Geoderma 146 (2008) 138–146

Contents lists available at ScienceDirect

Geoderma

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

Instance selection and classification tree analysis for large spatial datasets in digital soil mapping

Karsten Schmidt *, Thorsten Behrens, Thomas Scholten

Institute of Geography, Chair of Physical Geography, Eberhard Karls University Tübingen, Rümelinstraße 19-23, D-72074, Tübingen, Germany

ARTICLE INFO

Article history:
Received 19 October 2007
Received in revised form 25 April 2008
Accepted 15 May 2008
Available online 18 June 2008

Keywords:
Digital soil mapping
Classification trees
Instance selection
Grid learning
Pruning
Random sampling
Spatial data mining

ABSTRACT

Digital soil mapping is currently experiencing a tremendous increase in available environmental covariates and resolution for spatial soil predictions, resulting in computational problems in terms of limited data handling capabilities of machine learning approaches. This is of particular importance when gridded spatial soil class maps are used as a basis for predictions containing large amounts of redundant instances and noisy information.

In this study we systematically analyze the effect of instance selection, which aims at reducing sample size, while preserving or even increasing prediction accuracy. On a soil class dataset with 95,000 instances we tested two sampling approaches in relation to parameter settings of decision tree based learning: proportional and disproportional stratified random sampling. An automated grid search approach was used to find the best performing parameter settings of the decision tree.

The results show that an appropriate sampling method in combination with a grid search method returns better results than those obtained when grid learning is applied without instance selection. Instance selection increases prediction accuracy especially if the frequency distribution of the soil classes is low compared to the surrounding area. However, instance selection does not help in pedological interpretation. Nevertheless, it is a valuable pre-processing method to handle large spatial high resolution datasets in digital soil class prediction in terms of accuracy and computational costs.

As suggested on the basis of the results of this study, spatially constrained instance selection as well as boundary based digital soil mapping in terms of soil taxonomic contrast should be investigated in future pedometric research.

1. Introduction

Within the last decade an increase in the availability of explanatory variables to solve the function of soil forming factors (Jenny, 1941), reformulated for quantitative empirical modelling by McBratney et al. (2003), can be observed and is expressed in a dramatic increase of research publications in this field of soil science (McBratney et al., 2000, 2003). This is promising, as the global lack in the availability of soil data can now be counteracted using digital soil mapping approaches. However, data handling becomes more complex due to the high amount of available potential predictors as well as the large amount of pixels when existing rasterized soil maps are used as a basis for extrapolation. Thus, the lack of predictors is now replaced by the limitation of the data handling capabilities of the algorithms used for prediction, which limits the analysis of soil formation and distribution.

One of the most widely used and best performing inductive learning algorithms in terms of generating interpretable rules as well as prediction accuracy are classification tree algorithms (Breiman et al., 1984; Loh and Vanichsetakul, 1988; Mitchie et al., 1994; Behrens and Scholten, 2006b). Classification trees are non-parametric (Friedman, 1991; Mitchie et al., 1994), non-sensitive to the presence of missing data and to the inclusion of a large number of irrelevant features (Schafer, 1997; Hastie et al., 2001), and are described as a robust prediction technique (Loh and Vanichsetakul, 1988; Lagacherie et al., 2001; Scull et al., 2005). Applications in environmental sciences can thus be found in various disciplines like ecology (e.g. Geng et al., 2004; Munoz and Felicisimo, 2004), remote sensing (e.g. Hansen et al., 1996; Friedl and Brodley, 1997; Debeir et al., 2001; Gómez-Chova et al., 2003; He et al., 2003) and soil science (e.g. Lagacherie and Holmes, 1997; Zhang et al., 1999; Bui et al., 1999; Giasson et al., 2000; Moran and Bui, 2002; Bui and Moran, 2003; Zhou et al., 2004; Behrens and Scholten, 2006a,b; Geissen et al., 2007).

Yet, handling large datasets using decision trees can be inefficient in terms of learning time and prediction accuracy due to redundant and noisy information and often results in complex models (Liu and Motoda, 1998; Brighton and Mellish, 2002). This is particularly true when using gridded spatial datasets as a basis for predictions (Munoz and Felicisimo, 2004). In contrast to datasets originating from point

* Corresponding author. Tel.: +49 7071 29 77523.
E-mail address: karsten.schmidt@uni-tuebingen.de (K. Schmidt).

0016-7061/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2008.05.010

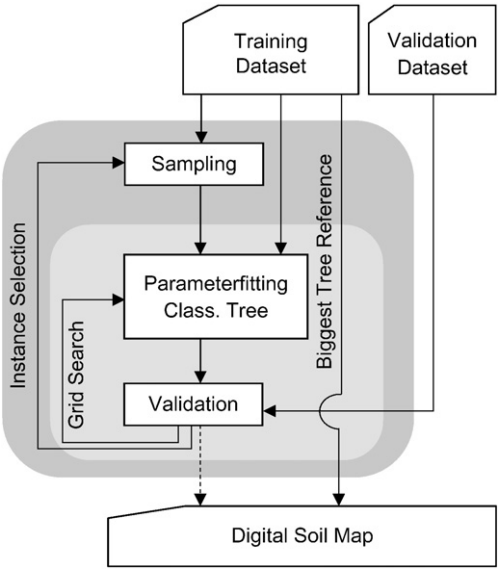

Fig. 1. Overview of methods.

sampling schemes, mostly counting only up to a view hundred samples, digital soil class prediction based on existing soil class maps can easily comprise several thousands to millions of training "samples" if the prediction is based on gridded spatial datasets where every pixel serves as a separate sample (Bui et al., 1999; Shrestha et al., 2004; Behrens et al., 2005). This leads to large amounts of redundant information, causing negative effects not only with regard to computation time but also to prediction accuracy (Qi, 2004). Lagacherie and Holmes (1997) showed that inconsistencies within the training dataset, such as noise, can greatly influence predictive accuracy.

To handle datasets containing redundant and/or noisy instances (samples) as well as multi-collinearity two main branches within statistical learning research do exist: instance selection (Liu and Motoda, 2001) and feature selection (John et al., 1994).

