Geoderma 219–220 (2014) 136–144

Contents lists available at ScienceDirect

Geoderma

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

Comparative performance of classification algorithms for the development of models of spatial distribution of landscape structures

Hocine Bourennane [a,*], Alain Couturier [a], Catherine Pasquier [a], Caroline Chartin [b], Florent Hinschberger [b], Jean-Jacques Macaire [b], Sébastien Salvador-Blanes [b]

[a] INRA, Unité de Science du Sol, 2163 Avenue de la Pomme de Pin CS 40001 Ardon, F-45075 Orléans Cedex 2, France
[b] Université François-Rabelais de Tours, EA 6293 GéHCO, Faculté des Sciences et Techniques, Laboratoire Géosciences et Environnement, Parc Grandmont, F-37200 Tours, France

ARTICLE INFO

Article history:
Received 5 September 2013
Received in revised form 31 December 2013
Accepted 1 January 2014
Available online xxxx

Keywords:
Factorial discriminant analysis
Multinomial logistic regression
Classification and regression trees
Soil thickness
Morphometric attributes
Landscape structures

ABSTRACT

This work aimed to evaluate whether different types of landscape structures (undulations, lynchets and undisturbed surfaces) can be discriminated by their morphometric attributes and the soil thickness. Three models based on the factorial discriminant analysis (FDA), the multinomial logistic regression (MLR) and the classification and regression trees (CART), respectively, were developed to classify different types of landscape structures. All these statistical techniques were performed using a training sample of 586 individuals over a 17 ha area located in the south-western Parisian Basin. The models developed by the CART and FDA revealed that in addition to soil thickness, the morphometric attributes slope and profile curvature significantly influence the spatial distribution of landscape structures. In addition to the variables selected by CART and FDA models, MLR model included elevation. An external validation of the classification models based on a validation sample of 148 individuals, revealed an overall well classification by CART model of 85% while those achieved with MLR and FDA models were 72% and 77%, respectively. As the predictor variables are known at all the nodes of a regular grid covering the study area; the three models developed were then used to map the landscape structures all over the 17 ha area. Resulting maps revealed a total disagreement between the three models for only 3% of the study area. For more than 50% of the study area the three models predicted a similar landscape structure. For the remaining surface, at least two of the three models predicted a similar landscape structure.

1. Introduction

One of the traditional tasks in soil survey is the allocation of individual sites in predefined classes of the existing systems of classification. To deal with this problem, surveyors have often developed classification approaches using a combination of experience and intuitive judgments to assign individual sites in predefined classes. However, it is generally difficult for soil surveyors to communicate precisely how they do it. Thus these classification approaches are difficult to be reproduced by users. In order to rationalize expertise of soil surveyors, different quantitative methods have been applied over time to study the spatial distribution of soils and their properties.

Among these methods, factorial discriminant analysis (FDA) was used very early and continues to be widely used in soil science (e.g. Anderson et al., 2009; Cox and Martin, 1937; Fernández-Getino et al., 2010; Hirmas et al., 2011; Jungmann et al., 2011; Taylor et al., 2009; Varol et al., 2012; Webster and Burrough, 1974) to attempt to solve assignment problem of soil profiles, soil horizons, etc. to different classes a priori defined.

The multinomial logistic regression (MLR) can also be used to deal with such problem. Indeed, this method was widely used for spatial modeling in land use and ecology studies as well as for digital soil mapping (e.g. Akgün and Türk, 2011; Bailey et al., 2003; Campling et al., 2002; Debella-Gilo and Etzelmüller, 2009; Hengl et al., 2007; Kempen et al., 2009; King et al., 1999; Marchetti et al., 2011; May et al., 2008; Müller and Zeller, 2002; Rhemtulla et al., 2007; Suring et al., 2008; Venkataraman and Uddameri, 2012).

The classification and regression trees (CART), introduced by Breiman et al. (1984), have also some potential to handle with the assignment problem of an individual such as soil profiles and soil horizons to different classes a priori defined. Algorithms of CART are non-parametric; so, no hypothesis is required regarding variable distribution (Friedman, 1991; Mitchie et al., 1994). In addition, several studies have shown that one of the most widely used and best performing inductive learning algorithms in terms of generating interpretable rules as well as prediction accuracy was classification tree algorithm (e.g. Behrens and Scholten, 2006; Loh and Vanichsetakul, 1988). These algorithms were also described as a robust prediction technique (e.g. Lagacherie et al., 2001; Scull et al., 2005). Applications

* Corresponding author at: Institut National de la Recherche Agronomique (INRA), France. Tel.: +33 2 38 41 48 28; fax: +33 2 38 41 78 69.
E-mail address: Hocine.Bourennane@orleans.inra.fr (H. Bourennane).

0016-7061/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2014.01.001

in environmental sciences can thus be found in various disciplines like ecology, remote sensing and soil science (e.g. Bater and Coops, 2009; Friedl and Brodley, 1997; Geissen et al., 2007; Hansen et al., 1996; Ließ et al., 2012; Mulder et al., 2011; Munoz and Felicisimo, 2004; Schmidt et al., 2008; Viscarra Rossel and Behrens, 2010).

The objective of this study was to compare three multivariate methods in the development of classification models for landscape structures and to elucidate the choices of multivariate techniques. For this purpose, we proposed to assess whether different types of anthropogenic landforms could be discriminated by their morphometric attributes and the soil thickness. To deal with this objective, accurate elevation measurements and a dense soil thickness survey were carried out over 17 ha in the center of France. Calibration and validation of the models were conducted from two sets of punctual measurements carried out in the study area. Finally, this paper examines the ability of the most powerful model, in regard to the validation results, to map the different types of anthropogenic landforms over the whole study area.

2. Materials and methods

2.1. Location of the study area and data acquisition

The study site and the data acquisition (Fig. 1) were widely presented in the paper by Chartin et al. (2011). We recall here the main points about these two aspects to help the readers of this work. The study site was carried out on a 17 ha southeast-facing hillslope located near the village of Seuilly (south-western Parisian Basin, 47°08.31′N, 0°10.97′ E). The main soils observed in the study area are calcaric Cambisols, epileptic calcaric Cambisols and colluvic Cambisols (Bellemlih, 1999; Boutin et al., 1990; FAO, 1998).

