Catena 182 (2019) 104138

Contents lists available at ScienceDirect

Catena

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

Bayesian data fusion for combining maps of predicted soil classes: A case study using legacy soil profiles and DEM covariates 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

Keywords:
World Reference Base (WRB)
Multinomial Logistic Regression (MLR)
Bayesian Maximum Entropy (BME)
Bayesian data fusion (BDF)

ABSTRACT

This paper tackles the issue of spatially predicting soil classes by combining at best soil information coming directly from legacy soil profiles with soil information indirectly obtained from spatial covariates. Based on Multinomial Logistic Regression (MLR) and Bayesian Maximum Entropy (BME) models, we first show that prediction models easily lead to very different soil maps while having at the same time quite comparable global performances. By relying afterwards on a Bayesian data fusion (BDF) approach, we emphasize the benefit of combining the output of these two prediction models in order to get a single final map that combines the major spatial features of the MLR and BME maps while at the same time improving the quality of the predictions.

The advocated methodology is illustrated with the mapping of World Reference Base (WRB) soil classes over a 10,480 km^2 area located in Iran. A set of 390 soil profiles allowed us to assign the WRB soil classes at these locations. In parallel, a set of potentially related covariates were computed from a 90 m resolution digital elevation model. Using MLR and BME models, predictions were obtained separately at the nodes of a 90 m resolution grid. Even if the performances of the MLR and BME models compare well, it is shown that a BDF procedure that combines both results yields improved performances, with spatial features that are a balanced combination of those found separately on the MLR and BME maps.

These results emphasize the benefit of data fusion in order to improve the quality of the final map. Though the study was conducted here using MLR and BME models for predicting WRB soil classes, we believe this methodology and the corresponding findings are relevant when it comes to handle the results of spatial prediction models that are making use of distinct information sources in other soil science mapping contexts.

1. Introduction

As soils play important roles in a wide variety of environmental and social fields (land use planning and management, ecosystem services, etc.), reliable soil maps have been recognized as important for management and sustainable development goals (Keesstra et al., 2016; Ayoubi et al., 2018).

Soils are the result of complex interactions between pedological and geomorphological processes that are acting at distinct landscape scales and positions, thus leading to a wide variety of soil types with different degrees of development. Traditional soil survey methods (TSM) which have been used for a long time aim at separating soil classes based on these relationships between soil properties and observable environmental features by studying soil profiles and accounting for field observations. As they heavily rely on expert judgment which is complex and qualitative, they lead to soil maps that are suffering from a low to medium efficiency accuracy and a lack of clearly quantified uncertainty assessment (Boettinger et al., 2010; Kempen et al., 2012). Over the last decades, this has triggered the development of digital soil mapping methods (DSM, see McBratney et al., 2003) that are now acknowledged as the fastest and easiest way of creating high resolution maps of soil properties and classes, even if TSM still remain a valuable option and can even lead to higher map accuracy in some circumstances. Numerous statistical, geostatistical, hybrid and machine learning methods have been used for DSM. The general principles and applications can be found in McBratney et al. (2003), Brungard (2009), and Zhang et al. (2017), and most popular methods over the last two decades are briefly reviewed by Zhang et al. (2017).

DSM mostly focus on the soil-landscape modeling, which quantifies the relationships between soil properties and environmental variables (Scull et al., 2003), with the idea of directly relating the soil properties at a specific location to the relevant landscape features and environmental covariates at

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

https://doi.org/10.1016/j.catena.2019.104138
Received 12 March 2019; Received in revised form 31 May 2019; Accepted 19 June 2019
0341-8162/

the same location. Soil maps obtained that way imply a set of rules about the soil and environmental covariates relationships that can be inferred from high-quality data on environmental variables and data analysis techniques (Smith et al., 2010; Yang et al., 2011). Most DSM researches used terrain attributes as the main predictor variables (McBratney et al., 2003) through the use of various indices computed both from digital elevation models (DEM's) and remotely sensed (RS) images (e.g., elevation, slope position, convexity, wetness indices as computed from DEM's, along with various spectral reflectances indices as obtained from RS images). Typically, soil properties are predicted using regression-like models from these ancillary information sources using a large panel of linear and non-linear models.

There is a fundamental difference between predicting soil properties and predicting soil classes. Typically, DSM methods have mainly focused on the prediction of a wide variety of soil properties that are viewed as spatial continuous variables. Among the most commonly predicted properties are soil texture, soil organic content, pH, horizons thickness and moisture content (see e.g. Were and Bui, 2015; Song et al., 2016b; Liu et al., 2012; Kumar et al., 2012; Brus et al., 2016; Odgers et al., 2015; Arrouays et al., 2014; Ließ et al., 2012). DSM techniques have been however much less frequently applied when it comes to directly predict soil classes, that are by essence spatial categorical variables. Even if soil classes are obviously defined from a finite set of soil properties (those properties largely depending on the classification system one is using), spatially predicting all soil properties that are needed for the classification step is a tedious (if not impossible) task, except for classifications that are based on a very limited set of properties (e.g. the USDA classification system based on grain size). As a consequence, one thus need to be able to directly predict these soil classes from covariates at hand. Kempen et al. (2009) used a multinomial logistic regression (MLR) approach for updating the Dutch soil map using legacy soil data. Jafari et al. (2012) and Afshar et al. (2018) have shown how soil great groups can be predicted using binomial and multinomial logistic regression from relief attributes, remote sensing indices and a geomorphology map. Jeune et al. (2018) compared MLR and random forest (RF) classifiers for mapping soil classes in Haiti. Zeraatpisheh et al. (2017) compared MLR and RF with conventional soil mapping to predict soil types in a semi-arid region in Iran. More generally, Brungard et al. (2015) assessed the performance of various machine learning approaches for mapping soil classes in three semi-arid landscapes, while Heung et al. (2016) provide an overview and comparison of machine-learning techniques for classification purposes.

Besides the use of DSM relying on a model for predicting soil classes from environmental covariates, another approach is the interpolation of soil classes based on spatial statistics models. This problem has been handled by the use of indicator kriging (Bierkens and Burrough, 1993; Carle and Fogg, 1996), Markov chain simulation models (Li et al., 2004; Wu et al., 2004; Park et al., 2007), regression kriging on memberships (Hengl et al., 2007) and Bayesian maximum entropy (BME) methods (Bogaert, 2002; D'Or and Bogaert, 2004). For DSM, these methods require a sufficient soil sampling density so that reliable and meaningful interpolations can be conducted, which is probably the main limitation of this approach.

As TSM is time-consuming and costly, it is unlikely that numerous soil profiles are available over large areas and the opportunities of carrying out new traditional soil surveys are low (Cambule et al., 2015). Legacy soil profiles are thus valuable and increasing attention has been paid to the use the legacy soil data in DSM without the need of additional expensive and time consuming field works (Rossiter et al., 2008; Odeh et al., 2012; Sulaeman et al., 2013). Legacy soil profiles are needed both for calibrating models and for assessing the performances of these models. Additionally, they convey information about the spatial organization of the soil classes and this information is not explicitly accounted for by DSM methods that directly link soil classes to collocated environmental covariates.

In this paper, we compare the performance of two DSM approaches for predicting soil classes from a set of legacy soil profiles and a DEM. A MLR approach will first be used to predict soil classes from collocated terrain attributes. A BME approach will then be used in order to directly interpolate soil classes from the same set of sampled locations without relying on these covariates. Both approaches will be compared and it will be shown how the corresponding results can be fused using a Bayesian data fusion (BDF, see Bogaert and Fasbender, 2007) approach, with the aim of obtaining a single final map that incorporates the features of both MLR and BME maps and that improves the quality of the predictions. The whole methodology will be illustrated using a legacy survey data set in Iran that consists of 390 soil profiles classified according to the World Reference Base (WRB), along with environmental covariates as obtained from a 90 m resolution DEM.

2. Material and methods

2.1. Study area

The study area is located in the East Zagros region of Iran between 31° 45′ to 33° 15′ North and 50° 0′ to 51° 30′ East (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. As a result, the study area includes different landform types, i.e., mountains, hills, alluvial gravelly fans, alluvial 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 (Mohammadi, 1999).

2.2. Legacy soil data

Based on ten legacy soil surveys that are available at the 1:50, 000 scale in the study area, 390 well documented legacy soil profiles were identified after applying rescue and renewal processes (not described here for the sake of brevity) based on the recommendations by Rossiter et al. (2008). Fig. 1 shows the location of these profiles, with different sampling densities between the North and South parts of the study area. These soil profiles are classified by the first author according to the World Reference Base (WRB) using the soil reference groups (RSG's), i.e. the first level of classification from the WRB (WRB, 2015).

Five soil RSG classes, i.e. Calcisols, Cambisols, Gleysols, Luvisols, and Regosols were identified over the study area (Fig. 2). As calcareous materials are the most common ones in this region, Calcisols (soils with accumulation of Calcium carbonate) are the most numerous and are mostly found in plains (48% of the profiles). Cambisols (moderately developed soils, with at least an incipient subsurface soil formation) are generally found on fans and higher parts of the plains (18% of the profiles). Gleysols (groundwater-affected soils) are formed in lowlands impacted by high groundwater tables (6% of the profiles). Luvisols (soils with a clay-enriched subsoil, high activity clays and high base status) are generally formed on alluvial plains (10% of the profiles). Regosols (soils with no significant profile development) are typically found in highlands and steep areas, i.e. fans and mountains (18% of the profiles).

