Geoderma 343 (2019) 280–293

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Contribution of EMI and GPR proximal sensing data in soil water content assessment by using linear mixed effects models and geostatistical approaches

Emanuele Barca[a], Daniela De Benedetto[b,*], Anna Maria Stellacci[c]

[a] Water Research Institute (IRSA), National Research Council (CNR), Bari, Italy
[b] Council for Agricultural Research and Economics, Research Centre for Agriculture and Environment, CREA-AA, Via Celso Ulpiani 5, 70125 Bari, Italy
[c] Department of Soil, Plant and Food Sciences, University of Bari "Aldo Moro", via Amendola 165/A, Bari, 70126, Italy

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Kriging with external drift (KED)
Linear mixed effects models (LMM)
Ground penetrating radar (GPR)
Electromagnetic induction (EMI)
Principal component analysis (PCA)

ABSTRACT

The estimation of topsoil water content is of primary interest in the framework of precision farming, but, in general, such assessment is costly and complicated by several interfering factors which do not allow an accurate prediction. Proximal sensing can provide suitable technological facilities to support researchers and technicians in this task. GPR and EMI sensors are valuable instruments as they can provide very informative covariates to be used for improving soil water content estimation. In the present work, it was explored the single (EMI or GPR) and the combined (EMI + GPR) contribution of these proximal data sources. Furthermore, geostatistical (Ordinary Kriging and Kriging with external drift) and linear mixed effects models were applied to compare their respective predictive capabilities. As a result, GPR demonstrated to be more effective in estimating topsoil water content with respect to EMI but, combining both the information, an improvement in the prediction accuracy was observed. Moreover, adding more covariates in the models (GPR outcomes or GPR + EMI outcomes) allowed filtering out the structured spatial component of soil water content. Finally, the statistical approaches proved to behave very similarly, with a slight better performance of Kriging with external drift.

1. Introduction

The assessment of topsoil water content (SWC) spatial variation has received considerable attention in many researches and applications, due to ever larger interest for land use planning, irrigation management, ecological and hydrological modelling. Agricultural and irrigation management practices, especially in semiarid and arid regions, largely depend on a timely and accurate characterization of temporal and spatial soil moisture dynamics in the root zone because of the impact of soil moisture on the production and health status of crops and on salinization (Vereecken et al., 2008). Knowing spatio-temporal distribution of soil moisture and soil water storage capacity is therefore an important asset for the optimization of irrigation under variable environment and for designing new optimal agricultural practices, in the framework of precision agriculture (Adamchuk et al., 2004).

Studies on spatial variability of SWC are typically based on values obtained through direct field sampling and laboratory analyses. Gravimetric method is inefficient in providing rapid data collection; it is also destructive, invasive and requires time-intensive sampling and heavy laboratory work (Kong et al., 2017; Neves et al., 2017). Moreover, assessment of SWC variability is complicated due to soil heterogeneity and interactions with other environmental factors.

A way to overcome the problem is to predict SWC on the basis of the relationships with other auxiliary covariates acquired by alternative soil moisture sensing techniques such as those derived from proximal sensing. In particular, geophysical sensors, such as electromagnetic induction (EMI) and ground penetrating radar (GPR) are gaining increasing popularity as tools for obtaining spatially distributed subsurface data that can be correlated with soil and hydrologic properties (Zhu et al., 2010; Minet et al., 2013; De Benedetto, 2012). Advantages of using these sensors lie in their non-invasive nature, the capability of collecting high resolution information in real time, the relative low cost of data acquisition and the possibility for a mobile survey configuration (Viscarra Rossel et al., 2011).

EMI methods are widely used for soil mapping given the high density "on-the-go" surveys of soil apparent electrical conductivity (EC_a) (Adamchuk et al., 2004) because the main cause of EC_a variability, in a non-saline soil, is the soil moisture content. This property has

* Corresponding author.
E-mail address: daniela.debenedetto@crea.gov.it (D. De Benedetto).

https://doi.org/10.1016/j.geoderma.2019.01.030
Received 25 May 2018; Received in revised form 20 September 2018; Accepted 9 January 2019
Available online 28 February 2019

driven researchers in using EMI data for SWC assessment (Martínez et al., 2010; Martínez et al., 2012; De Benedetto et al., 2013; Huang et al., 2016). GPR allows performing high spatial resolution measurements at the field scale and to bridge the scale gap in terms of spatial resolution and support scale (Grote et al., 2010; Hubbard et al., 2002). GPR amplitude data are related with some soil properties (such as porosity, bulk density, water content and texture), which vary along the soil profile at a very fine scale (Knight et al., 1997; De Benedetto et al., 2012). However, the development and acceptance of GPR as a soil water content sensor is still limited by the difficulty of its application under field conditions and by the complexity of data acquisition and processing.

Several case studies indicate that a successful prediction of soil water content with these techniques requires the use of local empirical or semi-theoretical petrophysical relationships that relate measured geophysical output to soil volumetric water content (e.g., Huisman et al., 2003; Grote et al., 2003; Lunt et al., 2005; Minet et al., 2012; Huang et al., 2017a, 2017b). However, petrophysical relationships, depending by specific local soil conditions (soil texture and/or porosity), need to be carefully calibrated. Steelman and Endres (2011) examined the effect of petrophysical relationships on the estimates of volumetric water content, for different soil textures and with a range of antenna frequencies, and showed that soil water content estimates varied significantly depending on the type of petrophysical relationship, on physical and chemical properties and the frequency used. Therefore, the difficulty of describing quantitatively all the processes occurring in the soil suggests the application of stochastic approaches, where the available sample data are viewed as the result of some random process (Isaaks and Srivastava, 1989). These approaches have significant advantages over deterministic ones: (i) a wide range of data types may be integrated; (ii) the final result does not suffer from artefacts of regularisation in an inversion algorithm; (iii) estimates of model uncertainty may be obtained.

In this framework, sensor datasets can be used as auxiliary information to effectively supplement a sparsely sampled target variable, such as SWC. However, there still remains the problem to identify the best combination of input variables from the available data and to estimate the values of these properties at un- sampled locations.

When using multivariate datasets deriving from different sources, often characterized by multiple outcomes, a critical issue is represented by collinearity and data redundancy that may reduce the efficiency of data analysis and interpretation. In addition, high dimensionality may increase the risk of overfitting (Hira and Gillies, 2015). Feature extraction (linear and non-linear) and feature subset selection methods (filters, wrappers, embedded, hybrid techniques) may allow to reduce data dimensionality by obtaining new uncorrelated variables, which synthesize the greatest part of the information, as well as identifying the variables most informative on the process under study (Hira and Gillies, 2015; Stellacci et al., 2016; Tang et al., 2013).

For the second problem, a number of 'hybrid' interpolation techniques (Hengl et al., 2004; McBratney et al., 2000; Berardi and Vurro, 2016), which combine kriging with exhaustive secondary information, has been developed and tested to obtain reliable predictions and prediction error variances. There are various methods to incorporate secondary information, for example a multivariate extension of kriging, known as cokriging, has also been used for merging primary and secondary information (Goovaerts, 2000; Hengl et al., 2004; McBratney et al., 2000; Berardi et al., 2016). However, this technique assumes intrinsic stationarity, both of the target variable and of more intensively measured variables, besides a strong spatial correlation between the variables (Webster and Oliver, 2001). Another way of taking into account the secondary variable is by assuming a spatial trend in the covariate, which is significantly related to the primary variable. The method, known as kriging with external drift (KED) in the scope of intrinsic random function of order k (IRF-k), introduces the concepts of generalized increments and generalized covariance and assumes that generalized increments produce a second-order stationary process (Bourennane et al., 2000; Hengl et al., 2003; Wackernagel, 2003; Cafarelli and Castrignanò, 2011; Cafarelli et al., 2015).

An alternative is to fit a linear mixed effects model (LMM) where the data are modelled as the additive combination of fixed effects (i.e., the unknown mean and coefficients of a trend model), random effects (the spatially dependent random variation in the geostatistical context) and independent random error (nugget variation in geostatistics) (Lark et al., 2006; Lark, 2012). Residual maximum likelihood (REML) is preferred to estimate variance parameters because they are less biased than both maximum likelihood estimates and method-of-moment estimates obtained from residuals of a fitted trend.

There is still a knowledge gap concerning the budget of information provided by the different proximal sensors for estimating the topsoil water content as well as the best suited statistical approach to be applied for fulfilling the task.

The objectives of this work were to evaluate the single or combined contribution of proximal sensing information provided by EMI and GPR data in estimating SWC and compare the performance of geostatistics and LMM approaches.

2. Materials and methods

2.1. Study area and data collection

The experiment was conducted on a bare surface (40 m × 20 m) in the experimental farm of CREA - Council for Agricultural Research and Economics, located in Rutigliano-Bari (40°59′48.25″N, 17°02′02.06″E), in South-eastern Italy. This study was deliberately confined to a small study area to minimize the influence of several factors other than soil water content.

Soil is classified as fine, mixed, superactive, thermic Typic Haploxeralfs according to the Soil Taxonomy (Soil Survey Staff, 2010). Soil texture is mainly clayey with the clay content ranging from 30% to 60% by weight and basically increasing in depth. Since high contents of clay may strongly attenuate the signal, a preliminary ERT survey was carried out which allowed us to estimate the resistivity of soil (values between 25 and 60 Ωm) and radar energy attenuation (values between 6 and 14 dB/m). These values show that Ground Penetrating Radar can be effectively and reliably used in this study area. The bedrock is constituted of a layered sequence of Cretaceous limestone with some dolomitic limestone level and occurs at variable depth due to its irregularly shaped boundary.

In October 2012, the experimental area was irrigated (drip irrigation) for a week until the saturation and surveys were carried out after water leaching for gravitation was almost negligible. One hundred and sixteen georeferenced samples (Fig. 1) were collected up to 0.30-m depth at the nodes of a square grid with spacing of 4 m and the water content was calculated with gravimetric method. The experimental area was deliberately oversampled in order to put all the involved methods under the optimal conditions.

2.2. Geophysical investigations

The study area was surveyed with an EMI sensor (EM38DD, Geonics Limited, Mississauga, Ontario, Canada), which composes of two units, consisting of a transmitter and receiver coil, mounted perpendicularly to each other, one orientated horizontally and the other one vertically (McNeill, 1980).

