European Journal of Soil Science

European Journal of Soil Science, 2018 doi: 10.1111/ejss.12553

Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database

F. Castaldi[a], S. Chabrillat[b], C. Chartin[a], V. Genot[c], A. R. Jones[d] & B. van Wesemael[a]

[a]*Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Universite Catholique de Louvain, Croix du Sud 2, L7.05.16, 1348 Louvain la neuve, Belgium,* [b]*Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum GFZ, Telegrafenberg, 14473 Potsdam, Germany,* [c]*Station Provinciale d'Analyses Agricoles, Rue de Dinant 110, 4557, Tinlot, Belgium, and* [d]*European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via E. Fermi, 2749, 21027, Ispra, Italy*

Summary

Quantification of the soil organic carbon (SOC) content over large areas is mandatory to obtain accurate soil characterization and classification, which can improve site-specific management at local or regional scales. In this context, soil spectroscopy is a well-consolidated and widespread method to estimate soil variables, and in particular SOC content, at a low cost for routine analysis. The increasing number of large soil spectral libraries collected worldwide reflects the importance of spectroscopy in soil science. These large libraries contain soil samples derived from a large number of pedological regions and thus from different parent materials and soil types. In the light of the huge variation in the spectral responses to SOC content and composition, a rigorous process is necessary to subdivide large spectral libraries to avoid calibration with global models that fail to predict local variation in SOC content. Here, we propose to classify the European LUCAS topsoil database with a cluster analysis based on a large number of soil properties. The soil samples collected from arable land in the LUCAS database were chosen to apply a standardized multivariate calibration approach, valid for large areas, to calibrate local models without the need for further field and laboratory work. Cluster analysis detected seven soil classes and the samples belonging to each class were used to calibrate specific partial least squares regression (PLSR) models to estimate SOC content in three spectral libraries collected in Belgium and Luxembourg. Soil organic carbon was predicted with good accuracy, both within each library (root mean square error (RMSE), 1.2–5.1 g kg^{-1}; ratio of performance to prediction (RPD), 1.41–2.24) and for the samples of the three libraries together (RMSE, 3.7 g kg^{-1}; RPD, 2.54). The proposed approach could enable SOC to be estimated for arable soils in Europe with only the spectra of soil samples and without the need for laboratory analyses.

Highlights

- We investigated the potential of the LUCAS database to estimate SOC in spectral libraries.
- We proposed a routine approach to estimate SOC with less laboratory work.
- The classification of the LUCAS topsoil dataset improved estimation accuracy of SOC
- SOC content can be predicted from soil spectra with models calibrated on the LUCAS database.

Introduction

Soil organic carbon (SOC) considerably influences fertility in arable and grassland soil (Sanchez *et al.*, 2009). The estimation of SOC content is not only important for agricultural purposes, but also because of its crucial role in the global carbon cycle. Therefore, quantification of the SOC stock has also become important for environmental monitoring and in relation to global warming (Paustian *et al.*, 2016). Recently, Minasny *et al.* (2017) reviewed the status of SOC stocks in 20 regions of the world, including the potential for improved management in the framework of the '4 per mille' initiative: 'Soils for Food Security and Climate'. They found that the potential for C sequestration was largest in arable land and that individual farmers face the challenge of developing techniques to increase the SOC status of their soil.

Correspondence: F. Castaldi. E-mail: fabio.castaldi@uclouvain.be
Received 14 March 2017; revised version accepted 4 February 2018

Soil spectroscopy is a well-established and widespread method to estimate soil properties by exploiting the interaction between chromophores and visible (vis, 400–780 nm) and near infrared (NIR, 780–2500 nm) radiation. This interaction is evident from absorption features associated with overtones (energy transition from fundamental to excited state) of functional groups (Clark, 1999). The increasing number of large soil spectral libraries worldwide testifies to the importance of VNIR spectroscopy (e.g. Knadel *et al.*, 2012; Peng *et al.*, 2013; Guerrero *et al.*, 2016). Recently, some of these libraries were grouped to form a global spectral library (Viscarra Rossel *et al.*, 2016), consisting of 23 631 soil samples from 92 countries. Moreover, in the framework of the European land use/cover area frame statistical survey (LUCAS), a large European soil spectral library was collected in 2009 (Tóth *et al.*, 2013). The LUCAS topsoil database consists of about 20 000 soil samples from the 25 member states of the European Union, and for each sample, in addition to spectral acquisitions, 12 chemical and physical variables were measured.

In spite of the existence of these large soil spectral libraries, most published research concerns the calibration of prediction models with local soil datasets (e.g. Casa *et al.*, 2013). To achieve a representative local dataset, which includes the entire variation in the study area, extensive sampling, spectral acquisition and laboratory analysis of properties (e.g. SOC) are necessary. All three procedures are time consuming and expensive. Moreover, even if validation of the local models within the same area of calibration often provides good results, the applicability of these empirical models in areas with different soil conditions (extrapolation) is difficult or impossible. Until physically based models are developed, care should be taken that the calibration dataset covers as much of the variation in the validation set as possible. A possible way to solve issues related to the collection of a representative *ad hoc* local dataset could be the use of large soil spectral libraries. These cover an extensive area with a range of soil conditions from which to derive calibration models that could be applied to different soil datasets, while at the same time ensuring standardization of the analytical techniques. To exploit a model calibrated with another library, however, it is necessary to use the same protocol in the field and laboratory, and the same spectral instrument for both datasets. Alternatively, when different instruments are used, a spectral transfer procedure can be applied to quantify and correct the differences between the spectral outputs provided by the two instruments, commonly called master and slave instruments (e.g. Shenk & Westerhaus, 1991a; Fearn, 2001; Kopačková & Ben-Dor, 2016).

The 'spiking' approach could increase the estimation accuracy of calibrations obtained from large libraries for use at local scales, but this means that some samples from the target site have to be added to the calibration matrix (Guerrero *et al.*, 2010; Wetterlind & Stenberg, 2010; Peng *et al.*, 2013). Moreover, this approach involves choosing which and how many target samples to add to the large library, and also to use for laboratory analyses to measure the target variables of the selected samples.

Large libraries, especially national and continental libraries, contain soil samples from a range of pedological regions with different parent materials and soil types. This heterogeneity entails variation of the mineralogical and organic contents (Clark, 1999; Chabrillat *et al.*, 2002; Castaldi *et al.*, 2015). This heterogeneity in large libraries is very relevant if the target variable is organic matter because of its structural complexity. Organic matter composition, and thus its interaction with the radiation along the electromagnetic spectrum, is influenced by the source of organic materials and their degree of mineralization, resulting in a very variable spectral response (Ben-Dor *et al.*, 1997). The same SOC content can be derived from many types of organic matter, which affect the spectral shape in different spectral regions. The spectral features related to organic matter occur along the entire spectral region, because of the heterogeneity of its components. Strong correlations between reflectance and organic matter content were detected in the visible region at 450, 590 and around 664 nm (Ben-Dor *et al.*, 1997). However, the large correlations in the visible region are mainly due to the direct relation between darkness of the soil and quantity of organic matter (Castaldi *et al.*, 2016). This relation is true, *ceteris paribus*, for the same soil moisture content and same parent material. Many important features for the estimation of SOC are in the NIR region (Viscarra Rossel & Behrens, 2010). They relate to specific chemical bonds: between carbon and hydrogen (C–H) at 1340–1380 nm, the amide nitrogen–hydrogen bond (N–H) and the bond between oxygen and hydrogen in the hydroxyl group, both at 1860–1900 nm. Lignin and cellulose, two of the main organic compounds, affect the reflectance between 1600–1800 nm and around 2100 nm, respectively (Ben-Dor *et al.*, 1997). Other features linked to cellulose, lignin and starch occur around 2300 nm, close to features related to phenolic, amide and aliphatic groups (Ben-Dor *et al.*, 1997).