It is worth noting that the RSG level of the WRB classification is intended for broad geographic groupings based on pedogenesis, while for national maps at least one major qualifier is recommended. Table 1 gives the number of soil profiles attached with the principal qualifiers for each of the five RSG classes. Most of the soil profiles for the Calcisols, Cambisols and Regosols classes are largely associated with a single qualifier. This is however not true for Gleysols and Luvisols. For all RSG classes other than Calcisols, the dominant qualifiers are Calcic and Calcaric. Since only 390 soil profiles are at hand, considering the 19 classes at the major qualifier level is impossible here as there would be a very limited number of soil profiles in most of them. Using only the

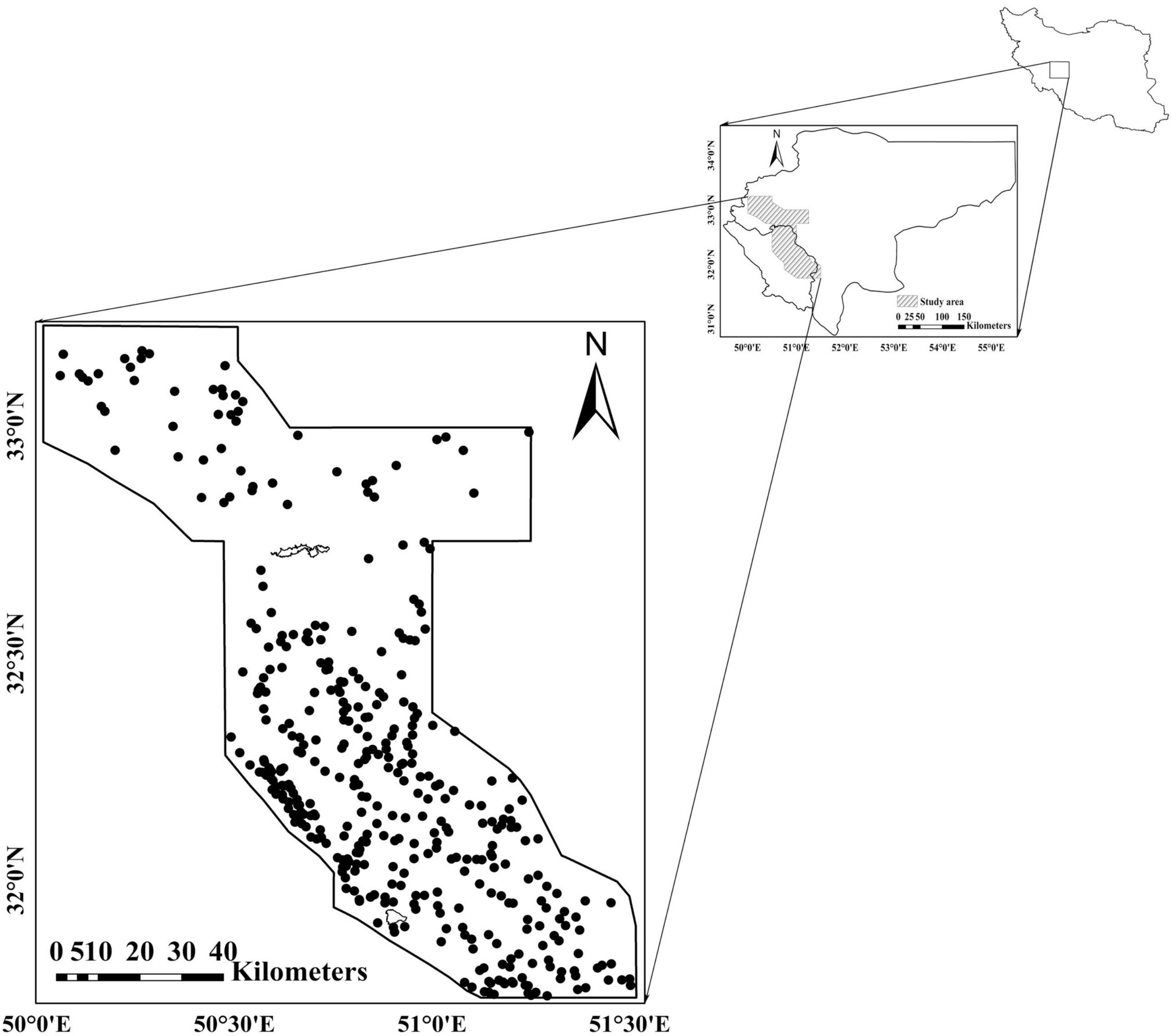

Fig. 1. Location of the study area in Iran, in the neighboring border of Isfahan (top right region of the study area) and Chaharmahal-va-Bakhtiari (down left region of the study area) provinces. Black dots correspond to the locations of the 390 sampled soil profiles from the legacy data sets.

five RSG classes is however expected to cause some trouble for some classes. For example, Calcaric Cambisols are transitional to Calcisols and one can expect that some confusion could occur when predicting these classes.

2.3. Ancillary spatial data

A DEM was compiled from the four DEM's scenes that were available at a 90 m spatial resolution over the study area (Jarvis et al., 2008) and was projected in Universal Transverse Mercator (UTM) coordinates (WGS 1984, UTM 39N). Nine covariates were computed from it and from its first and second-order derivatives, namely Elevation (EL), Slope (SL), Midslope position (MLS), Slope length and steepness factor (LSF), Multiresolution valley bottom flatness index (MRVBF), Valley depth (VD), Vertical distance to channel network (VDCHN), Topographic wetness index (TWI) and convexity (CON) (Olaya, 2004). The values of these covariates were computed at the same nodes as those of the composite DEM, thus yielding corresponding maps at a 90 m spatial resolution. Additionally, their values were also extracted at the coordinates of the 390 legacy soil profiles.

2.4. Multinomial logistic regression

Logistic regression (LR) belongs to the generalized linear models family (Lane, 2002) and handles the case where the response (dependent) variable is categorical. In a LR model, the log-odds of the probability of each category is assumed to be a linear regression model from a set of n explanatory variables $x_1,\ldots,x_n$. The simple binary logistic model (BLR) is used to estimate the probability π of a binary variable (i.e. two categories) with

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j \tag{1}$$

where $\pi/(1-\pi)$ is the odd-ratio, β_0 is the intercept and the $\beta_j's$ are the slopes associated with the various x_j variables. This BLR model is easily extended to an arbitrary number of m categories, yielding the multinomial logistic regression (MLR) model, with

$$\log\left(\frac{\pi_i}{\pi_m}\right) = \beta_{i0} + \sum_{j=1}^{n} \beta_{ij} x_j \qquad i = 1, \ldots, m-1 \tag{2}$$

where the π_i's are the probabilities for the m categories, with the mth

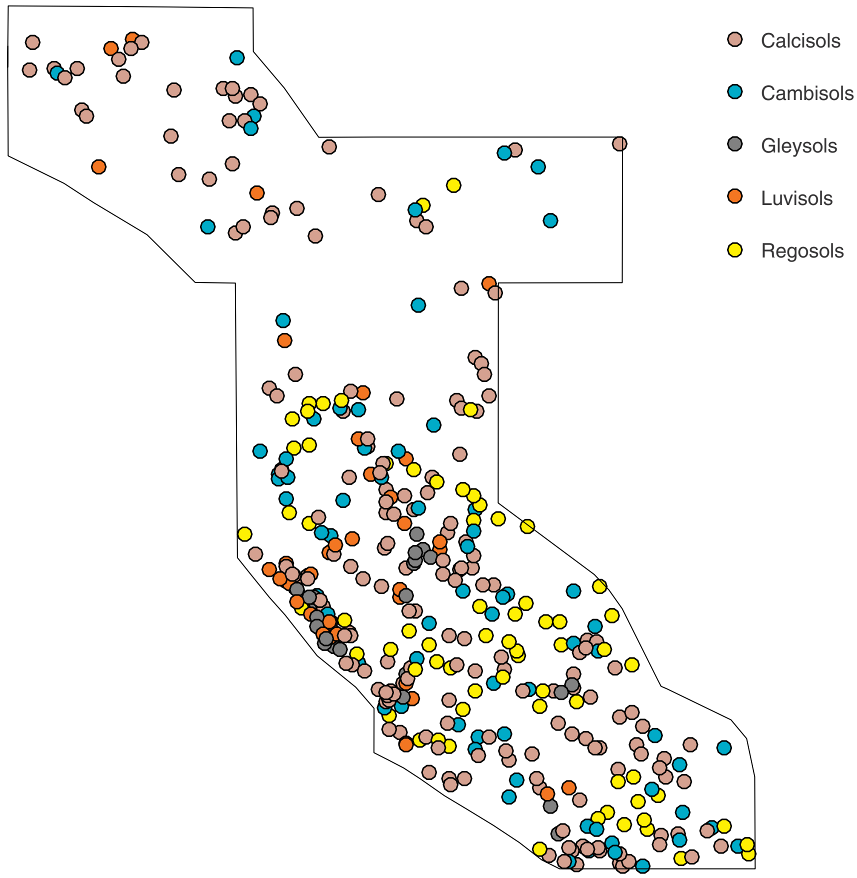

Fig. 2. Spatial distribution of the five observed soil classes (WRB classification) for the 390 legacy soil profiles.

category used here as the reference category without loss of generality. Though we will not detail the corresponding notations here for the sake of brevity, these BLR and MLR models can also make use of squared terms x_i^2 and interactions terms $x_i x_j$ if relevant.

LR models have been frequently used for modeling and predicting soil categorical properties (Lane, 2002; McBratney et al., 2000). E.g., they have been successfully used to (i) predict soil drainage classes (Campling et al., 2002; Giasson et al., 2006), (ii) model the relationship between soil types and terrain attributes (Giasson et al., 2006), and (iii) predict soil taxonomic (both USDA and WRB) classes (Hengl et al., 2007; Debella-Gilo and Etzelmüller, 2009; Jafari et al., 2012; Afshar et al., 2018). Jeune et al. (2018) and Zeraatpisheh et al. (2017) have shown that MLR stand comparison with more complex models like Random Forests for classifying soils and mapping soil classes. More specifically in Iran, MLR was successfully used by Jafari et al. (2012), Zeraatpisheh et al. (2017), Jafari et al. (2013), and Afshar et al. (2018) for mapping soil classes.