Technically, a transmitter and a receiver coil are placed on (or near) the soil surface at a fixed distance from each other, and the transmitter coil is energized with an alternating current. This generates a time-varying magnetic field, which induces electric fields in the soil, which in turn induce a secondary magnetic field. Both the primary and the secondary magnetic fields are sensed by the receiver coil and, under certain geometric conditions indicated as "low induction number"

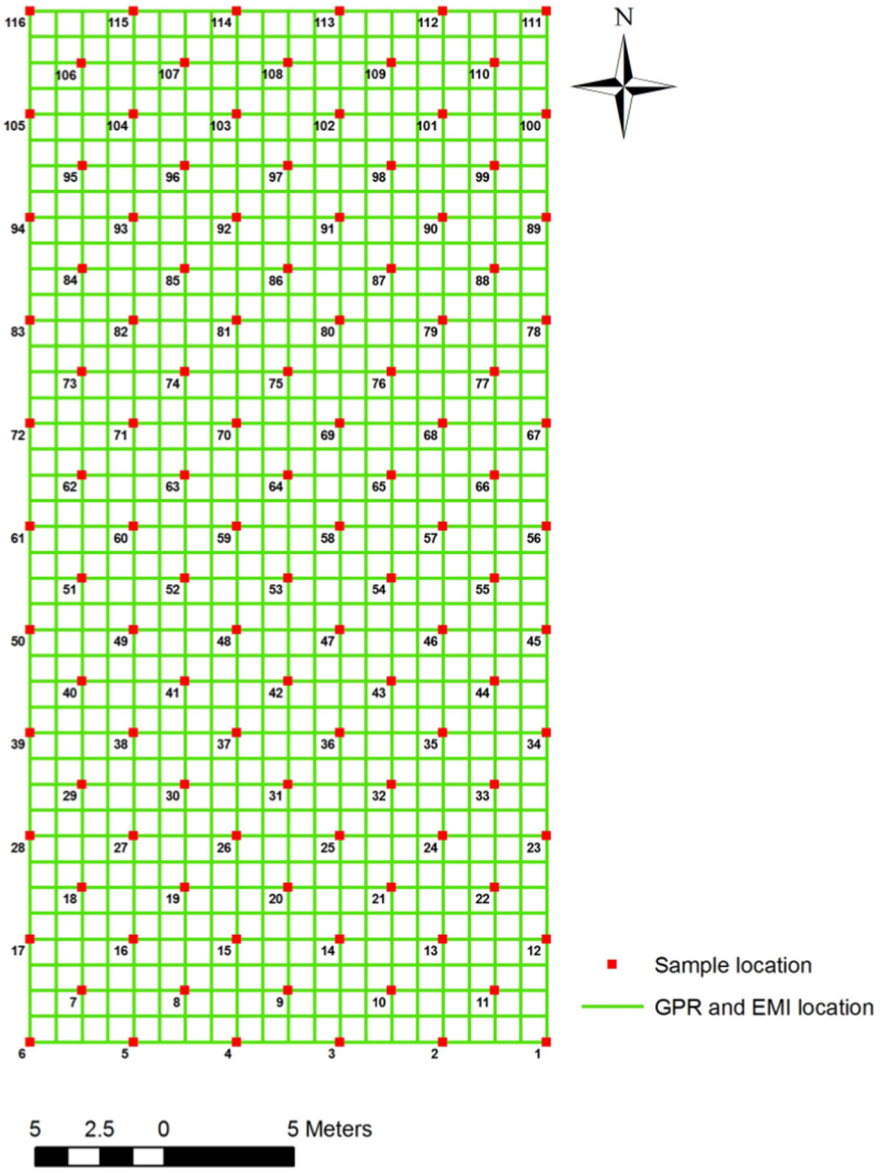

Fig. 1. The experimental area with the SWC sampling locations, at the nodes of a square grid with a 4 m spacing, shown with red dots and the EMI and GPR acquisitions along transects 1 m apart with green lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(McNeill, 1980), the ratio between the primary and the secondary magnetic field can be used to estimate the EC_a of the volume of soil under investigation.

The sensor, carried across the study area, simultaneously measures apparent electrical conductivity (EC_a, expressed in milli-Siemens per meter) in the two orientations: in vertical dipole mode (V), the theoretical maximum sensitivity corresponds to the depth of about 0.40 m and the theoretical maximum investigation depth to about 1.50 m. In horizontal dipole mode (H), the sensitivity of the device decreases with depth down to a theoretical maximum depth of investigation of about 0.75 m (McNeill, 1980).

EC_a measurements were geo-referenced using a Differential Global Positioning System (DGPS, HiPer® Pro, TOPCON) with planimetric and altimetric centimeter accuracies.

The survey was conducted using a non-metallic platform and a wood cover (to avoid any magnetic interference and thermal drift of the sensor) and the sensor was towed behind a tractor along longitudinal and transversal transects approximately 1 m apart (Fig. 1). The sensor was calibrated and zeroed according to the manufacturer instruction before starting measurements. The EC_a in both orientations (EC_a-H and EC_a-V) was recorded every second resulting in a spatial resolution of 0.5 m, on average, along the transect.

GPR data were collected on the EMI transects (Fig. 1) with the common offset reflection method, using two different monostatic GPR systems: one with a central frequency of 250 MHz (Noggin 250 MHz, Sensors & Software Inc., Mississauga, Ontario, Canada), the other with two central frequencies of 600 and 1600 MHz (IDS Ing-manufactured, RIS 2k-MF Multifrequency Array Radar-System).

The GPR produces a short pulse of high frequency (10–1000 MHz) electromagnetic energy, which propagates through the sub-surface materials at the velocity determined by the soil dielectric permittivity. When this propagating wave encounters any change in electrical properties, it is reflected or scattered back towards the surface (Davis and Annan, 1989). The amount of energy that is reflected by an interface is dependent upon the contrast in the relative dielectric permittivity of the two layers. Variations in the electrical properties of soils are usually associated with changes in volumetric water content which, in turn, gives rise to radar reflections (Davis and Annan, 1989). An

important parameter of a GPR system is the vertical (range or depth) resolution defined as the 1/4–1/8 wavelength of the incident radiation (Sheriff and Geldart, 1982).

The 250 MHz GPR system acquired the data using a time window of 100 ns with a sampling interval of 0.2 ns and spacing between the traces of 0.05 m. The 600 and 1600 MHz GPR system worked with a time window of 60 ns and a sampling interval of 0.05 ns; successive traces were collected every 0.024 m. The wavelengths calculated for propagation velocity of 0.06 m/ns were 2.4 m, 1 m and about 0.4 m for 250, 600 and 1600 MHz frequencies, respectively. The coordinates of the initial and final positions of GPR transects were recorded using the DGPS. Detailed discussions of the fundamental principles of GPR can be found in the publications by Daniels et al. (1988) and Davis and Annan (1989).

2.3. Pre-processing of geophysical data

The preliminary data analysis included a data quality check and data cleaning procedure. For EMI data, as the field computer records data every few seconds it is important to remove any points where the instrument was stationary. Then the second check examines the magnitude of the EC_a response, any negative values are removed.

For GPR, processing the raw data consisted in applying a set of filters (De Benedetto et al., 2015) and in extracting quantifiable variables, such as attenuation, to display GPR data in horizontal maps at a specified time (or depth), called amplitude maps or time slices. The enveloped amplitude maps (time slices) were built averaging the amplitude (or the square amplitude) of the radar signal, expressed in digital number (DN), within overlapping time windows of width Δt. Typically Δt must be of the order of the dominant period of the antennas (4 ns, 2 ns and 1 ns for 250, 600 and 1600 MHz antennas, respectively), because GPR reflections are normally taken over a time window of a microwave pulse length. The total time interval was of 10 and 20 ns for 600 and 250 MHz respectively, because this time was comparable with the depth of soil (at 0–0.30 m depth), and of 5.5 ns for 1600 MHz because of the attenuation of radar signal. The time slices were then transformed in depth slice maps using the velocity of the radar waves, equal to 0.06 mn s^{-1} up to 10 ns and 0.1 mn s^{-1} after 10 ns, determined through the analysis of Common-Mid Point measurements with a bistatic system at the frequency of 450 MHz (De Benedetto et al., 2013).

2.4. Principal component analysis

Before geostatistical and linear mixed effects model (LMM) analysis, the covariate variables were subjected to Principal Component Analysis (PCA). To perform the analysis on different sensors, the GPR data were collocated into the less numerous file (2559 points) containing EMI measurements by migrating them to the nearest EMI sample point.

PCA was performed in order to synthesize the information in the multivariate dataset and reducing/avoiding collinearity problems. In particular, to better deepen the relationships within different covariate groups and keep as much information as possible in the subsequent data analysis step, PCA was carried out separately on the following four subsets of variables: GPR data for each frequency (250 MHz, 600 MHz, 1600 MHz) and EMI data.

Principal Component Analysis (PCA) was carried out on the correlation matrix of the geophysical variables to obtain few new uncorrelated components explaining most of the variation of the initial data. The principal components (PCs) with an eigenvalue higher than one were retained according to the Kaiser criterion. Variable loadings were then examined to identify the variables that most contributed to each selected PC and investigate their relationships. PCA was implemented using the FACTOR procedure of the SAS/STAT software (SAS Institute Inc., 2017).

2.5. SWC modelling

With the aim of understanding the contribution of the information provided by the different geophysical sensors, singularly or combined, in estimating the soil water content, different models were fitted to the experimental data of SWC and to the factors extracted through PCA: (a) a model including only the factors extracted by EMI data; (b) a model including only the factors extracted by GPR data; (c) a model including the factors extracted by both EMI and GPR data.

Each model was analysed using linear mixed effects models (LMM) and geostatistical techniques (KED) in order to compare the performance of different approaches. In addition SWC was estimated using ordinary kriging (OK).

Previously to geostatistical and linear mixed model analysis, descriptive statistics were computed for the response variable (SWC) and for the covariate variables (GPR and EMI outcomes), in order to identify the presence of outliers and/or leverage data and synthesize the main features of data distribution. Hypothesis of normality was tested using Shapiro-Wilk and Anderson-Darling statistics. In addition, the relationships between the response variable and the factors extracted through PCA were investigated by means of Pearson correlation coefficients.