The large variation in spectral response to soil properties and characteristics means that a rigorous process is necessary to subdivide large spectral libraries and to avoid the failure of global models to predict local variation in SOC content (Stevens *et al.*, 2013). The subdivision of large spectral libraries to improve the estimation of soil variables can be carried out with two types of information: (i) auxiliary variables (Wijewardane *et al.*, 2016), parent material or land cover (Stevens *et al.*, 2013) to derive the classes and (ii) vis–NIR spectral data to decrease the number of samples in the calibration models by a local approach, whereby sets of samples spectrally similar to each sample of the validation dataset can be selected (Clairotte *et al.*, 2016; Viscarra Rossel *et al.*, 2016), or by cluster analysis to group spectrally similar samples together (Araújo *et al.*, 2014, Shi *et al.*, 2014). Two examples where the local approach with spectral data was enhanced by including an auxiliary variable are use of the LUCAS spectral data with sand content and geographical coordinates (Nocita *et al.,* 2014), and Shi *et al.* (2014) created local models from the Chinese vis–NIR spectral library and the soil geographical zones.

Here, we propose to subdivide the LUCAS topsoil database into soil classes based on all 12 available soil properties. The arable land soil samples within LUCAS were used to apply a

standardized multivariate calibration approach that can be applied to estimate variables of most agricultural soil types in Europe without additional field and laboratory work. We are not aware of any study that has used LUCAS topsoil or any other large spectral library to predict SOC by exploiting only the spectra of the samples.

To verify the potential of the LUCAS data to estimate SOC content in geographically local libraries, we followed five steps: (i) classification of the LUCAS samples collected from arable land, (ii) calibration of partial least squares regression (PLSR) models for each LUCAS class detected, (iii) comparison of the original LUCAS spectra with those obtained in a different laboratory with the same instrument, (iv) classification of three spectral libraries collected at the target sites (two in Belgium and one in Luxembourg) according to LUCAS classes and (v) estimation of SOC content by applying the LUCAS calibration models of a specific soil class only to samples of the target sites of the same class.

Materials and methods

Cluster analysis on LUCAS topsoil database

Each sample of the LUCAS topsoil database (Section S1, Supporting Information) was linked to the general LUCAS land-use and land-cover database (Eurostat, 2015) through an identification number (POINT_ID). This enabled selection of soil variables and spectra of the samples from arable land only (cereals, root crops, non-permanent industrial crops, dry pulses, vegetables, flowers, fodder crops, permanent crops and fallow land). This subset consists of 12 128 samples and will be referred to as LUCAS_crop. The cluster analysis of LUCAS_crop based on the 12 soil variables used the *k*-means clustering algorithm of Hartigan & Wong (1979) and Euclidean distance as a metric. This algorithm has been used successfully in the past to improve the prediction performances of large spectral libraries (e.g. Araújo *et al.*, 2014). There is some correlation among the 12 soil variables (e.g. pH in H_20 and $CaCl_2$, SOC and N), but they provide different information; therefore, removing any could lead to a loss of information (Table S1). Therefore, we carried out a principal component analysis (PCA) to obtain uncorrelated principal components, while still preserving most of the information (Shi *et al.*, 2014). The PCA was performed on the correlation matrix (Section S2, Supporting Information).

We set the optimal number of clusters by the 'gap' method (Tibshirani *et al.*, 2001); the gap value (Gap_c) was calculated for each number of clusters, *c*, and the largest was selected. The Gap_c is the difference between the logarithm of the within-cluster sum of squared distances from the cluster means $\ln W_{(c)}$) and its expected value under an appropriate null reference distribution $E_n^*[\ln(W_{(c)})]$ defined by bootstrapping, where *n* is the sample size.

We used the soil classes from cluster analysis with spectra to classify soil samples without any information about soil variables. For this, we used an artificial neural network (ANN) for discriminant analysis on the LUCAS_crop dataset, which is generally advantageous for categorical discrimination compared with traditional parametric techniques. A multilayer perceptron feed-forward neural network was created, where the input layer includes the soil spectral data and the output layer contains as many neurons as detected classes. The number of units in the hidden layer that provided the best cross-validation results of the classification was chosen to establish the final neural network architecture. The training dataset comprised 8530 samples (70%) extracted from LUCAS_crop. They were selected to take into account the relative frequency of each class detected by cluster analysis. We then tested the ANN on the remaining 3656 samples of LUCAS_crop. The overall process of discrimination (i.e. from selection of training and testing datasets to evaluation of the accuracy) was repeated 10 times, randomly extracting soil samples each time. The classification accuracy of the discriminant analysis was evaluated for each setting (different number of neurons in the hidden layer), both in general, including the weighted Cohen's kappa coefficient (*k*), and for each class. The iterative process allows a more robust validation, avoiding unrealistic results because of certain characteristics of the training and test datasets.

Geographically local spectral libraries

Three soil spectral libraries were collected in Belgium and Luxembourg (Figure 1).

Two libraries were collected in Belgium; the first comprises 54 samples from the northern part of Wallonia (Loam belt), along a 7-km-wide SW–NE strip of land from Gembloux to Sint-Truiden, whereas the second includes 61 samples from the entire Walloon region (Belgium). The northern part of Wallonia is mostly characterized by recent geological formations of the Quaternary, Tertiary and Secondary (Cretaceous) Eras. The area south of the Sambre–Meuse river axis is mainly characterized by geological formations from the Paleozoic Era (Devonian period), with outcrops from the Carboniferous, Cambrian and Silurian periods (Goidts & van Wesemael, 2007). The most productive soils are in the Loam belt where the Niveo-aeolian loess mainly occurs. Schist and psammite bedrock in association with limestone outcrops is frequent in the southern part. However, a thin loess cover persists (silt loam soils alternating with stony loam soils), which makes the soils suitable for agriculture. The average annual temperature decreases from the coast (around 10–11°C) to the Ardennes (around 8–9°C). The average annual precipitation shows the opposite trend (from 700 mm on the coast to 1200 mm in the Ardennes).

The third spectral library comprises 84 samples from the central and northern part of the Grand Duchy of Luxembourg within the Gutland–Oesling region. The northern region (Oesling) is characterized by a sub-horizontal plateau with a mean altitude of 450 m; the soils of this region are derived from Devonian slate, quarzite and sandstone substrates (Dystric Regosols and Dystric Cambisols (IUSS Working Group WRB, 2006)). The soil parent material in the southern region (Gutland) includes secondary marls, limestone, sandstone and dolomite. The Gutland region can be divided into two sub-regions: in the north the soils are mainly sandy loam, clay loam and clay (Dystric Cambisol (IUSS Working Group WRB, 2006)) and in the south the texture is sandy or sandy loam (Haplic Luvisols (IUSS Working Group WRB, 2006)).