Feature selection aims at reducing the feature space to the driving factors and at reducing multi-collinearity. In contrast, instance selection is applied to reduce the dataset by fitting out the relevant samples. The application of feature and/or instance selection depends on the structure of the dataset and the learning approach used. As the dataset used in this study contains large amounts of samples, and as decision trees are robust to correlated features, we focus on instance selection.

In pedometric research the discussion on instance selection has been limited. Even though some methods of classical spatial soil sampling designs (Cochran, 1977; Brus, 1993; Brus and de Gruijter, 1997; Domburg et al., 1997) and instance selection (Gu et al., 2001) are based on the same theoretical concepts, the aim of instance selection is contrary to the aim of soil sampling. In soil sampling it is most important to optimize sampling schemes to derive a sample set that is as sparse as possible to save labor and lab costs. In instance selection, more than enough samples are available. The challenge here is to extract a representative subset that is still large enough that no relevant information gets lost but is small enough that it can be handled easily by learning algorithms.

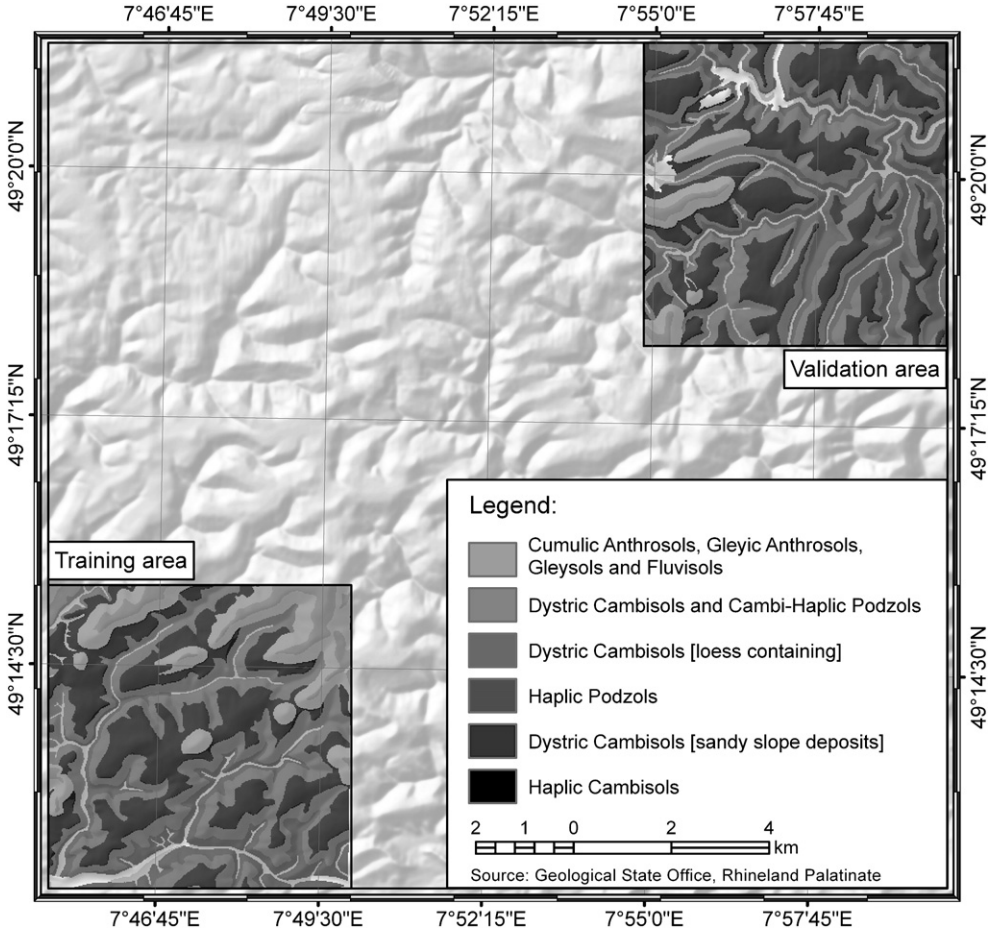

Fig. 2. Location of the investigation area in the Palatinate Forest, Germany.

Table 1
Description and coverage of the 6 soil types according to World Reference Base for Soil Resources (WRB) (IUSS Working Group, 2006) and the corresponding reference ID

Soil classes	ID	Coverage [%]
Haplic Cambisols	SC1	33
Dystric Cambisols and Cambi-Hablic Podzols	SC2	13
Dystric Cambisols [loess containing]	SC3	5
Haplic Podzols	SC4	33
Dystric Cambisols [sandy slope deposits]	SC5	12
Cumulic Anthrosols, Gleyic Anthrosols, Gleysols, and Fluvisols	SC6	4

Moran and Bui (2002) were the first to examine instance selection in a digital soil mapping approach by comparing two random sampling methods applied over all soil classes of the entire training dataset. In this study we go a step further in systematically analyzing instance selection on the basis of single soil classes, which is important to compare the outcome of different spatial sample distributions and relations to soil forming factors. Another approach reported by Qi (2004) addresses the task of noise reduction by selecting samples based on fitted histograms of the predictors. However, this approach is not applicable for large feature spaces (many predictors) as each feature has to be analyzed separately.

In this study we combine instance selection with an investigation of reasonable parameter settings for classification trees. The analysis of interactions between sample sizes, sample method, and tree parameters aims at stable digital soil mapping models with reduced complexity, faster computation times, as well as easier interpretability.

2. Rationale

In this study we systematically compare different random sampling designs and sample sizes to analyze the impact of instance selection on a digital soil class mapping task. As different model parameters of decision tree algorithms are directly related to sample size, i.e. the minimum samples in an end node as well as the cross validation pruning settings (Breiman et al., 1984; Mitchell, 1997), it is important to analyze these parameter settings in relation to instance selection approaches. In order to automatically test different parameter settings to find optimized model parameters, a so called grid search (grid- or hyper-learning) approach (cf., Gourieroux and Monfort, 1995; Gilardi and Bengio, 2001; Jin, 2006) is applied. In grid search procedures different parameter settings are systematically analyzed in p-dimensional parameter spaces (cf. Section 3.3).

The approach presented in this paper is illustrated in Fig. 1 and organized as follows:

First, biggest trees (no cross validation pruning) are calculated as references to analyze the effect of overfitting (Behrens and Scholten, 2006b) and for the subsequent analysis on parameter and instance selection settings.

