Geoderma 344 (2019) 50–62

Contents lists available at ScienceDirect

Geoderma

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

Spatial filtering and Bayesian data fusion for mapping soil properties: A case study combining legacy and remotely sensed data in Iran

Zahra Rasaei[a,*], Patrick Bogaert[b]

[a] College of Agriculture, Shahrekord University, Shahrekord, Iran
[b] Earth & Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Saturation percentage (SP) filtered kriging
Naive Bayes rule
Digital soil mapping (DSM)

ABSTRACT

In a spatial mapping context, we address in this paper the issue of combining at best soil data coming from legacy soil surveys with soil information that is indirectly obtained from remote sensing (RS). We show first how spatial scale issues need to beproperly addressed for soil mapping in order to increase the predictive performances of spatial prediction models, where the two prediction techniques that are compared here are kriging and random forests (RF). By relying afterwards on a Bayesian data fusion approach, we then emphasize the benefit of combining the output of these two prediction models in order to get a single prediction result that leads to an improved final map.

The advocated methodology is illustrated with the mapping of soil saturation percentage (SP) over a 10,480 km^2 area located in Iran, were SP is a key soil parameter for sustainable development and land management. A set of 396 soil profiles were obtained from legacy soil surveys and were used to compute vertically averaged SP values. In parallel, a set of RS covariates were obtained from a 90 meters resolution Landsat image and a digital elevation model. Based on the modeling of the spatialdependence both for soil SP data and for RS covariates, it is shown how the relationship between them is improved by using a filtered kriging technique that allows us to focus on their long range component only. Using these filtered SP value and RS covariates, predictions were obtained separately from kriging and from a RF model at the nodes of a 90 meters resolution grid. Even if the performances of these two models are almost identical with R^2 values equal to 0.66–0.67, it is shown afterwards that a Bayesian data fusion procedure that combines both predictions yields improved performance with a R^2 value equal to 0.86. These results clearly emphasize the importance of properly addressing scale issues and the benefit of data fusion inorder to improve the quality of the final map. Though the study was conducted here for SP values, we believe this methodology and these findings are relevant when it comes to handle other soil properties in a spatial mapping context that involves several data sources.

1. Introduction

Soils are considered as one of the main factors that need to be accounted for sustainable development and land management. Accordingly, there are high demands for reliable maps of soil characteristics (Keesstra et al., 2016; Montanarella and Vargas, 2012). Though traditional soil surveys can still be used alone in some instances for this goal, remote sensing (RS) technologies proved to be helpful due to their ability of mapping large geographical areas over a short time span and at a high spatial resolution, thus reducing the need of tedious and expensive sampling campaigns and laboratory measurements (Forkuor et al., 2017). In parallel, Digital Soil Mapping (DSM) offers new tools for creating such maps from various data sources and various combinations of spatial interpolation and linear/nonlinear models (Li and Heap, 2014; Minasny et al., 2010; Mulder et al., 2011). Numerous studies have explored the usefulness of RS data with various degrees of success depending on the soil property of interest (Forkuor et al., 2017; Mohamed et al., 2017; Mulder et al., 2011; Pinheiro et al., 2017; Yu et al., 2018).

As a consequence, combined advances in RS and DSM greatly helped producing maps of various soil properties at an affordable price for areas where traditional extensive surveys would be seriously impaired due to budget or accessibility issues (McBratney et al., 2003). However, even if RS data have the advantage of exhaustively covering the spatial area of interest, they do not reach the reliability of results obtained from direct field sampling and laboratory measurements, and direct measurements are still the only way to calibrate and validate the models that are used for predicting soil properties from these RS data.

* Corresponding author.
E-mail addresses: zahra.rasaei@gmail.com (Z. Rasaei), patrick.bogaert@uclouvain.be (P. Bogaert).

https://doi.org/10.1016/j.geoderma.2019.02.031
Received 12 February 2019; Accepted 20 February 2019

Due to time and money limitations, the opportunities of carrying out new traditional soil surveys are however low (Cambule et al., 2015). Hence, making the best possible use of sometimes multiple legacy soil surveys and properly addressing their harmonization is an important aspect of the problem. Some attention has already been paid to the use of legacy soil data in DSM without the need of additional expensive and time consuming field work (Rossiter, 2008; Sulaeman et al., 2013). As shown by (Odeh et al., 2012), past soil inventories are valuable sources of information for improving the sustainable management of soil resources. As an example, Gomez et al. (2016) conclude that legacy data can be useful for correcting soil clay content as predicted from RS data.

Among the numerous soil properties that could be mapped, soil moisture is a key environmental variable, as it is directly linked to nutrient accessibility for plants, along with drought environmental issues and more generally food and water security. Information about the spatial distribution of soil moisture is thus a useful tool for decision makers, especially in arid and semi-arid environments and regions (Legates et al., 2011; Zhan et al., 2007). The spatial distribution of soil moisture can bemodeled from point observations (Kumar et al., 2016), in combination with auxiliary covariates computed from a Digital Elevation model (DEM) or from RS images. Topographic indicators derived from DEM's such as elevation, slope, wetness index, and aspect proved to be useful (Natsagdorj et al., 2017; Takagi and Lin, 2012). In addition, satellite-based spectral reflectances in the visible and near-infrared (VNIR) bands are linked to the water content of soils (Zhan et al., 2007). Some studies indicate the benefit of combining VNIR, short wavelength infrared (SWIR) spectral bands (Liu et al., 2003) and land surface temperature (LST) (Mallick et al., 2009) for estimating soil moisture. For instance, Li et al. (2016) modeled the distribution of soil moisture using Landsat images, with an overall correlation coefficient 0.78 between predicted and observed values, while Natsagdorj et al. (2017) obtained a correlation coefficient of 0.65 bymaking use of a multivariate regression model based on LST, NDVI, elevation, aspect, slope and soil moisture index. Based on linear and nonlinear approaches, Aali et al. (2009) compared the performances of soil SP prediction using Adaptive Neural-based Fuzzy Inference System (ANFIS), artificial neural network (ANN) and multiple linear regression (MLR) models based on auxiliary soil properties (sand, silt, clay, and organic carbon percentage) with correlation coefficient of 0.59, 0.59, and 0.58, respectively.

From these previous studies, it is seen that efficiently predicting soil SP is a difficult task and none of the proposed approaches can be viewed so far as intrinsically superior to the others. This is of course due to our partial knowledge about the complex processes acting at the soil level and various sources of errors in the data (McBratney et al., 2003; Nelson et al., 2011). For ancillary RS data, notwithstanding that some of them are just proxies for the variable of interest, various geometric distortions effects (e.g., systematic error linked sensor displacement and deterioration, curvature and rotation of the earth, terrain relief, etc.) have adverse effects and correction procedures for DEM's and satellite images are strongly suggested (Temme et al., 2009). On the other side, even if direct soil measurements are less prone to major errors, they still remain unavoidable as a result of possibly different sampling and analytic procedures over time, e.g., various operators, various laboratories in charge and various sampling strategies (Nelson et al., 2011). Additionally, their usefulness for spatial prediction is highly dependent on the associated sampling density, with decreasing benefit as the distance between prediction and sampling locations is increasing (Brus and Heuvelink, 2007; Goovaerts, 1997).

Although our study will not investigate new procedures for reducing errors in the data, we will however show in this paper that properly accounting for the spatial dependence structure of these data can greatly help. We will first show how a techniquelike filtered kriging can reduce the effect of errors both for soil samples and RS data, thus allowing the user to improved results when modeling the relationship between soil SP and RS covariates. We will show afterwards that fusing maps obtained from distinct models (i.e. filtered kriging from legacy soil survey on one side and random forest modeling using RS covariates on the other side) can lead to an improved final map. We will rely on a Bayesian data fusion approach as advocated by Bogaert and Fasbender (2007), that proved to be efficient in various contexts (Fasbender et al., 2008; Gengler and Bogaert, 2016; Mattern et al., 2012; Peeters et al., 2010). This paper has therefore two main objectives: (i) to show how legacy soil data can be linked to RS covariates by properly accounting for the spatial structure of the data and to show how data fusion can help improve the prediction accuracy by combining a map obtained from filtered kriging based only on legacy soil data (i.e., using only neighboring soil data for predicting at unsampled locations) with another map obtained from a calibrated random forest model based only on RS covariates (i.e., only accounting for covariates sampled at the prediction locations whatever the spatial dependence of the data).

2. Material and methods

2.1. Study area

The study area is located between 31° 45′ to 33° 15′ N and between 50° 0′ to 51° 30′ E in the border of Isfahan and Chaharmahal-va-Bakhtiari provinces, in the East Zagros region of Iran (Fig. 1). It covers an area of 10,480 km^2, with mean annual temperature ranging from 9.5 °C to 12 °C and mean annual rainfall ranging from 320 mm to 410 mm. Major parent materials are sedimentary rocks (limestone, sandstone, shale and marl) and alluvial sediments. Topography is very diverse, with elevation ranging from 1802 m to 3835 m a.s.l. Accordingly, the study area includes different landform types, i.e., mountains, hills, alluvial gravelly fans, piedmont plains and lowlands. Soils formed on alluvial gravelly fans and high piedmont plains contain coarse parent materials and are shallow with coarse soil texture classes. When moving down toward lowlands containing fine Quaternary limestone deposits, soils are becoming deeper and heavier with a high clay content (Mohammadi, 1999).

2.2. Legacy soil data

Based on investigations for identifying historical soil survey data in the study area, ten legacy soil surveys are available at the 1:50000 scale. After applying rescue and renewal processes (not described here for the sake of brevity) that were conducted according to the recommendations proposed by Rossiter (2008), 396 usable legacy soil profiles were identified. Fig. 1 shows the location of these soil profiles, with sampling densities that largely differ over the study area. Though numerous soil properties are at hand for these profiles at various depths, this study will only focus on their corresponding Saturation Percentage (SP) values, where SP is defined as the ratio of the amount of water added to saturate dry soil samples to the total mass of the fully dried soil. Direct measurements of saturation percentage were done on initially dried soil samples that were saturated with deionized water and then oven dried at 105°C for a period 24 h.

Five standard depths, i.e., 0–5, 5–15, 15–30, 30–60 and 60–150 cm (Arrouays et al., 2014) where considered and the associated SP values where vertically averaged based on a weighted spline interpolation (Malone et al., 2009) using the aqp package in R software (R Development Core Team, 2017), thus leading to a single SP value for each profile. As the depth of the soil profiles varies over the study area and is ranging from less than 50 cm up to 200 cm, we considered that the relevant part of theprofile for SP related to plant growing is the top 150 cm. For profiles shallower than 150 cm, the SP values is thus averaged only over the existing standard depths. For profiles deeper than 150 cm, layers deeper than 150 cm were thus discarded.