As DSM can involve a large set of potentially useful covariates, feature subset selection is an important step that aims at identifying the best subset of covariates that achieve the highest possible predictive power for the model (Blanco et al., 2018; Ließ et al., 2016; Kuhn and Johnson, 2013). BLR models were used as a screening method to identify the most informative covariates extracted from the DEM. Using in turn each of the nine potential covariates, BLR models were thus first estimated from the 390 legacy soil profiles where soil classes are known. The Wald statistic was used to evaluate the significance of the regression coefficient of each predictor for each soil class (Hosmer and Hjort, 2002). Different MLR models were built afterwards to select the best subset of covariates. Calcisols (the most frequently observed soil class) was selected as the reference class. The performance of these MLR models were evaluated by using different criteria, namely, (i) the Akaike Information criterion (AIC) along with log-likelihood ratio testing to compare MLR models (Hosmer and Hjort, 2002), (ii) the McFadden R^2 that measures the reduction in maximized log-likelihood and which is mathematically close to the classical R^2 for least squares (Menard, 2000), and (iii) a 10-fold cross-validation to evaluate classification accuracy (Kuhn et al., 2008).

Table 1
List of the five observed RSG classes and major qualifiers classes, along with the corresponding number of soil profiles in each class.

Calcisols	188
Hyperskeletic	11
Hyperskeletic Leptic	2
Hyperskeletic Leptic Petric	2
Hyperskeletic Petric	4
Luvic	4
Petric	3
Skeletic	9
Haplic	153
Cambisols	**69**
Skeletic Calcaric	3
Calcaric	66
Gleysols	**24**
Calcaric	6
Calcic	18
Luvisols	**40**
Calcic Gleyic	4
Haplic	13
Calcic	23
Regosols	**69**
Calcaric Leptic	2
Calcaric Skeletic	1
Skeletic	6
Calcaric	60

2.5. Bayesian maximum entropy

To the opposite of MLR that aims at building first a model for predicting the probabilities of the various soil classes from a set of collocated proxies and then spatializes afterwards this model at the nodes of a prediction grid, we will focus here on a prediction of these probabilities that directly makes use of the spatial information brought from the soil samples only. It is indeed expected that soil classes exhibit some spatial organization, so that identical soil classes are more likely than different ones for locations that are close to each other. Under a stationarity assumption, this effect can be assessed using the set of $m \times m$ bivariate probability functions

$$\pi_{i,j}(\mathbf{h}) = P(C(\mathbf{x}) = c_i \cap C(\mathbf{x}+\mathbf{h}) = c_j) \qquad \forall\ i, j = 1, \ldots, m \tag{3}$$

that specify the probability that two soil classes c_i and c_j will be jointly observed at a distance $\mathbf{h}$ from each other, where $\mathbf{x}$ refers here to an arbitrary location over the study area. On the other side, if one aims at predicting the probabilities of the various soil classes at a specific location $\mathbf{x}_0$ based on a set of observed classes at neighboring locations $\mathbf{x}_1$, …,$\mathbf{x}_n$, this would require the joint probability table

$$\pi_{j_0,\ldots,j_n}(\mathbf{x}_0, \ldots, \mathbf{x}_n) = P\left(\bigcap_{i=0}^{n} C(\mathbf{x}_i) = c_{j_i}\right) \qquad \begin{array}{l}\forall\ i = 1, \ldots, n\\ \forall\ j_i = 1, \ldots, m\end{array} \tag{4}$$

where $j_i \in \{1,\ldots,m\}$ is the soil class at location $\mathbf{x}_i$, so that the conditional probabilities based on a set of observed classes at locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ is given by

$$\pi_{j_0}(\mathbf{x}_0|\{j_1, \ldots, j_n\}) = \frac{\pi_{j_0,\ldots,j_n}(\mathbf{x}_0,\ldots,\mathbf{x}_n)}{\sum_{j_0}\pi_{j_0,\ldots,j_n}(\mathbf{x}_0,\ldots,\mathbf{x}_n)} \qquad \forall\ j_0 = 1, \ldots, m \tag{5}$$

Although joint probabilities cannot be directly inferred from data at hand, it turns out that they can be estimated from the maximum entropy principle. For doing this, let us remark first that any bivariate probability $\pi_{j_{i'},j_{i''}}(\mathbf{x}_{i'}, \mathbf{x}_{i''})$ is directly obtained from the corresponding bivariate probability function $\pi_{j_{i'},j_{i''}}(\mathbf{x}_{i'} - \mathbf{x}_{i''})$ as defined in Eq. (3). Moreover, bivariate probabilities are also marginal probabilities from the joint probability table, with

$$\pi_{j_{i'},j_{i''}}(\mathbf{x}_{i'} - \mathbf{x}_{i''}) = \sum_{i \notin \{i', i''\}} \pi_{j_0,\ldots,j_n}(\mathbf{x}_0, \ldots, \mathbf{x}_n) \tag{6}$$

As a consequence, knowing the set of bivariate probability functions provides fragmentary information about the joint probability table. This table can be estimated from the maximum entropy principle by maximizing its Shannon's entropy $H(\pi_{j_0,\ldots,j_n}(.))$ subject to the constraints imposed on its set of bivariate probabilities as computed from Eq. (6), where

$$H(\pi_{j_0,\ldots,j_n}(.)) = -\sum_{j_0}\cdots\sum_{j_n}\pi_{j_0,\ldots,j_n}(.)\ln\pi_{j_0,\ldots,j_n}(.) \tag{7}$$

A detailed presentation of the maximum entropy method is not possible here and the reader is referred to Bogaert (2002) for a complete description of the theory.

The BME methodology for categorical variables has been successfully applied both in soil sciences (D'Or and Bogaert, 2004; Brus et al., 2008) and in other fields (Gengler and Bogaert, 2016b; Bogaert and Gengler, 2017). It is sufficient to remember here that one needs to first estimate the set of bivariate probability functions in Eq. (3) for optimally building the joint probability table in Eq. (4), that is needed afterwards for computing the conditional probabilities from Eq. (5). In practice, the method is computer intensive as it requires building at each prediction location a multivariate probability table of dimension m^{n+1}, where m is the number of soil classes and n is the number of soil samples accounted for in the spatial neighborhood of $\mathbf{x}_0$. For the $m = 5$ observed soil classes and using $n = 5$ soil samples in the prediction neighborhood for computational efficiency, this leads to a table having $5^6 = 15{,}625$ cells at each prediction location.

2.6. Bayesian data fusion

The general objective of data fusion is to combine various and possibly conflicting sources of information in order to obtain a single and harmonized final result. As such, the concept of data fusion encompasses various methods and approaches that are covering a wide field of applications (Haghighat et al., 2016; Castanedo, 2013). In the framework of soil sciences, numerous studies have highlighted the value of data fusion for mapping soil properties (see, e.g., Lu and Han, 2017; Grunwald et al., 2015). We will focus here more specifically on a statistical approach linked to Bayesian decision theory (Fassinut-Mombot and Choquel, 2004) known as Bayesian data fusion(BDF), that belongs to the set of high-level fusion methods (Castanedo, 2013; Das, 2008). This BDF method proved to be helpful in soil sciences (Brus et al., 2008) and in environmental studies (Gengler and Bogaert, 2016a; Mattern et al., 2012; Notarnicola and Posa, 2002). The method was originally advocated in a spatial context by Bogaert and Fasbender (2007). We will only use here a simplified version of BDF that is restricted to the fusion of two collocated information sources and that simplifies to the use of a naive Bayes fusion rule in that situation. The method was applied on a very similar soil mapping context by Rasaei and Bogaert (2019).

Let us remind that we have a set of categories $c_1,\ldots,c_m$ that correspond to the various possible soil classes at an arbitrary locations $\mathbf{x}_0$ where the soil class need to be predicted. Let us denote $\pi_i(\mathbf{x}_0|I_a) \equiv P(C(\mathbf{x}_0) = c_i|I_a)$ and $\pi_i(\mathbf{x}_0|I_b) \equiv P(C(\mathbf{x}_0) = c_i|I_b)$ as two models at hand for estimating the probability of observing category c_i at location $\mathbf{x}_0$, where I_a and I_b refer to the corresponding information used by these two models. What is sought for are the probabilities $\pi_i(\mathbf{x}_0|\{I_a, I_b\}) \equiv P(C(\mathbf{x}_0) = c_i|I_a \cap I_b)$ that rely on the joint use of both information sources. From Bayes theorem, one can write that

$$P(C(\mathbf{x}_0) = c_i|I_a \cap I_b) \propto P(I_a \cap I_b|C(\mathbf{x}_0) = c_i)\cdot P(C(\mathbf{x}_0) = c_i) \qquad \forall\ i = 1, \ldots, m \tag{8}$$

Assuming that $I_a \perp I_b|C(\mathbf{x}_0) = c_i$, i.e. the two information sources are independent conditionally to the true but unknown category $C(\mathbf{x}_0)$, and applying again Bayes theorem, one obtains the result

$$\begin{aligned} P(C(\mathbf{x}_0) = c_i|I_a \cap I_b) &\propto P(I_a|C(\mathbf{x}_0) = c_i)\cdot P(I_b|C(\mathbf{x}_0) = c_i)\cdot P(C(\mathbf{x}_0) = c_i)\\ &\propto \frac{P(C(\mathbf{x}_0) = c_i|I_a)\cdot P(C(\mathbf{x}_0) = c_i|I_b)}{P(C(\mathbf{x}_0) = c_i)} \end{aligned} \tag{9}$$

or, stated in a much more compact way,

$$\pi_i(\mathbf{x}_0|\{I_a, I_b\}) \propto \frac{\pi_i(\mathbf{x}_0|I_a)\cdot\pi_i(\mathbf{x}_0|I_b)}{\pi_i} \qquad \forall\ i = 1, \ldots, m \tag{10}$$

where $\pi_i \equiv P(C(\mathbf{x}_0) = c_i)$ is the marginal probability of the class. This result corresponds to the so-called naive Bayes fusion rule, where the π_i's play the role of the prior probabilities when no specific spatial information is at hand in the neighborhood, i.e., when the only information at hand is the marginal probabilities π_i's of these classes (that are estimated as the observed frequencies of the classes in the data set).

