.1016/j.geoderma.2008.12.018.

Heung, B., Hodúl, M., Schmidt, M.G., 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma 290, 51–68. https://doi.org/10.1016/j.geoderma.2016.12.001.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309. https://doi.org/10.1016/0034-4257(88)90106-X.

ISO 11464. (2006). Soil Quality — Pretreatment of Samples for Physico-chemical Analysis International Organization for Standardization, Genève, Switzerland (2006).

IUSS Working Group WRB, 2014. World reference base for soil resources. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106 (2014). https://doi.org/10.1017/S0014479706394902.

James, D.W., Wells, K.L., 1990. Soil sample collection and handling: Technique based on source and degree of field variability. Soil Testing and Plant Analysis 3, 25–44. https://doi.org/10.2136/sssabookser3.3ed.c3.

Ji, L., Peters, A.J., 2007. Performance evaluation of spectral vegetation indices using a statistical sensitivity function. Remote Sens. Environ. 106 (1), 59–65. https://doi.org/10.1016/j.rse.2006.07.010.

Jin, X., Song, K., Du, J., Liu, H., Wen, Z., 2017. Comparison of different satellite bands and vegetation indices for estimation of soil organic matter based on simulated spectral configuration. Agric. For. Meteorol. 244-245, 57–71. https://doi.org/10.1016/j.agrformet.2017.05.018.

Karmakar, R., Das, I., Dutta, D., Rakshit, A., 2016. Potential effects of climate change on soil properties: a review. Sci. Int. 4 (2), 51–73.

Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., Robbez-Masson, J.M., 2008. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote Sens. Environ. 112 (3), 825–835. https://doi.org/10.1016/j.rse.2007.06.014.

Liao, Q., Zhang, X., Li, Z., Pan, G., Smith, P., Jin, Y., Wu, X., 2009. Increase in soil organic carbon stock over the last two decades in China's Jiangsu Province. Glob. Change Biol. 15 (4), 861–875. https://doi.org/10.1111/j.1365-2486.2008.01792.x.

Liu, S., An, N., Yang, J., Dong, S., Wang, C., Yin, Y., 2015. Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China. Catena 133, 137–144. https://doi.org/10.1016/j.catena.2015.05.010.

Liu, Y.i., Liu, Y., Chen, Y., Zhang, Y., Shi, T., Wang, J., Hong, Y., Fei, T., Zhang, Y., 2019. The influence of spectral pretreatment on the selection of representative calibration samples for soil organic matter estimation using Vis-NIR reflectance spectroscopy. Remote Sensing 11 (4), 450. https://doi.org/10.3390/rs11040450.

Madeira, J., Bedidi, A., Cervelle, B., Pouget, M., Flay, N., 1997. Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: the application of a Thematic Mapper (TM) image for soil-mapping in Brasilia, Brazil. Int. J. Remote Sens. 18 (13), 2835–2852. https://doi.org/10.1080/014311697217369.

Mallarino, A.P., Wittry, D.J., 2004. Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter. Precis. Agric. 5 (2), 131–144. https://doi.org/10.1023/B:PRAG.0000022358.24102.1b.

Mevik, Wehrens, 2007. The pls package: principal component and partial least squares regression in R. J. Stat. Softw. 18 (2) https://doi.org/10.18637/jss.v018.i02.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C., 2014. e1071: Misc functions of the Department of Statistics (e1071). TU Wien. R package version 1 (3).

Miller, B.A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K., Wehrhan, M., Sommer, M., 2016. Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. Soil Tillage Res. 156, 194–208. https://doi.org/10.1016/j.still.2015.07.004.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., Van Meirvenne, M., 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Sci. Soc. Am. J. 73 (2), 614–621. https://doi.org/10.2136/sssaj2007.0410.

Mulla, D.J., McBratney, A.B., 2002. Soil spatial variability. Soil physics companion 343–373.

Muñoz, J.D., Kravchenko, A., 2011. Soil carbon mapping using on-the-go near infrared spectroscopy, topography and aerial photographs. Geoderma 166 (1), 102–110. https://doi.org/10.1016/j.geoderma.2011.07.017.

Nanni, M.R., Povh, F.P., Demattê, J.A.M., Oliveira, R.B.d., Chicati, M.L., Cezar, E., 2011. Optimum size in grid soil sampling for variable rate application in site-specific management. Scientia Agricola 68 (3), 386–392.

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., & Greve, M. H., 2015. Modeling soil organic carbon at regional scale by combining multispectral images with laboratory spectra. PloS one, 10(11), e0142295.

Pouget, M., Madeira, J., Le Floch, E., & Kamal, S., 1990. Caracteristiques spectrales des surfaces sableuses de la region cotiere nord-ouest de l'Egypte: application aux donnees satellitaires SPOT.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. 48 (2), 119–126. https://doi.org/10.1016/0034-4257(94)90134-1.