Second, different cross validation pruning settings are tested to analyze different parameter settings for the entire dataset to understand the behavior of the tree models in terms of overfitting and to obtain an additional reference to interpret the effect of instance selection.

Third, different instance selection approaches and sample sizes are analyzed on the basis of the same grid search approach as applied on the entire dataset to examine the impact on prediction accuracy.

All approaches are computed for binary (2-class) predictions, i.e. each soil class is tested individually to gain a deeper insight about the possible variations of results and as a basis for pedological interpretations. Furthermore, tree complexity in terms of the number of end nodes is discussed.

3. Materials and methods

3.1. Study area and datasets

The investigation area, a low mountain range within the south–west German–Lorraine Triassic escarpment (Rhineland–Palatinate; Germany), comprises 350 km^2 (Fig. 2).

The prediction approach is based on 40 terrain attributes derived from a digital elevation model with a resolution of 20 m (Behrens et al., 2005; Behrens and Scholten, 2006b). The local soils have formed from substrates of Upper Red Bed Sandstone, Bunter and Lower Limestones. The training as well as the validation area, each comprising 40 km^2, contain the same 6 soil classes (SC1–SC6) mapped at a scale of 1: 50 000 (Table 1). According to the DEM the soil map was rasterized to a resolution of 20 m. To compare the training with the validation area Fig. 3 shows feature spaces for elevation, slope, aspect, and curvature. As differences in feature spaces are marginal this setting serves as an optimal test bed for induction based learning (Behrens and Scholten, 2006b).

As data mining based predictions strongly rely on predictors and their feature spaces, predictions can only reasonably be extrapolated to areas where the same features with a similar feature space are available, i.e. similar soilscapes (Hole, 1978; Lagacherie et al., 2001).

As the study area is located in a low mountain range, relief plays an important role in pedogenesis in terms of erosion and accumulation. However, due to differences in parent material as well as a stratification with almost no dip, geology has a strong impact on soil formation in this region. Thus, including information on parent material in the prediction approach will lead to very high accuracies that will make the differences in prediction and instance selection approaches hard to interpret. Additionally, in terms of a "real world" example using digital elevation models only is more realistic as digital elevation models are widely available whereas large to medium scale geological maps are not (Behrens et al., in press). Hence, geological information was not included in this study as well as other spatial

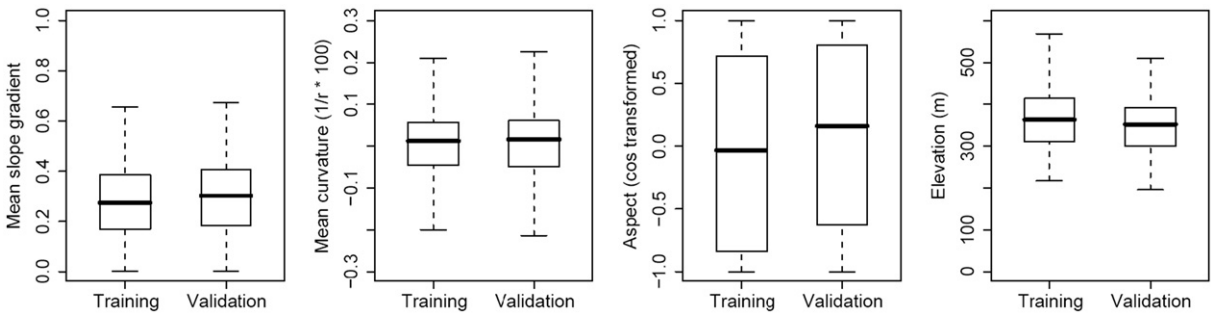

Fig. 3. Feature spaces of slope, curvature, aspect, and elevation of training and validation area.

predictors. Other ancillary variables like remote sensing or climate data were not available within this study.

3.2. Instance selection

Instance selection is a technique to reduce the size of a dataset by extracting relevant information, where the new subset resembles the original dataset in further analysis (Pal and Jain, 2005). Instance selection is generally applied for three technical reasons: *enabling*, *focusing*, and *cleaning* (Liu and Motoda, 2002) described below:

Enabling: As the capability to handle large datasets is limited for every data mining algorithm, instance selection enables these algorithms to function and work effectively (Moran and Bui, 2002; Grinand et al., 2008; Behrens et al., in press).

Focusing: Soil maps are generally not intended for the purpose of applying data mining techniques (Qi, 2004). For example, traditional soil mapping is based on a mental model by a soil surveyor (Bui and Moran, 2003; Bui, 2004) focused on a detailed soil profile description to subsequently construct a map of soil classes at a specific scale. These soil classes generally comprise larger regions with some hundreds to thousands of pixels of an underlying DEM resulting in highly redundant information. Thus, focusing aims at selecting the relevant information.

Cleaning: No soil map is perfect in terms of decision boundaries as well as spatial delineation (Burrough and McDonnell, 1998; McKenzie and Ryan, 1999; Zhou et al., 2004; Behrens et al., in press). A saying in computer science often used in the context of noise in datasets is "garbage-in-garbage-out". In this respect, cleaning a dataset helps removing, or at least reducing noisy data (Qi, 2004; Behrens et al., in press). In contrast to that, high quality soil data will lead to high quality results and reduced computational and labor costs.

Summarizing, the main aim of instance selection is to reasonably reduce large datasets for faster predictions while preserving accuracy (Liu and Motoda, 1998; Bui et al., 1999). Hence, the ideal outcome of instance selection is model independent and can be described as follows:

$$P(\mathrm{MeD}) = P(\mathrm{MsD}), \tag{1}$$

where P(MeD) is the predictive power of a model based on the entire dataset and P(MsD) is the predictive power of a subset (Liu and Motoda, 2001). On the one hand, in an optimistic case, "less is more" (Liu and Motoda, 1998), so that the resulting prediction accuracy based on the subset is higher than the one for the entire dataset. On the other hand, instance selection can lead to a trade-off between sample size and mining quality (Brighton and Mellish, 2002).

In this study on soil class prediction we apply proportional stratified random sampling (PS) and disproportional stratified random sampling schemes (DS) as instance selection approaches (Gu et al., 2001). The soil-sampling counterparts of these two approaches are well-known in spatial soil sampling design aiming to improve estimates (Brus, 1993, Domburg et al., 1997).