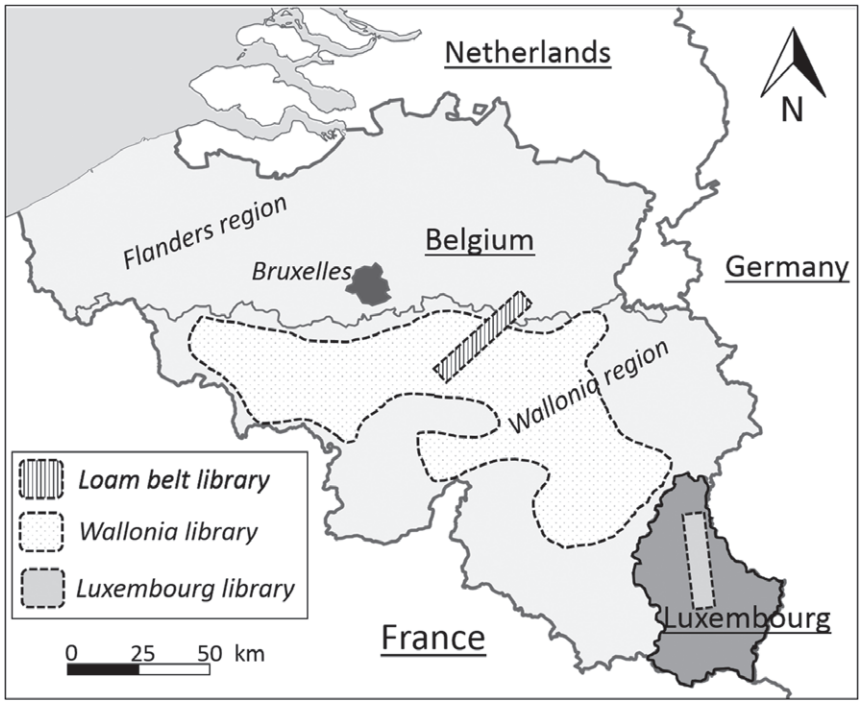

Figure 1 Map of the three sampling areas that relate to the soil spectral libraries.

Each sample was a composite of five subsamples from the 0–10-cm depth (0–20 cm for the Wallonia library), taken with a gouge auger within a 5-m radius. Soil samples were air-dried and passed through a 2-mm sieve in the laboratory. Each sample was scanned by the same instrument and the protocol was the same as that used for scanning the LUCAS dataset. The SOC values were measured by the same protocol and instrument as those in the LUCAS database (see Section S1) and they were used only to evaluate the accuracy of the prediction models.

Estimation of SOC in LUCAS_crop

To test the value of the clustering, we split each class detected within the LUCAS_crop dataset into calibration and validation subsets using the DUPLEX algorithm (Snee, 1977) on the spectral data, creating two subsets for each class with similar spectral properties and the same number of samples. The algorithm enables selection of calibration and validation datasets that represent the variation present. The DUPLEX algorithm was applied using the prospectr package (Stevens & Ramirez-Lopez, 2014) developed in R software (R Core Team, 2014). First, outliers were deleted from each class of LUCAS_crop.

Samples with an SOC content below or above 1.5 times the interquartile range (IQ) were considered outliers (Tukey, 1977) and were excluded from the calibration subsets used to form the partial least squares regression (PLSR) models. Each of these calibration models was validated on each validation subset, and not only on the subset of the same class. This enabled verification of whether a reasonable estimation accuracy was restricted to the validation subset of the same class, confirming goodness of the classification process.

Partial least squares regression (PLSR) is a well-known multivariate regression method in soil spectroscopy that enables an optimal linear model to exploit a limited number of uncorrelated components (i.e. the spectral bands) (Wold *et al.*, 2001). The components of the PLSR model are obtained by taking into account both the predictor matrix (in this case the spectral bands) and the dependent variable (SOC). We assessed the accuracy of each SOC prediction model for the calibration and the validation models with the root mean square error (RMSE; Equation (1)), ratio of the performance to deviation (RPD; Equation (2)), ratio of performance to interquartile range (RPIQ; Equation (3)) and coefficient of determination (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_\mathrm{o} - y_\mathrm{p}\right)^2}{n}}, \tag{1}$$

$$\mathrm{RPD} = \frac{\sigma}{\mathrm{RMSE}}, \tag{2}$$

$$\mathrm{RPIQ} = \frac{\mathrm{IQ}}{\mathrm{RMSE}}, \tag{3}$$

where y_o is the observed and y_p is the predicted value, n is the number of the samples, σ is the standard deviation and IQ is the interquartile range.

Estimation of SOC in local libraries

We estimated SOC content in the three geographically local libraries using two different approaches.

The first approach was based on the dataset collected by the laboratories of the REQUASUD agricultural extension service network; it comprises about 10 000 samples collected within the Wallonia region. The spectra of this dataset were acquired with an XDS Rapid Content Analyzer (FOSS NIR Systems Inc., Laurel, MD, USA) spectroradiometer between 400 and 2500 nm. They were used to calibrate local PLSR models (Genot *et al.*, 2011) to estimate SOC in the three spectral libraries (Luxembourg, Wallonia and Loam belt). The local approach entailed the selection of the spectra of the calibration model according to the similarity index. This index is based on the coefficient of determination (R^2) between the spectrum to be predicted and the spectra of the REQUASUD library. This procedure reduces the calibration dataset by selecting only the samples with a similarity index of 0.99. Thus, if there were not enough correlated spectra, the SOC of the sample would not be estimated (Genot *et al.*, 2011). For the samples of the three libraries, however, the number of correlated spectra was always sufficient.

The second approach entailed the classification of each sample of the three libraries by applying the ANN trained by the LUCAS_crop dataset. Then, we predicted SOC content of the three libraries with the PLSR model calibrated with the LUCAS_crop subset of the same class (Figure 2).

The estimation accuracy of the predictive models calibrated with large spectral libraries is not known in advance; therefore, possible outliers should be detected from the available data (i.e. the spectra). Therefore, before SOC was estimated for samples of the target sites, a modified Mahanalobis distance (GH) was calculated between each spectrum of the three spectral libraries and the average spectrum of REQUASUD to detect outliers (Shenk & Westerhaus, 1991a; Genot *et al.*, 2011; Gogé *et al.*, 2012). The GH values for LUCAS_crop were computed after cluster analysis in this case; GH was the distance between each spectrum of the spectral libraries and the average spectrum of LUCAS_crop belonging to the same class. For example, for a Belgian sample classified as B, the GH value was calculated as the distance between its spectrum and the average spectrum of class B in LUCAS_crop. Each spectrum of the dataset was transformed to a standard normal variate (SNV) and a PCA was carried out on these spectra. Mahanalobis distance was then calculated for each sample and divided by the number of principal components used (Shenk & Westerhaus, 1991b; Genot *et al.*, 2011). The GH value measures the similarity between a spectrum and all of the others. In other words, large values of GH indicate a large distance between the investigated spectrum and the average spectrum of REQUASUD or LUCAS_crop. Values of GH greater than three are considered to indicate outliers (Shenk & Westerhaus, 1991b). For spectra with a GH >3, spectroscopic prediction models were not used and samples were analysed by traditional laboratory analysis.

