https://doi.org/10.17221/97/2022-SWR

Prediction of the soil organic carbon in the LUCAS soil database based on spectral clustering

Baoyang Liu, Baofeng Guo, Renxiong Zhuo, Fan Dai, Haoyu Chi*

School of Automation, Hangzhou Dianzi University, Hangzhou, P.R. China

**Corresponding author: gbf@hdu.edu.cn*

Citation: Liu B.Y., Guo B.F., Zhuo R.X., Dai F., Chi H.Y. (2023): Prediction of the soil organic carbon in the LUCAS soil database based on spectral clustering. Soil & Water Res., 18: 43–54.

Abstract: The estimation of the level of the soil organic carbon (SOC) content plays an important role in assessing the soil health state. Visible and Near Infrared Diffuse Reflectance Spectroscopy (Vis-NIR DRS) is a fast and cheap tool for measuring the SOC. However, when this technology is applied on a larger area, the soil prediction accuracy decreases due to the heterogeneity of the samples. In this paper, we first investigate the global model performance in the LUCAS EU-wide topsoil database. Then, different clustering strategies were tested, including the k-means clustering based on the principal component analysis (PCA) and hierarchical clustering, combined with the partial least squares regression (PLSR) models, and a clustering based on a local PLSR approach. The best validation results were obtained for the local PLSR approach with R^2 = 0.75, root mean squared error of prediction (RMSEP) = 13.38 g/kg and ratio of performance to interquartile range (RPIQ) = 2.846, but the algorithm running time was 30.05 s. Similar results were obtained for the *k*-means clustering method with R^2 = 0.75, RMSEP = 14.61 g/kg and RPIQ = 2.844, at only 4.52 s. This study demonstrates that the PLSR approach based on k-means clustering is able to achieve similar prediction accuracy as the local PLSR approach, while significantly improving the algorithm speed. This provides the theoretical basis for adapting the spectral soil model to the needs of real-time SOC quantification.

Keywords: cluster analysis; regression analysis; retrieve; soil properties, Vis-NIR spectroscopy

Fundamental ecological services provided by soils include carbon sequestration, energy security, and food security (Kibblewhite et al. 2012). When it comes to carbon sequestration, soils often provide greater carbon storage capabilities to partially offset fuel emissions and thus reduce the danger of global warming (Conant et al. 2010). Soil organic carbon is essential for sustaining soil quality and food production, and its decline is one of the main threats of soil degradation (Lal 2004). Determining the soil organic carbon (SOC) content is a key step in assessing soil condition (Sanchez et al. 2009). As a result, there is an increasing demand to monitor the SOC content and other soil parameters (Lal 2004). Unfortunately, the costly and time-consuming nature of conventional soil sampling and analysis restricts the monitoring of soil properties on a large scale (Conant et al. 2010; Araújo et al. 2014).

Visible (Vis, 400–700 nm) and Near-Infrared (NIR, 700–2 500 nm) diffuse reflectance spectroscopy (DRS) was first applied to soil analyses in the 1980s and was shown to hold potential for predicting the soil SOC content and other properties (Dalal & Henry 1986; Viscarra Rossel et al. 2006). Therefore, diffuse reflectance spectroscopy provides a good quantitative alternative to soil properties (Islam et al. 2003). Visible near-infrared spectroscopy has been widely applied to detect heavy metal soil contamination, soil salinisation and water pollution due to its advantages, such as cost-efficiency, ease of opera-

Supported by the Fund of the National Natural Science Foundation of China, Project No. 42167039.

https://doi.org/10.17221/97/2022-SWR

tion, rapidity, minimal sample preparation and the development of chemometrics (Davies 2005). Soil spectroscopy is based on the assumption that absorption features in the visible range (400–700 nm) may be brought on by electron transitions (Ben-Dor et al. 1999). Another assumption is that the concentration of a certain soil attribute is linearly proportional to the combination of the absorption features within the spectrum (Ben-Dor et al. 1999; Bellon-Maurel & McBratney 2011). These absorption features are attributed to overtones and combination bands of fundamental vibrations of some of the molecules' functional groups, such as the hydroxyl groups (OH). Since the overtone and combination bands of each functional group are located at specific wavelengths of the spectrum, various materials can be recognised (Ben-Dor et al. 1999).

The prediction of soil properties requires the creation of a spectral library that links the spectra with the soil physical and chemistry data. Such a library should be designed to represent the variability of soil properties for the soil type of interest. Different mathematical modelling approaches are then used to infer soil properties (Shepherd & Walsh 2002; Viscarra Rossel & Behrens 2010; Stevens et al. 2013). In large and complex datasets, the relationship between soil properties and spectral data is highly non-linear and spatially dependent (Stenberg et al. 2010; Stevens et al. 2013). This usually leads to a decrease in the prediction accuracy, as well as increasing variances in the soil properties that finally produce larger prediction errors (Stevens et al. 2013; Nocita et al. 2014; Ward et al. 2019).

Nevertheless, there is local stability in the spectral variance associated with soil properties (Stevens et al. 2013). Thus, one possible approach for predicting soil properties in a large-scale database is the use of local regression (Nocita et al. 2014). Local regression has been widely applied to estimate soil properties in large-scale databases and has achieved promising results. In contrast to other local algorithms, Ramirez-Lopez et al. (2013) proposed a new algorithm for retrieving a set of nearest neighbours using an optimised principal component distance. This procedure selects the optimal number of principal components by considering the soil composition of the sample (Ramirez-Lopez et al. 2013). Nocita et al. (2014) first divided their cropland database into mineral and organic soils and obtained the best results for mineral soils using the local partial least squares regression (PLSR), which utilised a fixed number of nearest neighbours and set the PLS distance as a spectral distance measure. In addition, they used the sand content and geographic data as auxiliary distance measures for the further improvement of the prediction accuracy (Nocita et al. 2014). However, their method required prior knowledge not only from the spectra, but also from the additional SOC content, sand content and chemical data. Ward et al. (2019) used the exhaustive method to select the number of clusters that achieved the best PLSR model validation results as the input of the *k*-means algorithm. However, the process was time-consuming and lacks any theoretical basis (Ward et al. 2019).

The main points of this study include the following: (*i*) improving the performance of the SOC prediction model by converting the highly skewed SOC content to an approximately normal distribution through natural logarithms; (*ii*) using hierarchical clustering to determine the appropriate number of clusters, thus reducing the modelling time; and (*iii*) using only spectral data without using any geochemical soil information in building the SOC prediction model. The objectives of this paper are to (*i*) investigate whether spectral clustering has the potential to group large soil spectral libraries and thus improve the prediction accuracy compared to non-clustering-based models, and (*ii*) whether spectral clustering can achieve prediction accuracy similar to that of local PLSR and reduce the modelling time.

