Geoderma 155 (2010) 175–185

Contents lists available at ScienceDirect

Geoderma

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

Multi-scale digital terrain analysis and feature selection for digital soil mapping

Thorsten Behrens [a], A-Xing Zhu [b,c], Karsten Schmidt [a,*], Thomas Scholten [a]

[a] Institute of Geography, Physical Geography, University of Tübingen, Rümelinstraße 19-23, D-72074, Tübingen, Germany
[b] State Key Laboratory of Environment and Resources Information System, Institute of Geographical Science and Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA

ARTICLE INFO

Article history:
Received 21 May 2008
Received in revised form 10 June 2009
Accepted 18 July 2009
Available online 28 August 2009

Keywords:
Multi-scale digital terrain analysis
Feature selection
Spatial data mining
Digital soil mapping
ANOVA
Principal components analysis
Random subsets
Decision trees

ABSTRACT

Terrain attributes are the most widely used predictors in digital soil mapping. Nevertheless, discussion of techniques for addressing scale issues and feature selection has been limited. Therefore, we provide a framework for incorporating multi-scale concepts into digital soil mapping and for evaluating these scale effects. Furthermore, soil formation and soil-forming factors vary and respond at different scales. The spatial data mining approach presented here helps to identify both the scale which is important for mapping soil classes and the predictive power of different terrain attributes at different scales. The multi-scale digital terrain analysis approach is based on multiple local average filters with filter sizes ranging from 3×3 up to 31×31 pixels. We used a 20-m DEM and a 1:50 000 soil map for this study. The feature space is extended to include the terrain conditions measured at different scales, which results in highly correlated features (terrain attributes). Techniques to condense the feature space are therefore used in order to extract the relevant soil forming features and scales. The prediction results, which are based on a robust classification tree (CRUISE) show that the spatial pattern of particular soil classes varies at characteristic scales in response to particular terrain attributes. It is shown that some soil classes are more prevalent at one scale than at other scales and more related to some terrain attributes than to others. Furthermore, the most computationally efficient ANOVA-based feature selection approach is competitive in terms of prediction accuracy and the interpretation of the condensed datasets. Finally, we conclude that multi-scale as well as feature selection approaches deserve more research so that digital soil mapping techniques are applied in a proper spatial context and better prediction accuracy can be achieved.

1. Introduction

Terrain attributes are the most widely used predictors in digital soil mapping (McBratney et al., 2003). This is because relief is one of the most important soil-forming factors (Jenny, 1941) and digital elevation models (DEM) are widely available at different resolutions. These range from <1 m for LIDAR data up to 1 km for datasets with global coverage. Scale is an important consideration when maps of soil classes are produced. The map-user requires information at a particular scale and available covariates have a particular spatial resolution.

The impact of scale and resolution on soil property and soil class mapping is analyzed and described in various publications. The influence of the resolution (grid size) of digital elevation models on soil landscape modelling as well as applications has been widely discussed (e.g. Thompson et al., 1999; Schoorl et al., 2000; Valeo and Moin, 2000; Vázquez et al., 2002; Claessens et al. 2005; Smith et al., 2006; Zhu, 2008).

Concerning the influence of scale on the spatial distribution of soils, most authors focus on soil properties (Burrough, 1983; Goovaerts and Webster, 1994; German, 1999; Lark and Webster, 2001; Nemes et al., 2003; Lark, 2005; Zhu et al., 2004; Bartoli et al., 1991, 2005; Zhao et al., 2006). Methods for incorporating analyses of different scales of variation into techniques for mapping the distribution of soil classes have not been widely reported in the literature (Hupy et al., 2004).

Classical machine learning approaches like classification trees, artificial neural networks, support vector machines etc. have no mechanisms to detect spatial relationships (Moran and Bui, 2002). However, the need to develop straightforward approaches to deal with multi-scale variations in digital soil mapping is of great importance (McBratney et al., 2003; Lagacherie, 2008). To analyze the effect of scales and to incorporate the spatial domain in data mining or prediction processes Moran and Bui (2002), Behrens (2003), Behrens et al. (2005), Smith et al. (2006), and Grinand et al. (2008) use contextual spatial information, whereas Lark (2006) and Mendonca-Santos et al. (2006) demonstrate the use of wavelets.

This paper demonstrates how contextual information from predictor variables can be incorporated into statistical digital soil mapping. It helps to analyze both the scale of soil classes and the

* Corresponding author. Tel.: +49 7071 29 78943.
E-mail address: karsten.schmidt@uni-tuebingen.de (K. Schmidt).

0016-7061/$ – see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2009.07.010

predictive power of different terrain attributes at different scales. A larger number of terrain attributes than is commonly used in such studies are analyzed. A multi-scale digital terrain analysis approach based on multiple local average filters (Behrens et al., 2005; Grinand et al., 2008) is proposed, which produces separate hierarchical spatial components (McBratney et al., 2003). This can be applied to various kinds of spatial predictor variables. The core of this approach is to compute terrain variables over a range of filter sizes, which not only increases the feature space, that is, the number of variables used as predictors, but also produces a set of highly correlated features (terrain attributes). Dimension reduction and feature selection techniques (cf. Liu and Motoda, 1998) are then used to reduce the feature space and select proper features in order to extract the relevant soil-forming features and scales for each soil class in a map.

The methodological framework introduced in this study offers the possibility to answer the following questions:

- Is prediction accuracy effected by incorporating predictors with explicit information on scale dependence?
- Which terrain attributes are most important to predict certain soil classes?
- Do optimal scales in terms of filter size exist to best predict specific soil classes?
- Are there differences in prediction accuracies based on different feature selection approaches?
- How does prediction accuracy change with respect to the number of features?
- Can the selected features be interpreted in terms of soil formation?
- Which approach for selecting the relevant features would best answer these questions?

The approach is introduced and applied to datasets typically available for DSM approaches in Germany: a 20-m DEM and a 1:50 000 soil map.

2. Material and methods

2.1. Study area and datasets

The study area, Palatinate Forest, is located in southern Rhineland-Palatinate, Germany. It is a low mountain range area within the southwest German–Lorraine Triassic escarpment. The soils in the Palatinate Forest were formed from substrates of the Upper Red Bed Sandstone, Bunter, and Lower Limestone. According to the current German mapping standard (Ad-hoc-AG Boden, 2005) the soil maps within the Palatinate Forest were produced at a scale of 1:50 000 relating soil classes to geology and terrain conditions.

The entire study area covers about 300 km². All modelling approaches are based on a DEM with a resolution of 20 m. For training and validation, two spatial subsets were selected from the soil map, each consisting of the same six soil classes (Table 1, Fig. 1) and each covering an area of about 40 km². This dataset is similar to the one used by Behrens and Scholten (2006a,b), a comparative study of different data mining approaches. The two soil datasets are located about 50 km apart.

2.2. Digital terrain analysis and terrain attributes

We used 19 terrain attributes in this study. As some of them were calculated on the basis of different thresholds, a total of 38 terrain datasets were created (Table 2). The attributes are classified into 9 different groups based on how they are computed and their postulated effects on soil formation.

All attributes are explained briefly below, starting with *local terrain attributes* calculated on the basis of moving window techniques over the same spatial extent for each pixel (e.g. Zevenbergen and Thorne, 1987), followed by *regional terrain attributes* based on contributing area concepts (e.g. Quinn et al., 1991), and *combined terrain attributes* based on local and regional attributes such as the compound topographic index (e.g. Beven and Kirkby, 1979).

2.2.1. Local terrain attributes

2.2.1.1. Non-curvature attributes.

Elevation (EL)

Elevation is the original values of the DEM.

Mean slope (SLD)

Mean slope is the average slope over a neighbourhood. Following the approach of Dietrich and Montgomery (1998), mean slope is calculated based on the geometric mean of the slopes between the two cardinal and the two diagonal slopes in a 3×3 pixel neighbourhood.

Steepest slope (SLT)

Steepest slope is the steepest decent within a neighbourhood. It is the slope gradient between two adjacent cells forming the steepest line orientated in a down slope direction over a 3×3 pixel neighbourhood (Tarboton, 1997).

