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

DOI: http://doi.org/10.1016/j.geoderma.2019.07.010

A remote sensing adapted approach for soil organic carbon prediction based on the spectrally clustered LUCAS soil database

Kathrin J. Ward [a], Sabine Chabrillat [a], Carsten Neumann [a], Saskia Foerster [a]

[a] GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

Corresponding author: K.J. Ward, E-Mail: ward@gfz-potsdam.de

Abstract

The estimation of the soil organic carbon (SOC) content plays an important role for carbon sequestration in the context of climate change, food security and soil degradation. Reflectance spectroscopy has proven to be a promising technique for SOC quantification in the laboratory and increasingly from air- and spaceborne platforms, where hyperspectral imagery provides great potential for mapping SOC on larger scales with regular updates. When applied on larger scales, soil prediction accuracy decreases due to the inhomogeneity of samples. In this paper, we examined if spectral clustering of the LUCAS EU-wide topsoil database is successful without using other covariates than the spectral database and can improve SOC model performance compared to a reference model that was calibrated on the whole database without clustering. Different clustering methodologies were tested, including a k-means clustering based on principal component analyses or based on spectral feature variables, combined with partial least squares regression (PLSR) models, and a clustering based on a local PLSR approach which builds a different multivariate model for each sample to be predicted. Furthermore, in order to allow for subsequent application to hyperspectral remote sensing data, atmospheric water wavelengths were removed from the analyses. The local PLSR approach achieved best results and was additionally applied to LUCAS spectra resampled to the upcoming hyperspectral EnMAP sensor which led to good results: R^2 = 0.66, RMSEP = 5.78 g kg^{-1} and RPIQ = 1.93. The k-means clustering approach showed slightly better results than the reference model. Overall, our results showed similar performances for SOC prediction models

compared to other approaches using PLSR with a larger spectral range and other soil parameters as covariates. This study shows that (i) it is possible to transfer the local PLSR approach onto a wavelengths reduced spectral library and to predict estimations of SOC at low-cost with reasonable accuracy based on large scale soil databases; and (ii) that the local regression approach is a valuable tool for SOC prediction models based solely on spectral data without the use of other soil covariates.

Keywords: soil organic carbon, reflectance spectroscopy, cluster analysis, soil spectral library, Europe

Abbreviations:

AF	Absorption feature
CF	Curve feature
CR	Continuum removal
EnMAP	Environmental Mapping and Analysis Programme
HF	Hull feature
LUCAS	Land Use/Land Cover Area Frame Survey
LV	Latent variables of PLSR
PCs	Principal components of a PCA
PCA	Principal component analysis
PLSR	Partial least squares regression
SAM	Spectral angle mapper
SFV	Spectral feature variables
SOC	Soil organic carbon
SWIR	Shortwave infrared
VNIR	Visible and near-infrared

Highlights

- Reduced need for ground truth data by large-scale spectral clustering and modelling
- Adapting existing approaches in preparation for future spaceborne SOC estimation
- Comparison of clustering approaches with reference models on complete LUCAS data
- Local PLSR outperforms other approaches and reference model in SOC quantification

1. Introduction

Soils provide essential ecosystem services such as food production, flood prevention and carbon sequestration (Kibblewhite et al., 2012). With regard to carbon sequestration, soils generally hold the potential of intensified carbon uptake to partially offset fossil fuel emissions and thereby attenuating climate change (e.g. Conant et al., 2011; Lal, 2004). This potential is especially high on degraded soils where improved agricultural management practices can additionally lead to increased crop yields and thus enhanced food security (Denton et al., 2014; Lal, 2004). A key parameter to determine the state of soils is the soil organic carbon (SOC) content (Sanchez et al., 2009). In order to mitigate the risks of degrading soils and thus threatened appropriation of ecosystem services, a monitoring of SOC content and other soil parameters is essential. However, due to high costs and the time consuming nature of conventional soil sampling and analysis this can hardly be achieved on larger scales (Araújo et al., 2014; Conant et al., 2011; Sanchez et al., 2009).

Therefore, diffuse reflectance spectroscopy of soils in the visible and near- infrared (VNIR) to the shortwave infrared (SWIR) (400-2500 nm) provides a good alternative for the quantification of soil properties (Islam et al., 2003). The spectral properties of soils can be measured in a cheap and rapid way and thus provide a trade-off between costs and accuracy (Bellon-Maurel and McBratney, 2011; O'Rourke and Holden, 2011; Viscarra Rossel and Behrens, 2010). Soil spectroscopy is based on the assumption that the concentration of a specific soil property is

78 linearly related to a combination of absorption features within the spectrum (Bellon-Maurel and
79 McBratney, 2011; Ben-Dor et al., 1999). These absorption features are induced by overtones
80 and combination bands of fundamental vibrations of some of the molecules' functional groups,
81 e.g. the hydroxyl group (OH). As each functional group's overtones and combination bands are
82 located at specific wavelengths of the spectrum, different materials can be identified (Ben-Dor
83 et al., 1999; Davies, 2005). Absorption features in the visible range (400-700 nm) may also be
84 caused by electron transitions (Ben-Dor et al., 1999).

85 Soil reflectance spectra consist of broad and weak absorption features that are partly
86 superimposing each other (Stenberg et al., 2010). To extract quantitative information of
87 potentially small amounts of soil constituents, different mathematical modelling approaches are
88 applied (comparisons e.g. in Stevens et al., 2013; Viscarra Rossel and Behrens, 2010). One of
89 the most commonly used techniques is partial least squares regression (PLSR) which accepts a
90 large number of predictor variables with high collinearity (Stenberg et al., 2010) which is the
91 case with diffuse reflectance spectroscopy.

92 Diffuse reflectance spectroscopy in the VNIR-SWIR range has been applied in soil science for
93 more than 20 years (Bellon-Maurel and McBratney, 2011; Stenberg et al., 2010). It is most
94 often used in the laboratory but in-situ as well as airborne applications are increasingly utilized
95 (Ben-Dor et al., 2009). A large number of studies have been conducted in the laboratory that
96 prove successful estimation of soil properties on local and regional scales with high accuracies
97 (overview in Viscarra Rossel et al., 2016). Numerous models have been calibrated out of many
98 local spectral soil libraries with different measurement protocols leading to a large number of
99 independent small scale models (Stevens et al., 2013). More recently there is the tendency to
100 develop national and international or even global soil spectral databases and to build global
101 prediction models (e.g. Araújo et al., 2014; Brown et al., 2006; Tóth et al., 2013; Viscarra
102 Rossel et al., 2016). On larger areas, the prediction accuracies tend to decrease, which is mainly

caused by different, non-linear relationships between soil properties and spectra as well as increasing variances of soil properties that lead to larger prediction errors (Nocita et al., 2014; Stenberg et al., 2010; Stevens et al., 2013).

With current, e.g., PRISMA (Loizzo et al., 2018), and upcoming hyperspectral spaceborne missions, e.g., EnMAP (Guanter et al., 2015) and SHALOM (Feingersh and Ben-Dor., 2015), the quantification of soil properties on larger scales comes into reach. These satellites will have the potential to periodically update existing SOC maps in bare soil areas that currently can be surveyed only with low spectral resolution satellites or where SOC estimations are often based on outdated point-wise information (Sanchez et al., 2009). In preparation for these upcoming new data from spaceborne sensors, currently SOC modelling approaches are looking at the potential of large-scale soil spectral libraries to be used as an alternative to local ground databases. The overall aim of these new approaches is to build soil prediction models that can be applied universally on large scales to become more independent of local ground truth data that are currently needed for model calibration. Therefore, we use the European LUCAS topsoil database (Land Use/Land Cover Area Frame Survey) (Tóth et al., 2013) as a basis to develop general, robust prediction models for the quantification of SOC. Previous work done by Stevens et al. (2013) which was based on the LUCAS database to develop SOC prediction models using PLSR modelling and dividing the database according to land cover types, obtained a good prediction accuracy (RMSE of 4.9 g kg^{-1}). Similarly, Nocita et al. (2014) used a local PLSR approach with the LUCAS database, so that locally the relationship between a soil property and spectral data can be stable, allowing for linear modelling (Ramirez-Lopez et al., 2013).

In this paper, we intend to adapt and expand existing approaches in preparation for future SOC estimation from spaceborne sensors. For this, we investigate the accuracy of SOC predictions using the LUCAS soil spectral database and considering a remote sensing adapted approach where (i) the LUCAS database is spectrally reduced to the wavelengths that can be used from

spaceborne sensors, cutting out the larger atmospheric water bands, (ii) the LUCAS database is clustered solely based on spectral data avoiding the use of any geochemical soil information, and (iii) the model input for SOC predictions also consists of spectral data only without using other soil properties as covariates. Our analyses focus on the comparison of different spectral clustering approaches in combination with PLSR modelling. The objective is to investigate whether spectral clustering has the potential to group the large soil spectral database LUCAS in such a way that the links between SOC and spectral data become approximately linear, and would therefore improve prediction accuracies compared to models that were built based on the non-clustered database.