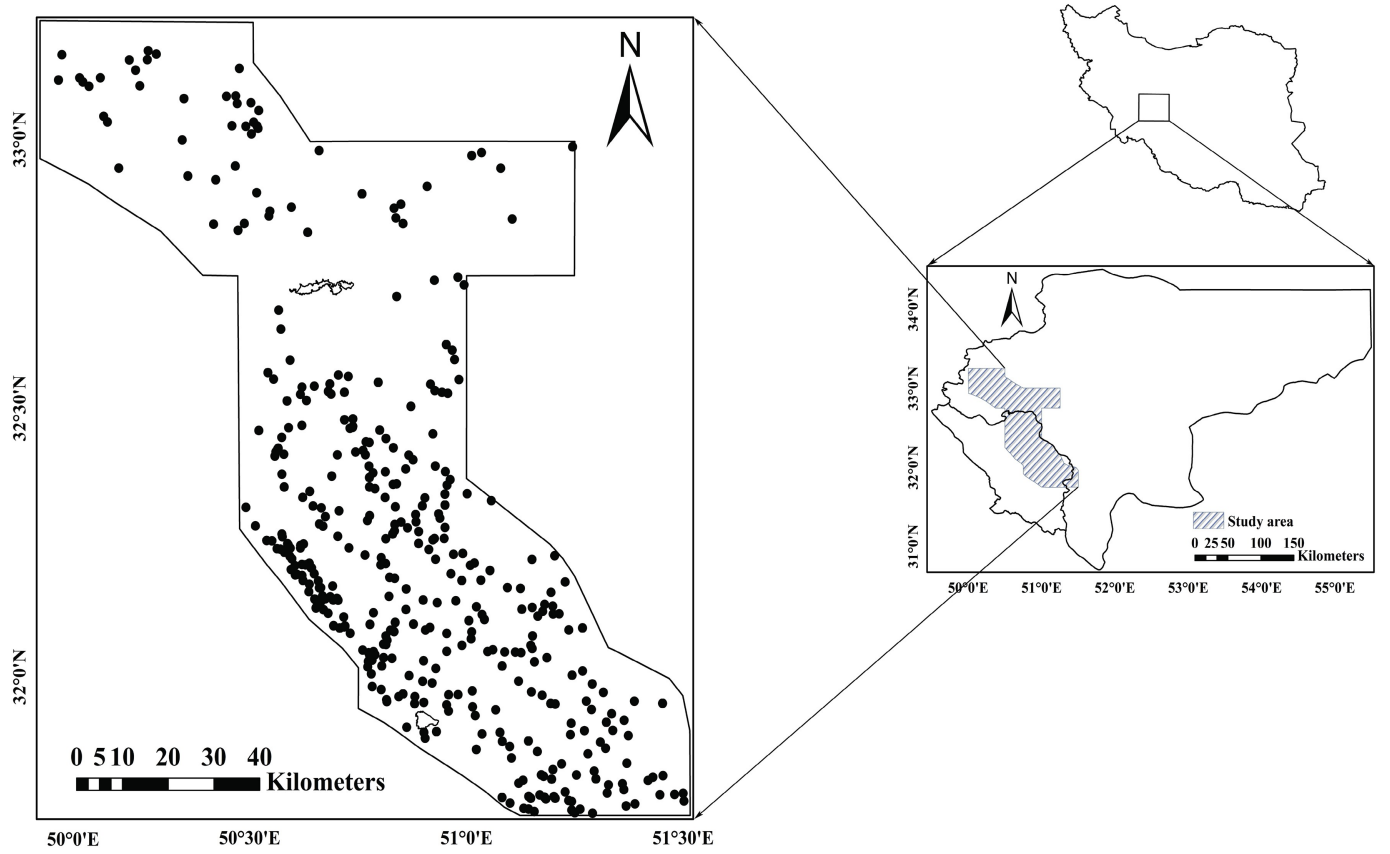

Fig. 1. Location of the study area in Iran, in the neighboring border of Isfahan (top right region of the study area) and Chaharmahal-va-Bakhtiari (down left region of the study area) provinces. Black dots correspond to the location of the 396 sampled soil profiles coming from the legacy data sets. The two major water bodies in the study area (Zayandehrud dam in the north and Choghakhor lagoon in the south) are also delineated.

2.3. Remotely sensed data

Four DEM's where available at a 90 m spatial resolution and covering various parts of the study area (Jarvis et al., 2008). They were merged into a single composite DEM which was then projected in Universal Transverse Mercator (UTM) coordinates. From this composite DEM and its first and second-order derivatives, six variables were obtained, namely Elevation, Slope, Aspect, Valley depth (Abdel-Kader, 2011), Midslope position (Böhner and Antonić, 2009) and Topographic wetness index (Sörensen et al., 2006). Accordingly, six corresponding maps of these variables were obtained at a 90 m resolution at the same nodes than those of the composite DEM. Additionally, all variables were computed at the coordinates of the 396 legacy soil profiles.

In parallel, four Landsat 8 Operational Land Imager images at 30 m spatial resolution were available in July 2016. They were stacked together in order to calculate NDVI (Rouse Jr et al., 1974), Clay index (Boettinger et al., 2008), Gypsum index (Nield et al., 2007), Carbonate index (Boettinger et al., 2008), and Soil moisture index (Dupigny-Giroux and Lewis, 1999). These maps were then resampled at a 90 m resolution and projected to the same coordinate system (WGS 1984, UTM 39N) for the DEM and its derived variables. Additionally, all values of these indices were also extracted at the coordinates of the 396 legacy soil profiles by overlying the soil profiles over the image and by extracting the values of covariates in the corresponding pixels.

2.4. Filtered kriging

When it comes to mapping soil properties from a finite subset of sampled locations irregularly located over space, kriging techniques are standard approaches that are routinely used as a part of the DSM methods. The spatial dependence of any spatial random variable can be first estimated from the data and modelled afterwards by relying on a covariance function $C(\boldsymbol{h})$ or semivariogram $\gamma(\boldsymbol{h})$, where $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}'$ is the distance vector separating two arbitrary locations $\boldsymbol{x}$ and $\boldsymbol{x}'$. This in turn allows us to obtain the best linear unbiased predictor (BLUP) of this variable at any arbitrary location $\boldsymbol{x}_0$, with $Z^p(\boldsymbol{x}_0) = \boldsymbol{\lambda}'\boldsymbol{Z}$, where $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_n)$ is the vector of weights associated with the vector of values $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))'$ that are in the neighbourhood of the prediction location $\boldsymbol{x}_0$. These weights are the solution of a linear system of equations that guarantees $Z^p(\boldsymbol{x}_0)$ is the linear predictor that minimizes the variance of the prediction error $Var[Z(\boldsymbol{x}_0) - Z^p(\boldsymbol{x}_0)]$ subject to the unbiasedness condition $E[Z(\boldsymbol{x}_0) - Z^p(\boldsymbol{x}_0)] = 0$. From classical statistical theory, this predictor is nothing else that the conditional expectation $E[Z(\boldsymbol{x}_0)|\boldsymbol{Z}]$ if one assumes the hypothesis of a joint Gaussian distribution for $(Z(\boldsymbol{x}_0), \boldsymbol{Z}')$.

Although this is a standard way of mapping a spatial variable from non-gridded samples, one might want to predict only part of the variability of this variable by filtering out some of its spatial components. Let us consider that our variable can be decomposed as

$$Z(\boldsymbol{x}) = Z_s(\boldsymbol{x}) + Z_\ell(\boldsymbol{x}) \quad (1)$$

where $Z_s(\boldsymbol{x})$ is referring to short scale components (this including a possible nugget effect) while $Z_\ell(\boldsymbol{x})$ is referring to large scale components (this including the mean $E[Z(\boldsymbol{x})]$, so that $E[Z_s(\boldsymbol{x})] = 0$). We can additionally assume that $Z_s(\boldsymbol{x}) \perp Z_\ell(\boldsymbol{x})$ so that $C_Z(\boldsymbol{h}) = C_{Z_s}(\boldsymbol{h}) + C_{Z_\ell}(\boldsymbol{h})$. A linear predictor of the long range components only with $Z_\ell^P(\boldsymbol{x}_0) = \boldsymbol{\lambda}'\boldsymbol{Z}$, where now $\boldsymbol{\lambda}$ refers to a new set of weights ensuring that $Var[Z_\ell(\boldsymbol{x}_0) - Z_\ell^P(\boldsymbol{x}_0)]$ is minimized, is then given by

$$Var\left[Z_\ell(\boldsymbol{x}_0) - Z_\ell^p(\boldsymbol{x}_0)\right] = Var\left[(1 - \lambda')\begin{pmatrix} Z_\ell(\boldsymbol{x}_0) \\ \boldsymbol{Z} \end{pmatrix}\right]$$
$$= (1 - \lambda')\begin{pmatrix} \sigma^2_{Z_\ell} & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}' & \Sigma_{\boldsymbol{Z}} \end{pmatrix}\begin{pmatrix} 1 \\ -\lambda \end{pmatrix}$$
$$= \sigma^2_{Z_\ell} - 2\lambda'\boldsymbol{\sigma} + \lambda'\Sigma_{\boldsymbol{Z}}\lambda' \quad (2)$$

where ${\sigma_{Z_\ell}}^2$ is the variance of $Z_\ell(\boldsymbol{x}_0)$, $\Sigma_{\boldsymbol{Z}}$ is the covariance matrix of $\boldsymbol{Z}$ and $\boldsymbol{\sigma}$ is the vector of covariances between $Z_\ell(\boldsymbol{x}_0)$ and $\boldsymbol{Z}$. Similarly to ordinary kriging, minimizing Eq. (2) by assuming a constant but unknown mean leads to the linear system of equations

$$\begin{pmatrix} \Sigma_{\boldsymbol{Z}} & \mathbf{1} \\ \mathbf{1}' & \mathbf{0} \end{pmatrix}\begin{pmatrix} \lambda \\ \eta \end{pmatrix} = \begin{pmatrix} \sigma \\ 1 \end{pmatrix} \quad (3)$$

where $\mathbf{1} = (1, \ldots, 1)'$ and $\mathbf{0}$ are the unit vector and the null matrix of appropriate sizes, respectively. By comparison with ordinary kriging used for predicting $Z(\boldsymbol{x}_0)$, the right-hand side of Eq. (3) now only accounts for the covariances $Cov[Z_\ell(\boldsymbol{x}_0), Z(\boldsymbol{x}_i)]$ $(i = 1, \ldots, n)$ through the vector $\boldsymbol{\sigma}$. According to the independence hypothesis $Z_s(\boldsymbol{x}) \perp Z_\ell(\boldsymbol{x})$, we have however $Cov[Z_\ell(\boldsymbol{x}_0), Z(\boldsymbol{x}_i)] = Cov[Z_\ell(\boldsymbol{x}_0), Z_\ell(\boldsymbol{x}_i)]$ so that these covariances correspond to the values of the covariance function $C_{Z_\ell}(\boldsymbol{h})$ where $\boldsymbol{h} = \boldsymbol{x}_0 - \boldsymbol{x}_i$. As long as $C_{\boldsymbol{Z}}(\boldsymbol{h})$ is modelled by adding models for $C_{Z_s}(\boldsymbol{h})$ and $C_{Z_\ell}(\boldsymbol{h})$, Eq. (3) allows us to directly predict the long range component only by filtering out the measurement errors, nugget effect and short scale components.