2.5.1. Geostatistical analysis

The geostatistical analysis aimed at evaluating spatial heterogeneity of the soil properties by producing maps of the collected data. Soil water content was estimated using ordinary kriging (OK) on covariates locations (2559 points). It is one of the most basic form of kriging in which the unknown value $z(\mathbf{x_0})$ of a given realization of $Z(\mathbf{x_0})$ is predicted from the known values $z(x_i)$ $i = 1, 2, \ldots, N$, at the support points $\mathbf{x_i}$. The ordinary kriging estimator can be written as:

$$z^*_{OK}(\boldsymbol{x}_0) = \sum_{i=1}^{N} \lambda_i z(\boldsymbol{x}_i)$$

where λ_i are weights associated with the N sampling points. The weights are chosen in such a way that the estimator is unbiased, the values are continuous and the estimation error is minimized:

$$\min E\left[(Z^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0))^2\right]$$

under the unbiasedness condition: $\sum_{i=1}^{N} \lambda_i = 1$.

This ensures that kriging is an exact interpolator because the estimated values are identical to the observed values when a kriged location coincides with a sample location.

2.5.2. Spatial linear mixed effects models

Linear mixed effects model (LMM) is a generalization of the classic regression approach which includes, along with the usual explanatory variables called fixed effects, also random effects accounting for data auto-correlation. LMMs are usually applied to observations which violate the residual independence assumption trying to fix it.

The form of a mixed effect model is the following:

$$\mathbf{Y} = \boldsymbol{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

where, $\mathbf{Y}$ is a vector of response observations with the property, $E[\mathbf{Y}|\mathbf{X}] = \boldsymbol{X\beta}$; $\boldsymbol{\beta}$ is a vector of parameters related to the fixed effects; $\boldsymbol{X}$ and $\boldsymbol{Z}$, called *design matrices*, account for the non-random and random predictors, respectively; $\boldsymbol{u}$ is a vector of parameters related to the (spatial) random effects with the property: $E[\boldsymbol{u}] = 0$ and associated variance-covariance matrix $var(\boldsymbol{u}) = \sigma_s^2 \boldsymbol{H}(\phi)$; $\boldsymbol{\varepsilon}$ is a vector of errors with the property $E[\varepsilon] = 0$ and associated variance matrix $\tau^2 \cdot I$.

The spatially correlated random effects vector u is Gaussian and $H(\cdot)$ is defined by a suitable correlation function $f(h, \phi)$, where σ_s^2 is the *partial sill*, ϕ is the *range* and h the *lag* or distance (Pollice and Bilancia, 2002). The random error term ε is $N(0, \tau^2 \cdot I)$, τ^2 is called *nugget effect*. ε is uncorrelated with respect the spatial random effect $\boldsymbol{u}$, that is *Cov*

Table 1
Summary statistics for soil water content (SWC, g 100 g^{-1}).

Variable	N	Mean	Sd	I Qu.	Median	III Qu.	Min	Max	Skewness	Kurtosis
SWC	113	24.89	1.11	24.31	24.95	25.66	21	27.09	−0.64	0.65

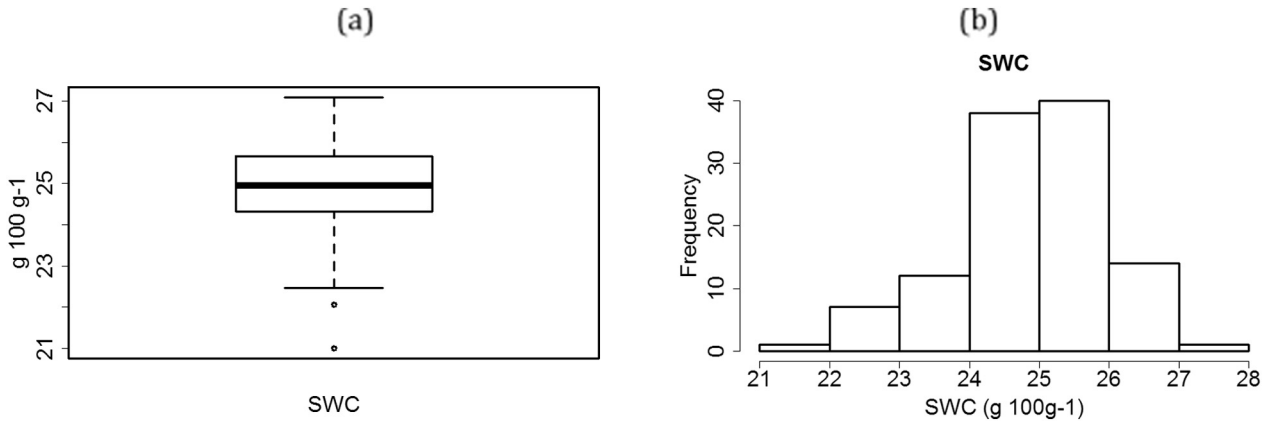

Fig. 2. Box plot (a) and histogram (b) of the soil water content (SWC, g 100 g^{-1}).

Table 2
Testing the normality for soil water content (SWC).

Variable	Test	*p*-Value
SWC	Shapiro-Wilk	0.022
SWC	Anderson-Darling	0.055 (*)

(*) confirmed H_0: hypothesis of normality.

$(\boldsymbol{u}, \varepsilon) = 0$.

A standard method for LMM parameters estimation is maximum likelihood (ML). However, this method leads to the underestimation of the variance components (Davidian and Giltinan, 1995; Ripley, 1988) when the correlation structure is not known a priori. Therefore, the simultaneous estimates of correlation parameters and fixed effect coefficients were obtained by restricted (or residual) maximum

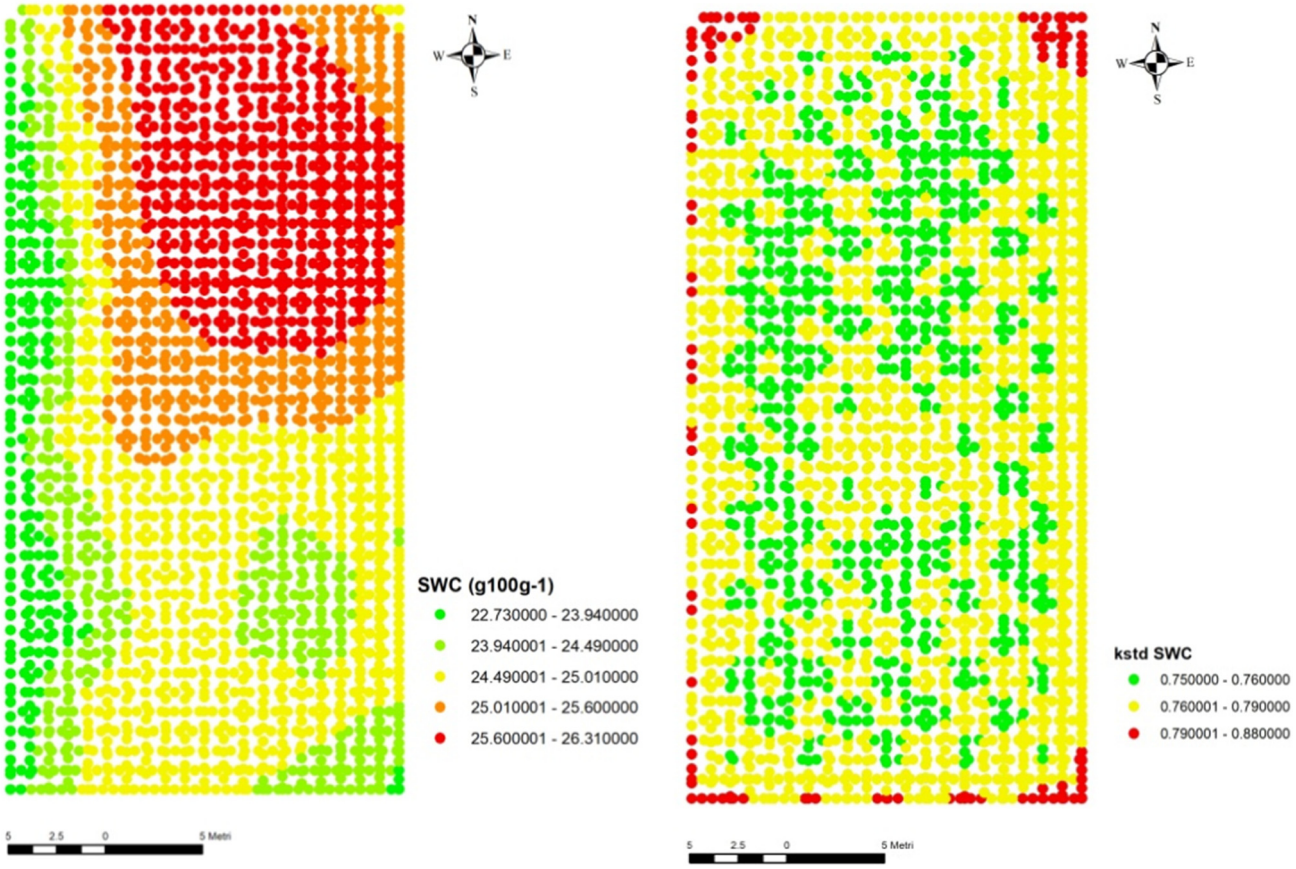

Fig. 3. Spatial estimates of SWC (a) and estimation kriging standard deviation –kstd- (b) obtained with ordinary kriging (OK).

Table 3
Eigenvalues and variance explained by each component from PCA carried out on GPR variables: (a) 250 MHz (10 variables); (b) 600 MHz (10 variables); (c) 1600 MHz (6 variables). n = 2559 cases.

(a) Eigenvalues of the correlation matrix: Total = 10 average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.76117827	0.94502390	0.4761	0.4761
2	3.81615437	3.01284560	0.3816	0.8577
3	0.80330877	0.52784742	0.0803	0.9381
4	0.27546135	0.05946137	0.0275	0.9656
5*	0.21599998	0.11440027	0.0216	0.9872

(b) Eigenvalues of the correlation matrix: Total = 10 average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.29425950	1.78518995	0.4294	0.4294
2	2.50906954	0.77241195	0.2509	0.6803
3	1.73665759	0.75740579	0.1737	0.8540
4	0.97925180	0.64058795	0.0979	0.9519
5	0.33866386	0.23838541	0.0339	0.9858

(c) Eigenvalues of the correlation matrix: Total = 6 average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.36349236	3.55755836	0.7272	0.7272
2	0.80593400	0.40423097	0.1343	0.8616
3	0.40170303	0.14369176	0.0670	0.9285
4	0.25801126	0.10040256	0.0430	0.9715
5	0.15760871	0.14435806	0.0263	0.9978

*Only the first 5 PCs are reported in the table.

likelihood estimation (REML) (Pollice and Bilancia, 2002), to consider the loss of degrees of freedom due to the estimation of the fixed effects coefficients (Cressie, 1993).