The landscape is composed of three types of morphological elements. Two types correspond to anthropogenic linear landforms, lynchets and undulations, located along former field borders, removed during previous campaigns of land consolidation, and along present field borders, respectively (Chartin et al., 2011; Houben, 2008). The geometrical characteristics (shape and size) of both lynchets and undulations are different and widely presented in the paper by Chartin et al. (2011). In addition, they are distinguishable infield from "undisturbed surfaces", i.e., areas which morphology was not affected by the presence of any present or former field borders.

Soil thickness was measured by manual augering at 734 locations (Fig. 1b) by considering the spatial distribution of considered linear landforms and undisturbed areas. Twenty percent of the observations (148 points) were randomly selected to constitute the validation set. The remaining 80% of the dataset (586 points) was used as the training set of the model.

A topographical survey was performed using two DGPS (Trimble ® ProXRS) as a base and a mobile recorder, respectively. Coordinates and elevations of 1550 points were obtained by post-treatment of the data and used to estimate a Digital Elevation Model (DEM) on a two-meter grid. Topographic attributes such as slope gradient (Slope), curvature (Curve), planform and profile curvatures (Planc and Profc) were derived (Fig. 2) from the DEM through the algorithms implemented in the GIS ArcGis 9.3.1.

Finally, each point of the soil sampling scheme (Fig. 1b) was informed about values of soil thickness and topographic attributes, and assigned to one of the three categories of landscape structures (lynchets 'LY', undulations 'UN' or undisturbed surface 'US') on the basis of its geographic coordinates.

2.2. Principles of factorial discriminant analysis (FDA)

This section is devoted to a brief presentation of FDA used to establish the classification model of landscape structures on the study area. For a detailed presentation, the reader can refer to books on the subject, such as Tabachnick and Fidell (1996) and Tomassone et al. (1988).

FDA is a statistical method for describing and forecasting. Its purpose is to study the relationship between a qualitative variable and a set of quantitative variables. Three main objectives can be assigned to the discriminant analysis:

1. determine the variables most discriminating with regard to specific category,
2. determine the category of an individual based on its characteristics,
3. validate a classification or make a choice between several classifications to determine which is most relevant.

The discriminant analysis comes at a posteriori classification. The FDA can be considered as an extension of the problem of regression where the dependent variable is qualitative. The data consist of n observations divided into k classes or categories and described by p variables. Traditionally, one can distinguish two aspects in discriminant analysis:

1. a descriptive aspect which consists in finding linear combinations of variables that separate in the best way the k categories and gives a graphic representation that well reflects this separation,
2. a decisional aspect where a new individual arises and for which we know the values of the predictors, it is then to decide in which category it should affect it. In such cases, this is a classification problem.

Two models of FDA are possible based on a fundamental assumption: if we assume that the covariance matrices are identical, one is in the case of linear factorial discriminant analysis. Assuming that the covariance matrices are different for at least two categories, we are then in the case of a quadratic model. The test of Box allows checking this hypothesis (Bartlett's approximation allows the use of a chi-square law for the test).

2.3. Multinomial logistic regression (MLR)

Multinomial logistic regression is the extension for the binary logistic regression when the categorical dependent outcome has more than two levels.

The goal of multinomial logistic regression is to estimate the probability of each class using a same set of influencing variables. The model is similar to the binomial logistic regression in the sense that the logarithm of the odds ratio is assumed to be a linear function of the influencing variables. However, one of the classes is taken as the baseline and odds ratios are developed for all other classes with respect to this baseline. For a thorough presentation, the reader can refer to Agresti (2002) or Hosmer and Lemeshow (2000). Nonetheless, a brief presentation is given below concerning the binomial logistic model and its generalization to the multinomial case.

In the binomial logistic regression, the probability (p_1) that an object belongs to group 1, and the probability (p_2) that it belongs to group 2, according to a set of predictor variables, are given by the logit link function:

$$\text{logit}(p_1) = Ln(p_1/p_2) = Ln(p_1/1-p_1) = \mathbf{x}\beta \tag{1}$$

where $\mathbf{x}$ is a vector of predictor variables, and β is a vector of model coefficients that are usually estimated by maximum likelihood.

The expression (Eq. (1)) can be rewritten as:

$$\frac{p_1}{1-p_1} = \exp(\eta). \tag{2}$$

The left term in Eq. (2) is called the odds ratio. From expression (2) it follows that:

$$p_1 = \frac{\exp(\eta)}{1+\exp(\eta)}. \tag{3}$$

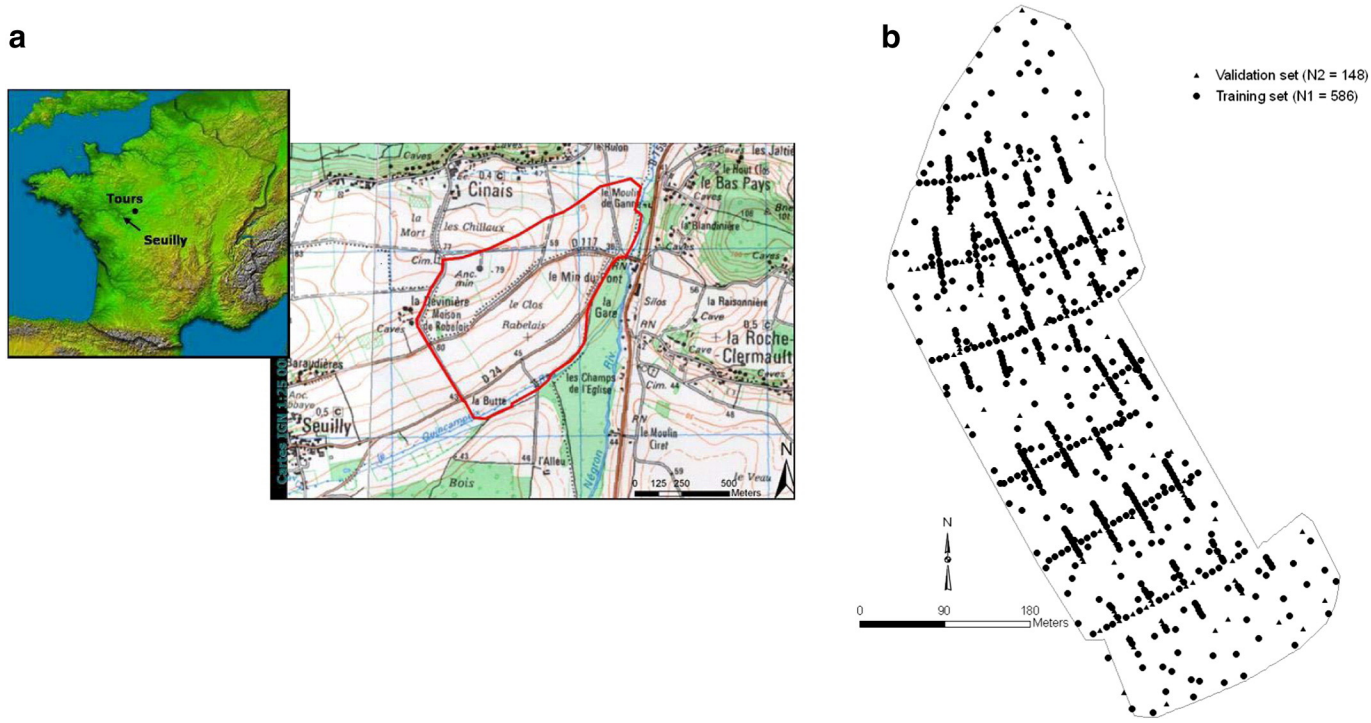

Fig. 1. (a) Localization of the study site; (b) sampling pattern of the training set and the validation set.

The binomial logistic regression model can be generalized to the multinomial case where the number of logistic functions is one less than the number of groups. For example if there are three groups, one of the groups is taken to be a reference group (say group 0), so that the first logistic function can be used to predict the probability that an object will belong to group 1 rather than group 0, and the second logistic function can be used to predict the probability that an object will belong to group 2 rather than group 0.

The significance of the logistic regression model is assessed with the likelihood ratio test (G statistic). This test is used to determine the improvement of the full model over the intercept-only model (Hosmer and Lemeshow, 2000). According to the same authors, the significance of an individual model coefficient is assessed with the Wald statistic (W), which is obtained by comparing the estimated coefficient to an estimate of its standard error.

2.4. Classification and regression trees (CART)

A classification tree is used to predict the group membership of objects on the basis of one or more predictor variables. The tree consists of a set of decision rules, applied in a sequential manner, until each object has been assigned to a specific group. The first decision rule, applied at the 'root node' of the tree to the values of all objects along one or more predictor variables, has two possible outcomes: objects are either sent to a terminal node (leaf), which assigns a class, or to an intermediate node, which applies another decision rule. Ultimately, all objects are sent to a terminal node and assigned a class. In the simplest type of classification tree, the splits are binary (each parent node is attached to two daughter nodes) and the decision rules are univariate (based on a single variable). Classification tree can be based on continuous or discrete predictor variables, or on a mixture of both (when univariate splits are used), and the trees are generally constructed by recursive partitioning (i.e. a given predictor variable can be used in more than one decision rule).

One commonly used algorithm for constructing classification tree is classification and regression trees (CART), developed by Breiman et al. (1984). CART find optimal univariate splits by carrying out an exhaustive search of all possible splits. CART are non-parametric classifiers: no assumptions are made about the distributions of the variables.

3. Results and discussion

3.1. Descriptive statistics on the training sample and spatialization of soil thickness

The training sample includes 586 individuals. These individuals were assigned to three distinct categories of landscape structures. Thus, 319 individuals were assigned to category US representing undisturbed surfaces, 167 individuals in category UN representing the undulations and 100 collected on the lynchets were assigned to LY category. In addition, each individual in the training sample was informed of the thickness of the soil (ST) and five morphometric attributes: altitude (Elv), the intensity of the slope (Slope), curvature (Curve), and plan and profile curvatures (Planc and Profc) derived from the DEM by using the algorithms implemented in the GIS ArcGis 9.3.1.

The variable to be predicted is a categorical variable to three terms. This prediction is made from six quantitative variables observed on a training sample of 586 individuals. From this training sample, classification models of landscape structures were developed by applying FDA, MLR and CART which were presented in the previous section.

Table 1 shows not only significant differences between the averages of variables of the three categories, but also values significantly different for standard deviations.

The Wilks's Lambda test allowed assessing whether the vectors of averages for the various categories are equal or not. Results of this test (Table 2) showed that at least one of the mean vectors was different from another since the calculated p-value was inferior to the significance level $\alpha = 0.05$. However, the box plots (not shown) of the distribution of variables within each category and their comparisons showed that a number of recoveries can take place between the three categories for some of the measured variables.

Prior to map the anthropogenic landforms using the different models developed, ST was mapped (Fig. 3a) by ordinary kriging on a

2 m regular grid over the whole study area. The spatial autocorrelation of ST (Fig. 3b), quantified through the semivariogram, was quite strong. This experimental semivariogram was fitted by a nested model. The latter consists of a nugget model plus a spherical model. In an attempt to validate the variogram model, cross validation was used on the original data. Every known point was estimated by using a neighborhood around it, but not itself. Having made such calculations, the results using a moving neighborhood are revealing. They show that the mean error was close to zero (0.002), and the ratio of the mean squared error to the kriging variance (the variance ratio) was close to 1 (1.005). Thus, using the fitted variogram functions and the kriging equation system, ST was estimated (Fig. 3a) over the study area from the 586 punctual measurements of ST.

3.2. Mapping of anthropogenic landforms using the factorial discriminant analysis model

Two models of FDA are possible based on a fundamental assumption. If we assume that the covariance matrices are identical, we should deal with linear factorial discriminant analysis. On the contrary, if we