Filtered kriging has proved to be a valuable tool for dampening either the effect of measurement errors or positioning errors for the sampled locations (Christensen, 2011; Grohmann and Steiner, 2008; Hamzehpour and Bogaert, 2017). As an example, Bourennane et al. (2017) found that filtering nugget effect and short scales spatial components of auxiliary data may enhance their correlation with the target variable. Their results also implied an improvement of the prediction precision using spatially filtered covariates instead of raw covariates. Indeed, we will show in this paper how using filtered SP values instead or raw SP values as coming directly from the samples helps improve the relationship with DEM and RS covariates.

From now on, we will refer to these fSP values as if they were the observable values of the long range component, though they are obtained from filtered kriging here. Only these fSP values will be used in all subsequent computations and comparisons. One could argue that these fSP values are not the true values of the long range component, but these true values will never be at hand, unfortunately. We do however have the guarantee that these fSP values are predicted at best thanks to the properties of filtered kriging, i.e., they are the best linearly predicted values of the long range component in the mean square sense.

2.5. Random forest regression

Random Forests (RF) belongs to the set of machine learning algorithms and is used to model the relationship between a target variable and various possible covariates. RF's rely on the idea of building a multitude of decision trees based on a training data set in order to get the mode (in a classification framework) or the mean prediction (in a regression framework) of these individual trees. RF's are considered as a way to correct for decision trees tendency of overfitting to their training set. As proposed by Breiman (2001), their goal is thus to reduce the variance of the predictions by accounting for the results obtained from numerous deep decision trees that are trained on different parts of the same training set.

In RF's, many unpruned and independent trees are grown, instead of just one tree for a traditional decision tree model. A bootstrap sample function is used to randomly split the dataset into homogeneous subsets with respect to the target variable. A random subset of the samples is thus used to grow and train each tree while the remaining part (the so-called out-of-bag sample) is used to validate the tree and to estimate the model accuracy. The performance of the RF model is estimated from the MSE for the out-of-bag (OOB) samples, with

$$MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{Z}_{\ell,i} - \widehat{Z}_{\ell,i,OOB})^2 \quad (4)$$

where N is the number of observations, $\widehat{Z}_{\ell,i}$ is the average prediction of Z_ℓ over all decision trees for the ith observation and $\widehat{Z}_{\ell,i,OOB}$ is the average of all out-of-bag predictions.

An interesting property of RF's is their ability to assess the importance of the variables that are used as input for the decision trees by alternatively building decision trees with and without each variable and assessing their impact from MSE and node purity. The RF algorithm estimates the importance of a variable by looking at how much the prediction error increases when OOB data for that variable are permuted while all others are left unchanged. The necessary calculations are carried out tree bytree as the random forest is built (Breiman, 2001). In regression random forest, the node purity is measured by the residual sum of squares. OOB samples are thus needed to estimate the performance of the model as well as to select the best input variables.

Thanks to the random bootstrap sampling and random subset selection of variables at nodes, the correlation between decision trees decreases and consequently the accuracy of the RF model increases (Gislason et al., 2006). It is considered that RF's correct for the inherent overfitting that adversely affects decision trees as well as many other machine learning techniques like support vector machines and neural networks. Compared to linear regression, RF's require no assumption about the probability distribution of the target variable and are considered as robust both with respect to non-linearity and overfitting (Statnikov et al., 2008). Furthermore, fast computations, high accuracy predictions along with their ability to handle a very large number of covariates place them among the most accurate available machine learning algorithms. The number of trees that need to be build in the forest (n_{tree}) and the number of input variables used to grow each single tree (m_{try}) are the main parameters that need to be set by the user for building the RF regression model.

When using RF regression in our spatial context, it provides a nonlinear estimate of the conditional expectation $E[Z_\ell(\boldsymbol{x}_0)|\,\boldsymbol{Y}_0]$ at any prediction location $\boldsymbol{x}_0$ where the set of k covariates $\boldsymbol{Y}_0 = (Y_1(\boldsymbol{x}_0), \ldots, Y_k(\boldsymbol{x}_0))'$ is at hand. In the context of spatially predicting soil properties, the high accuracy of RF models has been emphasized in various studies (Blanco et al., 2018; Forkuor et al., 2017; Yang et al., 2016). Accordingly, a RF regression (Liaw and Wiener, 2002) was used to explore the relation between SP values and various covariates as derived from the DEM and RS data. A 10-fold cross-validation (Jung and Hu, 2015) was used to find the optimal parameters n_{tree} and m_{try}. The model was trained on a subset of 296 values (about 75% of the whole data) that were randomly selected from the fSP values. The model performance was evaluated based on R^2 values computed from theremaining 100 fSP values that were used as test data, were R^2 is the squared estimated correlation coefficient between these one hundred $Z_\ell(\boldsymbol{x}_i)$ values and their corresponding predictions as given by $E[Z_\ell(\boldsymbol{x}_i)|\,mathbfY_i]$.

2.6. Bayesian data fusion

Data fusion is a generic term that encompasses various methods and approaches sharing the same goal of combining various and possibly conflicting sources of information in order to obtain a single and harmonized final result. Data fusion is covering a wide variety of possible application fields (Castanedo, 2013; Haghighat et al., 2016). Clearly, any type of parameter estimation involving multiple information sources can benefit from data fusion methods and soil properties are no exception. Numerous studies have highlighted the values of data fusion for mapping soil characteristics (Grunwald et al., 2015; Lu and Han, 2017). We will focus here more specifically on a classical approach

related to Bayesian decision theory (Fassinut-Mombot and Choquel, 2004). Bayesian data fusion (BDF) belongs to the set of high-level fusion methods (Castanedo, 2013; Das, 2008) that proved to be helpful in environmental studies (Gengler and Bogaert, 2016; Mattern et al., 2012) and soil sciences (Brus et al., 2008; Notarnicola and Posa, 2002). Although the method as originally advocated by Bogaert and Fasbender (2007) combines data fusion and geostatistical concepts in a very general way, we will only present here a simplified version ofit that is restricted to the fusion of two collocated information sources and that reverts to the use of a naive Bayes fusion rule in that case.

Focusing on this simplest possible case, let us consider that Z_ℓ, ${}_a{}^P(\boldsymbol{x}_0) = E[Z_\ell(\boldsymbol{x}_0)|I_a]$ and Z_ℓ, ${}_b{}^P(\boldsymbol{x}_0) = E[Z_\ell(\boldsymbol{x}_0)|I_b]$ are two predictors at hand for the same random variable $Z_\ell(\boldsymbol{x}_0)$ at a same location $\boldsymbol{x}_0$ (i.e., the long range component of SP in our study), where I_a and I_b are the corresponding information from which the conditional expectations have been obtained, with associated distributions $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$. What is sought for is the conditional distribution $f(z_\ell(\boldsymbol{x}_0)|I_a, I_b)$ that relies on the joint use of both information sources. From Bayes theorem, one can write that

$$f(z_\ell(\boldsymbol{x}_0) \mid I_a, I_b) \propto f(I_a, I_b \mid z_\ell(\boldsymbol{x}_0)) f(z_\ell(\boldsymbol{x}_0)) \tag{5}$$

Assuming that $I_a \perp I_b \mid z_\ell(\boldsymbol{x}_0)$, i.e. the two information sources are independent conditionally to the true but unknown value $z_\ell(\boldsymbol{x}_0)$, and applying again the Bayes theorem, one thus obtains the result

$$\begin{aligned} f(z_\ell(\boldsymbol{x}_0) \mid I_a, I_b) &\propto f(I_a \mid z_\ell(\boldsymbol{x}_0)) f(I_b \mid z_\ell(\boldsymbol{x}_0)) f(z_\ell(\boldsymbol{x}_0)) \\ &\propto \frac{f(z_\ell(\boldsymbol{x}_0) \mid I_a) f(z_\ell(\boldsymbol{x}_0) \mid I_b)}{f(z_\ell(\boldsymbol{x}_0))} \end{aligned} \tag{6}$$

that corresponds to the so-called naive Bases fusion rule, where $f(z_\ell(\boldsymbol{x}_0))$ plays the role of the prior distribution when no information is at hand. Although no distribution hypotheses are required in general, one can show that if both conditional distributions $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$ are Gaussian, their product is Gaussian too. Based on the same general result, it can also be shown that assuming additionally a Gaussian hypothesis for the prior distribution $f(z_\ell(\boldsymbol{x}_0))$, the fused distribution $f(z_\ell(\boldsymbol{x}_0)|I_a, I_b)$ is then also Gaussian.

Despite its apparent simplicity and the use of a conditional independence hypothesis which is most of the time difficult to assess, the naive Bayes fusion rule proved to be robust (Zhang, 2004) and is widely used in the field of machine learning (Witten et al., 2016).

3. Results and discussion

3.1. Summary statistics

The spatial distribution of the 396 vertically-averaged SP values are shown in Fig. 2a. Lowest and highest SP values are respectively equal to 23% and 63%. The corresponding histogram shows that SP values are in good agreementwith the assumption of a Gaussian distribution (Fig. 2b), with estimated mean and standard deviation equal to 43.4% and 6.84%, respectively. A low coefficient of variation of 16% also indicates a moderate spatial variation of SP values over the area, with highest SP values that are mainly encountered in plains and lowlands.

3.2. Spatial statistics

To the opposite of SP values that are only at hand at 396 locations, the 11 covariates that were obtained both from the DEM (6 of them) or from the RS images (5 of them) are all available over the whole study area at a spatial resolution of 90 m. As an illustration, the maps for Elevation, Midslope position, Valley depth (all computed from the DEM) and for Carbonate index (as computed from RS images) are presented in Fig. 3.

