Laboratory-based Vis–NIR spectroscopy and partial least squareregression with spatially correlated errors for predicting spatial variation of soil organic matter content

Massimo Conforti [a], Annamaria Castrignanò [b], Gaetano Robustelli [c], Fabio Scarciglia [c], Matteo Stelluti [b], Gabriele Buttafuoco [a,]*

[a] CNR, Institute for Agricultural and Forest Systems in the Mediterranean (ISAFOM), Rende, CS, Italy

[b] CRA — Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Bari, Italy

[c] DiBEST, University of Calabria, Rende, CS, Italy

Abstract

Soil organic matter (SOM) has beneficial effects on soil properties for plant growth and production. Moreover, SOM changes carbon dioxide concentrations in the atmosphere and can influence climate warming. Conventional methods for SOM determination based on laboratory analyses are costly and time consuming. Use of soil reflectance spectra is an alternative approach for SOM estimation and has the advantage of being rapid, non-destructive and cost effective. This method assumes that residuals are independent and identically distributed. However, in most cases this assumption does not hold owing to spatial dependence in soil samples. The aim of the paper was to test the potential of laboratory Vis–NIR spectroscopy to develop an approach of partial least square regression (PLSR) with correlated errors for estimating spatially varying SOM content from laboratory-based soil Vis–NIR spectra and producing a continuous map using a geostatistical method.

The study area was the Turbolo watershed (Calabria, southern Italy), which is representative of Mediterranean areas being highly susceptible to soil degradation. Topsoil samples were collected at 201 locations. To reduce the lack of linearity that may exist in the spectra, reflectance (R) spectra were transformed in absorbance spectra (log (1 / R)). Partial least squared regression (PLSR) analysis was then used to predict SOM from reflectance spectra. To take into account spatial correlation between observations, the significant latent variables from PLSR were used as regressors in a linear mixed effect model with correlated errors of SOM. The spatial approach and traditional PLSR were compared through the calculation of root mean square prediction error (RMSPE). In order to pro- duce a continuous map, the estimated SOM data were interpolated by ordinary kriging. The approach is particularly advantageous when the data exhibit a pronounced spatial autocorrelation and could be used in digital soil mapping.

1. Introduction

Soil organic matter (SOM) is a key attribute of soil and environmental quality because it is an important sink and source of main plant and microbial nutrients (Nieder and Benbi, 2008). Moreover, SOM exerts an important influence on the physical, chemical and biological properties and functions of soil (McBratney et al., 2014; Nieder and Benbi, 2008), because its depletion may reduce aggregate stability, resulting in crusting and compaction, as well as nutrient supply (Mabit and Bernard, 2009). Moreover, organic matter increases the soil's nutrient cycling capability (McBratney et al., 2014) and provides a large pool of macronutrients such as nitrogen, phosphorous and sulfur, which are very important for soil fertility. In addition, SOM has a positive influence on water retention capacity, porosity and cation exchange capacity (CEC).

On the global scale, carbon stored in soils represents one of the largest reservoirs of organic carbon and consequently, by either sequestering or releasing carbon in the atmosphere, soil can alter the terrestrial carbon balance and thereby the greenhouse effect (Lal, 2004; Lützow et al., 2006).

In recent decades, visible, near-infrared (Vis–NIR) reflectance spectroscopy has been found to be useful in measuring soil properties because the techniques are rapid, relatively inexpensive, and require minimal sample preparation and no hazardous chemicals; furthermore, they are non-invasive and several soil properties can be measured from a single scan (e.g. Demattê et al., 2006; McBratney et al., 2006; Reeves et al., 2001, 2002; Shepherd and Walsh, 2002; Stenberg et al., 2010; Viscarra Rossel et al., 2006).

There is widespread interest in Vis–NIR reflectance spectroscopy, even though soil Vis–NIR spectra are largely non-specific, resulting from overlapping absorptions of constituents often present in small concentrations in the soil (Viscarra Rossel and Behrens, 2010). The method is based on the simplified assumption that the soil reflectance in the 350–2500 nm spectral region is a linear combination of the spectral signatures of its compositional components weighted by their abundance (Ben-Dor, 2002; Curran, 1994; Ge et al., 2007). Therefore, changes in the chemical, physical and mineralogical properties of the soil produce distinct spectral features that can be detected through reflectance spectroscopy (Aïchi et al., 2009; Conforti et al., 2013a; Nanni and Demattê, 2006; Shepherd and Walsh, 2002; Viscarra Rossel et al., 2006). In particular, soil reflectance spectra are heavily dependent on SOM, as well as on other properties such as soil moisture and texture (Aïchi et al., 2009; Stevens et al., 2008).

Vis–NIR reflectance spectroscopy requires only a few seconds to measure a soil sample, but the relevant information needs to be mathematically extracted from the spectra so that it can be correlated with soil properties. To analyze soil reflectance spectra chemometrics techniques and multivariate calibrations (Martens and Næs, 1989; Stenberg et al., 2010; Viscarra Rossel and Behrens, 2010), such as multiple linear regression (MLR), principal components regression (PCR), partial least-squares regression (PLSR) and artificial neural networks (ANN) (e.g. Aïchi et al., 2009; Conforti et al., 2013b; Farifteh et al., 2007; Shepherd and Walsh, 2002; Viscarra Rossel et al., 2006) are generally used.

However, these techniques assume that SOM residuals (measured SOM minus predicted SOM) are identically and independently distributed: in other words, SOM observations should be independent of each other to guarantee optimality of the prediction model (Ge et al., 2007). Since soil properties generally exhibit significant spatial correlation with different degrees of spatial dependence, the use of PLSR combined with a linear mixed effect model (LMEM) (Lark, 2009; Stein, 1999) is expected to produce more accurate estimates. LMEM uses the significant latent variables from PLSR as fixed effects and the spatial covariance function of residuals as the stochastic (random) component to predict SOM.

Moreover, in the perspective of site-specific management, SOM content needs to be estimated spatially in order to produce accurate continuous maps, which can improve the information on local variation required by land managers and farmers (Viscarra Rossel and McBratney, 1998). However, from this point of view, the combined approach still leaves the task unfinished because the SOM predictions are made only at the sampled locations. A geostatistical analysis allows to map the spatial pattern of SOM prediction (Brown et al., 2006; Mouazen et al., 2007; Sarkhot et al., 2011; Viscarra Rossel et al., 2011), which is much more informative than the map of sparse observations for estimating carbon storage in the soil.