Results

Cluster analysis and SOC estimation in the LUCAS_crop dataset

Cluster analysis was carried out with eight principal components, which explained 97% of the variance of the 12 soil variables available in the LUCAS topsoil dataset (Table 1 and Section S1). The largest Gap_c value in the cluster analysis was for seven clusters; the classes are labelled with letters from A to G (Figure 3, Table 2). The soils of Class A are mainly a slightly basic loam with a large silt content (47.3%) and a mean SOC content of 29.67 g kg^{-1} (Table 2). They are also characterized by large values of P, K, N and CEC. Class B consists of slightly acid soil (pH, 5.6) with a sandy loam texture (sand, 70.6%), small $CaCO_3$ content (0.4 g kg^{-1}) and small macroelement (N, P, K) and SOC contents. The soils classified as C are also slightly acid with small N, P, K, SOC and $CaCO_3$ contents; the texture is silty loam with a large silt content (57%). Class D consists mainly of acid organic soils with a large organic carbon content that ranges between 134.9 and 570 g kg^{-1} and with a mean of 318.6 g kg^{-1}. The soils of class E have a clay loam texture, a basic reaction, large $CaCO_3$ content (391.8 g kg^{-1}), small mean SOC content (16.9 g kg^{-1}) and are generally rich in extractable K (283.4 mg kg^{-1}). Class F has soil with a slightly basic pH, a sandy loam texture (sand, 54.3%) and small nutrient content (SOC, N, P, K). The soils of class G have a silty clay loam texture with small sand content (15.6%), and consistent quantities of $CaCO_3$ (60.8 g kg^{-1}) and K (326.4 mg kg^{-1}).

Table 1 Eigenvalues and percentages of explained variance from the principal component analysis (PCA) of the LUCAS_crop dataset

Order	Eigenvalue	Percentage	Accumulated percentage
1	3.61	30.11	30.11
2	2.60	21.68	51.79
3	1.45	12.12	63.91
4	1.23	10.24	74.15
5	0.97	8.08	82.23
6	0.76	6.37	88.60
7	0.59	4.95	93.55
8	0.48	4.00	97.55
9	0.20	1.65	99.20

The overall accuracy obtained by the ANN discriminant analysis using five neurons in the hidden layer was 0.80 with a kappa coefficient of 0.75 (Table 3). The accuracy for each class ranged from 0.72 (F) to 0.91 (E), except for class A (in this case the accuracy was poor (0.39)).

The best PLSR calibration models for the prediction of SOC, in terms of RPD, RPIQ and R^2, were obtained for classes C (RPD, 2.0) and G (RPD, 2.1) (Table 4). The smallest RMSE values were for E and F (4.0 g kg^{-1}). The good results obtained with the calibration models for C and G were confirmed by the validation (RPD > 2). Nevertheless, SOC was also predicted well for the other classes (RPD ranges between 1.6 and 2.0), in particular class A (RPD, 2.0; RMSE, 6.6; RPIQ, 2.6; R^2, 0.8). The validation against a class different from the one for which the models were calibrated (e.g. model calibrated on A and validated on C) always provided poorer results than the validation on the same class (Tables 5). However, we obtained satisfactory results with the PLSR model of class A to predict SOC in class C (RPD, 1.9) and with the model for class C to predict samples of B (RPD, 1.8). For both cases, no

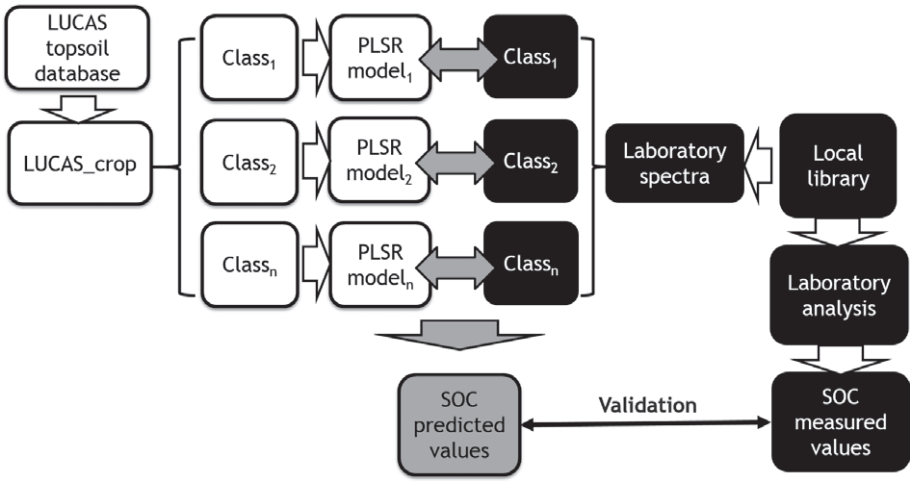

Figure 2 Flow chart showing the methodology for estimating soil organic carbon content in the three local libraries with the LUCAS topsoil database.

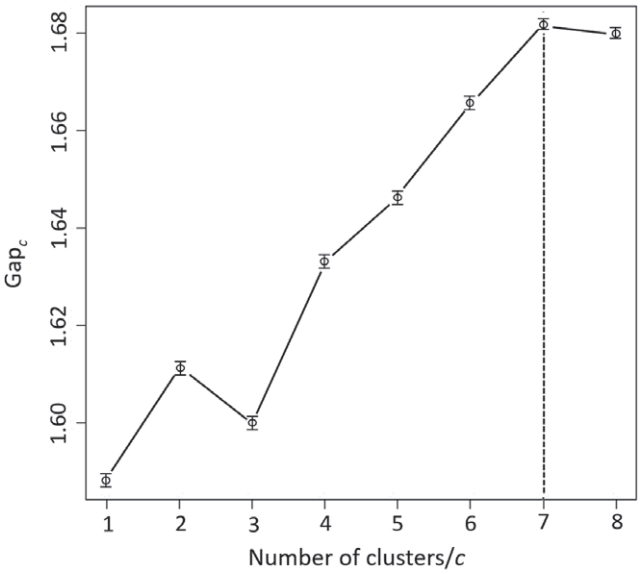

Figure 3 Plot of the gap value (Gap$_c$) for each number of clusters obtained by the *k*-means algorithm. The vertical dotted line indicates the number of clusters relating to the maximum Gap$_c$ value. The bars show the standard error of Gap$_c$.

satisfactory results were obtained by inverting the calibration and validation datasets.

Estimation of SOC in geographically local libraries

The SOC contents of soil samples of the Luxembourg dataset varied between 6.6 and 49.8 g kg^{-1}, with a mean value of 23.4 g kg^{-1} (Table 6). Although the Luxembourg samples were collected in a more confined area than the ones in the Walloon dataset, their variability is much greater. The SOC content in the Walloon samples ranged from 7 to 43.1 g kg^{-1}, but most of the values were close to the mean (11.8 g kg^{-1}). The SOC values in the Loam belt were quite small (mean, 10.2 g kg^{-1}; standard deviation, 1.69 g kg^{-1}).