In this study, different LMMs were estimated and compared. In particular, the distributional assumptions were assessed by testing the Gaussianity by Shapiro-Wilk and the independence of the residuals by Runs test (Wu and Zhao, 2013) and Moran's I statistic. The LMM models were estimated using the lme functions of the R library {nlme} (Pinheiro et al., 2012).

2.5.3. Kriging with external drift

Kriging with external drift (KED) is a particular formulation of kriging under non stationary conditions which allows the use of secondary information to account for the spatial variation of the primary variable's local mean. The secondary variables are chosen for their significant correlation with the variable of interest and should be available at every location of the primary variable and every estimation grid point (Chilès and Delfiner, 1999; Buttafuoco and Castrignanò, 2005; Cafarelli and Castrignanò, 2011). KED is applied here on intrinsic random functions of order k (IRF-k). Such theory was introduced by Matheron (1973) to avoid the difficulties resulting from the simultaneous estimation and modelling of the drift and variogram.

In KED, to account for the multi-scale behavior of a spatial attribute Z, its value at the location $\boldsymbol{x}$ (the coordinates vector) is decomposed as follows

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x})$$

where $m(\boldsymbol{x})$ is a deterministic function known as the drift and $Y(\boldsymbol{x})$ is the spatially correlated random component with zero mean (stochastic part).

A first assumption is that $m(\boldsymbol{x}) = \sum_{l=0}^{k} a_l f^l(x_\alpha)$ where a_l are unknown coefficients, $f^l(x)$ are known functions of the coordinates ($\boldsymbol{x}$ = (x, y)), in the practice $f^l(x) = x^l$ are monomials of the coordinates and \or of the covariates. In the general case, $m(\boldsymbol{x})$ can be decomposed into two sums: i) a linear combination of coordinates' monomials; ii) a linear combination of covariates' monomials. The first is called “internal drift”, the second “external drift”.

In case of irf-k, data are transformed by means of linear combinations

$$Z(\lambda) = \sum_{\alpha=1}^{N} \lambda_\alpha Z(\boldsymbol{x}_\alpha)$$

Subjected to the following constraint

$\sum_{\alpha=1}^{N} \lambda_\alpha f^l(x_\alpha) = f^l(x_0)$
for $l = 0, \ldots, k$; with x_0 as unknown estimation point.

Defined $Z(\lambda, x) = \Sigma \lambda_\alpha Z(x_\alpha + x)$, if $Z(\lambda, x)$ is a stationary random function with respect to x then Z(x) is said to be an intrinsic random function of order k.

The correlation structure associated with the random part is expressed by a generalized covariance (GC) function of the distance $\mathbf{h}$ between two observations x_α and x_β, $K(x_\alpha - x_\beta)$, denoted by $K(\mathbf{h})$. The second assumption is that the correlation structure $K(\mathbf{h})$ can be expressed as a linear combination of a given set of known basic structures under some conditions on the coefficients. The generalized covariance is the polynomial GC model under constraints on the coefficients b:

$$K(|\mathbf{h}|) = C_0\delta(|\mathbf{h}|) - b_0\ |\mathbf{h}| + b_1\ |\mathbf{h}|^2 \log |\mathbf{h}| + b_2\ |\mathbf{h}|^3$$

where $\delta(|\mathbf{h}|) = 0$ for $|\mathbf{h}| > 0$ else $\delta(|\mathbf{h}|) = 1$. The coefficients C_0, b_0, b_1 and b_2, in a two dimensional space R^2, must satisfy the following conditions (Chilès and Delfiner, 1999):

$$C_0 \geq 0, \quad b_0 \geq 0, \quad b_1 \geq 0, \quad b_2 \geq -\frac{3}{2}\sqrt{b_0 b_1}.$$

In intrinsic random function kriging, the structural analysis is

Table 4
Variable loadings of the selected components from PCA carried out on GPR variables; (a) 250 MHz (10 variables); (b) 600 MHz (10 variables); (c) 1600 MHz (6 variables). n = 2559 cases. Values are multiplied by 100 and rounded to the nearest integer.

(a)			(b)				(c)	
250 MHz	Factor1	Factor2	600 MHz	Factor1	Factor2	Factor3	1600 MHz	Factor1
Amp 0.06 m	89*	25	Amp 0.03 m	50*	56*	−53*	Amp 0.015 m	94*
Amp 0.12 m	92*	31	Amp 0.06 m	77*	30	−54*	Amp 0.045 m	97*
Amp 0.18 m	91*	40*	Amp 0.09 m	89*	−12	−39*	Amp 0.075 m	79*
Amp 0.24 m	80*	55*	Amp 0.12 m	83*	−47*	−16	Amp 0.105 m	89*
Amp 0.30 m	44*	81*	Amp 0.15 m	75*	−64*	6	Amp 0.135 m	88*
Amp 0.60 m	−27	83*	Amp 0.18 m	64*	−60*	39*	Amp 0.165 m	59*
Amp 0.70 m	−75*	55*	Amp 0.21 m	58*	3	68*		
Amp 0.80 m	−67*	60*	Amp 0.24 m	51*	64*	43*		
Amp 0.90 m	−34	80*	Amp 0.27 m	53*	75*	25		
Amp 1.00 m	−49*	72*	Amp 0.30 m	34	35	35		

* Indicates the signifcance of the variable loadings.

Table 5
Eigenvalues and variance explained by the single components from PCA carried out on EMI variables. n = 2559 cases.

Eigenvalues of the correlation matrix: Total = 2 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.80175060	1.60350119	0.9009	0.9009
2	0.19824940		0.0991	1.0000

Table 6
Variable loadings of the selected component from PCA carried out on EMI variables. n = 2559 cases. Values are multiplied by 100 and rounded to the nearest integer.

	Factor1	Factor2
EC_a-H	95*	31
EC_a-V	95*	−31

* Indicates the signifcance of the variable loadings.

performed in two stages: first the order, *k*, of the drift is established and, secondly, the generalized covariance, K(**h**), is estimated by fitting a parametric model. To determine the degree of drift and the best combination of auxiliary variables, the least-squares errors are ranked in ascending magnitude for each target point and for the various options; the smallest rank, averaged over the different target points, corresponds to the optimal model for the drift. Cross-validation (leave-one-out) was used to select the optimal generalized covariance (Carroll and Cressie, 1996); the model that led to a standardized error closest to 1, was finally selected.

2.5.4. Models comparison

Evaluation of LMM models was performed comparing respective information criteria indices, residual analysis and correlation between observed and predicted SWC values using Pearson correlation coefficient (r).

Comparison among the KED models with the different extracted factors was done using the cross-validation statistics (Carroll and Cressie, 1996). Let Y[i] be the observed response value removed at the i-th iteration; let [i] be its corresponding prediction obtained by fitting the model to the remaining n-1 points; let e[i] = Y[i] − [i] be the difference between the observed and estimated values and let σ[i] be the mean squared prediction error of [i]; the three cross-validation statistics are:

$$CV_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{e_{[i]}}{\sigma_{[i]}};$$

$$CV_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{e_{[i]}^2}{\sigma_{[i]}^2}\right)^{\frac{1}{2}};$$

$$CV_3 = \left(\frac{1}{n}\sum_{i=1}^{n}e_{[i]}^2\right)^{\frac{1}{2}}$$

CV_1 (mean standardized error) was used to assess the unbiasedness of prediction and the optimal value of CV_1 should be approximately zero; CV_2 (root mean square standardized error) was used to assess the accuracy of mean squared prediction error and should be approximately 1; CV_3 (root mean square error) was used to check the goodness of prediction, and models with smaller values should be preferred, because this means that fitted values are close to observed values (Carroll and Cressie, 1996). Moreover, to assess the degree of association between observations and estimates, the Pearson coefficients were calculated for these methods. Comparison among the different modelling approaches, ordinary kriging (OK), LMM and KED, was done using the root mean square error (CV_3).

2.5.5. Cross-correlogram

Given two different prediction maps of the same variable, M_A and M_B, provided by the different estimation methods, let's say A and B, there are several methods to compare them (Stein et al., 1997; Barca and Passarella, 2008). To account for the inherently spatial characteristics of map representation, the cross-correlogram, which measures the correlation as a function of the distance between observations, is particularly well-suited. The analytical formulation of the cross-correlogram is the following:

$$\rho_{A,B}(h) = \frac{E\left[z_{i,jA}, z_{i',j'B}\right] - m_A \bullet m_B}{s_A \bullet s_B}$$

where $z_{i,\,jA}$ and $z_{i',\,j'B}$ represent the values at locations $(i,\ j)$ and (i',j') of the two maps separated by the h distance, respectively; $h = \sqrt{(i-i') + (j-j')}$ represents the distance between the two locations, E denotes the mathematical expectation, m_A and m_B represent the populations means and s_A and s_B represent the populations standard deviations. If patterns are completely similar, apart from a constant, $\rho_{A,\,B}(0)$ should be equal to 1. To estimate $\rho_{A,\,B}(h)$ from the available data, the following equation can be used:

$$r_{A,B}(h) = \frac{\sum_{i,j=1}^{N(h)} z_{i,jA} \bullet z_{i',j'B} - \hat{m}_A \bullet \hat{m}_B}{\hat{s}_A \bullet \hat{s}_B}$$

To compute $r_{A,\,B}(h)$, the procedure is as follows: from both the maps are collected all the couples whose locations are separated by the distance h. Indices $\hat{m}_A$ and $\hat{m}_B$ and $\hat{s}_A$ and $\hat{s}_B$ represent the mean and the standard deviation of mapped $z_{i,\,jA}$ and $z_{i',\,j'B}$, respectively. $N(h)$ is the total number of these pairs with respect to the h value. If the methods

Table 7
REML estimates of fixed effects, standard errors and significance (t-test) and residual analysis for the models including: (a) EMI factor; (b) GPR factors; (c) both EMI and GPR factors. S-W indicates the Shapiro-Wilk test.