The variogram analysis conducted for SP shows that about 50% of the total variability can be associated with a long range spatial component. Due to the spacing between soil profiles and the lack of repeated measurements at the same locations, it is however impossible to separately identify the parts of the short scale components that are either associated with measurement errors, nugget effect and possibly real short scale effects (i.e. spatial effects acting at distances lower than about two kilometers in our case). Accordingly, the variogram for SP values was modeled using a combination of nugget effect and exponential models, with

$$\gamma_{SP}(\boldsymbol{h}) = \sigma^2_{Z_s} Nug(\|\boldsymbol{h}\|) + \sigma^2_{Z_\ell} Exp(\|\boldsymbol{h}\|; r_{Z_\ell}) \tag{7}$$

with short and long scale variances ${\sigma_{Z_s}}^2$ and ${\sigma_{Z_\ell}}^2$ equal to 27 %^{2} and 30 %^{2}, respectively, and with a long scale range r_{Z_ℓ} equals to 50 km (i.e. the Euclidean distance $\|\boldsymbol{h}\|$ needed for reaching 95% of the ${\sigma_{Z_\ell}}^2$ value).

For all DEM and RS covariates, thanks to the 90 m resolution of the grid and the 1,267,420 corresponding values, the variogram analysis allows us to get reliable information at much lower distances. As seen from Fig. 4b to e, the spatial characteristics are widely varying from one covariate to another. While a covariate like Carbonate Index (rCaI) exhibits a behaviour quite similar to SP, other covariates are including various scale effects in different proportions. Focusing on rCaI only, Fig. 4b shows that it can be modeled as the sum of two exponential models associated with ranges that differ by more than one order of magnitude, with

$$\gamma_{CaI}(\boldsymbol{h}) = \sigma^2_{Y_s} Exp(\|\boldsymbol{h}\|; r_{Y_s}) + \sigma^2_{Y_\ell} Exp(\|\boldsymbol{h}\|; r_{Y_\ell}) \tag{8}$$

with short scale and long scale variances ${\sigma_{Y_s}}^2$ and ${\sigma_{Y_\ell}}^2$ equal to 0.01 and 0.007, respectively, and with short and long scale ranges r_{Y_s} and r_{Y_ℓ} equals to 2.5 km and 50 km, respectively.

From the first line of Table 1, it can be seen that the correlation between raw SP values (rSP) and the aforementioned covariates are rather moderate (discarding at this stage the results referring to filtered SP and filtered Carbonate Index that will be explained later on), with the highest observed correlations equal to -0.30 for rCaI and 0.32 for Valley depth (Val). All other covariates that were not presented here (7 of them) exhibit correlations with rSP that are lower than 0.1 in absolute value. The negative relationship between SP and Carbonate index is corroborating the relationship between reflectance in the red bands and soil moisture content, as in Zhan et al. (2007).

From the limited number of SP samples, it is clear that one cannot decide if the variance ${\sigma_{Z_s}}^2$ is really corresponding to a pure nugget effect (as modeled here in Eq. (7)) or is rather associated with a possible short scale component that cannot be captured in the analysis due to the distances between samples. From the correlation between rSP and rCaI and the fact that rCaI is clearly including such a short scale component with no nugget effect (Fig. 4b), one can indeed suspect that a short scale component is a possibility for rSP too. On the other side, our ability to map SP from spatial interpolation techniques is restricted to the spatially correlated part of the variability, thus excluding the prediction of thenugget effect component. In order to investigate these aspects, it is suggested to proceed with a filtering step, both with the aim of focusing only on the useful part of variability when it comes to mapping SP values and in order to see how filtered values can improve our ability to relate SP with other covariates. Indeed, (i) as we are interested in a mapping at a large scale (so that short scale variations are not of real interest for this map), (ii) as we expect a nugget effect along with measurement and positioning errors, (iii) as even if there is a real short scale component, we cannot assess its importance and it will add to the true nugget effect at the modeling stage, this justifies the fact that only the long-range component is useful for practical purposes here.

3.3. Filtered kriging

Based on filtered kriging (Eq. (3)) and the variogram model for SP (Eq. (7)), it is possible to split rSP values as in Eq. (1) at the 396 locations in order to get estimates for $Z_\ell(\boldsymbol{x})$ only. The corresponding

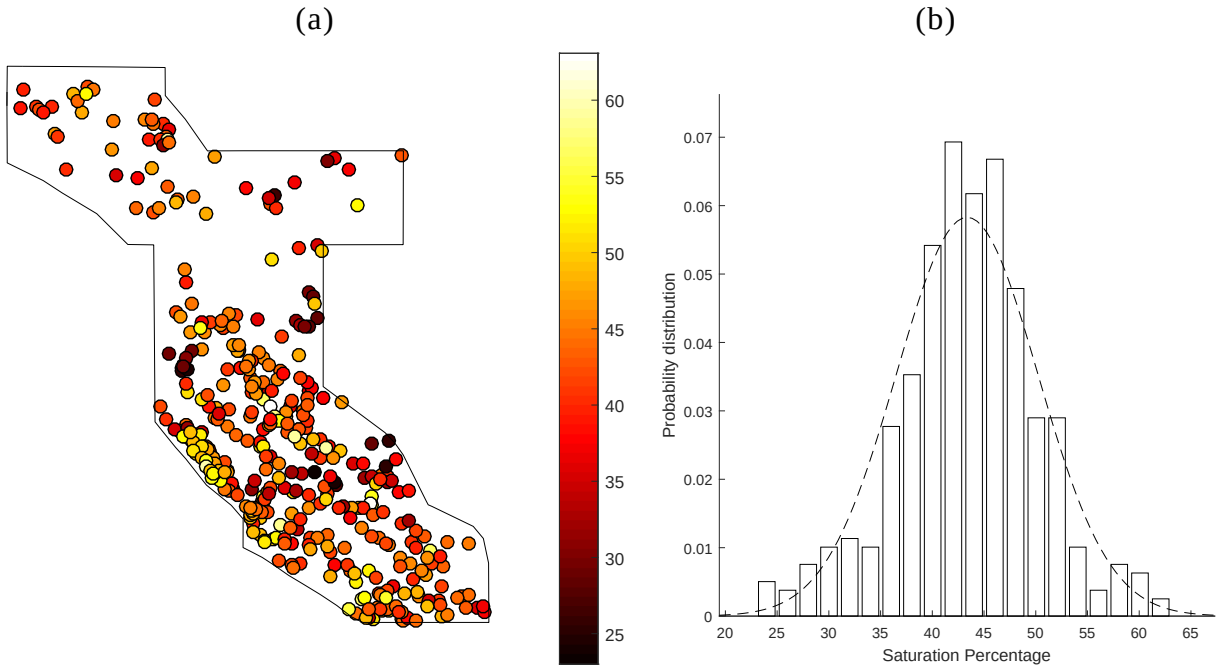

Fig. 2. Profile-averaged values of Saturation Percentage (SP) for the 396 legacy soil profiles over the study area (a). As seen from (b), SP values are Gaussian distributed, with estimated mean and standard deviation equal to 43.4% and 6.84%, respectively.

values in Fig. 5a now only retains the part of the variability related to the long range component, as seen from a comparison with Fig. 2a. From the comparison between the first and second lines in Table 1, it can be seen that all correlations between these filtered SP values (fSP) and the other covariates have been increased, with correlations now equal to -0.47 and 0.45 for rCaI and Val. For covariates that did not include short scale components, these increased correlations is an expected result as fSP values are now related to more similar spatial scales with these covariates. For rCaI, the reason is less obvious, but one might suspect that only the long range components of rSP and rCaI are clearly related, while their corresponding short scale components are not. In order to confirm this hypothesis, the same filtering procedure was applied for rCaI by retaining only the long range component both at the nodes of the 90 meters resolution grid and at the 396 locations that were sampled for SP. The filtered values for Carbonate index (fCaI) are presented in Fig. 5b, which is to be compared to the initial rCaI values as presented in Fig. 3a. From Table 1, one can see that the correlation between fCaI and fSP is now equal to -0.62, i.e. twice the correlation that was initially observed between rSP values rCaI. The correlations between fCaI and the other covariates are slightly improved as well.

The previous analyses confirm that scale effects can play adverse effects when one tries to relate spatial variables together. From the raw values of SP and other covariates, one can only observe weak global correlations due to the fact that these variables may be related differently at distinct spatial scales. Focusing on more comparable spatial scales allows us to improve the correlations with (and thus the usefulness of) these covariates. A procedure like filtered kriging thus helps us to reach two objectives. The first one is to map the relevant part of the spatial variability for SP by removing nuisance effects like possible measurement and positioning errors (see (Hamzehpour and Bogaert, 2017) for a thoroughful discussion about these aspects)that are likely to appear when processing legacy data. This allows us to obtain a map of fSP values over the whole area, as presented in Fig. 8a. The second objective is to enhance the relationship with other covariates that can in turn be used in alternate spatial regression models for predicting SP. The next section will address specifically this last issue.

3.4. Random forest regression

As RF regression aims at predicting at best SP values based on the set of eleven covariates as obtained from DEM and RS images, it is thus of some concern here to identify the best subset of these covariates to be used in order to achieve the highest possible predictive power for the model. This step is also referred to as feature subset selection in machine learning literature (Kuhn, 2011; Kuhn and Johnson, 2013). In a DSM context, the importance of feature selection has been acknowledged by Blanco et al. (2018), Ließ et al. (2016), Xiong et al. (2014), among others. Though numerous algorithms are possible for this task (Forkuor et al., 2017; Ließ et al., 2016; Taghizadeh-Mehrjardi et al., 2015; Xiong et al., 2014), only stepwise regression, Elastic net regression (Friedman et al., 2016) and the Boruta method (Kursa et al., 2010; Kursa and Rudnicki, 2010) have been considered and compared here. To the best of our knowledge, Elastic net has not been used yet for DSM, while Boruta has been rarely used although its ability to select the best features has been shown by Ließ et al. (2014).

Table 2 summarizes the results for the covariates selection, where for each selection algorithm the covariates are vertically sorted in descending order of importance in the corresponding RF model. In order to emphasize the effect of focusing on filtered SP values, the results are also presented both for rSP and for their filtered counterpart (fSP) as obtained from the filtered kriging procedure in Section 3. For all feature subset selection algorithms, the 396 legacy soil profiles were randomly split into a training set (296 profiles) and a test set, with the RF model fitted on the training set only. The whole procedure was repeated 20 times so that the R^2 values presented in Table 2 are the averagedR^2 values over these twenty runs. For each RF model, 500 trees ($n_{tree} = 500$) where built and 2 randomly selected predictors ($m_{try} = 2$) where considered each time.