MATERIAL AND METHODS

LUCAS soil database. This study is based on the pan-European Land Use/Land Cover Area Frame Survey (LUCAS) 2015 topsoil database which is managed by EUROSTAT together with the European Commission's Directorates-General for Environment and the Joint Research Centre at Ispra, Italy (Tóth et al. 2013; Orgiazzi et al. 2017; Jones et al. 2020; https://esdac.jrc.ec.europa.eu). In 2015, the LUCAS survey was carried out in all the EU-28 Member States (MS) and included 21 859 top-soil samples (0–20 cm) collected on different land use types. In the countries sampled in 2009 and 2012, 90% of the locations were maintained while the remaining 10% of the points were substituted by new sampling locations, including points above 1 000 m in elevation, which were out of the scope of the LUCAS 2009 and LUCAS 2012 surveys. To ensure the comparability of the data, all the survey teams followed a single soil sampling protocol. The sampling protocol consisted of collect-

https://doi.org/10.17221/97/2022-SWR

ing a composite of five subsamples, the first of which was collected at a selected location, while the other four subsamples were collected at a distance of 2 m along the main direction. All the soil samples were air-dried in a drying chamber at a temperature of 40 °C for an average of 3 days. The soil samples were crushed and sieved, and fractions smaller than 2 mm were retained for further analysis. The following properties were then analysed: the particle size distribution, pH of H_2O, pH of $CaCl_2$, organic carbon, carbonate, nitrogen, phosphorus, and potassium, cationic exchange capacity (CEC), and visible-NIR diffuse reflectance.

After heating the soil to 900 °C, the total carbon content was measured using a VarioMax CN Analyser (Elemental Analysis, Germany) and the SOC was obtained by subtracting the carbonate content (measured according to ISO 10693:1995) from the total carbon amount. The Vis-NIR absorbance of the soil was measured using a FOSS XDS Rapid Content Analyser (FOSS NIR Systems Inc., Denmark), which operates at the wavelength range of 400–2 500 nm with a spectral resolution of 2 nm and a spectral data interval of 0.5 nm. Each sample was placed on a 140 × 40 × 50 mm cuvette and scanned twice in both directions. The average spectrum of the two replicates was calculated and any samples with absorbance repeat standard deviations greater than 1% were removed.

In order to reduce the redundancy of the spectral data and the number of model calculations, we resampled the spectral data interval to 2 nm. Similarly, we selected 4 239 soil samples as the research objects, which were collected for the first time in the LUCAS 2015 database.

Database pre-processing. In several studies, the 1st derivative led to the best modelling results (e.g., Stevens et al. 2013; Nocita et al. 2014). Thus, several pre-processing techniques were applied: 1st derivative, Savitzky-Golay smoothing with 2nd order polynomial and a window size of 20 data points which corresponds to 40 nm (Savitzky & Golay 1964).

The distribution of the SOC content in the subset studied in this paper is highly skewed (skewness = 4.072), so we transformed it to approximately normally distributed values using the natural logarithm (new skewness = 0.61). The dataset was then divided into subsets for calibration (70%) and validation (30%) using the Kennard-Stone algorithm (Kennard & Stone 1969), which selects samples based on a distance metric to produce a typical subset. The clustering and model calibration are only based on the calibration subset, and the validation subset is solely used to assess the model performance.

Methodological overview. In this study, we compared three different modelling approaches (Figure 1): (*i*) a PLSR approach was used based on the complete database without clustering, called the reference model (Figure 1A). (*ii*) the *k*-means algorithm based on principal component analysis (PCA) and hierarchical clustering was used (Figure 1B), then SOC predictions were performed on each spectral cluster using PLSR; (*iii*) a local PLSR method was used, where for each validation sample, a set of the most similar calibration samples was selected based on the distance metric for calibration of the independent PLSR model (Figure 1C). For comparability reasons, all the models are calibrated and validated on the exact same subset of LUCAS.

Reference model without clustering using PLSR. As an initial stage, the PLSR model for the SOC prediction was built based on the spectral data of the newly collected samples in the LUCAS 2015 database. PLSR integrates the advantages of principal component regression, a typical correlation analysis and multilinear linear regression, which has been used in many disciplines such as social sciences, bioinformatics and economics (Wold et al. 2001). PLSR uses a few factors and orthogonal factors, called latent variables (LVs) as new predictor variables of the response Y. The number of LVs is unknown and has a significant effect on the prediction results of the model. In this paper, a combination of two commonly used methods was chosen to achieve good model accuracy without over-fitting. (*i*) We used 10-fold cross-validation to estimate the root mean square error (RMSE) for different numbers of LVs

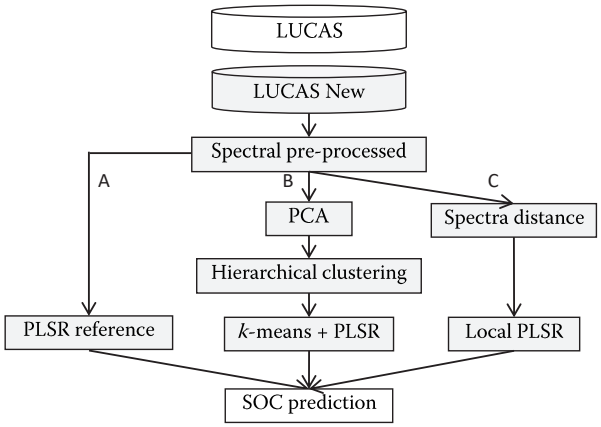

Figure 1. Overview of the general processing structure
PLSR – partial least squares regression; PCA – principal component analysis; SOC – soil organic carbon

https://doi.org/10.17221/97/2022-SWR

and chose the smallest number of LVs within one standard deviation of the minimum RMSE (Stevens et al. 2013). (*ii*) We used the adjusted coefficient of determination (*adj*. R^2, see Equation (1)) which takes the number of components used in a model into account.

$$adj.\ R^2 = 1 - (1 - R^2) \times (n - 1)/(n - k - 1) \quad (1)$$

where:
n – the number of samples;
k – the number of LVs.

Hierarchical clustering and k-means clustering. As shown in Figure 1B, we first divided the spectral data into several different clusters using the clustering algorithms, and then built the PLSR model for each spectral cluster. In order to eliminate noise, reduce multicollinearity and improve the computational speed, a principal component analysis based on spectral variance was applied to the pre-processed spectral data before the hierarchical clustering and k-means clustering in this study. The PCA method was used only for the clustering process. The k-means algorithm demands the number of clusters as an input, and here we used the hierarchical clustering algorithm to obtain the appropriate number of clusters instead of the exhaustive method commonly adopted in previous studies.