Proportional stratified random sampling accounts for the frequency distribution of each soil class in the entire dataset. In the disproportional approach the same amount of instances is selected for all classes. This approach is recommended by Kohonen et al. (1995) for supervised classification applications, even when the a priori probabilities are skewed.

We compare the original dataset comprising 95,000 pixels with 6 different subsets sizes of 500, 1000, 2500, 5000, 7500 and 10,000 samples for both sampling approaches. For each sample size three independent iterations were carried out to provide information on sampling variance. The results are discussed on the average of these three predictions.

3.3. Prediction and validation

3.3.1. Classification tree

Classification tree analysis is a supervised non-parametric statistical classification approach based on binary recursive partitioning techniques (Breiman et al., 1984; Friedman, 1991; Loh and Shih, 1997). The step-wise constant partitioning scheme (Friedman, 1991) provides increasingly homogeneous subsets in terms of the dependent variable—in our case the soil classes—based on existing observations. It is used to identify rules which can subsequently be applied for extrapolation. Partitioning is stopped if a minimum tolerated amount of samples (pixels) in a node of the tree is reached. This threshold influences the size of the tree in terms of end nodes and is thus strongly related to overfitting and generalization (Mitchie et al., 1994). For each terminal node the majority class label is assigned for the final classification results.

To construct a classification tree we apply the CRUISE algorithm as introduced by Loh and Shih (1997). Compared to the greedy search approach of CART (Breiman et al., 1984), CRUISE is based on an analysis of variance which is reported to be stable and unbiased in terms of variable and split point selection (Kim and Loh, 2001).

The 1D CRUISE—algorithm used in this study separates the tree growing process into two steps: first selecting the split variable and second selecting the split point. To select the most important variable for each split an ANOVA F-Test is computed. The variable with the smallest p-value, which determines the significance of the features, is chosen to separate between two classes. Once the variable with the smallest p-value is selected, a threshold that defines the split point is computed applying a linear discriminant analysis (LDA) (Kim and Loh, 2001). As LDA is most effective when the data are normally distributed with the same covariance matrix, a Box–Cox transformation is calculated to adjust the response variable (cf. Qu and Loh, 1992).

3.3.2. Validation

Predicted categorical spatial data are validated using a confusion matrix (Van Rijsbergen, 1979; Ishioka, 2003) by deriving measures of effectiveness such as recall, precision and the F-measure which are important measures for information retrieval (Van Rijsbergen, 1979; Raghavan et al., 1989; Manning and Schütze, 1999; Giasson et al., 2000; Zhu, 2000; Behrens et al., 2005). Table 2 shows a confusion matrix.

Recall (rc) describes the relation between positive (tt) and negative predicted pixels (ft) and thus the probability that a mapped soil class sample (pixel) is actually predicted, i.e. it represents underestimation:

$$\mathrm{rc} = \frac{\mathrm{tt}}{\mathrm{tt} + \mathrm{ft}}, \tag{2}$$

whereas precision (pc) describes the probability that a predicted soil class pixel is actually mapped, i.e. it represents overestimation:

$$\mathrm{pc} = \frac{\mathrm{tt}}{\mathrm{tt} + \mathrm{tf}}. \tag{3}$$

To quantify the overall model quality in a composite measure the F_1-measure (Van Rijsbergen, 1979), representing the harmonic mean

Table 2
Confusion matrix

		Original soil unit	
		True	False
Predicted soil unit	True	tt	tf
	False	ft	ff

of under- and overestimation, is frequently used. It is calculated as follows:

$$F_1 = \frac{2*\mathrm{rc}*\mathrm{pc}}{\mathrm{rc}+\mathrm{pc}}. \quad (4)$$

3.4. Grid-search

Grid-search or hyper-learning (Ensor and Glynn, 1997; Bergez et al., 2004) is a simple, yet time consuming procedure to find the best parameter settings or combinations for fitting a model in a *p*-dimensional parameter space systematically. Wrapped around a decision tree algorithm it leads to *p* trees based on *p* different parameter settings. The most important parameters controlling the tree size as well as overfitting, which consequently influence validation accuracy, are the minimum data in a terminal node (mindat) and the standard error (SE) of a 10-fold cross validation pruning (Breiman et al., 1984). Thus, we apply a 2D grid learning scheme. Based on the size of the soil dataset, the F_1-measure was calculated for each combination of a mindat of $m = 5, 10, 50, 100, 500, 1000, 5000, 10\,000$ and the entire training set size as well as SE-values of SE = 0, 1,…, 10. Additionally, un-pruned (biggest) trees based on a mindat of 5 were calculated as reference models for evaluating each combination of mindat and SE.

3.5. Model complexity

As an additional basis to interpret the model performance, the results of the instance selection approaches as well as the impact of noise and redundancy on the prediction algorithm we investigate model complexity. With regard to tree based approaches the number of end nodes can be used as a simple measure which is mainly determined by three parameters: sample size, mindat, and the cross validation pruning settings. The more samples and the smaller mindat the larger a tree can grow. The more it overfits this way (i.e. learns noise instead of relevant information) the more it will be reduced by the subsequent pruning.

The more end nodes in a tree the more complex the model and the more complex the interpretation of the results (Munoz and Felicisimo, 2004).

We compare the model complexities of all approaches: biggest trees, grid search over the entire dataset, as well as the instance selection approaches with their three underlying random sampling iterations.

4. Results and discussion

4.1. Classification tree analysis

4.1.1. Biggest tree reference predictions

The analysis of the biggest tree models for each soil class shows varying results in terms of predictability (Table 3). Soil classes SC2, SC3 and SC5 show prediction accuracies in the validation area below 0.5 which is due to two major reasons: First, they show a weak relation to relief, and second they contain noisy information in terms of imprecise delineations, resulting from a high degree of taxonomic similarity of soils mainly differentiated by altering Loess contents.

The comparatively high prediction accuracy for these soil classes in the training area serves as a first indicator for overfitting, due to too complex tree models based on noisy data.

The biggest tree results serve as primary reference measures to evaluate the validation accuracy obtained by parameter fitting and/or instance selection.