The ANN trained with spectral data of the classified LUCAS_crop dataset was applied to the spectral libraries of the target sites: 86 samples were in class A (72 in Luxembourg and 14 in Wallonia) and 18 in class B (five in Luxembourg and 13 in Wallonia), whereas all samples of the Loam belt belonged to class F, together with seven samples in Luxembourg and 34 in Wallonia (Table 6). Because most of the samples of class A came from the Luxembourg dataset (84%), this class was the most variable in terms of SOC content (Table 6), whereas the standard deviations of the other two classes were <3 g kg^{-1}.

The FOSS spectral measurements had strong repeatability (see Section S1; Table S2; Figure S1); therefore, the PLSR models calibrated with LUCAS_crop were applied directly to the geographically local spectral libraries (i.e. without any spectral transfer function). The GH values from the REQUASUD dataset showed 37 outliers: 27 in Luxembourg, two in Wallonia and eight in the Belgian Loam belt dataset. The GH value from LUCAS_crop enabled removal of 18 outliers in the Luxembourg dataset, whereas GH was below 3 for all samples from the other two libraries. The overall accuracy using the REQUASUD model improved consistently after removal of outliers (RMSE, 3.4 g kg^{-1}; RPD, 3.13; RPIQ, 2.20; R^2, 0.8) (Figure 4a; Table 7). Overall, the LUCAS_crop models provided satisfactory accuracy both before (RMSE, 4.0 g kg^{-1}; RPD, 2.47; RPIQ, 2.20; R^2, 0.8) and after (RMSE, 3.7 g kg^{-1}; RPD, 2.54; RPIQ, 1.46; R^2, 0.8) the removal of outliers (Figure 4b). The best estimation accuracy was obtained with the Luxembourg libraries for both approaches (RPD > 2).

Discussion

Systematic reduction of the heterogeneity in the LUCAS_crop dataset by cluster analysis

Classification of the LUCAS crop dataset resulted in more effective models being fitted to estimate soil properties at the local scale (Genot *et al.*, 2011; Nocita *et al.*, 2014; Guerrero *et al.*,

Table 2 Descriptive statistics of soil variables for the seven classes detected by cluster analysis (n is the number of samples)

Class	Coarse /%	Clay /%	Silt /%	Sand /%	pH in H_2O	pH in CaCl	OC /g kg^{-1}	CaCO3 / g kg^{-1}	N /g kg^{-1}	P /mg kg^{-1}	K /g kg^{-1}	CEC /cmol(+) kg^{-1}
A ($n = 554$)												
Min	0	6.0	10.0	2.0	4.6	4.0	8.3	0	0.9	0	58.7	1.0
Mean	14.6	26.1	47.3	26.7	7.2	6.7	29.7	54.5	2.8	106.5	764.2	21.1
Max	68.0	66.0	82.0	81.0	8.8	7.8	178.5	549.0	12.7	789.8	7342.0	58.2
SD	10.7	10.2	14.7	16.7	0.7	0.7	20.8	95.6	1.7	64.6	689.3	8.5
B ($n = 2475$)												
Min	0	0	1.0	35.0	3.6	3.0	2.0	0	0.2	0	0	1.0
Mean	9.8	8.7	20.7	70.6	5.6	5.0	21.5	0.4	1.8	43.8	111.9	7.0
Max	67.0	38.0	54.0	97.0	7.2	6.8	183.1	7.0	11.2	179.3	797.8	48.6
SD	8.7	5.5	10.0	13.5	0.6	0.6	19.9	0.7	1.3	30.0	86.7	4.6
C ($n = 2805$)												
Min	0	2.0	17.0	1.0	3.7	3.2	2.8	0	0.3	0	0	1.0
Mean	12.1	21.6	57.0	21.4	6.0	5.5	25.7	1.2	2.3	33.0	168.1	12.2
Max	62.0	70.0	90.0	56.0	8.0	7.5	181.4	95.0	12.5	128.2	919.7	46.9
SD	9.4	8.9	13.8	12.9	0.7	0.8	20.1	3.9	1.4	23.0	101.8	5.4
D ($n = 140$)												
Min	0	0	0	0	3.7	3.0	134.9	0	8.2	0	0	13.6
Mean	12.5	5.1	7.0	5.0	5.3	4.9	318.6	11.3	18.0	34.4	193.5	69.5
Max	55.0	70.0	57.0	64.0	7.4	7.1	567.0	406.0	38.6	226.3	1885.0	190.4
SD	12.6	13.0	16.1	13.0	0.9	1.0	104.8	47.8	6.4	34.1	209.8	44.2
E ($n = 1565$)												
Min	1.0	1.0	15.0	1.0	6.8	6.4	1.0	114.0	0.3	0	0	1.0
Mean	16.5	28.3	46.3	25.4	8.0	7.4	16.9	391.8	1.5	21.5	283.4	18.1
Max	73.0	65.0	86.0	73.0	8.9	8.1	123.0	909.0	11.0	136.8	1061.0	61.8
SD	10.9	8.7	10.8	13.5	0.3	0.2	11.2	140.2	0.9	21.2	160.2	6.5
F ($n = 2424$)												
Min	0	1.0	1.0	21.0	6.0	5.1	1.0	0	0	0	0	1.0
Mean	12.3	16.2	29.5	54.3	7.4	6.8	13.4	30.9	1.5	27.9	165.8	13.5
Max	67.0	40.0	71.0	97.0	9.7	8.1	33.5	390.0	9.5	130.2	1039.0	55.7
SD	9.7	6.7	10.8	13.7	0.6	0.5	12.0	56.2	0.9	23.9	111.9	6.0
G ($n = 2223$)												
Min	0	7.0	13.0	1.0	4.8	4.3	1.0	0	0.6	0	0	1.0
Mean	13.3	36.4	48.0	15.6	7.5	7.0	25.1	60.8	2.4	25.8	326.4	27.9
Max	66.0	79.0	81.0	71.0	10.1	9.3	160.3	380.0	13.2	124.6	1276.0	137.0
SD	10.6	11.8	13.0	10.2	0.6	0.6	19.3	72.1	1.6	21.5	189.3	10.3

Rows report basic statistics, including minimum (Min), mean, maximum (Max) and standard deviations for each soil variable (OC, organic carbon; CaCO3, calcium carbonate; N, nitrogen; P, phosphorus; K, potassium; CEC, cation-exchange capacity).

Table 3 Confusion matrix for the accuracy of classification obtained by the ANN discriminant analysis on the LUCAS_crop validation dataset

Class	A	B	C	D	E	F	G
A	0.39	0	0.12	0	0.05	0.18	0.26
B	0.01	0.86	0.08	0.01	0	0.05	0.00
C	0.01	0.06	0.84	0	0	0.06	0.03
D	0	0.11	0	0.82	0	0	0.07
E	0.01	0	0	0	0.91	0.04	0.05
F	0.02	0.08	0.07	0	0.04	0.72	0.07
G	0.03	0.01	0.05	0	0.03	0.09	0.80
Overall accuracy	0.80						
k[a]	0.75						

[a]k, Cohen's kappa coefficient.
ANN, artificial neural network.