As seen from a comparison between algorithms, Boruta selection outperforms both stepwise and Elastic net selections by using a lower number of covariates into the RF modelling while leading to higher performance. The selected covariates from Boruta (namely fCaI, El, Val and MidS in descending order of importance) are those already presented in Fig. 3 and in Fig. 5b. Clearly, the ability of predicting rSP values is very limited whatever the algorithm which is used, while

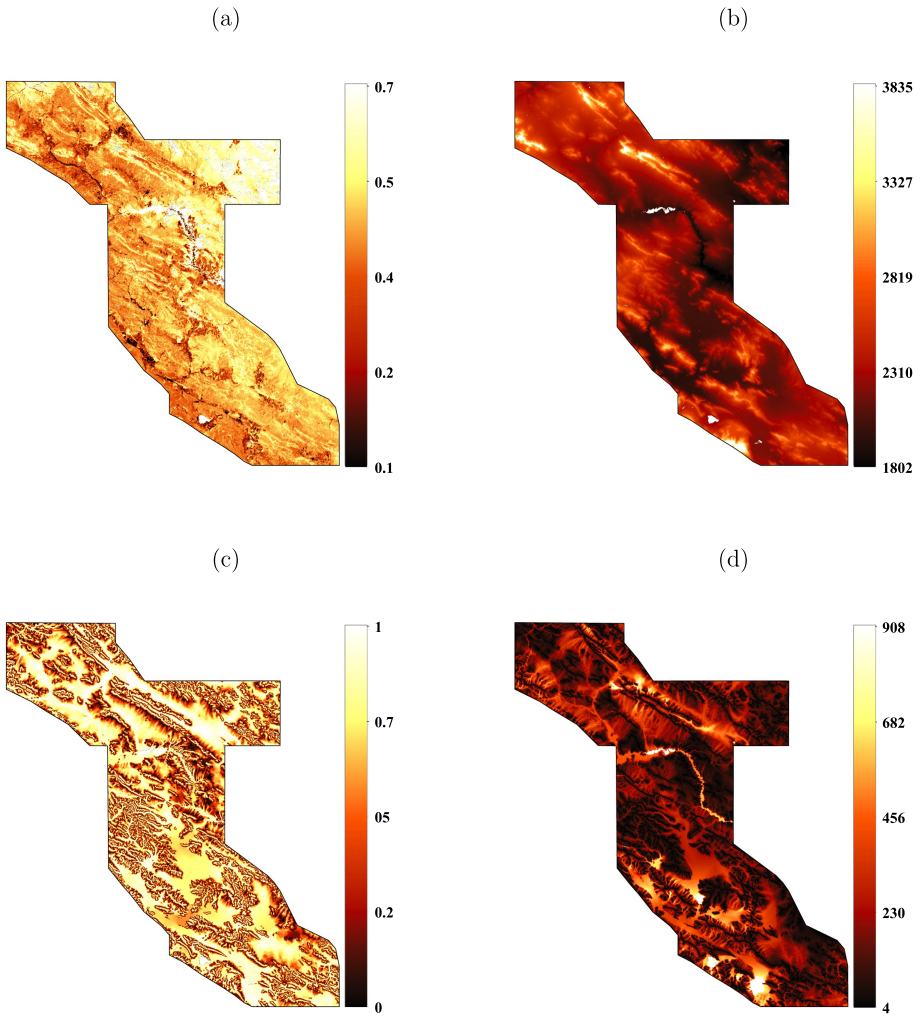

Fig. 3. Maps of the four covariates that are used in the Random Forest model for predicting SP values over the study area, with (a) Carbonate index, (b) Elevation, (c) Midslope position and (d) Valley depth. White areas that are identical on all mapscorrespond to the main water bodies.

performances are increased when focusing on fSP and when using fCaI instead of rCaI. Out of the four variables selected by Boruta, three of them are related to surface topography while a single one (fCaI) is related to the reflectance of the soil surface. Itcan be concluded that the spatial distribution of SP values can be jointly explained in part by spectral properties and topographic features, as supported by other studies (McBratney et al., 2003; Natsagdorj et al., 2017; Takagi and Lin, 2012). The global average performance of the corresponding RF models for fSP is however still moderate, with a R^2 equal to 0.61. One can also note the discrepancies between R^2 values obtained for the training set and for the test set, suggesting thatoverfitting still remains an issue here, with average R^2 values that are all close to 0.9 for the training set, whatever the selection algorithm.

Based on the four selected variables from the Boruta algorithm, the RF regression model that yielded the best performance out of the twenty runs was finally selected, with R^2 equal to 0.67 for the 100 locations of the corresponding test set. This result is consistent with the $R^2 = 0.65$ as reported by Natsagdorj et al. (2017) while fitting a multiple linear regression model to predict soil moisture based on a combination of topographic and spectral covariates.

3.5. Bayesian data fusion

At this stage, we now have at hand two spatial predictors for the long range component $Z_\ell(\boldsymbol{x}_0)$ for SP over the study area. The first one, say $Z_{\ell,\ a}{}^{p}(\boldsymbol{x}_0) = E[Z_\ell(\boldsymbol{x}_0)|I_a]$, corresponds to the filtered kriging predictor applied at the nodes of the same grid, where $I_a = \boldsymbol{Z}$ is the information conveyed by neighbouring locations where rSP values are at hand. The second one, say $Z_{\ell,\ b}{}^{p}(\boldsymbol{x}_0) = E[Z_\ell(\boldsymbol{x}_0)|I_b]$, correspondsto the RF predictor applied at the nodes of the 90 m resolution grid as presented in Section 4, where $I_b = \boldsymbol{Y}_0$ is referring to the information conveyed by the four selected covariates at the same locations. Both predictors thus lead to distinct maps of the long range component for SP, as seen from Fig. 8a and b. Using the BDF framework as presented in Section 6, the idea is now to conciliate both maps in order to obtain a better single map of the estimate $E[Z_\ell(\boldsymbol{x}_0)|I_a, I_b]$, i.e. the expectation of the fused distribution $f(z_\ell(\boldsymbol{x}_0)|I_a, I_b)$ as obtained from Eq. (6), that accounts for the varying quality of both predictors depending on the location $\boldsymbol{x}_0$.

For the prior distribution $f(z_\ell(\boldsymbol{x}_0))$, we will consider it as Gaussian with mean equal to 43% (i.e. the mean of the fSP values at the 396 locations) and standard deviation equal to $\sigma_{Z_\ell} = 5.48\%$ (i.e. the square root of the variance $\sigma_{Z_\ell}{}^2$ for the long-range component in Eq. (7)). For the RF regression model, we will consider a Gaussian distribution too with a standard deviation equal to 2.29%, as estimated from the errors of the prediction at the 100 randomly selected locations that were used for validating the selected model (see Fig. 6b). While the prior distribution is the same for all possible locations $\boldsymbol{x}_0$, the mean of the RF distribution $f(z_\ell(\boldsymbol{x}_0)|I_a)$ differs from place to place, with however the same standard deviation for all $\boldsymbol{x}_0$. Finally, for the distribution $f(z_\ell(\boldsymbol{x}_0)|I_b)$ of the filtered kriging predictor, both its mean and its standard deviation depend on the prediction location.

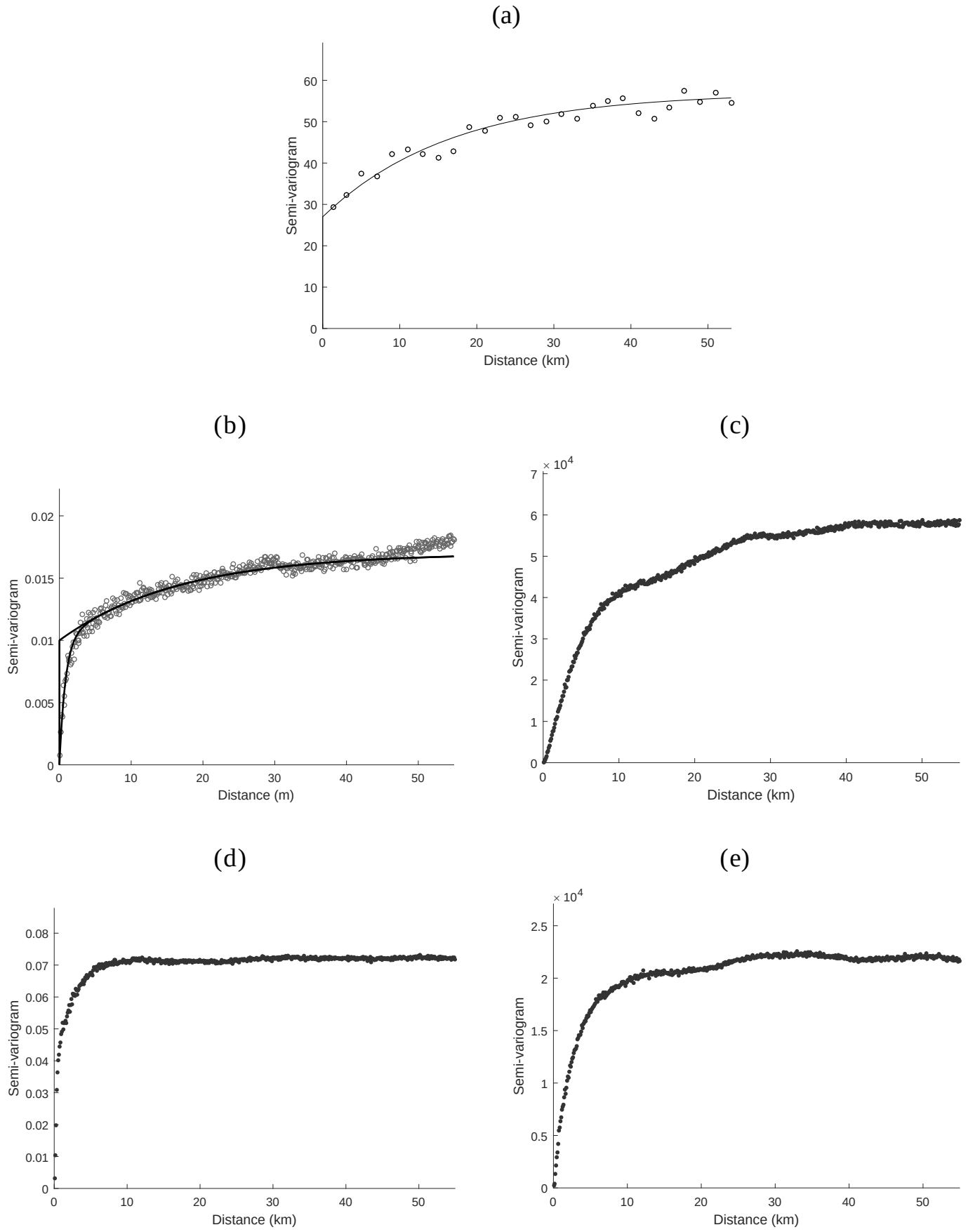

Fig. 4. Semi-variograms for (a) Saturation Percentage (SP) based on the 396 legacy soil profiles, along with semi-variograms for (b) Carbonate index, (c) Elevation, (d) Midslope position and (e) Valley depth, respectively.