The objective of the paper was to develop an approach of partial least square regression (PLSR) with correlated errors for estimating spatially varying soil organic matter from laboratory-based soil Vis–NIR spectra and producing a continuous map using a geostatistical method. To estimate SOM, PLSR was combined with a linear mixed effect model (LMEM), which used the significant latent variables from PLSR as fixed effects, whereas spatial correlation between residuals as stochastic (random) component.

2. Materials and methods

2.1. Study area

The study area was the Turbolo watershed, located in the north of Calabria (southern Italy) between 39°32′25″N and 39°29′51″N latitude, 16°12′57″E and 16°05′21″E longitude (Fig. 1), and covers an area of 29.2 km2. Elevation ranges from 75 to 1015 m a.s.l., and slopes from 0° to 56.5°, then the landscape is characterized by large variability. The streams have a sub-dendritic drainage pattern, and the length of the main channel is about 13 km.

The climate is sub-humid, with average annual precipitation of 1200 mm and average air temperature of 16 °C (Conforti, 2009; Conforti et al., 2011). Rainfall mostly occurs from November to February, with frequent high-intensity rainstorms. The pedoclimatic regime is xeric and thermic, shifting to udic and mesic in the upper reaches (ARSSA, 2003).

The western part of the Turbolo watershed is characterized by steep slopes shaped on Paleozoic metamorphic rocks (mainly gneiss and schist), intensely fractured and weathered and in many places covered by a thick regolith (Fig. 1). In a wide eastern part of the study area, the morphology is characterized by gentle slopes and terraces cut on sedimentary terrains of Neogene–Quaternary ages mostly clays, sands and conglomerates (Lanzafame and Zuffa, 1976).

The main soil groups occurring in the study area (Fig. 2a), according to the soil map of Calabria (ARSSA, 2003), are Luvisols, Cambisols, Vertisols and Fluvisols (IUSS Working Group WRB, 2006).

The soil profiles frequently appear truncated or severely degraded by water erosion and gravitational processes (Conforti et al., 2012; Conforti et al., 2014; Lucà et al., 2011; Scarciglia et al., 2012). The prevailing soil textural classes are sandy loam and sandy clay loam (Buttafuoco et al., 2012; Conforti, 2009).

From the point of view of land use (Fig. 2b), about half of the study area is characterized by agriculture, mainly crops and olive groves, whereas more than 20% has a shrubby and herbaceous cover often left to pasture (Conforti, 2009). The remaining area consists of woodland, especially in the western part of the basin (Fig. 2b). Finally, erosion may be extreme on bare slopes.

2.2. Soil sampling and analysis

Composite soil samples were collected at 201 locations within the study area (Fig. 1) by using an auger sampler; soil sampling depth was set at 0.20 m, because this represents the most frequent value of A- horizon depth in the area.

The sampling sites were selected by subdividing the study area into 300 m × 300 m cells and within each of which, one point was chosen to be representative of the cell area on the basis of main soil–landscape features (geological substrate, topographic characteristics, soil types, land use/cover and development/degradation conditions of the topsoil). The locations of the sampling sites were recorded with a GPS Garmin eTrex30, with an accuracy of 3–5 m.

To ensure a soil homogeneous mixture, soil samples were dried, ground and sieved at 2 mm prior to analysis. Each sample was then split into two sub-samples: one was used for the laboratory spectral measurements, while the determination of SOM content was conducted on the other sub-sample, by the Walkley–Black method (Sequi and De Nobili, 2000).

2.3. Measurement and pre-treatment of Vis–NIR spectroscopy data

Vis–NIR reflectance of soil samples was measured in the laboratory, under artificial light, using an ASD FieldSpec Pro 350–2500 nm spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA), which combines three spectrometers to cover the spectrum portion (350 and 2500 nm), with a sampling interval of 1.4 nm for the 350–1000 nm region and 2 nm for the 1000–2500 nm region. FieldSpec Pro provided output at spectral resolution of 1 nm through a weighted cubic spline algorithm for interpolation, thus producing 2151 spectral bands. Two 100 W halogen lamps with a zenith angle of 30°, located at a distance of approximately 0.50 m from the soil sample were used as light sources. The soil samples, which were gently pressed and leveled with a spatula to obtain a smooth surface, were set inside a black cylinder of 10-cm diameter and 1-cm height during the measurements. The spectroradiometer was located in a nadir position with a distance of 10 cm from the sample, allowing the radiance measurements within a circular area of approximately 4.5-cm diameter. The noise level in the spectral signal was reduced through averaging 30 spectra for each soil sample. In addition, to eliminate any possible spectral anomalies due to geometry of measurement, four replicate scans were acquired by rotating the soil sample by 90° and were averaged in post-processing. A Spectralon panel (30 × 30 cm^2, Labsphere Inc., North Sutton, USA) was used as white reference to compute reflectance values. A reference spectrum under the same conditions of measurement was acquired immediately before the first scan and after every set of eight samples.

The spectral reflectance curves were finally averaged at 10 nm, so reducing the number of wavelengths from 2151 to 216, to smooth the spectra and keep down the risk of over-fitting (Shepherd and Walsh, 2002).

In order to further reduce residual noise and enhance the absorption frequencies, a number of spectral data pre-processing techniques were applied before statistical analysis:

- The measured reflectance (R) spectra were transformed into apparent absorbance through log (1 / R) to reduce noise, offset effects, and to enhance the linearity between measured absorbance and SOM concentration.
- The absorbance spectra were mean-centered to ensure that all results would be interpretable in terms of variation around the mean.
- Subsequently, the absorbance spectra were smoothed through a median filter algorithm with a first derivative to remove an additive baseline (Viscarra Rossel, 2008).
- Finally, absorbance spectra were normalized through the multiplicative scatter correction (MSC) (Geladi et al., 1985) to reduce the effect of scattering.

Details on pre-processing methods can be found in Martens and Næs (1989) and in Næs et al. (2004).

