Dipartimento di Economia e Finanza

METODI E ANALISI STATISTICHE

2020

Università degli Studi di Bari Aldo Moro

Toma E., d'Ovidio F. (a cura di) (2020). *Metodi e Analisi Statistiche*, Dipartimento di Economia e Finanza, Università degli studi di Bari *Aldo Moro*.

Prima edizione: dicembre 2020

ISBN 978-88-6629-023-0

Gli articoli qui presentati sono stati oggetto, oltre che di valutazione interna, anche di revisione anonima (in "doppio cieco").

Editing finale: F.D. d'Ovidio, E. Toma

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2020
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-023-0
pp. 95-118

Geostatistical analysis of soil reflectance spectra for field-scale digital soil mapping. A case study

Natalia Leone[1*], Valeria Ancona[1], Davide Fragnito[2], Domenico Vitale[3], Massimo Bilancia[4]

[1]Water Research Institute, National Research Council, Viale Francesco de Blasio, 5, Bari,
[2]Master's Graduate in Statistical, Actuarial and Financial Sciences,
[3]CMCC Foundation, Euro-Mediterranean Center on Climatic Change, Viterbo,
[4]Ionian Department DJSGE – University of Bari A. Moro, Via Duomo 259, Taranto.

Abstract: Knowledge of field-scale soil variability is essential for sustainable soil management. Traditional techniques, based on soil analysis, are costly and time-consuming. An alternative method would be the use of visible-infrared reflectance spectroscopy coupled with multivariate analysis, specifically principal component analysis (PCA) and geostatistics.
In this study, after brief reviews regarding reflectance spectroscopy, PCA, and geostatistics, we presented a methodological approach for digital soil mapping in a study area of Southern Italy. Reflectance spectra of 240 surface soil samples collected at geo-referenced sites, were decomposed by PCA. The first three components (PC1, PC2, PC3) explained most (98%) of the total variance of the initial data set, therefore, they were considered for the assessment of soil spatial variability by variography and kriging (geostatistics). The resulting PC1, PC2 and PC3 kriging maps were interpreted in the light of the information contents on reflectance spectra and compared with the results of a previous, conventional soil survey. The presented strategy seems to be efficient and reliable for mapping soil spatial variability.

Keywords: Soil reflectance; Principal component analysis (PCA); Geostatistics; Digital Soil Mapping.

[*] Corresponding author: natalia.leone@ba.irsa.cnr.it
All authors reviewed and revised the manuscript, equally contributing to this work, approved the final version, and agreed to submit the revised manuscript for publication. The authors state that they have no disclosure to declare.

1. Introduction

Soils are rarely homogeneous at different spatial scales (Odlare et al., 2005). Variations in the soil properties, which are particularly evident over a large scale, may occur markedly also within a few hectares of farmland (field scale), due to small changes in topography and thickness of parent material layers or the effects of past human management (Brady and Weil, 2002). Despite this, soils are traditionally treated as homogeneous, with possible adverse effects on crop yield, management costs, and the environment. However, these effects can be contained, if not completely avoided, by adapting soil management to the site's specific conditions, as assessed through the correct knowledge of the within-field variability. This is the purpose of the set of agricultural techniques, better known as "precision agriculture". A way to investigate within-field soil variability could be the production of detailed maps, based on many traditional chemical and physical analyses. However, these analyses are expensive and time-consuming. Therefore, this approach is unsuitable when soils need to be analysed, as in precision farming. Hence, the need to investigate alternative techniques. Recently, particular interest has been shown towards reflectance spectroscopy in the visible and near-infrared visible domain (vis-NIR spectroscopy) (Leone et al., 2012). Vis-NIR reflectance spectroscopy is a rapid, cost-effective, non-invasive, and non-destructive technique that requires only minimal sample preparation and does not require the use of hazardous chemicals (Viscarra Rossel et al., 2006). Vis-NIR reflectance spectroscopy is defined as the ratio between radiation reflected from the surface of a material (the soil, in our case) and the radiation incident on it, at different wavelengths, between 350 and 2500 nm (Drury, 1993).

In the above-mentioned spectral region, each soil constituent has specific absorption properties, due to energy transitions, either electronic (in the visible) or vibrational (in the near-infrared; Leone, 2000a). Therefore, soils with different chemical, physical and mineralogical properties show various spectral features. The latter can be conveniently analysed to acquire either qualitative or quantitative information on these properties (Leone et al., 2012), or to analyse and map the spatial distribution of the soil mantle (digital soil mapping) (Odlare et al., 2005; Viscarra Rossel and Behrens, 2010), in combination with multivariate and geostatistical data analysis. Although promising, the use of vis-NIR spectroscopy, combined with multivariate and geostatistical analysis, has been little used, also due to the lack of knowledge about the basic concepts of vis-NIR spectroscopy, multivariate statistical and geostatistical methods. The present work aims to provide a methodological contribution

to digital soil mapping based on soil spectral reflectance measurements, multivariate statistical methods, and geostatistics.

2. Some basic concepts

2.1 Soil spectral reflectance

Soil is a semi-infinite medium relative to electromagnetic radiation. In other words, electromagnetic radiation incident on the soil is either absorbed within it or is reflected from its surface. The latter can be measured and then related to soil properties (Irons et al., 1989). The fraction of incident flux that is reflected is referred to as spectral reflectance is; it is usually measured in the spectral domains of visible (vis, 350-780 nm) and near infrared (NIR, 780-2500 nm). For this, the spectral reflectance in the 350-2500 nm domain is commonly referred to as vis-NIR reflectance. However, sometimes, the spectral domain between 700 and 2500 nm is further divided into near infrared (NIR, 780-1100 nm) and short-wave infrared (1100-2500 nm).

The vis–NIR spectral reflectance of a soil is affected by several chemical and physical properties, referred to as "chromophores". A given soil sample consists of a variety of chromophores, which vary with environmental conditions. In many cases, the spectral signals related to a given chromophore overlap with those of other chromophores and thereby hinder the assessment of the effect of a given chromophore (Ben-Dor et al., 1999).

Water, organic matter, and minerals are the main chemical chromophores of a soil. Their influence on soil reflectance is related to vibrational motions and electron transitions. The vibrational motions consist of oscillations in the relative positions of bonded atomic cores. The oscillations either stretch molecular bond lengths or bend interbond angles. Energy level transitions involving nuclear vibrations typically result in the absorption or emission of radiation within the infrared portion of the spectrum (Irons et al., 1989). The electronic transitions involve changes in the energy levels of the electrons in soil atoms and molecules. Electronic processes produce absorption bands readily distinguishable from these produced by vibrational processes based on their appearance, and from their general location in the spectrum. These bands occur mostly in the ultraviolet, and extend with diminishing frequency into the visible, but rarely appear in the infrared. The usual limit is an iron band near 1000 nm (Irons et al., 1989; Hunt and Salisbury, 1970). On the other hand, very sharp bands in the near-infrared region are also observed. The frequency of occurrence and intensity of these bands decreases towards the visible range (Hunt and Salisbury,

1970). Below we briefly illustrate the effects of the main chemical chromophores on soil reflectance.

