Available online at www.sciencedirect.com
ScienceDirect
Geoderma 143 (2008) 180–190
GEODERMA
www.elsevier.com/locate/geoderma

Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context

Clovis Grinand, Dominique Arrouays *, Bertrand Laroche, Manuel Pascal Martin

INRA, InfoSol, US 1106, Avenue de la Pomme de Pin, BP 20619, Ardon, F-45166 Olivet Cedex, France

Received 18 December 2006; received in revised form 22 July 2007; accepted 3 November 2007
Available online 3 December 2007

Abstract

This paper aims to investigate the potential of using soil-landscape pattern extracted from a soil map to predict soil distribution at unvisited location. Recent machine learning advances used in previous studies showed that the knowledge embedded within soil units delineated by experts can be retrieved and explicitly formulated from environmental data layers However, the extent to which the models can yield valid prediction has been little studied. Our approach is based on a classification tree analysis which has underwent a recent statistics advance, namely, stochastic gradient boosting. We used an existing soil-landscape map to test our methodology. Explanatory variables included classical terrain factors (elevation, slope, curvature plan and profile, wetness index, etc.), various channels and combinations of channels from LANDSAT ETM imagery, land cover and lithology maps. Overall classification accuracy indexes were calculated under two validation schemes, either taken within the training area or from a separated validation area. We focused our study on the accuracy assessment and testing of two modelling parameters: sampling intensity and spatial context integration. First, we observed strong differences in accuracy between the training area and the extrapolated area. Second, sampling intensity, in proportion to the class extent, did not largely influence the classification accuracy. Spatial context integration by the use of a mean filtering algorithm on explanatory variables increased the Kappa index on the extrapolated area by more than ten points. The best accuracy measurements were obtained for a combination of the raw explanatory dataset with the filtered dataset representing regional trend. However, the predictive capacity of models remained quite low when extrapolated to an independent validation area. Nevertheless, this study offers encouragement for the success of extrapolating soil patterns from existing soil maps to fill the gaps in present soil map coverage and to increase efficiency of ongoing soil survey.

Keywords: Digital soil mapping; Boosted classification tree; Extrapolation; Sampling intensity; Validation; France

1. Introduction

The increasing amount of numerical data combined with fast development of new information processing tools change significantly the way in which information on soils is acquired and managed. The use of Digital Soil Mapping (DSM), based on geographical information science, statistics and pedology (McBratney et al., 2003), is continuously increasing. The generic framework of Digital Soil Mapping has been defined by McBratney et al. (2003) as scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method. It is based on the seven predictive scorpan factors, a generalisation of Jenny's five factors, namely: (1) *s*: soil, other or previously measured attributes of the soil at a point; (2) *c*: climate, climatic properties of the environment at a point; (3) *o*: organisms, including land cover and natural vegetation; (4) *r*: topography, including terrain attributes and classes; (5) *p*: parent material, including lithology; (6) *a*: age, the time factor; (7) *n*: space, spatial or geographic position. Interactions between these factors are also considered. Digital Soil Mapping has been tested in a wide range of soil and scale mapping contexts throughout the world (McBratney et al., 2003; Grunwald, 2006; Dobos et al., 2006). It has been used to understand and quantify the relationships between soils and their environmental attributes, mostly derived from exhaustive and easy-to-access datasets such as Digital Elevation Models (DEM) and remote sensing imagery. Recent soil landscape predictive algorithms such as neural networks, fuzzy logic or tree model tools arose mainly from data-mining and

* Corresponding author. Tel.: +33 238 41 48 02.
E-mail address: dominique.arrouays@orleans.inra.fr (D. Arrouays).

0016-7061/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2007.11.004

machine learning fields, also referred to as knowledge discovery in a database in its overall process (Fayyad et al., 1996).

Digital soil mapping is computer-assisted production of a digital map of soil type and soil properties (Dobos et al., 2006). Predictive soil-landscape mapping usually involves two modelling approaches: (i) grouping observations (pixels) that present homogenous attributes without prior knowledge (unsupervised classification), or (ii) training the model on known soil type observations (supervised classification). Irvin et al. (1997) compared those techniques for soil-mapping purposes. Supervised approaches usually produce more suitable results as they use available knowledge (training data) to fit the model and thus lead to more interpretable maps. As related by Lagacherie (2002), in a supervised classification process one makes the implicit hypothesis that a soil-landscape is structured in such way that it can be repeatedly predicted from a specific combination of soil-forming factors. Therefore, a forthcoming question is to what extent this hypothesis is true? An attempt to digitally identify a relevant area for extrapolation was carried out by thresholding a measure of distance to delineate a representative area where the model could be applied, based on elevation, slope and geology layers (Lagacherie et al., 2001). Another approach involved the rule induction process to detect a relevant physiographic region using entire reference areas as single separate target classes (Bui and Moran, 2003). A key issue is to have an independent and representative test sample to get relevant estimates of the extrapolation accuracy (Dobos et al., 2006). Land cover mapping by remote sensing faces similar issues (Foody, 2002). More generally, soil landscape prediction from existing maps involves recovering the mental model used by the soil surveyor to set up the map (Lagacherie et al., 1995; Bui, 2004). This is a reverse soil mapping process and has broad relevance to any other application of knowledge discovery from natural resource maps (Qi and Zhu, 2003).

Classification tree analysis (CTA) is a modelling technique that is being used increasingly (Lawrence et al., 2004). CTA has several advantages that seem to suit well soil-landscape modelling applications. One of the most interesting features is that they are non-parametric, which means that no assumption is made regarding variable distribution. Thus, it avoids variable transformation caused, for instance, by bi-modal or skewed histograms, which are frequent in soil class signatures. They are non-sensitive to missing data, perform automatic variable subset selection, are not sensitive to the inclusion of a large number of irrelevant variables, and finally, they can handle quantitative and categorical data, making it possible to integrate DEM-derived variables, remote sensing bands or indexes together with geology or land cover categorical layers. Efficiency of using CTA for predictive soil landscape mapping was demonstrated in a few studies at regional and subregional scale (Moran and Bui, 2002; Scull et al., 2005). Recent studies showed their potential for land cover mapping from remote sensing images analysis (Lawrence et al., 2004; Friedl and Brodley, 1997) and geomorphological mapping (Luoto and Hjort, 2005).