Aspect (D)

Aspect is the average orientation of the neighbourhood centred around a given pixel. The calculation is based on the finite differences approach introduced by Horn (1981). There are two concepts to transform the circular character of aspect into the linear space: one is the sine and cosine transformation, which result in parameters presenting eastness and northness; the other is deviation from bearings. We prefer the deviations from bearings because more directional derivates can be calculated and thus the influence of aspect on soil can be easily interpreted. We use the deviations from 0°, 45°, 90° and 135° in this study.

2.2.1.2. Curvatures. Curvature is a measure of convexity or concavity of terrain surface. It is often used to indicate areas of material removal and areas of material accumulation. Curvature can be defined and so computed in various ways. We used two major categories of curvature measure.

a) Based on finite differences

The curvatures in this group are based on finite difference equations given by Evans (1980). The curvature attributes used in this study were selected from a system of curvatures introduced by Shary et al. (2002): *Mean curvature (MEC), Profile curvature (PRC), Horizontal curvature (HOC), Minimal curvature (MIC), and Maximum*

Table 1
Analyzed soil classes and their distribution in the training and the validation area; Tr: coverage in the training area; VA: coverage in the validation area; (Behrens and Scholten, 2006a).

	Soil class	Parent material	TR [%]	VA [%]
S1	Cumulic anthrosols, gleyic anthrosols, gleysols and fluvisols	Fluvic soil material in valleys and adjacent to water courses	5	4
S2	Haplic cambisols	Loess-containing sandy Pleistocene periglacial slope deposits covering weathered sandstone (Middle Bunter)	32	34
S3	Dystric cambisols	Loess-containing sandy Pleistocene periglacial slope deposits covering elevation weathered sandstone (Middle Bunter)	5	2
S4	Haplic podzols	Sandy slope deposits rich in debris covering weathered sandstone (Middle Bunter)	32	32
S5	Dystric cambisols	Blocky, sandy slope deposits covering deep-weathered sandstone (Lower Bunter)	11	20
S6	Dystric cambisols and Cambi-haplic podzols	Blocky, sandy slope deposits covering weathered blocky sandstone (Middle Bunter)	12	5

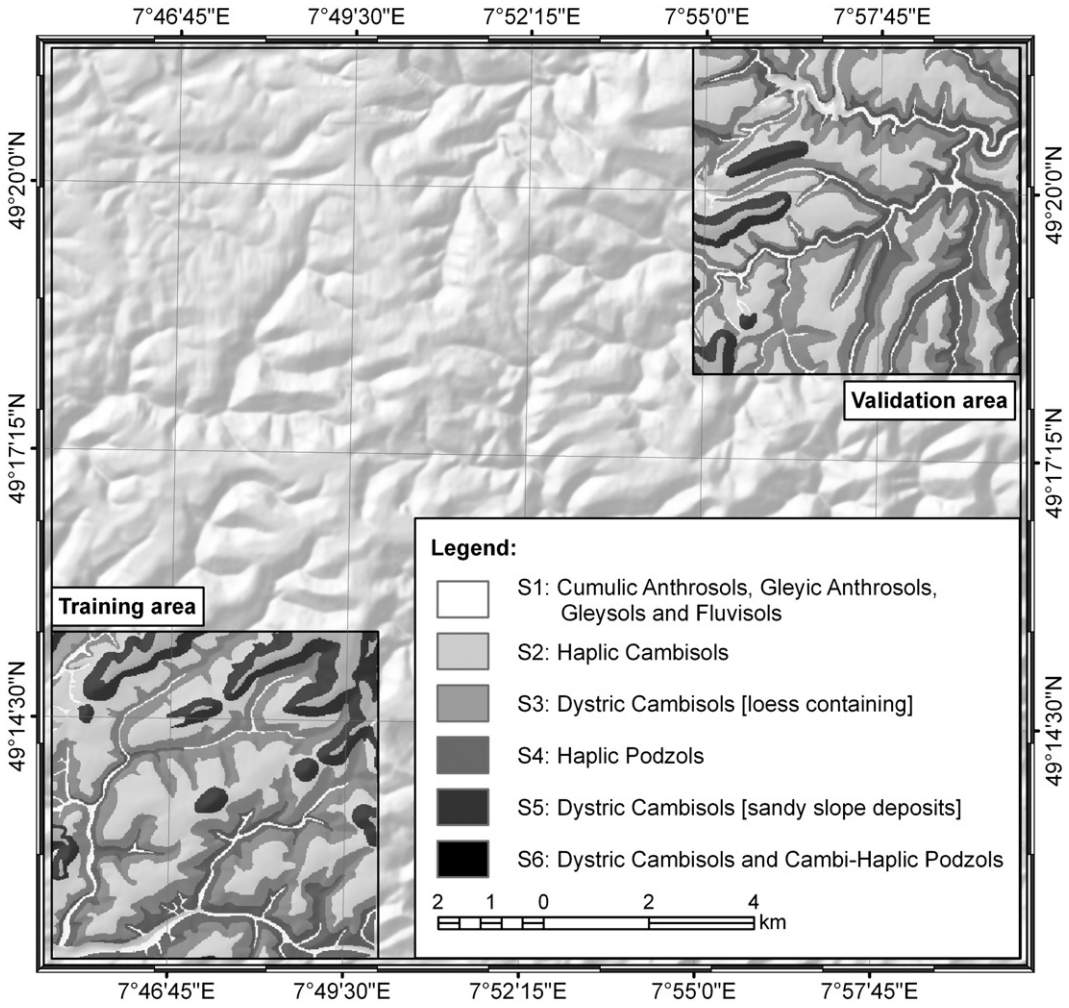

Fig. 1. Investigation area and spatial distribution of soil types in the study area, Palatinate Forest, Germany.

curvature (MAC). Interested readers are referred to Shary et al. (2002) for details on these curvature measures.

b) Based on context approaches

Relative profile curvature (RPC)

Relative profile curvature is the ratio of two mean slope values: the mean slope of all cells higher than the centre cell divided by the mean slope of all cells lower than the centre cell (Behrens, 2003). Thus, it is a measure of local mass balance.

Relative horizontal curvature (RHC)

Relative horizontal curvature is based on an analysis of the aspect of the neighbouring pixels around a centre cell (Kleefisch and Köthe, 1993). If the aspect of a neighbouring grid cell is directly pointing into the centre cell, a value of 100 is assigned to the corresponding cell, whereas −100 is assigned if it points to the opposite direction. A value of 0 is assigned if the aspect of a neighbouring cell is 90° to the direction to the neighbouring centre cell. For all other angles, the difference between these two angles is calculated and normalized within the range of −100 and 100. Finally, the average is calculated over all differences resulting in recognition of convergent positions if the average is greater than 0, and divergent positions if the average is below 0.

Waxing and waning slopes (WW)

Huber (1994) introduced the concept of waxing and waning slopes on the analysis of neighbourhood statistics. The approach is based on a comparison of the difference in elevation between the centre pixel and the minimum in a local neighbourhood and that of the centre pixel and the maximum in the same neighbourhood. If the difference

Table 2
Groups of terrain attributes, based on thresholds and underlying algorithms.

	Group	Description	Thresholds	Parameters
Local attributes	EL	Elevation a.s.l.		EL
	D	Difference from aspect angle	0, 45, 90, 135 [°]	D0, D45, D90, D135
	SL	Slope		SLT, SLD
	CP	Primary curvatures		MEC, PRC, HOC MIC, MAC
	CS	Secondary curvatures		WW, RPC, RHC, TR
Regional and combined attributes	RHP	Relative hillslope position	1, 2, 5, 10, 20, 50 [ha]	RHP, RHP1, RHP2, RHP5, RHP10, RHP20, RHP50
	PD	Projected distance to stream	1, 2, 5, 10, 20, 50 [ha]	PD1, PD2, PD5, PD10, PD20, PD50
	LE	Local elevation above stream	1, 2, 5, 10, 20, 50 [ha]	LE1, LE2, LE5, LE10, LE20, LE50
	RA	Attributes based on contributing area		LS, CTI, CA

from the maximum is smaller than that from the minimum, it is an upper or waxing slope; otherwise, it is a toe or a waning slope. Waxing and waning slopes (WW) are generally calculated on neighbourhoods larger than 3×3 pixels (Huber, 1994). In this study, WW was calculated on the basis of a 7×7 pixel neighbourhood.