2.1.1 Water

Reflectance spectra of moist soils show prominent absorption bands centred at 1400 and 1900 nm. These bands, along with weaker bands at 970, 1200 nm, and 1777 nm, are attributable to overtones and combinations of fundamental vibrational frequencies of water molecules in the soil. In addition to absorption bands, increasing moisture content generally decreases soil reflectance across the entire reflectance spectrum (Irons et al., 1989).

2.1.2 Organic matter

Organic matter has spectral activity throughout the entire VNIR-SWIR region, especially in the visible region. In general, the spectral reflectance decreases in the entire wavelength range between 400–2500 nm as the organic matter content increases (Hoffer and Johannsen, 1969). Baumgardner et al. (1970, 1985) observed that organic matter plays an important role on soil reflectance when its content exceeds 2% and that the reflectance spectra of soils rich in organic matter often have a concave shape between 500 and 1300 nm, compared with the convex shape of the spectra of soils with low organic matter contents. Due to the strong influence of organic matter in the visible region, a soil becomes darker with increasing organic matter. However, many other soil properties, such as texture, structure, moisture, and mineralogy, can influence this (Hummel et al., 2001), implying that darkness would only be a useful discriminator within a limited geological variation.

Absorptions band by organic in the vis–NIR are often weak and not readily apparent to the naked eye (Stenberg et al., 2010). These bands result from the stretching and bending of NH, CH, and CO groups (Ben-Dor et al., 1999; Bokobza, 1998; Goddu and Delker, 1960).

Bands around 1100, 1600, 1700 to 1800, 2000, and 2200 to 2400 nm have been identified as being particularly important for soil organic carbon (Ben-Dor and Banin, 1995; Dalal and Henry, 1986; Krishnan et al., 1980; Henderson et al., 1992; Morra et al., 1991; Malley et al., 2000; Stenberg et al., 2010). Clark et al. (1990) assigned bands near 2300, 1700, and 1100 nm to combination bands and first and second overtones, respectively, of the C–H stretch fundamentals near 3400 nm.

2.1.3 Minerals

Clay minerals, iron-oxides, and carbonates are the most important minerals affecting

spectral reflectance.

Kaolinite, smectite, and illite are the most abundant clay minerals in soils, particularly those from the Mediterranean region (Torrent, 1995). Kaolinite has characteristic absorption doublets near 2200 and 1400 nm. The absorptions wavelengths near 1400 nm (1393 and 1415 nm) are due to overtones of the O–H stretch vibration near 2778 nm, while those near 2200 nm (2165 and 2207 nm) are attributed to Al–OH bend and O–H stretch combinations.

Smectite has sharp characteristics absorptions bands near 1400, 1900, and 2200 nm. The band near 1400 nm is partly due to the overtone of structural O–H stretching in the octahedral layer of this clay mineral. This band, along with that near 1900 nm, is also attributed to vibrational motions in the water molecules bound in the interlayer lattices, in the form of water adsorbed on particle surfaces and hydrated cations (Bishop et al., 1994). Such water is not present in kaolin. Therefore, the presence of a weak absorption band near 1900 nm may be used as a diagnostic feature for kaolinitic dry soils. Illite also shows absorption bands near 1400, 1900, and 2200, which, however, are much weaker than those of smectite as well as near 2340 and 2445 nm (Post and Noble, 1993). The latter could be used to distinguish between illitic and smectitic soils (Post and Noble, 1993). However, they are weak, and especially the one near 2445 nm may be confused with absorptions due to organic matter. The spectral response of soils is strongly affected by the presence and abundance of iron oxyhydroxides (Leone, 2000a, 2000b, 2011).

Goethite (α–FeOOH) and haematite (α–Fe_2O_3) are, by far, the most common Fe-oxide mineral in soils (Torrent et al., 2007; Zhao et al., 2017). Both these minerals show broad and smooth absorption features in the visible-near infrared region, due to electronic transition. The spectra of goethite exhibit absorption bands in the near-infrared, near 920 nm and four absorption bands in the visible, near 420, 480, 600 nm (Leone, 2011). A band near 1700 nm can also be observed (Zheng et al., 2016). The spectra of haematite are characterised by three main absorption bands near 520, 650 and 880 nm (Viscarra Rossel and Behrens, 2010; Viscarra Rossell et al., 2010; Leone, 2011). The absorptions in the visible region cause the vivid colours of Fe oxides, for example, yellow goethite and red haematite (Stenberg et al., 2010).

Carbonates have several absorption bands in the short-wave infrared, due to overtones and combination bands CO_3 fundamental (Clark et al., 1990). The strongest band occurs near 2335 nm, but some weaker absorptions occur near 2160, 1990 and 1870 nm. In addition to chemical chromophores, as discussed above, the reflectance of light from the soil surface is dependent on several physical chromophores. Among these, soil texture, that is the size distribution of the soil mineral particles, plays a

significant role. As reported in Irons et al. (1989), the reflectance generally increases, and contrasts of absorption features decrease as particle size decreases. This behaviour is characteristic of transparent materials, and most silicate minerals behave transparently in the short-wave region. In contrast, the reflectance of opaque materials decreases as particle size decreases.

In common experience, clayey soils often appear darker than sandy soils even though primary clay particles are much smaller than sandy grains. The difference may be explained in part by the different mineralogies of clay and particles but may also be due to the tendency of clay particles to aggregate.

2.2 Principal Component Analysis

For reasons of completeness, we include a brief and modern treatment of principal component analysis (PCA). In what follows the main reference is Hastie et al. (2009), while references more specifically oriented to the type of problem we are dealing with are Odlare et al. (2005) and Viscarra Rossel and Chen (2011).

Suppose that the normalized spectra have been arranged into a rectangular matrix $X \in \mathbb{R}^{N\times p}$, in which N represents the number of sampled spatial locations, while p is the number of wavelengths at which the spectra have been sampled. The data matrix is written in extended form as:

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}, \tag{1}$$

which can be interpreted as a collection of N points embedded into a p-dimensional Euclidean space ($p \leq N$). We want to represent these data points through the following reduced rank (rank $q \leq p$) affine model:

$$f(\lambda) = \mu + V_q\lambda, \tag{2}$$

where $\mu \in \mathbb{R}^p$, $V_q \in \mathbb{R}^{p\times q}$ and $\lambda \in \mathbb{R}^q$ is a q-dimensional parameter. In this representation, columns of V_q are assumed to be orthonormal; in other words, they are an

orthonormal basis of $\mathrm{span}(V_q)$. This affine approximation must be optimal with respect to the Euclidean normalization, minimizing the reconstruction error:

$$\min_{\mu,\{\lambda_i\},V_q} \sum_{i=1}^{N} \left\| x_i - \mu - V_q \lambda_i \right\|^2. \tag{3}$$