2. Material and Methods

2.1 LUCAS soil database

This study is based on the pan-European Land Use/Land Cover Area Frame Survey (LUCAS) 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 (Orgiazzi et al., 2017; Tóth et al., 2013). LUCAS is the first attempt to build a consistent soil database to support policy making. The sampling for this survey took place in 2009 in 25 Member States of the European Union and includes 19,967 top-soil samples (0-20 cm) collected on different land use types. The database consists of 12 different soil properties, including SOC as well as spectral measurements in the VNIR-SWIR range. A particular advantage of the LUCAS database is that all physical and chemical as well as spectral measurements have been conducted using harmonized standards and protocols (Tóth et al., 2013). The SOC content has been measured by dry combustion using a vario Max CN Analyzer (Elementar Analysensysteme GmbH, Germany). Before taking spectral measurements, the samples were dried at 40°C, crushed and sieved (< 2 mm). The absorbance spectra were measured using a FOSS XDS Rapid

Content Analyzer within a range of 400.0-2499.5 nm with a spectral resolution of 0.5 nm, resulting in 4200 wavelengths (Tóth et al., 2013).

To exclude the ranges of strong atmospheric attenuation that are not useful in remote sensing analyses, we excluded the spectral ranges of strong water absorptions around 1400 nm and 1900 nm, precisely we excluded 1350-1500 nm and 1800-1950 nm. Furthermore, as observed by Stevens et al. (2013), the spectral range 400-500 nm shows instrumental artefacts and was removed from further analyses.

Also, we subset the LUCAS database to agricultural areas based on land use and land cover classes provided within the database. We focused on agricultural areas as these areas are temporarily free of vegetation and can therefore be used for subsequent mapping of soil properties from air- and spaceborne platforms.

2.2 Database pre-processing

In several studies the 1st derivative led to best modelling results (e.g. Araújo et al., 2014; Nocita et al., 2014; Stevens et al., 2013). Thus, we used the 1st derivative of the absorbance spectra after applying a Savitzky-Golay smoothing (Savitzky and Golay, 1964) filter using a 2nd order polynomial and a window size of 41 bands which corresponds to 20.5 nm.

The distribution of the SOC content in the agricultural subset is highly skewed (skewness = 4.64), so we transformed it to approximately normally distributed values using the natural logarithm (new skewness = 0.12). Subsequently, the dataset was divided into subsets for calibration (70%) and validation (30%) using the Kennard-Stone algorithm (Kennard and Stone, 1969). This algorithm chooses samples based on a distance measure to produce representative subsets. Clustering and model calibration is solely based on the calibration subset, and the validation subset is only used to assess clustering and model quality.

2.3 Methodological overview

In this study, we tested two different clustering-modelling approaches (Fig. 1): (i) the k-means algorithm was used based on either a Principal Component Analyses (PCA; Fig. 1, B1) or Spectral Feature Variables (SFV; Fig. 1, B2), then SOC predictions were performed on each spectral cluster using PLSR; (ii) a local PLSR approach (Fig.1, C) was used where for each validation sample a separate PLSR model was calibrated on the basis of a set of most similar calibration samples that was selected based on distance metrics. Afterwards, we compared the SOC prediction accuracies obtained by the different approaches with a reference model (Fig. 1, A) which was calibrated based on the complete database without previous clustering. The reference model was used to investigate the performances of the clustering approaches in improving the model accuracy. For reasons of comparability all models were calibrated and validated on exactly the same LUCAS subsets. The detailed workflow for each of the two clustering-modelling approaches is given in Fig. 2.

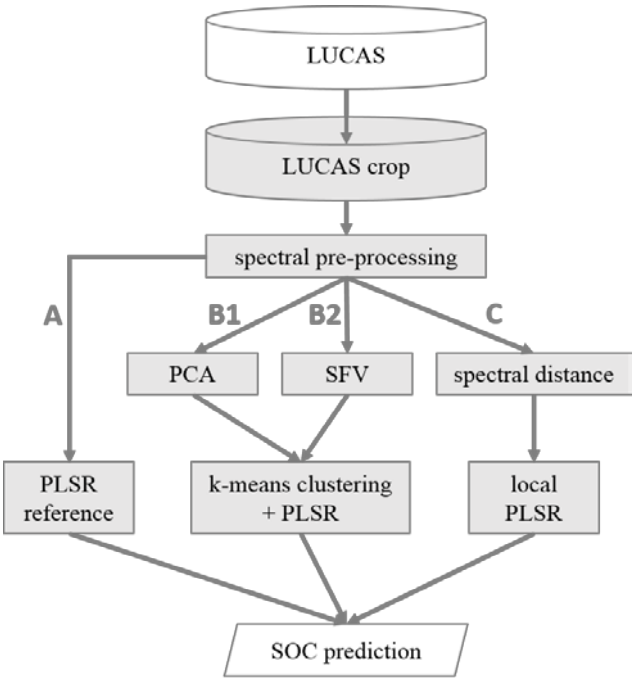
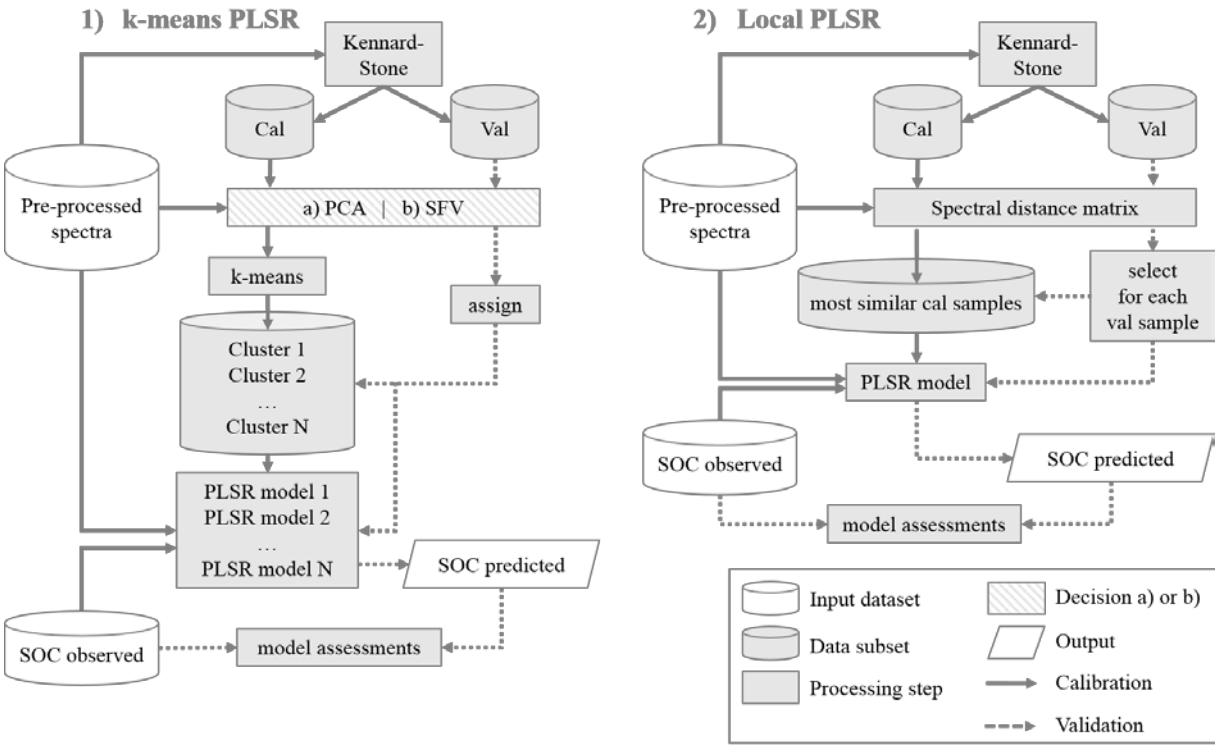


Fig. 1: Overview of general processing structure.



191

192 Fig. 2: Detailed processing overview of the two clustering-modelling approaches: 1) k-means
193 clustering based either on a) PCA or b) SFV and PLSR (left), and 2) local PLSR (right).

194 2.4 Reference model without clustering using PLSR

195 As initial stage, a reference SOC prediction model was built based on our whole agricultural
196 and spectrally-reduced LUCAS dataset, using PLSR (Fig. 1, A). We applied the R package pls
197 (Mevik et al., 2016) for PLSR analyses. PLSR is suitable for data that consist of a matrix of
198 many highly collinear predictor variables X that are used to predict the response variable(s) Y.
199 Both X and Y are projected into a new dataspace in such a way that the covariance between X
200 and Y is maximized. A few orthogonal regression coefficients, called latent variables (LV) are
201 then used as predictors for Y. The number of LV is unknown and needs to be determined (Wold
202 et al., 2001). Here we chose a combination of three common methods to achieve good model
203 accuracies without over-fitting the models to the data. The results of the three methods were
204 averaged to automatically select the best number of latent variables which leads to better
205 validation accuracies than just using one of the methods. (i) We used a 10-fold cross-validation

206 to estimate the root mean squared error (RMSE) for different numbers of LV and chose the
207 minimal number of LV within one standard deviation of the minimal RMSE (comparable to
208 Stevens et al. (2013)). (ii) We used the commonly applied adjusted coefficient of determination
209 ($adj.\ R^2$, see Eq. (1)) which takes into account the number of components used in a model. (iii)
210 We used the adjusted Wold's R with a threshold of > 0.95 (following Li et al., 2002). It is based
211 on the ratio of the predicted error sum of squares of the PLSR LV m+1 and the LV m, with m
212 as the number of LV. The additional LV m+1 will only be included in the PLSR model if it
213 provides significantly better predictions.