(a)						Residual analysis			
Effect	Estimate	St error	t value	Pr > |t|	AIC	Mean	St dev	S-W	Moran I (p > Z)
Intercept	25.0140	0.09216	271.41	< 0.0001	316.5	0	0.9507	0.2816	< 0.0001
Factor1_EMI	0.4997	0.07985	6.26	< 0.0001					

(b)						Residual analysis			
Effect	Estimate	St error	t value	Pr > |t|	AIC	Mean	St dev	S-W	Moran I (p > Z)
Intercept	24.8952	0.08994	276.79	< 0.0001	302.1	0	0.8337	0.4409	0.5175
Factor1_250 MHz	−0.3891	0.09968	−3.90	0.0002					
Factor2_250 MHz	0.3708	0.08747	4.24	< 0.0001					
Factor1_600 MHz	−0.3720	0.09288	−4.01	0.0001					
Factor2_600 MHz	0.03378	0.08693	0.39	0.6983					
Factor3_600 MHz	0.1279	0.09594	1.33	0.1853					
Factor1_1600 MHz	0.3055	0.09071	3.37	0.0011					

(c)						Residual analysis			
Effect	Estimate	St error	t value	Pr > |t|	AIC	Mean	St dev	S-W	Moran I (p > Z)
Intercept	24.9769	0.08838	282.60	< 0.0001	293.3	0	0.787	0.1272	0.4330
Factor1_250 MHz	−0.2998	0.09787	−3.06	0.0028					
Factor2_250 MHz	0.3120	0.08463	3.69	0.0004					
Factor1_600 MHz	−0.2891	0.09117	−3.17	0.0020					
Factor2_600 MHz	0.01799	0.08262	0.22	0.8280					
Factor3_600 MHz	0.1107	0.09118	1.21	0.2276					
Factor1_1600 MHz	0.2687	0.08671	3.10	0.0025					
Factor1_EMI	0.2726	0.07653	3.56	0.0006					

being compared produced similar results, a decreasing cross-correlogram for increasing values of h is expected.

3. Results

3.1. Descriptive analysis

A preliminary statistical analysis on the response variable, SWC, highlighted the presence of three outliers, which were excluded from the subsequent data analysis.

Descriptive statistics for SWC are reported in Table 1 and in Fig. 2. SWC varied between 21 and 27.09 g 100 g^{-1}, with an average value of 24.89 g 100 g^{-1}.

Fig. 2 showed that the middle part of the distribution (the box) was symmetrical with respect to the I and III quartiles, the left tail (the whisker below) was little more pronounced. The left asymmetry was confirmed by the sign of the skewness coefficient (Table 1). The histogram representation showed that the SWC distribution did not strongly deviate from the normal distribution (Fig. 2).

As confirmed also by the Anderson-Darling statistics ($p = 0.055$, Table 2), SWC data can be considered approximately normal, thus they were not subjected to a normal transform.

3.2. Estimation of SWC

An isotropic variogram model was fitted to the experimental variogram of SWC including the following basic structures: a nugget effect and a spherical model. The fitted variogram had a range of 29.7 m, a nugget of 0.4287 and a partial sill of 1.1611. The nugget-to-sill ratio was of approximately 27%, indicating a strong spatial dependence of the target variable observations (Cambardella et al., 1994). The results of cross-validation test (mean error, ME, and the mean square standardized error, MSSE) were satisfactory because mean error was quite close to 0 (ME = 0.02) and the mean square standardized error was almost 1 (MSSE = 1.077).

The spatial distribution of SWC estimated by OK (Fig. 3a) showed highest values in the north central portion of the study area that extended to north-eastwards. The western border and the southern area were characterized by lowest values.

Since there were not topographic variations in the study area and the relationship between SWC and textural properties (in particular clay content) was unexpectedly negative, as observed in a previous study, the spatial patterns of SWC could have been influenced by the variations in soil depth and soil/bedrock interface (De Benedetto et al., 2013). Using a combination of different sensors and integrating their data could provide more accurate estimation of soil water content. Fig. 3b shows the estimation kriging standard deviation (kstd) map which indicates an edge effect due to a lesser number of observations involved in the prediction. This is coherent with the fact that kstd is a precision rather than an accuracy index.

3.3. Results of PCA

Principal Component Analysis, carried out on GPR data, showed that for the 250 MHz frequency, the first two factors had an eigenvalue higher than 1 and were able to cumulatively synthesize a percentage of total variance > 85% (Table 3a). On the first factor, the amplitudes in the range 0.06 m–0.24 m and 0.7 m–0.8 m showed the highest positive and negative loadings, respectively (Table 4,a); on the second factor the amplitudes in the range 0.3 m–0.6 m and 0.9 m–1 m weighted more and positively (Table 4a). The maps at the different depths, estimated in a previous study (De Benedetto et al., 2013), looked quite similar with the northern area characterized by less attenuation, whereas the PCA results showed a discontinuity in the radar signal (change in the sign for the loadings) between 0.3 m and 0.6 m depths. This behaviour might be related to an interface in the soil, probably due to a shallow ploughing, consistent with a pre-existing pedological profile (De Benedetto et al., 2012).

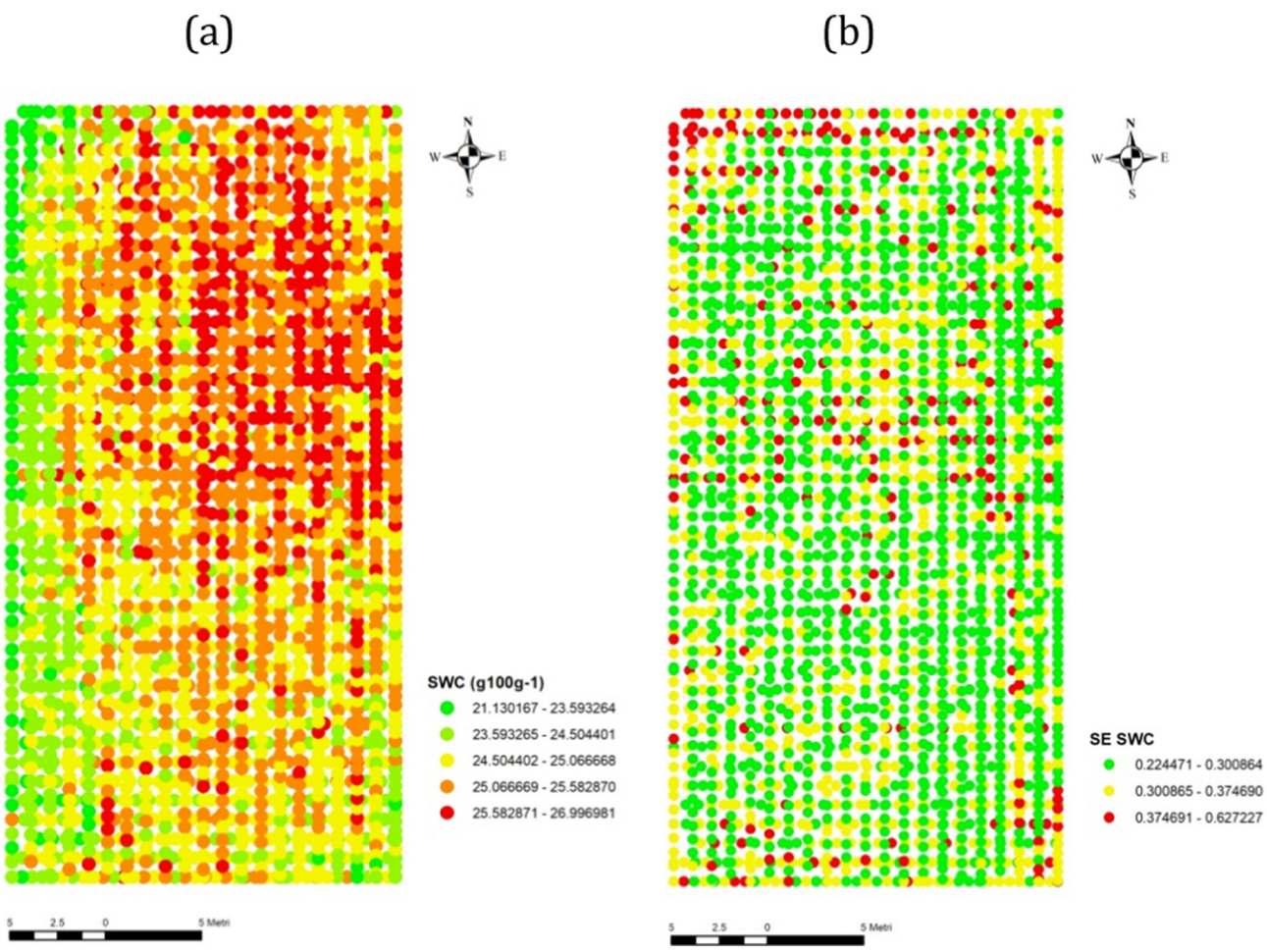

Fig. 4. Estimates of SWC obtained with linear mixed effects model (LMM) with the model including both EMI and GPR factors (a) and standard error (SE) (b).

Table 8
Selected generalized covariance functions for SWC, and corresponding cross validation statistics, for each variable group using KED.

	Basic structures	CV_1	CV_2	CV_3	r
(a) EMI factor	Nugget effect Functions 1st order $R = 4.6$ m	0.009	1.11	0.79	0.689
(b) GPR factors	Nugget effect	−0.017	0.94	0.80	0.682
(c) All factors	Nugget effect	0.011	0.94	0.76	0.72

With regard to the 600 MHz frequency, the first three factors showed an eigenvalue > 1 and explained 85.4% of the total variance (Table 3b). The first factor synthetized the role of the amplitudes in the range 0.06 m–0.18 m, while the second factor highlighted an inverse relationship between the amplitudes in the ranges 0.12 m–0.18 m and 0.24 m–0.27 m, indicating a discontinuity as observed for 250 MHz frequency. On the third factor the highest weights were shown by the amplitude at 0.03 m and 0.21 m depths.

In the PCA carried out on the 1600 MHz frequency, only the first factor had an eigenvalue > 1 and explained 72.7% of the total variance (Table 3c). All the variables had significant and high loadings with particular regard to those in the interval 0.015 m–0.135 m. The 1600 MHz maps revealed a high level of spatial continuity along the soil profile at least to 0.15 m depth, because the maps were quite similar (De Benedetto et al., 2013).