Table 1
Correlation between SP values and selected covariates (rSP: raw SP, fSP: filtered SP, rCaI: raw carbonate index, fCaI: filtered carbonate index, El: elevation, MidS: midslope position, Val: valley depth).

Variable	rSP	fSP	rCaI	fCaI	El	MidS	Val
rSP	1	0.77	−0.30	−0.38	−0.11	0.22	0.32
fSP		1	−0.47	−0.62	−0.18	0.26	0.45
rCaI			1	0.39	0.01	−0.18	−0.29
fCaI				1	−0.10	−0.20	−0.37
El					1	−0.19	−0.30
MidS						1	0.50
Val							1

Table 2
Results of various feature selection algorithms for prediction SP values (CI: clay index, GI: gypsum index, Sl: slope. Refer to Table 1 for the other abbreviations).

Algorithm		Stepwise		Elastic net		Boruta	
Selected covariates		rSP	fSP	rSP	fSP	rSP	fSP
		CI	fCaI	GI	fCaI	Val	fCaI
		rCaI	CI	CI	GI	El	El
		GI	GI	rCaI	CI	rCaI	Val
		Val	Val	Val	Val	MidS	MidS
		Sl	Sl	MidS	MidS		
R^2	Training	0.90	0.89	0.90	0.91	0.93	0.92
	Test	0.28	0.36	0.30	0.48	0.36	0.61

In order to illustrate the way BDF works four our data, Fig. 7 illustrates few situations that can be encountered (out of the 100 randomly selected locations considered in Fig. 6). For each of these situations, one needs to consider two things. The first one is related to the information conveyed by I_b and I_b, as measured by how much the distributions $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$ differ from the prior distribution $f(z_\ell(mathbfx_0)$. The second thing is the agreement between the two distributions $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$, that are obtained from distinct information I_b and I_b.

Typically, Fig. 7a corresponds to the case where the prediction location is far away from any sampled location for SP, so that the filtered kriging predictor is close to the prior distribution with $f(z_\ell(\boldsymbol{x}_0)|I_a) \simeq f(z_\ell(\boldsymbol{x}_0)$ and thus I_a does not convey much information. As a result, from Eq. (6), the fused distribution is close to the RF distribution $f(z_\ell(\boldsymbol{x}_0)|I_b)$ at that location, where I_b is the only valuable information at hand.

In Fig. 7b, both I_a and I_b are informative (i.e. $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$ differ from the prior $f(z_\ell(\boldsymbol{x}_0))$ and they are also similar to each other. As a consequence, the fused distribution leads to a lowered uncertainty about the predicted value (i.e. a lower variance) while having same expectation as for $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$. Stated otherwise, the predictedvalue from the fused distribution is the same than for filtered kriging and RF, but with a lowered prediction variance.

In contrary, Fig. 7c corresponds to filtered kriging and RF predictors that disagree (their distributions are different) but are both informative (their distributions both differ from the prior one). As a consequence, the fused distribution is a compromise between both predictors, with an expectation in between those for $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$. It is worth noting however that the expectation of the fused distribution is not merely an average of the expectations for $f(z_\ell(\boldsymbol{x}_0)|I_a)$ and $f(z_\ell(\boldsymbol{x}_0)|I_b)$. Indeed, the result of the fusion also depends on the position of these distributions with respect to the prior. This is illustrated in Fig. 7d, where the expectation of the fused distribution is higher that those of each predictor. This is a logical consequence of the fact that both predictors are in agreement and are informative. As a consequence, the fused distribution emphasizes thefact that SP values should definitively be high, thus leading to an expectation that is higher than for each predictor.

In order to illustrate the performance of BDF for our data set, Fig. 6a and b present the result of the prediction at the same 100 locations for the RF predictor and for a cross-validation of the filtered kriging predictor, leadingto very similar R^2 values so that both predictors have comparable average performances. As seen from Fig. 6c, using BDF leads to improved predictions, with a $R^2 = 0.86$, i.e. an average gain of 20% by comparison with each predictor considered separately. From Fig. 8c, it can be seen that the final fused map is merging the features of the RF and filtered kriging maps, according to the naive Bayes rule as illustrated in Fig. 7. From the results in Fig. 6, there is a also strong evidence that this fused map is globally a better picture of the true long range component for SP values than the maps obtained either from RF regression or from filtered kriging alone.

4. Conclusions

When addressing the issue of mapping soil properties in the best

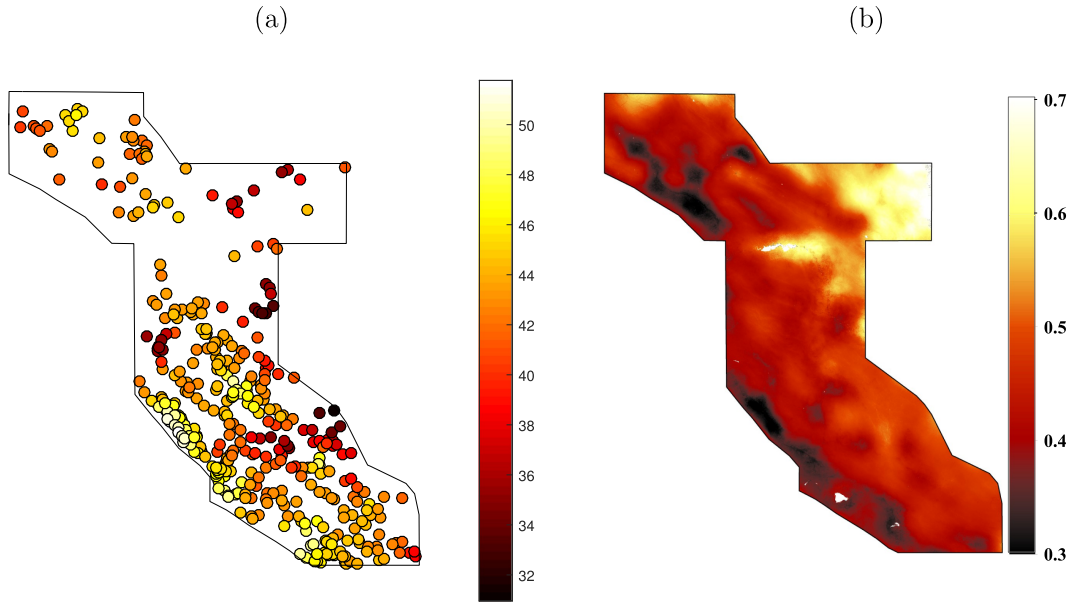

Fig. 5. Spatially filtered values for (a) SP from the 396 legacy soil profiles and (b) Carbonate index from RS images, with an estimated negative correlation of −0.62 between filtered values for these two variables.

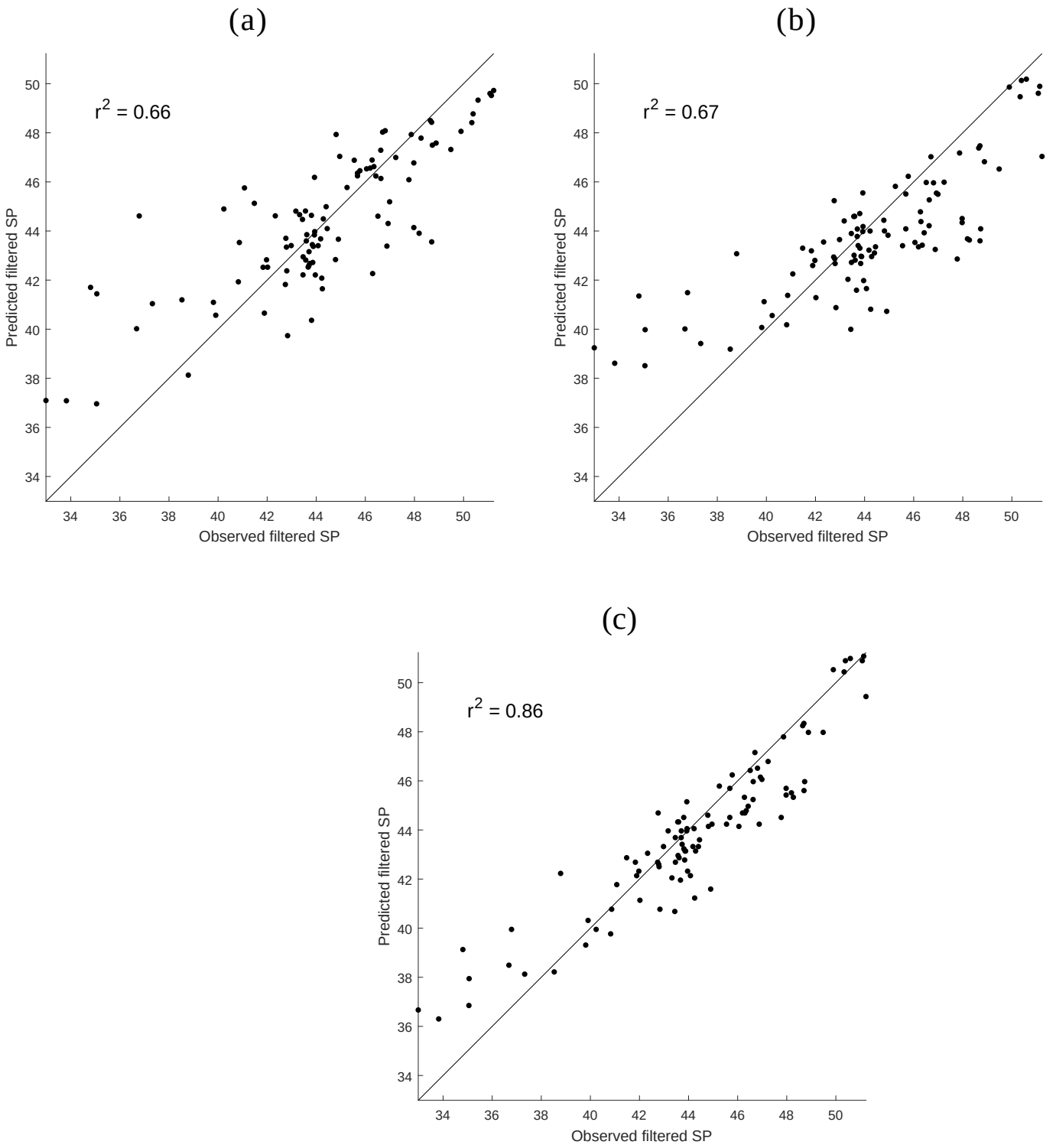

Fig. 6. Cross-validation results for SP prediction at the 100 locations selected for validating the selected Random Forest (RF) model. Parts (a) and (b) refer to the results for filtered kriging and random forest regression, respectively. Part (c) refers to the results for the Bayesian data fusion at the same 100 locations.

possible way, it is common that several sources of information are at hand for this goal. In this paper, we focused on the prediction of SP values over an area of 10,480 km^2 locatedin Iran by combining legacy soil profiles coming from various surveys and more recent remotely sensed images. We have shown that properly accounting for the spatial scales at which the process is acting is an important aspect. We also emphasized the benefit of fusing maps obtained from these distinct information sources in order to obtain a single and better final product.