214 $$adj.R^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1) \quad (1)$$

215 with n as the number of samples and k as the number model components.

216 2.5 Method 1: k-means clustering and PLSR

217 In the first spectral clustering approach (Fig. 1, B1 & B2 and Fig. 2, left), we used the k-means
218 algorithm to cluster the data prior to applying the PLSR algorithm to each spectral cluster. K-
219 means starts with randomly selected initial cluster centres and assigns the closest samples to
220 these centres. Based on these clusters it calculates new cluster centres and reassigns all samples.
221 This step is repeated until the algorithm converges (Hartigan, 1975). To remove noise, reduce
222 collinearity and to increase the computational speed, we tested the performance of k-means
223 clustering for two independent spectral reduction methods: (a) based on spectral variance using
224 Principal Component Analysis (PCA) and; (b) based on the direct analyses of spectral features
225 using a set of Spectral Feature Variables (SFV) following Bayer et al. (2012). The spectral
226 reduction methods are used for the clustering processes only, whereas the PLSR models are
227 calibrated on the pre-processed spectra. The k-means algorithm demands the number of clusters
228 as an input and here we based this choice on the best PLSR model validation results. Therefore,
229 we tested different numbers of clusters between 2 and 15.

The PCA method focuses on the reduction of spectral variance in the data based on the projection of the dataspace in the principal component bands ordered in terms of decreasing variance. The SFV method focuses on the physical analyses of spectral shape and characteristic absorption bands directly linked to soil chromophores (e.g. Ben-Dor et al., 2009). Although less commonly used as spectral reduction, the SFV method presents the advantage that it is based solely on the direct analyses of spectral features related to soil properties, and carries different information than spectral variance.

The following procedure was adopted for both sets of clusters independently: for each cluster a separate PLSR model was calibrated based on the calibration dataset of the pre-processed spectra. As the clustering process was based solely on the calibration subset, each validation sample had to be assigned to one of those clusters. This was conducted using the shortest distance to the cluster centres in the multidimensional PCA- resp. SFV-dataspace. Therefore, it was necessary to also calculate the Principal Components (PCs) of the PCA resp. SFV for the validation samples. As the PCA was solely calculated based on the calibration subset, the PCs for the validation samples were predicted using the same dataspace transformation. To validate this clustering-modelling approach, for each validation sample the PLSR model was applied, that is corresponding to the cluster to which the sample had previously been assigned.

For the PCA, we used the first 20 PCs that explained more than 99.5% of the spectral variance. For the SFV approach, we used an expanded selection of SFV following Bayer et al. (2012), focusing on spectral features associated with main soil chromophores such as SOC, clay, iron oxides, carbonates and gypsum. Three types of SFV are considered, as shown in Fig. 3: absorption features (AF), curve features (CF), and hull features (HF), associated with diagnostic spectral absorptions, spectral shapes, and spectral continuum. We adapted the approach of Bayer et al. (2012) and used five AF, one CF and two HF.

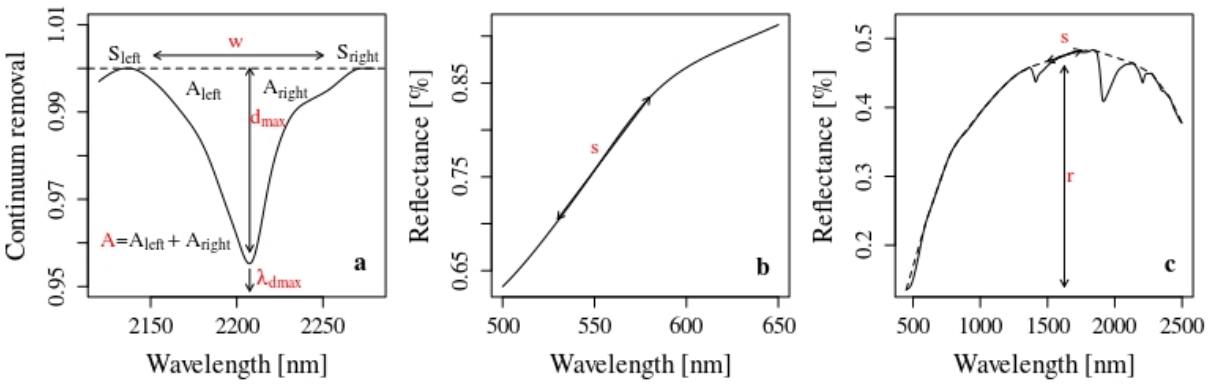


Fig. 3: Overview of SFV used in this study. **a** absorption features (AF), **b** curve feature (CF) and **c** hull features (HF) shown at the example of a mean spectrum of the LUCAS database; variables potentially used in the analysis are marked in red; modified after Bayer et al. (2012).

As AF we calculated the maximum depth of the absorption feature (d_{max}) and the corresponding wavelength (λ_{dmax}), the width between the shoulders of the feature (w), the area of the feature ($A = A_{left} + A_{right}$) and the asymmetry of the feature ($AS = A_{left} / A_{right}$; AS is not shown in Fig. 3a). Therefore, we calculated the continuum removal (CR) of each feature's spectral range (Table 1) and searched for the minimum to determine d_{max} and λ_{dmax}. To detect the left / right shoulder of each feature, we searched for the last / first wavelength left / right of the maximum absorption (λ_{dmax}) which lies on the convex hull (CR = 1). The SFV width w is the difference in wavelengths between the two shoulders. To calculate the left and right area of the feature A_{left} / A_{right}, the area under the curve (function auc from R package flux (Jurasinski et al., 2014)) is subtracted from the total area of the corresponding side of the feature. The total area is the sum of the area below and above the curve within zero and one and within the wavelengths of the corresponding shoulder and the maximum absorption.

The CF was calculated based on a line fit of the reflectance values within the spectral range under study. This line fit was used to derive the mean slope (s).

For the calculation of the HF, we used the continuum removal of the reflectance spectra. Based on a line fit of the convex hull within the spectral range under study, the mean slope (s) and mean reflectance (r) were calculated. The line fit was based on the reflectance values of those points lying on the convex hull (CR = 1) within the spectral range. Additionally, the range of points used to calculate the line fit was extended to the first points on the left and right side of the spectral range if possible, to account for changes at the margins of the spectral range. Based on this line fit the reflectance values of the two bordering wavelengths of the spectral range were predicted and used as a basis to calculate the two SFV. Thereby, the mean slope is the difference in reflectance of the two bordering wavelengths divided by the difference in wavelengths, and the mean reflectance is the mean value of the two bordering wavelengths.

The SFV were calculated for each significant spectral range separately which were taken from the literature. We included spectral absorption features of several spectrally important soil properties in our calculations of SFV as they have primary correlations to spectral absorptions (Stenberg et al., 2010). As some of the spectral absorptions used by Bayer et al. (2012) are very similar, we selected the more unique ones (see Table 1). Prior to usage, we normalized the SFV by subtracting the mean and dividing by the standard deviation. We also checked the SFV for constant values (standard deviation divided by mean < 0.001), redundant variables and variables very highly correlated to other variables ($r > 0.9$), and removed them.

Table 1: Spectral absorptions based on spectrally active soil properties which are initially used to calculate SFV. *Numbers in brackets are the original values from (Bayer et al., 2012) which are adapted due to removed wavelengths at water bands and at 400-500 nm.

Name	Type	Range *	Associated soil property	References
SFV1	AF	1600, 1799.5 (1815)	SOC	(e.g. Ben-Dor et al., 1997; Viscarra Rossel and Behrens, 2010)
SFV2	AF	2240, 2410	SOC, clay	(e.g. Ben-Dor et al., 1997; Viscarra Rossel and Behrens, 2010)
SFV3	HF	(450) 500, 740	SOC, clay, iron	(e.g. Bartholomeus et al., 2008; Baumgardner et al., 1986; Hill and Schütt, 2000)
SFV4	HF	(1460) 1500, 1750	SOC, clay	(e.g. Bartholomeus et al., 2008; Baumgardner et al., 1986; Hill and Schütt, 2000)
SFV5	AF	(450) 500, 680	iron	(e.g. Grove et al., 1992; Hunt, 1970; Viscarra Rossel and Behrens, 2010)
SFV6	AF	580, 800	iron	(e.g. Grove et al., 1992; Viscarra Rossel and Behrens, 2010)
SFV7	AF	750, 1300	iron	(e.g. Ben-Dor and Banin, 1994; Clark, 1999; Viscarra Rossel and Behrens, 2010)
SFV8	CF	550, 590	iron	(e.g. Clark, 1999)
SFV9	AF	2100, 2290	clay	(e.g. Chabrillat et al., 2002; Viscarra Rossel and Behrens, 2010)
SFV10	AF	2300, 2400	carbonate	(e.g. Gaffey, 1987)
SFV11	AF	1690, 1800	gypsum	(e.g. Milewski et al., 2018)

2.6 Method 2: local PLSR

Locally weighted PLSR models belong to the memory-based learning approaches which can outperform machine learning algorithms such as artificial neural networks and decision trees (Ramirez-Lopez et al., 2013). Basically, the local PLSR approach (Fig. 1, C and Fig. 2, right) selects a set of samples (nearest neighbours) out of a calibration database which are spectrally most similar to a new sample, and this set of nearest neighbours is then used to calibrate a prediction model for the new sample (Ramirez-Lopez et al., 2013). The process is repeated for each validation sample. This approach can be thought of as a kind of adaptive clustering because it creates tailor-made calibration sets for each new sample. It has not been applied in soil spectroscopy very often (Ramirez-Lopez et al., 2013) but recently Nocita et al. (2014) used it

for SOC estimations in the LUCAS database and showed that it is a promising approach. This is consistent with the observation that in the past the PLSR approach has shown to be very promising and delivering high performance models in the soil spectroscopy and remote sensing community for the prediction of SOC content when it was applied on local scale, which is often associated with spectrally similar signatures. Nocita et al. (2014) first divided their cropland database in mineral and organic soils, based on chemical data, and obtained the best results for the mineral soils using the 250 nearest neighbours with the pls distance as spectral distance measure. The pls distance is based on the Euclidean distance of the scores of the PLSR which are relating SOC content and the spectra and therefore requires prior knowledge not only of the spectra but additionally of the SOC content. Furthermore, they used sand content as auxiliary distance measure as it was improving the results.