The hierarchical agglomerative clustering algorithm is a hierarchical approach that gradually merges existing clusters until the desired number of clusters is reached (Wishart 1969). We used Ward association as the association criterion and Euclidean distance as the distance metric. Finally, we determined the number of clusters based on the dendrogram generated by the hierarchical clustering algorithm, which is then used as the input to the *k*-means algorithm.

k-means starts with randomly selected initial cluster centres and assigns the closest samples to these centres. Based on these clusters, it calculates new cluster centres and reassigns all the samples. This process continues until the set number of clusters is reached. The following steps are applied to the set of clusters: for each cluster, the individual PLSR models are calibrated in the calibration dataset of the pre-processed spectra. As the clustering process was based exclusively on the calibration subset, each validation sample had to be assigned to one of those clusters. To validate this clustering modelling approach, a PLSR model was built based on the clusters in the calibration set to which the validation sample belonged, in order to predict the SOC content of this validation sample.

Local partial least square regression. Locally weighted PLSR models are the memory-based learning approaches and create a specific calibration set for each sample to be predicted, which outperform machine learning algorithms such as artificial neural networks and decision trees (Ramirez-Lopez et al. 2013). The local PLSR selects the most similar set of spectral samples for each validation sample in the calibration set based on the similarity metric, and builds an independent PLSR model for each validation sample based on this set of spectral samples (Ward et al. 2020). As shown in Figure 1C, a basic local PLSR approach can be described with the following pseudo-code:

1. Given a set of n reference samples

$$(Xr, Yr) = \{xr_i, yr_i\}^n{}_{i=1}$$

and a set of m samples to predict

$$(Xp, Yp) = \{xp_i, yp_i\}^m{}_{i=1}$$

where:
Xr, Yr – spectral data matrix and SOC content in the calibration set, respectively;
Xp, Yp – spectral data matrix and SOC content in the validation set, respectively.

2. for each sample to predict p_i i = 1, 2, …, m do
3. compute d_i, the distance vector between Xr and xp_i
4. find the most similar samples in Xr as the k ones minimising d_i, i.e., the k-nearest neighbours
5. fit a PLSR model with the k nearest neighbours
6. choose the optimal model parameters for the prediction of p_i, e.g., appropriate number of LVs for a PLSR model
7. predict the SOC content of sample p_i and compute the square error $(yp_i - \hat{y}p_i)^2$
8. end
9. evaluate the model performance using the root mean square error (RMSE)

Model assessments. To assess the model accuracy, the ln-transformed SOC values (measured and predicted) were used for dimensionless measurements, whereas for measurements with units (g/kg), the original SOC values and back-transformed predicted values were used. The coefficient of determina-

https://doi.org/10.17221/97/2022-SWR

tion (R^2) (Equation (2)), the root mean squared error of prediction (RMSEP, Equation (3)), the relative RMSEP (rRMSEP) (Equation (4)), the ratio of performance to deviation (RPD) (Equation (5)) and the ratio of performance to interquartile range (RPIQ, Equation (6)) were used to evaluate and compare the model performance:

$$R^2 = 1 - \sum_{i=1}^{n} \left(yp_i - yo_i \right)^2 / \sum_{i=1}^{n} \left(yo_i - \bar{yo}_i \right)^2 \quad (2)$$

where:
yo_i – the observed value of sample i;
yp_i –the predicted value of sample i;
$\bar{y}o$ – the mean of the observed SOC value.

$$\text{RMSEP} = \left(\sum_{i=1}^{n} \left(yp_i - yo_i \right)^2 / n \right)^{1/2} \quad (3)$$

where:
n – the number of samples.

$$\text{rRMSEP} = 100 \times RMSEP/\bar{y}o \quad (4)$$

$$\text{RPD} = \text{SD}(yo)/RMSEP \quad (5)$$

where:
SD – the standard deviation.

$$\text{RPIQ} = IQ(yo)/RMSEP \quad (6)$$

where:
IQ – the interquartile range.

The rRMSEP, RPD, and RPIQ are ways of normalising the RMSE's of the prediction to compare calibration models and datasets where the measured variables have different ranges or variances (Bellon-Maurel & McBratney 2011).

In this paper, we applied Matlab 2018a to implement the PLSR model, cluster algorithms and local PLSR model. Excel 2016 was applied for the spectral data resampling and pre-processing.

RESULTS AND DISCUSSION

LUCAS database and pre-processing. We used the samples that were first sampled in the LUCAS 2015 database as the study object, so the number of samples was reduced to 4 239 values. Within the selected LUCAS 2015 subset, the SOC content ranges between 0.4 and 517.2 g/kg, with a mean value of 46.42 g/kg and a standard deviation of 59.34 g/kg. The percentage of clay ranges between 0 and 62% and is 19.15% on average. The $CaCO_3$ content varies between 0 and 898 g/kg with a mean of 85.34 g per kg and a standard deviation of 167.22 g/kg. The SOC content and $CaCO_3$ content both showed high standard deviations, indicating considerable variation among the soil samples. This variation may have been caused by the fact that the soil samples were gathered from various land cover and use types. As shown in Table 1, the standard deviation of the SOC in the LUCAS 2015 database was slightly higher than the standard deviation of the SOC for the newly sampled data, but the other statistical results were similar. This indicates that the newly sampled soil samples used in this paper are representative.

The spectra show a large variation in the absorbance due to the influence of the SOC content and mineralogical composition (Figure 2). In the visible region (400–700 nm), the difference in the reflectance for higher SOC content classes is clearer than in the NIR region (Baumgardner et al. 1986). While this was not true for longer wavelengths, the darkest spectrum, which corresponded to the sample with the greatest SOC value, displayed the maximum absorbance between 500 and 750 nm. This was in line with the findings of Stenberg et al. (2010), who discovered that organic soils had decreased absorbance in the NIR wavelength region. In the NIR region, clear absorption features were recorded near 1 440 and 1 915 nm, which were attributed to the OH-soil hygroscopic moisture in the clay minerals (Stenberg et al. 2010). At 2 100 nm, the absorption is determined by the nitrogen content, and due to the relationship between

Table 1. Descriptive statistics for the soil organic carbon (SOC) of the LUCAS 2015 database and the newly sampled data

Dataset	N	No. of LC	No. of LU	SOC (g/kg)							
				min	max	Q25	Q75	mean	SD	med	skew
LUCAS 2015	21 859	68	23	0.1	560.2	12.5	38.6	42.2	76.6	20.4	4.3
New samples	4 239	60	19	0.4	517.2	15.2	54.1	46.4	59.3	29.2	4.1