As mentioned by Luoto and Hjort (2005), CTA was practically used in two linked but distinct purposes: induction and prediction. Induction-oriented studies used CTA to uncover the relationship between soil units or properties and environmental attributes, to

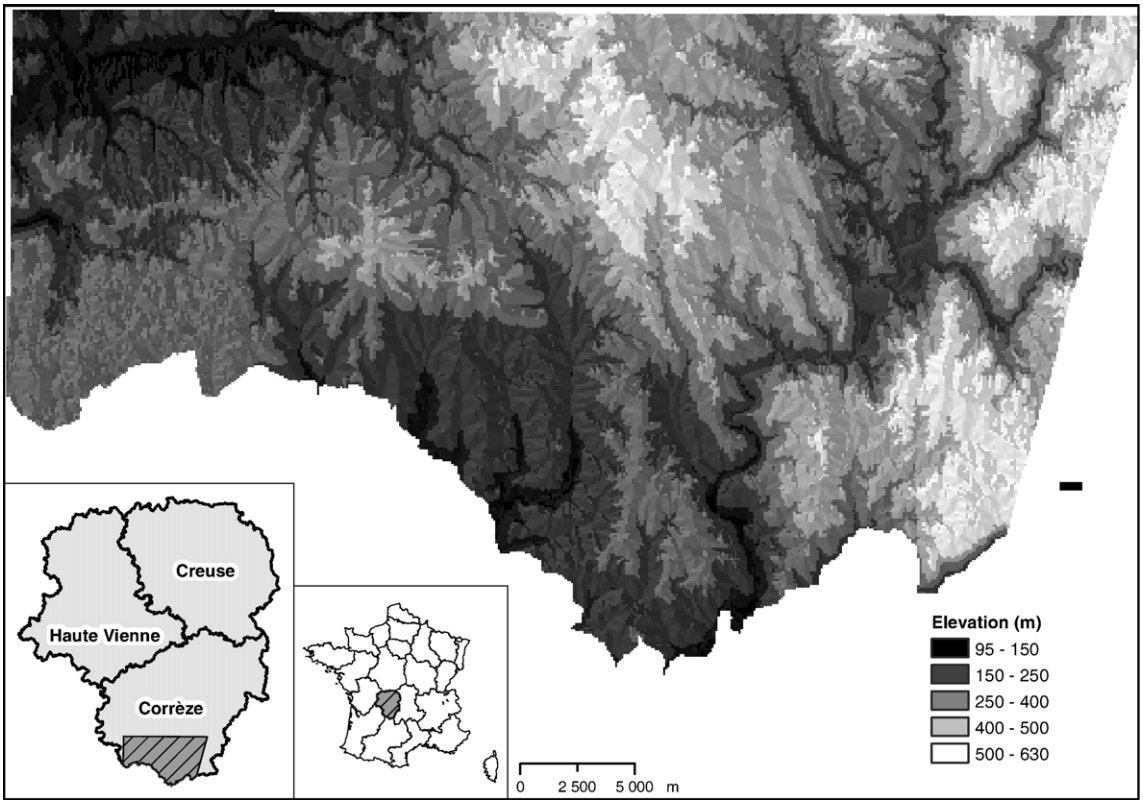

Fig. 1. Location of the study area.

identify the discriminant variables and to compare rules determined by the model with expert knowledge-based rules (McKenzie and Ryan, 1999; Qi and Zhu, 2003; Bui, 2004; Bui et al., 2006). On the other hand, prediction-oriented studies used quantitative relationships between the soil response variables and the environmental soil-forming factors to predict soil landscape patterns over unvisited areas (Lagacherie et al., 1995, Moran and Bui, 2002, Scull et al., 2005).

In this study, CTA is implemented with a recent statistic advance, namely, stochastic gradient boosting (Freidman, 1999). The boosted tree model is called Multiple Additive Regression Tree (MART). Boosting aims to recursively create trees focusing on the most misclassified observations, using a weighting system. This technique showed significant improvements in the classification accuracy compared to unboosted classification and regression trees (Freidman, 1999). The purpose of this paper is to evaluate the ability of this model to provide accurate soil landscape prediction at an unsampled area. Thus, we focus our analysis on a comparative study of two validation procedures. In the first validation procedure, we separate two pixel populations within a training area. The calibration of MART model is made on one population and the validation on the other one. In the second validation procedure, we test the predictive capacity of MART on pixels located outside the training area. Under this validation scheme, we also discuss the effect of sampling intensity and the effect of integration of spatial context.

2. Method

2.1. Study area

The study area is located in France, on the western part of the "Massif Central" (Fig. 1). It covers 90 000 ha and represents 25 soil landscape units mapped at 1/250 000 scale. These soil-landscape units were derived from a synthesis of a pre-existing soil map at 1/100 000 scale (Bonfils, 1976). The density of soil observations by auger borings was ca 1/50 ha. This region is characterised by a high morphological and parental material heterogeneity. Four subregions can be identified. The Brive Basin is located north-west to south, mainly below 250 m of elevation, covering 41% of the study area. In this subregion, the parent materials are mainly on sandstone, marl and, more locally sand and gravel from alluvial terraces. The calcareous Causse plateau (7%) is a relatively flat area of 300 m mean elevation, located in the south-western part of the reference area. The large eastern part of the study area (the Limousin plateau), ranging from 250 m to 630 m, is characterised by gently undulating relief, and includes metamorphic (41%) and eruptive (10%) soil-forming lithological contexts.