Despite its apparent simplicity and the use of a conditional independence hypothesis that might be difficult to assess, the naive Bayes fusion rule proved to be robust (see, e.g., Zhang, 2004) and is widely used in the field of big data and machine learning (Witten et al., 2016).

3. Results and discussion

3.1. Multinomial logistic regression

Each of the nine covariates were used separately for building BLR models for the five soil classes. Results from Table 2 show that the performances of these models largely differ for the various soil classes and for the various covariates at hand. E.g., for Cambisols, only two covariates (MRVBF and VD) have a p-value lower than 0.01, while for Regosols all covariates are (most of the time largely) below that

Table 2
BLR models for each soil class using each covariate separately, with the corresponding *p*-values. For each soil class, the most significant covariate is underlined (SL: Slope, MSL: Midslope position, LSF: Slope length and steepness factor, MRVBF: Multiresolution valley bottom flatness index, VD: Valley depth, VDCHN: Vertical distance to channel network, TWI: Topographic wetness index, CON: Convexity, EL: Elevation).

Soil class	Covariates								
	SL	MSL	LSF	MRVBF	VD	VDCHN	TWI	CON	EL
Calcisols	$5.2e^{-07}$	$9.9e^{-05}$	$6.2e^{-06}$	$1.1e^{-03}$	$2.3e^{-01}$	$2.8e^{-06}$	$2.6e^{-03}$	$6.4e^{-02}$	$7.0e^{-03}$
Cambisols	$3.1e^{-01}$	$1.9e^{-01}$	$7.5e^{-01}$	$1.5e^{-03}$	$3.6e^{-03}$	$9.7e^{-01}$	$9.8e^{-02}$	$6.6e^{-02}$	$9.2e^{-02}$
Gleysols	$4.9e^{-03}$	$1.7e^{-02}$	$9.0e^{-04}$	$2.2e^{-07}$	$7.4e^{-08}$	$5.3e^{-02}$	$4.0e^{-01}$	$6.9e^{-05}$	$1.5e^{-05}$
Luvisols	$2.2e^{-03}$	$2.5e^{-02}$	$3.6e^{-03}$	$2.0e^{-04}$	$1.5e^{-05}$	$5.5e^{-02}$	$6.3e^{-01}$	$7.0e^{-04}$	$8.0e^{-04}$
Regosols	$3.9e^{-15}$	$4.6e^{-12}$	$1.9e^{-13}$	$1.2e^{-11}$	$2.4e^{-09}$	$2.2e^{-16}$	$9.2e^{-03}$	$5.5e^{-09}$	$3.2e^{-12}$

threshold. Comparing now between covariates, none of them are relevant for all soil classes, and the number of predictable soil classes from each covariate (with *p*-values below 0.01) is ranging from two (for TWI and VDHCN) up to five (for MRVBF). However, it can be seen that SL and VD seem to be the most relevant covariates, except for Cambisols. Corresponding maps are given in Fig. 3.

From these first findings, all MLR models based on pairs of covariates where estimated and compared. Table 3 confirms that the best model that maximizes R^2 and minimizes AIC is obtained with SL and VD as covariates (all models in Table 3 involve the same number of $k = 12$ parameters). Clearly too, the performance of these models remains limited, with the highest R^2 equal to 0.2 and the highest accuracy equal to 0.56 (to be compared to the lowest accuracy equal to 0.48 for the worst model). Using additional covariates did not improve significantly the model (details not shown). Testing for squared terms SL^2, VD^2 and interaction effect SL*VD shows that only VD^2 is statistically significant (p-value $= 4.0e^{-5}$ for the log-likelihood ratio test). Our final model MLR model involving $k = 16$ parameters is thus

$$\log\left(\frac{\pi_i}{\pi_5}\right) = \beta_{i0} + \beta_{i1}{*}SL + \beta_{i2}{*}VD + \beta_{i3}{*}VD^2 \qquad i = 1, \ldots, 4 \tag{11}$$

(where π_5 refers to the Calcisols reference soil class) with associated values $R^2 = 0.248$ and AIC = 854 (to be compared with the fourth line in Table 3). The confusion matrix of the predicted soil classes as estimated from cross-validation at the 390 sampled locations is given in Table 4a (the predicted soil class is the maximum probability class). The accuracy of the model is equal to 57% and it can be seen that this model is unable to predict the Cambisols and Luvisols classes, while the model is overestimating the Calcisols class (76% of the predicted classes). The corresponding soil map obtained by applying the model at the nodes of the prediction grid is given in Fig. 5a (the predicted soil class is the maximum probability class). This map confirms that whatever the values for SL and VD over the area, the MLR model will never predict Cambisols and Luvisols classes. The spatial patterns of this map are of course related to those observed for SL and VD (see Fig. 3).

3.2. Bayesian maximum entropy

From the 390 sampled soil profiles at hand, the bivariate probability functions $\pi_{i,j}(\mathbf{h})$ were estimated for the five observed soil classes. As seen from Fig. 4, soil classes exhibit a spatial dependence up to 5–10 km depending on the pair of soil classes. For each prediction location on the grid, BME conditional probabilities were computed by accounting for the 5 closest soil profiles. Computations were done using Matlab and the BMElib2.0 library (see https://mserre.sph.unc.edu/BMElib_web/BMELIB.htm) on a Windows 10 computer with Intel Core I7-6700 K processors at 4 GHz. The computing time was about 32 h for processing the 1,267,420 nodes of the prediction grid.

As a consequence of the limited range of the spatial dependence between soil classes, BME predictions are likely to be useful only in places where the spatial density of the soil profiles is sufficient enough for ensuring that some soil samples will be within the neighborhood of the prediction location at distances that do not exceed this spatial range. From Fig. 2, this will not be the case for a large extent of the

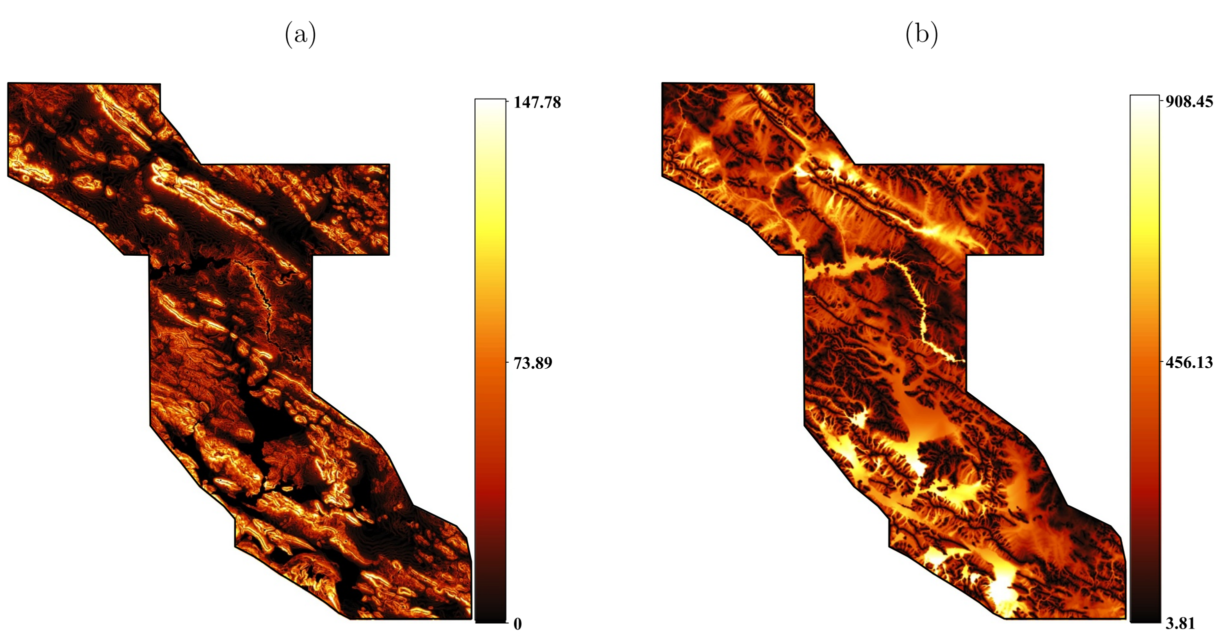

Fig. 3. Maps of (a) Slope and (b) Valley depth used in the MLR model for predicting soil classes over the study area.

Table 3
MLR models for each possible pair of covariates, with the corresponding McFadden R^2, Accuracy and AIC values. Underlined values corresponds to the three best values (i.e., highest R^2, highest Accuracy and lowest AIC).