New samples – a new soil sample added to the LUCAS 2015 database; N – No. of samples; LC – land cover class type; LU – land use type; SD – standard deviation; med – median; skew – skewness; Q25 – the first quartile; Q75 – the third quartile

https://doi.org/10.17221/97/2022-SWR

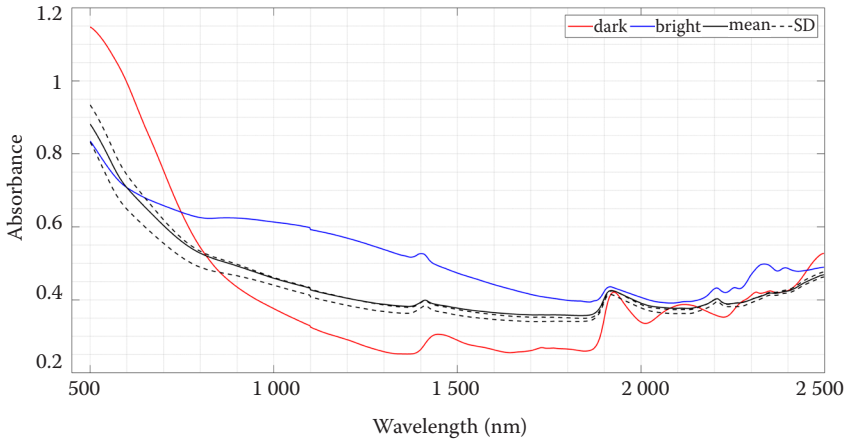

Figure 2. Spectral variability in the LUCAS 2015 subset showing the mean and standard deviation, as well as the darkest and the brightest spectrum
SD – standard deviation

nitrogen and organic matter, we observed an increase in the absorption depth with the SOC content. The absorption feature at 2 207 nm is usually caused by Al-OH, and this absorption feature associated with clay mineralogy is critical for soils with a low SOC content. In addition, there is an organic matter-related C-H characteristic peak near 2 300 nm. In spite of the large-scale soil diversity and the multiple interactions among the soil absorbance and other soil properties (e.g., texture, structure, and mineralogy), the spectral curves corresponding to different SOC contents showed the same trends as those observed in previous studies (Viscarra Rossel et al. 2006).

The PLSR model, which fitted the original values of the SOC, resulted in R^2 of 0.61, RPD of 1.61, and RMSEP of 17.60 g/kg in the validation subset. The PLSR reference model, which fitted the natural log values of the SOC, led to R^2 of 0.73, RPD of 1.93, and RMSEP of 14.99 g/kg. This demonstrated that standardising the SOC prior to modelling significantly improves the model performance.

Local PLSR. Figure 3 is an illustration of the local PLSR approach showing an example for the validation sample. Sample 11104 contains SOC (63.3 g/kg), clay (19%) and $CaCO_3$ (26 g/kg). The nearest neighbours of this sample contains SOC, clay and $CaCO_3$ with mean values of 43.39 g/kg, 14.49% and 41.36 g/kg and standard deviations of 35.34 g/kg, 9.07% and 88.19 g/kg, respectively. Although no significant differences can be observed in the pre-processed absorption spectra of sample 11104 and its nearest neighbours, there were significant differences in their soil properties. It supported the observation by Nocita et al. (2014) that in sizable datasets, even spectrally identical neighbours can have wildly dissimilar soil characteristics.

In this paper, Euclidean distance was adopted as the distance metric and a fixed number of calibration samples were used to calibrate the model for each validation sample. The number of nearest neighbours strongly influenced the SOC prediction accuracy. To determine the best fixed number of calibration samples, we selected 30% of the calibration dataset as the test validation set, using the Kennard-Stone algorithm (Kennard & Stone 1969), and iteratively tested different numbers. The performance of the

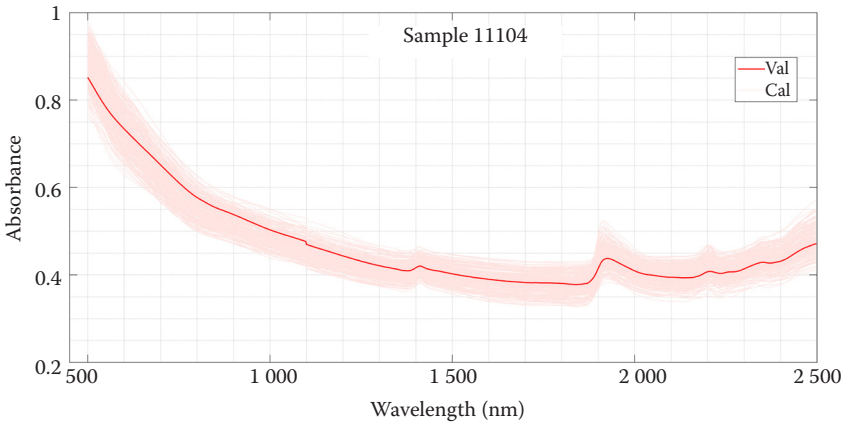

Figure 3. An example of the local partial least squares regression (PLSR) approach showing the absorbance spectra of the validation samples (Val) and calibration samples (Cal)

https://doi.org/10.17221/97/2022-SWR

Table 2. The test validation result for the local partial least squares regression (PLSR) model with different number of nearest neighbours

No. of neighbours	Model performance					Model data		
	R^2	RMSEP (g/kg)	rRMSEP	RPD	RPIQ	LV	NvalT	NcalT
50	0.7221	416.4828	18.8895	1.8971	2.6261	17	890	2 077
100	0.7760	**221.6711**	16.9590	2.1131	2.9250	17	890	2 077
250	**0.8038**	228.7834	**15.8737**	**2.2575**	**3.1250**	17	890	2 077
500	0.7840	255.9636	16.6545	2.1517	2.9785	17	890	2 077
750	0.7654	346.1274	17.3556	2.0648	2.8582	17	890	2 077

RMSEP – root mean squared error of prediction; rRMSEP – relative RMSEP; RPD – ratio of performance to deviation; RPIQ – ratio of performance to interquartile range; LV – latent variables; NvalT – No. of test validation samples; NcalT – No. of test calibration samples; bold – optimal results

local model with different numbers of nearest neighbours in the test validation set is shown in Table 2. From 50 to 250 nearest neighbours, an improvement in the model performance was observed except for the RMSEP. From 250 to 750 nearest neighbours, a slow decrease in the model performance was observed. At last, we chose the number of neighbours which led to the best results within the test validation set. Applying the local PLSR method with a distance metric of Euclidean distance and 250 nearest neighbours to the validation dataset, we were able to calibrate good prediction models with R^2 = 0.7557, RMSEP = 13.3817 g/kg, rRMSEP = 10.9486, RPD = 2.0231 and RPIQ = 2.8466. Since the local PLSR algorithm built a unique calibration model for each validation sample, the algorithm ran a total of 30.05 s.

