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 25 January 2014

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:

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

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:

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

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

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