Covariates	R^2	Accuracy	AIC
SL + MSL	0.192	0.554	887
SL + LSF	0.193	0.556	884
SL + MRVBF	0.192	0.559	884
SL + VD	0.201	0.562	872
SL + VDHCN	0.191	0.551	886
SL + TWI	0.193	0.556	891
SL + CON	0.194	0.567	1067
SL + EL	0.192	0.554	885
MSL + LSF	0.006	0.480	1070
MSL + MRVBF	0.003	0.482	1063
MSL + VD	0.014	0.523	979
MSL + VDHCN	0.006	0.482	1070
MSL + TWI	0.005	0.482	1080
MSL + CON	0.013	0.482	1064
MSL + EL	0.012	0.482	1060
LSF + MRVBF	0.006	0.482	1063
LSF + VD	0.017	0.536	968
LSF + VDHCN	0.007	0.487	1064
LSF + TWI	0.008	0.477	1070
LSF + CON	0.016	0.474	1058
LSF + EL	0.016	0.477	1054
MRVBF + VD	0.013	0.533	962
MRVBF + VDHCN	0.005	0.480	1058
MRVBF + TWI	0.004	0.482	1064
MRVBF + CON	0.012	0.485	1053
MRVBF + EL	0.010	0.482	1048
VD + VDHCN	0.017	0.533	971
VD + TWI	0.016	0.513	981
VD + CON	0.022	0.528	962
VD + EL	0.051	0.554	885
VDHCN + TWI	0.007	0.480	1070
VDHCN + CON	0.018	0.482	1054
VDHCN + EL	0.020	0.482	1057
TWI + CON	0.013	0.482	1067
TWI + EL	0.009	0.482	1062
CON + EL	0.007	0.482	1050

northern part of the study area, where few soil profiles are separated by large distances. In these areas, BME predicted probabilities will tend to the marginal probabilities for the various soil classes, i.e., $\pi_{j_0}(\mathbf{x}_0|\{j_1, \ldots, j_n\}) = \pi_{j_0}\ \forall\ j_0 \in \{1, \ldots, m\}$. As Calcisols is the most frequent class, BME predictions will thus mainly lead to a map where Calcisols is largely dominant if one selects the most probable soil class as the predicted one. This effect is clearly visible in Fig. 5b, where most of the northern part is classified as Calcisols, with few small patches of other predicted soil classes when prediction locations are very close to sampled soils that belong to these other classes. For the southern part of the area, the situation is however different thanks to the higher spatial density of the soil profiles. From a comparison between Figs. 2 and 5b, one can see that indeed BME yields a spatial segmentation into soil classes that is in good agreement with the spatial organization of the observed soil samples. To the opposite of MLR that was unable to predict Cambisols and Luvisols classes as the most probable ones whatever the place, BME is accounting for what is observed at neighboring locations and thus leads to improved prediction for these classes.

In order to assess the performance of BME, a leave-one-out cross-validation (LOOCV) was conducted at the 390 samples locations and the corresponding confusion matrix is presented in Table 4b (where the predicted class is the class having the highest predicted probability of occurrence). Overall, the global performance of BME is on par with MLR, with a total accuracy of 54% as seen from Table 4b. While total accuracy was equal to 57% for MLR, this higher value must be tempered by the inability of MLR to predict both Cambisols and Luvisols classes, as mentioned earlier. Should the number of sampled soils be increased, this would increase the reliability of the estimation for the MLR model, but it would not dramatically change its overall performance. On the opposite, increasing the sample size would clearly improve the performance of BME, as this would be translated into a higher sampling density and thus into more reliable spatial predictions.

3.3. Bayesian data fusion

At this stage, we now have at hand two methods for estimating the probabilities of the various soil classes over the whole extent of the study area. For a given arbitrary location $\mathbf{x}_0$, the first set of probabilities $\pi_i(\mathbf{x}_0|I_a)$ (with $i = 1,\ldots,m$) corresponds to the results of the MLR model applied at that location, where I_a is the information conveyed by the collocated covariates at $\mathbf{x}_0$. The second set of probabilities $\pi_i(\mathbf{x}_0|I_b)$ corresponds to the BME prediction that makes use of the available sampled soils in a spatial neighborhood around $\mathbf{x}_0$. Both sets of probabilities are thus making use of distinct information and thus led to very different maps. Using the BDF framework that was presented in Section 2.6, the idea is now to conciliate using Eq. (9) these probabilities in order to obtain a more reliable set of probabilities $\pi_i(\mathbf{x}_0|\{I_a,I_b\})$.

In order to illustrate the way BDF works for our data, Fig. 6 presents few cases that can be encountered (among many other possible ones). Typically, Fig. 6a corresponds to the case where both MLR and BME agree about the fact that Calcisols is the most probable soil class. Accordingly, the fused probabilities will confirm this choice, with a Calcisols class that has now a higher probability than for MLR and BME. In Fig. 6b, BME leads to predicted probabilities that are very close to the marginal ones, i.e. when the spatial neighborhood of the prediction

Table 4
Confusion matrices for (a) MLR, (b) BME and (c) BDF based on the 390 samples where soil classes are known (labels for the lines are in the same order as for the columns). The predicted soil class corresponds to the soil class having the highest predicted probability. Corresponding total accuracy is equal to 57%, 54% and 61% respectively.

	(a) Predicted					(b) Predicted					(c) Predicted				
	Ca	Cb	Gl	Lu	Re	Ca	Cb	Gl	Lu	Re	Ca	Cb	Gl	Lu	Re
Observed	174	0	4	0	10	155	12	3	2	16	151	16	8	4	9
	49	0	3	0	17	51	6	0	1	11	42	12	0	1	14
	18	0	5	0	1	5	1	14	4	0	5	0	16	3	0
	34	0	4	0	2	26	0	2	8	4	23	1	7	8	1
	23	0	2	0	44	39	1	0	1	28	15	5	0	0	49

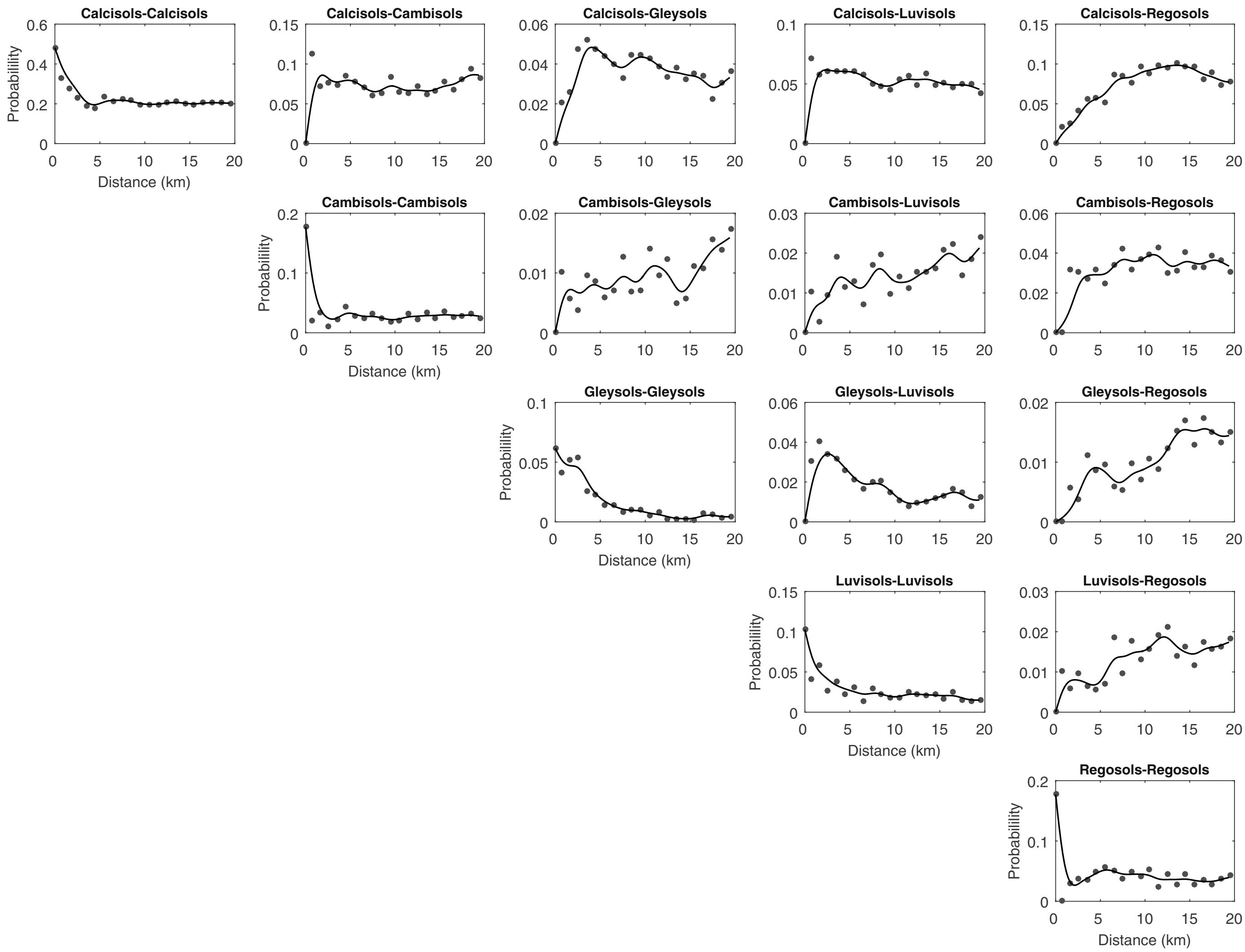

Fig. 4. Estimated (dots) and smoothed (line) bivariate probabilities for the various pairs of soil classes as a function of the distance between locations.

location does not contain valuable information, either due to a limited number of samples or due to the fact that these samples are far away. We thus have $\pi_i(\mathbf{x}_0|I_b) \simeq \pi_i$ in Eq. (9) and so $\pi_i(\mathbf{x}_0|\{I_a,I_b\}) \simeq \pi_i(\mathbf{x}_0|I_a)$, i.e. fused probabilities will correspond to those for MLR. To the opposite, Fig. 6c illustrate the case where both MLR and BME are informative but are in clear disagreement, as MLR is favouring Calcisols while BME is strongly favouring Regosols. As a consequence of these conflicting results, the fused probabilities are balancing the probabilities of these two classes that are now close to each other, thus translating the idea that the uncertainty about the correct class has increased.