4.1.2. Grid search

The prediction results over the different settings for mindat and SE as analyzed by the grid search approach are shown in Fig. 4, revealing varying impacts to the different soil classes.

Table 3
Prediction accuracy (F_1) and the number of terminal nodes (nodes) for the biggest tree settings (BT, no cross validation pruning, minimum instances in the end node—mindat = 5), for the optimized tree model parameters SE (Standard Error) and mindat based on a grid search method (GS), for the optimized settings for the tree model parameters (SE and mindat) and sample sizes (S) for proportional stratified random sampling (PS) and disproportional stratified random sampling (DS) averaged over three independent runs

		SC1	SC2	SC3	SC4	SC5	SC6
BT	F_1T	0.81	0.80	0.77	0.73	0.66	0.69
	F_1V	0.72	0.41	0.25	0.52	0.43	0.55
	Nodes	622	358	265	625	477	344
GS	SE	1	0	9	4	2	0
	Mindat	100	5	100	10	1000	5000
	F_1T	0.78	0.81	0.65	0.67	0.50	0.52
	F_1V	0.72	0.42	0.28	0.52	0.48	0.57
	Nodes	62	23	14	45	28	5
DS	SE	4	0	0	1	1	2
	Mindat	5	5	5	10	5	5
	S	10,000	7500	10,000	5000	7500	7500
	F_1T	0.77	0.73	0.65	0.70	0.58	050
	F_1V	0.78	0.36	0.19	0.62	0.61	0.54
	SD_T	0.004	0.017	0.012	0.003	0.004	0.008
	SD_V	0.019	0.010	0.005	0.008	0.005	0.003
	Nodes	30	126	145	47	58	21
PS	SE	0	0	1	0	2	0
	Mindat	1000	5	5	10	50	500
	S	10,000	10,000	10,000	10,000	10,000	7500
	F_1T	0.73	0.74	0.68	0.70	0.54	0.52
	F_1V	0.75	0.40	0.31	0.54	0.47	0.58
	SD_T	0.007	0.029	0.029	0.009	0.004	0,0001
	SD_V	0.005	0.026	0.018	0.008	0.023	0,001
	Nodes	14	48	24	60	22	2
	ΔF_1V [%]	5.49	−2.45	1.24	9.86	12.90	0.53

Based on the optimized settings in sample size, mindat, and SE the variation of the reiterations is expressed through the standard deviation (SD) for both training (T) and validation (V) area.
The information gain for the optimal instance selection method compared to parameter optimization (GS) was finally calculated by the difference (ΔF_1V [%]) between the single validation results, thus positive values indicate better results for instance selection, negative values indicate better results for parameter optimization.

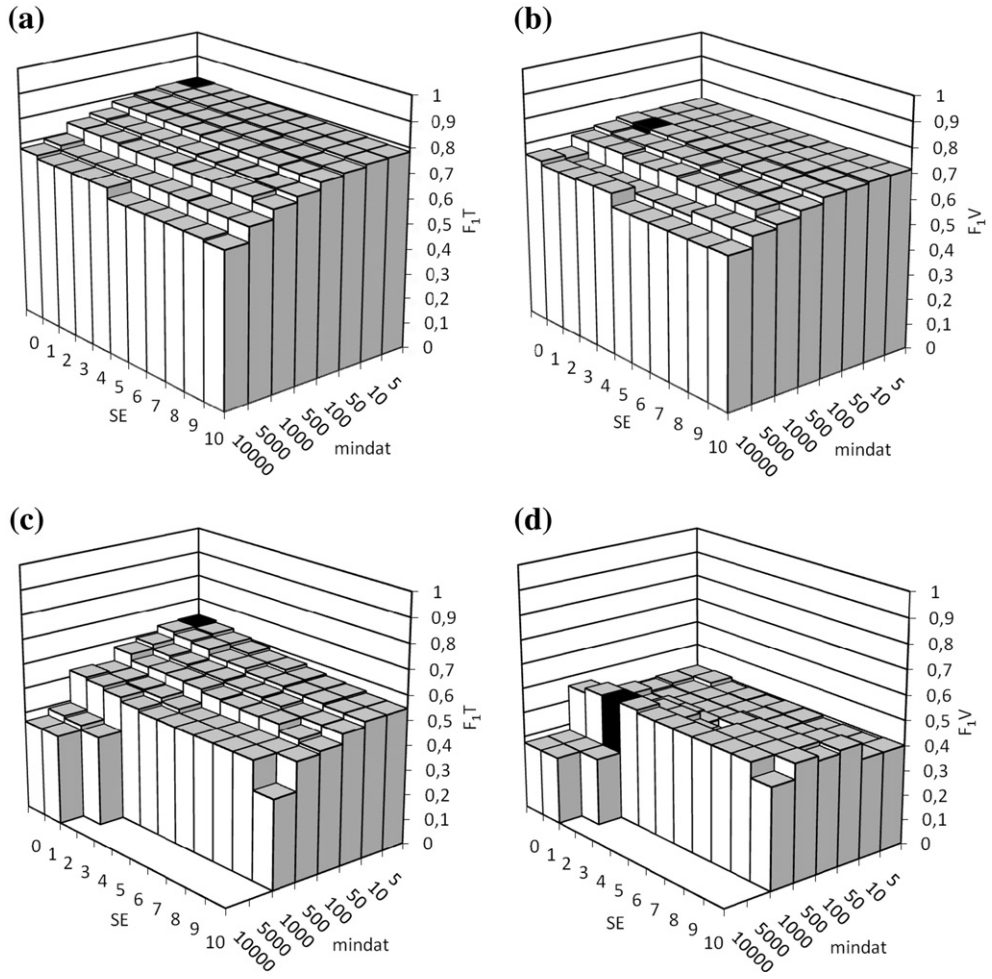

Fig. 4. Prediction accuracy (F_1) for the training (T) and validation (V) area of different model settings for SC1 (a, b) and SC5 (c, d). The best results are marked in black.

It can be seen that SC 1 produces relatively stable results across the parameter space of the grid-learning scheme, whereas SC 5 shows unstable results. As SC1 can be extrapolated accurately into the validation area (Fig. 4b) it seems to be highly correlated to relief and accurately mapped, whereas SC5 shows a decrease in prediction accuracy of 11% between training and validation area. In this case, less complex models in terms of higher mindat return better results in the validation area.