2.4. Analysis of Vis–NIR data

The approach aims at establishing a mathematical relationship between the response variable y (measured values of SOM) and the set of predictors X (spectral data). Among the available multivariate statistical methods, partial least squares regression (PLSR) (Geladi and Kowalski, 1986) was preferred. PLSR is a common chemometrics meth- od in Vis–NIR analysis (Martens and Næs, 1989; Viscarra Rossel et al., 2006). The idea behind PLSR is to find a few linear combinations (com- ponents or factors) of the original X-values and to use only these linear combinations in the regression equation (Næs et al., 2004). In this way, the irrelevant and unstable information is discarded and only the most relevant part of the X-variation is used for regression; the problem of collinearity is solved and more stable regression equations obtained (Næs et al., 2004). PLSR reduces the Vis–NIR matrix (reflectance by observation) to a small number of statistically significant components and is based on latent variable decomposition of two sets of variables: the set X of predictors (matrix n × N, where n is the number of observations and N is the number of wavelengths) and the set y of response variable (vector n × 1 of SOM measurements). The latent variables are orthogonal factors that maximize the covariance between independent (X) and dependent variables (y), and explain most of the variations in both predictors and responses. For more details on the PLSR method, see e.g. Martens and Næs (1989).

The optimal number of latent variables was chosen through a leave- one-out cross-validation (Efron and Tibshirani, 1993) as the number that minimizes the predicted residual sum of squares (PRESS).

The best prediction of the leave-one-out cross-validation model was evaluated using the coefficient of determination (R2) and root mean square error of prediction (RMSE).

Besides centering the predictors and the response variable, they were also scaled to standard deviation equal to one. Scaling serves to place all predictors and response on an equal footing relative to their variation in the data. Pre-treatment of data was performed with PArLeS v. 3.1 software developed by Viscarra Rossel (2008), and PLSR with the procedure PLS of SAS/STAT statistical package software (SAS Institute Inc., 2013 release 9.3).

2.5. REML-estimation of SOM with spatially correlated errors

The regression method implemented in the PLS procedure fits the observed data through the use of the ordinary least squares (OLS) method, which assumes that residuals of prediction are independent and identically distributed. Since SOM observations are expected to be autocorrelated, the variogram estimated from the residuals is biased because its point estimates depend in a non-linear way on the estimates of the coefficients of regressors (Lark et al., 2006). The state of the art for this problem is to use the residual maximum likelihood (REML) estimation of the spatial variance model in combination with the empirical best linear unbiased predictor (E-BLUP) (Patterson and Thompson, 1971). According to this approach, SOM is computed from a linear mixed effect model (LMEM) comprising an additive combination of the factors extracted with PLS as fixed effects, one random effect, which is the spatially dependent random variable in a geostatistical context and an independent random variable. The advantage of REML, to estimate variance parameters for the random effect, is that it reduces the bias found in maximum likelihood or OLS estimates (Cressie, 1993). Spatial covariance models, originally developed for Geostatistics, are also used in the mixed effect model approach (Diggle et al., 1998); therefore, correlation structure can be described by a variogram of spatial residuals.

The LMEM may be written as:

$$z = X\beta + Zu + \varepsilon \qquad 1$$

where the vector z contains the SOM observations, X is an n × p design matrix consisting of the n observations of the p fixed effects (the factors extracted with PLS), the vector β contains the p fixed-effect coefficients; u is the spatially dependent random variable; Z is the design matrix and the term ε is a vector of independent random errors. The random terms u and ε are assumed to be jointly Gaussian and independent of each other. The term ε, in particular, represents both independent measurement errors and variation at a spatial scale smaller than the one of sampling and is the nugget effect in geostatistics. If u is assumed to be drawn from a second-order stationary random process, its correlation matrix will depend only on the relative locations of the observations, and its covariance function will be an authorized mathematical model of the distance between observations used in geostatistics. The parameters of such a function will be estimated by REML because this removes dependence of the estimates on the fixed-effect coefficients. These coefficients are the ones that maximize the residual log-likelihood function and are found numerically through the use of a ridge-stabilized Newton–Raphson algorithm (Lindstrom and Bates, 1988). Once the parameters of covariance function and the coefficients of fixed effects are estimated, the predictions are computed at the sites where the factors are known.

The spatial association of the residuals from PLSR was tested in different ways:

- Calculating the Moran's I (Moran, 1950) and Geary's c (Geary, 1947) spatial autocorrelation statistics and comparing these to their expect- ed values under a null spatial (completely randomized) model;
- Fitting a mathematical model to the experimental variogram of the residuals;
- Performing a likelihood ratio test to assess whether the simplifications used in the non-spatial correlation model are still applicable with spatially correlated errors (Oman, 1991; Wolfinger, 1993).

This test requires computation of the restricted log-likelihood (LLR) for each model, evaluated at the REML estimates of parameters. The likelihood ratio statistic for comparing the reduced (non-spatial) model to the full (spatial) model is:

$$\chi 2 = -[LL_R(\text{reduced model}) - LL_R(\text{full model})]: \qquad 2$$

Under the null hypothesis that the reduced model is no different from the full one; the likelihood ratio statistic is distributed as Chi- squared with the number of freedom degrees equal to the difference in the number of parameters of each of the two models. Because the fixed part is the same for the two models, only the number of parameters in the variance–covariance structure needs to be considered.

Since REML estimation entails an explicit assumption that ε has a Gaussian distribution, the distributional assumptions for the mixed effect model are tested by calculating the descriptive statistics of residuals and comparing residuals with the corresponding quantiles of the standard normal variable.

The two procedures, PLSR and the combination of PLSR with linear mixed effect model, are also compared by root mean square prediction error (RMSPE) calculated through cross-validation.

The linear mixed effect model approach was implemented using MIXED procedure of SAS/STAT software (SAS

Institute Inc., 2013 release 9.3).

To form the SOM predictions at an unsampled site in order to produce a continuous map, the estimates were interpolated by ordinary kriging (Webster and Oliver, 2007). All geostatistical analyses were carried out with the software package ISATIS®, release 2014 (Géovariances, 2014).

3. *Results and discussion*

Table 1 presents summary statistics for SOM data. The SOM contents varied spatially from a minimum value of 0.30% to a maximum of 6.50%, with a mean value of 2.62% (Table 1). The SOM dataset was characterized by a positively skewed distribution (0.84) (Table 1, Fig. 3).