A comparison of the predicted soil classes (i.e. maps of the classes having the maximum conditional probabilities) clearly shows how BDF is combining results from MLR and BME. In the northern part of the study area, where soil samples are scarce and where accordingly BME is providing poor predictions, one can see that the fused map in Fig. 5c is essentially identical to the MLR map of Fig. 5a. The situation is largely different in the southern part, where most of the soil samples are found. In the areas where samples are concentrated, the fused map does differ significantly from the MLR map. Patches of Regosols are clearly identified and the fused map shows the occurrence of Cambisols and Luvisols and that could not be predicted at all from the MLR model. It can be seen from Fig. 5c that the fused map is merging the features of MLR and BME maps, and there is some evidence that this fused map is globally a better overall picture of the true soil classes in the study area than those provided either from MLR or BME. In order to quantitatively assess the benefit of BDF, Table 4c presents the confusion matrix for the predicted soil classes from BDF at the 390 sampled locations, where the predicted class is the class having the highest fused probability of occurrence. While total accuracy (i.e. the frequency of diagonal counts) was equal to 57% and 54% for MLR and BME, respectively, it reaches 61% for the fused probability. As one might wonder if this increase is statistically significant, one can also check that BDF leads indeed to more assertive choices about the most probable soil class. Fig. 7 gives the distribution of these most probable classes at the 390 sampled locations for MLR, BME and BDF. From a comparison of these figures, it can be seen that this maximum probability is increased for the fused data as well, with average maximum probabilities $\bar{p}_{\max}$ equal to 0.58, 0.56 and 0.64 for MLR, BME and BDF, respectively. Though this increase does not appear to be dramatic here, this is also explained by the fact that both MLR and BME have limited prediction performances. Accordingly, for our specific case study, there is thus little hope to dramatically increase the performance of the final map by fusing MLR and BME results, but this limitation is not linked to the efficiency of the fusion procedure itself.

In order to compare the uncertainty attached to the MLR, BME and BDF maps, one can compute the entropy of the predicted distributions of the soil classes at the nodes of the prediction grid as done by Kempen et al. (2009). For Shannon entropy, the values are in the range $[0, \ln m]$, where m is the number of soil classes. The lower bound is reached when a single class has a predicted probability equal to 1 while the upper bound is reached when all classes have predicted probabilities equal to

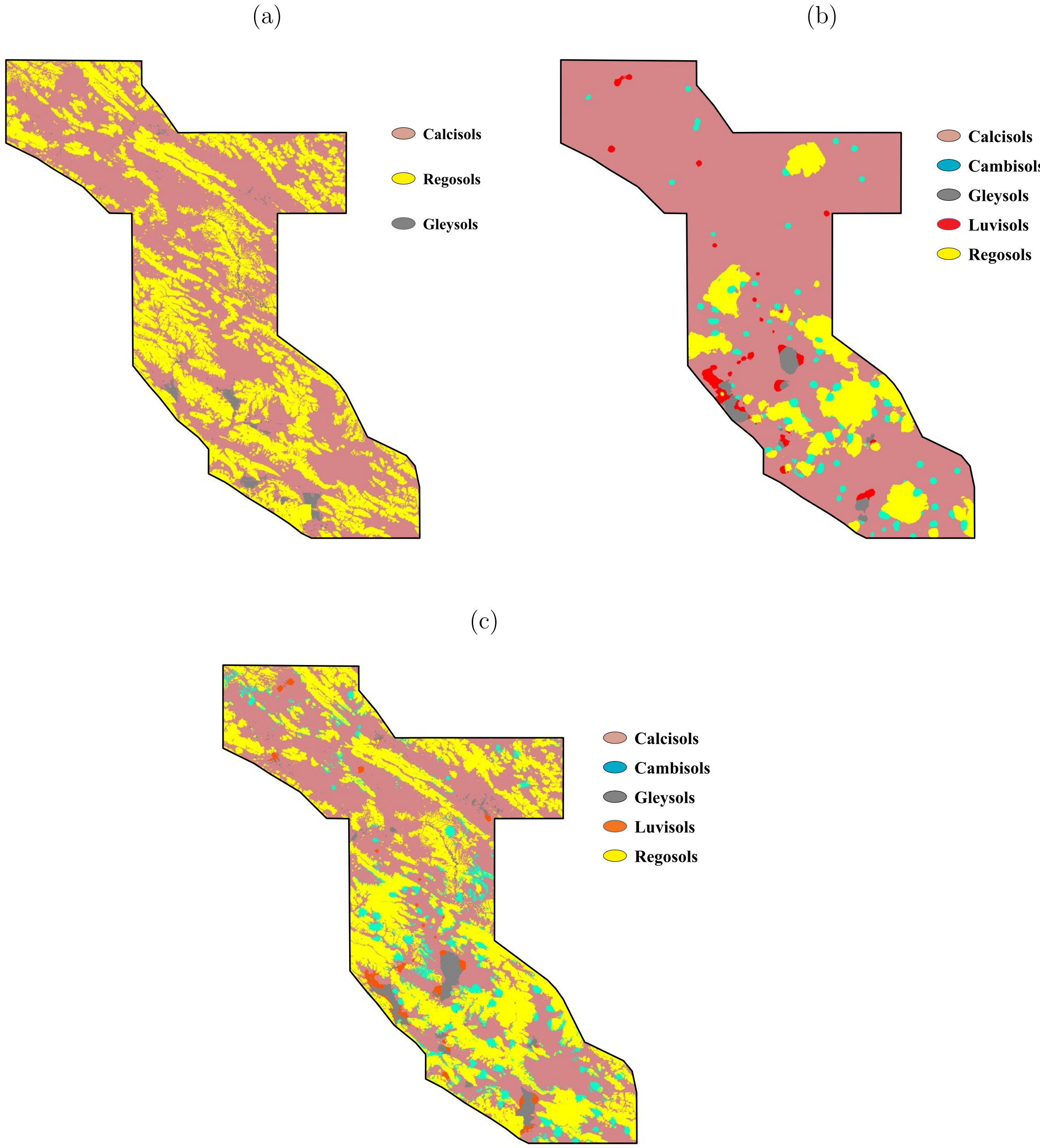

Fig. 5. Maps of the predicted (i.e. maximum probability) soil classes over the study area as obtained from (a) Multinomial Logistic Regression, (b) Bayesian maximum entropy, and (c) Bayesian data fusion.

$1/m$. Fig. 8 shows the relative frequencies of these entropies computed over the whole map (with $m = 5$ for our study). These results are in agreement with the previous findings, where one can see that BDF yields higher frequencies of lower entropy values, with a mean entropy over the map equal to 0.84, to be compared with the values 0.93 and 1.22 for MLR and BME, respectively.

It is worth remembering that BDF as obtained from Eq. (9) relies on a conditional independence hypothesis between the information sources that are used separately for MLR and BME predictions. It is impossible to prove that this hypothesis holds true in practice from the data at hand. However, it possible to check that, conditionally to the true soil class, the confusion matrices for BME and MLR can indeed be considered as statistically independent. Let us consider only the soil classes that are predicted at least once for MLR and BME (thus excluding Cambisols and Luvisols that cannot be predicted at all from the MLR model). For these three soil classes, one can easily build the three-dimensional $3 \times 3 \times 3$ table of joint counts for MLR, BME and true soil classes. Assessing the hypothesis of independence between the counts for MLR and BME predicted classes conditionally to the true soil classes is done using a standard χ^2 test of conditional independence in a three-dimensional contingency table. With an observed statistics $\chi^2_{obs} = 11.01$ for a χ^2 distribution with $(3-1)(3-1)3 = 12$ degrees of freedom, one obtains a p-value $p_v = 0.53$ for the null hypothesis of conditional independence. The fact we accept the null hypothesis (i.e., counts in the various soils classes for MLR and BME are statistically independent conditionally to the true soil class) can be viewed as indirectly evidencing that MLR and BME are making use of unrelated information, in agreement with the hypothesis used in Eq. (9). It also explains why the fused probabilities provide better results than using either MLR or BME, as this fusion procedure is making use of non-redundant and thus complementary information, i.e., those provided by covariates for MLR on one side and those provided by spatially neighboring soil samples for BME on the other side. Fusing them thus makes sense and logically leads to improved prediction results.