We adapted this approach to fit our study by testing other spectral distance measures and avoiding the use of auxiliary variables as we aim to develop an approach that is based on spectral data only. Also, a pre-clustering based on chemical data is thus not applied. As we have removed the water bands from the spectra and therefore have a differing spectral coverage, we also tested a sequence of fixed numbers of nearest neighbours. Additionally, we investigated if applying a sequence of thresholds within the distance measure instead of using a fixed number of nearest neighbours can improve the results.

In order to find a suitable distance measure as a basis for the local PLSR approach, we tested four different measures. The pls distance (plsDist), as used by Nocita et al. (2014) is not suitable for the basic idea of our study which is to use spectral information as input only. It additionally requires knowledge of the SOC content, but is applied here for reasons of comparability. The correlation distance (corDist) is based on the correlation coefficient between two spectra which is subtracted from 1. Here we used the corDist function available in the MKmisc package in R (Kohl, 2018). The Mahalanobis distance (MDist) and the spectral angle mapper (SAM), which

is the angle between vectors in the hyperspectral space, were calculated using the fDiss function in the resemble package in R (Ramirez-Lopez and Stevens, 2016). For all spectral measures, besides the pls distance, we used the first 20 PCs of a PCA based on pre-processed spectra as input to remove noise, reduce collinearity and to increase the computational speed. The pls distance was calculated directly on pre-processed spectra.

2.7 Application to simulated EnMAP spectra

Simulated EnMAP spectra were produced based on the LUCAS database. Therefore, the agricultural subset of LUCAS spectra was resampled to EnMAP's spectral resolution using the spectralResampling function in the hsdar package in R (Lehnert et al., 2017). EnMAP is designed to measure in the 420-2450 nm range with more than 240 bands. It consists of two spectrometers that have a spectral overlap between 900 and 1000 nm (Segl et al., 2010). Here we excluded the bands of the first spectrometer within the overlapping range. The same pre-processing and processing steps were performed as for the original LUCAS resolution. The following water bands were removed: 1358.50-1499.40 nm and 1803.50-1951.00 nm and 16 PCs explaining more than 99% of the spectral variance were used. The best modelling approach found in this study was applied to the simulated EnMAP spectra.

2.8 Model assessments

To assess the model accuracy, the ln-transformed SOC values (measured and predicted) were used for dimensionless measures, whereas for measures with units (g kg^{-1}) original SOC values and back-transformed predicted values were used. As performance indicators, we calculated the coefficient of determination (R^2) (Eq. 2), the root mean squared error of prediction (RMSEP, Eq. 3), the relative RMSEP (rRMSEP) (Eq. 4), the ratio of performance to deviation (RPD) (Eq. 5), the ratio of performance to interquartile range (RPIQ, Eq. 6) and the bias (Eq. 7), (following e.g. Nocita et al., 2014; Steinberg et al., 2016; Stevens et al., 2013):

$$R^2 = 1 - \sum_{i=1}^{n}(yp_i - yo_i)^2 \,/\, \sum_{i=1}^{n}(yo_i - \overline{yo})^2 \quad (2)$$

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

$$rRMSEP = 100 * RMSEP/\overline{yo} \quad (4)$$

$$RPD = sd(yo)/RMSEP \quad (5)$$

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

$$bias = 1/n * \sum_{i=1}^{n}(yp_i - yo_i)^2 \quad (7)$$

with yo_i being the observed SOC value of sample i and yp_i being the predicted SOC value of sample i. $\overline{yo}$ is the mean of the observed SOC values, n is the number of samples, sd is the standard deviation and IQ is the interquartile range. The rRMSEP, RPD and RPIQ are ways to standardize the RMSEP to be able to compare datasets and clusters with different ranges and variances (Nocita et al., 2014).

3. Results

3.1 LUCAS database and pre-processing

We subset the LUCAS database to agricultural areas which reduced the number of samples to 8294. This subset contains mainly mineral soils as well as 41 samples classified as organic soils. Within the selected LUCAS subset, the SOC content ranges between 0 – 193.9 g kg^{-1} with a mean value of 17.5 g kg^{-1}. The percentage of clay goes up to 79% with 22% on average. The $CaCO_3$ content varies between 0 – 882 g kg^{-1} with a mean of 85 g kg^{-1}. The spectra show a large variation in absorbance due to the influence of SOC content and mineralogical composition (Fig. 4).

The PLSR reference model without clustering, which was calculated to assess the potential of improvement of the clustering approaches, led to an R^2 of 0.59, RPIQ of 1.76 and RMSEP of 7.37 g kg^{-1}.

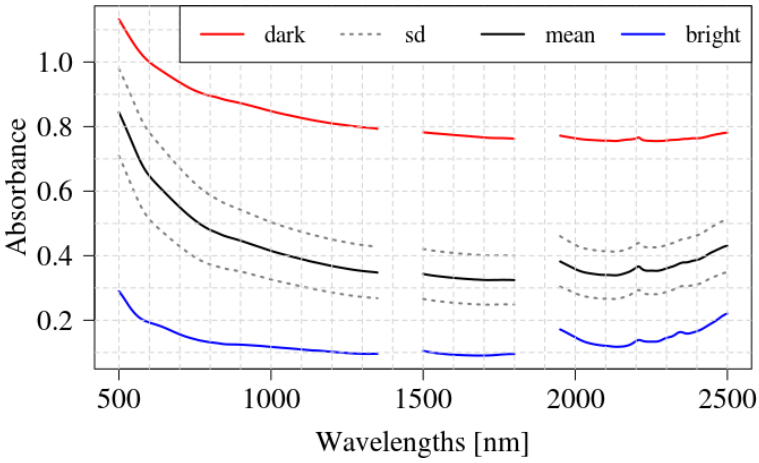


Fig. 4: Spectral variability in the LUCAS agricultural areas subset showing mean and standard deviation, as well as the darkest and the brightest spectrum.

3.2 k-means clustering and PLSR

All validation results for the two clustering approaches are shown in Table 2. The k-means PCA approach resulted in seven clusters with calibration sizes ranging from 249 to 1391 samples. In addition, there was one very small cluster with 23 calibration samples and as only two validation samples were assigned to this cluster it was not included in the final assessment. Very variable results depending on the spectral clusters were achieved, ranging from very poor (R^2 = 0.32, pca6) to very good (R^2 = 0.84, pca2). Except for these two spectral clusters and another one with poor performance (R^2 = 0.45, pca1), in general a medium performance is achieved with R^2 between 0.54-0.64 in all five other spectral clusters. The RPD values underline this statement with values above 1.4 for the aforementioned five clusters and below 1.4 for the two models with a poor performance. All clusters, except the excluded small one, showed a highly skewed SOC distribution with skewness values above one reaching to a maximum of 5.8. This underlines the need for SOC normalization prior to modelling which was done here.

For the k-means SFV approach some of the SFV were removed due to high correlations ($r > 0.9$) which led to a basis of 33 SFV. All SFV11 variables were removed as they showed high correlations to the SFV1 variables. The AF variable A was always highly correlated to dmax

and was therefore excluded from all spectral ranges except for the SFV10 range. Here dmax was removed instead of A, as dmax of SFV10 was highly correlated to dmax of SFV2.

The best results for the k-means SFV approach were achieved for four clusters with calibration cluster sizes ranging from 218 to 2526 samples. In addition, there was again a very small cluster of 22 calibration samples with only two assigned validation samples which was excluded. The rest of the samples was distributed mainly to two large clusters (sfv3, sfv5). The R^2 values of most clusters show a medium prediction performance of above 0.5 with one exception showing a good accuracy with a R^2 of 0.7 (sfv1). This cluster with the best performance also concerning RPIQ values includes comparably more samples with higher SOC values which also shows in a higher standard deviation. All RPD values are above 1.4 which indicates fair models or above 1.8 which indicates good models. The original SOC values of all clusters were highly skewed with a maximum skewness of 4.3.