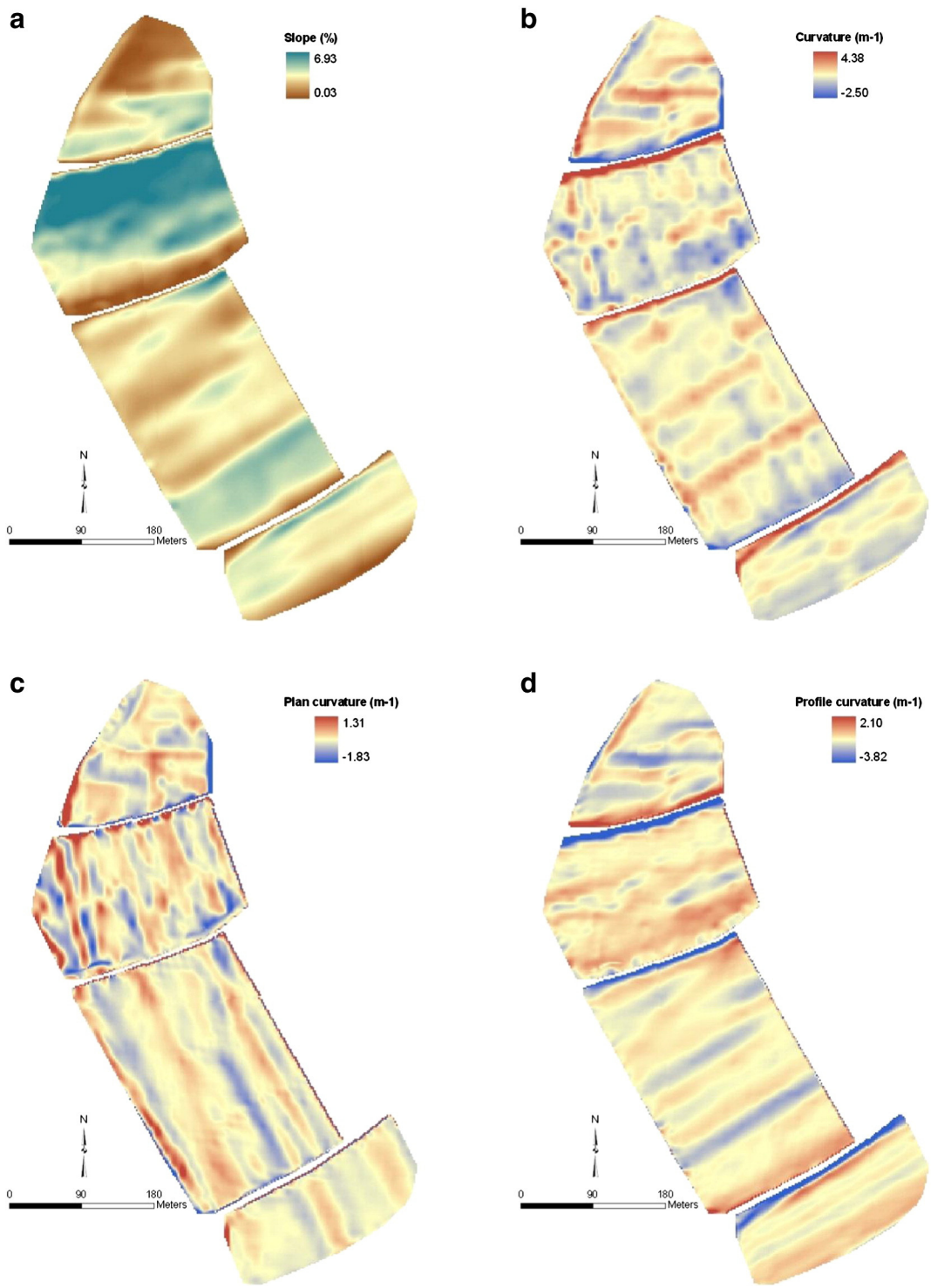

Fig. 2. (a) Slope gradient; (b) curvature; (c) plan curvature; (d) profile curvature, all derived from a Digital Elevation Model.

Table 1
Basic statistics of six quantitative descriptors of the qualitative variable: statistics by category.

Category	Variable	Unit	Mean	Std
LY: Lynchet	ST	m	1.05	0.33
($N_1 = 100$)	Elv	m	59.04	12.83
	Curve	m^{-1}	−0.20	0.22
	Planc	m^{-1}	−0.01	0.12
	Profc	m^{-1}	0.19	0.21
	Slope	%	1.49	1.19
UN: Undulation	ST	m	0.61	0.19
($N_2 = 167$)	Elv	m	56.96	6.36
	Curve	m^{-1}	0.05	0.11
	Planc	m^{-1}	0.01	0.05
	Profc	m^{-1}	−0.05	0.10
	Slope	%	3.14	1.25
US: Undisturbed	ST	m	0.45	0.18
surface ($N_0 = 319$)	Elv	m	57.95	11.59
	Curve	m^{-1}	−0.02	0.17
	Planc	m^{-1}	0.01	0.06
	Profc	m^{-1}	0.02	0.15
	Slope	%	3.44	1.40

assume that the covariance matrices are different for at least two groups, we should deal with a quadratic model. The Box test can assess this hypothesis (Bartlett's approximation allows the use of a chi-square law for the test). The results of this test presented in Table 3 confirm that we cannot make the assumption of equal variance–covariance matrices between the three categories.

FDA (backward stepwise) revealed that the most discriminant variables for the three categories were ST, Slope, and Profc. Table 4 shows the standardized coefficients of canonical discriminant functions. These coefficients allow measuring the relative contribution of the initial variables to discrimination for a given category. For the first discriminant function the intensity of the slope is opposed to the other discriminant variables. The contribution of ST and Profc was greater with respect to that of the Slope in the first discriminant function. For the second discriminant function, the contribution of the Profc was greater and was opposed to the contributions of the two other variables. Coefficients of ranking functions (Table 5) can be used to directly determine which category must be assigned an observation based on the values taken for the various explanatory variables. An observation is assigned to the category for which the ranking function is highest.

The results of classification (Table 6a) show an overall percentage of well-classification in the order of 77%. In other words, the apparent error rate on the training sample is 23%. This error rate is mainly due to the US category and to a lesser degree to the UN category.

However, the same sample (training sample) was used to determine the coefficients of ranking functions and to evaluate the success of the assignment rules. This approach can still give an overly optimistic picture.

The ideal is to perform an external validation. This procedure has the advantage of providing unbiased estimates of the percentages of well-sorted and poorly-sorted. It consists of estimating the coefficients of discriminant functions on a training sample whereas the allocation rule and calculating the percentage of well classified are performed from another sample set (validation sample).

Table 2
FDA: test of Wilks's Lambda (Rao approximation).

Lambda	0.379
F (observed value)	120.984
F (critical value)	2.106
df1	6
df2	1162
p-Value	<0.0001
α	0.05

Table 3
FDA: test of Box (chi-square asymptotic approximation).