2.2. Digital data

Relief attributes were derived using a DEM at a resolution of 50 m from the French Geographical Institute (IGN, 2006). Parent materials were derived from a 1/250 000 lithological map synthesised from all geological surveys available over the entire region. Spectral reflectance in various wavelengths represents an integrated response of soil-forming environment from bedrock to vegetation (Dobos et al., 2000). Hence, one scene of Landsat Enhanced Thematic Mapper (ETM) acquired on 9th November 2000 and covering the study area was downloaded (University of Maryland, 2005). A land cover map derived from satellite image interpretation and provided by the freely available Corine Land Cover 2000 European database was also used (Commission of the European Community, 1993). We did not use climatic data in this study because meteorological stations were rather

Table 1
Data used and derived geographical layers

Name	Scale Resolution	Nature of the soil variable derived	Name	Description	Scorpan name	Type
Target variable						
Soil map	1/250 000	Reference area	pedo	25 classes representing soil-landscape units of Brive-la-gaillarde region	S_c	categorical
Environnementals dataset		Nature of soil-forming factors derived				
DEM	50 m	Topography	alt	Elevation (m)	R	quantitative
			slope	Local slope angle (%)	R	
			curvv	Profile curvature	R	
			curvh	Plan curvature	R	
		Hydrology	cti	Compound Topographic Index	R	
			dppr	Relative hydrological distance to the nearest river (m)	R	
			hppr	Relative height to the nearest river (m)	R	
		Spectral reflectance	band1-band7	All bands of the Landsat ETM scene except the panchromatic band and band 6	O, S, P	quantitative
Landsat ETM	30 and 60 m		band6	Thermal Infra-red band of Landsat ETM	C	
		Vegetation	ndvi	Normalized Difference Vegetation Index	O	
Geological Map	1/250 000	Lithology	geol	Lithological units (15 classes)	P	categorical
Corine Land Cover 2000	1/100 000	Land Cover	clc	Third level Corine Land Cover description (16 classes)	O	Categorical

Scorpan factors: R: relief, O: Organisms, S: soil, P: parent material, Sc: soil classes (McBratney et al., 2003).

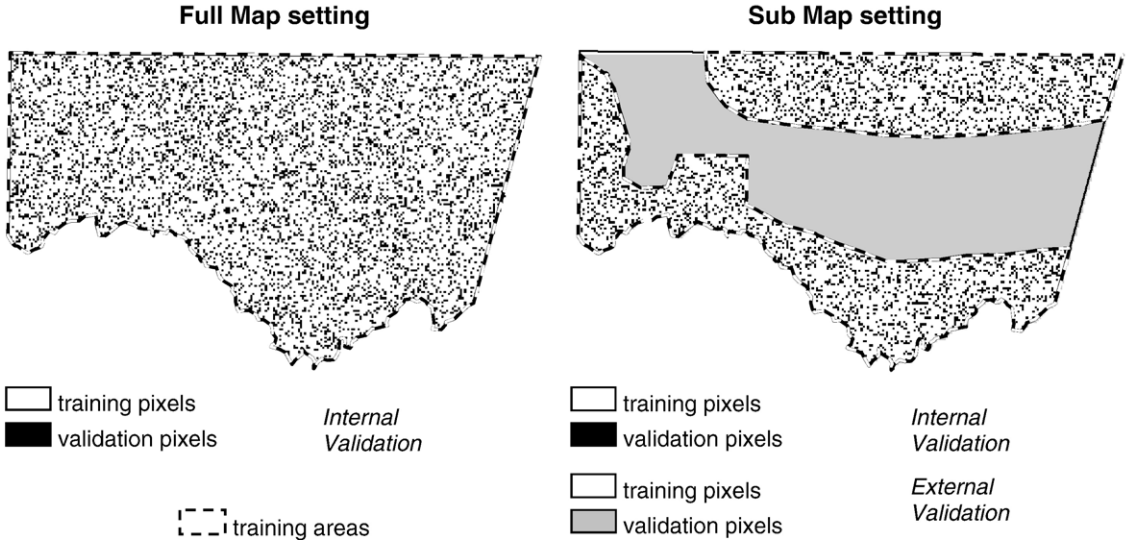

Fig. 2. Validation procedures. In the full map setting (left) training and validation pixels that were sampled within whole map (internal validation). In the sub map setting (right) the training area is restricted to dotted areas. The internal validation is applied on the training area, the external validation is applied on the grey area.

few. Moreover, the main factor controlling climate in this region is elevation, which was retained in the MART model.

2.3. Boosted classification tree

Initially developed by Breiman et al. (1984), the Classification and Regression Tree algorithm (CART) was first adjusted to predictive soil mapping from a reference area by Lagacherie (1992). It allows the training dataset to be split recursively into increasingly homogeneous subsets. Each split is defined as a set of conditional rules based on the explanatory variables. The tree created is usually very large with multiple terminal nodes, meaning that the model is intimately fitted on the training data (Lagacherie et al., 1995). This adverse effect is called overfitting. To avoid overfitting, trees can be pruned back by tuning a splitting rule algorithm. One of the most interesting features of CART is that it gives quantitative insight into the dataset by stating explicit splitting rules. Besides, relatively important variables can be pointed out by counting the times the variable was used in nodes (Bui et al., 2006). However, inconsistencies within the training dataset, such as noise or outliers, can greatly affect the classifier's accuracy (Lagacherie and Holmes, 1997).

Recent statistical advances were implemented on decision tree models, namely stochastic gradient boosting (Freidman, 1999). Boosting is an iterative process that looks at errors from previous classifier steps to decide how to focus on the next iteration over data (Freund and Schapire, 1996). The implementation of boosting in the tree model used here is an "off-the-shelf" application called Multiple Additive Regression Tree (MART) coupled with *R* statistical software (R Development Core Team, 2005). The MART model iteratively builds small trees (often six to eight nodes) from a randomly sampled fraction of observations until satisfactory and stable error rate is achieved. At each iteration, the tree is built to best fit the errors associated with the linear combination of previous trees. In the final classifier, each observation is classified according to the most common classification among the trees (Lawrence et al., 2004). Recent studies identified advantages of using boosted trees compared with simple trees: improvement of accuracy (Moran and Bui, 2002; Lawrence et al., 2004), little tuning needed, and high robustness (Friedman and Meulman, 2003).

2.4. Data pre-processing

The data collected were used to derive environmental variables taken as surrogates of soil-forming factors (Table 1).