The results shown in Table 3 reveal a high variability of the optimal model settings over the soil classes. The differences between the optimized settings are too high to derive optimal global settings, which might be applied for all soil classes as almost the entire range of mindat and SE occurs. Generally, the differences in prediction accuracy between the soil classes can be related to missing predictors (e.g. geology) or noise in terms of mapping precision (Fig. 5).

Grid search offers the possibility to analyze the effect of overfitting. Generally, overfitting is indicated by large differences between the prediction accuracy in the training and validation areas. Table 3 shows that for all soil classes except SC2 the accuracy in the training area decreases due to the grid search approach. In no case the validation accuracy decreases. Thus, the generalization rate is higher for all soil classes except for SC2. Concerning SC2 it is remarkable that the training accuracy increases even though the computed model is more than 15 times less complex in terms of terminal nodes.

Even if the tree size is generally expected to correlate with the amount of samples provided (Munoz and Felicisimo, 2004), the values for mindat und SE cannot be related to sample size and also for larger soil classes higher values for mindat cannot be found (Tables 1 and 3).

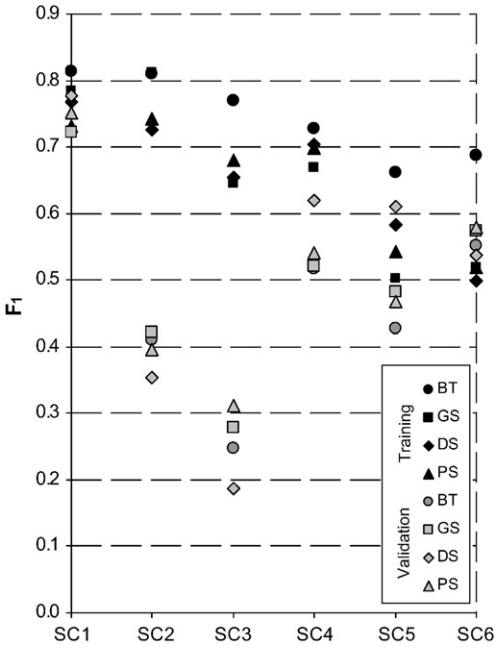

Fig. 5. Optimized prediction results for disproportional stratified random sampling (DS), proportional stratified random sampling (PS) and grid search based optimized parameter settings (GS) for each soil class compared to the biggest tree results (BT) for both training and validation area.

4.2. Instance selection

The optimized sampling sizes for the instance selection approaches range from 5000 up to 10,000 pixels (Table 3). Values below 5000 return lower validation accuracies for all soil classes, indicating that relevant information gets lost. Most remarkably—except for the F_1-measure in the validation area of SC2—at least one of the two sampling approaches returns better results as obtained by the grid learning approach for the entire dataset. For some soil classes (SC4, SC5) the effect of instance selection boosts the validation accuracy up to 13%.

The general relation between sample size and model parameters can be seen by analyzing the optimized values for mindat, which are lower for the reduced sample sets, based on the instance selection approaches than for the grid search approach based on the entire dataset (Table 3). The only exception is SC1-based proportional stratified random sampling, resulting from different settings of the standard error, which are typically higher for the grid search approach based on the entire dataset.

The differences between the instance selection methods PS and DS are partly related to the frequency of the soil classes, such that smaller soil classes return better validation results when proportional stratified random sampling is applied (Fig. 5). This seems contrary to what is expected as less pixels of a small soil class are sampled compared to its surrounding area (Chen et al., 2004).

The high positive impact of instance selection in digital soil class mapping with an average increase of 7% for the F_1-measure in the validation area is shown in Table 4.

This positive effect holds true for all three independent random selections carried out for every prediction in this study, which is shown by low standard deviations in Table 3.

The average variation in prediction accuracy of the three random sampling iterations for the optimized DS settings is 1.6% in the training area and 1.8% in the validation area in terms of the *F*-measure. For PS the corresponding values are 4% and 3%. Thus, variation does not increase in the validation area.

As the variation of the three independent PS samplings is within the range of the average improvement in accuracy achieved for SC2, SC3, and SC6, it can be stated that PS is not very useful in terms of increasing accuracy compared to grid learning solely (Table 3). Nevertheless, it speeds up computation.

Concerning DS, which is superior to PS for SC1, SC4, and SC5, variation is smaller than for PS and—more important—is below the rate of increase in accuracy (Table 3). Thus, for these soil classes DS is recommend both for faster computation and for increasing prediction accuracy. This confirms the finding of Kohonen et al. (1995) who are recommending DS.

Yet, for the some units with the lowest proportion (SC3, SC6) PS provides up to 12% better results than DS. Thus, the recommendation of Kohonen et al. (1995) is not true for extremely skewed a priori proportions or imbalanced data. This is difficult to explain, as boosting soil classes with small proportions via sampling should produce better results at first sight (Chen et al., 2004). An explanation might be that important information of spatially neighbouring regions gets lost (Chen et al., 2004). This is especially the case if soil classes are mapped imprecisely and more information about spatially neighbouring regions is needed to average the noise between the soil class and its surrounding and thus to increase the ability of the model to generalize. This is emphasized by the high mindat of 500 for the optimized PS settings of SC6 compared to a mindat of 5 for DS, indicating overfitting by a too complex model. Thus the problem seems to be correlated to noise rather than spatial coverage. This can be explained soil scientifically for the valley bottom soils (SC6). They are mapped according to their location in the geological map which is problematic because of three reasons: first, even if there are overlaps between geological maps and soil maps, the main purpose of the geological map is not to indicate soil distribution. Second, the valley bottoms in geological maps are mostly generalized and thus do not exactly fit to the location as shown in current DEMs. Third, geological maps for this region are drawn on the basis of older topographical maps resulting again in slightly different spatial delineations. Thus, predictions based on terrain attributes only return weaker results than soil classes mapped on the basis of current topographical maps.

The prediction results might also be affected by missing features. As the formation of SC2 and SC3 is strongly related to geology (Behrens and Scholten, 2006b) information on relief is not sufficient for good predictions. For these classes which produce the worst prediction results PS is again better. The explanation is the same: more information is needed in the direct neighborhood to enable the model to generalize and not to overestimate, as reported by Behrens and Scholten (2006b).