Finally, in the PCA carried out on the EMI variables, the first factor explained about 90% of the total variance and was able to synthesize the greatest part of the information provided by the horizontal and the vertical EC_a (Tables 5 and 6).

The Pearson's correlation coefficients, highlighting the relationship between SWC and the selected factors, are reported in the following lines: −0.325 (Factor1_250 MHz), 0.484 (Factor2_250 MHz), −0.282 (Factor1_600 MHz), 0.215 (Factor2_600 MHz), 0.249 (Factor3_600 MHz), 0.157 (Factor1_1600 MHz), 0.510 (Factor1_Eca).

3.4. Linear mixed effects models

To understand the contribution of the information provided by the different geophysical sensors, the following LMM models were fitted to the experimental data of SWC and to the factors extracted through PCA: (a) a model including only the EMI factor; (b) a model including only the GPR factors; (c) a model including both EMI and GPR factors. For each of the above mentioned groups, both no-spatial (null models) and spatial models were built.

The model including only EMI was significant and showed an Akaike information criterion (AIC) of 316.5 (Table 7); Pearson correlation coefficient between predicted and observed data was 0.5107 ($p < 0.0001$).

In the model including the GPR factors, four out of the six factors extracted showed a significant effect in estimating SWC: the first two factors at 250 MHz, the first factor at 600 MHz and the factor extracted at 1600 MHz (Table 7). The model had a lower AIC value (302.1) and the correlation between predicted and observed data was of 0.657. These results seem to indicate a greater contribution of the information provided by GPR in estimating SWC in comparison with that deriving only by EMI sensor.

In the model including the factors extracted from both EMI and GPR data, the same four GPR factors (Factor 1 and Factor 2 at 250 MHz;

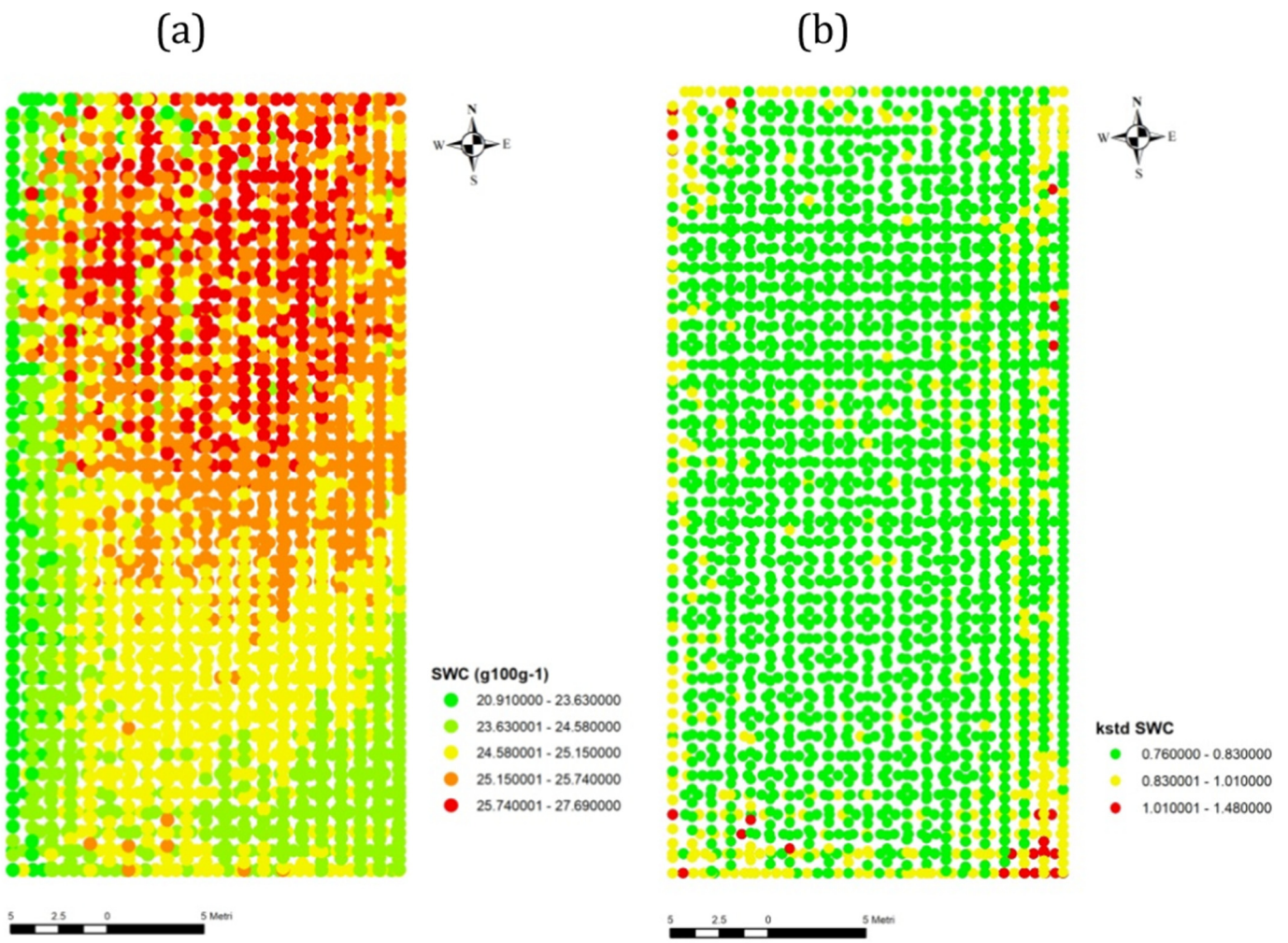

Fig. 5. Spatial estimates of SWC obtained with kriging with external drift (KED) with the model including all geophysical factors (a) and estimation kriging standard deviation (kstd) (b).

Table 9
Root mean square error (CV_3) of the estimation of SWC using OK, LMM and KED.

Estimation method	CV_3
OK	0.83
LMM	0.84
KED	0.76

Factor 1 at 600 MHz; Factor 1 at 1600 MHz) and EMI factor were significant (Table 7). Factor 2 at 250 MHz and EMI factor showed the greatest contribution to the model ($p = 0.0004$ and $p = 0.0006$). This model had the lowest AIC value (293.3) and the highest correlation between predicted and observed data ($r = 0.702$) indicating that combination of information deriving from different sources was able to improve the prediction of the SWC.

Results of residual analyses confirmed the goodness of fitting, with a mean close to 0, a standard deviation close to 1 (0.9507, 0.8337 and 0.787 for the three models, respectively), and a distribution not significantly different from the normal as confirmed by the Shapiro-Wilk test results (Table 7). Runs test suggested the global independence of model residuals and the Moran's I test for the GPR and the full models indicated a not significant overall spatial autocorrelation (Table 7). These results were definitely confirmed after fitting the corresponding spatial LMM models for each group of covariates as the spatial component of the models resulted non-significant.

The model including both EMI and GPR factors resulted to be the best among the compared LMM models; consequently, it was used to estimate the SWC at covariates locations (2559 points). The map reported in Fig. 4 looked much noisier than that obtained with OK, where the higher level of smoothness of the latter was due to OK characteristics. In general, the maps revealed a wide northern area of highest values, though the LMM map looked more locally variable in the southern area. Variation in absolute values was registered in comparison with the OK estimations and, in particular, larger values of SWC values were observed in the southern part of the field. Finally, the map of the standard error (Fig. 4b) was more confused with respect to the corresponding OK map. This is due to the influence of the random component of the LMM formulation. Also in this case, the edge effect is evident.

3.5. Kriging with external drift

Kriging with external drift (KED) was applied, like LMM, to SWC and the following groups of covariates: a) EMI factor, b) all GPR factors and finally c) both EMI and GPR factors.

For drift identification, several models were compared obtained including the intercept and an increasing number of covariates; the models with the smallest mean error rank and with the mean squared error equal to one were chosen within each group. The estimated generalized covariance for the model selected in each group was reported in Table 8 with the three cross-validation statistics and the

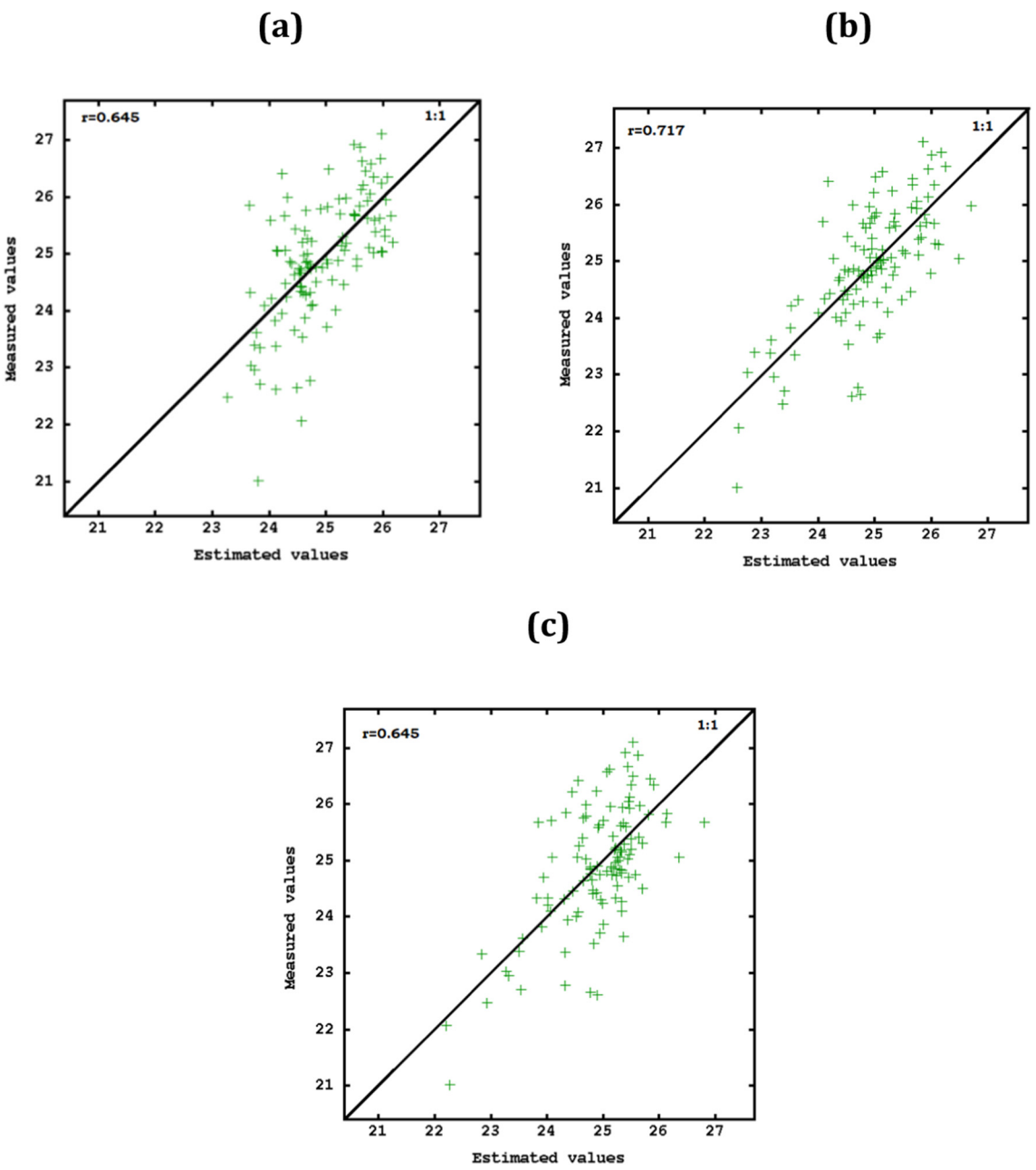

Fig. 6. Scatter diagram of the measured data of SWC versus the estimated values of SWC using OK (a) KED (b) and LMM (c).

correlation coefficient between predicted and observed data.