Table 2
Number of available training and validation pixels for both internal and external validation procedures

Soil-Landscape Units	Full Map	Sub Map			
	Training pixels	Training pixels		Validation pixels	
UC1	6627	5506	*83%*	1121	*17%*
UC2	2629	1355	*52%*	1274	*48%*
UC3	788	552	*70%*	236	*30%*
UC5	5296	3056	*58%*	2240	*42%*
UC6	5281	2146	*41%*	3135	*59%*
UC7	3998	1583	*40%*	2415	*60%*
UC8	4157	1864	*45%*	2293	*55%*
UC9	7536	3010	*40%*	4526	*60%*
UC10	5361	5061	*94%*	300	*6%*
UC11	1321	455	*34%*	866	*66%*
UC12	1273	254	*20%*	1019	*80%*
UC13	1708	725	*42%*	983	*58%*
UC14	4260	1861	*44%*	2399	*56%*
UC15	6520	4813	*74%*	1707	*26%*
UC17	347	334	*96%*	13	*4%*
UC20	1662	799	*48%*	863	*52%*
UC21	2691	1527	*57%*	1164	*43%*
UC23	5744	2666	*46%*	3078	*54%*
UC24	2288	327	*14%*	1965	*86%*
UC26	2033	860	*42%*	1173	*58%*
UC28	1240	923	*74%*	317	*26%*
UC29	2576	2230	*87%*	346	*13%*
UC30	746	623	*84%*	123	*16%*
UC31	2894	2599	*90%*	295	*10%*
TOTAL	78,976	45,129	*57%*	33,851	*43%*

In this study, we considered easy to derive and interpret terrain parameters often used in digital soil mapping applications (Dobos et al., 2006). Pre-processing was carried out using ArcGis/ArcInfo GIS platforms (ESRI, 2005). DEM-derived variables include surfaces of first derivative (slope) and secondary derivatives (profile and plan curvatures). These variables are direct descriptors of the landforms representing morphological attributes of the terrain. Secondary attributes were calculated so as to characterise specific hydrological processes that were identified to be important soil-forming discriminators. The compound topographic index (CTI), also referred to as wetness index, was derived from the slope and upslope contributing area (Wilson and Gallant, 2000). Further "raster-based" computation was carried out to create surface representing the height of each pixel above the closest waterway and the distance along the flow direction to its closest river outlet. These two layers are thought to provide information on the surface drainage. Each band of Landsat ETM was taken as a separate input variable. Vegetation status and biomass were also integrated through the normalised difference vegetation index (NDVI), a well-used vegetation index based on red (r) and near infrared (nir) bands (NDVI=(nir−r)/(nir+r)). Lithology and land cover vector maps were converted into "pixel-based" maps. We chose a working pixel size of 100 m which is considered to be a relevant and detailed enough geographical support for a 1/250 000 predictive mapping scale (Dobos et al., 2006). Derived terrain parameters and remote sensing bands and indices were resampled according to this spatial support, using the nearest neighbour resampling method. Other resampling techniques could have been suitable (e.g., bilinear sampling, cubic convolution), but they were not tested as the focus of our study was not to test resampling techniques, but to study sampling intensity and validation procedure effects.

We considered that, within the training area, soil unit boundaries were subjected to positional errors due to the "expert-based" delineation and its legibility constraints. A positional error of 100 m at soil unit boundaries was applied to remove uncertain pixels from the training process, using a buffer operation. Following this process, the soil landscape map was then converted into a raster map. The final database (Table 1) was composed of 17 environmental variables and one target variable representing 25 soil landscape units.

2.5. Accuracy assessment procedure

The basis of the validation techniques we used is to exclude a fraction of the sample from the modelling process and to compare the predicted value of these samples with their reference value (Foody, 2002). This accuracy assessment is summarised in the error matrix or contingency table. It was used in this study to derived two classification accuracy indexes: overall accuracy rate and the Kappa index. The former is a simple ratio between the correctly allocated number of pixels (confusion matrix diagonal) and the overall number of classified pixels. The Kappa index is a robust index which takes into account the probability that a pixel is classified by chance (Girard and Girard, 1999). The Kappa index is, therefore, always slightly lower than the overall accuracy measurement.

As mentioned by Muchoney and Strahler (2002) there are two underlying issues when defining "statistically valid independent estimate of accuracy": spatial autocorrelation and representativeness of sampling. These authors considered that samples within the same site, in our case, taking validation pixels from the same area than the training one, cannot be considered as being independent as those pixels are autocorrelated. Therefore, such "pixel-based" ways of separating the training and validation sample could lead to biased estimates of accuracy. They suggested that sample independency can be achieved when two samples are separated by a certain distance.

Thus, we applied two distinct validation procedures in order to assess to what extent the classifier provided accurate prediction within and outside the training area. These sampling procedures are referred to as internal and external validation respectively. The former uses training and validation pixels that were sampled within the training area, whereas the latter uses geographically distinct training and validation areas. These validation schemes are illustrated in Fig. 2. The internal validation scheme was first applied on the whole map (scheme referred as "full map setting in Fig. 2, left), to test the effect of sampling intensity (see Section 2.6). Then, we separated training and validation areas (scheme referred as "submap setting" in Fig. 2, right), to test both validation schemes. To separate the training area from the validation one, we manually divided the existing map so that all lithological formations were present in both areas.

2.6. Sampling Intensity

Sampling design and size is of major importance in the accuracy assessment process (Foody, 2002). One must set up the location and the intensity of sampling to capture the whole variability of the classes, in order to get a representative class signature over the study area. However, practical constraints often limit the realisation of this statistical requirement (Foody, 2002). As recommended by Moran and Bui (2002), we applied a random weighted area sampling scheme. This procedure samples a number of pixels within each class, proportional to the extent of the soil landscape units. The underlying hypothesis below this sampling scheme is that more individuals are needed to characterise large units than smaller units. Besides, this method makes sure that small units are not under-represented, which may be the case with simple random or systematic sampling.

We investigated whether an appropriate sampling intensity existed, that is to say a trade-off between having too much training data leading to a high-computational fitting exercise and too little data with the risk that the model is unable to characterise adequately soil landscape relationships. We tested a range of sampling intensity from 10 to 90% of the class extent. For each sampling intensity value, the modelling process was performed thirty times in order to mitigate and assess the random sampling effect. We did not study lower sampling intensities as the main objective of this study was to test the MART model for DSM using an already mapped training area. Accuracy measurements were carried out on the full map and the submap setting schemes (Fig. 2) according to the validation

procedures presented previously (see Section 2.5). Overall available numbers of samples are presented in Table 2.