Topographic roughness (TR)

The approach to calculate topographic roughness is based on the variation of slope and aspect within a local neighbourhood (Behrens, 2003). As generally all neighbouring pixel values are slightly different from the values of the slope and the aspect grid are rounded to the nearest integer. The variety (number of different values) is calculated within a 7×7 pixel neighbourhood for both slope and aspect. Topographic roughness value for a given pixel is the product of the number of different slope gradient values and the number of different aspect values within the window.

2.2.2. Regional terrain attributes

2.2.2.1. Contributing area (CA). In this study, CA is computed based on the multi-flow algorithm introduced by Quinn et al. (1991). This variable is used both as a stand-alone attribute and as the basis to calculate relative hillslope position, compound topographic index and USLE LS-factor (see below). The algorithm uses the neighbouring slopes in a 3×3 pixel neighbourhood to weigh the outflow of each cell to its neighbours (Dietrich and Montgomery, 1998).

2.2.2.2. Projected distance to stream (PD). Projected distance to stream is the Euclidian distance between the local pixel and the closest pixel on the streamline network. The streamline network is derived from the DEM using a so-called single-flow algorithm (Jenson and Domingue, 1988). The level of detail of the stream network depends on the threshold of flow accumulation used to define the stream network (Behrens, 2003). We used 6 different thresholds, namely 1, 2, 5, 10, 20, and 50 ha, to derive the streamlines for comparison and for scale analysis, based on the assumption that different catchment sizes and hydrological processes have different impacts on soil formation.

2.2.2.3. Local elevation (LE). Local elevation is the elevation difference between the elevation of a local pixel and the elevation of its closest stream pixel. After delineating the streamline from the DEM, the pixel values of the stream are replaced by the original elevation values of the DEM. Afterwards, a nearest neighbour approach is used to assign the elevation on the streamline to other non-stream pixels. This elevation data layer is referred to as streamline elevation data layer (sEL). Finally, the streamline DEM is subtracted from the original DEM, resulting in a grid containing local elevation (e.g., MacMillan et al., 2000; Behrens, 2003).

2.2.2.4. Relative hillslope position (RHP). Relative hillslope position is a measure of relative distance between valley bottoms and ridges. Two approaches for determining RHP were used in this study. The first approach, as introduced by Hatfield (1999), is related to local elevation. The calculation is based on the original elevation data layer (EL), the streamline elevation data layer (sEL) and the crest elevation data layer (cEL), which is created similarly to sEL, but using the elevation of ridge lines (crest). The final equation is:

$$RHP = \{(EL - sEL) / (cEL - sEL)\} \times 100 \quad (1)$$

The other approach (RHP_{CA}), is to subtract the CA of the original DEM and the CA calculated over an inverted DEM (Behrens, 2003). Thus, in mid slope positions the values are close to zero. In upper slopes the values are negative, whereas in toe slopes the resulting values are positive.

2.2.3. Combined terrain attributes

2.2.3.1. Compound topographic index (CTI). The compound topographic index, as originally introduced by Beven and Kirkby (1979), is defined as:

$$CTI = \ln(CA / \tan(SLT)) \quad (2)$$

Where SLT (Tarboton, 1997) is the steepest (maximum) slope gradient over a neighbourhood centred around a local pixel and CA is the contributing area (Dietrich and Montgomery, 1998) of that pixel. This index measures the balance of water accumulation and drainage over a local neighbourhood around a given pixel.

2.2.3.2. USLE LS-factor (LS). The USLE LS-factor (Wischmeier and Smith, 1978) is a well-known and important parameter. In this study we use an approach modified for German mid-latitude landscapes (Feldwisch, 1995).

For slopes <9% the *S*-factor is:

$$S = 0.063 + 10.461 \times \sin(SLD) \quad (3)$$

whereas for slopes >= 9 % the RUSLE equation (McCool et al., 1987) is used:

$$S = 16.8 \times \sin(SLD) - 0.5 \quad (4)$$

The L-factor is then based on:

$$b = 11.16 \times \sin(SLD) / (0.765 + 2.625 \times \sin(SLD)) \quad (5)$$

$$m = b / (b + 1) \quad (6)$$

$$L = (CA / 22.13)m \quad (7)$$

Finally the LS-factor is calculated by multiplying the L- and the S-factor grids.

2.3. The multi-scale approach

In terms of specific geomorphometry (Evans, 1972) and the numerical description of continuous surface forms (Pike, 1993), most techniques to derive terrain attributes are constrained by the resolution of the DEM (Wood, 1996). Wood (1996) argues that the scale implied by the resolution of DEM is not always appropriate in the context of landscape characterization. We assume that this holds true for digital soil mapping, as has been shown in other works (Behrens, 2003; Behrens et al., 2005, 2008; Smith et al., 2006; Zhu, 2008).

We applied square local neighbourhood averaging filters with sizes of $n \times n$ pixels, where n is an odd number ranging from 3 to 31, to all terrain attributes to investigate the influence of scale in digital soil mapping systematically. This local averaging extends the feature space by a factor of 14. Filtering is also applied to terrain attributes based on different thresholds of contributing area (Section 2.2.2). Examples of filtered terrain attributes are given in Fig. 2.

Other options to analyze and incorporate scale exist but were not used in this study. One alternative method is to calculate terrain attributes based on DEMs with different resolutions. This approach can result in errors of the surface models (Wood, 1996) as well as resampling problems within the prediction approach. Another alternative approach is to fit multi-scale quadratic approximations to elevation values within windows of varying dimensions (e.g. Haralick, 1983; Wood, 1996; Smith et al. 2006). Compared to these two alternatives, the main technical advantage of the approach of local average filtering used in the present study is that it can easily be applied to any terrain attribute, regardless of the underlying algorithm.

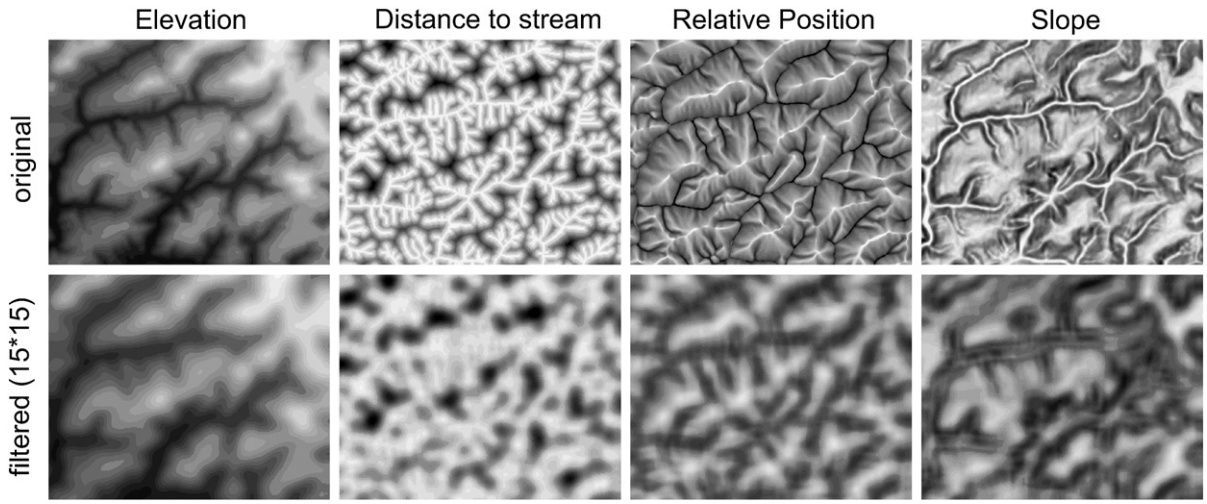

Fig. 2. Examples to show the effect of filtering terrain attributes from a subset of the study area. Original terrain attributes (elevation, distance to stream (1 ha), relative position, and slope) as well as filtered versions based on a 15 × 15 pixel window are shown.