To analyze the relationship of SOM with soil type and land use, the measured SOM data were classified into four classes (i.e. high, medium, low and very low) based on the USDA textural classes (Table 2 and Fig. 2) (Soil Survey Staff, 2010) and then compared with soil types and land use (Fig. 4).

The comparison between the classes of topsoil SOM content and the ones of soil type and land use showed that high SOM contents were prevalently recorded in the Cambisols and Luvisols (Fig. 4a) and in woodland areas (Fig. 4b). Low SOM content values were measured in topsoil samples of cropland, which are often characterized by intense water erosion and tillage-induced erosion due to unsustainable agricultural practices (Conforti, 2009). Moreover, topsoil samples with very low SOM content were associated with barren lands, mostly on land with intense erosive processes (Conforti et al., 2013a, b).

A visual inspection of the set of spectra allowed us to detect that they are affected by variations in SOM content. The mean reflectance spectra of the four classes of SOM content (Fig. 5) showed a tendency to de- crease with SOM, as reported by other authors (Ben-Dor, 2002). The overall shape of the Vis–NIR spectra was generally similar for all sam- ples and most displayed some degree of steep slope between 400 and 900 nm. All soil reflectance spectra exhibited high absorption peaks around 1400 nm, 1900 nm and 2200 nm (Fig. 5). These features may be associated with clay minerals, OH features of free water at 1400 and 1900 nm, and lattice OH features at 1400 and 2200 nm (Ben-Dor, 2002). The spectra also showed a small absorption peak around 2200 nm, which may be due to organic molecules (e.g., CH_2, CH_3, and NH_3), Si\OH bonds, cation\OH bonds in phyllosilicate minerals (e.g., kaolinite, montmorillonite) (Clark et al., 1990).

We retained eight PLSR factors (latent variables) since they resulted to be significant by cross-validation and explained more than 80% of variation in both predictors and response. We deemed acceptable a loss of less than

20% of the information for the construction of a prediction model of SOM.

The spatial autocorrelation of the residuals from PLSR was verified with both Moran's I and Geary's c, tests (Table 3). The observed Moran's I coefficient (Table 3) was statistically greater (0.217) than the expected value (− 0.005) indicating a positive spatial autocorrelation of the residuals. The Geary's c index (Table 3) confirmed the positive spatial autocorrelation of the residuals and was less (0.681) than the expected value (1).

An exponential model with a practical range equal to 600 m was fitted to the experimental variogram of residuals. The non-spatially correlated component (nugget effect) was about twice (0.23%^{2}) the structured component (partial sill = 0.13%^{2}), which may be due to the rather coarse sampling scale of soil. The estimated parameters of the variogram model were used as input values to initialize the iterative procedure of fitting in the mixed effect model estimator.

The REML estimated parameters (partial sill, range, nugget effect) of the exponential model of covariance function of residuals and the estimates of the intercept (β_0) and the coefficients (β_i) of the eight latent variables (fixed effects) are shown in Table 4. The exponential spatial variance model was preferred to other authorized variance models on the basis of the residual likelihood, because all the LMEMs had the same fixed-effect structure.

All the fixed effects and residual (nugget effect) were highly significant; the parameters of the covariance function were significant at a probability level of about 0.10. The weak stochastic component, related to spatial autocorrelation, was probably due to a too coarse sampling scale. However, the likelihood ratio test:

$$\chi 2 = -[LL_R(nonspatialmodel) - LL_R(spatialmodel)]$$

$$= -(-419.9 + 413.4) = 6.5$$

was significant at probability level of $p < 0.05$, which means that the residuals of SOM estimation were spatially correlated; therefore, the use of the mixed effect model approach after PLS regression is justified and expected to improve the prediction of SOM.

Table 5 shows the summary statistics of the residuals from the fitted LMEM calculated with cross-validation and Fig. 6 displays the q–q plot. The residuals were symmetrically distributed and showed no evident departure from the normality assumptions of the model, supporting the assumption of a Gaussian random process superimposed on an ex- ternal drift represented by the spectral latent variables. Moreover, the RMSPE of LMEM was 0.59, smaller than RMSPE found for traditional PLSR model (0.69), which is further evidence of the advantages of the proposed approach.

The utility of using reflectance data, synthetized in eight latent variables as fixed effects, for spatial prediction of SOM was also proved with the Akaike information criterion (Akaike, 1973), which was small- er for the estimated LMEM compared with the one of the no-external drift models, in which the one fixed effect was an overall mean (intercept) (419.4 against 913.9).

The above results are quite promising; however, the estimated relation between SOM and spectroradiometric data needs to be tested further across a wider range of soils, characterized by different properties, texture, parent material and age of landscape, to confirm its wider applicability.

To produce a continuous map of LMEM SOM predictions, a bounded isotropic variogram model was estimated, after checking the occurrence of anisotropy with a variogram map (not shown), including a nugget effect (1.07%2) and two spherical models with ranges of about 753 m and 2066 m, respectively. The results of cross validation were quite satisfactory because the mean of the estimation error (−0.01%) and the mean squared deviation ratio (1.06) were close to 0 and 1, respectively.

The interpolated map of the LMEM SOM predictions, obtained using ordinary kriging, is reported in Fig. 7. The map shows that high contents of SOM (N 5%) can be observed along the slopes in the western part of the study area, which is characterized by Cambisols formed on metamorphic rocks; in addition, the SOM shows higher values where Fluvisols are developed and with scrub/herbaceous land cover and olive groves. Low values of SOM (on average b 2%) were mapped in the central and eastern portion of the Turbolo catchment, where there are Luvisols and Cambisols and land use characterized mainly by crops and olive groves (Fig. 2). In these areas, the low content of SOM could be due to tillage erosion caused by mechanized agriculture, which promotes the oxidation of SOM and leads to increased soil erosion (Rasmussen et al., 1998). Spatial distribution of SOM shows that the low contents were found in the areas where soils (e.g. Vertisols and Fluvisols) developed on clayey and sandy parent materials (Fig. 1), often truncated by erosive processes (Conforti et al., 2011). Moreover, a visual inspection of the map shows that the lowest values of SOM are located in hilly barren lands, where clay lithology outcrops.