2.7. Spatial context integration

Classification trees, as many classification algorithms, have no mechanisms to integrate spatial relationships within the model (Moran and Bui 2002). However, environmental attributes at neighbouring locations can be of great interest in predicting soil pattern (McBratney et al., 2003). Therefore, some authors investigated the potential of incorporating local neighbourhood information into the training pixels using convolution filtering operations (Switzer, 1980; Schetselaar et al., 2000; Moran and Bui 2002). Filtering is achieved by passing a floating window

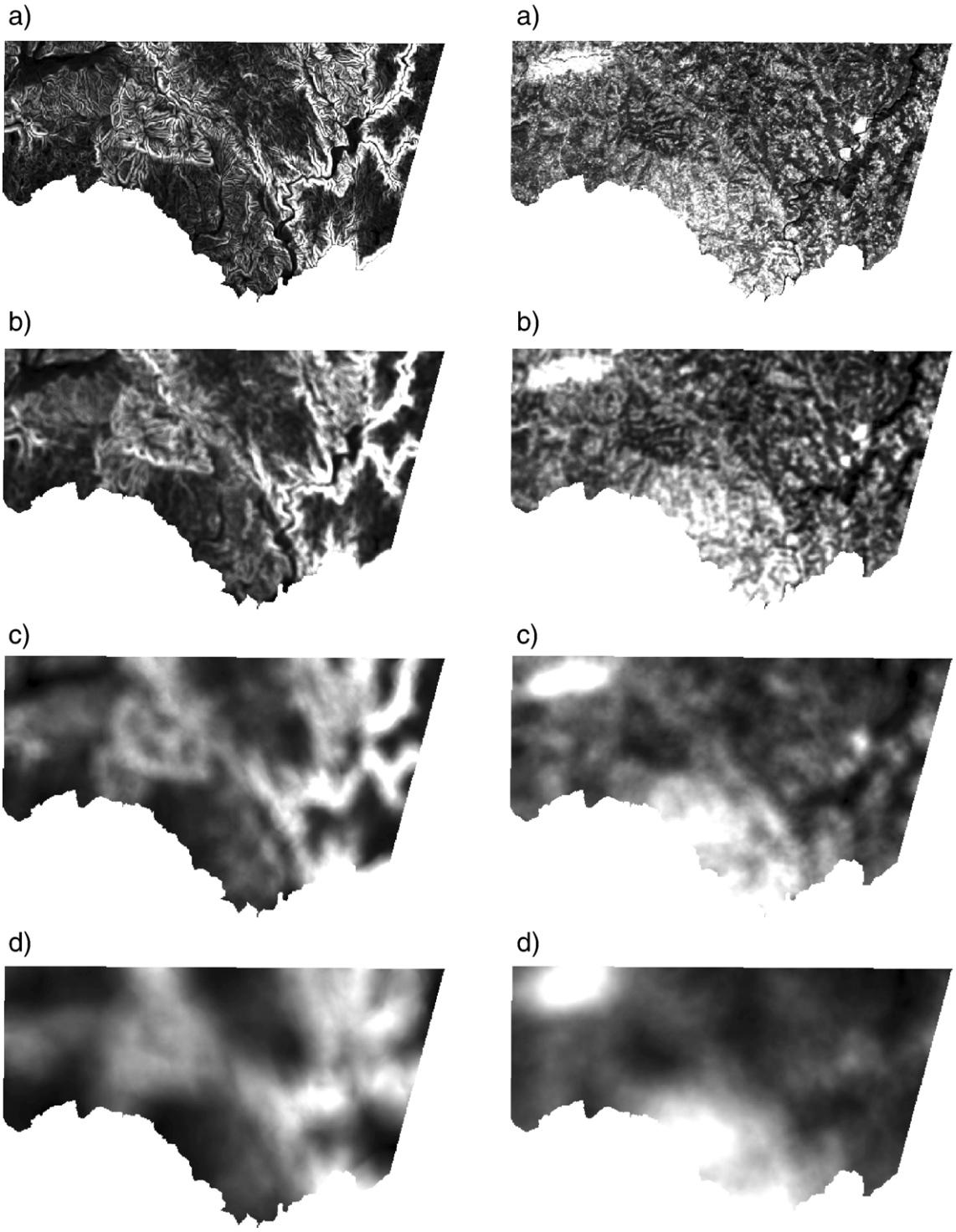

Fig. 3. Illustration of the integration of the spatial context by mean filtering. Slope factor (on left) and thermal infrared canal of Landsat ETM (right) for a) raw data, b) data filtered with a radius of 300 m, c) data filtered with a radius of 1200 m, d) data filtered with a radius of 2500 m.

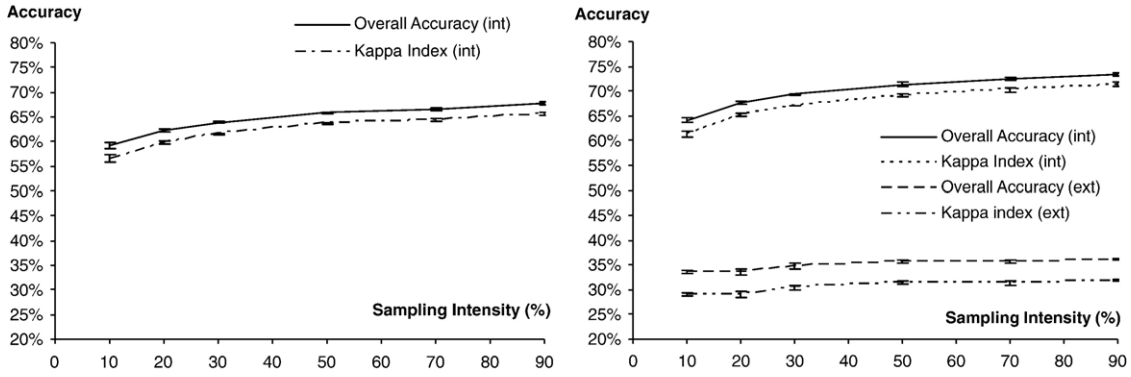

Fig. 4. Overall estimates of accuracy obtained for the test on sampling intensity. Int: internal validation, ext: external validation.

over the variable to calculate a value of the processing cell (central pixel) using the values of its neighbouring cells (Bonn and Rochon, 1992). Filtering requires setting up two essentials parameters: the size and the type of the floating window. Moran and Bui (2002) showed that adaptively filtered data combined with raw data can improve the overall accuracy of the classification from 49 to 70%. In this study, we explore the interest of using mean convolution circular windows for filtering. The test is carried out by increasing the floating windows radius. A visual inspection of the filtering output conducted us to choose five radius values: 300, 700, 1200, 2500 and 5000 metres. An illustration of spatial filtering effect on variables is presented in Fig. 3.