Clustering approaches and PLSR. The results of the hierarchical clustering and k-means clustering are shown in Figure 4. As shown on the left in Figure 4, the *x*-axis in the tree diagram represents the number of sub-nodes in each node, and the splitting stops when the number of sub-nodes is less than 30. This is because the samples included when the number of sub-nodes is less than 30 are already very similar. The *y*-axis in the tree diagram indicates the distance between the different clusters. Based on the dendrogram obtained by the hierarchical clustering, we set the number of clusters to four as the input parameter of the *k*-means algorithm. As shown on the right in Figure 4, a PCA was applied to the pre-processed spectral data prior to the hierarchical clustering and *k*-means clustering, resulting in a first principal component (PCA1) contribution of 86.51% and a second principal component (PCA2) contribution of 10.63%, with a total contribution of 97.14%. The *k*-means clustering of the soil spectral data was performed mainly based on the magnitude of PCA1, i.e., the intensity of the absorbance. Secondly, it was based

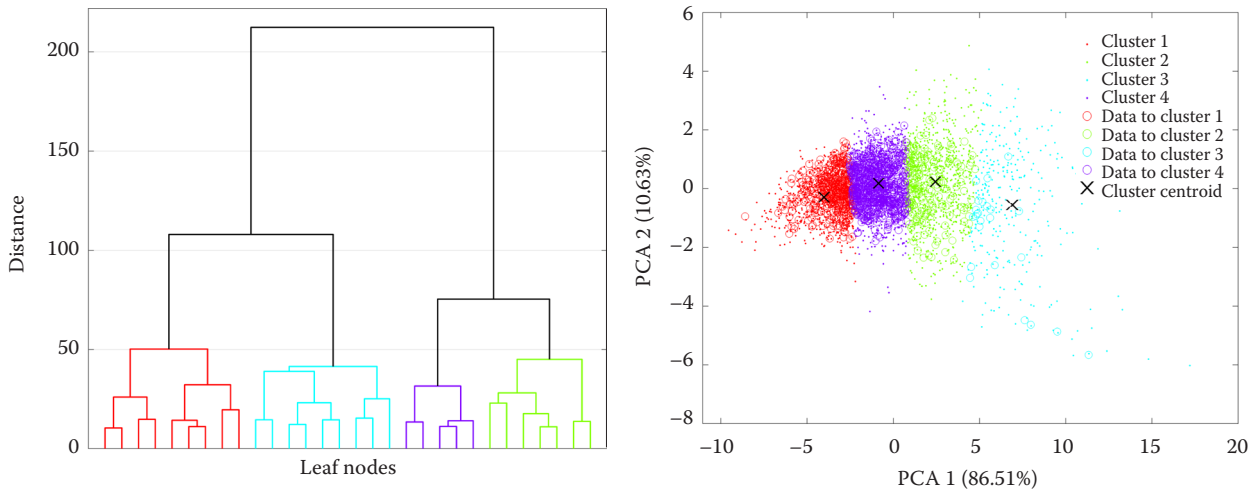

Figure 4. The results of hierarchical clustering (left) and *k*-means clustering (right)
PCA – principal component analysis

https://doi.org/10.17221/97/2022-SWR

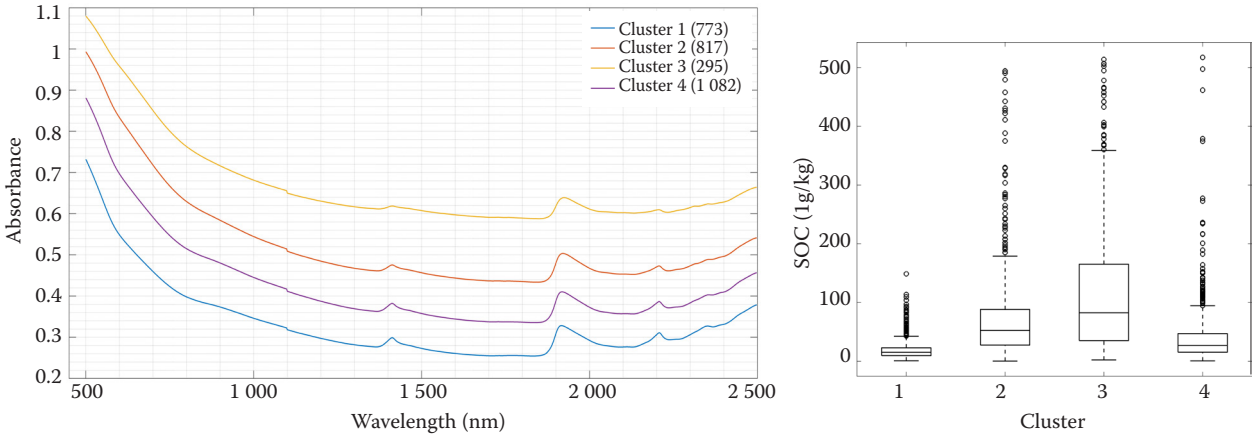

Figure 5. Mean spectra of the clusters showing the principal component analysis (PCA) and hierarchical cluster based approaches; numbers in brackets are the number of samples within each cluster including the calibration samples (left); boxplots showing the soil organic carbon (SOC) distribution within the clusters (right)

on the shape characteristics of the spectral curve, which was reflected in the differences in PCA2 of the four clustering centres. The PCA2 of each clustering centre (–0.5535, –0.2920, 0.1819 and 0.2351) was basically around 0, so there was little difference in the shape of the spectral curves.

Figure 5 shows the mean reflectance and SOC range for each spectral cluster in the calibration subset. Since PCA-based clustering focuses on spectral variance, it can be seen in Figure 5 that there are some differences in the spectra and SOC contents between the clusters. The mean spectra of the clusters show differences mainly in the absorbance with the smallest, cluster 3, showing the darkest mean spectrum and, cluster 1, showing the brightest, related to the lowest SOC values within this cluster. This indicated the absorbance increased with an increase in the SOC content.