It can be easily proved that a partial solution of the above optimization problem is given by the following expression (Hastie et al., 2009):

$$\begin{aligned} \hat{\mu} &= \bar{x}, \\ \hat{\lambda}_i &= V_q^T (x_i - \overline{x}), \end{aligned} \tag{4}$$

where $\bar{x} \in \mathbb{R}^p$ is the p-dimensional vector of column means of the data matrix X. This solution is partial in the sense it depends on the orthogonal matrix V_q; it can be proved that an explicit expression for V_q can be obtained exploiting the following Singular Value Decomposition (SVD):

$$X = UDV^T. \tag{5}$$

In this decomposition we assume that X has been preliminarily centred with respect to column means, $U \in \mathbb{R}^{N \times p}$ and $V \in \mathbb{R}^{p \times p}$ are matrices with orthonormal columns $\left(U^T U = I_p = V^T V\right)$, with $I_p \in \mathbb{R}^{p \times p}$ the identity matrix) and $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix of non-negative entries. The columns of V are called the right singular vectors of X and are the eigenvectors of the matrix $X^T X \in \mathbb{R}^{p \times p}$ associated with its non-zero eigenvalues. The columns of U are called the left singular vectors of X and are the eigenvectors of the matrix $XX^T \in X^{N \times N}$ that correspond to its non-zero eigenvalues. The diagonal elements of matrix D are called the singular values of X and are the non-negative square roots of the (common) non-zero eigenvalues of both matrix $X^T X$ and matrix XX^T. We assume that the diagonal elements of D, are in decreasing order and this uniquely defines the order of the columns of U and A (except for the case of equal singular values). Principal components are the columns of the following matrix:

$$\mathbb{Z} = XV = UD. \tag{6}$$

Columns of the $\mathbb{Z}$ matrix are also known as "scores", as they represent the coordinates of original spectra re-expressed in a new coordinate system obtained through

a rotation of the Euclidean space. Given the rank $q \leq p$, the solution V_q is given by the first q columns of V, and the columns of XV_q are denoted as PC1, PC2, and so on. Orthogonality of principal components is an immediate consequence of orthogonality of right singular vectors (Golyandina and Korobeynikov, 2014).

Let $S_{\mathbb{Z}}$ denote the sample variance-covariance matrix of $\mathbb{Z}$. The trace of $\mathbb{Z}$ coincides with the sum *v* of the variances of the original variables:

$$\mathcal{s} = \sum_{i=1}^{p} s_i = trace(S_{\mathbb{Z}}), \tag{7}$$

in which s_i denotes the sample variance of the *i*-th principal components. It is also well known that $s_1 \geq s_2 \geq \cdots \geq s_p$; based on this property, if we define the cumulative variance as:

$$\mathcal{s}_q = \sum_{i=1}^{q} s_i, \tag{8}$$

the following percentage measures the quota of variance of original variables explained by the first $q \leq p$ principal components:

$$\frac{\mathcal{s}_q}{\mathcal{s}} \times 100\%. \tag{9}$$

There are several heuristics to select a proper number q^{opt} of principal components to retain. The most common consists in taking the first q^{opt} right singular vectors to capture at least a fixed quota *f* of the total variance (e.g.: $f = 0.95$ or $f = 0.98$). Other more formal methods do indeed exist, even though they are not used here; the interested reader is sent back to the literature on the subject (Cadima and Jolliffe, 2001; Jolliffe and Cadima, 2016; Orestes Cerdeira et al., 2020).

A physical interpretation of principal components can be based on right singular vectors (or simply eigenvectors) contained in the matrix V, whose elements are also referred to as 'loadings'. The *i*-th principal component is a linear combination of

original variables, with weights taken equal to the loadings of the corresponding eigenvector:

$$Xv_i = \left(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_p\right)\begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{ip} \end{pmatrix} = v_{i1}\tilde{x}_1 + v_{i2}\tilde{x}_1 + \cdots + v_{ip}\tilde{x}_1, \qquad (10)$$

where the data matrix X has been written in terms of its p columns (variables).

By analysing the profile of the loadings associated with each principal component it is often possible to reconstruct their meaning, since loadings represents the weight assumed by a specific wavelength in contributing to the determination of the corresponding principal component. However, sometimes it may not be easy to extract a clear interpretation from loadings, and if additional variables measuring soil composition are available it is preferable to assess the association between the scores of principal components and soil variables. This is the approach followed, for example, by Odlare et al. (2005), but see also Viscarra Rossel and Behrens (2010). Sparse principal components (Zou et al., 2006) is another interesting possibility to improve interpretability, reducing the number of explicitly used variables by artificially setting to zero the loadings having absolute values smaller than a predetermined tolerance.

2.3 Geostatistics

The objective of this section is to introduce a model that describes the spatial variation of each principal component $Z_j, j = 1,2, \ldots, q^{opt}$, that simultaneously allows to predict with minimum mean square prediction error its value at spatial locations that have not sampled. For this purpose, each principal component is considered as a regionalized variable:

$$Z_j \equiv \left\{Z_j(s_i)\ ;\ i = 1,2, \ldots N,\ \ s \in D \subset \mathbb{R}^2\right\}, \quad j = 1,2, \ldots, q^{opt}. \qquad (11)$$

where D is the study area. In this way, observed principal components are considered as a realization of the random function:

$$\{Z_j(s): s \in D \subset \mathbb{R}^2\}. \qquad (12)$$

The model used here to describe the spatial variation is the linear model of

regionalization (Wackernagel, 2013):

$$Z_j(s) = \mu_j(s) + W_j(s) + \varepsilon_j(s), \tag{13}$$

where $\mu_j(\cdot) = \mathsf{E}(Z_j(\cdot))$ is a deterministic function of spatial coordinates and represents the large scale expected variation, whereas component $\varepsilon_j(\cdot)$ represents the irreducible error, modelled as a spatial White Noise with null expected value and uncorrelated with $W_j(\cdot)$. The function $W_j(\cdot)$ is an intrinsically stationary random function, whose increments are second-order stationary, for which it is possible to define a semi-variogram:

$$\begin{aligned} \gamma_j(h) &= \frac{1}{2}\mathsf{Var}\left[W_j(s+h) - W_j(s)\right] = \\ &= \frac{1}{2}\mathsf{E}\left[\left(W_j(s+h) - W_j(s)\right)^2\right]. \end{aligned} \tag{14}$$

The semi-variogram $\gamma_j(h)$ changes as the length and direction of vector *h* change, but it does not depend on its point of application. If $\gamma_j(h)$ depends on *h* only through its length $||h||$ (Euclidean norm), the random function $W_j(\cdot)$ is said to be isotropic. Properties of theoretical variograms $2\gamma_j(\cdot)$ are well known, particularly those for whom a function of a spatial increment *h* is a valid theoretical variogram (Gaetan and Guyon, 2010). We will also always assume that the theoretical semi-variogram is continuous at zero in the isotropic case. This assumption is equivalent to assume that the random function $W_j(\cdot)$ is L_2-continuous (or mean-square continuous). Without insisting on mathematical details, L_2-continuity allows to treat $W_j(\cdot)$ as a random function modelling the local variation on small scale of soil properties, and guarantees the existence of the spatial correlation function as a convenient tool to characterize the soil microstructure (Cressie, 2015).