From what previously shown, it results that, by simply adding two columns of spatial coordinates to reflectance data and modifying the regression method, it is possible to improve SOM prediction and produce a continuous representation of SOM spatial variation.

4. Conclusions

In this study, a combined method (PLSR-regression with correlated errors) was used with Vis–NIR spectra to determine organic matter in soil within the context of digital soil mapping. The key objective was to develop an approach, which accounted for spatial dependence, should it occur, whereas it is generally ignored in regression methods. The results showed that the approach proposed can improve the prediction of SOM and that soil reflectance spectra, if treated with prop- er analytical procedures, can serve as excellent co-variables for SOM estimation. The proposed methodology could be incorporated into remote/proximal sensing for digital soil-property mapping by using remotely or proximally sensed hyperspectral images as exhaustive variables, known at each node of an interpolation grid, where only a small number of reference measurements would be needed to estimate calibration function. The use of geostatistical techniques, such as multicollocated cokriging or kriging with external drift (Castrignanò et al., 2011), could extend SOM prediction to the whole area monitored by the remote or proximal sensor.

Acknowledgments

The authors thank the reviewers for their critical comments and suggestions, which greatly improved the quality of our manuscript. We are grateful to Kevin O'Connel for his help in polishing the English of this paper.

References

Aïchi, H., Fouad, Y., Walter, C., Viscarra Rossel, R.A., Lili Chabaane, Z., Sanaa, M., 2009. Regional predictions of soil organic carbon content from spectral reflectance measurements. Biosyst. Eng. 104, 442–446.

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N., Csaki, F. (Eds.), 2nd International Symposium on Informa- tion Theory. Akademia Kiado, Budapest, pp. 267–281.

ARSSA. 2003. Carta dei suoli della regione Calabria — scala 1:250,000. Monografia divulgativa. ARSSA — Agenzia Regionale per lo Sviluppo e per i Servizi in Agricoltura, Servizio Agropedologia. Rubbettino, 387 pp. (In Italian)

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Adv. Agron. 75, 173–243.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil char- acterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.

Buttafuoco, G., Conforti, M., Aucelli, P.P.C., Robustelli, G., Scarciglia, F., 2012. Assessing spatial uncertainty in mapping soil erodibility factor using geostatistical stochastic simulation. Environ. Earth Sci. 66, 1111–1125.

Castrignanò, A., Buttafuoco, G., Comolli, R., Castrignanò, A., 2011. Using digital elevation model to improve soil pH prediction in an Alpine doline. Pedosphere 21, 259–270.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. 95, 12653–12680.

Conforti, M., 2009. Studio geomorfopedologico dei processi erosivi nel bacino del T. Turbolo (Calabria settentrionale) con il contributo della spettrometria della riflettenzaPhD Thesis University of Calabria, Italy (310 pp).

Conforti, M., Aucelli, P.P.C., Robustelli, G., Scarciglia, F., 2011. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo Stream catchment (Northern Calabria, Italy). Nat. Hazards 56, 881-898.

Conforti, M., Buttafuoco, G., Leone, A.P., Aucelli, P.P.C., Robustelli, G., Scarciglia, F., 2012. Soil erosion assessment using proximal spectral reflectance in VIS–NIR–SWIR region in sample area of Calabria region (Southern Italy). Rend. Online Soc. Geol. Ital. 21 (Part 2), 1202–1204.

Conforti, M., Buttafuoco, G., Leone, A.P., Aucelli, P.P.C., Robustelli, G., Scarciglia, F., 2013a. Studying the relationship between water-induced soil erosion and soil organic matter using Vis–NIR spectroscopy and geomorphological analysis: a case study in a southern Italy area. Catena 110, 44-58.

Conforti, M., Froio, R., Matteucci, G., Caloiero, T., Buttafuoco, G., 2013b. Potentiality of laboratory visible and near infrared spectroscopy for determining clay content in for- est soils: a case study from high forest beech (Fagus sylvatica) in Calabria (southern Italy). EQA Int. J. Environ. Qual. 11, 49–64.

Conforti, M., Pascale, S., Robustelli, G., Sdao, F., 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena 113, 236–250.

Cressie, N., 1993. Statistics for Spatial Data (Revised Edition). Wiley, New York. Curran, P.J., 1994. Imaging spectrometry. Program. Phys. Geogr. 18, 247–266.

Demattê, J.A.M., Sousa, A.A., Alves, M.C., Nanni, M.R., Fiorio, P.R., Campos, R.C., 2006. Deter- mining soil water status and other soil characteristics by spectral proximal sensing. Geoderma 135, 179–195.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. J. Appl. Stat. 47, 299–350.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. vol. 57. Chapman and Hall, London, UK (436 pp).

Farifteh, J., Van Der Meer, F., Atzberger, C., Carranza, E.J.M., 2007. Quantitative analysis of salt-affected soil reflectance spectra: a comparison of two adaptive methods (PLSR and ANN). Remote Sens. Environ. 110, 59–78.

Ge, Y., Thomasson, J.A., Morgan, C.L., Searcy, S.W., 2007. VNIR diffuse reflectance spectros- copy for agricultural soil property determination based on regression-kriging. T ASABE 50, 1081–1092.

Geary, R.C., 1947. Testing for normality. Biometrika 34, 209-242.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta. 185, 1–17.

Geladi, P., MacDougall, D., Martens, H., 1985. Scatter correction for near-infrared reflectance spectra of meat. Appl. Spectrosc. 39, 491–500.

Géovariances, 2014. Isatis technical references, version 2014. Avon Cedex, France.

IUSS Working Group WRB, 2006. World Reference Base for Soil Resources 2006. World Soil Resources Reports 103. FAO, Rome.

Lal, R., 2004. Soil carbon sequestration to mitigate climate change. Geoderma 123, 1–22. Lanzafame, G., Zuffa, G., 1976. Geologia e petrografia del foglio Bisignano (Bacino del Crati, Calabria). Geol. Romana 15, 223–270.

Lark, R.M., 2009. Kriging a soil variable with a simple nonstationary variance model. J. Agric. Biol. Environ. St. 14, 301–321.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57, 787–799.