All the validation results for the k-means clustering approach combined with the hierarchical clustering are shown in Table 3. All the validation results for the different clusters are shown in Figure 6. The coefficient of variation of the SOC content for all the validation samples was 0.78. The coefficients of variation of the SOC in cluster 1, cluster 2, cluster 3, and cluster 4 were 0.64, 0.65, 0.99, and 0.59, respectively. The higher coefficient of variation in cluster 3 may be caused by the small number of samples in cluster 3. The coefficients of variation in all the other clusters were lower than the coefficients of variation in the entire validation set. This may be one of the reasons why the k-means-based modelling approach can improve the prediction accuracy. The *k*-means approach combined with hierarchical clustering resulted in four clusters with calibration sizes ranging from 295 to 1 082 samples. Variable validation results were obtained depending on the spectral clustering, ranging from moderate (R^2 = 0.6117, cluster 1) to very good (R^2 = 0.8025, cluster 3). Except for these two

Table 3. Overall validation results for the *k*-means clustering approaches

	N	R^2	RMSE (g/kg)	rRMSE	RPD	RPIQ	LV	SD	Mean	SOC range
								(g/kg)		
k-means PLSR	1 272	0.7553	14.6126	12.3967	2.0214	2.8442	17.5	28.4	36.0	1.2–334.7
Cluster 1	385	0.6117	**9.2015**	13.8643	1.6049	2.0528	17	13.4	20.7	1.2–81.1
Cluster 2	186	0.7384	19.2741	**7.8888**	1.9419	2.4764	18	40.1	61.4	8.3–334.7
Cluster 3	24	**0.8025**	40.0221	11.6491	**2.2499**	**3.4719**	**15**	68.1	68.5	7.3–325.7
Cluster 4	677	0.6802	13.9354	10.4707	1.7684	2.3107	20	21.5	36.2	4.0–144.9

PLSR – partial least squares regression; RMSE – root mean squared error; rRMSE – relative RMSE; RPD – ratio of performance to deviation; RPIQ – ratio of performance to interquartile range; LV – latent variables; SD – standard deviation; SOC – soil organic carbon; for the row *k*-means PLSR, the R^2, RPD, RPIQ and rRMSE use combined predicted values; for the column LV, the mean values are calculated; bold – optimal results

https://doi.org/10.17221/97/2022-SWR

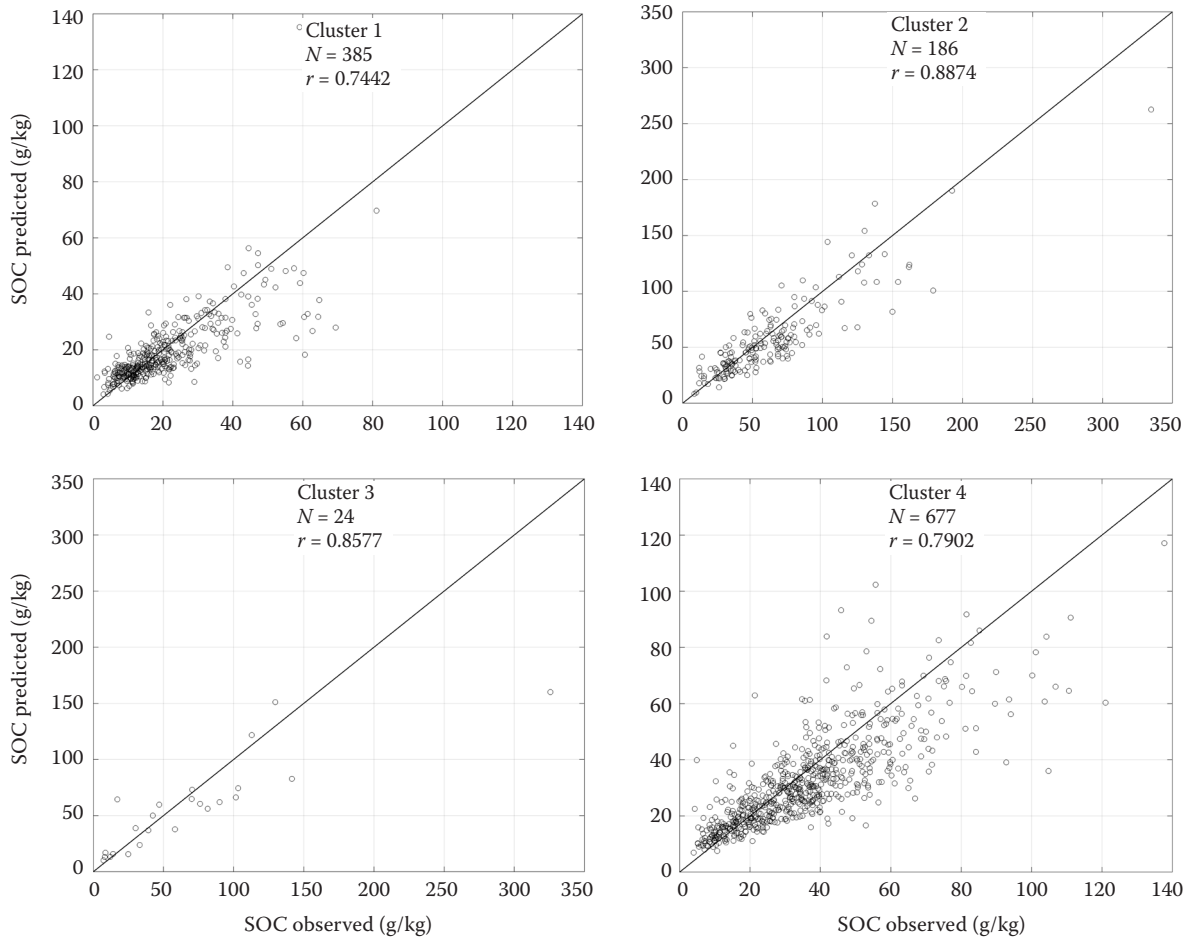

Figure 6. Observed vs. predicted soil organic carbon (SOC) values of the validation samples for the *k*-means clustering approaches; Pearson's correlation coefficient *r* is given

spectral clusters, in general, a good performance is achieved with R^2 of 0.6802 and 0.7384 in the other two spectral clusters. With values over 1.8 for the two models with strong model performance and below 1.8 for the two models with medium performance, the RPD values highlighted this claim. Although cluster 3 had the best modelling performance including the highest R^2, RPD, and RPIQ values (Table 3), cluster 3 had the highest RMSEP of 40.02 g/kg. This cluster has the highest mean SOC value of 68.5 g/kg and the highest standard deviation of 68.1 g/kg. This result is in line with Stenberg et al. (2010), who stated that the prediction error of the spectral model increases with an increase in the standard deviation of the predicted soil properties. Therefore, it is important to consider the distribution of the SOC values when comparing the RMSEP of different study sites and clusters. The RPD, RPIQ, or rRMSEP are more suitable because these three metrics consider different ranges and differences (Ward et al. 2019).

Compared to the reference PLSR model, the *k*-means clustering was able to improve the organics (OM) prediction results because it could handle non-linearities in large heterogeneous datasets. Similar conclusions were reached by Araújo et al. (2014), who compared clustering results with augmented regression trees and support vector machines as the reference model and found that both models performed in the same range. Thus, *k*-means clustering combined with the PLSR model alone seems to improve the performance of the PLSR reference model. Based on the experimental results in this paper, we also validated this idea.