2.4. Feature selection

Feature selection aims to reduce the dimensionality of datasets by eliminating redundant and/or noisy features (Liu and Motoda, 1998) and is conducted for two main reasons: First, it gives a deeper insight into the driving forces of a prediction – i.e. the soil-forming processes of each soil class – and second, regarding data mining techniques, a reasonable selection of features mitigates the "curse of dimensionality". The curse of dimensionality, a term coined by Bellmann (1961), points to the problem of highly dimensional data, yielding weaker prediction results compared to datasets with a reduced feature space (Pechenizkiy et al., 2003). Thus, these methods optimally fit our problem of analyzing a high dimensional and highly correlated set of terrain attributes that are produced using our multi-scale approach.

The importance of a feature, or the predictive power of a feature, can be calculated in two ways – based on so-called *filter* or *wrapper* approaches (John et al., 1994). A filter approach is carried out as a pre-processing step of a prediction. It identifies the relevant features that have a significant impact on the corresponding spatial distribution of soil classes (Lai et al., 2006). In contrast to filter approaches, the importance of features in wrapper approaches is calculated as the result of multiple runs of a prediction algorithm based on changes in the features training dataset. The most effective subset is then generally used in the final predictions (Breiman, 1996).

Feature selection can be achieved in either supervised or unsupervised manners. In the unsupervised case, classification for feature set reduction is based on statistical similarity of the features. In the case of supervised feature selection the influence of each feature is tested against an existing sample set based on prediction results. Hence, supervised feature selection is based on the feature importance, whereas the unsupervised case is better described with statistical dimension reduction.

In this study we compare a selection of supervised and unsupervised filter approaches of varying complexity as well as one wrapper approach.

2.4.1. Tested feature selection approaches

A simple filter method to reduce dimensionality in an unsupervised fashion is the Karhunen–Loeve-expansion or principal components analysis (PCA). It is one of the most common feature extraction approaches (Pechenizkiy et al., 2003). In addition to the classical PCA we applied a semi-supervised PCA approach. In this approach we calculated a separate PCA for each soil class in such a way that only instances containing the corresponding soil class were used. We call this approach class-PCA (cPCA).

To describe the ability of a numerical attribute to separate nominal features like soil classes, an analysis of variance approach (ANOVA) can be used as a linear, univariate, and supervised filter technique. Classification tree algorithms based on this concept were developed by Loh and Shih (1997) and Kim and Loh (2001) (c.f. Section 2.5). In order to analyze each terrain attribute independently, we implemented a decision stump (Iba and Langley, 1992) based on an ANOVA *F*-ratio. For feature selection each terrain attribute is tested against each soil class. The final ranking of the features for each soil class is based on the resulting *F*-values.

A more complex supervised feature selection method for both filter and wrapper approaches applied in this study is based on multiple predictions with varying feature subsets. Thus, the prediction accuracy varies in each iteration based on the different combinations of features used. A full-featured search over the entire feature space (N) using heuristic search algorithms (Blum and Langley, 1997) would require 2^N iterations. We applied a stochastic optimization approach using random subsets of the total number of features because a full-featured search was not possible given the 608 features we had to work with. This method is part of various ensemble prediction approaches (Ho, 1998; Bay, 1999; Breiman, 2001; Skurichina and Duin, 2002; Lai et al., 2006).

Different approaches are possible to select an optimal performing subset as the final set of predictors. The most straightforward approach is to select the best performing subset (Kohavi and John, 1997), which is the classical wrapper approach. Another approach is to average prediction accuracies over the features and to select the best performing features for the final prediction. Additionally, for reasons of completeness, we also compare a ranking based on the lowest prediction performances (filter approach), where we also picked the best performing features.

In contrast to the supervised ANOVA filter approach described above, which is applied on single features only, the random subspace models can reveal non-linear relationships and possible interactions between features (Zhao and Liu, 2007) when decision trees, artificial neural networks or similar methods are used. A major disadvantage of this method is the high computational costs (Forman, 2003; Blum and Langley, 1997).

In this study, the parameter settings for the random subspace approaches are:

- 500 iterations which seems to be enough as it shows that very high accuracies can be obtained before all possible weak classifiers are computed (Ho, 2002),

- a minimum of 100 instances in each terminal-node, an average derived from Schmidt et al. (2008),
- unpruned trees as suggested for random forests by Breiman (2001) and Breiman and Cutler (2004), and
- the square root of the number of features as suggested for feature subset for random forests by Breiman (2001) and Breiman and Cutler (2004).

All feature selection approaches used in this study return ordered lists of features. Hence, criteria need to be applied to select a reduced subset. In the case of PCA we used the Kaiser criterion or Eigenvalue-rule (Kaiser, 1960). As we want to analyze terrain attributes and their behaviour at different scales, we selected original terrain attributes instead of the derived PCA components. Therefore, we picked all terrain attributes with the highest Eigenvalue for each component as final features for the prediction. Hence, the evaluation of PCA cannot directly be compared with other studies using PCA as a feature extraction technique.

For the ANOVA approach we analyzed all soil classes with the aim of finding a threshold, where each soil class could be described by at least two terrain features. The resulting *F*-value threshold was selected to be 4000. This ensures that not too many features are selected for soil classes, which reduces the problem of correlated features.

In the random subset approach, the number of features chosen in the random subset method determines the size of the subset. In this case it is the square root of 608, resulting in 25 predictors for each soil class. Consequently, we compared the three possible random subset approaches of feature ordering based on the first 25 top ranked features. Thus, for the approach based on the lowest prediction results, the "best of the worst" set is chosen, which might then contain important features. To analyze possible further reductions we tested the best features of each group of terrain attributes as described in Section 2.2 to further reduce multicollinearity for the best performing random subset approach.

2.4.2. Generated feature sets

Based on the approaches introduced above, the datasets used to compare the different feature selection approaches and to analyze scale dependency are:

Original datasets:
- All 38 unfiltered terrain attributes (*oTA*)
- All 608 terrain multi-scale attributes (*msTA*)

Datasets based on feature selection approaches:
- Principal components analysis based on *msTA*:
  - Entire dataset (*PCA*)
  - Separately for each soil class area only (*cPCA*)
- ANOVA approach based on *msTA*:
  - All attributes for each soil class [*F*-value>4000] (*ANOVA1*)
  - The best attribute of each terrain attribute group for each soil class [*F*-value>4000] (*ANOVA2*)
- Random subset approach based on *msTA*:
  - Best subset in the training area [maximum] (*RSS1*)
  - Selection based on averaged prediction results [mean] (*RSS2*)
  - Selection based on worst prediction results [minimum] (*RSS3*)
  - The best attribute of each terrain attribute group for each soil class based on the best performing approach out of RSS1, RSS2, and RSS3 approach (*RSS4*)

2.5. Prediction and validation

For predictions we used a 10-fold cross-validated 1D CRUISE classification tree approach (Kim and Loh, 2001), which, among artificial neural networks, was shown to be one of the best data mining based prediction approaches in mapping soil classes (Behrens and Scholten, 2006a,b). CRUISE is based on an ANOVA approach for selecting the best feature to split a branch of a tree and on a linear discriminant analysis for split point selection. It was found to be superior to the well-known CART (Breiman et al., 1984) algorithm (Kim and Loh, 2001). We decided to use the CRUISE classification tree in this study to have a consistent methodological framework for the linear and non-linear supervised feature selection as well as for the prediction approach, as both methods are based on an analysis of variance.

To validate the results in the spatial domain we used the F_1-measure (van Rijsbergen, 1979) to compare the predictions for training as well as for independent validation.

The F_1-measure is calculated as the harmonic mean of precision and recall based on the confusion matrix (Table 3):

Precision:

$$P = tt / (tt + ft) \tag{8}$$

Recall:

$$R = tt / (tt + tf) \tag{9}$$

F_1-measure:

$$F = 2 \times (P \times R) / (P + R) \tag{10}$$

To provide a single measure for the ability to generalize, i.e. a small difference between the F_1-measures in the training (F_1T) and the validation area (F_1V), as well as for the validation accuracy, we propose the prediction quality measure *Q*, calculated as follows:

$$Q = F_1V \times (1-(F_1T-F_1V)) \tag{11}$$

Thus, *Q* weighs the validation accuracy against the generalization ability. If both the training and validation F_1-measure are high, the prediction approach can be recommended. All measures range between 0 and 1. High values indicate good model performance.