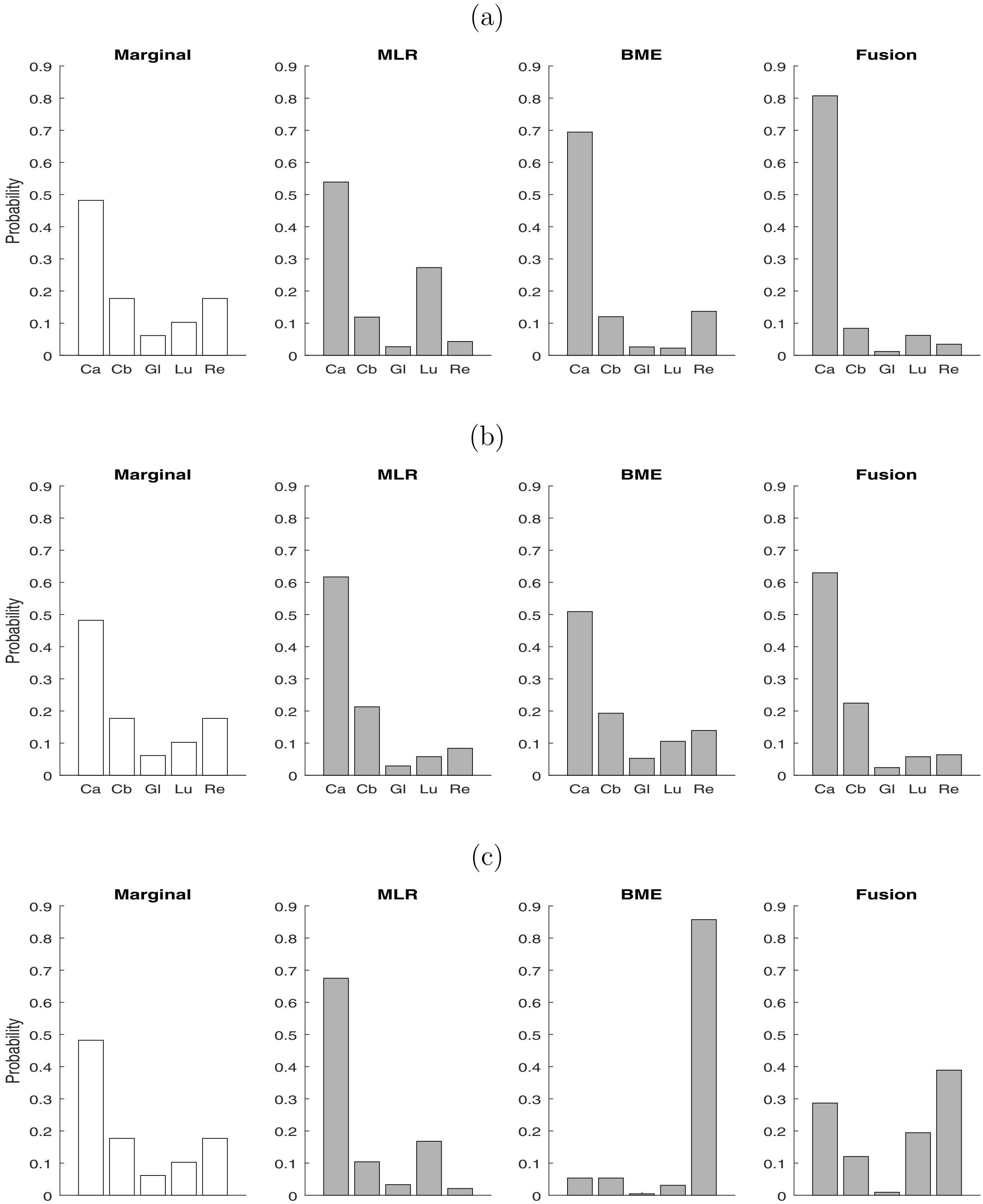

Fig. 6. Illustration of the data fusion procedure for various cases. Part (a) corresponds to an agreement between MLR and BME about the most probable category. Part (b) corresponds to a non-informative BME prediction (same values as for marginal probabilities). Part (c) corresponds to conflicting results for MLR and BME (Ca: Calcisols, Cb: Cambisols, Gl: Gleysols, Lu: Luvisols, Re: Regosols).

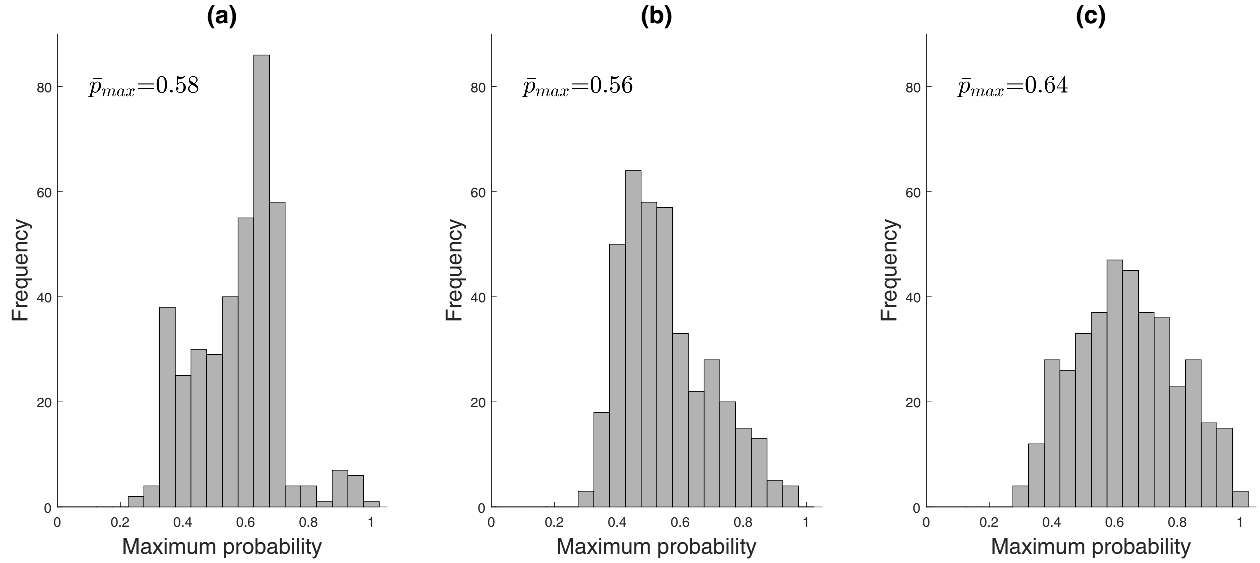

Fig. 7. Distribution of the maximum predicted probability for all soil classes at the 390 sampled locations. Parts (a), (b) and (c) refer to MLR, BME and BDF, respectively, along with the corresponding average maximum probabilities $\bar{p}_{\max}$ for these distributions.

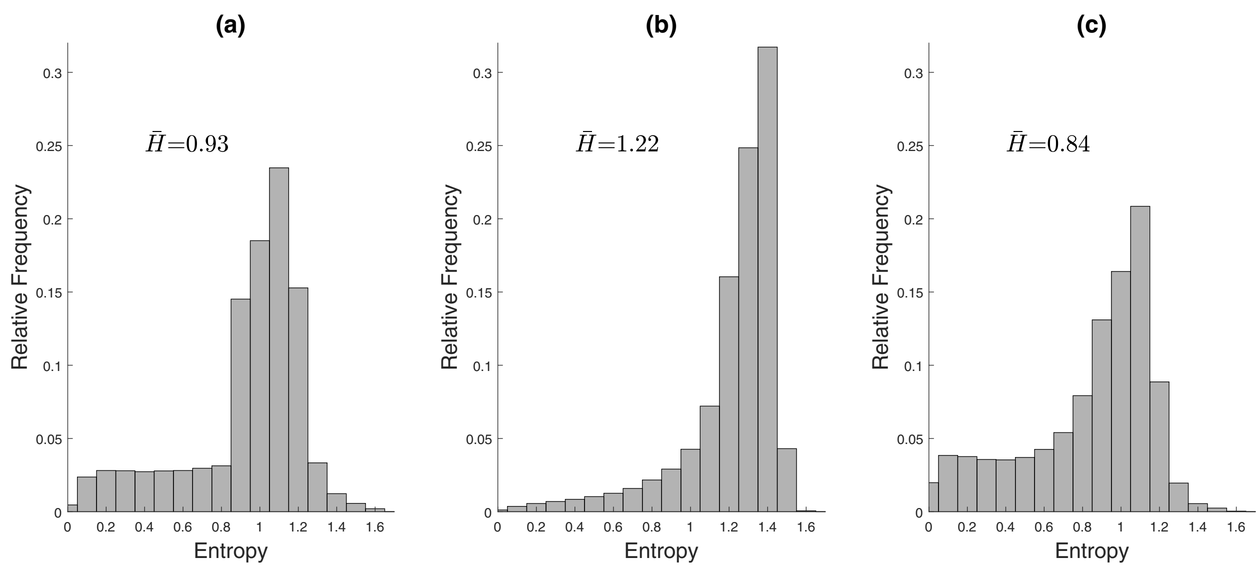

Fig. 8. Relative frequencies of the entropy at the nodes of the prediction grid. Parts (a), (b) and (c) refer to MLR, BME and BDF, respectively, along with the corresponding average entropy $\bar{H}$.

3.4. Limitations of the study

As previously mentioned, the RSG level of the WRB classification is intended for broad geographic groupings based on pedogenesis, while for national maps at least one major qualifier is recommended (or even two for local maps). In our case, as only 390 legacy soil profiles were at hand (with 48% of them classified as Calcisols), it is not realistic to reliably infer either the MLR model or the bivariate probability functions used in the BME model beyond this first level of the WRB. Indeed, this would involve a considerably higher number of classes with few observations in each of them. It is expected that using only the RSG's will lead to possible confusion between classes. For example, Calcaric Cambisols are transitional to Calcisols. Calcic Luvisols and Calcic Cambisols could also wrongly be classified as Calcisols, and this could partly explain the inability of the MLR model to predict these Cambisols and Luvisols classes. Unfortunately, due to the inherent limitations of the data set, there is little hope that this issue can be fixed here.

It is worth remembering that the results and findings in this study are based on the classified legacy soil profiles, by ignoring potential classification errors. Those errors are expected to negatively impact the quality of the predictions. Accounting additionally for these errors in the prediction process would be theoretically feasible both for MLR and BME (e.g., using a Monte-Carlo approach), but this would require first than these errors can be properly assessed. This was clearly not possible for our study.

A LOOCV was used to compare the performances of the MLR, BME and BDF approaches by considering a simple binary correct/incorrect loss function, without accounting for the specificity of the classification. As it is expected that some classes are more easily confused than others, some errors are more expected than others too. A weighted accuracy assessment could thus be more relevant (see, e.g., Rossiter et al. (2017) who rely on the use of taxonomic distances). This however requires expert judgments for assigning the corresponding weights, which is a non-trivial task.

Finally, it is worth noting that MLR and BME are two approaches among an increasingly larger set of available DSM methods. Depending on the properties of the variables to be mapped, the covariates at hand and the goal of the study, this includes among others generalized linear geostatistical models (Diggle and Ribeiro, 2007), geographically weighted regression (Song et al., 2016a), regression kriging (Keskin and Grunwald, 2018), kriging with external drift, (quantile) regression trees, random forests, support vector machines and artificial neural networks (see, e.g., Keskin et al., 2019; Pouladi, 2019; Szatmári and Pásztor, 2019 for a recent comparison of various subsets of these methods in different contexts). Our goal here was not to compare all of them but to emphasize how a Bayesian data fusion can improve the final prediction by merging distinct models are using distinct information (namely collocated covariates for MLR and spatial information for BME). Moreover, the results of these alternated DSM methods could similarly be fused using the BDF approach by relying on a similar conditional independence hypothesis, that is likely to hold true as long as the information used in these various methods are clearly different, as it was the case for our study.

4. Conclusions

In this paper, WRB soil classes were predicted over an area of 10,480 km^2 located in Iran by using two DSM approaches, namely MLR and BME. For this goal, we relied on 390 soil profiles coming from legacy soil surveys, where the northern part of the area is much less densely sampled than the southern part.