Overall results. The overall results in Table 4 show that the normalising the SOC before calibrating the prediction model could significantly improve the model performance compared to the PLSR model. Nevertheless, so far, only few studies have transformed skewed SOC contents before the spectral predictions (e.g., Viscarra Rossel et al. 2016, Ward et al. 2019). The *k*-means clustering approach could also improve

https://doi.org/10.17221/97/2022-SWR

Table 4. The overall validation results for the reference model and the clustering approaches (No. of validation samples = 1 272)

	Model performance					Model data	
	R^2	RMSEP (g/kg)	rRMSEP	RPD	RPIQ	LV	time (s)
PLSR	0.6179	17.6007	48.8521	1.6179	1.7215	19	**1.68**
PLSR reference	0.7316	14.9927	11.4753	1.9303	2.7159	**17**	1.70
k-means PLSR	0.7553	14.6126	12.3967	2.0214	2.8442	17.5	4.52
Local PLSR	**0.7557**	**13.3817**	**10.9486**	**2.0231**	**2.8466**	**17**	30.05

PLSR – partial least squares regression; RMSEP – root mean squared error of prediction; rRMSEP – relative RMSEP; RPD – ratio of performance to deviation; RPIQ – ratio of performance to interquartile range; LV – latent variables; bold – optimal results

the RMSEP, R^2, RPD and RPIQ compared to the PLSR reference. In the validation subset, the overall best results were achieved by the local PLSR approach. It was able to improve the prediction accuracy visible in all the model parameters, e.g., the RMSEP could be reduced by > 1.6 g/kg compared to the PLSR reference model. Similar results to the local PLSR were obtained by the *k*-means PLSR, but the running time of the *k*-means PLSR algorithm was reduced substantially, from 30.05 s to 4.52 s. The main reason that k-means PLSR significantly improves the speed of the algorithm operation is the relatively small number of clusters (four in this study). However, the local PLSR approach builds a prediction model for

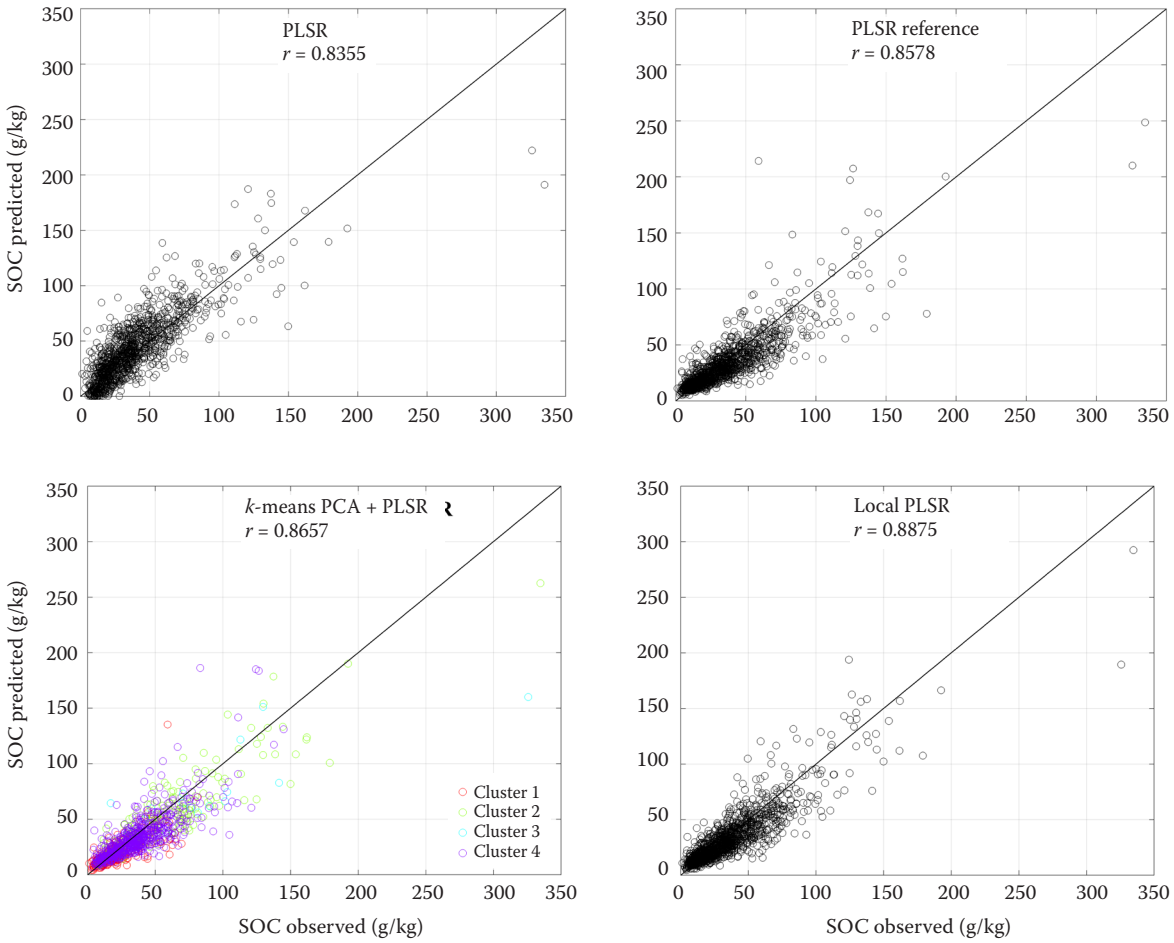

Figure 7. Observed vs. predicted soil organic carbon (SOC) values of the validation samples for the partial least squares regression (PLSR) reference and the clustering approaches

The colour represent the four *k*-means clusters; Pearson's correlation coefficient *r* is given; PCA – principal component analysis

https://doi.org/10.17221/97/2022-SWR

each validation sample, which leads to about 1200 different calibration subsets considered as clusters.