The optimal linear predictor $Z(\cdot)$ has known expression, either when $\mu(\cdot) = \mathsf{E}\big(Z(\cdot)\big) = b_0$ or when it has a non-stationary structure, such as $\mu(\cdot) = b_0 + b_1 s_x + b_2 s_y$, where $s = (s_x, s_y)$ are spatial coordinates. The empirical counterparts of the optimal linear predictor, under each one of these two scenarios, correspond to ordinary kriging and universal kriging, respectively. The kriging equations for estimating the optimal linear predictor presuppose that the functional form of the semi-variogram is known, except for a finite number of parameters. The theoretical semi-variogram has therefore to be replaced by a consistent estimate, and this fact causes several mathematical difficulties, because we do not have theoretical guarantees that the empirical predictor remains optimal in the sense of mean square prediction error

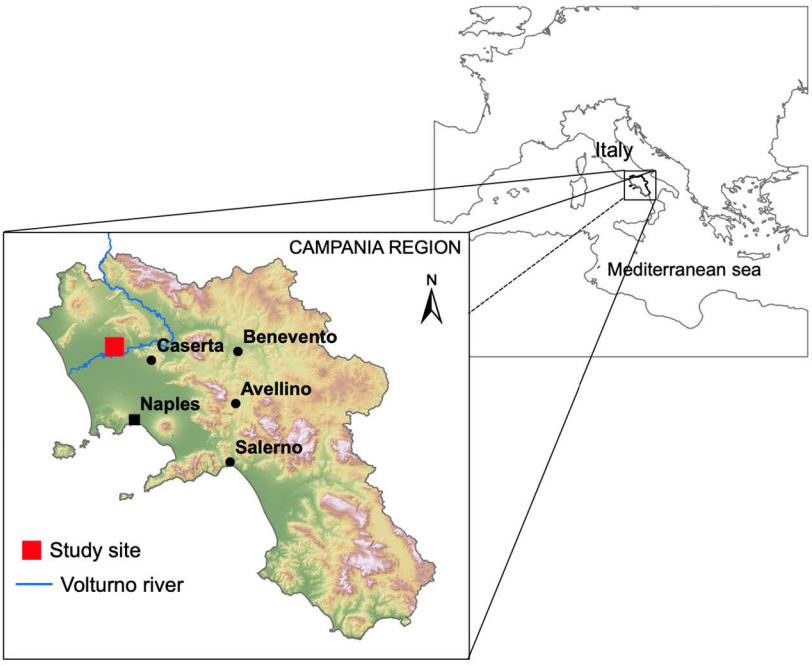
(Bivand et al., 2013). A second difficulty is that kriging has not excellent predictive performances when the likelihood of the data is not Gaussian. In this case it is often convenient to transform the underlying random process, taking for example the logarithm of observed values (as principal components can assume negative values, an offset might eventually be added to ensure that all scores become strictly positive; Varouchakis et al., 2012).

3. Materials and methods

3.1 Site description and sampling

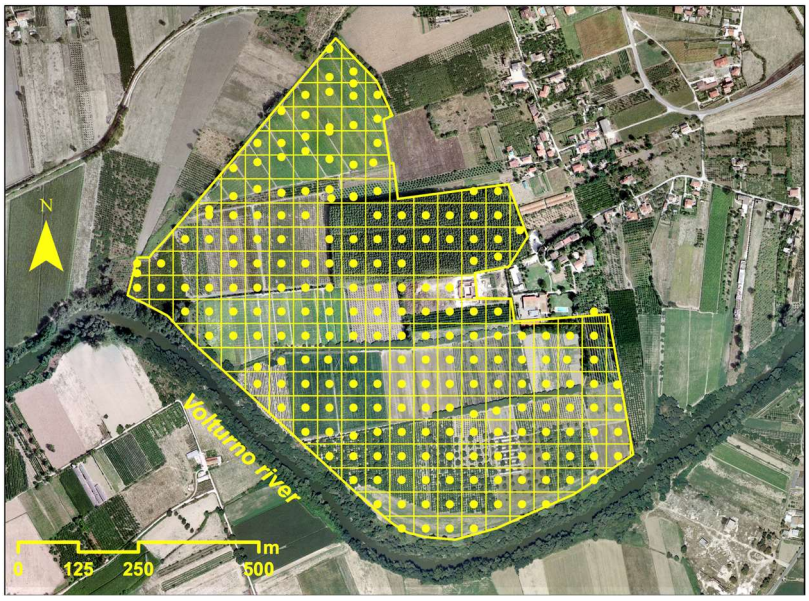
The study covered the entire area (60 ha) of a farm located in the north-western part of the Campania region (Fig. 1), within the municipality of Capua (province of Caserta). This area falls within an abandoned meander of the lower course of the F. Volturno (Aucelli et al., 2014). The production system is mainly oriented towards fruit and cereal growing. The main soil types are Haplic and Fluvic Cambisols and Haplic Luvisos (Grilli wt al. 2014; FAO-WRBSR, 2014).

Figure 1. Localisation of the study area (41°06'09" N, 14°11'20" E).



Within the farm, surface soil samples were collected, at a depth of 20-30 cm, at 240 geo-referenced sites (Fig. 2), more or less regularly spaced, falling approximately in the centre of a 50 × 50 m grid. After collection, the soil samples were transported to the laboratory, air-dried, sieved (2 mm), and finely ground before being subjected to reflectance measurements.

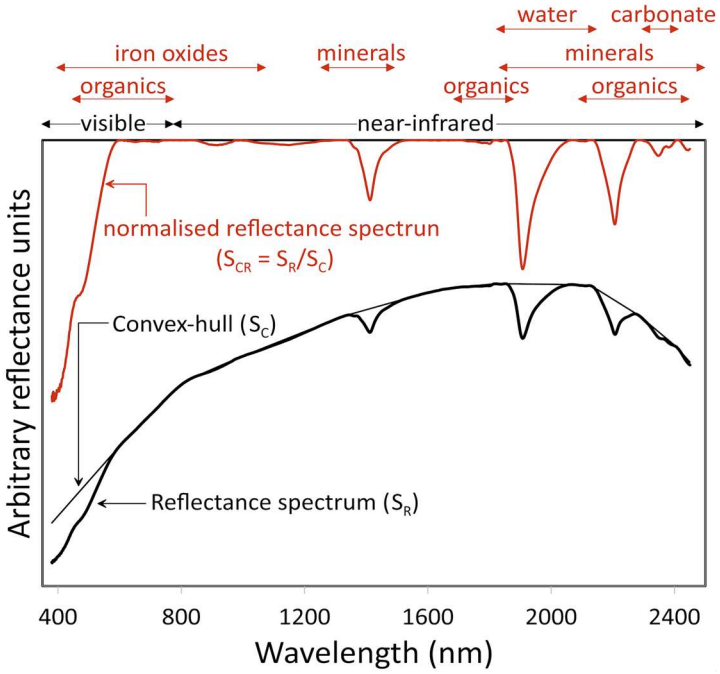
Figure 2. Sampling grid of the investigated field.