With respect to the spatial scale, we emphasized that trying to directly relate SP values with covariates (using a RF regression model here) can lead to disappointing results, with R^2 ranging here from 0.28 to 0.36 depending on the covariates selection algorithm, so that the benefit of using such a model for predicting SP values would be quite limited. However, it was also shown that the relationship between SP and covariates is scale-dependent. Indeed, based on a filtered kriging approach that aimat accounting only for the long range variability of SP

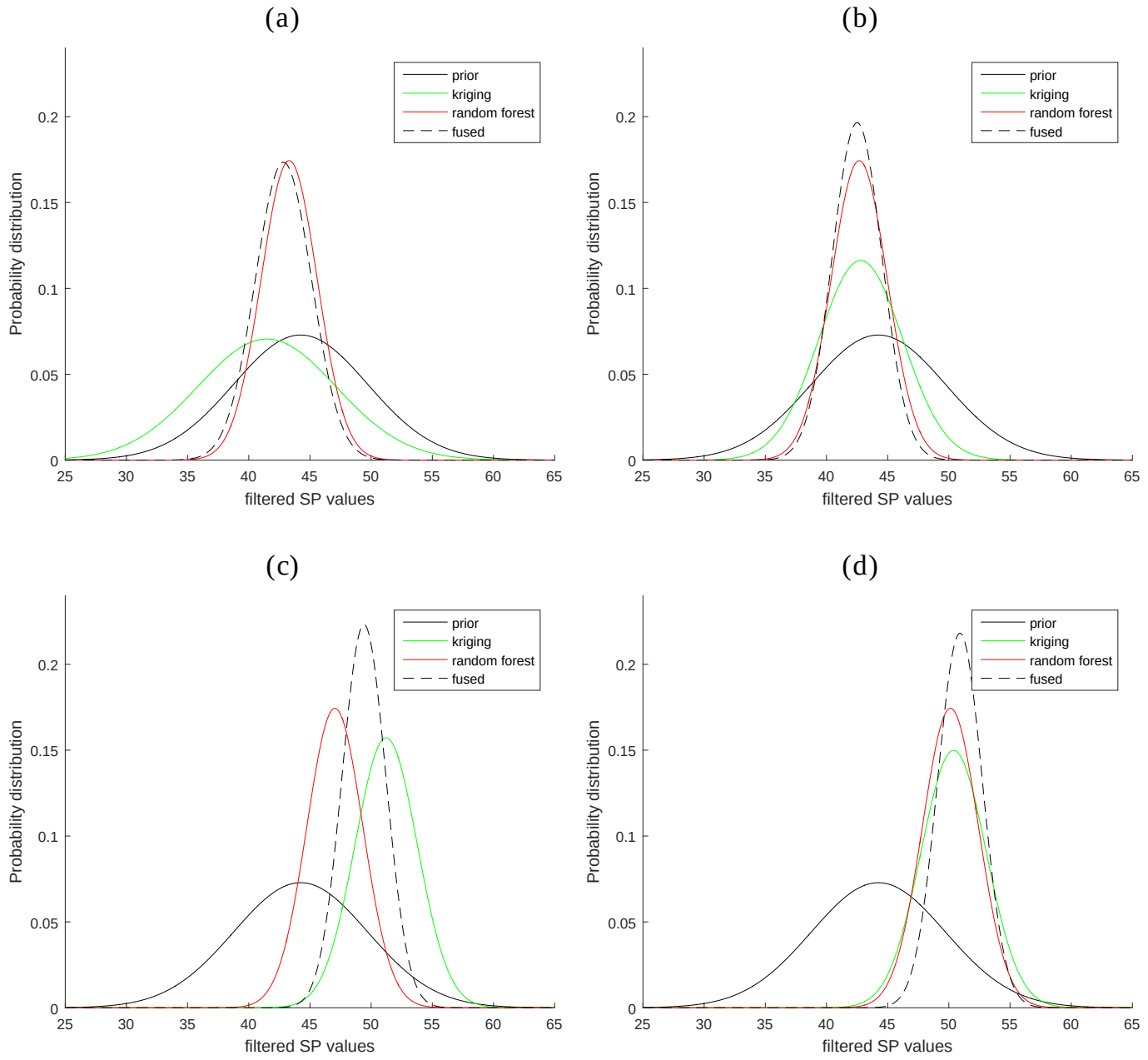

Fig. 7. Illustration of the data fusion procedure for various situations. Part (a) corresponds to a weakly informative filtered kriging prediction. Part (b) corresponds to predictions from filtered kriging and RF that are both informative and in agreement with each other, while for Part (c) filtered kriging and RF predictions are both informative but disagree to some extent with each other. Part (d) illustrates that the mean of the fused distribution is not necessarily convex with respect to themean of each prediction.

and Carbonate index values, the performances of our model was largely increased, with R^2 ranging from 0.36 to 0.61 depending again on the covariates selection algorithm. We thus found that while performance of regression-like models can be disappointing at a first sight, this issue might disappear when one is focusing on the appropriate range for the spatial variability. In our case, this increase of performances when dealing with spatiallyfiltered values is also an expected result if one considers that the various errors and uncertainties that can impact our data are mainly acting at short spatial scales (i.e., measurement errors, positioning errors, along with the effect of mixing point-based soil profiles with pixel-based covariates), so that there is little point in incorporating them into the prediction process itself.

With respect to data fusion, we emphasized the benefit of the BDF approach that aims at merging the output of the filtered kriging method using legacy soil profiles on one side and a RF regression model built from our selected covariates on the other side. While the performances for filtered kriging and RF regression where quite similar (with $R^2 = 0.66$ and 0.67, respectively), the final fused map is improved (with $R^2 = 0.86$), thanks to the property of the naive Bayes fusion rule that properly accounts for the information content conveyed at each location by each prediction method, where this information content differs from one location to another one.

Although the importance of the spatial scales and the benefit of data fusion have been emphasized here for the specific case of SP mapping, it is clear that these main general results are potentially relevant for any other soil property that needs to bemapped from various data sources and using possibly different methods than ours. We thus expect these findings to be beneficial for the soil science community in general.

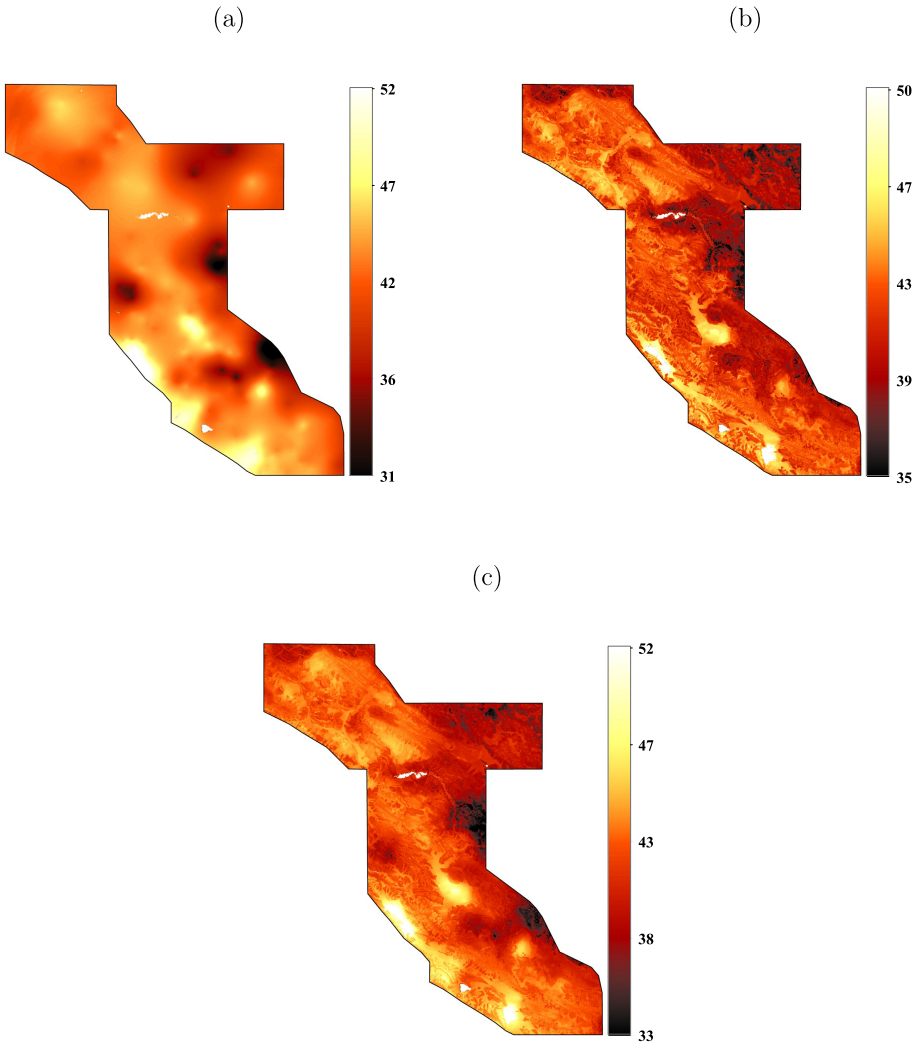

Fig. 8. Maps of the predicted values for filtered SP over the study area, with (a) predictions from filtered kriging, (b) predictions from random forest regression, and (c) Bayesian data fusion of the two previous maps.

Acknowledgment

Authors are indebted to two anonymous reviewers for their detailed and numerous comments that greatly helped improving the manuscript. The first author also thanks the Iranian Ministry of Science, Research and Technology for fully supporting her research work.