2016). Guerrero *et al.* (2016) showed that a non-systematic reduction of library size does not ensure reliable results. They randomly decreased a Spanish national library and the accuracy of SOC estimates consistently worsened. Wetterlind & Stenberg (2010) used PCA to select 50 samples from the Swedish national library that were spectrally most similar to each target site (reduced national dataset). In this case, the reduced dataset usually outperformed the national dataset, but to obtain reliable results they 'spiked' both datasets with local samples. Classification of large spectral libraries, and thus the use of local models, has always appeared to improve the accuracy of estimation compared with global models (e.g Araújo *et al.*, 2014; Clairotte *et al.*, 2016).

Although the use of spectral data for classification is attractive and it has been widely used (e.g. Gogé *et al.*, 2012; Nocita *et al.*, 2014),

Table 4 The statistics for the PLSR calibration models and their validation for each LUCAS_crop class

Class	Number of samples	Number of outliers removed	PLSR components	Calibration				Validation			
				RMSE /g kg^{-1}	RPD	RPIQ	R^2	RMSE /g kg^{-1}	RPD	RPIQ	R^2
A	554	34	9	7.4	1.7	2.3	0.6	6.6	2.0	2.6	0.8
B	2475	174	16	5.6	1.8	2.3	0.7	5.7	1.8	2.2	0.7
C	2805	166	20	5.7	2.0	2.6	0.8	5.3	2.2	2.9	0.8
D	140	0	4	73.1	1.4	2.0	0.4	62.0	1.6	2.3	0.6
E	1565	85	16	4.0	1.6	2.0	0.6	3.8	1.7	2.1	0.7
F	2424	157	20	4.0	1.6	2.0	0.6	3.9	1.7	2.2	0.7
G	2223	161	20	5.0	2.1	2.6	0.8	5.0	2.1	2.7	0.8
All								5.6	6.5	2.2	0.8

PLSR, partial least squares regression; RMSE, root mean square error; RPD, ratio of performance to prediction; RPIQ, ratio of performance to interquartile range.

Table 5 The RPD values of validation for each LUCAS class obtained using the PLSR calibration models of all classes

Class	A	B	C	D	E	F	G
A	2.01	0.69	1.87	0.38	0.51	0.45	1.44
B	0.88	1.75	0.97	0.35	0.49	0.84	0.93
C	0.79	1.80	2.16	0.35	0.49	0.81	0.82
D	0.05	0.04	0.04	1.60	0.01	0.02	0.03
E	1.24	0.72	1.28	0.35	1.72	0.86	1.28
F	0.93	1.39	1.09	0.33	0.96	1.71	0.99
G	1.55	0.55	1.49	0.36	0.85	0.61	2.14

RPD, ratio of performance to prediction; PLSR, partial least squares regression.

Table 6 Descriptive statistics of the three local libraries (Luxembourg, Loam belt and Wallonia) and the three detected classes (A, B and F)

Dataset	*n*	Minimum /g kg^{-1}	Maximum /g kg^{-1}	Mean /g kg^{-1}	SD /g kg^{-1}
Library					
Luxembourg	84	6.61	49.85	23.39	11.12
Wallonia	61	7.00	43.10	11.77	4.64
Loam belt	54	7.06	15.49	10.18	1.69
Class					
A	86	6.62	49.85	23.55	11.28
B	18	8.00	19.90	11.45	2.89
F	95	7.00	15.49	10.53	1.79
All	199	6.62	49.85	16.24	9.63

the values of soil properties derive from chemical analysis, whereas spectra provide only indirect information on soil characteristics and only for those variables that interact with vis–NIR radiation (chromophores); therefore, prediction could be weak (Shi *et al.*, 2014). Nocita *et al.* (2014) used vis–NIR spectra of the LUCAS soil database to develop local PLSR models and, although the results for SOC prediction were good, the accuracy improved by adding sand content as a covariate in the local PLSR models in croplands, grasslands and woodlands. Sand content, however, has disadvantages as a covariate when, as here, the goal is to predict SOC content of samples for which there is no *a priori* knowledge. We tried another approach to reduce heterogeneity of the LUCAS dataset by cluster analysis with soil properties and by not using spectral data at this first step. This approach improved the accuracy of estimation (Tables 5 and 8).

The discrimination process by ANN

Class A was the only one among the seven classes with a poor accuracy of classification (Table 3). This was partly because class A had similar mean values of some soil variables (sand, silt and $CaCO_3$) to those of other classes, especially classes E and G (Table 2). Nevertheless, the PLSR model calibrated with samples of class A could satisfactorily predict the samples of class G (Table 5). This was not surprising because the main difference between classes A, E and G was the large values of P and K in the former; these two macro-elements are known to have little effect on the prediction of SOC. However, the main cause of the poor accuracy of classification of Class A was probably the small number of samples within this class. In other words, this type of soil was under-sampled (only 554 samples) in the LUCAS database compared with many more in other classes.

Classification and SOC estimation in the three target sites

The LUCAS topsoil database was used mainly to estimate SOC or other soil properties at the national or continental scale (Panagos *et al.*, 2013). Nocita *et al.* (2014), however, implemented a local PLSR with the LUCAS topsoil database to improve the accuracy of predicting SOC. Although they obtained good cross-validation results by including sand as a covariate in the croplands data (RMSE, 3.9 g kg^{-1}; RPD, 2.5), their models were not used to predict SOC in other spectral libraries. Gogé *et al.* (2012) and Clairotte *et al.* (2016) used a national spectral library to develop local PLSR models. The validation datasets, however, were selected from the same library as the one used for the calibration. Although the validation dataset in Viscarra Rossel *et al.* (2016) was derived from

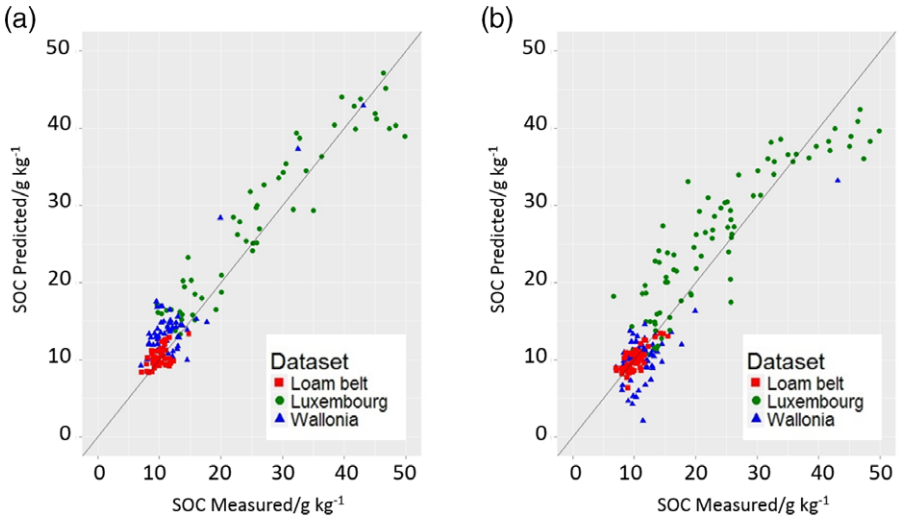

Figure 4 Plots of measured against predicted soil organic carbon content with (a) REQUASUD and (b) LUCAS_crop datasets.