Lindstrom, M.J., Bates, D.M., 1988. Newton–Raphson and EM algorithms for linear mixed- effects models for repeated-measures data. J. Am. Stat. Assoc. 83, 1014–1022.

Lucà, F., Conforti, M., Robustelli, G., 2011. Comparison of GIS-based gullying susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. Geomorphology 134, 297–308.

Lützow, M.V., Kögel-Knabner, I., Ekschmitt, K., Matzner, E., Guggenberger, G., Marschner, B., Flessa, H., 2006. Stabilization of organic matter in temperate soils: mechanisms and their relevance under different soil conditions — a review. Eur. J. Soil Sci. 57, 426–445.

Mabit, L., Bernard, C., 2009. Spatial distribution and content of soil organic matter in an agricultural field in eastern Canada, as estimated from geostatistical tools. Earth Surf. Proc. Land 35, 278-283.

Martens, H., Næs, T., 1989. Multivariate Calibration. John Wiley & Sons, Chichester, United Kingdom, UK.

McBratney, A.B., Minasny, B., Viscarra Rossel, R.A., 2006. Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis. Geoderma 136, 272–278.

McBratney, A.B., Stockmann, U., Angers, D., Minasny, B., Field, D., 2014. Challenges for soil organic carbon research. In: Alfred, E., Hartemink, A.E., McSweeney, K. (Eds.), Soil Carbon. Springer, New York, pp. 3–16.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

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

Næs, T., Isaksson, T., Fearn, T., Davies, T., 2004. A User-Friendly Guide to Multivariate Calibration and Classification. Reprinted with Corrections. NIR Publications, Chichester.

Nanni, M.R., Demattê, J.A.M., 2006. Spectral reflectance methodology in comparison to traditional soil analysis. Soil Sci. Soc. Am. J. 70, 393–407.

Nieder, R., Benbi, D.K., 2008. Carbon and Nitrogen in the Terrestrial Environment. Springer.

Oman, S.D., 1991. Multiplicative effects in mixed model analysis of variance. Biometrika 78, 729–739.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Rasmussen, P.E., Goulding, K.W.T., Brown, J.R., Grace, P.R., Janzen, H.H., Körschens, M., 1998. Long-term agroecosystem experiments: assessing agricultural sustainability and global change. Science 282, 893–896.

Reeves III, J.B., McCarty, G.W., Reeves, V.B., 2001. Mid-infrared diffuse reflectance spec- troscopy for the quantitative analysis of agricultural soils. J. Agric. Food Chem. 49, 766-772.

Reeves III, J., McCarty, G., Mimmo, T., 2002. The potential of diffuse reflectance spectros- copy for the determination of carbon inventories in soils. Environ. Pollut. 116, S277–S284.

Sarkhot, D.V., Grunwald, S., Ge, Y., Morgan, C.L.S., 2011. Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. Geoderma 164, 23–32.

SAS Institute Inc, 2013. SAS® 9.3 Guide to Software Updates. SAS Institute Inc., Cary, NC, USA.

Scarciglia, F., Conforti, M., Buttafuoco, G., Robustelli, G., Aucelli, P.P.C., Morrone, F., Casuscelli, F., Palumbo, G., 2012. Integrated study of a soil catena in the Turbolo watershed (Calabria, southern Italy): soil processes, hydrology and geomorphic dynamics. Rend. Online Soc. Geol. Ital. 21 (Part 2), 1215–1217.

Sequi, P., De Nobili, M., 2000. Determinazione del carbonio organico. In: Violante, P. (Ed.), Metodi di analisi chimica del suolo, VII.3. Franco Angeli, Roma, pp. 18–25 (in Italian).

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66, 988–998.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

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

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

Soil Survey Staff, 2010. Keys to Soil Taxonomy, 11th Edit., USDA — United States Department of Agriculture. Natural Resources Conservation Service, Washington, DC (338 pp).

Viscarra Rossel, R.A., 2008. ParLeS: software for chemometrics analysis of spectroscopic data. Chemom. Intell. Lab. 90, 72–83.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil dif- fuse reflectance spectra. Geoderma 158, 46–54.

Viscarra Rossel, R.A., McBratney, A.B., 1998. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. Geoderma 85, 19–39.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.

Viscarra Rossel, R.A., Chappell, A., de Caritat, P., McKenzie, N.J., 2011. On the soil informa- tion content of visible–near infrared reflectance spectra. Eur. J. Soil Sci. 62, 442–453. Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd ed. Wiley, Chichester.

Wolfinger, R.D., 1993. Covariance structure selection in general mixed linear models.Commun. Stat.-Theor. M. 22, 1076-1106.

Figures and Table

Fig. 1. Location of the study area and topsoil samples. The lithologic map of study area is also reported

Fig. 2. Soil (a) and land use (b) maps. A posting of the measured SOM content values using four classes is also reported.

Fig. 3. Histogram of measured SOM data.

Fig. 4. Soil samples distribution in the SOM classes for different soil types (a) and land use (b).

Fig. 5. Mean reflectance curves of soils for different classes of SOM.

Fig. 6. q–q plot of residuals for the fitted spatial linear mixed effects model.

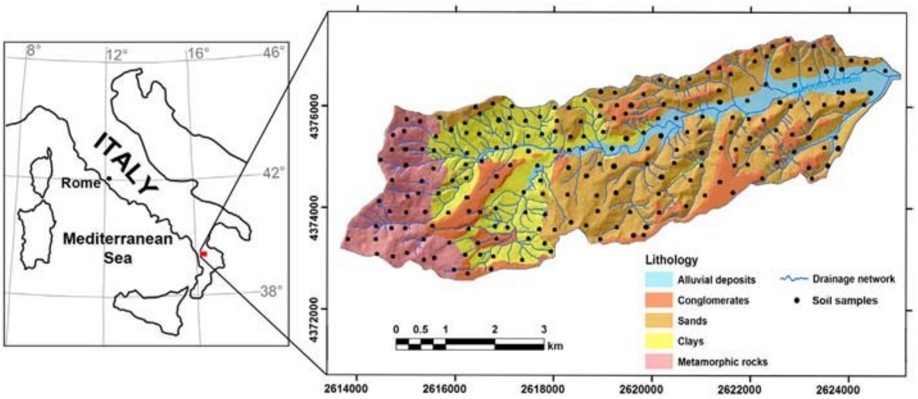
Table 1 - Basic statistics of measured SOM content data.