3. Results and discussion

The discussion of results is primarily based on the following three tables (Tables 4–6). Table 4 contains the selected features and their corresponding scales based on the different feature selection approaches. Tables 5 and 6 show the prediction results for each selected feature subset as well as the number of selected features.

All approaches were evaluated on the basis of prediction accuracy, which will be discussed first (Section 3.1). The results of the feature selection approaches will then be discussed in terms of selected features, algorithms, scale dependency, and soil classes (Sections 3.2–3.4).

3.1. Overall prediction accuracy

Based on the F_1-measures in the validation area, the RSS4 approach is judged to offer the best predictive performance (Table 5). ANOVA2 and RSS4 are the best feature subset selection approaches based on the prediction quality measure *Q* (Table 6). Consequently, most interpretations are based on these two datasets. The prediction results of the RSS4 dataset range from 0.44 to 0.75 in the validation area

Table 3
Confusion matrix to compare mapped and predicted pixels as a basis for prediction accuracy, precision, and recall (t = true; f = false).

		Mapped	
		True	False
Predicted	True	tt	tf
	False	ft	ff

Table 4
Best performing features and corresponding spatial filter sizes resulting from the applied feature selection approaches.

	ANOVA1 / ***ANOVA2***						PCA	cPCA						RSS4					
	S1	S2	S3	S4	S5	S6		S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
EL			***1***	***7***	***3***	***31***	31		1, 31	1	5	1	1, 31		7	11	15	7	3
SLT		9	11	7		***7***			21		1, 21			3					9
SLD		***7***	***9***	***7***		7	31			17	1, 19		1, 3		9	5	11		
D0								9, 23		1		1							
D45							1,15,17	25	15	15	15	21	15		23			25	1
D90								9, 11, 25	15			31	19						
D135							19		17	21	17	1, 19	17				15		
RHP		11	21	***25***					21	21			21						
RHP1		13	21	25	31								1						
RHP2		11	19	23	27											11	29		
RHP5		***9***	21	21	23		17	23	1						5				
RHP10		5	19	21	17						13								1
RHP20		3	17	17	***13***			13			31	5						31	
RHP50		3	***11***	15	7							29		3					
PD1			19	25		31											25		
PD2		7	15	23		31													
PD5		3	15	19	19	***31***													
PD10		***1***	13	17	***15***			31				3							
PD20		1	***9***	***15***	11			9						9	9				15
PD50		1	3	11	9							19	31						
LE1						29	31		1		1, 25								
LE2						31		9								1			27
LE5				13		***31***		17									7		
LE10			7	15	13	31								1					
LE20			7	15	11	31													
LE50		***3***	***5***	***15***	**9**	31									17			1	
CA			21	23						31		25					27		
CTI	***3***					***15***				1		1		1				27	15
LS		***5***	***19***	***17***		7	7, 31	7	3							29			
MEC		17	25	***25***	19			23											
PRC		15	***31***	29	***19***	***31***					3	21	3, 31			11		15	
HOC			31		31		1, 3	3, 5	23, 29	5, 19, 31	1, 3, 11	1, 7	31				19		17
MIC		***15***	27	23	15				1	1, 17	31	19	29						
MAC							1, 3, 9, 23, 25	27	31		3, 5	3, 19	23, 31		1				
RPC	***3***	***15***	***27***	25	21				1	7				1	23		15		
RHC	3	15	25	***25***	23	31		5				31							
WW		17	27	29	***21***													1	7
TR						***9***	1, 29		31	1, 13, 21	7, 23	3	1, 25			13			

PCA: Entire dataset (max. 20 attributes); cPCA: class-PCA for each soil class (max. 20 attributes); ANOVA1: All significant attributes (F-value>4000); ANOVA2 (bold, italic): best attributes for each group of terrain attributes based on ANOVA1; RSS4: best attribute of each terrain attribute group based on best performing RSS 1–3 approaches.; S1–S6: soil classes. The shading separates the terrain attribute groups and feature selection approaches for better readability.

Table 5
Prediction results (F_1-measure) for the different datasets for all soil classes (S1–S6).

	S1		S2		S3		S4		S5		S6		Mean	
	TR	VA	TR	VA	TR	VA	TR	VA	TR	VA	TR	VA	TR	VA
msTA	0.6	0.51	0.78	0.74	0.83	0.35	0.78	0.6	0.76	0.49	0.82	0.50	0.76	0.53
oTA	0.56	0.31	0.77	0.73	0.84	0.50	0.79	0.61	0.75	0.41	0.81	0.45	0.75	0.50
PCA	0.03	0.00	0.77	0.67	0.77	0.22	0.67	0.48	0.55	0.26	0.65	0.10	0.57	0.29
cPCA	0.55	0.40	0.76	0.66	0.81	0.28	0.78	0.58	0.76	0.43	0.80	0.45	0.74	0.47
ANOVA1	0.45	0.46	0.73	0.69	0.80	0.45	0.77	0.58	0.75	0.42	0.81	0.45	0.72	0.51
ANOVA2	0.46	0.47	0.72	0.67	0.80	0.48	0.77	0.58	0.72	0.47	0.80	0.48	0.71	0.53
RSS1	0.64	0.51	0.80	0.72	0.87	0.31	0.81	0.57	0.76	0.43	0.84	0.35	0.79	0.48
RSS2	0.59	0.47	0.76	0.73	0.84	0.48	0.75	0.53	0.77	0.46	0.81	0.47	0.75	0.52
RSS3	0.57	0.51	0.75	0.7	0.82	0.47	0.73	0.54	0.77	0.49	0.81	0.46	0.74	0.53
RSS4	0.56	0.50	0.81	0.75	0.83	0.44	0.77	0.56	0.75	0.49	0.80	0.48	0.75	0.54
Mean	0.50	0.41	0.77	0.71	0.82	0.40	0.76	0.56	0.73	0.44	0.80	0.42	0.73	0.49

msTA: all 608 terrain attributes from multi-scale digital terrain analysis; oTA: all 38 original unfiltered terrain attributes; PCA: entire Dataset (max. 20 attributes); cPCA: class-PCA for each soil class (max. 20 attributes); ANOVA1: all significant attributes; ANOVA2: best attributes for each group of terrain attributes (if significant); RSS1: 25 most predictive attributes based on prediction maximum; RSS2: 25 most predictive attributes based on prediction mean; RSS3: 25 most predictive attributes based on prediction minimum; RSS4: best attribute of each terrain attribute group based on best performing RSS 1–3 approaches.; TR: training dataset; VA: validation dataset.

Table 6
Ranked mean prediction quality (Q) and number of selected features for all datasets tested.

	ANOVA2	RSS4	msTA	RSS3	ANOVA1	RSS2	oTA	RSS1	cPCA	PCA
Q	0.43	0.43	0.42	0.42	0.41	0.41	0.39	0.35	0.35	0.22
Features	2–8	6–9	608	25	3–29	25	38	25	16–20	19
% of msTA	0.3–1.3	1–1.5	100	4.1	0.5–5.8	4.1	6.3	4.1	2.6–3.3	3.1

msTA: all 608 terrain attributes from multi-scale digital terrain analysis; oTA: all 38 original unfiltered terrain attributes; PCA: entire Dataset (max. 20 attributes); cPCA: class-PCA for each soil class (max. 20 attributes); ANOVA1: all significant attributes; ANOVA2: best attributes for each group of terrain attributes (if significant); RSS1: 25 most predictive attributes based on prediction maximum; RSS2: 25 most predictive attributes based on prediction mean; RSS3: 25 most predictive attributes based on prediction minimum; RSS4: best attribute of each terrain attribute group based on best performing RSS 1–3 approaches.

(Table 6). With an average of 0.54 for validation, the results are almost 20% better than those reported by Behrens and Scholten (2006a,b), which were based on a different and unscaled terrain attribute set. Yet, both accuracies (training and validation) are within the range of accuracy achieved by similar prediction approaches and better than those based on conventional field soil surveys (Zhu et al., 2001).