−2Log(M)	272.172
Chi^2 (observed value)	269.584
Chi^2 (critical value)	21.026
df	12
p-Value	<0.0001
α	0.05

In the absence of an external validation sample, it is essential to achieve at least a systematic cross-validation. This operation consists in excluding systematically the individual from the training sample that one wants to classify. The calculation of posterior probabilities used in the allocation rule is not based on the value of the generalized distances of the centroids of individual groups; this procedure does not require to run N discriminant analyzes but to adjust the calculation of distances of Mahalanobis. In our study, systematic cross-validation (Table 6b) leads to the same results with the training sample (Table 6a) but this is not always the case.

Finally, the results of external validation (Table 6c) go along the same lines as those of the confusion matrix based on the training sample and the confusion matrix of systematic cross-validation (Table 6a and b) namely that the risk of confusion between individuals of categories US and UN is higher in contrast to the LY category where all the results show that it is easier to discriminate them from other categories (the undulations and undisturbed areas) based on soil and morphometric criteria measured.

As a sequel, the coefficients of ranking functions presented in Table 5 are applied to the pixels of slope and Profc presented in Fig. 2 as well as for all pixels of ST (Fig. 3a). Thus for each category (LY, UN, US) we obtain a grid and then each pixel of the study area is assigned to the category for which the ranking function is highest. We obtain thus (Fig. 4a) the spatial distribution of the three categories of landscape structures across the study area.

3.3. Regionalizing anthropogenic landforms through multinomial logistic regression model

The Likelihood Ratio, Score, and Wald tests (Table 7a) were examined to determine the improvement of the MLR model over the intercept-only model (also called the null model). All three tests yielded similar results ($p < 0.0001$, Table 7a), namely, the MLR Model was more effective than the null model. It was therefore inferred that at least one explanatory variable was a significant predictor of landscape structures.

To assess the strength of multinomial logistic regression relationship, Cox & Snell R Square and the Nagelkerke R square values were used. These statistics provide an indication of the amount of variation in the dependent variable. These are described as pseudo R square. Table 7b reveals that the values are 0.57 and 0.66 respectively; suggesting that between 57% and 66% of the variability is explained by the set of variables used in the model.

The evaluation of the usefulness for logistic models was assessed by computing the proportional accuracy rate due to chance. This statistic was computed by calculating the proportion of cases for each category based on the number of cases on each category and then by squaring and summing the proportion of cases in each category. The value of this statistic is equal to 0.406 in our case.

To characterize the model as useful, we compared the overall percentage accuracy rate produced versus the proportional accuracy rate due to chance. If the latter is 25% less than overall percentage accuracy rate, we can conclude that the model is helpful.

The classification accuracy rate was equal to 74.06% (Table 8a) which was greater than the proportional by chance accuracy of 50.75% ($1.25 \times 40.6\% = 50.75\%$), suggesting that the model was useful.

The validation results using the external validation sample of 148 individuals are summarized in Table 8b. They revealed that the overall

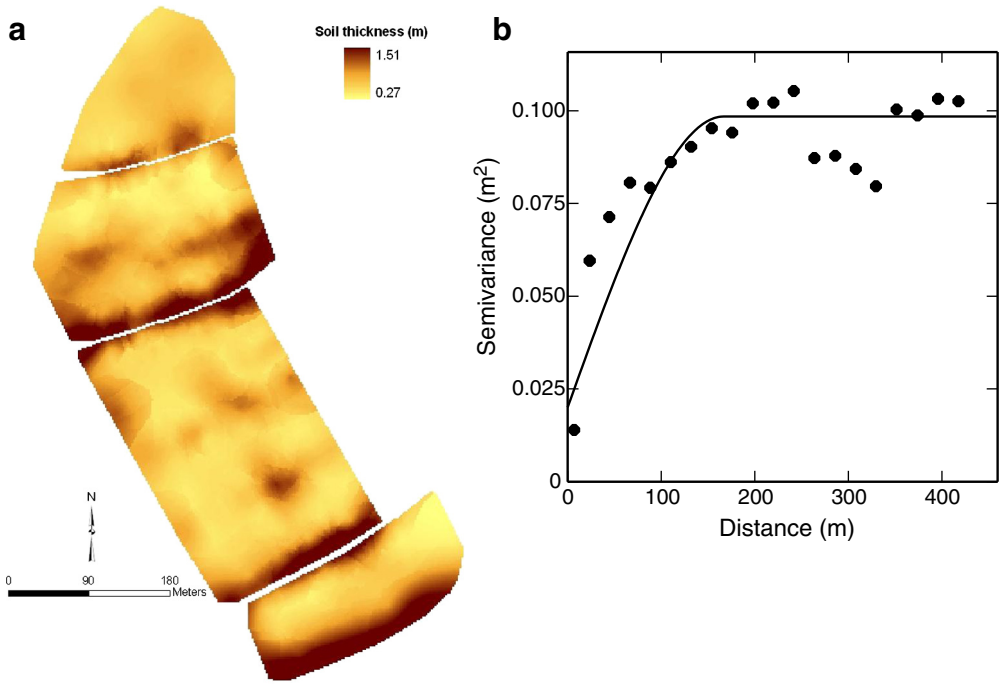

Fig. 3. (a) Soil thickness (ST) predicted by ordinary kriging from punctual measurement (training set); (b) experimental variogram of ST (dots) and the theoretical model fits (solid line).

well classification by the MLR model was less than that obtained by FDA (72% versus 77% of correct classification). The discrepancy between the results of the two classification models is owing to category UN where the individuals are less well discriminated from the US category when using the MLR model.

The intercept of the multinomial logistic regression model and slope coefficients for each predictor variables, along with their *p*-values for categories US and UN of the training sample are shown in Table 9. The model was developed with the probability of occurrence of category LY (Lynchets) as reference for evaluating logits. The Wald test evaluates whether or not, the independent variable is statistically significant in differentiating between two groups in each of embedded binary logistic comparisons.

In comparing the probabilities of occurrence of category US to category LY, it is evident that all variables except for Planc were significant predictors. A unit increase in Slope increased the logit of occurrence of category US as opposed to category LY. It also appears that a unit decrease in ST would increase the logit of category US over category LY.

The statistically significant variables for category UN with respect to category LY were also all variables except for Planc. A unit decrease in ST increases the odds of category UN as opposed to category LY. A unit decrease in Profc seems to favor the occurrence of category UN as opposed to category LY.