Concluding, if the prediction works quite well and PS produces better results for skewed distributions, the soil class is mapped imprecisely and/or important predictors are missing. Thus, Kohonen et al. (1995) are right for datasets which are not skewed, do not contain noise, but all relevant features. The possibility to overcome this problem and to achieve better generalization results for DS on skewed distributions and noisy datasets might be to use a spatially constrained sampling approach, where more samples are taken near the boundary. This should be based on spatial distance density function schemes to average out the delineation noise in the soil data and therefore to increase validation accuracy. This might lead to a new paradigm in digital soil class mapping focusing on the boundaries instead of concentrating on the more homogeneous cores of the class areas. Following the concept described by Hole (1978), different types of boundaries can be characterized by the taxonomical contrast of the adjacent soil classes. In this concept of nine orders, order 9 represents a boundary separating orders of soil and order 1 represents a boundary separating soil phases. Thus, theoretically, higher order boundaries should be easier to predict as low order boundaries, due to a higher contrast which should be expressed by a higher contrast in state factors. Further studies have to prove this concept.

Table 4
Average prediction accuracy (F_1) of the approaches tested (*T*=training area, *V*=validation area, BT—biggest tree, GS—optimized parameter settings based on a grid search method, S—optimized sampling schemes)

Method	F_1T	F_1V
BT	0.75	0.48
GS	0.66	0.50
S	0.67	0.55

4.3. Model complexity

Model complexity analysis in terms of the number of end nodes offers the possibility to explain why instance selection in combination with grid search approaches produces better results than grid search over the entire dataset only.

Table 3 shows that in average the tree complexities obtained from the grid search approach over the entire dataset are comparable to the instance selection approaches. This is contrary to the general expectation that smaller datasets return less complex models. It can be explained with the different mindat and pruning settings and shows the high dependency between sample size and parameter settings. As random sampling is not a procedure designed to remove noise from a dataset as for example special approaches like Wilson editing (Behrens et al., in press) and as the different randomized sampling tests returned very similar prediction results (cf. Section 4.2) the surprisingly good results of the instance selection approaches are not based on dataset *cleaning* as no noise is removed. As a consequence, this phenomenon must be an effect of *enabling*

(cf. Section 3.2) in terms of removing redundant information. In this case, CRUISE must be regarded unstable in terms of redundancy. More detailed insight can be gained from the three independent random sampling tests. Here we see that the average node size for DS and PS differs. The node size for the approaches where DS returned the best results is similar for the entire dataset, whereas the node size for the approaches where PS returned better results is larger. As PS returns better results for skewed distributions (i.e., the soil class comprises a relatively small area in the training area) it can be stated that the prediction results obtained with CRUISE are only affected by redundancy if the distributions are much skewed. Thus, it is an effect of *focusing*.

In general, even though the tree complexity compared to the biggest tree models is reduced by parameter fitting for about 94% for the entire dataset, as well as 93% and 92% for all DS and PS approaches, the resulting models must be regarded too complex for interpretation in most of the cases. Thus, in this study instance selection does not help easing pedological interpretation. Additional approaches to analyze feature importance might be more efficient for this purpose (John et al., 1994, Liu and Motoda, 1998).

4.4. Summarizing conclusions

This study systematically analyzes the influence of instance selection schemes and grid learning in data mining based digital soil mapping approaches. Based on 3702 tree inductions for 6 soil classes, the results can be summarized and discussed as follows:

Biggest trees generally lead to higher prediction accuracies in the training area and lower generalization rates compared to grid learning and instance selection due to overfitting and effects of enabling and focusing.

Optimizing sensitive parameters increases the generalization rate due to a reduced effect of overfitting which is trivial and as expected (e.g. Breiman et al., 1984).

An appropriate sampling method in combination with grid search returns better results than obtained when grid learning is applied on the entire dataset. This enabling effect is remarkable as the results are better than the ideal outcome of instance selection where the predictive power of a model based on the entire dataset equals the predictive power of a subset (Liu and Motoda, 2001).

Instance selection increases prediction accuracy especially if the frequency distribution of the soil classes is low compared to the surrounding area which must be regarded as an effect of *focusing* (Liu and Motoda, 2002).

For small soil classes (skewed frequencies) proportional sampling returns better results. This effect shows that adaptive sampling in terms of handling characteristic soil distributions and frequencies differently will lead to higher prediction accuracies compared to global approaches (Moran and Bui, 2002).

Based on the analysis of model complexity instance selection does not necessarily boost pedological interpretations.

Instance selection and grid learning are important tools in digital soil class mapping as they both help speeding up computation and increasing prediction accuracy. Yet, it is not easy to recommend a sampling scheme a priori. To handle skewed, noisy, and redundant data in digital soil class mapping based on large datasets additional approaches like Wilson editing (Wilson, 1972; Behrens et. al, in press), latin-hypercube sampling (Carre et al., 2007), and prototype generation (Wai et al., 2001) should be tested in further studies. As suggested on the basis of the results of this study, spatially constrained instance selection as well as boundary mapping in terms of soil taxonomic contrast (Hole, 1978) should be investigated in future pedometric research.

References

Behrens, T., Scholten, T., 2006a. Digital soil mapping in Germany—a review. Journal of Plant Nutrition and Soil Science 169, 434–443.

Behrens, T., Scholten, T., 2006b. 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.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168, 21–33.

Behrens, T., Schmidt, K., Scholten, T., in press. An approach to removing uncertainities in nominal environmental covariates and soil class maps. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer, 440 pp.

Bergez, J.E., Garcia, F., Lapasse, L., 2004. A hierarchical partitioning method for optimizing irrigation strategies. Agriculture Systems 80, 235–253.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression trees. Wadsworth.

Brighton, H., Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6, 153–172.

Brus, D.J., 1993: Incorporating models of spatial variation in sampling strategies for soil. Doctoral thesis. Wageningen. Netherlands.

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. Geoderma 80, 1–59.

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

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darlin basin of Australia. Geoderma 111, 21–44.

Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Australian Journal of Soil Research 37, 495–508.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information System, 2nd ed. Oxford University Press. 356 pp.

Carre, F., McBratney, A.B., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141, 1–14.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. Technical Report, vol. 666. Department of Statistics, University of California, Berkely.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. John Wiley & Sons. 428 pp.