For the drift determination, the included variables were: for the first group, the intercept and EMI factor; for the second one the intercept, the Factor 1 for 1600 MHz frequency, the Factor 1 and 2 for 250 MHz and Factor 1 for 600 MHz frequency variables; and finally for the third group the intercept, Factor 1 and 2 for 250 MHz frequency, Factor 1 for EMI and Factor 1 for 600 MHz frequency variables.

These results showed the statistical significance of the geophysical variables in soil water content prediction. In particular, in the model including only GPR factors, the results showed that the factors related to GPR at 250 MHz frequency provided information along the soil profile up to 1 m depth, while the factors related to GPR at 600 MHz and 1600 MHz frequencies provided information on the shallower layer (up to 0.18 m).

In the model including all geophysical factors, differently from what observed in the previous LMM modelling and in the model with only GPR variables, the selected factors were related to the shallower layers. The estimated generalized covariance, for both GPR and all factors models, consisted of Nugget Effect, in agreement with LMM results. Therefore the stochastic variation, described by the generalized covariance function, was not spatially structured and the external drift filtered out all the structured component of spatial variation of SWC.

The estimated generalized covariance, for the model including EMI factor, consisted of Nugget Effect and one function of the first order with range of 4.6 m. In this case, the stochastic variation was spatially structured and the external drift was not sufficient to filter out all the structured component of spatial variation of SWC.

The model including all geophysical factors had the highest correlation between predicted and observed data ($r = 0.72$) and the smaller value of CV_3 (Table 8). Therefore, this model was used to estimate SWC with KED (Fig. 5).

Comparing the estimates of SWC obtained with KED (Fig. 5) with those obtained with OK (Fig. 3) and LMM (Fig. 4), the maps seem to reproduce the same main structures of spatial dependence. The high

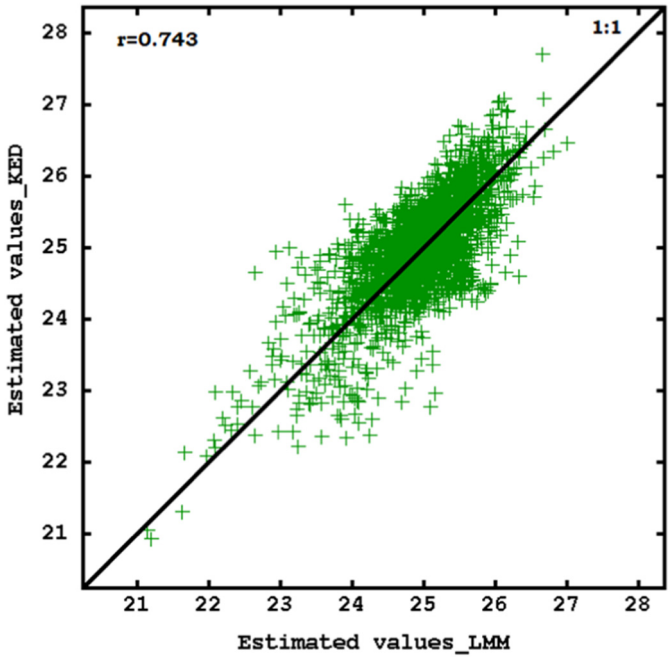

Fig. 7. Scatter diagram of the estimated values of SWC using LMM versus the estimated values of SWC using KED.

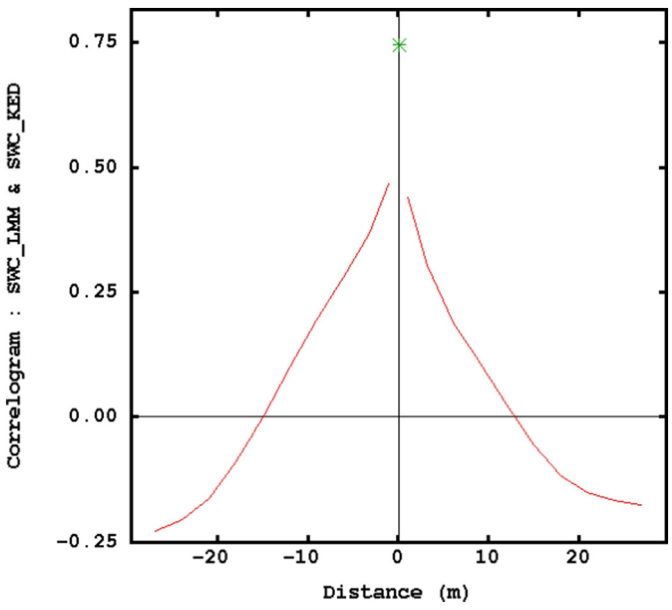

Fig. 8. Cross-correlogram between SWC estimated with LMM and KED.

level of smoothness observed in kriging estimates was due to the coarser sampling scale compared with the geophysical surveys. The increased variability, observed in the KED and LMM estimates, was due to the sub-meter scale information in the geophysical data used to downscale the SWC variation. However, the estimates obtained with the KED reproduced more accurately the structures observed in OK estimates. Variation in absolute values was registered in comparison with the OK and LMM estimations and, in particular, lower SWC values were observed in the southern part of the field and larger in the northern area.

Map of the estimation kriging standard error (kstd) for the KED (Fig. 5b) confirms the edge effect observed for OK (Fig. 3b), although less evident due to the larger standard deviation range (max-min). Differently from LMM (Fig. 4b), the central part of the map shows a greater uniformity.

3.6. Comparison among the estimates

The results of cross-validation test (CV_3), used to check the goodness of fit of the different models and thus their accuracy, showed that KED outperformed LMM and OK because lower root mean squared error was observed (Table 9).

For a direct comparison, the estimates garnered with the different techniques were plotted with respect to the measured SWC values (Fig. 6). KED provided a better fit showing a quite good agreement between the observations and estimations as confirmed by the larger correlation coefficient (r = 0.72) and points more gathered around the 1:1 line, in comparison with OK and LMM (r = 0.65).

Although the correlation coefficients for OK and LMM were equal, it was noticed an overestimation of OK, due to the well-known smoothing effect of this predictor.

Comparing the mapped estimates of SWC using LMM and KED (Fig. 7), it must be stressed that it is not possible a direct visual comparison due to the different scales of the methods predictions as a consequence of the smoothing in geostatistics approaches; notwithstanding, a quantitative comparison can be performed (Stein et al., 1997).

From a visual inspection of the scatterplot of the two estimates of SWC, it can be observed that the values were evenly distributed around the 1:1 line, showing a quite good point-wise agreement between the estimations and consequently a great similarity between the two methodologies (Lark, 2012).

From the spatial standpoint, the cross-correlogram of KED and LMM SWC estimates (Fig. 8) made more objective the visual comparison between maps. In particular, the correlogram computed at lag (h) zero was 0.743 and corresponded to the green point at the apex of the curve in Fig. 8. Finally, the figure displayed an approximately symmetrical decreasing form, confirming that the two maps had a similar spatial pattern.

4. Conclusions

The present paper investigates two different aspects of topsoil SWC prediction and mapping. The first one is a methodological comparison among LMM, OK and KED. The second tries to rank the contribution of different proximal geophysical information in SWC prediction.

Concerning the first issue, LMM and OK showed very close outcomes, whereas KED, particularly for the EMI data case, slightly outperformed the other approaches. The similarity between LMM and OK could appear unexpected, in principle. In fact, LMM model includes both covariates information and spatial structure, therefore, with respect the OK, which models only the stochastic information, it should be outperforming. However, for the presented case, LMM was unable to find a significant residual spatial structure, except for EMI data. Therefore, it seems that the deterministic modelling was able to filter out the spatial information described in OK and that the information brought by the investigated covariates was almost equivalent to the spatial auto-correlated structure.

Differently from LMM, KED was able to model a residual spatial structure for EMI dataset and, for the other cases, to better predict SWC. In conclusion, under the experimental conditions implemented, KED demonstrated to be better suited and more efficient for modelling SWC data and geophysical covariates.

Concerning the second issue, observing the correlation values between observed/predicted of the three LMM groups, it can be indirectly derived the strength of the contribution of the information provided by the different sensors. Covariate information, particularly when GPR data and GPR + EMI data were used, filtered out entirely the spatial component. GPR evidently outperformed EMI, however the combined use of both data showed an even larger explanatory capability. At a first

glance, this difference is less perceivable for KED models, because by comparing the correlation coefficients of EMI and GPR models, they appear to be almost equal. However, EMI model contains also a spatial component (first order function), therefore this seems to indicate that the lower information provided by EC_a covariate is integrated by such function. After the above consideration, it can be concluded that both the approaches confirm the larger informative contribution of GPR with respect to EMI. From a physical standpoint, as previously highlighted, the above result can be explained by the different nature of EMI and GPR outcomes. The former are integrated values over all soil layers, while the latter gives information with different spatial resolution using different frequency antenna. Consequently, GPR information results more sensitive to near-surface effect than EMI. A further result is that the combination of the two sources of proximal data provides more accurate predictions of topsoil water content, proving that information brought by EMI is not completely contained in the GPR data. The improvement is not quantitatively very large but significant, anyway. Testing that the additional variability of SWC prediction corresponds to real features of moisture distribution in the soil is a challenge for the future research. However, as further future development, it can be checked if introducing non-linear covariates of proximal information can improve the topsoil water content prediction.