The MLR models summarized in Table 9 were used to estimate the probabilities of occurrence of the spatial distribution of the three categories of landscape structures over the study area. In practice, the MLR models were applied to the pixels of ST, Profc, Slope and Elv predictor variables. Having made such calculations, we obtained maps showing the probability of occurrence at each pixel for each category. Finally, the category with the largest probability was used to construct a predictive map (Fig. 4b) of the spatial distribution of landscape structures over the study area.

Table 4
FDA: standardized coefficients of canonical discriminant functions.

	F1	F2
ST	0.817	0.437
Slope	−0.346	0.108
Profc	0.410	−0.886

Table 5
FDA: coefficient of ranking functions.

	US	LY	UN
Constant	−4.839	−9.734	−2.280
ST	18.536	17.210	12.618
Slope	2.345	3.655	1.379
Profc	1.363	6.743	−8.026
ST × ST	−16.757	−6.563	−18.990
ST × Slope	−1.055	−2.016	3.112
ST × Profc	1.726	−2.178	−20.639
Slope × Slope	−0.272	−0.508	−0.489
Slope × Profc	−0.299	−0.105	4.791
Profc × Profc	−23.567	−11.456	−61.082

Table 6
Confusion matrices resulting from the factorial discriminant analysis.

Category	US	LY	UN	Total	% correct
(a) FDA: confusion matrix for the training sample					
US	231	17	71	319	72.41
LY	8	87	5	100	87.00
UN	29	4	134	167	80.24
Total	268	108	210	586	77.13
(b) FDA: confusion matrix for the results of systematic cross-validation					
US	230	17	72	319	72.10
LY	10	85	5	100	85.00
UN	29	4	134	167	80.24
Total	269	106	211	586	76.62
(c) FDA: confusion matrix for the results of external validation					
US	52	7	11	70	74.29
LY	6	32	1	39	82.05
UN	8	1	30	39	76.92
Total	66	40	42	148	77.03

3.4. Application of CART model to anthropogenic landform mapping

As already stated by Chartin et al. (2011), the overall prediction performance of the CART model was more than 80% when applied to morphometric attributes and soil thickness values of the training sample (Table 10a). The confusion matrix showed that the resulting classification and regression tree performed well for categories US and LY. Categories US and LY have 87.77% and 85.0% of their respective points rightly classified. Approximately three quarters of the misclassified points from category US are classified in category UN. Concerning category LY, the main errors of the model appeared to involve the category US. In category UN, 24.0% of points were incorrectly classified; they are all allocated to category US by the model. The most important risk of confusion during the application of the CART model then, involves the category UN and the category US.

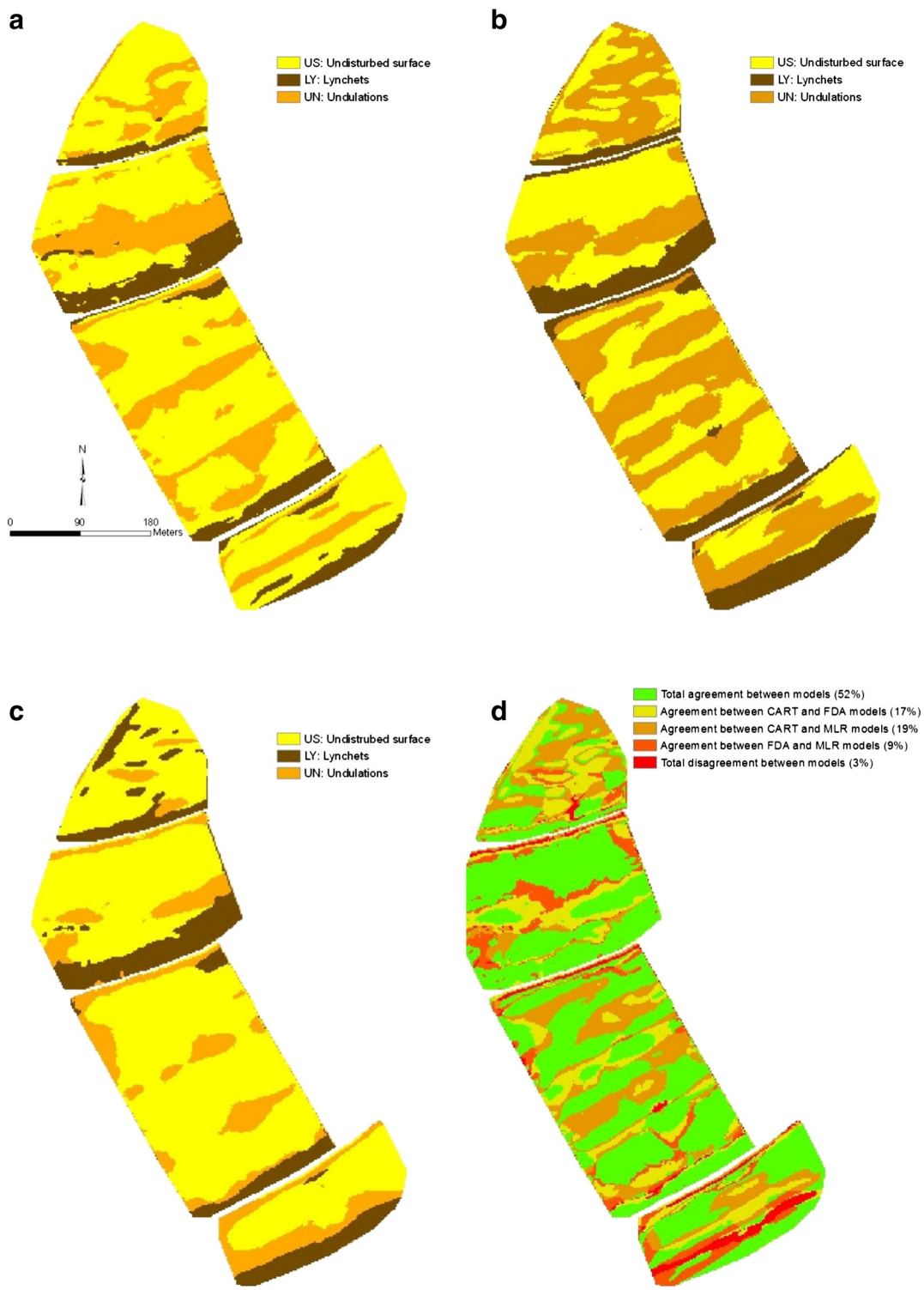

Fig. 4. Spatial distribution of the three categories of landscape structures carried out from: (a) the ranking functions of the FDA; (b) MLR model; (c) CART model; and (d) matching of the prediction models.