3.2 VIS-NIR spectroscopy

The diffuse vis–NIR spectral reflectance was measured in the laboratory, on a residual fraction of soil samples, under controlled light conditions, using the procedure described in Leone et al. (2019). Noisy portions of the measured reflectance spectra, between 350 and 399 nm and between 2451 and 2500 nm, were removed, leaving spectra in the range of 400-2450 nm for the analysis. The resulting reflectance spectra were normalized, using the continuum removal approach (Clark and Roush, 1984), see Fig. 3. To this end, a convex hull was fitted over the original spectral curve, and the absorption spectrum was then calculated considering the ratio between the original reflectance spectrum and the enveloping curve (de Jong, 1992; van der Meer, 1999).

Figure 3. Sample soil vis–NIR spectrum displayed as percent reflectance, convex-hull and continuum removed reflectance. The plot shows regions of the spectrum that hold important information on soil constituents.



3.3 Data pre-processing

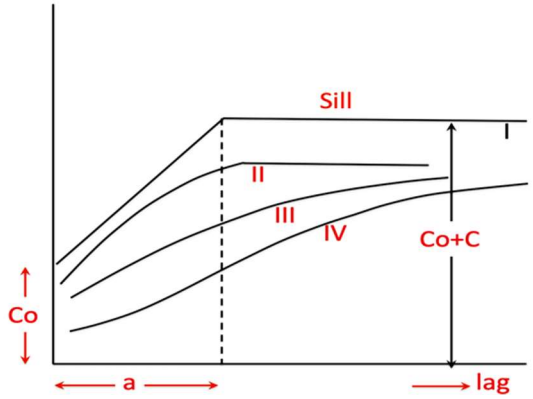
The vis-NIR spectra hold redundant information due to the high degree of correlation between neighbouring wavelengths. For this reason, PCA was performed on the normalized spectra, from which their means (centred data) were subtracted (Viscarra Rossel and Chen, 2011). The initial data were not standardized to the unit of variance since all wavelengths were referred to the same unit, and the differences in their variability were relevant in themselves.

The results of PCA condense the information contained in the spectra. The loadings describe how much each variable contributes to a particular principal component. The PCA reduces the dimensionality of the data, in this case, the reflectance, in a few components, which describe most of the original variance. The first component explains most of the variance, while the subsequent components explain a smaller, progressively decreasing portion.

3.4 Spatial data analysis

The spatial patterns of the first three principal components were analysed using geostatistical analysis. For each principal component, a semi-variogram $\gamma_j(h)$ was estimated which provides a means of quantifying the spatial variation of a variable by measuring the degree of correlation between sampling points separated by a given distance (Webster and Oliver, 2007). The typical parameters on which a theoretical semi-variogram model depends are nugget, range, and sill (Fig. 4).

Figure 4. Examples of semi-variograms: I Linear; II Spherical; III exponential; IV Gaussian. The model parameters nugget (Co), sill (Co+C) and range (a) are shown.



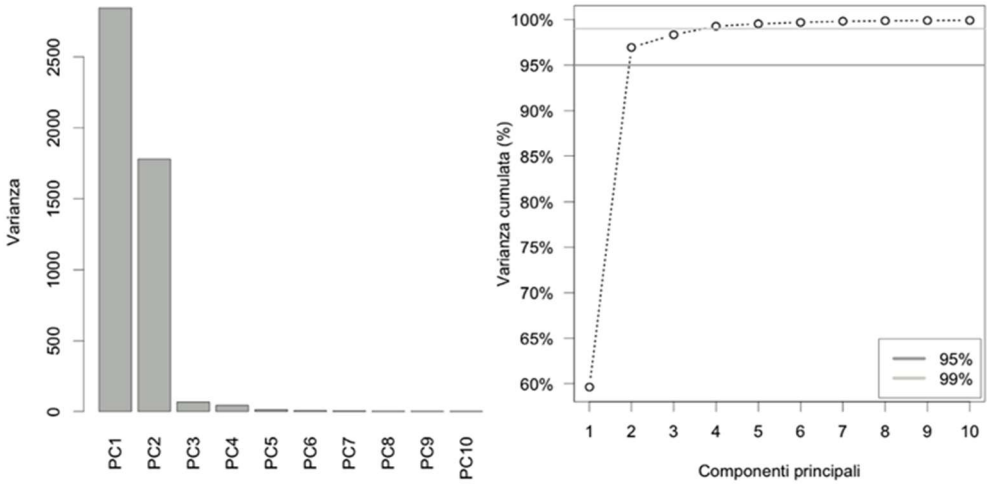
The nugget corresponds to the positive intercept with the y-axis. It is interpretable as the effect of measurement errors, also due to the finite scale at which the phenomenon is observed. Therefore, it increases as the inaccuracy of the measurements increases, i.e., when the sampling interval is too wide. The range is the minimum separation distance at which observations no longer exhibit any spatial correlation. The sill is the point where the theoretical semi-variogram model reaches a limit value (possibly asymptotically) and measures the total variability of the phenomenon. The estimation of the parameters of the theoretical model of semi-variogram is carried out by placing the model in question next to an empirical semi-variogram calculated on data (Cressie, 2015). The estimated theoretical semi-variogram is used to produce a digital map in which the phenomenon under study (in this case, the first three main components analysed separately) is spatially interpolated even outside the sampling sites, through the optimal linear predictor known as kriging.

All the analyses presented next paragraph were carried out using the R 3.6.3 software (R Core Team, 2020).

4. Results and discussions

About 98% of the variance of the original spectral data set is explained by the first three principal components (Fig. 5). In particular, PC1 and PC2 are capable to describe 96.94% of the overall variance, while passing to the first three components together, the overall variance increases to 98.34% (against a modest increase to 99.27% obtained using four components). For this reason, only the scores of the first three principal components were used as input for subsequent analyses.

Figure 5. Variances of the first 10 principal components (a), and percentage of the cumulative variance described by the first 10 principal components. Horizontal bars indicate the number of components which reproduce at least 95% or 99% of the original variability of the normalized vis–NIR spectra (b).



The frequency distribution of the scores of the first three principal components is shown in Fig. 6; from this figure, it is evident that the scores of PC2 present a moderate degree of negative skewness and kurtosis. This reminds us of the probable need for a preliminary logarithmic transformation. It is also evident the presence of some presumable anomalous values (particularly in PC2) located in the left tail of the distribution.

Figure 6. Frequency distribution of scores of the principal components PC1, PC2 and PC3. The empirical coefficient of asymmetry (Skew) and that of kurtosis (Kurt) is also reported above each graph.