2.8. prediction accuracy and soil scape variability

In order to assess the effect of soil scape variability on classification rate, moving windows of 10×10 pixels were applied on the validation area. By overlaying the windows and the existing soil scape map, we calculated the number of "real" soil scapes per window, and plotted it versus the number of well classified validation pixels.

3. Results

3.1. Sampling intensity

In the full map setting, overall accuracy measurements increase with increasing number of training pixels (Fig. 4a). This is an expected result as the more data is used to fit the model, the better the soil class variability is represented and the better the prediction on juxtaposed test pixels. Overall accuracy ranges from 59% to 67%, which means for this latter value that 2 test pixels out of 3 have been allocated to the correct class. The largest increase in predictive capability occurs between 10 and 20% of the training data and evolves thereafter linearly with increasing proportion of training data. We note a small difference of about 2% between the overall accuracy and the Kappa index. This observation points out the low probability of allocating pixels to correct classes by chance. The highest standard deviation measurement is observed for 10% sampling intensity value. This is consistent in the sense that when few pixels are sampled, the more the random sampling effect is expressed. However, this standard deviation is fairly low (0.7), showing that the predictive ability of the model is quite stable even for low sampling intensities.

In the submap setting, both internal and external validations procedures were carried out. In the internal mode, we note a similar pattern as observed in the full map setting, even though both overall accuracy and Kappa index are levelled up at 5.4% on average (Fig. 4b). Using an external validation scheme induces a marked decrease on overall accuracy and Kappa index. Mean values of overall accuracy (34%) and Kappa index (30%) suggest that the predictive ability of the models is quite unsatisfactory as only one pixel out of three may be correctly classified in the average. Higher discrepancies are observed between overall accuracy and Kappa index suggesting that chance factor has a larger influence. However, random sampling

Table 3
Classification accuracy measurements using raw input dataset or filtered dataset using different radius sizes

Nature of variables	Floating window radius (meters)	Integration area (ha)	Internal validation		External validation	
			Overall Accuracy (%)	Kappa index (%)	Overall Accuracy (%)	Kappa index (%)
RAW			69	67	35	30
FILTERED	300	28	81	80	37	33
	700	154	85	84	37	33
	1200	452	87	86	40	36
	2500	1963	88	87	38	34
	5000	7854	88	87	37	33

Table 4
Classification accuracy measurements using both raw input dataset and filtered dataset using different radius sizes

Combination of variables	Internal validation		External validation	
	Overall accuracy (%)	Kappa index (%)	Overall accuracy (%)	Kappa index (%)
RAW + FILTERED (r = 300)	78	77	37	33
RAW + FILTERED (r = 700)	85	83	38	34
RAW + FILTERED (r = 1200)	87	86	42	38
RAW + FILTERED (r = 2500)	89	88	42	38
RAW + FILTERED (r = 5000)	89	88	44	40

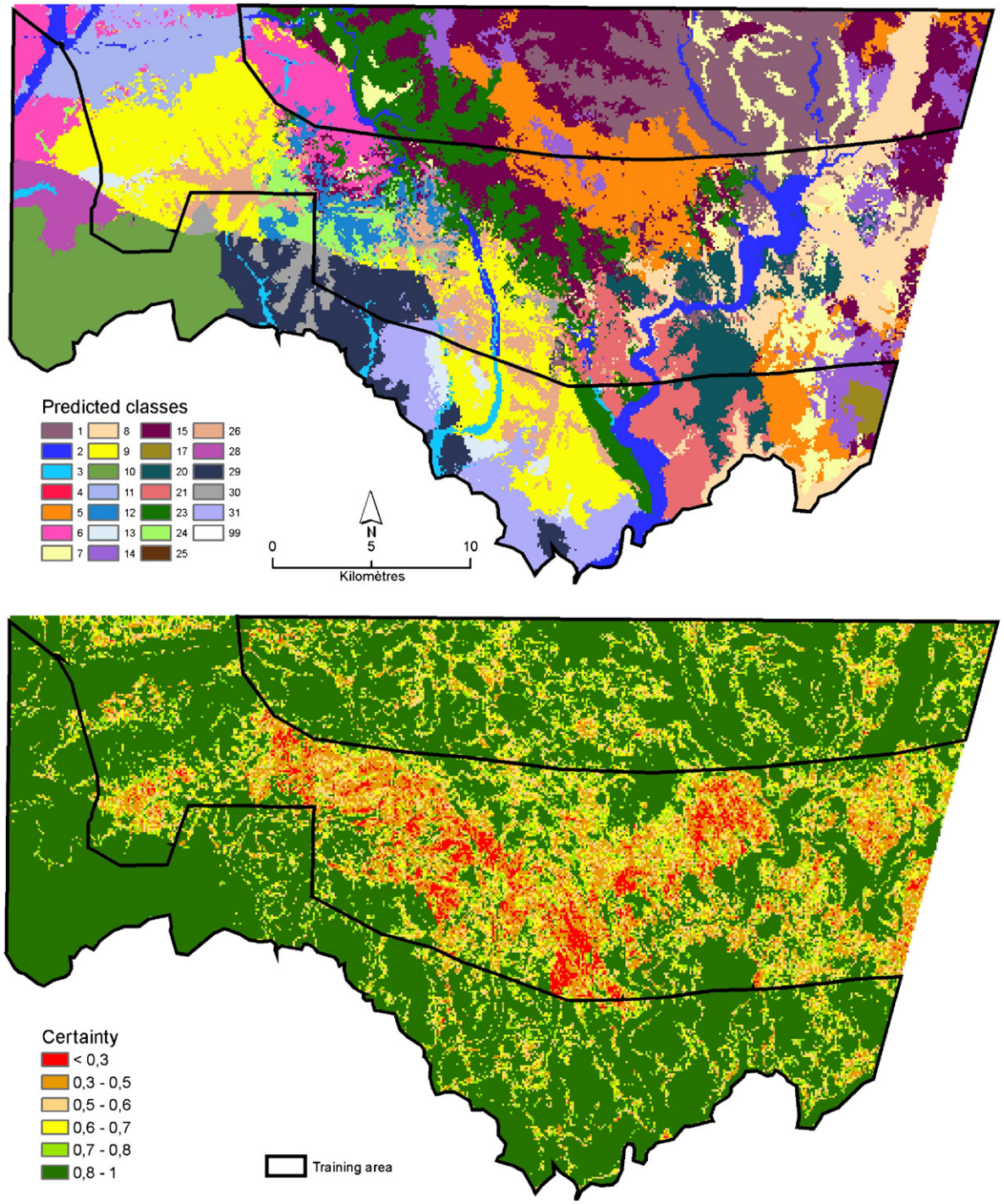

Fig. 5. Output class prediction and probability (external validation, sampling intensitity of training data of 30%, raw dataset + filtered data with a radius of 5000 m.

effect still does not appear to affect classification results significantly.