Table 7
Goodness of fit measures for the multinomial logistic regression model.

(a) MLR: overall model evaluation			
Tests	χ^2	*df*	*p*
Likelihood ratio test	497.462	10	<0.0001
Score test	412.368	10	<0.0001
Wald test	155.910	10	<0.0001
(b) MLR: pseudo R-square			
Cox & Snell R^2			Nagelkerke R^2
0.57			0.66

Table 8
Confusion matrices resulting from the multinomial logistic regression model.

Category	US	LY	UN	Total	% correct
(a) Classification result for the training sample using multinomial logistic regression model					
US	274	7	38	319	85.89
LY	11	86	3	100	86.00
UN	91	2	74	167	44.31
Total	376	95	115	586	74.06
(b): Confusion matrix for the results of external validation using multinomial logistic regression model					
US	55	6	9	70	78.57
LY	4	33	2	39	84.18
UN	20	1	18	39	46.15
Total	79	40	29	148	71.62

Table 10b presents validation results for the CART model performed through the validation sample. According to these results, 85% of the points from the validation sample are well classified. Categories US, LY and UN had 86%, 87% and 82% of their points well classified, respectively. These proportions are slightly different from those presented in Chartin et al. (2011) since the CART model, in this paper, considered less morphometric attributes when running the model compared to the model in Chartin et al. (2011). In any way, the present CART model appeared significantly relevant even if the differences between categories US and UN could be delicate in some situations. The application of the decision rules of the CART model to the grids of Slope, ST and Profc allowed mapping (Fig. 4c) the different landscape structures over the study area.

A total agreement between the three maps, achieved by the FDA, MLR and CART models, was observed for 52% of the total area (Fig. 4d) against 3% for a total disagreement. The confidence level of the assignment of the structure types over these areas can be considered as very high for 52% of the total area and very low for 3% of the total area.

Table 9
Multinomial logistic regression model parameters.

Category	Source	Coefficient	SD	χ^2 Wald	$p > \chi^2$
US	Intercept	13.518	2.191	38.081	<0.0001
	ST	−10.888	1.184	84.518	<0.0001
	Elv	−0.103	0.025	17.117	<0.0001
	Planc	3.783	2.666	2.013	0.156
	Profc	−13.197	2.480	28.313	<0.0001
	Slope	0.817	0.165	24.452	<0.0001
UN	Intercept	11.383	2.185	27.142	<0.0001
	ST	−6.013	1.076	31.236	<0.0001
	Elv	−0.116	0.025	21.156	<0.0001
	Planc	4.145	2.729	2.308	0.129
	Profc	−17.027	2.498	46.459	<0.0001
	Slope	0.712	0.161	19.636	<0.0001

Table 10
Confusion matrices resulting from the classification and regression trees model.

Category	US	LY	UN	Total	% correct
(a) CART: matrix of confusion for the training sample					
US	280	10	29	319	87.77
LY	12	85	3	100	85
UN	40	0	127	167	76.05
Total	332	95	159	586	83.96
(b) CART: confusion matrix for the results of external validation					
US	60	2	8	70	85.71
LY	4	34	1	39	87.18
UN	7	0	32	39	82.05
Total	71	36	41	148	85.14

As the overall well classification by the CART model (85%) was higher compared to that achieved with the MLR and FDA models (72% and 77% of accurate classification, respectively) based on the external validation, it can be assumed that for 36% of the area (Fig. 4d) the confidence level of the assignment of the structure can be considered as high. Indeed, for these areas agreements between the CART model and FDA model on the one hand and on the other hand between the CART model and MLR model were observed. For the remaining area (9%), the confidence level of the assignment of the structure can be considered as medium since the results of the CART model were different from those achieved using the models FDA and MLR which are obviously in agreement.

4. Summary and conclusions

The focus of this paper has been on the comparison of three multivariate methods in the development of classification models for landscape structures. The three methods used for developing the classification model were factorial discriminant analysis (FDA), multinomial logistic regression (MLR) and classification and regression trees (CART). The models were constructed using morphometric attributes and soil thickness that explain the occurrence of three landscape structures. The results of this study showed that in addition to soil thickness, the morphometric attributes that significantly influenced the spatial distribution of landscape structures were slope and profile curvature when using the CART and FDA models. The MLR model uses the elevation in addition to the variables selected by the CART and FDA models. The external validation revealed that the CART model appeared more appropriate in the assignment of the objects to the three categories of landscape structures compared to the FDA and MLR models. The overall well classification for the three models ranged from 72% (MLR model) to 85% (CART model). Nevertheless, the improvement of the assignment of objects reached a maximum of 15% between the CART model and the MLR model. In addition, the mapping performed using each of the three models revealed a total disagreement between the models for only 3% of the study area and for more than 50% of the study area the three models predicted a similar landscape structure. For the remaining area, two of the three models predicted a similar landscape structure.

The question therefore arises as to which method should be employed in a given context. All the three methods can be used to assign objects to two or more groups, and all the three methods can employ one or more predictor variables. FDA and MLR use all statistically significant predictor variables simultaneously in the model, whereas CART use the predictor variables in a hierarchical and recursive manner. An advantage of CART analysis lies in its use as a nonparametric classifier; in contrast to FDA and MLR, which both make assumptions about the nature of the data. In addition, the CART method gives rules in natural language. From a practical and scientific point-of-view, it is always desirable to choose the simplest model that has a satisfactory predictive performance. Accordingly, the final message from our findings is that

the CART method is the simplest and the best for spatial distribution of a categorical variable as it leads to a model which: (*i*) gives the best results for both cross validation and the external validation; (*ii*) gives classification rules in natural language handy to non-statistician users; and (*iii*) requires no assumptions about the nature of the data distribution.