Debeir, O., Latinne, P., van den Steen, I., 2001. Remote sensing classification of spectral, spatial and contextual data using multiple classifier systems. Proccedings 8th ECS Image Analysis, Bordeaux, pp. 584–589.

Domburg, P., de Gruijter, J.J., van Beek, P., 1997. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma 75, 183–201.

Ensor, K.B., Glynn, P.W., 1997. Stochastic optimization via grid search. Mathematics of Stochastic Manufacturing Systems. American Mathematical Society. 399 pp.

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

Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Annals of Statistics 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.

Geng, W., Cosman, P., Berry, C.C., Feng, Z., Schafer, W.R., 2004. Automatic tracking, feature extraction and classification of *C. elegans* phenotypes. IEEE Transactions on Biomedical Engineering 10/51, 1811–1820.

Giasson, E., van Es, C., van Wambeke, A., Bryant, R.B., 2000. Assessing the economic value of soil information using decision analysis techniques. Soil Science 165/12, 971–978.

Gilardi, N., Bengio, S., 2001. Local machine learning models for spatial data analysis. Journal of Geographic Information and Decision Analysis 4/1, 11–28.

Gómez-Chova, L., Calpe, J., Soria, E., Camps-Valls, G., Martin, J.D., Moreno, J., 2003. Cart-based feature selection of hyperspectral images for crop cover classification. IEEE International Conference on Image Processing 3, 589–592.

Gourieroux, Ch., Monfort, A., 1995. Statsitcs and Econometric Models: General Concepts, Estimation, Prediction, and Algorithms. Cambridge University Press, Cambridge. 504 pp.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180–190.

Gu, B., Hu, F., Liu, H., 2001. Sampling: knowing whole from its part. In: Liu, H., Motoda, H. (Eds.), Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Boston. 448 pp.

Hansen, M., Dubayah, R., DeFries, R., 1996. Classification trees: an alternative to traditional land cover classifiers. International Journal of Remote Sensing 17, 1075–1081.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York. 533 pp.

He, P., Fang, K.T., Xu, C.-J., 2003. The classification tree combined with SIR and Ist application to classification of mass spectra. Journal of Data Science 1, 425–445.

Hole, F.D., 1978. An approach to landscape analysis with emphasis on soils. Geoderma 21/1, 1–23.

Ishioka, T., 2003. Evaluation of criteria for information retrieval. IEEE/WIC International Conference on Web Intelligence WI 2003, Sponsored by IEEE Computer Society and Web Intelligence Consortium, Halifax, Canada, pp. 425–431.

IUSS Working Group, 2006. World reference base for soil resources 2006, World Soil Resources Reports No. 103, 2nd ed. FAO, Rome.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York. 281 pp.
Jin, Y., 2006. Multi-Objective Machine Learning. Springer Verlag. 660 pp.
John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. Proceedings of the 11th International Conference on Machine Learning, pp. 121–129.
Kim, H., Loh, W.-Y., 2001. Classification trees with unbiased multiway splits. Journal of the American Statistical Association 96, 589–604.
Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1995. The learning vector quantization program package; version 3.1.Online Puplication http://www.cis.hut.fi/research1995.
Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. International Journal of Geographical Information Science 11, 183–198.
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.
Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Boston. 214 pp.
Liu, H., Motoda, H., 2001. Data reduction via instance selection. In: Liu, H., Motoda, H. (Eds.), Instance Selection and Construction for Data mining. Kluwer Academic Publishers, Boston. 448 pp.
Liu, H., Motoda, H., 2002. On issues of instance selection. Data Mining and Knowledge Discovery, vol. 6. Kluwer Academic Publishers, pp. 115–130.
Loh, W.Y., Shih, Y.S., 1997. Split selection methods for classification trees. Statistica Sinica 7, 815–840.
Loh, W.Y., Vanichsetakul, N., 1988. Tree-structured classification via generalized discriminant analysis. Journal of the American Statistical Association 83, 715–728.
Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press Cambridge, London. 620 pp.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometrics techniques for use in soil survey. Geoderma 97, 293–327.
McBratney, A.B., Mendonça Santos, 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.
Mitchell, T.M., 1997. Machine Learning. Singapore. 414 pp.
Mitchie, D., Spiegelhalter, D.J., Taylor, C.C., 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York. 298 pp.
Moran, J.C., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Science 16/6, 533–549.
Munoz, J., Felicisimo, A.M., 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15, 285–292.
Pal, N.R., Jain, L., 2005. Advanced Techniques in Knowledge Discovery and Data Mining. Springer Verlag, London. 264 pp.
Qi, F., 2004. Knowledge discovery from area-class resource maps: data preprocessing for noise reduction. Transactions in GIS 8/3, 297–308.
Qu, P., Loh, W.Y., 1992. Application of Box–Cox transformations to discrimination for the two-class problem. Communications in Statistics. Theory and Methods 21, 2757–2774.
Raghavan, V., Bollmann, P., Jung, G.S., 1989. A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems 7, 205–229.
Schafer, J., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London. 430 pp.
Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a dessert landscape. Ecological Modelling 181, 1–15.
Shrestha, D.P., Zinck, J.A., Van Ranst, E., 2004. Modelling land degradation in the Nepalese Himalaya. Catena 135–156.
Van Rijsbergen, C.J., 1979. Information Retrieval. Butterworths. 153 pp.
Wai, L., Keung, C.-K., Ling, C.X., 2001. Learning via prototype generation and filtering. In: Liu, H., Motoda, H. (Eds.), Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Boston. 448 pp.
Wilson, D.L., 1972. Asymptotic properties of nearest neighbour rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics 2, 408–420.
Zhang, J., Guo, D., Wan, Q., 1999. Geospatial data mining and knowledge discovery using decision tree algorithm—a case study of soil data set of the Yellow River Delta.—Geoinformatics and Socioinformatics. Proceedings of Geoinformatics'99 Conference, Ann Arbor, Michigan, pp. 1–8.
Zhou, B., Zhang, X., Wang, R., 2004. Automated soil resources mapping based on decision tree and Bayesian predictive modeling. Journal of Zhejiang University Science 5, 782–795.
Zhu, A.X., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resources Research 36/3, 663–677.