Sampling intensities higher than 30% did not increase the prediction accuracy outside the training area. The important gap between internal and external validation results raises concerns about the ability of the classification tree model to predict classes that are not spanned within the domain of the training data. Further modellings were carried out using a 30% sampling intensity value in the submap setting only.

3.2. Spatial context integration

We first analysed the effect of using the spatially filtered input dataset compared to using the raw dataset (Table 3). For the internal validation design, there was a clear improvement of prediction accuracy. Although overall accuracy with the raw dataset did not exceed 70%, using the filtered dataset yielded values ranging from 81% fora small floating window size up to 88% for large windows. This is an illustration of spatial

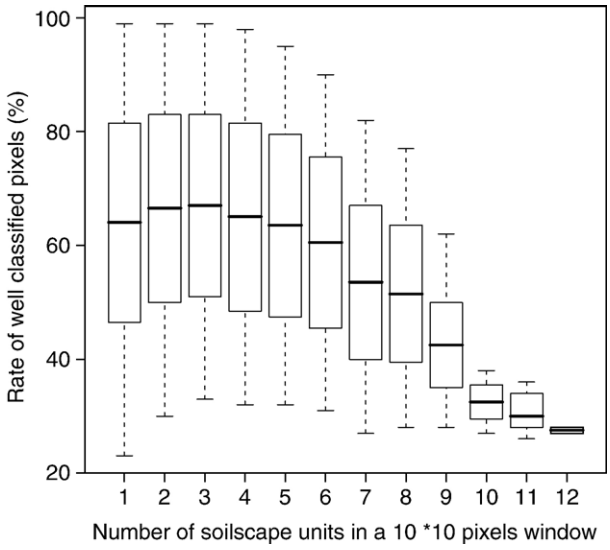

Fig. 6. Boxplot of the rate of well classified pixels vs number of soilscape units in a 10×10 pixel window.

autocorrelation effect when training and validation datasets are sampled from the same area. On the contrary, when using external validation, only a slight improvement was observed by filtering compared to using the raw dataset. The greatest prediction accuracy (40%) was achieved for a filter radius of 1200 m. The best results were obtained when both raw and filtered explanatory variables were combined (Table 4). This result highlights the interest of mixing fine scale variables of high local variability with the same variables transformed to subregional trends.

3.3. Extrapolation

Output maps from the best model are presented in Fig. 5. We observed large differences between noise distribution in the predicted soil pattern for the training and validation areas. The low noise rate on training areas, confirmed by the numerical classification accuracy measurements, proved that the model retrieved relevant and accurate soil landscape relationships. On the other hand, the high level of noise at certain locations within the validation area rendered soil patterns hardly distinguishable. This suggests that noise density can be used as a spatial indicator of prediction accuracy. The probability map, considered as an expression of prediction uncertainty, gives coherent information to this respect. Indeed, we found a strong relation between pixels of high uncertainty their location with mixed soil distribution (Fig. 6). In a soil survey approach, this could be an important support in building an efficient soil sampling scheme, concentrating soil sampling in uncertain prediction locations.

4. Discussion

The results presented in this paper give insight into the way the knowledge contained in existing soil maps can be used to predict soil landscape patterns at unvisited areas for regional soil mapping purposes.

4.1. Sampling intensity

Sampling intensity did not influence the classification accuracy observed in the training area. As observed by Moran and Bui (2002), rather low sampling rates were sufficient to capture the entire class variability. This statement is from an overall point of view as class homogeneity or purity does vary from class to class. Very low sampling rates (<10%) were not tested because our aim was not to simulate field sampling but to test the use of a pre-existing soil map for training; however, we suspect that accuracy would have decreased substantially. Indeed, in an analogy with conventional soil survey, the lowest tested value of 10% would represent a field sampling density of 1 sample per 10 ha, which is far more intense than classical soil sampling recommendations in a 1/250 000 scale mapping context (Finke et al., 2001).

4.2. Spatial context integration

Spatial context integration by the use of a mean filtering algorithm on explanatory variables increased the Kappa index on the extrapolated area by more than ten points. The best accuracy measurements were obtained for a combination of raw explanatory dataset with the same dataset filtered with a circle integration window of 5 km radius. The overall accuracy increased as the window radius increased. This surprising result could reflect the effect of a global regional trend (i.e. elevation) on soil spatial distribution. The spatial prediction on training areas displayed a substantial decrease of noise compared with using raw variables only. Our approach remains rather simple in the sense that a fixed radius window was used on terrain-derived variables and remote sensing bands with no assumptions regarding local variability of each explanatory variable, compared to the approach of Moran and Bui (2002), who used variograms to adjust the size of the filter. However, our results address the issue raised by McKenzie and Ryan (1999) about the mismatch of scales between the target variable and the explanatory variables. The authors mentioned that the appropriate level of details of variables should depend on the processes controlling soil formation which are strongly landscape dependent. When dealing with existing soil maps, the level of detail described by the derived variables may not be the one that was integrated in the mental model of the surveyor that delineated soil units. Thus, we found that spatial filtering was an efficient means for integrating various scales of explanatory variables.

4.3. Validation procedures and DSM perspectives

The validation procedures that were used for both sampling intensity and spatial context integration tests revealed the importance of the location of the test pixels when evaluating the predictive ability of the model. Indeed, we observed a gap of 40% of overall accuracy between internal and external

validation schemes. This obviously faces the issue of test pixels independence. The underlying question is to determine whether spatial autocorrelation between training and test pixels occurs which, in consequence, may overestimate the classification accuracy. The internal accuracy is likely to be prone to overestimation as training and test samples are located close to each other. Then, spatial filtering increases spatial autocorrelation and thus increases classification accuracy. On the other hand, external accuracy may be underestimated as test pixel takes into account only a subpopulation of the soil class.