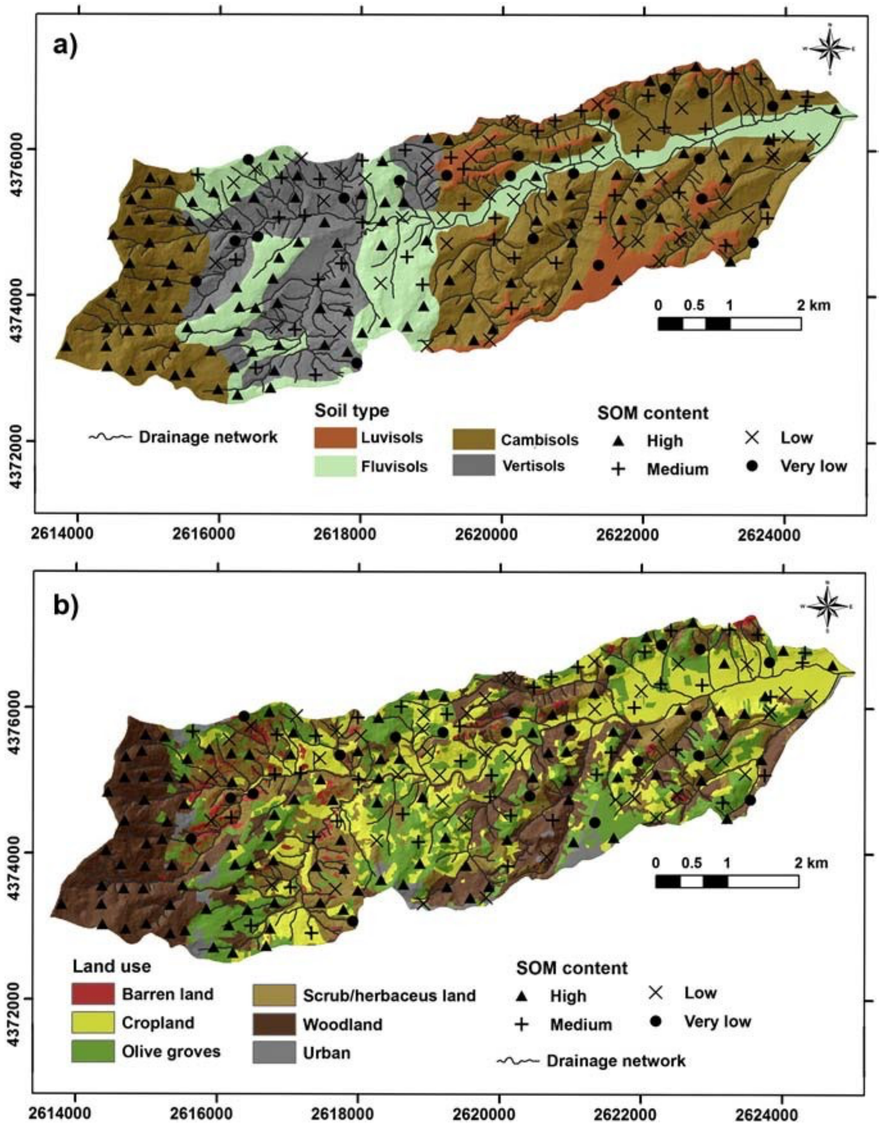
Table 2 - SOM content classified according to the USDA textural classes.

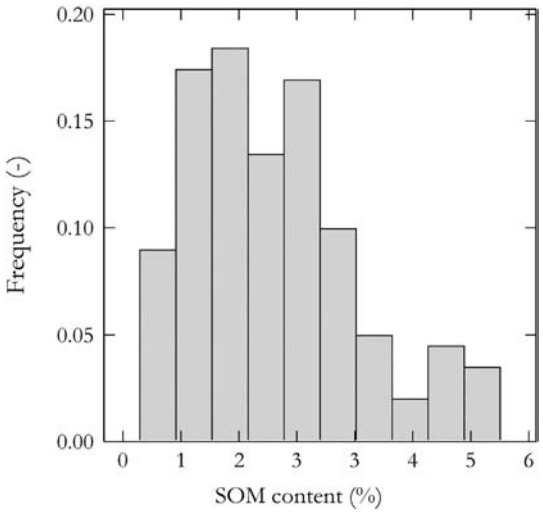
Table 3 - Results for the autocorrelation statistics.

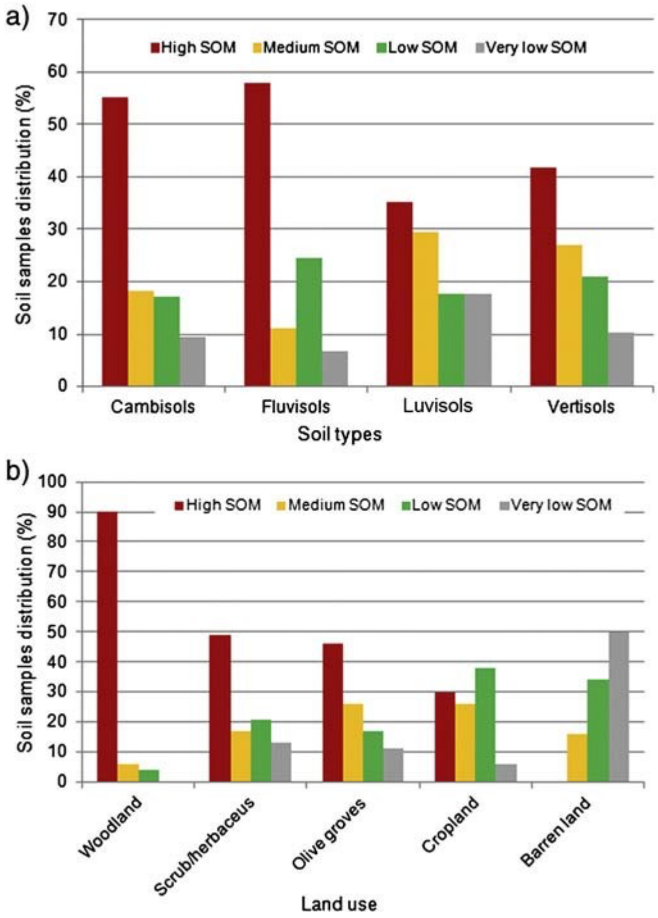
Table 4 - Results of linear mixed model estimation.