Fig. 5 shows the mean reflectance and SOC range for each spectral cluster. The mean spectra of the SFV clusters are more spectrally differentiated than when based on PCA. They show differences in albedo and in absorption features, with the smallest cluster sfv4 showing the brightest mean spectrum and cluster sfv1 the darkest, related to the highest SOC values within this cluster. For the PCA clustering, that is based on spectral variance and less focused on spectral features, there are only marginal differences in spectra and SOC content between the clusters. For the SFV clusters the geographical distribution reveals spatial patterns (map not shown): the two largest clusters sfv3 and sfv5 have a tendency to be located towards the north resp. south of Europe. The cluster including high SOC contents (sfv1) is mainly located in northern Germany and Denmark, and the very small cluster (sfv4) is spatially restricted to Spain. For the PCA clusters no spatial patterns were visible (map not shown).

Table 2: Validation results for the k-means clustering approaches combined with PLSR; Nval (= number of validation samples in cluster), Ncal (= number of calibration samples in cluster), and LV (= latent variables)

* validation results using combined predicted values from all clusters but pca3, resp. sfv4; for the column LV the mean values are calculated

	Model performance						**Model data**				
	R^2	**RMSEP [g kg^{-1}]**	**rRMSEP**	**RPD**	**RPIQ**	**Bias**	**LV**	**Nval**	**Ncal**	**SD [g kg^{-1}]**	**SOC range [g kg^{-1}]**
PCA*	**0.60**	**8.48**	**52.9**	**1.60**	**1.80**	**-0.42**	**12.1**	**2488**	**5806**	**13.4**	**0-193.9**
pca1	0.44	5.80	37.6	1.34	2.00	0.28	9	336	934	11.4	0-121.5
pca2	0.84	17.76	50.6	2.52	3.25	-2.95	17	110	881	26.0	3.8-193.9
pca3	NA	NA	NA	NA	NA	NA	3	2	23	6.0	2.5-28.2
pca4	0.61	3.50	24.5	1.61	1.93	-0.64	14	685	1131	6.6	0-55.3
pca5	0.57	12.09	71.1	1.53	1.85	0.03	19	735	1391	9.9	0-113.4
pca6	0.32	5.68	52.1	1.22	1.23	-0.64	9	97	249	9.2	0-65.6
pca7	0.64	4.55	32.8	1.67	2.06	-0.44	16	156	520	6.9	2.5-62.2
pca8	0.54	5.41	37.1	1.47	1.96	-0.70	10	367	677	10.1	0-84.2
SFV*	**0.63**	**6.68**	**41.6**	**1.64**	**1.85**	**-0.58**	**11.6**	**2488**	**5806**	**13.4**	**0-193.9**
sfv1	0.70	25.27	56.49	1.85	3.03	-4.26	12	87	633	29.7	5.6-193.9
sfv2	0.54	5.04	32.05	1.49	1.45	-0.31	9	72	218	10.4	2.9-121.4
sfv3	0.58	5.21	31.89	1.54	2.04	-0.36	19	1051	2526	9.8	2.3-93.6
sfv4	NA	NA	NA	NA	NA	NA	2	2	22	6.1	2.5-28.2
sfv5	0.56	4.41	31.85	1.51	1.69	-0.50	16	1276	2407	7.2	0-66.3

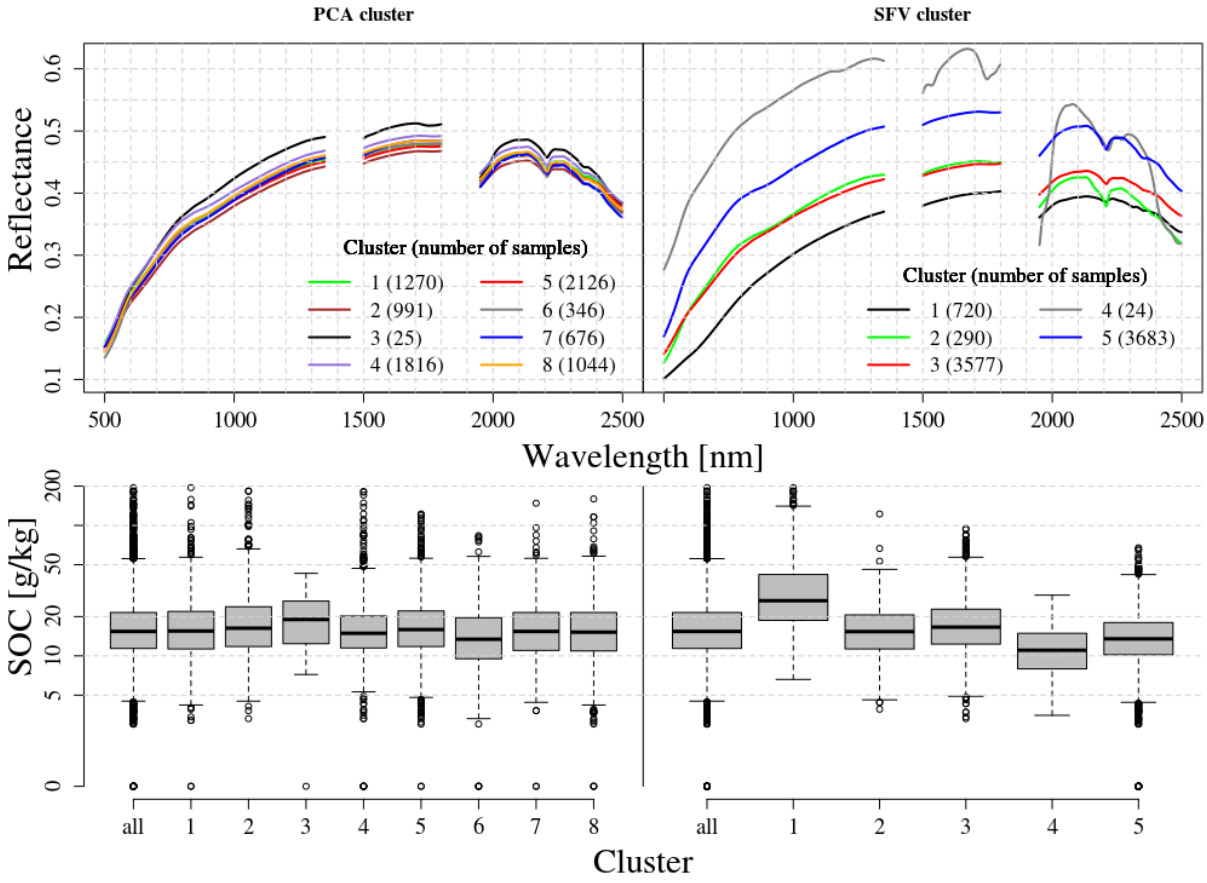


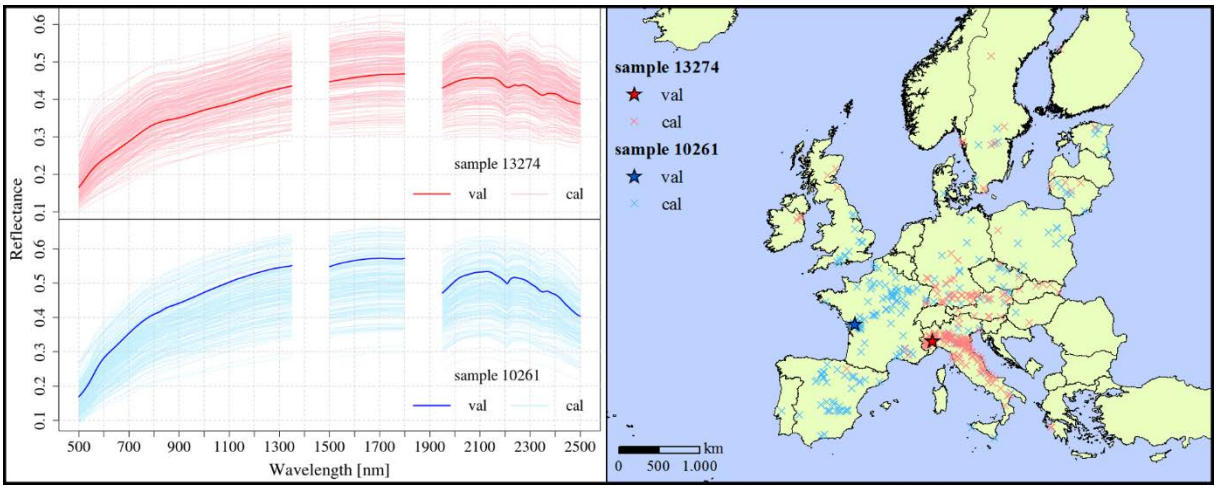
Fig. 5: Mean spectra of the clusters (top) showing the PCA based (left) and the SFV based (right) approaches. For clarity these spectra are shown in reflectance. Numbers in brackets are number of samples within each cluster including calibration and validation samples. Boxplots showing the SOC distribution (bottom) within the whole LUCAS subset (all) and within the clusters of the PCA approach (left) and the SFV approach (right). The SOC content is shown on a logarithmic scale.