Table 7 Independent validation statistics of the PLSR models calibrated with LUCAS_crop and REQUASUD datasets and applied to the three local libraries. The table gives the results for each library after the removal of outliers (n GH > 3)

Calibration dataset	Validation dataset	n	n GH > 3	RMSE /g kg^{-1}	RPD	RPIQ	R^2
Lucas_crop	Luxembourg	84	18	5.1	2.24	2.97	0.8
	Wallonia	61	0	3.2	1.45	0.93	0.6
	Belgian Loam Belt	54	0	1.2	1.41	1.69	0.6
	All	199	0	4.0	2.47	2.20	0.8
	All (GH < 3)	181	18	3.7	2.54	1.46	0.8
REQUASUD	Luxembourg	84	27	4.4	2.82	4.53	0.9
	Wallonia	61	2	3.4	1.61	0.88	0.6
	Belgian Loam Belt	54	8	1.3	1.10	1.37	0.5
	All	199	0	4.7	2.10	3.48	0.8
	All (GH < 3)	162	37	3.4	3.13	2.20	0.8

GH, modified Mahanalobis distance; PLSR, partial least squares regression; RMSE, root mean square error; RPD, ratio of performance to prediction; RPIQ, ratio of performance to interquartile range.

Table 8 The root mean square error (RMSE, g kg^{-1}) of validation of the PLSR SOC models calibrated with each class of the LUCAS_crop and applied to samples of three local libraries according to class

	Classes of the local samples		
Calibration dataset	A RMSE/g kg^{-1}	B RMSE/g kg^{-1}	F RMSE/g kg^{-1}
LUCAS A	5.4	4.8	2.6
LUCAS B	11.6	4.1	9.7
LUCAS C	11.6	8.7	9.7
LUCAS D	153.2	67.2	41.9
LUCAS E	9.0	10.3	11.6
LUCAS F	10.8	5.1	1.7
LUCAS G	6.4	5.0	4.8
LUCAS_crop	8.7	5.8	4.8

PLSR, partial least squares regression; SOC, soil organic carbon.

different laboratories across the world, the prediction model was not applied to a geographically local dataset, but rather to a large dataset (3585 samples) selected randomly from the global library with considerable variability in terms of SOC. These characteristics meant that a comparison between our results and those obtained by Viscarra Rossel *et al.* (2016) was not appropriate.

To our knowledge no previous research has satisfactorily used a continental or national vis–NIR spectral library to predict SOC by completely independent validation on geographically local datasets, which involved considerably reduced laboratory effort. Our independent validation of the PLSR models calibrated with the LUCAS_crop dataset on the geographically local libraries showed the best results for the Luxembourg samples (RPD, 2.24; RPIQ, 2.97). Most of the samples collected in Luxembourg, especially in the Oesling region, have large SOC content and this is mainly due to the wet and cold conditions of this plateau area. Although most of the outliers detected in the Luxembourg data belonged to class A, they did not strongly affect the RMSE value (Table 7). The estimation accuracy within the Loam belt and Walloon library was sufficient: RPD >1.40 and a small RMSE, especially in the Loam belt dataset (1.2 g kg^{-1}). The good results obtained for the Luxembourg samples were probably because of the large variability in SOC contents of the samples belonging to class A in Luxembourg (Figure 4; Table 7). The overall RMSE and those obtained for each spectral library (Luxembourg, Loam belt and Wallonia) were better than that reported by Guerrero *et al.* (2016), both with and without 'spiking'. However, the standard deviation and the range of SOC values in the three datasets were consistently less than in Guerrero *et al.* (2016); therefore, these two sets of results are not strictly comparable. The classification of both the LUCAS (Nocita *et al.*, 2014) and French national spectral libraries (Clairotte *et al.*, 2016) with local PLSR models resulted in an RPD of validation that was similar to the one we obtained for our three target sites (RPD, 2.54). Moreover, our RPD and RPIQ values were similar to those obtained in other studies where geographically local libraries were used to estimate SOC (Nocita *et al.*, 2014), or more frequently

where only cross-validation results were shown (Castaldi *et al.*, 2016). We always obtained the best estimation accuracy (RMSE) with the class provided by the ANN trained on the LUCAS_crop dataset (Table 8). This encourages the use of our approach and, in particular, it emphasizes the goodness of the classification by ANN. Moreover, the large RMSE values were obtained when the entire LUCAS_crop dataset was used, which affirms the advantage of classification for estimating SOC at the three target sites (Table 8).

The REQUASUD model provided slightly more accurate predictions than those obtained using our approach based on the LUCAS_crop data (Figure 4a,b); but the REQUASUD results required the exclusion of 37 samples (i.e. 19% of the three datasets and 32% of the samples in Luxembourg) (Table 7). The large variability in the LUCAS_crop dataset meant that fewer samples had to be removed than for the prediction models based on the REQUASUD dataset. Both approaches (LUCAS_crop and REQUASUD) suggest a methodology for the routine application of soil spectroscopy to estimate soil properties that avoids chemical or physical laboratory analyses.

We used the same instrument for the spectral acquisitions in both the geographically local and LUCAS spectral libraries (XDS Rapid Content Analyzer, FOSS NIR Systems Inc., Laurel, MD, USA). Other instruments (e.g. field spectroradiometers) are widely used to record spectra for spectral libraries across the world; therefore, there is a need to develop a spectral transfer method. Recently, Kopačková & Ben-Dor (2016) showed the importance of soil samples with stable spectral performances (internal soil standard samples) in soil spectroscopy to normalize and align all other soil spectral measurements.

Conclusions

We investigated the capacity of the European LUCAS topsoil database to predict SOC in three spectral libraries. A cluster analysis subdivided the LUCAS dataset into seven classes based on selected principal components from 12 soil properties. The spectral data were linked with SOC values of the LUCAS dataset to calibrate PLSR models for each class to predict SOC content with good accuracy both within each geographically local library (RMSE, 1.2–5.1 g kg^{-1}; RPD, 1.41–2.24) and for samples of the three libraries together (RMSE, 3.7 g kg^{-1}; RPD, 2.54). The large variability of the data in the LUCAS_crop library enabled SOC content to be predicted by a routine chemometrics approach. Our approach represents a reproducible tool to estimate SOC in arable soils with a satisfactory level of accuracy with only the spectra of soil samples and without the need for further laboratory analyses. To generalize the previous statement, however, other tests need to be carried out in different regions across Europe. The LUCAS soil database was resampled in 2015 and this will be repeated in 2018, adding new soil samples and properties. This should make the proposed methodology even more suitable for the quantitative estimation of SOC or other soil properties in different pedological regions across Europe.

Supporting Information

The following supporting information is available in the online version of this article:

S1. LUCAS topsoil dataset.

S2. Principal component analysis of the LUCAS_crop dataset.

Table S1. Descriptive statistics of soil variables of the LUCAS_crop dataset.

Table S2. The average sum of squared deviations (ASDS) computed between the new and old spectral data of the sub_crop dataset.

Figure S1. Example of the repeatability of spectral acquisition of three soil samples included in the sub_crop dataset.

Acknowledgements

The research was funded by the Belgian Federal Science Policy Office (BELSPO) as part of the PROSOIL project 'The evaluation of forthcoming satellites for mapping topsoil organic carbon in croplands' (Contract NR SR /10/327) and by the Service Public de Wallonie as part of the Carbiosol project. The support is gratefully acknowledged. We are grateful to Marco Bravin of the Earth and Life Institute of the Université Catholique de Louvain (UCL) for the essential organic carbon measurements.