These accuracy errors due to spatial autocorrelation were also discussed by Friedl et al. (2000) and Muchoney and Strahler (2002) for land cover classification. Internal and external classification can be referred to respectively as pixel and site-based calibration/validation used by these authors and more generally by the remote sensing community. The low external classification accuracies we obtained are consistent with values reported by these authors, i.e. around 50% for the overall accuracy.

We observed in this paper that "true" accuracy can only be approximated correctly if test samples are collected at a certain distance from the training samples.

Recent papers (Dobos et al., 2000; Moran and Bui, 2002; Scull et al., 2005) outlined the trend of predictive soil models to shift from research to operational phase. This trend leads to more and more enlarged study areas and consequently to an increasing number of soil units involved in the modelling process. Therefore, fitting models is more and more challenging due to a larger number of soil classes, and larger predicted areas are usually linked to a decrease of thematic and spatial accuracy. Nevertheless, this trend meets the regional and national needs such as the 1/250 000 French and European soil mapping programmes (Finke et al., 2001).

5. Conclusion

This study provides new insight into the way knowledge from soil maps could be retrieved to perform extrapolation. We applied a methodology based on the boosted classification tree model that follows three main steps: pre-processing, modelling, and extrapolation. We focused our study on the accuracy assessment and testing of two modelling parameters: sampling intensity and spatial context integration. First, we observed strong differences in accuracy between the training area and the extrapolated area. Second, sampling intensity, in proportion to the class extent, did not largely influence the classification accuracy. Spatial context integration by the use of a mean filtering algorithm on explanatory variables increased the Kappa index on the extrapolated area by more than ten points. The best accuracy measurements were obtained for a combination of the raw explanatory dataset with the filtered dataset representing regional trend. However, the predictive capacity of models remained quite low when extrapolated to an independent validation area. Nevertheless, this study offers encouragement for the success of extrapolating soil patterns from existing soil maps to fill the gaps in present soil map coverage and to increase efficiency of ongoing soil survey.

Acknowledgements

The soil mapping programme of France is granted by the French Ministry for Agriculture. We thank Michel Baffet for his advice on local soil geography. We thank the two anonymous reviewers for their helpful comments.

References

Bonfils, P., 1976. Carte pédologique de France à 1/100 000. Feuille de Brive. 1 carte et 1 notice explicative. Institut National de la Recherche Agronomique, Versailles, France.

Bonn, F., Rochon, G., 1992. Précis de télédétection, Volume 1- Principes et méthodes. Presses de l'Université du Québec/AUPELF, Sainte-Foy.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression trees- Wadsworth & Brooks Wadsworth statistics/probability series.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17–26.

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties. Ecol. Model. 191, 431–446.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111, 21–24.

Commission of the European Community, 1993. CORINE Land Cover. Bruxelles, Belgium (Data download at http://dataservice.eea.europa.eu/dataservice last accessed 11/26/2007).

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97, 367–391.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., Tòth, G., 2006. Digital Soil Mapping as a support to production of functional maps — EUR 22123 EN. Office for Official Publication of the European Communities, Luxemburg.

ESRI® (Environmental Systems Research Institute), 1994–2005. http://www.esri.com.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence Press, Menlo Park. pp. 1–34.

Finke, P., Hartwich, R., Dudal, R., Ibanez, J., Jamagne, M., King, D., Montanarella, L., Yassoglou, N., 2001. Bases de données géoréférencée des sols pour l'Europe. Manuel de Procédures version 1.1 — EUR 18092 FR. European Community, JRC Ispra, Italy.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.

Freidman, J.H., 1999. Stochastic gradient boosting. Technical Report. Department of Statistics, Stanford University.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148–156.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemilogy. Stat. Med. 22, 1365–1381.

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.

Friedl, M.A., Woodcock, C., Gopal, S., Muchoney, D., Strahler, A.H., Barker-Schaaf, C., 2000. A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data. Int. J. Remote Sens. 21 (5), 1073–1077.

Girard, M.C., Girard, C.M., 1999. Traitements des données de télédétection. Dunod, Paris. 529p.

Grunwald, S., 2006. Environmental Soil-Landscape Modelling, Geographic Information technologies and Pedometrics. Taylor and Francis Group.

IGN (Institut National Géographique), 2006. BDCarto. http://www.ign.fr/.

Irvin, B.J., Ventura, S.J., Slater, K.B., 1997. Fuzzy and isodata classification of landform elements form digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77, 137–154.

Lagacherie, P., 1992. Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme

référence. Mémoire de Thèse, Université de Montpellier. Institut National de la Recherche Agronomique, France. 175p.

Lagacherie, P., 2002. Cartographie des sols et de leurs propriétés a un niveau sub-régional. UMR INRA-ENSAM Sol et Environnement, Montpellier, France. 48p.

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geogr. Inf. Sci. 11, 183–198.

Lagacherie, P., Legros, J.P., Burrough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65, 283–301.

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.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sens. Environ. 90, 331–336.

Luoto, M., Hjort, J., 2005. Evaluation of current statistical approaches for predictive geomorphological mapping. Geomorphology. 67, 299–315.

McBratney, A.B., Mendonça, M.L., Minasny, B., 2003. On digital Soil Mapping. Geoderma 117, 3–52.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

Moran, J.M., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16 (6), 533–549.

Muchoney, D.M., Strahler, A.H., 2002. Pixel-and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. Remote Sens. Environ. 84, 290–299.

Qi, F., Zhu, A.X., 2003. Knowledge discovery from soil maps using inductive learning. Int.J. Geogr. Inf. Sci. 17 (8), 771–795.

R Development Core Team., 2005. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schetselaar, E.M., Chung, C.J.F., Kim, K.E., 2000. Integration of Landsat TM, Gamma-Ray, Magnetic, and Field Data to discriminate Lithological Units in Vegetated Granite-Gneiss Terrain. Remote Sens. Environ. 71, 89–105.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model. 181, 1–15.

Switzer, P., 1980. Extensions of Linear Discriminant Analysis for Statistical Classification of Remotely Sensed Satellite Imagery. Math. Geol. 12 (4), 367–376.

University of Maryland, 2005. Global Land Cover Facility, University of Maryland, URL: http://glcf.umiacs.umd.edu/index.shtml, (last date accessed: 22 may 2006).

Wilso, P.J, Gallan, C.J (Eds.), 2000. Terrain analysis:principles and applications. John Wiley & Sons, Ltd., New York. 303p.