3.2. Feature selection algorithms and feature count

Technically, the most interesting result of the feature selection evaluation is the number of selected features in relation to the prediction results (Table 6). The most condensed datasets (ANOVA2 and RSS4) return the highest Q-measures. Furthermore, RSS4 contains about 2/3 of the features chosen by ANOVA1 (Table 4) revealing a relatively high degree of similarity. However, the prediction results for the entire dataset are similar to the ones obtained using RSS3, RSS4, and ANOVA2 (Table 4), which return the smallest amount of features (Table 6).

As the complete extended feature set performs competitively, the CRUISE prediction approach must be regarded as very stable with respect to highly correlated and irrelevant features – at least for datasets with many instances. Thus, it can be stated that concerning prediction accuracy, feature selection is not important when stable prediction approaches like CRUISE are used. On the other hand, the ANOVA1 approach, containing the same attributes as ANOVA2, plus some additional (correlated) ones, returns (slightly) weaker results. The same pattern can be seen with the RSS approaches. First, the condensed datasets selected based on the RSS4 approach, perform best. Second, the maximum (wrapper) approach (RSS1) appears to achieve a good fit to the training data but poor prediction of the validation data, as indicated by the respective F_1 measure. This can be interpreted as overfitting.

With respect to PCA, the Eigenvalue-rule (Kaiser, 1960) returned about 30 features for each soil class. Because the relevance describing the dataset based on the Eigenvalues is relatively low after 15 to 20 features, we generally used the first 20 features only. Additionally, as some terrain attributes appeared more than once, the final number of features can be less than 20 (Table 6).

As found in other studies (Pechenizkiy et al., 2003), the unsupervised PCA approach turned out to be the worst technique in terms of selecting optimal features for all six soil classes. The semi-supervised class-PCA approach performed slightly better (Tables 5 and 6). As shown in Table 4, both PCA approaches select more local terrain attributes than regional ones (cf. Section 2.2) compared to the supervised approaches. This is evident especially for the aspect attributes (D0, D45, D90, and D135) and indicates that PCA has a serious drawback, as it gives high weights to features with a high variability, irrespective of whether they are useful for classification or not. As a result, the chosen features might have poor discriminating power (Pechenizkiy et al., 2003). The ANOVA approaches are recommended for evaluating feature importance as they require the least processing time and also respect Occam's razor by identifying the fewest number of predictor variables. The selection of attributes and the prediction achieved using ANOVA are faster than a prediction approach using all attributes and much faster than using RSS.

In summary, it can be stated that only about 1% of the 608 terrain datasets was needed to achieve the most effective predictions. Furthermore, all supervised approaches worked much better than the alternative unsupervised and semi-supervised PCA approaches. All linear approaches were competitive with the non-linear feature selection approaches. Thus, non-linearity and feature interaction appear to not be pronounced in this dataset. Interestingly, the RSS3 approach – based on the "best of the worst" features of the ensemble approach, is comparable to the best approaches, whereas the best feature subset, based on the wrapper approach, is overfitted.

3.3. Scale

The most important result of this study is that scale has an influence on prediction accuracy, as illustrated in Table 4, where the selected features and the corresponding relevant scales are listed. It is remarkable that only 7% of the parameters chosen are unfiltered (ANOVA1, RSS4).

As both PCA approaches are not fully supervised and perform significantly worse (Section 3.2), it is not appropriate to discuss the relevant scales for soil prediction using these two approaches. Thus, the discussion of scale focuses on the best performing approaches, ANOVA and the RSS.

Although filter sizes are different in most cases for single attributes between the ANOVA approaches and RSS4 (Table 4), the mean spatial ranges for the soil classes (except for S6) are similar for both approaches (Fig. 3). To some degree the mean spatial ranges correlate with the spatial shapes of the soil classes. This is shown by the corresponding area / perimeter ratios in Fig. 3. For example, soil classes with an elongated shape, such as S1, correspond to small filter sizes, whereas larger units with a more "areal" character, such as S6, respond to larger filter sizes.

However, even if the mean spatial ranges can be interpreted, it is important to filter each terrain attribute with different filter sizes as each attribute has its highest predictive power at different filter sizes (Table 4) for different soil classes and for different prediction methods.

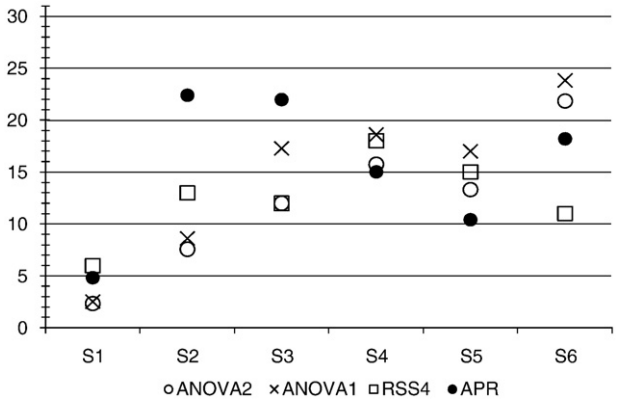

Fig. 3. Mean relevant filter sizes for each soil class based on the ANOVA approaches and RSS4. Additionally the area / perimeter ratio (APR, divided by 5) for each soil class is shown.

This is one of the most important results. It shows that even within one soil class different filter sizes are needed for different terrain attributes to achieve optimized predictions. Thus, it is important to test all filtered attributes on all soil classes of a map separately.

The ANOVA approach is considered most suited to visualizing the changes in predictive power over different filter sizes as the ANOVA *F*-value is reported for each attribute individually. Fig. 4 shows the ANOVA *F*-values plotted against the filter sizes in separate diagrams for each soil class. This clearly demonstrates the importance of multi-scale digital terrain analysis in predicting soil classes. For each soil class different scales and different attributes are identified as most important.

Table 4 reveals that the neighbourhood range in the ANOVA dataset is, in most cases, negatively correlated to the contributing area thresholds in hectares used to calculate RHP, PD, and LE. For smaller contributing areas larger filter sizes are required. Hence, it can be stated that a larger threshold for the contributing area has the same effect as a larger filter size. This emphasizes the existence of the influence of scale. Furthermore, this also indicates that the influence of scale does not exist in a random fashion. Instead, it is related to soil-forming processes and the general geomorphometry at the landscape scale.

3.4. Soil

From a soil science perspective, the interpretation of the selected features, their corresponding scales and the general prediction performance is most interesting. As shown in Fig. 4 and Table 4, fluvial soils (S1) can be predicted best using only two attributes: the compound topographic index and the relative profile curvature.