Acknowledgements

The authors would like to thank the European Commission (EU) and the Italian Ministry for Education, University and Research (MIUR) for funding the present research and the methodological contribution, in the frame of the collaborative international consortium DESERT (ID-217) financed under the ERA-NET Cofund WaterWorks2014 Call. This ERA-NET is an integral part of the 2015 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI).

References

Adamchuk, V.I., Hummel, J.W., Morgan, M.T., Upadhyaya, S.K., 2004. On-the-go soil sensors for precision agriculture. Comput. Electron. Agric. 44 (1), 71–91.
Barca, E., Passarella, G., 2008. Spatial evaluation of the risk of groundwater quality degradation. A comparison between disjunctive kriging and geostatistical simulation. Environ. Monit. Assess. 137 (1–3), 261–273.
Berardi, M., Vurro, M., 2016. The numerical solution of Richards' equation by means of method of lines and ensemble Kalman filter. Math. Comput. Simul. 125, 38–47.
Berardi, M., Andrisani, A., Lopez, L., Vurro, M., 2016. A new data assimilation technique based on ensemble Kalman filter and Brownian bridges: an application to Richards' equation. Comput. Phys. Commun. 208, 43–53.
Bourennane, H., King, D., Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97, 255–271.
Buttafuoco, G., Castrignanò, A., 2005. Study of the spatio-temporal variation of soil moisture under forest using intrinsic random functions of order k. Geoderma 128, 208–220.
Cafarelli, B., Castrignanò, A., 2011. The use of geoadditive models to estimate the spatial distribution of grain weight in an agronomic field: a comparison with Kriging with external drift. Environmetrics 22, 769–780.
Cafarelli, B., Castrignanò, A., De Benedetto, D., Palumbo, A. D., Buttafuoco, G., 2015. A linear mixed effect (LME) model for soil water content estimation based on geophysical sensing: a comparison of an LME model and kriging with external drift. Environ. Geol. 73(5), 1951–1960.
Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., Konopka, A.E., 1994. Field-scale variability of soil properties in central Iova soils. Soil Sci. Soc. Am. J. 58, 1501–1511.
Carroll, S.S., Cressie, N.A., 1996. Comparison of geostatistical methodologies used to estimate snow water equivalent. Water Resour. Bull. 32 (2), 267–278.
Chilès, J.P., Delfiner, P., 1999. Geostatististics: Modelling Spatial Uncertainty. Wiley, New York, NY.
Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley Series in Probability and Statistics.
Daniels, D.J., Gunton, D.J., Scott, H.F., 1988. Introduction to subsurface radar. IEE Proc. 135, 278–316.
Davidian, M., Giltinan, D.M., 1995. Nonlinear Models for Repeated Measurement Data. Chapman and Hall, New York.
Davis, J.L., Annan, A.P., 1989. Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy. Geophys. Prospect. 37, 531–551.
De Benedetto, D., Castrignanò, A., Sollitto, D., Modugno, F., Buttafuoco, G., Lo Papa, G., 2012. Integrating geophysical and geostatistical techniques to map the spatial variation of clay. Geoderma 171–172, 53–63.
De Benedetto, D., Castrignanò, A., Quarto, R., 2013. A geostatistical approach to estimate soil moisture as a function of geophysical data and soil attributes. Procedia Environ. Sci. 19, 436–445.
De Benedetto, D., Quarto, R., Castrignanò, A., Palumbo, D.A., 2015. Impact of data processing and antenna frequency on spatial structure modelling of GPR data. Sensors 15, 16430–16447.
Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. J. Hydrol. 228, 113–129.
Grote, K., Hubbard, S., Rubin, Y., 2003. Field-scale estimation of volumetric water content using ground-penetrating radar ground wave techniques. Water Resour. Res. 39, 1321–1335.
Grote, K., Anger, C., Kelly, B., Hubbard, S., Rubin, Y., 2010. Characterization of soil water content variability and soil texture using GPR groundwave techniques. J. Environ. Eng. Geophys. 15 (3), 93–110.
Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust. J. Soil Res. 41, 1403–1422.
Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93.
Hira, Z.M., Gillies, D.F., 2015. A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinforma. 2015, 198363. https://doi.org/10.1155/2015/198363.
Huang, J., Scudiero, E., Choo, H., Corwin, D.L., Triantafilis, J., 2016. Mapping soil moisture across an irrigated field using electromagnetic conductivity imaging. Agric. Water Manag. 163, 285–294.
Huang, J., Scudiero, E., Clary, W., Corwin, D.L., Triantafilis, J., 2017a. Time-lapse monitoring of soil water content using electromagnetic conductivity imaging. Soil Use Manag. 33, 191–204.
Huang, J., McBratney, A.B., Minasny, B., Triantafilis, J., 2017b. Monitoring soil water dynamics using electromagnetic conductivity imaging and the ensemble Kalman filter. Geoderma 285, 76–93.
Hubbard, S., Grote, K., Rubin, Y., 2002. Mapping the volumetric soil water content of a California vineyard using high-frequency GPR ground wave data. Lead. Edge 25, 552–559.
Huisman, J.A., Hubbard, S.S., Redman, J.D., Annan, A.P., 2003. Measuring soil water content with ground penetrating radar: a review. Vadose Zone J. 2, 476–491.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University press, New York.
Knight, R., Tercier, P., Jol, H., 1997. The role of ground-penetrating radar and geostatistics in reservoir description. Lead. Edge 16 (11), 1576–1582.
Kong, Q., Chen, H., Mo, Y.L., Song, G., 2017. Real-time monitoring of water content in sandy soil using shear mode piezoceramic transducers and active sensing—a feasibility study. Sensors 17 (10), 2395.
Lark, R.M., 2012. Towards soil geostatistics. Spatial Statistics 1, 92–99.
Lark, R.M., Cullis, B.R., Welham, S., 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. https://doi.org/10.1111/j.1365-2389.2005.00768.x.
Lunt, I.A., Hubbard, S.S., Rubin, U., 2005. Soil moisture estimation using ground-penetrating radar reflection data. J. Hydrol. 307, 254–269.
Martínez, G., Vanderlinden, K., Giráldez, J.V., Espejo, A.J., Muriel, J.L., 2010. Fieldscale soil moisture pattern mapping using electromagnetic induction. Vadose Zone J. 9, 871. https://doi.org/10.2136/vzj2009.0160.
Martínez, G., Vanderlinden, K., Pachepsky, Y., Giráldez Cervera, J.V., EspejoPérez, A.J., 2012. Estimating topsoil water content of clay soils with data from time-lapse electrical conductivity surveys. Soil Sci. 177, 369–376. https://doi.org/10.1097/SS.0b013e31824eda57.
Matheron, G., 1973. The intrinsic random functions and their applications. Adv. Appl. Probab. 5, 239–465.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327.
McNeill, J.D., 1980. Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers. Technical Note TN-6. Geonics Ltd., Mississauga, Ontario, Canada.
Minet, J., Bogaert, P., Vanclooster, M., Lambot, S., 2012. Validation of ground penetrating radar full-waveform inversion for field scale soil moisture mapping. J. Hydrol. 424–425, 112–123.
Minet, J., Verhoest, N.E.C., Lambot, S., Vanclooster, M., 2013. Temporal stability of soil moisture patterns measured by proximal ground-penetrating radar. Hydrol. Earth Syst. Sci. Discuss. 10, 4063–4097.
Neves, H.H.D., Mata, M.G.F.D., Guerra, J.G.M., Carvalho, D.F.D., Wendroth, O.O., Ceddia, M.B., 2017. Spatial and temporal patterns of soil water content in an agroecological production system. Sci. Agric. 74 (5), 383–392.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., the R Development core team, 2012. Nlme: Linear and Nonlinear Mixed Effects Models (R Package Version 3.1-103).
Pollice, A., Bilancia, M., 2002. Kriging with mixed effects models. In: LXII: Statistica, pp. 405–429.
Ripley, B.D., 1988. Statistical Inference for Spatial Processes. Cambridge University Press, Cambridge.
SAS Institute Inc, 2017. SAS/STAT Software Release 9.4. (Cary, NC, USA).
Sheriff, R.E., Geldart, R.E., 1982. Exploration Seismology, Volume 1: History, Theory, and Data Acquisition. Cambridge University Press, New York (253 pp).
Soil Survey Staff, 2010. Keys to Soil Taxonomy, 11th ed. USDA-Natural Resources Conservation Service, Washington, DC.
Steelman, C.M., Endres, A.L., 2011. Comparison of petrophysical relationships for soil moisture estimation using GPR ground waves. Vadose Zone J. 10, 270–285.

Stein, A., Brouwer, J., Bouma, J., 1997. Methods for comparing spatial variability patterns of millet yield and soil data. Soil Sci. Soc. Am. J. 61 (3), 861–870.

Stellacci, A.M., Castrignanò, A., Troccoli, A., Basso, B., Buttafuoco, G., 2016. Selecting optimal hyperspectral bands to discriminate nitrogen status in durum wheat: a comparison of statistical approaches. Environ. Monit. Assess. 188 (3), 1–15. https://doi.org/10.1007/s10661-016-5171-0.

Tang, J., Alelyani, S., Liu, H., 2013. Feature Selection for Classification: A Review. Data Classification: Algorithms and Applications. CRC Press.

Vereecken, H., Huisman, J.A., Bogena, H., Vanderborght, J., Vrugt, J.A., Hopmans, J.W., 2008. On the value of soil moisture measurements in vadose zone hydrology: a review. Water Resour. Res. 44, W00D06.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time. Adv. Agron. 113, 243–291.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer-Verlag, Berlin.

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. Wiley, Chichester.

Wu, X., Zhao, B., 2013. Nonparametric Statistics, 4th ed. China Statistics Press.

Zhu, Q., Lin, H., Doolittle, J., 2010. Repeated electromagnetic induction surveys for determining subsurface hydrologic dynamics in an agricultural landscape. Soil Sci. Soc. Am. J. 74 (5), 1750–1762.