For MLR, Slope and Valley Depth (as computed from a 90 m resolution DEM) were the most informative covariates for predicting soil classes but the model fails to correctly predict Cambisols and Luvisols classes. On the opposite, BME can correctly predict these classes and has a comparable global accuracy, but the quality of the prediction directly depends on the sampling density around the prediction location. Using a BDF procedure that aims at combining MLR and BME by accounting for their local conditional distributions, the final map has an increased global accuracy of 61%. Even if this improvement might appear as marginal here due to the limited performances of both MLR and BME for our case study, it is nevertheless clear from a comparison of the distributions of the maximum conditional probabilities that BDF increases the chance to correctly predict soil classes. As this data fusion depends on the probability distributions of MLR and BME that vary over space, BDF is thus able to balance the results from these two distinct prediction models on a local basis. As a result, the final fused map is a combination of the spatial features that are separately found in the MLR and BME maps. This is especially true in the southern part of the area where the sampling density is higher (thus yielding better BME predictions), while in the northern part the fused map is mainly driven by the results obtained from MLR only due to the low sampling density.

Although the respective benefits of MLR, BME and data fusion have been emphasized here for the case of WRB soil classes mapping in a semi-arid region of Iran, it is clear that the methodology is relevant for other situations were possibly other models than BME and MLR and other classifications than WRB are used. We thus expect these findings to be beneficial for the soil science community in general. This is particularly relevant when it comes to fuse the results obtained from various prediction models that are relying on distinct sets of information, as building a single prediction model that would makes use of all information at the same time could turn to be intractable. In these situations, a fusion of the prediction results obtained separately from these models is a necessary step.

Acknowledgements

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

Afshar, F.A., Ayoubi, S., Jafari, A., 2018. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. Geoderma 315, 36–48.
Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., 2014. Globalsoilmap: toward a fine-resolution global grid of soil properties,. In: Advances in Agronomy. 125. Elsevier, pp. 93–134.
Ayoubi, S., Mokhtari, J., Mosaddeghi, M.R., Zeraatpisheh, M., 2018. Erodibility of calcareous soils as influenced by land use and intrinsic soil properties in a semiarid region of central Iran. Environmental Monitoring and Assessment 190, 192.
Bierkens, M., Burrough, P., 1993. The indicator approach to categorical soil data: I. Theory. Eur. J. Soil Sci. 44, 361–368.
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., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S., 2010. Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Science & Business Media.
Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. Stochastic Environ. Res. Risk Assess. 16, 425–448.
Bogaert, P., Fasbender, D., 2007. Bayesian data fusion in a spatial prediction context: a general formulation. Stochastic Environ. Res. Risk Assess. 21, 695–709.
Bogaert, P., Gengler, S., 2017. Bayesian maximum entropy and data fusion for processing qualitative data: theory and application for crowdsourced cropland occurrences in Ethiopia. Stochastic Environ. Res. Risk Assess. 31, 1.
Brungard, C.W., 2009. Alternative Sampling and Analysis Methods for Digital Soil Mapping in Southwestern Utah 2. Master of Science Thesis.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68–83.
Brus, D., Yang, R.M., Zhang, G.L., 2016. Three-dimensional geostatistical modeling of soil organic carbon: a case study in the Qilian Mountains, China. Catena 141, 46–55.
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.
Campling, P., Gobin, A., Feyen, J., 2002. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Sci. Soc. Am. J. 66, 1390–1401.
Carle, S.F., Fogg, G.E., 1996. Transition probability-based indicator geostatistics. Math. Geol. 28, 453–476.
Castanedo, F., 2013. A review of data fusion techniques. Sci. World J. 2013, 1–19.
Das, S.K., 2008. High-level Data Fusion. Artech House Publishers.
Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77, 8–18.
Diggle, P., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer, New York.
D'Or, D., Bogaert, P., 2004. Spatial prediction of categorical variables with the Bayesian maximum entropy approach: the Ooypolder case study. Eur. J. Soil Sci. 55, 763–775.
Fassinut-Mombot, B., Choquel, J.B., 2004. A new probabilistic and entropy fusion approach for management of information sources. Info. Fusion 5, 35–47.
Gengler, S., Bogaert, P., 2016a. Bayesian data fusion applied to soil drainage classes spatial mapping. Math. Geosci. 48, 79–88.
Gengler, S., Bogaert, P., 2016b. Integrating crowdsourced data with a land cover product: a Bayesian data fusion approach. Remote Sens. 8, 545–552.
Giasson, E., Clarke, R.T., Inda Junior, A.V., Merten, G.H., Tornquist, C.G., 2006. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Sci. Agric. 63, 262–268.
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. IEEE Trans. Inf. Forensics Secur. 11, 1984–1996.
Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140, 417–427.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.
Hosmer, D.W., Hjort, N.L., 2002. Goodness-of-fit processes for logistic regression: simulation results. Stat. Med. 21, 2723–2738.
Jafari, A., Ayoubi, S., Khademi, H., Finke, P., Toomanian, N., 2013. Selection of a taxonomic level for soil mapping using diversity and map purity indices: a case study from an Iranian arid region. Geomorphology 201, 86–97.
Jafari, A., Finke, P., Vande Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63, 284–298.
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).
Jeune, W., Francelino, M.R., de Souza, E., Inácio, E., 2018. Multinomial logistic regression

and random forest classifiers in digital mapping of soil classes in western Haiti. Rev. Bras. Cienc. Solo 42 e0170133.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdá, A., 2016. The significance of soils and soil science towards realization of the United Nations sustainable development goals. Soil 2, 111–128.
Kempen, B., Brus, D.J., Heuvelink, G.B., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151, 311–326.
Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76, 2097–2115.
Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41.
Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York.
Kuhn, M., et al., 2008. Caret package. J. Stat. Softw. 28, 1–26.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189-190, 627–634.
Lane, P., 2002. Generalized linear models in soil science. Eur. J. Soil Sci. 53, 241–251.
Li, W., Zhang, C., Burt, J.E., Zhu, A., Feyen, J., et al., 2004. Two-dimensional Markov chain simulation of soil type spatial distribution. Soil Sci. Soc. Am. J. 68, 1479–1490.
Ließ, M., Glaser, B., Huwe, B., 2012. Making use of the world reference base diagnostic horizons for the systematic description of the soil continuum application to the tropical mountain soil-landscape of southern Ecuador. Catena 97, 2–30.
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, F., Geng, X., Zhu, A.X., Fraser, W., Waddell, A., 2012. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. Geoderma 171-172, 44–52.
Lu, Z., Han, D., 2017. Multi-source hydrological soil moisture state estimation using data fusion optimisation. Hydrol. Earth Syst. Sci. 21, 3267–3285.
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 Resour. Prot. 4, 929–943.
McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327.
McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Menard, S., 2000. Coefficients of determination for multiple logistic regression analysis. Am. Stat. 54, 17–24.
Mohammadi, M., 1999. Correlation Study of Soils in East Zagros. Iranian Soil and Water Research Institute, pp. N.1062.
Notarnicola, C., Posa, F., 2002. Extraction of soil parameters: two case studies using Bayesian fusion of multiple sources data. In: 2002. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 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.
Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237-238, 190–198.
Olaya, V., 2004. A Gentle Introduction to SAGA GIS. The SAGA User Group eV, Gottingen, Germany, pp. 208.
Park, E., Elfeki, A.M., Song, Y., Kim, K., 2007. Generalized coupled Markov chain model for characterizing categorical variables in soil mapping. Soil Sci. Soc. Am. J. 71, 909–917.
Pouladi, N., 2019. Møller, A.B., Tabatabai, S., Greve, M.H. Mapping soil organic matter contents at field level with cubist, random forest and kriging. Geoderma 342, 85–92.
Rasaei, Z., Bogaert, P., 2019. Spatial filtering and Bayesian data fusion for mapping soil properties: a case study combining legacy and remotely sensed data in Iran. Geoderma 344, 50–62.
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., Mendona-Santos, M.d.L. (Eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht, pp. 69–80.
Rossiter, D.G., Zeng, R., Zhang, G.L., 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma 292, 118–127.
Scull, P., Franklin, J., Chadwick, O., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geog. Earth Environ. 27, 171–197.
Smith, S., Bulmer, C., Flager, E., Frank, G., Filatow, D., 2010. Digital soil mapping at multiple scales in British Columbia, Canada. In: Program and Abstracts, 4Th Global Workshop on Digital Soil Mapping Rome, Italy, pp. 17–25.
Song, X.D., Brus, D.J., Liua, F., Li, D.C., Zhao, Y.G., Yang, J.L., Zhang, G.L., 2016a. Mapping soil organic carbon content by geographically weighted regression: a case study in the Heihe River basin, China. Geoderma 261, 11–22.
Song, X.D., Liu, F., Zhang, G.L., Li, D.C., Zhao, Y.G., 2016b. Estimation of soil texture at a regional scale using local soil-landscape models. Soil Sci. 181, 435–445.
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.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geo-statistics and machine learning algorithms. Geoderma 337, 1329–1340.
Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403.
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.
WRB, I.W.G., 2015. World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports. 106. pp. 192.
Wu, K., Nunan, N., Crawford, J.W., Young, I.M., Ritz, K., 2004. An efficient Markov chain model for the simulation of heterogeneous soil structure. Soil Sci. Soc. Am. J. 68, 346–351.
Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J.E., Qi, F., et al., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. 75, 1044–1053.
Zeraatpisheh, M., Ayoubi, S., Jafari, A., Finke, P., 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. Geomorphology 285, 186–204.
Zhang, G. l., Feng, L., SONG, X. d., 2017. Recent progress and future prospect of digital soil mapping: a review. J. Integr. Agric. 16, 2871–2885.
Zhang, H., 2004. The optimality of naive Bayes. Proceedings of the the 17th International FLAIRS Conference (FLAIRS2004). pp. 562–567.