3.3 Local PLSR

Fig. 6 is an illustration of the local PLSR approach showing two examples for the validation samples. Sample 10261 contains more SOC (37.7 g kg^{-1}), clay (36%) and $CaCO_3$ (431 g kg^{-1}) compared to sample 13274 (SOC: 13.1 g kg^{-1}, clay: 6%, $CaCO_3$: 36 g kg^{-1}. We compared four different distance measures that could potentially be used as a basis for the local PLSR approach: correlation distance (corDist), Mahalanobis distance (MDist), pls distance (plsDist),

spectral angle mapper (SAM). For each validation sample a fixed number of calibration samples was used to calibrate the model. To determine the best fixed number of calibration samples we selected 30% of the calibration dataset as test validation set, using the Kennard-Stone algorithm (Kennard and Stone, 1969), and iteratively tested different numbers. We chose the same number for all distance measures to have a fair comparison. The best compromise was 450 samples. All distance measures led to test results in a comparable range as shown in Fig. 7. The MDist attains the lowest RMSEP and the lowest bias, whereas the plsDist reaches the highest R^2 and RPIQ values. As plsDist does not fit to the basic ideas of this study, MDist was chosen as adequate distance measure.

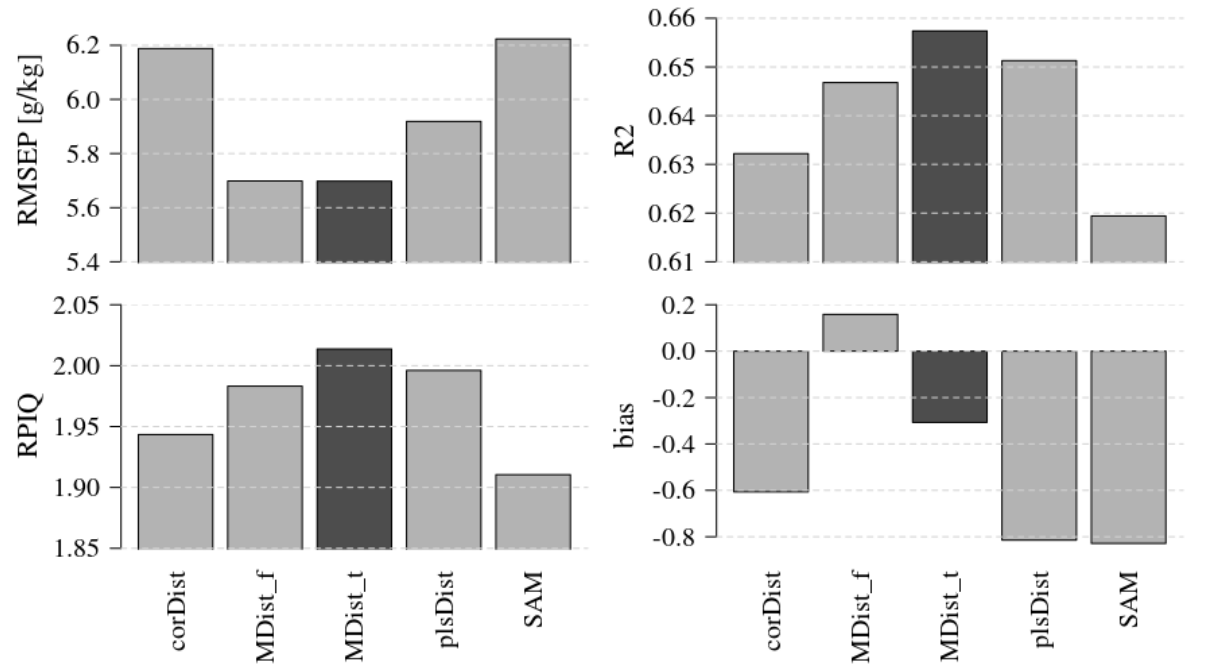
In a next step we tested if the usage of a threshold within the distance measure to define the calibration datasets could improve the results. The advantage of this approach is that we abstain from using the same fixed number of calibration samples for each validation sample but allow for a larger number of samples in the calibration subsets. We used a minimum size of calibration samples of 200 to ensure that enough samples were used for model calibration. Here we tested different sequences of thresholds and again chose the threshold which led to the best results within the test set. For the MDist a threshold of 0.19 could improve the test results for R^2 and RPIQ (Fig. 7). Applying the local PLSR approach using MDist with threshold to the validation dataset we were able to calibrate good prediction models with $R^2 = 0.67$, RMSEP = 5.16 g kg^{-1}, RPD = 1.74, RPIQ = 1.96 and a low bias = 0.1.



462

463 Fig. 6: Two examples of the local PLSR approach showing two validation samples (val) and

464 their calibration samples (cal): reflectance spectra (left) and geographical distribution (right).



465

466 Fig. 7: Barplots comparing the model quality of different distance measures tested in the local

467 PLSR approach based on an independent test set: correlation distance (corDist), Mahalanobis

468 distance (MDist), pls distance (plsDist), spectral angle mapper (SAM). Distance measures are

469 based on a fixed number of calibration samples (light grey), and on a threshold in the distance

470 measure of the MDist result (MDist_t, dark grey).

3.4 Overall results

The overall results in Table 3 show that the two k-means clustering approaches could slightly improve the model accuracy in terms of R^2 and RPD compared to the reference model. The k-means SFV approach could also improve the RMSEP compared to the PLSR reference. The overall best results were achieved by the local PLSR approach. It was able to improve the prediction accuracy visible in all model parameters, e.g. the RMSEP could be reduced by more than 2 g kg^{-1} compared to the reference. Regarding Fig. 8 and corresponding to the previous results, the best fit is achieved by the local PLSR approach which shows the highest correlation of 0.9 between observed and predicted SOC values.

Table 3: Overall validation results for the reference model and the clustering approaches (k-means and local PLSR). LV = latent variables, Nval = number of validation samples, Ncal = number of calibration samples. For the k-means approaches the model performance parameters are the validation results using combined predicted values from all clusters but pca3, resp. sfv4 and the LV are averages.

	Model performance						**Model data**		
	R^2	**RMSEP [g kg-1]**	**rRMSEP**	**RPD**	**RPIQ**	**Bias**	**LV**	**Nval**	**Ncal**
PLSR reference	0.59	7.37	45.9	1.56	1.76	-0.70	19	2488	5806
k-means PCA	0.60	8.48	52.9	1.60	1.80	-0.42	12.1	2488	5806
k-means SFV	0.63	6.68	41.6	1.64	1.85	-0.58	11.6	2488	5806
Local PLSR	0.67	5.16	32.2	1.74	1.96	0.10	12.8	2488	5806

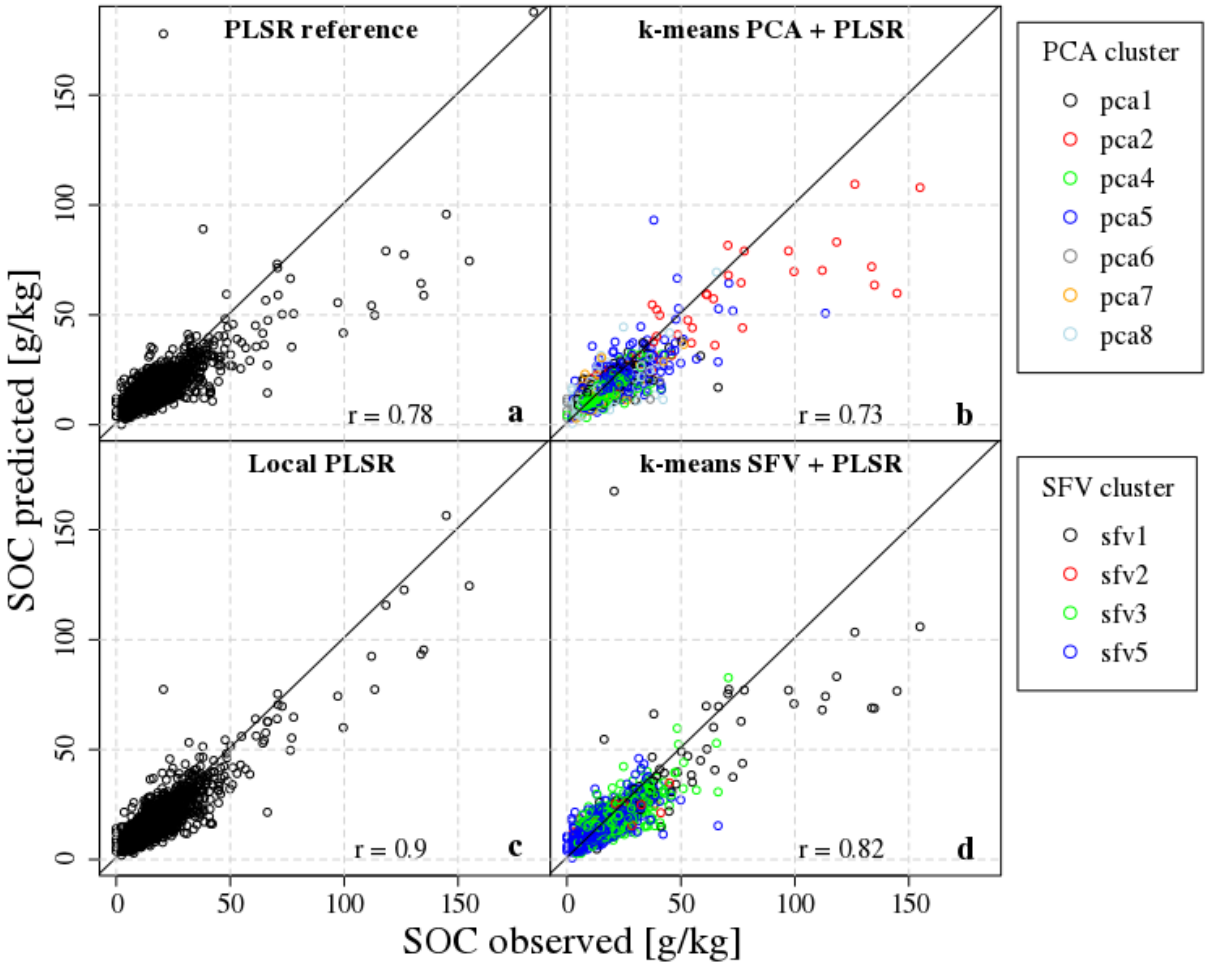


Fig. 8: Observed vs. predicted SOC values of the validation samples for the PLSR reference (a) and the clustering approaches (b-d). The colours in b and d represent the seven respectively four k-means clusters. Pearson's correlation coefficient r is given. Notice: outlier are not shown in b (183/263, 20.7/299), c (183/222) and d (183/232).