R Development Core Team, 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Richardson, A.J., Wiegand, C.L., 1977. Distinguishing vegetation from soil background information. Photogramm. Eng. Remote Sens. 43 (12), 1541–1552.

Roels, J.M., Jonker, P.J., 1983. Probability sampling techniques for estimating soil erosion. Soil Sci. Soc. Am. J. 47 (6), 1224–1228. https://doi.org/10.2136/sssaj1983.03615995004700060032x.

Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: A new paradigm for agriculture. Agric. Syst. 148, 71–74. https://doi.org/10.1016/j.agsy.2016.07.001.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the Great Plains with ERTS. NASA special publication 351 (1974), 309.

Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., Schüler, G., 2010. Retrieval of chlorophyll and nitrogen in Norway spruce (Picea abies L. Karst.) using imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 12 (1), 17–26. https://doi.org/10.1016/j.jag.2009.08.006.

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T., 2010. A method to generate soilscapes from soil maps. J. Plant Nutr. Soil Sci. 173 (2), 163–172. https://doi.org/10.1002/jpln.200800208.

Shadish, W.R., 2002. Revisiting field experimentation: field notes for the future. Psychol. Methods 7 (1), 3–18. https://doi.org/10.1037/1082-989X.7.1.3.

Shi, T., Wang, J., Chen, Y., Wu, G., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. Int. J. Appl. Earth Obs. Geoinf. 52, 95–103. https://doi.org/10.1016/j.jag.2016.06.002.

Signal Developers, (2013). Signal: signal processing URL: http://r-forge.r-project.org/projects/signal (2013).

Stamper, D.J., Agouridis, C.T., Edwards, D.R., Purschwitz, M.A., 2014. Effect of soil sampling density and landscape features on soil test phosphorus. Appl. Eng. Agric. 30 (5), 773–781. https://doi.org/10.13031/aea.30.10328.

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agric. Ecosyst. Environ. 94 (1), 31–47. https://doi.org/10.1016/S0167-8809(02)00013-0.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. In: Advances in agronomy, Vol. 107. Academic Press, pp. 163–215.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144 (1–2), 395–404. https://doi.org/10.1016/j.geoderma.2007.12.009.

Tsui, C.-C., Tsai, C.-C., Chen, Z.-S., 2013. Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan. Geoderma 209-210, 119–127. https://doi.org/10.1016/j.geoderma.2013.06.013.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.

Vapnik, V., 2000. The Nature of Statistical Learning Theory, second ed. Springer-Verlag, New York.

Vasat, R., Boruvka, L., Jaksik, O., 2012. Number of sampling points influences the parameters of soil properties spatial distribution and kriged maps. In: Digital Soil Assessments and Beyond. CRC Press, London, pp. 251–256.

Vašát, R., Kodešová, R., Klement, A., Borůvka, L., 2017. Simple but efficient signal preprocessing in soil organic carbon spectroscopic estimation. Geoderma 298, 46–53. https://doi.org/10.1016/j.geoderma.2017.03.012.

Wang, H., Zhang, X., Wu, W., Liu, H., 2021. Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. Remote Sensing 13 (7), 1229. https://doi.org/10.3390/rs13071229.

Wilding, L.G., 1985. Spatial Variability: Its Documentation, Accommodation and Implication to Soil Surveys. D.R. Nielsen, J. Bouma (Eds.), Soil Spatial Variability, Proceedings of a Workshop of ISSS and the SSSA, Les Vegas USA. Nov. 30–Dec. 1, 1984 (1985). Wageningen, The Netherlands.

Wollenhaupt, N.C., Wolkowski, R.P., 1994. Grid soil sampling. Better Crops with Plant Food 78 (4), 6–9.

Worsham, L., Markewitz, D., Nibbelink, N.P., West, L.T., 2012. A comparison of three field sampling methods to estimate soil carbon content. Forest Science 58 (5), 513–522. https://doi.org/10.5849/forsci.11-084.

Yang, T.R., Kershaw Jr., J.A., Weiskittel, A.R., Lam, T.Y., McGarrigle, E., 2019. Influence of sample selection method and estimation technique on sample size requirements for wall-to-wall estimation of volume using airborne LiDAR. Forestry: Int. J. Forest Res. 92 (3), 311–323. https://doi.org/10.1093/forestry/cpz014.

Zhao, Y., Xu, X., Tian, K., Huang, B., Hai, N., 2016. Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. Environ. Earth Sci. 75 (1), 4. https://doi.org/10.1007/s12665-015-4895-4.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66 (3), 535–547. https://doi.org/10.1111/ejss.12244.