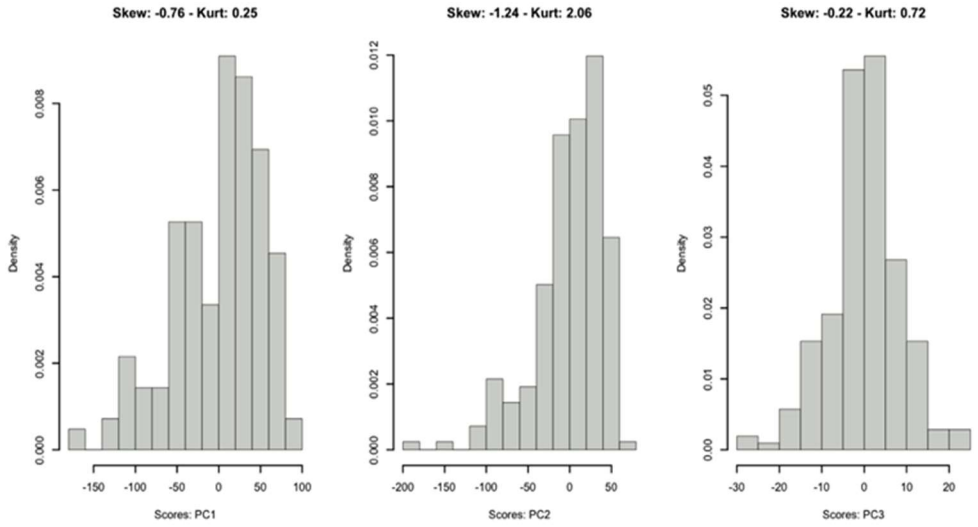
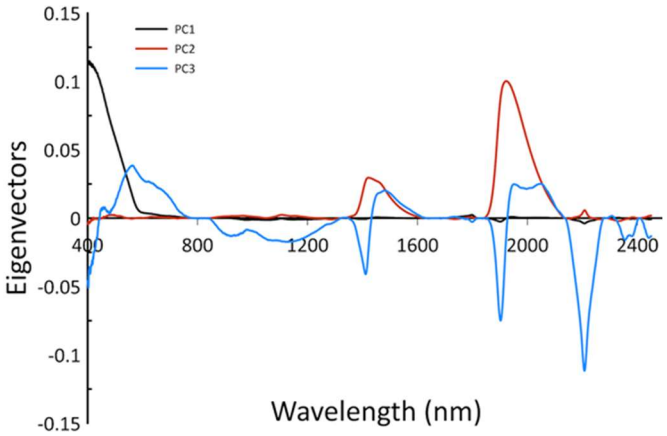


Figure 7 shows the eigenvectors of the first three principal components. The eigenvector of the first principal component showed a steep increase toward the blue and ultraviolet wavelengths, mainly due to a strong iron–oxygen charge transfer band, associated with the presence of iron–oxides, that extend into the ultraviolet (Hunt, 1980). The higher values of loadings in the visible range of the first principal component might be partly due to soil organic carbon (McCauley et al., 1993; Shonk et al., 1991). The eigenvector of the second principal component was dominated by positive loadings near 1400 and 1900 nm, which might be due to 1:1 layer clay minerals (mainly, smectite), specifically to structural O–H stretching mode in its octahedral layer (1400 nm) and combination vibrations of water bound in the interlayer lattices as hydrated cations and water adsorbed on particle surfaces (1400 and 1900 nm) (Bishop et al., 1994; Clark et al., 1990).

Finally, the eigenvector of the third principal component had negative loadings near 1400 and 1900 nm (mainly due to smectite, as previously discussed) and near 2200 nm, due to Al-OH bend in the lattice of 1:2 layer clay minerals (mainly kaolinites) (Clark et al., 1990). Table 1 shows the model parameters of the semi-variograms used for the spatialization of the first three principal components.

Figure 7. Eigenvectors (loadings) of the first three principal components (PC_1, PC_2, PC_3).



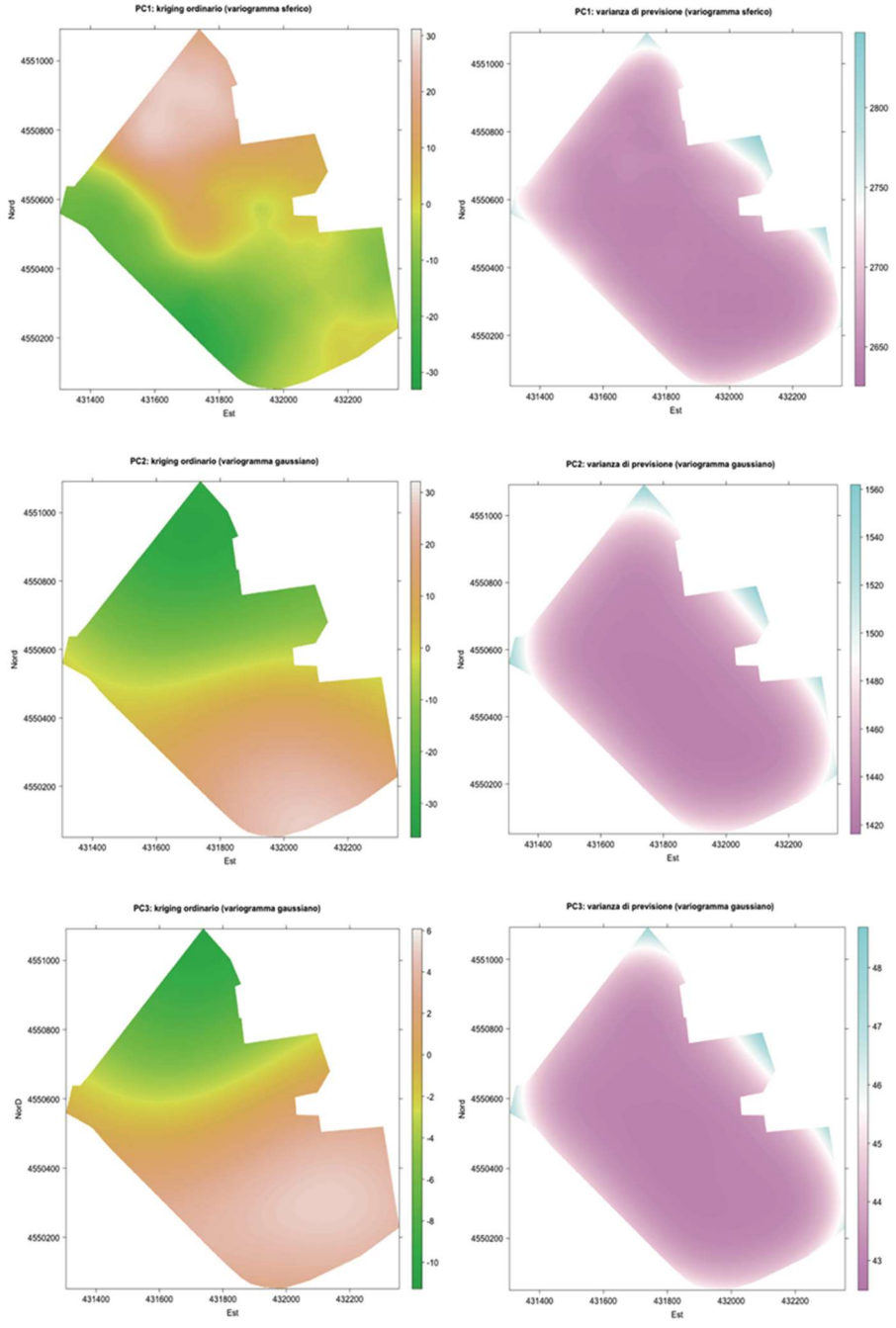
The estimated parameters (using the generalized least squares method; Cressie, 2015) of the theoretical semi-variogram models chosen for the three components were reported in Table 1: the empirical semi-variogram was estimated up to 80% of the maximum theoretically possible distance within the map being studied. The choice of theoretical semi-variogram models, has been largely justified on the basis of an assessment of the goodness of adaptation made 'by eye', even though more sophisticated approaches are indeed possible. To map the information content of the main components, we performed the spatial prediction by means of ordinary kriging on a regular discrete grid containing about 2.23 x 106 points.