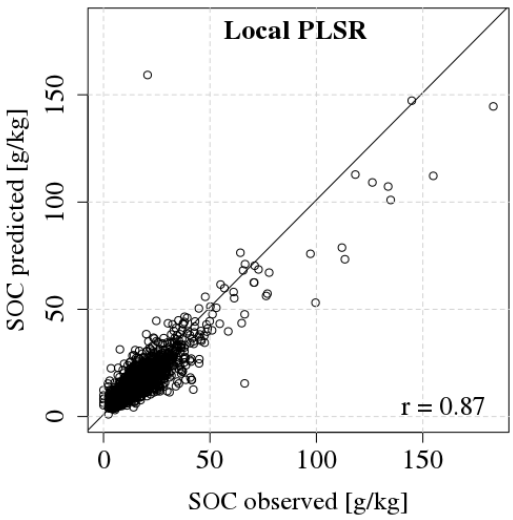
3.5 Simulated EnMAP spectra

Best results were delivered by the local PLSR and consequently this approach was applied to the simulated EnMAP dataset using the best configurations investigated in the previous steps, using the threshold in the Mahalanobis distance. The results in Table 4 and Fig. 9 show that the validation results using the simulated EnMAP spectra decrease only slightly compared to the full spectral LUCAS resolution.

498 Table 4: Validation results for the local PLSR applied to simulated EnMAP spectra. LV = latent
499 variables, Nval = number of validation samples, Ncal = number of calibration samples

	Model performance						Model data		
	R^2	**RMSEP [g kg-1]**	**rRMSEP**	**RPD**	**RPIQ**	**Bias**	**LV**	**Nval**	**Ncal**
Local PLSR	0.66	5.78	36.0	1.71	1.93	-0.20	11.9	2488	5806

500



501

502 Fig. 9: Observed vs. predicted SOC values of the validation samples for the local PLSR
503 approach applied to simulated EnMAP spectra.

504 **4. Discussion**

505 4.1 K-means clustering and PLSR

506 We applied different clustering techniques to investigate whether they were suitable to improve
507 the prediction accuracy of a reference model that was based on the whole non-clustered dataset.
508 The k-means clustering approach was tested for two different versions. Both of them could
509 improve the overall model accuracy compared to the reference. Araújo et al. (2014) come to
510 the conclusion that a k-means clustering is able to increase the organic matter (OM) prediction
511 results compared to a reference PLSR model as the former can cope with non-linearity in large
512 and heterogeneous datasets. In their study they compared the clustering results with boosted

regression trees and support vector machines as reference models and found out that they performed in the same range. Thus, a k-means clustering combined with separate PLSR models seems to be able to improve a PLSR reference model similar to our observation. In our study, especially using a set of SFV for the basis of the k-means clustering could improve the modelling results. A preliminary set of SFV was selected based on the approach of Bayer et al. (2012) focusing on SOC, clay and iron contents. In the LUCAS database, a much higher heterogeneity in the spectral database is observed compared to the spectral heterogeneity of the spectral data from Bayer et al. (2012), and therefore, we added SFV based on the features of carbonates and gypsum.

The first SFV cluster showed the highest RMSEP of 25.3 g kg^{-1} and a very high bias although it shows the best modelling performance with the highest R^2, RPD and RPIQ values (Table 2). This cluster has a very high mean SOC value of 35.4 g kg^{-1} (the second highest mean SOC value is 17.9 g kg^{-1} in SFV cluster 3) and a high standard deviation. This result is conform to Stenberg et al. (2010) who stated that the prediction errors of spectroscopic models increase with increasing standard deviation of the predicted soil property. Therefore, it is important to consider the distribution of SOC values when comparing the RMSEP of different study sites or clusters. The RPD, RPIQ or the rRMSEP are better suited as they account for different ranges and variances.

The k-means clustering based on SFV resulted in more differentiated clusters compared to the PCA approach. The SFV clusters showed differences between their mean spectra in terms of albedo and spectral features (Fig. 5). The SOC distributions including mean values differ between the SFV clusters (Fig. 5) and the SOC standard deviation decreased for most of the SFV clusters compared to the LUCAS dataset, which is conform to the findings by Araújo et al. (2014) who also observed this behaviour for many of their clusters. Additionally, there are slight patterns visible in the spatial distribution of the SFV clusters which is confirming Stevens

et al. (2013) who stated that the link between soil properties and their spectra can be very complex and that it can vary in space.

The number of LV in a PLSR model is an essential component leading to very different model results. We advise to include the number of LV in the results of future studies using PLSR to make results more comparable and transparent.

4.2 Local PLSR

We show in this paper that the local PLSR approach significantly improved the SOC modelling results compared to the reference PLSR model. It showed an increase in model accuracy relative to the reference PLSR of +14% R^2, -30% RMSEP, +11% RPIQ, and the largest improvement with -86% was the bias. The Mahalanobis distance was an adequate alternative for the pls distance which was used in the local PLSR approach in the study by Nocita et al. (2014). Additionally, we could slightly improve the performance of the local PLSR approach by using a threshold to derive the calibration dataset, instead of a fixed number of samples. This allowed more samples to be selected for the individual calibration datasets for some validation samples.

Clearly, the local PLSR approach outperformed the k-means approaches which were combined with classic SOC prediction models based on PLSR for each cluster. Thus, the local PLSR approach that is based on spectral distance is better able to perform spectral clustering linked to SOC modelling than the k-means classification algorithms based on statistical multivariate (PCA) or spectral feature based (SFV) methods. A major difference between k-means and local PLSR is that for the k-means methods a comparably small number of clusters is formed (5 resp. 8 in this study) whereas with the local PLSR approach one model is computed per validation sample which leads to ~2,500 different calibration subsets which can be seen as clusters.

Comparing the results in our study to the study by Nocita et al. (2014) who also applied the local PLSR approach on the LUCAS database, a slight reduction in model performance is

562 observed. This can be explained as we modified some important parameters of the approach to
563 make it more generic and to be applicable for remote sensing. First, the removal of the water
564 bands excluded information from the spectra which could not be used for model calibration and
565 validation any more, thus, reducing the prediction accuracy. Second, we did not apply a pre-
566 discrimination between mineral and organic soils as our study is based on the sole use of the
567 spectral data. For organic soils Nocita et al. (2014) already demonstrated that for these soils the
568 model performance was much lower than for mineral soils. They had derived a very much
569 higher RMSEP (51.1 g kg^{-1}) with their local PLSR approach on the organic soils, which are
570 included in our study. Additionally, Nocita et al. (2014) modified the local regression procedure
571 by including other covariates (geographical and texture information) in the computation of the
572 distance between samples. We considered solely the spectral data, reducing the input
573 information for the modelling.

574 Our results in general compare well to other studies using large soil spectral libraries for the
575 prediction of soil properties (e.g. Araújo et al., 2014; Stevens et al., 2013; Viscarra Rossel et
576 al., 2016), although with slightly reduced accuracy. In the literature comparable studies are for
577 laboratory purposes and based on the whole spectral database, including the water bands. In our
578 case, removing the water bands that are important predictors for soil properties accordingly
579 seems to slightly reduce the prediction accuracy which has to be expected for large scale SOC
580 modelling. As such, the prediction errors in our study are comparatively large due to the higher
581 standard deviation in the large scale LUCAS database in comparison to local studies (Nocita et
582 al., 2014). Nevertheless, the prediction errors are still in an accepted reasonable range when
583 applied for remote sensing purposes.

584 Another issue is that underestimation of higher SOC values as shown in Fig. 8 is a well-known
585 issue in PLSR modelling as shown in the results for the reference model and the k-means
586 approaches. Reasons are the under-representation of higher SOC values in the calibration set

(Brown et al., 2005) caused by the skewed distribution of the SOC content and changes in the relationship between SOC and spectra for higher SOC values due to a saturation of the SOC spectral response (Nocita et al., 2014). Nevertheless, the local approach as shown in Fig. 8 seems to be able to deal with the prediction of high SOC values, which would show that spectral distance can cope well with higher amount of SOC content to group high-SOC content samples and perform a PLSR SOC prediction model with reasonable accuracy, also for these samples.