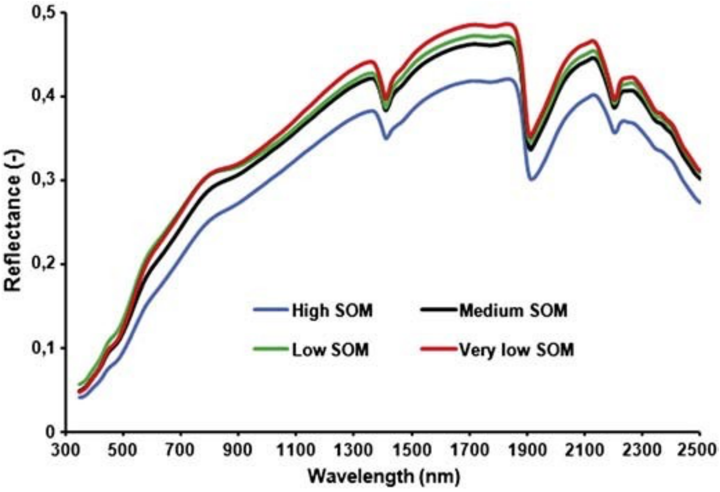
Table 5 - Basic statistics of residuals for the fitted spatial linear mixed effects model.

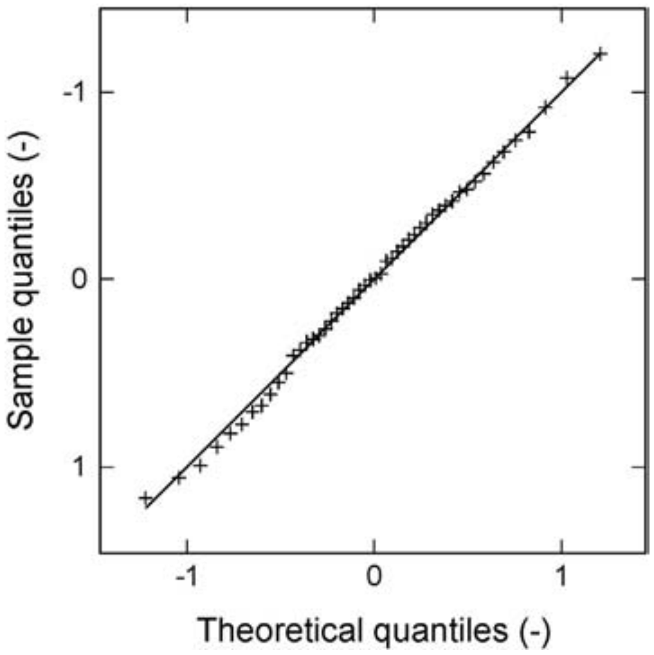



a)
4376000
4374000
4372000
2614000
2616000
2618000
2620000
2622000
2624000
0 0.5 1 2 km
Drainage network
Soil type
Luvisols
Fluvisols
Cambisols
Vertisols
SOM content
High
Medium
Low
Very low
b)
Land use
Barren land
Cropland
Olive groves
Scrub/herbaceus land
Woodland
Urban
SOM content
High
Medium
Low
Very low
Drainage network


Frequency (-)
0.00
0.05
0.10
0.15
0.20
SOM content (%)
0
1
2
3
4
5
6


a)
High SOM
Medium SOM
Low SOM
Very low SOM
Soil samples distribution (%)
0
10
20
30
40
50
60
70
Cambisols
Fluvisols
Luvisols
Vertisols
Soil types
b)
High SOM
Medium SOM
Low SOM
Very low SOM
Soil samples distribution (%)
0
10
20
30
40
50
60
70
80
90
100
Woodland
Scrub/herbaceus
Olive groves
Cropland
Barren land
Land use


Reflectance (-)
0
0,1
0,2
0,3
0,4
0,5
300
500
700
900
1100
1300
1500
1700
1900
2100
2300
2500
Wavelength (nm)
High SOM
Medium SOM
Low SOM
Very low SOM



1

Count	201
Mean (%)	2.62
Minimum (%)	0.30
Lower quantile (%)	1.50
Median (%)	2.40
Upper quantile (%)	3.30
Maximum (%)	6.50
Standard deviation (%)	1.43
Variance (%)	2.04
Skewness (−)	0.84

2

Assumption	Coefficient	Observed	Expected	Stand. dev.	Probability
Randomization	Moran's *I*	0.217	−0.005	0.049	<0.0001
Randomization	Geary's c	0.681	1.000	0.061	<0.0001

3

USDA texture class	SOM content (%)			
	Very low	Low	Medium	High
Sand, loamy sand, sandy loam	<0.8	0.8–1.4	1.5–2.0	>2.0
Loam, sandy clay, sandy clay loam, silty loam, silt	<1.0	1.0–1.8	1.9–2.5	>2.5
Clay, clay loam, silty clay, silty clay loam	<1.2	1.2–2.2	2.3–3.0	>3.0

4

Covariance parameter estimates

Covariance parameters	Subject	Estimate	Standard error	Probability
Exponential model	Partial sill	0.1117	0.1029	0.1389
	Range	377.99	348.72	0.1392
Residual	Nugget effect	0.2552	0.1040	0.0071

Solution for fixed effects

Effect	Estimate	Standard error	Probability
β_0	2.6255	0.0639	<0.0001
β_1	0.1548	0.0068	<0.0001
β_2	0.0538	0.0070	<0.0001
β_3	0.1690	0.0154	<0.0001
β_4	0.0521	0.0088	<0.0001
β_5	0.0375	0.0074	<0.0001
β_6	0.0761	0.0133	<0.0001
β_7	0.1326	0.0191	<0.0001
β_8	0.1115	0.0219	<0.0001

Count	201
Mean (%)	−0.01
Minimum (%)	−1.47
Median (%)	0.00
Maximum (%)	1.50
Standard deviation (%)	0.59