References

Araújo, S.R., Wetterlind, J., Demattê, J.A.M. & 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. https://doi.org/10.1111/ejss.12165.

Ben-Dor, E., Inbar, Y. & Chen, Y. 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sensing of Environment*, **61**, 1–15. https://doi.org/10.1016/S0034-4257(96)00120-4.

Casa, R., Castaldi, F., Pascucci, S., Palombo, A. & Pignatti, S. 2013. A comparison of sensor resolution and calibration strategies for soil texture estimation from hyperspectral remote sensing. *Geoderma*, **197–198**, 17–26.

Castaldi, F., Palombo, A., Pascucci, S., Pignatti, S., Santini, F. & Casa, R. 2015. Reducing the influence of soil moisture on the estimation of clay from hyperspectral data: a case study using simulated PRISMA data. *Remote Sensing*, **7**, 15561–15582. https://doi.org/10.3390/rs71115561.

Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S. & 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. https://doi.org/10.1016/j.rse.2016.03.025.

Chabrillat, S., Goetz, A.F., Krosley, L. & Olsen, H.W. 2002. Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution. *Remote Sensing of Environment*, **82**, 431–445. https://doi.org/10.1016/S0034-4257(02)00060-3.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M. *et al.* 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma*, **276**, 41–52. https://doi.org/10.1016/j.geoderma.2016.04.021.

Clark, R.N. 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. In: *Manual of Remote Sensing, Volume 3, Remote Sensing for the Earth Sciences* (ed. A.N. Rencz), pp. 3–58. John Wiley and Sons, New York, NY.

Eurostat (2015) LUCAS 2015 *Land Use/Cover Area Frame Survey –Land Use and Land Cover Nomenclature* [WWW document]. Technical Reference Document C-3. URL http://ec.europa.eu/eurostat/documents/205002/6786255/LUCAS2015-C3-Classification-20150227.pdf/969ca853-e325-48b3-9d59-7e86023b2b27 [accessed on 11 August 2017], https://doi.org/10.1016/j.bmhimx.2015.05.002.

Fearn, T. 2001. Standardisation and calibration transfer for near infrared instruments: a review. *Journal of Near Infrared Spectroscopy*, **9**, 229–244.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y. & Dardenne, P. 2011. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. *Journal of Near Infrared Spectroscopy*, **19**, 117–138. https://doi.org/10.1255/jnirs.923.

Gogé, F., Joffre, R., Jolivet, C., Ross, I. & Ranjard, L. 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics and Intelligent Laboratory Systems*, **110**, 168–176. https://doi.org/10.1016/j.chemolab.2011.11.003.

Goidts, E. & van Wesemael, B. 2007. Regional assessment of soil organic carbon changes under agriculture in southern Belgium (1955–2005). *Geoderma*, **141**, 341–354. https://doi.org/10.1016/j.geoderma.2007.06.013.

Guerrero, C., Zornoza, R., Gómez, I. & Mataix-Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. *Geoderma*, **158**, 66–77. https://doi.org/10.1016/j.geoderma.2009.12.021.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D. *et al.* 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, **155**, 501–509. https://doi.org/10.1016/j.still.2015.07.008.

Hartigan, J.A. & Wong, M.A. 1979. Algorithm AS 136: a *k*-means clustering algorithm. *Applied Statistics*, **28**, 100–108. https://doi.org/10.2307/2346830.

IUSS Working Group WRB 2006. *World Reference Base for Soil Resources 2006*. World Soil Resources Reports no 103. FAO, Rome. https://doi.org/10.1096/fj.06-1005ufm.

Knadel, M., Deng, F., Thomsen, A. & Greve, M. 2012. Development of a Danish national Vis-NIR soil spectral library for soil organic carbon determination. In: *Digital Soil Assessments and Beyond. Proceedings of the 5th Global Workshop on Digital Soil Mapping* (ed. A.B. McBratney), pp. 403–408. CRC Press, Boca Raton, FL.

Kopačková, V. & Ben-Dor, E. 2016. Normalizing reflectance from different spectrometers and protocols with an internal soil standard. *International Journal of Remote Sensing*, **37**, 1276–1290. https://doi.org/10.1080/01431161.2016.1148291.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A. *et al.* 2017. Soil carbon 4 per mille. *Geoderma*, **292**, 59–86. https://doi.org/10.1016/j.geoderma.2017.01.002.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B. & Montanarella, L. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology & Biochemistry*, **68**, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.

Panagos, P., Ballabio, C., Yigini, Y. & Dunbar, M.B. 2013. Estimating the soil organic carbon content for European NUTS2 regions based on LUCAS data collection. *Science of the Total Environment*, **442**, 235–246. https://doi.org/10.1016/j.scitotenv.2012.10.017.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G.P. & Smith, P. 2016. Climate-smart soils. *Nature*, **532**, 49–57. https://doi.org/10.1038/nature17174.

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., Wollesen de Jonge, L. *et al.* 2013. Predicting soil organic carbon at field scale using a national soil spectral library. *Journal of Near Infrared Spectroscopy*, **21**, 213–222. https://doi.org/10.1255/jnirs.1053.

R Core Team 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ [accessed on 1 February 2017]. https://doi.org/10.4067/S0370-41062014000500018.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J. *et al.* 2009. Digital soil pap of the world. *Science*, **325**, 6–7.

Shenk, J.S. & Westerhaus, M.O. 1991a. New standardization and calibration procedures for Nirs analytical systems. *Crop Science*, **31**, 1694–1696.

Shenk, J.S. & Westerhaus, M.O. 1991b. Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. *Crop Science*, **31**, 469–474.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X. *et al.* 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, **57**, 1671–1680. https://doi.org/10.1007/s11430-013-4808-x.

Snee, R.D. 1977. Validation of regression models: methods and examples. *Technometrics*, **19**, 415–428. https://doi.org/10.1080/00401706.1977.10489581.

Stevens, A. & Ramirez-Lopez, L. 2014. *An Introduction to the Prospectr Package* [WWW document]. URL https://cran.r-project.org/web/packages/prospectr/prospectr.pdf [accessed on 15 June 2016], https://doi.org/10.1002/jcop.21663.

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.

Tibshirani, R., Walther, G. & Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 411–423. https://doi.org/10.1111/1467-9868.00293.

Tóth, G., Jones, A. & Montanarella, L. 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environmental Monitoring and Assessment*, **185**, 7409–7425. https://doi.org/10.1007/s10661-013-3109-3.

Tukey, J.W. 1977. Some thoughts on clinical trials, especially problems of multiplicity. *Science*, **198**, 679–684. https://doi.org/10.1126/science.333584.

Viscarra Rossel, R.A. & Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, **158**, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D. *et al.* 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*, **155**, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.

Wetterlind, J. & Stenberg, B. 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, **61**, 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.

Wijewardane, N.K., Ge, Y., Wills, S. & Loecke, T. 2016. Prediction of soil carbon in the conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. *Soil Science Society of America Journal*, **80**, 973–982. https://doi.org/10.2136/sssaj2016.02.0052.

Wold, S., Sjöström, M. & Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.