References

Aali, K.A., Parsinejad, M., Rahmani, B., 2009. Estimation of saturation percentage of soil using multiple regression, ANN, and ANFIS techniques. Comput. Inform. Sci. 2, 127–136.
Abdel-Kader, F.H., 2011. Digital soil mapping at pilot sites in the northwest coast of Egypt: a multinomial logistic regression approach. Egyptian J. Remote Sens. Space Sci. 14, 29–40.
Arrouays, D., McKenzie, N., de Forges, A.R., Hempel, J., McBratney, A.B., 2014. GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC press.
Blanco, C.M.G., Gomez, V.M.B., Crespo, P., Ließ, M., 2018. Spatial prediction of soil water retention in a páramo landscape: methodological insight into machine learning using random forest. Geoderma 316, 100–114.
Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., et al., MendonÃa-Santos, M., 2008. Landsat spectral data for digital soil mapping. In: Hartemink, A.E., McBratney, A.B. (Eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht, pp. 193–202.
Bogaert, P., Fasbender, D., 2007. Bayesian data fusion in a spatial prediction context: a general formulation. Stoch. Env. Res. Risk A. 21, 695–709.
Böhner, J., Antonić, O., 2009. Chapter 8 land-surface parameters specific to topo-climatology. In: Hengl, T., Reuter, H.I. (Eds.), Geomorphometry Concepts, Software, Applications. Developments in Soil Science. Elsevier, pp. 195–226.
Bourennane, H., Hinschberger, F., Chartin, C., Salvador-Blanes, S., 2017. Spatial filtering of electrical resistivity and slope intensity: enhancement of spatial estimates of a soil property. J. Appl. Geophys. 138, 210–219.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86–95.
Brus, D.J., Bogaert, P., Heuvelink, G.B.M., 2008. Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. Eur. J. Soil Sci. 59, 166–177.
Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2015. Rescue and renewal of legacy soil resource inventories: a case study of the limpopo national park, mozambique. Catena 125, 169–182.
Castanedo, F., 2013. A review of data fusion techniques. Sci. World J. 2013, 1–19.
Christensen, W.F., 2011. Filtered kriging for spatial data with heterogeneous measurement error variances. Biometrics 67, 947–957.
Das, S.K., 2008. High-Level Data Fusion. Artech House Publishers.
Dupigny-Giroux, L.A., Lewis, J.E., 1999. A moisture index for surface characterization over a semiarid area. Photogramm. Eng. Remote. Sens. 65, 937–946.
Fasbender, D., Radoux, J., Bogaert, P., 2008. Bayesian data fusion for adaptable image pansharpening. Trans. Geosci. Remote Sens. 46, 1847–1857.
Fassinut-Mombot, B., Choquel, J.B., 2004. A new probabilistic and entropy fusion approach for management of information sources. Inform. Fusion 5, 35–47.
Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS One 12, 0170478.
Friedman, J., Hastie, T., Simon, N., 2016. Tibshirani, R., 2016. Lasso and Elastic-Net Regularized Generalized Linear Models. r-package Version 2.0–5.
Gengler, S., Bogaert, P., 2016. Bayesian data fusion applied to soil drainage classes spatial mapping. Math. Geosci. 48, 79–88.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recogn. Lett. 27, 294–300.

Gomez, C., Gholizadeh, A., Boruvka, L., Lagacherie, P., 2016. Using legacy data for correction of soil surface clay content predicted from VNIR/SWIR hyperspectral airborne images. Geoderma 276, 84–92.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.
Grohmann, C.H., Steiner, S.S., 2008. SRTM resample with short distance-low nugget kriging. Int. J. Geogr. Inf. Sci. 22, 895–906.
Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of soil and remote sensing data to model soil properties. In: Grunwald, S. (Ed.), Advances in Agronomy. Elsevier, pp. 1–109.
Haghighat, M., Abdel-Mottaleb, M., Alhalabi, W., 2016. Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition. Trans. Inform. For. Sec. 11, 1984–1996.
Hamzehpour, N., Bogaert, P., 2017. Improved spatiotemporal monitoring of soil salinity using filtered kriging with measurement errors: An application to the west urmia lake, iran. Geoderma 295, 22–33.
Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled srtm for the Globe Version 4. Available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi. cgiar. org).
Jung, Y., Hu, J., 2015. A k-fold averaging cross-validation procedure. J. Nonpara. Stat. 27, 167–179.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., et al., 2016. The significance of soils and soil science towards realization of the united nations sustainable development goals. Soil 2, 111–128.
Kuhn, M., 2011. Variable Selection Using the Caret Package.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York.
Kumar, K., Arora, M.K., Hariprasad, K.S., 2016. Geostatistical analysis of soil moisture distribution in a part of solani river catchment. Appl Water Sci 6, 25–34.
Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36, 1–13.
Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta a system for feature selection. Fund. Inform. 10, 271–285.
Legates, D.R., Mahmood, R., Levia, D.F., DeLiberty, T.L., Quiring, S.M., Houser, C., Nelson, F.E., 2011. Soil moisture: a central and unifying theme in physical geography. Prog. Phys. Geogr. Earth Environ. 35, 65–86.
Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: A review. Environ. Model Softw. 53, 173–189.
Li, B., Ti, C., Zhao, Y., Yan, X., 2016. Estimating soil moisture with landsat data and its application in extracting the spatial distribution of winter flooded paddies. Remote Sens. 8, 38–50.
Liaw, A., Wiener, M., 2002. Classification and Regression by Randomforest. R news 2. pp. 18–22.
Ließ, M., Hitziger, M., Huwe, B., 2014. The sloping mire soil-landscape of southern ecuador: influence of predictor resolution and model tuning on random forest predictions. Appl. Environ. Soil Sci. 2014, 603132.
Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. PLoS One 11, 0153673.
Liu, W., Baret, F., Gu, X., Zhang, B., Tong, Q., Zheng, L., 2003. Evaluation of methods for soil surface moisture estimation from reflectance data. Int. J. Remote Sens. 24, 2069–2083.
Lu, Z., Han, D., 2017. Multi-source hydrological soil moisture state estimation using data fusion optimisation. Hydrol. Earth Syst. Sci. 21, 3267–3285.
Mallick, K., Bhattacharya, B.K., Patel, N.K., 2009. Estimating volumetric surface moisture content for cropped soils using a soil wetness index based on surface temperature and NDVI. Agric. For. Meteorol. 149, 1327–1342.
Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.
Mattern, S., Raouafi, W., Bogaert, P., Fasbender, D., Vanclooster, M., 2012. Bayesian data fusion (BDF) of monitoring data with a statistical groundwater contamination model to map groundwater quality at the regional scale. J. Water Resou. Protect. 4, 929–943.
McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the world reference base. Geoderma 155, 132–139.
Mohamed, E.S., Saleh, A.M., Belal, A.B., Gad, A.A., 2017. Application of near-infrared reflectance for quantitative assessment of soil properties. Egyptian J. Remote Sens. Space Sci. 21, 1–14.
Mohammadi, M., 1999. Correlation Study of Soils in East Zagros. Iranian Soil and Water Research Institute, pp. 1062.
Montanarella, L., Vargas, R., 2012. Global governance of soil resources as a necessary condition for sustainable development. Curr. Opin. Environ. Sustain. 4, 559–564.
Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping a review. Geoderma 162, 1–19.
Natsagdorj, E., Renchin, T., Kappas, M., Tseveen, B., Dari, C., Tsend, O., Duger, U.O., 2017. An integrated methodology for soil moisture analysis using multispectral data in mongolia. Geo-spatial Inform. Sci. 20, 46–55.
Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. Eur. J. Soil Sci. 62, 417–430.
Nield, S.J., Boettinger, J.L., Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using landsat ETM data. Soil Sci. Soc. Am. J. 71, 245–252.
Notarnicola, C., Posa, F., 2002. Extraction of soil parameters: two case studies using bayesian fusion of multiple sources data. In: Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International. IEEE, pp. 905–907.
Odeh, I.O.A., Leenaars, J., Hartemink, A., Amapu, I., 2012. The challenges of collating legacy data for digital mapping of nigerian soil. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. Taylor & Francis, London, pp. 453–458.
Peeters, L., Fasbender, D., Batelaan, O., Dassargues, A., 2010. Bayesian data fusion for water table interpolation: incorporating a hydrogeological conceptual model in kriging. Water Resour. Res. 46, W08532.
Pinheiro, É.F.M., Ceddia, M.B., Clingensmith, C.M., Grunwald, S., Vasques, G.M., 2017. Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central amazon. Remote Sens. 9, 293–315.
R Development Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rossiter, D.G., 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. In: Hartemink, A.E., McBratney, A.B., MendonÃa-Santos, M.D.L. (Eds.), Digital Soil Mapping With Limited Data. Springer, Dordrecht, pp. 69–80.
Rouse Jr., J., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the great plains with ERTS. In: NASA. Goddard Space Flight Center 3d ERTS-1 Symposium. Texas A&M University, College Station, TX, United States, pp. 309–317.
Sörensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol. Earth Syst. Sci. Discussions 10, 101–112.
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319–329.
Sulaeman, Y., Minasny, B., McBratney, A.B., Sarwani, M., Sutandi, A., 2013. Harmonizing legacy soil data for digital soil mapping in indonesia. Geoderma 192, 77–85.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in baneh region, iran. Geoderma 253 (245), 67–77.
Takagi, K., Lin, H.S., 2012. Changing controls of soil moisture spatial organization in the shale hills catchment. Geoderma 173–174, 289–302.
Temme, A.J.A.M., Heuvelink, G.B.M., Schoorl, J.M., Claessens, L., 2009. Geostatistical simulation and error propagation in geomorphometry. Dev. Soil Sci. 33, 121–140.
Witten, I.H., Eibe, F., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier.
Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environ. Model Softw. 57, 202–215.
Yang, R.M., Zhang, G.L., Liu, F., Lu, Y.Y., Yang, F., Yang, F., et al., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol. Indic. 60, 870–878.
Yu, H., Kong, B., Wang, G., Du, R., Qie, G., 2018. Prediction of soil properties using a hyperspectral remote sensing method. Arch. Agron. Soil Sci. 64, 546–559.
Zhan, Z.M., Qin, Q.M., Ghulan, A., Wang, D.D., 2007. NIR-red spectral space based new method for soil moisture monitoring. Sci. China Ser. D Earth Sci. 50, 283–289.
Zhang, H., 2004. The optimality of naive Bayes. In: Proceedings of the the 17th International FLAIRS conference (FLAIRS2004), pp. 562–567.