Table 1. Variogram model parameters of the first three principal components. The Gaussian model reaches its sill asymptotically.

Variable	Model	Nugget	Sill	Range	C/(C+Co)
PC1	Spherical	2510	3022	710	0.17
PC2	Gaussian	1400	2390	590	0.41
PC3	Gaussian	42	105	600	0.60

Fig. 8 shows the spectral maps resulting from the spatialization of the scores of the first three principal components. We have reported the optimal linear forecast calculated on the grid (on the left), as well as the corresponding variance of the prediction error (on the right). The quality of the forecasts is somewhat stable across the map, and, as expected, only a modest decline towards the edge of the area under study is highlighted.

Figure 8. On the left: kriging maps of the first three principal components (using variogram models as reported in Table 1). On the right: maps of the relative kriging prediction variance.



The first principal component, as previously discussed, mainly represents the content of iron oxides (Fig. 7). Therefore, we can affirm that the soils of the southernmost area of the investigated company surface (higher PC1 scores), morphologically higher, have higher contents of oxides than iron. This hypothesis is consistent with the geochemical dynamics of soil Fe, strongly influenced by the redox conditions of the medium. In oxidizing conditions, more frequent in the morphologically higher areas of the study area, Fe tends to become insoluble, forming oxyhydroxides. Under reducing conditions, determined by conditions of prolonged water stagnation, more frequent in the morphologically more depressed areas of the study area, the iron compounds dissolve readily, freeing Fe2+.

The second principal component represents above all the clay mineral contents of the smectite group. Therefore, the northernmost areas (higher values of the scores) of the study area are likely to be those with the highest clay mineral contents. The third principal component, as already mentioned, is instead inversely related to the clay mineral contents, both of the smectite group and the kaolinite group. This result further confirms the increasing trend of these minerals and, probably, of the finer particle size fractions proceeding from south to north of the company surface. The spectral patterns of the maps relating to the first three principal components (Fig. 8) are consistent with the variability of the soils and their properties, as previously outline in a traditional soil study (Grilli et al., 2014).

5. Conclusions

Vis-NIR spectroscopy, coupled with multivariate statistics and geostatistics, is a useful tool for mapping the spatial variability of soils (digital soil mapping). The reflectance spectra in the Vis-NIR domain contain relevant information on the chemical, physical and mineralogical properties of soils. Multivariate statistical analysis, particularly principal component analysis, is an important tool to condense the highly interrelated reflectance values into a few synthetic variables (principal components), uncorrelated to each other. The geostatistical analysis allows spatializing the principal components and produces spectral maps, which can be interpreted in the light of the known relationships between reflectance and soil properties.

In this study, the spectral maps of the first three principal components have been realized and interpreted. Future studies will be necessary to combine the information contained in the principal component maps, possibly in combinations with other

digital maps (e.g. morphometric maps) using classification and/or data-fusion methods, to produce discretized maps of soil variability, most useful for practical uses.

References

Aucelli P., Filocamo F., Leone N., Leone A.P. (2014). I paesaggi del Basso Volturno, in Leone, A.P., Buondonno A., Aucelli, P.P.S. (Eds): *Paesaggi e suoli del Basso Volturno per una frutticultura innovativa*. Iuorio Edizioni, Benevento, pp. 8–40

Baumgardner, M., Kristof, F.S., Johannsen, C.J., Zachary, A.L. (1970). Effects of organic matter on the multispectral properties of soils. *Proceedings of Indian Academy of Sciences*, 79:413–422

Baumgardner, M., Silva, L., Biehl, L., Stoner, R. (1985). Reflectance properties of soils. *Advances in Agronomy*, 38:1–44

Ben-Dor, E., Banin, A. (1995). Near infrared analysis is a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, 59: 364–372

Ben-Dor, E., Irons, J.R., Epema, G. (1999). Soil reflectance, in Renzc, A. N. (Ed.): *Remote Sensing for the Earth Sciences*. John Wiley & Sons, New York, pp. 111–188

Bishop, J.L., Pieters, C.M., Edwards, J.O. (1994). Infrared spectroscopic analyses on the nature of water in montmorillonite. *Clays and Clay Minerals*, 42:702–716

Bivand, R.S., Pebesma, E., Gomez-Rubio, V. (2013). *Applied spatial data analysis with* R, *2nd Edition*. Springer New York

Bokobza, L. (1998). Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 6:3–17

Brady, N.C., Weil, R.R. (2002) *The Nature and Properties of Soils, 13th Edition*. Pearson Education, Inc., Upper Saddle River, 960

Cadima, J., Jolliffe, I.T. (2001). Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(1):62–79

Clark, R.N., Roush, T.L. (1984). Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, 98(B7): 6329–6340

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

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons (Wiley Series in Probability and Statistics)

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

de Jong, S.M. (1992). The analysis of spectroscopical data to map soil types and soil crusts of Mediterranean eroded soils. *Soil Technology*, 5(3):199–211

Drury, S.A. (1993). *Image Interpretation in Geology*, *2nd Edition*. Chapman & Hall, London

FAO-WRBSR (2014). International union of soil science (IUSS) working group world reference base. World reference base for soil resources. In: *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. FAO, Rome, Italy, pp. 192

Gaetan, C., Guyon, X. (2010). *Spatial Statistics and Modeling*, Springer-Verlag New York

Goddu, R.F., Delker, D.A. (1960). Spectra-structure correlations for the near-infrared region. *Analytical Chemistry*, 32:140–141

Golyandina, N., Korobeynikov, A. (2014). Basic Singular Spectrum Analysis and forecasting with R. *Computational Statistics & Data Analysis*, 71:934–954