The LUCAS dataset used in this study and also most of the clusters in both k-means approaches have highly skewed SOC distributions. Therefore, it is important to transform the SOC values to approximately normal distribution. The transformation improved the model accuracies for all approaches. Nevertheless, so far only few studies transform skewed SOC contents before spectral predictions (e.g. Viscarra Rossel et al., 2016).

4.3 Simulated EnMAP spectra

Using the local PLSR approach on a LUCAS database that was spectrally reduced to match EnMAP's spectral characteristics only leads to a slight decrease of model accuracy compared to using the full spectral range of LUCAS. The validation results are still significantly better than those of the PLSR reference and the k-means approaches.

5. Conclusion

The objective of this study was to investigate (i) the potential of large soil spectral libraries for the modelling of SOC adapted to remote sensing applications, using the LUCAS EU-wide topsoil database and (ii) if spectral clustering of the large inhomogeneous spectral database helps to improve the quantification of SOC compared to SOC predictions based on the whole non-clustered database, by testing different clustering methodologies.

We tested a k-means clustering and explored two approaches that were either based on a PCA of the spectra or based on SFV. The SFV approach delivered better results, and both methods

could slightly improve the results of the PLSR reference model. Secondly, we tested a local PLSR approach which selected for each validation sample a set of most similar samples out of the pool of calibration samples that were used for model calibration. This approach achieved the overall best results and could clearly improve the SOC prediction accuracy compared to the reference model. We used the Mahalanobis distance as distance measure and a threshold instead of a fixed number of samples which further improved the results. The local PLSR, as the best model in our study, was applied to simulated EnMAP data based on the LUCAS database and model accuracy was almost as good as for the original LUCAS spectral resolution.

We noted that the number of LV in a PLSR model is very crucial for the accuracy and should therefore be specified in future work to encourage discussions on reasonable numbers of LV. Additionally, the highly skewed SOC content should be transformed into an approximately normal distribution prior to model calibration.

With this study we make a step towards the adaption of spectral soil models to the needs of air- and spaceborne SOC quantification. Our results are in the same range as other studies using large scale databases, with a slight reduction in accuracy considering a spectrally-reduced data set, not applying pre-clustering of the database, and conducting all analyses based on spectral information only without any prior knowledge of the SOC content or other soil covariates as in other studies. This study indicates that it is possible to improve the prediction accuracy of SOC by portioning the database into smaller groups. But it also shows that overlapping, individual groups are preferred over fixed ones. We demonstrate that the local PLSR approach is a valuable tool for SOC prediction based on large soil spectral databases that can be used without additional covariates than the spectral data. The usage of simulated hyperspectral data based on LUCAS led to good results which is very promising for current and future hyperspectral missions and ought to be tested on imagery spectral data for an area-wide quantification of

SOC. To do so, some challenges need to be faced like bridging the gap between laboratory and field resp. image spectra.

Acknowledgement

The study is funded within the EnMAP scientific preparation program under the DLR Space Administration with resources from the German Federal Ministry of Economic Affairs and Energy.

Literature

- 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(5), 718-729.
- Bartholomeus, H., Schaepman, M., Kooistra, L., Stevens, A., Hoogmoed, W., Spaargaren, O., 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma 145(1-2), 28-36.
- Baumgardner, M.F., Silva, L.F., Biehl, L.L., Stoner, E.R., 1986. Reflectance properties of soils, Advances in agronomy. Elsevier, pp. 1-44.
- Bayer, A., Bachmann, M., Müller, A., Kaufmann, H., 2012. A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem. Applied and Environmental Soil Science 2012.
- 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(7), 1398-1410.
- Ben-Dor, E., Banin, A., 1994. Visible and near-infrared (0.4–1.1 μm) analysis of arid and semiarid soils. Remote Sensing of Environment 48(3), 261-274.
- Ben-Dor, E., Chabrillat, S., Demattê, J., Taylor, G., Hill, J., Whiting, M., Sommer, S., 2009. Using imaging spectroscopy to study soil properties. Remote Sensing of Environment 113, S38-S55.
- 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), 1-15.
- Ben-Dor, E., Irons, J.R., Epema, G., 1999. Soil reflectance. In: A.N. Rencz (Ed.), Remote sensing for the Earth Science. Wiley, New York, pp. 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(3), 251-267.
- Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132(3), 273-290.
- 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(2-3), 431-445.
- Clark, R.N., 1999. Spectroscopy of rocks and minerals and principles of spectroscopy: Chapter 1. Remote Sensing for the Earth Sciences, 3. John Wiley & Sons, New York, NY, USA.

- Conant, R.T., Ogle, S.M., Paul, E.A., Paustian, K., 2011. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Frontiers in Ecology and the Environment 9(3), 169-173.
- Davies, T., 2005. An introduction to near infrared spectroscopy. NIR news 16(7), 9-11.
- Denton, F., Wilbanks, T.J., Abeysinghe, A.C., Burton, I., Gao, Q., Lemos, M.C., Masui, T., O'Brien, K.L., Warner, K., 2014. Climate-resilient pathways: adaptation, mitigation, and sustainable development, Cambridge, United Kingdom and New York, NY, USA.
- Feingersh, T., Ben-Dor, E., 2015. SHALOM–A commercial hyperspectral space mission. Optical payloads for space missions, 247-263.
- Gaffey, S.J., 1987. Spectral reflectance of carbonate minerals in the visible and near infrared (0.35–2.55 um): Anhydrous carbonate minerals. Journal of Geophysical Research: Solid Earth 92(B2), 1429-1440.
- Grove, C., Hook, S.J., Paylor III, E., 1992. Laboratory reflectance spectra of 160 minerals, 0.4 to 2.5 micrometers.
- Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., 2015. The EnMAP spaceborne imaging spectroscopy mission for earth observation. Remote Sensing 7(7), 8830-8857.
- Hartigan, J.A., 1975. Clustering algorithms, 209. Wiley New York.
- Hill, J., Schütt, B., 2000. Mapping complex patterns of erosion and stability in dry Mediterranean ecosystems. Remote Sensing of Environment 74(3), 557-569.
- Hunt, G.R., 1970. Visible and near-infrared spectra of minerals and rocks: I silicate minerals. Modern geology 1, 283-300.
- 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(6), 1101-1114.
- Jurasinski, G., Koebsch, F., Guenther, A., Beetz, S., 2014. flux: Flux rate calculation from dynamic closed chamber measurements. R package version 0.3-0. https://CRAN.R-project.org/package=flux.
- Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11(1), 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(5), 573-577.
- Kohl, M., 2018. _MKmisc: Miscellaneous functions from M. Kohl_, R package version 1.0, stamats.de.
- Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. science 304(5677), 1623-1627.
- Lehnert, L.W., Meyer, H., Bendix, J., 2017. hsdar: Manage, analyse and simulate hyperspectral data in R. R package version 0.7.0.
- Li, B., Morris, J., Martin, E.B., 2002. Model selection for partial least squares regression. Chemometrics and Intelligent Laboratory Systems 64(1), 79-89.
- Loizzo, R., Guarini, R., Longo, F., Scopa, T., Formaro, R., Facchinetti, C., Varacalli, G., 2018. PRISMA: the Italian hyperspectral mission, IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 175-178.
- Mevik, B.-H., Wehrens, R., Liland, K.H., 2016. pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0. https://CRAN.R-project.org/package=pls.
- Milewski, R., Chabrillat, S., Brell., M., Schleicher, A., Guanter, L., 2018. Assessment of the 1.75 µm Absorption Feature for Gypsum Estimation Using Laboratory, Air- and Spaceborne Hyperspectral Sensors. International Journal of Applied Earth Observation and Geoinformation (in rev.).
- 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.

- O'Rourke, S., Holden, N., 2011. Optical sensing and chemometric analysis of soil organic carbon–a cost effective alternative to conventional laboratory methods? Soil Use and Management 27(2), 143-155.
- 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.
- 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.
- Ramirez-Lopez, L., Stevens, A., 2016. resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. R package version 1.2.2.
- 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(5941), 680-681.
- Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry 36(8), 1627-1639.
- Segl, K., Guanter, L., Kaufmann, H., Schubert, J., Kaiser, S., Sang, B., Hofer, S., 2010. Simulation of spatial sensor characteristics in the context of the EnMAP hyperspectral mission. IEEE Transactions on Geoscience and Remote Sensing 48(7), 3046-3054.
- Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., Foerster, S., 2016. Prediction of Common Surface Soil Properties Based on Vis-NIR Airborne and Simulated EnMAP Imaging Spectroscopy Data: Prediction Accuracy and Influence of Spatial Resolution. Remote Sensing 8(7), 613.
- Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter five-visible and near infrared spectroscopy in soil science. Advances in agronomy 107, 163-215.
- Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PloS one 8(6), e66409.
- Tóth, G., Jones, A., Montanarella, L., 2013. LUCAS Topsoil Survey: Methodology, Data and Results. JRC Technical Reports. Luxembourg. Publications Office of the European Union, EUR26102 – Scientific and Technical Research series – ISSN 1831-9424 (online).
- Viscarra Rossel, R., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1), 46-54.
- 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.
- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems 58(2), 109-130.