Acknowledgments

Financial support provided by the ANR (Agence Nationale de la Recherche) VMCS project LANDSOIL is gratefully acknowledged.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd edition. John Wiley and Sons, New York.
Akgün, A., Türk, N., 2011. Mapping erosion susceptibility by a multivariate statistical method: a case study from the Ayvalık region, NW Turkey. Comput. Geosci. 37, 1515–1524.
Anderson, R.H., Farrar, D.B., Thoms, S.R., 2009. Application of discriminant analysis with clustered data to determine anthropogenic metals contamination. Sci. Total Environ. 408, 50–56.
Bailey, N., Clements, T., Lee, J.T., Thompson, S., 2003. Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach. J. Environ. Manag. 67, 395–407.
Bater, C.W., Coops, N.C., 2009. Evaluating error associated with lidar-derived DEM interpolation. Comput. Geosci. 35, 289–300.
Behrens, T., Scholten, T., 2006. A comparison of data-mining techniques in predictive soil mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam, pp. 353–364.
Bellemlih, S., 1999. Stocks particulaires holocènes et bilans de matières dans un bassin fluviatile en domaine sédimentaire - Le bassin du Négron, Sud-ouest du Bassin Parisien, France. Ph.D. thesis Université de Tours, France.
Boutin, D., Froger, D., Rassineux, J., 1990. Feuille Loudun (1724–1624), Carte des sols du Département de la Vienne et de la région Centre au 1:50000. Chambre d'Agriculture de la Vienne - IGN - INRA.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth.
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.
Chartin, C., Bourennane, H., Salvador-Blanes, S., Hinschberger, F., Macaire, J.-J., 2011. Classification and mapping of anthropogenic landforms on cultivated hillslopes using DEMs and soil thickness data — example from the SW Parisian Basin, France. Geomorphology 135, 8–20.
Cox, G.M., Martin, W.M., 1937. Use of a discriminant function for differentiating soils with different azotobacter populations. Iowa State Coll. J. Sci. 11, 323–332.
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.
FAO, 1998. World reference base for soil resources. World Soil Resources Report No 84. FAO, Rome.
Fernández-Getino, A.P., Hernández, Z., Piedra Buena, A., Almendros, G., 2010. Assessment of the effects of environmental factors on humification processes by derivative infrared spectroscopy and discriminant analysis. Geoderma 158, 225–232.
Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.
Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Ann. Stat. 19 (1), 1–82.
Geissen, V., Kampichler, C., López-de Llergo-Juárez, J.J., Galindo-Acántara, A., 2007. Superficial and subterranean soil erosion in Tabasco, tropical Mexico: development of a decision tree modeling approach. Geoderma 139, 277–287.
Hansen, M., Dubayah, R., DeFries, R., 1996. Classification trees: an alternative to traditional land cover classifiers. Int. J. Remote Sens. 17, 1075–1081.
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.
Hirmas, D.R., Graham, R.C., Kendrick, K.J., 2011. Soil-geomorphic significance of land surface characteristics in an arid mountain range, Mojave Desert, USA. Catena 87, 408–420.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, Second edition. John Wiley and Sons, New York.
Houben, P., 2008. Scale linkage and contingency effects of field-scale and hillslope-scale controls of long-term soil erosion: anthropogeomorphic sediment flux in agricultural loess watersheds of Southern Germany. Geomorphology 101, 172–191.
Jungmann, M., Kopal, M., Clauser, C., Berlage, T., 2011. Multi-class supervised classification of electrical borehole wall images using texture features. Comput. Geosci. 37, 541–553.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., 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.
King, D., Bourennane, H., Isambert, M., Macaire, J.-J., 1999. Relationship of the presence of a non-calcareous clay-loam horizon to DEM attributes in a gently sloping area. Geoderma 89, 95–111.
Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., Barthes, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101, 105–118.
Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture Comparison of regression tree and Random Forest models. Geoderma 170, 70–79.
Loh, W.Y., Vanichsetakul, N., 1988. Tree-structured classification via generalized discriminant analysis. J. Am. Stat. Assoc. 83, 715–728.
Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., Francaviglia, R., 2011. Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data. Catena 85, 267–273.
May, R., Van Dijk, J., Wabakken, P., Swenson, J.E., Linnell, J.D.C., Zimmermann, B., Odden, J., Pedersen, H.C., Andersen, R., Landa, A., 2008. Habitat differentiation within the large-carnivore community of Norway's multiple-use landscapes. J. Appl. Ecol. 45, 1382–1391.
Mitchie, D., Spiegelhalter, D.J., Taylor, C.C., 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York (298 pp.).
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.
Müller, D., Zeller, M., 2002. Land use dynamics in the central highlands of Vietnam: a spatial model combining village survey data with satellite imagery interpretation. Agric. Econ. 27, 333–354.
Munoz, J., Felicisimo, A.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15, 285–292.
Rhemtulla, J.M., Mladenoff, D.J., Clayton, M.K., 2007. Regional land-cover conversion in the U.S. upper Midwest: magnitude of change and limited recovery (1850–1935–1993). Landsc. Ecol. 22, 57–75.
Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146, 138–146.
Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a dessert landscape. Ecol. Model. 181, 1–15.
Suring, L.H., Goldstein, M.I., Howell, S.M., Nations, C.S., 2008. Response of the cover of berry-producing species to ecological factors on the Kenai Peninsula, Alaska, USA. Can. J. For. Res. 38, 1244–1259.
Tabachnick, B.G., Fidell, L.S., 1996. Using Multivariate Statistics. Harper Collins, New York.
Taylor, J.A., Coulouma, G., Lagacherie, P., Tisseyre, B., 2009. Mapping soil units within a vineyard using statistics associated with high-resolution apparent soil electrical conductivity data and factorial discriminant analysis. Geoderma 153, 278–284.
Tomassone, R., Danzart, M., Daudin, J.J., Masson, J.P., 1988. Discrimination et Classement. Masson, Paris.
Varol, M., Gökot, B., Bekleyen, A., Şen, B., 2012. Spatial and temporal variations in surface water quality of the dam reservoirs in the Tigris River basin, Turkey. Catena 92, 11–21.
Venkataraman, K., Uddameri, V., 2012. Modeling simultaneous exceedance of drinking-water standards of arsenic and nitrate in the Southern Ogallala aquifer using multinomial logistic regression. J. Hydrol. 458–459, 16–27.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.
Webster, R., Burrough, P.A., 1974. Multiple discriminant analysis in soil survey. J. Soil Sci. 25, 120–134.