Ramirez-Lopez et al. (2013) used a spectral-based learner to predict the SOC from the regional soil spectral library (R-SSL) of the State of São Paulo and the global soil spectral library (G-SSL). The results showed that the model prediction performance was R^2 of 0.59 and RMSE of 0.25 g/kg in the R-SSL and R^2 of 0.68 and RMSE of 0.8 g/kg in the G-SSL. Nocita et al. (2014) used a local regression model with the PLS distance and sand content as a covariates to achieve a high accuracy soil organic carbon prediction in the LUCAS 2012 database, with R^2 of 0.84, RMSE of 3.6 g per kg, RPD of 2.5, and RPIQ of 2.3. Ward et al. (2019) used the local PLSR method for the SOC prediction based on the spectral data only in the LUCAS 2012 database, with the R^2 of 0.67, RMSE of 5.16 g/kg, RPD of 1.74, and RPIQ of 1.96. Previous studies have demonstrated that the prediction error of spectral models increases with an increasing standard deviation of the predicted soil properties (Stenberg et al. 2010; Nocita et al. 2014). Since the soil samples used in this paper differ from the above studies, model prediction performance metrics other than RMSE were compared. Comparing the results of our study with those of Ramirez-Lopez et al. (2013) and Ward et al. (2019), we found some improvement in the performance of the model in this paper. Although comparing our results with the results of Nocita et al. (2014), we found a slight decrease in the performance of the model in this paper. This is mainly because Nocita et al. (2014) improved the local regression process by including other covariates (geographical and texture information) in the calculation of the distance between samples and using PLS distances. The acquisition of covariate information can consume a lot of human, time, and financial resources, thus failing to reflect the characteristics of the hyperspectral technology. In this paper, only spectral data are considered, thus reducing the input information for modelling and making it more general. Meanwhile, the prediction error of the SOC in this paper is within a reasonable range in a large-scale soil database.

As shown in Figure 7, the underestimation of higher SOC values is a well-known problem in PLSR modelling as shown in the results for the PLSR reference and clustering approach. The reasons for this are the skewed distribution of SOC content leading to under-representation of higher SOC values in the calibration set (Brown et al. 2005) and changes in the relationship between higher SOC values and the spectra due to saturation of the SOC spectral response (Nocita et al. 2014).

CONCLUSION

We tested a *k*-means clustering algorithm that was based on a PCA of the spectra. The number of clusters of the *k*-means algorithm was determined by using hierarchical clustering instead of exhaustive enumeration, thus significantly reducing the algorithm runtime. Compared to the local PLSR approach, the *k*-means clustering method was able to reduce the algorithm runtime from 30.05 s to 4.52 s while achieving similar results. Compared with the local PLSR method, the *k*-means method can save 84.96% of the modelling time for the SOC prediction modelling. Both approaches could improve the results of the PLSR reference model. We noted that the distribution of the SOC content in a large soil spectral library was severely skewed, which had a negative impact on the prediction accuracy. Therefore, it was essential to convert the highly skewed SOC content to an approximate normal distribution before the model calibration.

With this study, we have taken a step forward in adapting the spectral soil model to the need for real-time SOC quantification. The results of this study show that: (*i*) it is possible to improve the SOC prediction by dividing the large soil database into smaller groups compared to the global model; (*ii*) compared with the local PLSR approach, the k-means clustering approach based on a PCA and hierarchical clustering achieved similar results while significantly improving the algorithm running speed.

Acknowledgement: The authors would like to thank the surveyors and the SOIL Action of the Institute for Environment and Sustainable – Joint Research Centre of the European Commission for their great efforts in preparing, managing and maintaining the LUCAS dataset, and for sharing the LUCAS dataset.

REFERENCES

Araújo S., Wetterlind J., Demattê J., Stenberg B. (2014): 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. European Journal of Soil Science, 65: 718–729.

Baumgardner M.F., Silva L., Biehl L.L., Stoner E.R. (1986): Reflectance properties of soils. Advances in Agronomy, 38: 1–44.

https://doi.org/10.17221/97/2022-SWR

Bellon-Maurel V., McBratney A. (2011): Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils critical review and research perspectives. Soil Biology and Biochemistry, 43: 1398–1410.

Ben-Dor E., Irons J.R., Epema G. (1999): Soil reflectance. In: Rencz A.N. (ed.): Remote Sensing for the Earth Science. New York, Wiley: 111–188.

Brown D.J., Bricklemyer R.S., Miller P.R. (2005): Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma, 129: 251–267.

Conant R.T., Ogle S.M., Paul E.A., Paustian K. (2010): Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Frontiers in Ecology and the Environment, 9: 169–173.

Dalal R.C., Henry R.J. (1986): Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. Soil Science Society of America Journal, 50: 120–123.

Davies T. (2005): An introduction to near infrared spectroscopy. NIR News, 16: 9–11.

Islam K., Singh B., McBratney A. (2003): Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Soil Research, 41: 1101–1114.

Jones A., Fernandez-Ugalde O., Scarpa S. (2020): LUCAS 2015 Topsoil Survey. Presentation of Dataset and Results, EUR 30332 EN, Luxembourg, Publications Office of the European Union.

Kennard R.W., Stone L.A. (1969): Computer aided design of experiments. Technometrics, 11: 137–148.

Kibblewhite M.G., Miko L., Montanarella L. (2012): Legal frameworks for soil protection: Current development and technical information requirements. Current Opinion in Environmental Sustainability, 4: 573–577.

Lal R. (2004): Soil carbon sequestration impacts on global climate change and food security. Science, 304: 1623–1627.

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

Orgiazzi A., Ballabio C., Panagos P., Jones A., Fernández-Ugalde O. (2017): LUCAS soil, the largest expandable soil dataset for Europe: A review. European Journal of Soil Science, 69: 140–153.

Ramirez-Lopez L., Behrens T., Schmidt K., Stevens A., Demattê J.A.M., Scholten T. (2013): The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. Geoderma, 195 (Supplement C): 268–279.

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. (2009): Digital soil map of the world. Science, 325: 680–681.

Savitzky A., Golay M.J.E. (1964): Smoothing and differentiation of data by simplified least squares procedures. Analysis Chemistry, 36: 1627–1639.

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

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

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

Tóth G., Jones A., Montanarella L. (2013): LUCAS Topsoil Survey: Methodology, Data and Results. JRC Technical Reports. EUR26102-Scientific and Technical Research Series. Luxembourg, Publications Office of the European Union.

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

Viscarra Rossel R.A., 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: 59–75.

Viscarra Rossel R., Behrens T., Ben-Dor E., Brown D., Demattê J., Shepherd K., Shi Z., Stenberg B., Stevens A., Adamchuk V. (2016): A global spectral library to characterize the world's soil. Earth Science Reviews, 155: 198–230.

Ward K.J., Chabrillat S., Neumann C., Foerster S. (2019): A remote sensing adapted approach for soil organic carbon prediction based on the spectrally clustered LUCAS soil database. Geoderma, 353: 297–307.

Ward K.J., Chabrillat S., Brell M., Castaldi F., Spengler D., Foerster S. (2020): Mapping soil organic carbon for airborne and simulated EnMAP imagery using the LUCAS soil database and a local PLSR. Remote Sensing, 12: 3451.

Wishart D (1969): Note: An algorithm for hierarchical classifications. Biometrics, 25: 165–170.

Wold S., Sjöström M., Eriksson L. (2001): PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory, 58: 109–130.

Received: July 1, 2022
Accepted: December 16, 2022
Published online: January 23, 2023