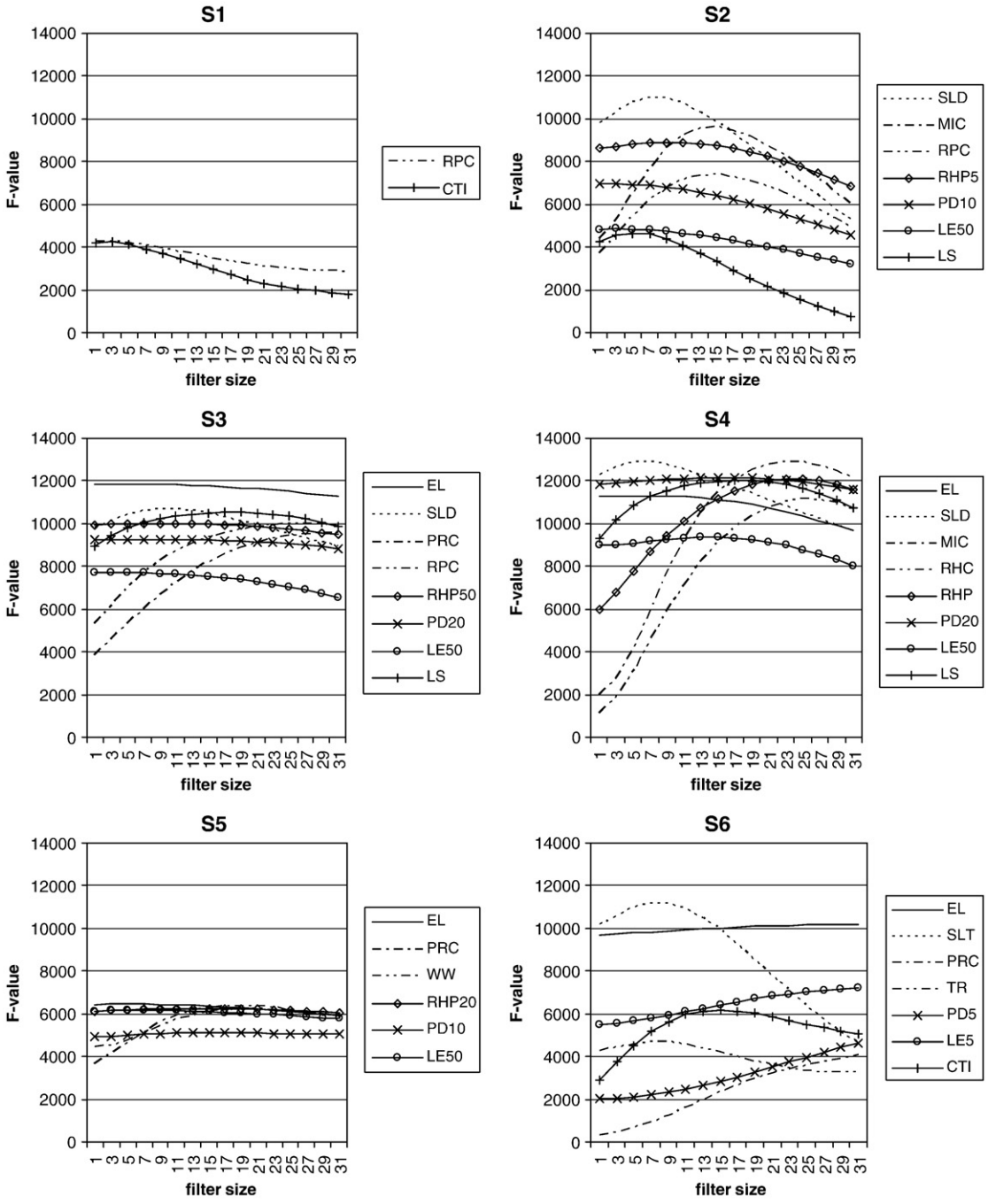

Fig. 4. ANOVA *F*-values plotted against filter sizes (*n* *n* pixel) for all terrain attributes selected by the ANOVA2 approach and all soil classes (S1–S6).

Interestingly, the accuracy in the validation area only increases slightly when all 608 predictor data sets are used. This demonstrates the high predictive power of these two attributes alone. The importance of these two variables can easily be interpreted in terms of their role in soil formation, as both were developed to describe soil–water and mass-movement relationships (Beven and Kirkby, 1979; Behrens, 2003). Most of the other soil classes can be described by a mixture of regional, local, and combined terrain attributes (Fig. 4). As the study area is a cuesta landscape, terrain position attributes (LE, RHP, and PD) become most important as they discern the vertical structure of the landscape. The vertical structure of the area is highly influenced by geology, which was not used as a predictor in this study (Schmidt et al., 2008). Slope plays an important role for 3 soil classes mainly occurring on steeper slopes.

The PCA approaches for feature selection returned comparatively weak prediction accuracies for four soil classes. Only for soil classes S2 and S4 PCA shows reasonable results. Two reasons might explain this fact: First, both soil classes are generally easy to predict as they show the highest F_1-measures in the validation area for all feature selection approaches. S2 are Haplic Cambisols occurring on steep slopes, while S4 are Haplic Podzols, which are located on the flat hilltops and thus contain a higher Loess component due to lower erosion rates. Second, both soil classes comprise about 30% of the training and the validation area each (Table 1). Thus, due to this large spatial extent, the selection of features, which describe the structure of the dataset (as derived by PCA) seems to be relevant for prediction, or at least does not seem to have a negative influence.

Concerning soil classes S3, S4, S5, and S6, all of which occur on steeper slopes, terrain attribute EL returns the highest ANOVA *F*-values (Table 4, Fig. 4) indicating their relation to geological settings in this area.

Another interesting aspect concerning filter sizes is, that for some soil classes (i.e. S2 and S4), nearly all relevant attributes vary substantially in optimal scale in terms of the ANOVA *F*-value, whereas, for some other soil classes (i.e. S1 and S3), this effect is comparatively moderate. This again seems to be related to the shape of areas occupied by these soil classes. S1 and S5 have the lowest area/perimeter ratio of all soil classes (Fig. 3); the smallest soil classes S1 and S3 (Table 1) return the weakest prediction results (Table 5). Hence, the filter effect increases in cases where the area of the soil class is large compared to the resolution of the DEM. In cases where the shape of the soil class is elongated and/or the soil class is comparatively small, the beneficial effect of the multi-scale approach is limited.

Based on the highest prediction accuracies obtained for soil classes 2 and 4, it can be hypothesized that these soil classes can be predicted well on the basis of terrain attributes only, whereas soil classes S1 and S3 might need other additional predictors to achieve best prediction results.

4. Conclusions

Several conclusions can be drawn based on the spatial data mining framework introduced in this study:

- multi-scale digital terrain analysis is an important tool for data mining based digital soil mapping approaches as it helps increase prediction accuracy compared to standard digital terrain analysis,
- each soil class is best predicted by different combinations of features filtered at different scales,
- only a small number of features are typically required to achieve good predictions,
- the selection of features in this way appears to aid the interpretation of the data with respect to soil formation,
- supervised feature selection approaches are superior to unsupervised ones,
- simple linear feature selection approaches are superior to much more complex approaches, and
- for straightforward predictions, multi co-linearity does not decrease prediction results substantially if robust algorithms, such as CRUISE, are used.

With respect to future digital soil mapping studies, it is important to validate these results in different soil landscapes. In the context of scale, wavelet analysis (Lark and Webster, 2001; Lark, 2006) should also be considered as an analytical tool. Furthermore, the proposed framework needs to be tested on soil attributes where similar results are expected (Behrens, 2003). Additional data mining and feature selection approaches need to be tested and compared. The interesting point in this respect is that the feature selection approaches applied can be combined with "black box" approaches like artificial neural networks. Feature selection, in combination with instance selection approaches (Schmidt et al., 2008), might help to further reduce model complexity while decreasing computation time and improving prediction accuracy.

Finally, we conclude that multi-scale and feature selection approaches have not yet received the attention they deserve in soil science literature. Both offer the possibility to learn more about soil formation in a spatial context and to increase the accuracy of digital soil mapping.

Acknowledgements

We are grateful to the Federal Geological Survey of Rhineland-Palatinate for providing the data.

A-Xing Zhu appreciates funding from the National Basic Research Program of China 2007CB407207, the Chinese Academy of Sciences International Partnership Project "Human Activities and Ecosystem Changes" CXTD-Z2005-1, the National Key Technology R&D Program of China (2007BAC15B01), and the support by the Vilas Trustees via its Vilas Associate Program to the University of Wisconsin-Madison.

References

Bartoli, F., Philippy, R., Doirisse, M., Niquet, S., Dubuit, M., 1991. The mass and porosity fractal dimensions (Dm and Dp) of silty and sandy soils were determined on in situ soils using a variety of soil sections (thin, very-thin and ultra-thin), by image analysis on a continuous scale from m to 10–9 to 10–1 m. European Journal of Soil Science 42, 167–185.

Bartoli, F., Genevois-Gomendy, V., Royer, J.J., Niquet, S., Vivier, H., Grayson, R., 2005. A multiscale study of silty soil structure. European Journal of Soil Science 56, 207–224.

Bay, D., 1999. Nearest neighbor classification from multiple feature subsets. Intelligent Data Analysis 3, 191–209.

Behrens, T., 2003. Digitale Reliefanalyse als Basis von Boden-Landschafts-Modellen – am Beispiel der Modellierung periglaziärer Lagen im Ostharz. Boden und Landschaft 42 42, 189 pp. Giessen.

Behrens, T., Scholten, T., 2006a. A comparison of data mining approaches in predictive soil mapping. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping – An Introductory Perspective. In: Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Behrens, T., Scholten, T., 2006b. Digital Soil Mapping in Germany – a review. J. Plant Nutr. Soil Sci. 169, 434–443 invited paper.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmidt, 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., 2008. An approach to removing uncertainties in nominal environmental covariates and soil class maps. In: Hartemink, A., McBratney, A., Mendoca-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer.