Grilli, E., Leone, A.P., Buondonno, A. (2014). I suoli dell'Azienda GiòSole, in Leone, A.P., Buondonno, A. and Aucelli, P.P.C. (Eds.): *Paesaggi e Suoli del Basso Volturno per una Frutticoltura Innovativa*. Grafiche Iuorio, Benevento.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*, *2nd Edition*. Springer New York

Henderson, T.L., Baumgardner, M.F., Franzmeier, D.P., Scott, D.E., Coster, D.C. (1992). High dimensional reflectance analysis of soil organic matter, *Soil Science Society of America Journal*, 56:865–872

Hoffer, R.M., Johannsen, C.J. (1969). Ecological potentials in spectral signature analysis, in Johnson, P.L. (Ed.): *Remote Sensing in Ecology*. University of Georgia Press, Athens, pp. 1–29

Hummel, J.W., Sudduth, K.A., Hollinger, S.E. (2001) Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. *Computers and Electronics in Agriculture*, 32:149–165

Hunt, G.R., Salisbury, J. (1970). Visible and near-infrared spectra of minerals and rocks. Silicate minerals. *Modern Geology*, pp. 283–300

Hunt, G.R., (1980) - Electromagnetic radiation: the communications link in remote sensing, in Siegal, B.S. and Gillespie, A.R. (Eds.): *Remote Sensing in Geology*. John Wiley & Sons, New York, pp. 5–45

Irons, J., Weismiller, R., Petersen, G. (1989). Soil reflectance, in Asrar, G. (Ed.): *Theory and Application of optical remote sensing*. Wiley, New York, pp. 66–106

Jolliffe, I.T., Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 374(265):2015.0202

Krishnan, P., Alexander, D.J., Butler, B., Hummel, J.W. (1980). Reflectance technique for predicting soil organic matter, *Soil Science Society of American Journal*. 44:1282-1285

Leone, A.P. (2000a). Spettrometria e valutazione della riflettanza spettrale dei suoli nel dominio ottico 400 – 2500 nm. *Italian Society of Remote Sensing*, 19:1-26

Leone, A.P. (2000b). 'Bi-directional reflectance spectroscopy of Fe-oxides minerals in Mediterranean Terra Rossa soils: a methodological approach'. *Agricoltura Mediterranea*. 130:144–154

Leone, N. (2011) *Uso della spettrometria VNIR e della diffrattometria Rx per la caratterizzazione delle forme pedologiche del ferro: analisi comparativa su modelli di sintesi*, unpublished Bachelor Degree thesis, University of Sannio, Benevento, Italy.

Leone, A.P., Viscarra-Rossel, A.R., Amenta, P., Buondonno, A. (2012). Prediction of Soil Properties with PLSR and vis-NIR Spectroscopy: Application to Mediterranean Soils from Southern Italy. *Current Analytical Chemistry*. 8(2):283–299

Leone, A.P., Leone, G., Leone, N., Galeone, C., Grilli, E., Orefice, N., Ancona, V. (2019). Capability of Diffuse Reflectance Spectroscopy to Predict Soil Water Retention and Related Soil Properties in an Irrigated Lowland District of Southern Italy. *Water*, 11(8):1712

Malley, D. F., Martin, P. D., McClintock, L. M., Yesmin L., Eilers, R. G., Haluschak P. (2000). Feasibility of analysing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy, in Davies, A.M.C. and Giangiacomo, R. (Eds.): *Near Infrared Spectroscopy: Proceedings of the 9th International Conference*. NIR Publications, Chichester, UK, pp. 579–585

McCauley, J.D., Engel, B.A., Scudder, C.E., Morgan, M.T., Elliot, P.W. (1993). Assessing the spatial variability of organic matter. *St. Joseph: American Society of Agricultural Engineers*, ASAE Paper No. 93–1555

Morra, M.J., Hall, M.H., Freeborn, L.L., (1991). Carbon and nitrogen analysis of soil fractions using near-infrared reflectance spectroscopy. *Soil Science Society of American Journal*, 55:288–291

Odlare, M., Svensson, K. Pell, M. (2005). Near Infrared Reflectance Spectroscopy for Assessment of Spatial Soil Variation in an Agricultural Field. *Geoderma*, 126(3-4):193–202

Orestes Cerdeira, J., Duarte Silva, P., Cadima, J., Minhoto, M. (2020). subselect: Selecting Variable Subsets. Retrieved from https://cran.r-project.org/package=subselect

Post, J.L., Noble, P.N. (1993). The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. *Clays and Clay Minerals*, 41: 639–644

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Shonk, G.A., Gaultney, L.D., Schulze, D.G., Van Scoyoc, G.E. (1991).Spectroscopic sensing of soil organic matter content. *Transactions of the American Society of Agricultural Engineers*, 34:1978–1984

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

Torrent, J. (1995) *Genesis and properties of the soil of the Mediterranean regions*- University of Naples Federico II, Department of Agro-Chemical Sciences. Arti Grafiche Licenziato, Naples, p. 111

Torrent J., Liu Q., Bloemendal, J., Barron V. (2007). Magnetic enhancement and iron oxides in the upper luochuan Loess-paleosol sequence, Chinese Loess plateau. *Soil Science Society of America Journal*, 71(5):1570–1578

van der Meer, F. (1999). Can we map swelling clays with remote sensing? *International Journal of Applied Earth Observation and Geoinformation*. 1(1):27–35

Varouchakis, E. A., Hristopulos, D.T., Karatzas, G.P. (2012). Improving kriging of groundwater level data using nonlinear normalizing transformations – a field application. *Hydrological Sciences Journal*, 57(7):1404–1419

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(1–2): 59–75.

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

Viscarra Rossel, R.A., Rizzo, R., Demattе, J.A.M. and Behrens, T. (2010). Spatial modelling of a soil fertility index using vis–NIR spectra and terrain attributes. *Soil Science Society of America Journal*, 74:293–1300

Viscarra Rossel, R. A., Chen, C. (2011). Digitally Mapping the Information Content of Visible–near Infrared Spectra of Surficial Australian Soils. *Remote Sensing of Environment*, 115(6):1443–1455

Wackernagel, H. (2013). *Multivariate Geostatistics: An Introduction with Applications*, *3rd edition*. Springer, Berlin Heidelberg

Webster, R. and Oliver, M.A. (2007). *Geostatistics for Environmental Scientists*, *2nd Edition*. John Wiley & Sons, New York

Zhao, L., Hong, H., Fang Q., Yin, K., Wang, C., Li, Z., Torrent, J., Cheng, F., Algeo, T.J. (2017). Monsoonal climate evolution in southern China since 1.2 Ma: new constraints from Fe-oxide records in red earth sediments from the Shengli section, Chengdu Basin. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 473:1–15

Zheng, G,, Jiao, C., Zhou, S., Shang, G. (2016). Analysis of soil chronosequence studies using reflectance spectroscopy. *International Journal of Remote Sensing*, 37(8): 1888–1901

Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(3):265–286