Bellmann, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton Univ, Press.

Beven, K., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Bulletin of Hydrologic Sciences 24, 4369.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 12.

Boden, Ad-Hoc-AG, 2005. Bodenkundliche Kartieranleitung. Schweitzerbart'sche Verlagsbuchhandlung, Hannover.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman L. and Cutler, A., 2004. Random Forests – Manual. http://oz.berkeley.edu/users/breiman/RandomForests/cc_manual.htm.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth, California.
Burrough, P.A., 1983. Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. Journal of Soil Science 34, 577–597.
Claessens, L., Heuvelink, G.B.M., Schoorl, J.M., Veldkamp, A., 2005. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. Earth Surface Processes and Landforms 30, 461–477.
Dietrich, W.E., Montgomery, D.R., 1998. A digital terrain model for mapping shallow landslide potential. http://socrates.berkeley.edu/~geomorph/shalstab.
Evans, I.S., 1972. General geomorphometry, derivatives of altitude, and descriptive statistics. In: Chorley, R.J. (Ed.), Spatial Analysis in Geomorphology. Methuen, London, pp. 17–90.
Evans, I.S., 1980. An integrated system of terrain analysis and slope mapping. Zeitschrift für Geomorphologie 36, 274–295.
Feldwisch, N., 1995. Hangneigung und Bodenerosion. Boden und Landschaft 3, Giessen.
Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305.
German, P., 1999. Scale dependence and scale invariance in hydrology. Soil Science 164, 694–695.
Goovaerts, P., Webster, R., 1994. Scale-dependent correlation between topsoil copper and cobalt concentrations in Scotland. European Journal of Soil Science 45, 79–95.
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.
Haralick, R.M., 1983. Ridge and valley detection on digital images. Computer Vision, Graphics and Image Processing 22, 28–38.
Hatfield, D.C., 1999. TopoTools – A Collection of Topographic Modeling Tools for ArcINFO. http://www.giscafe.com/GISVision/TechPaper/TopoTools.html.
Ho, K.L., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20. New York, pp. 832–844.
Ho, K.L., 2002. A data complexity analysis of comparative advantages of decision forest constructors. In: Pattern Analysis & Applications, vol. 5. London, pp. 102–112.
Horn, B.K.P., 1981. Hillshading and the reflectance map. Proceedings of the IEEE 69, 14–47.
Huber, M., 1994. The digital geo-ecological map – concepts, GIS-methods and case studies: Physiogeographica. Basel, 144 p.
Hupy, C.M., Schaetzl, R.J., Messina, J.P., Hupy, J.P., Delamater, P., Enander, H., Hughey, B.D., Boehm, R., Mitrok, M.J., Fashoway, M.T., 2004. Modeling the complexity of different, recently deglaciated soil landscapes as a function of map scale. Geoderma 123, 115–130.
Iba, W., Langley, P., 1992. Induction of one-level decision trees. Proc. of the 9th International Workshop on Machine Learning.
Jenny, H., 1941. Factors of Soil Formation. New York, McGraw-Hill. 281.
Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for geographic information systems analysis. Photogrammetric Engineering and Remote Sensing 54, 1593–1600.
John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. Proceedings of the Eleventh International Machine Learning Conference. Morgan Kaufmann, New Brunswick, pp. 121–129.
Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educational and psychological measurement 20, 141–151.
Kim, H., Loh, W.-Y., 2001. Classification trees with unbiased multiway splits. Journal of the American Statistical Association 96, 598–604.
Kleefisch, B., Köthe, R., 1993. Wege zur rechnergestützten bodenkundlichen Interpretation digitaler Reliefdaten. Geologisches Jahrbuch F, pp. 59–122.
Kohavi, R., John, G.H., 1997. Wrappers for features subset selection. Artificial Intelligence 97, 1–2.
Lagacherie, P., 2008. Digital soil mapping: a state of the art. In: Hartemink, A., McBratney, A. and Mendoca-Santos, M.L.: Digital Soil Mapping with Limited Data. Springer.
Lai, C., Reinders, M.J.T., Wessels, L., 2006. Random subspace method for multivariate feature selection. In: Patter Recognition Letters, vol. 27. New York, pp. 1067–1076.
Lark, R.M., 2005. Exploring scale-dependent correlation of soil properties by nested sampling. European Journal of Soil Science 56, 307–317.
Lark, R.M., 2006. Decomposing digital soil information by spatial scale. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping – An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.
Lark, R.M., Webster, R., 2001. Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. European Journal of Soil Science 52, 547–562.
Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery And Data Mining. Kluver.
Loh, W.-Y., Shih, Y.-S., 1997. Split selection methods for classification trees. Statistica Sinica 7, 815–840.
MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113, 81–109.
McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
McCool, D.K., Brown, L.C., Foster, G.R., Mutchler, C.K., Meyer, D., 1987. Revised slope steepness factor for the universal soil loss equation. Transactions of the ASAE 30, 1387–1396.
Mendonca-Santos, M.L., McBratney, A.B., Minasny, B., 2006. Soil prediction with spatially decomposed environmental factors. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping – An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.
Moran, C., Bui, E., 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science 16, 533–549.
Nemes, A., Schaap, M.G., Wösten, J.H.M., 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Science Society of America Journal 67, 1093–1102.
Pechenizkiy, M., Puuronen, S., Tsymbal, A., 2003. Feature extraction for classification in knowledge discovery systems. Proc. 7th Int. Conf. on Knowledge-based Intelligent Information & Engineering Systems, Berlin, pp. 526–532.
Pike, R.J., 1993. A bibliography of geomorphometry. United States Geological Survey Open-File Report 93-262-A, p. 132. Menlo Park, CA.
Quinn, P., Beven, K., Chevallier, P., Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes 5, 59–79.
Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146, 138–146.
Schoorl, J.M., Sonneveld, M.P.W., Veldkamp, A., 2000. Three-dimensional landscape process modelling: the effect of DEM resolution. Earth Surface Processes and Landforms 25, 1025–1034.
Shary, P.A., Sharaya, L.S., Mitusov, A.V., 2002. Fundamental quantitative methods of land surface analysis. Geoderma 107, 1–35.
Skurichina, M., Duin, R.P.W., 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications 5, 121–135.
Smith, M.P., Zhu, A.-X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137, 58–69.
Tarboton, D.G., 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research 33, 309–319.
Thompson, J.A., Bell, J.C., Butler, C.A., 1999. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modelling. Geoderma 100, 67–89.
Valeo, C., Moin, S.M.A., 2000. Grid-resolution effects on a model for integrating urban and rural areas. Hydrological Processes 14, 2505–2525.
Van Rijsbergen, C.J., 1979. Information Retrieval. Butterworth, London.
Vázquez, R.F., Feyen, L., Feyen, J., Refsgaard, J.C., 2002. Effect of grid size on effective parameters and model performance of the MIKE-SHE code. Hydrological Processes 16, 355–372.
Wischmeier, W.H., Smith, D.D., 1978. Predicting Rainfall Erosion Losses – A Guide to Conservation Planning. USDA Handbook 537. InUS Government Printing Office, Washington, DC.
Wood, J., 1996. The geomorphological characterization of digital elevation models. PhD Thesis. University of Leichester, UK.
Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes Landforms 12, 47–56.
Zhao, Z., Liu, H., 2007. Searching for interacting features. Proc. of International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India.
Zhao, Y., Shi, X., Weindorf, D.C., Yu, D., Sun, W., Wang, H., 2006. Map scale effects on soil organic carbon stock estimation in North China. Soil Science Society of America Journal 70, 1377–1386.
Zhu, A.X., 2008. Spatial scale and neighborhood size in spatial data processing for modeling the natural environment. In: Mount, N.J., Harvey, G.L., Aplin, P., Priestnall, G. (Eds.), Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13. InCRC Press, Florida.
Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65, 1463–1472.
Zhu, J., Morgan, C.L.S., Norman, J.M., Yue, W., Lowery, B., 2004. Combined mapping of soil properties using a multi-scale tree-structured spatial model. Geoderma 118, 321–334.