European Journal of Soil Science

European Journal of Soil Science, February 2010, **61**, 133–143 doi: 10.1111/j.1365-2389.2009.01205.x

The ConMap approach for terrain-based digital soil mapping

T. Behrens[a], K. Schmidt[a], A. X. Zhu[b,c] & T. 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, and* [c]*Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA*

Summary

We present a new digital terrain analysis framework for digital soil mapping, referred to as contextual elevation mapping (ConMap). In contrast to common regression approaches based on features from digital terrain analysis, ConMap is not based on standard terrain attributes, but on elevation differences from the centre pixel to each pixel in circular neighbourhoods only. These differences are used as features in random forest regressions. We applied and validated the framework by predicting topsoil silt content in a loess region of 1150 km^2 in Rhineland-Palatinate and Hesse, Germany. Three hundred and forty-two samples and a 20-m resolution digital elevation model were used for this illustration and validation. We compared ConMap with standard and multi-scale terrain analysis approaches as well as with ordinary kriging interpolations. Cross-validation root mean square error (*RMSE)* decreased from 16.1 when the standard digital terrain analysis was used to 11.2 when ConMap was used. This corresponds to an increase in variance explained (R^2) from 15 to 61%. Even though ordinary kriging out-performed standard terrain analysis as well, the variance explained was 6% smaller compared with that using ConMap. The results show that the geomorphic settings in the study area must have induced the spatial trend, which can be accounted for by ConMap over different scales. We conclude that ConMap shows great potential for digital soil mapping studies.

Introduction

There are a number of studies on digital soil mapping that document successful attempts to interpolate and extrapolate soil properties and soil classes (Zhu *et al*., 2001; McBratney *et al*., 2003; Behrens & Scholten, 2006). Many of these studies rely on terrain attributes (McBratney *et al*., 2003), indicating the importance of topography as a soil-forming factor (Jenny, 1941; Gerrard, 1981). However, a discussion on integrating topography at larger spatial scales has been limited in digital soil mapping research (Lagacherie, 2008).

In pedology, soilscapes are characterized by a typical spatial pattern and typical taxonomic relationships of soils, as well as by their relationship to landform or landscape characteristics at different scales (Hole, 1978; Gerrard, 1981). This relationship is of great importance for digital soil mapping at the landscape scale, as geomorphic settings in larger neighbourhoods might influence local climate conditions, such as precipitation patterns. Hence, soil properties might be different even if all other state factors such as local relief and parent material do not change.

In order to take the geomorphic settings over larger neighbourhoods into account, measures of local and regional geomorphic context are necessary (MacMillan, 2004). For extrapolating soil mapping units or soil properties using pedometric approaches it is therefore important to incorporate information about the soil-forming factors not only from the specific point where a soil class or a specific soil property value is measured, but also from its larger spatial arrangement.

There are two major reasons for developing a new terrain-based approach for digital soil mapping. The first is the variety of algorithms for deriving terrain attributes such as slope, curvature or contributing area available. Slope, as the most prominent example, can be calculated in many different ways on the basis of grid datasets (e.g. Evans, 1980; Zevenbergen & Thorne, 1987; Wood, 1996). All approaches show different results. Hence, in applied digital soil mapping several problems arise and different soil properties might be sensitive to different algorithms of the 'same' terrain attribute. Thus, multiple versions of one terrain attribute have to be tested in one regression approach.

Correspondence: T. Behrens. E-mail: thorsten.behrens@uni-tuebingen.de
Received 18 March 2009; revised version accepted 16 September 2009

Additionally, various special software packages are needed and various data formats have to be handled.

The second reason for developing a new terrain-based concept is that straightforward approaches for handling multi-scale variations in digital soil mapping are missing (Lagacherie, 2008). In contrast to local terrain attributes such as slope, calculated on the basis of local 3 × 3 pixel neighbourhoods, values of regional terrain attributes, such as contributing area, are determined over larger, irregular surface areas. Hence, a combination of both types of terrain attributes is generally recommended. However, these types of terrain attributes can only be used to describe soil variations caused by gravitational processes on a hillslope (i.e. only the site and the catena scale are covered). Therefore, standard terrain analysis is limited when larger geomorphic arrangements influence soil properties at a specific point in a landscape.

Existing approaches used to account for multiple scales can be divided into three groups: (i) those that derive terrain attributes from different neighbourhood sizes, (ii) those that apply filters to digital elevation model (DEM)-derived attributes, and (iii) those based on wavelet analysis. These approaches are now briefly described.

Some terrain attributes based on context filters such as 'elevation percentile' (Gallant & Hutchinson, 2008) can easily be calculated on the basis of different neighbourhood sizes and thus account for larger scales (Behrens, 2003). Gallant & Hutchinson (2008) recommend setting the neighbourhood radius to the size of the average hillslope length. Smith *et al*. (2006), in a similar method to Wood's (1996), used variable neighbourhood sizes to derive local terrain attributes such as slope and aspect based on quadratic approximations. Zhu (2008) and Zhu *et al*. (2008) examined the sensitivity of computed terrain derivatives to the combined effect of DEM resolution and neighbourhood size. Moran & Bui (2002) applied an approach to incorporate the spatial context by using an adaptive filtering scheme based on the analysis of local variograms. The resulting adaptive filter with variable window sizes was then applied to the DEM. Behrens *et al*. (2005) and Grinand *et al*. (2008) used averaging filters on terrain attributes to incorporate contextual spatial information. Behrens *et al*. (2009) used this approach for a systematic analysis of scale. Lark (2007) and Mendonça-Santos *et al*. (2007) demonstrated the ease of use of wavelets to integrate different spatial scales in the mapping process. However, these studies were not directly focussing on multi-scale terrain variations. All studies show that incorporating spatial context improves prediction results and that the accuracy of the digital soil mapping models can vary significantly depending on scale.

The methods mentioned above cannot be compared with each other directly. The averaging-filter approaches and the wavelets approach can be applied on any terrain attribute, whereas the extension of quadratic approximations is only suitable for local terrain attributes such as slope, aspect or curvature (Wood, 1996). A general problem that all three groups of approaches have in common when integrating larger scales is that the approaches are either complex or specific. Additionally, only indirect and non-directional information is generated at different scales, and thus there are no direct measurements of surface properties and processes in larger areas. However, directional information may be important if a factor influencing soil spatial distribution has a directional component such as wind direction. The interaction between geomorphology and such factors could result in a trend in soil properties.

To overcome these constraints, we developed a new digital terrain analysis framework for digital soil mapping, which is based on spatial data-mining using contextual elevation data, and named ConMap. The main objective of ConMap is that a range of spatial neighbourhoods can be analysed on the basis of very simple topographic indicators. Therefore, it comprises complex surface function approximations described using standard terrain analysis as well as factors in the geomorphic arrangement over larger neighbourhoods.

Materials and methods

Study area and datasets

To test the ConMap approach we predicted topsoil silt content for an area of approximately 1150 km^2. The study area is located in Rhineland-Palatinate at the border with Hesse in Germany (Figure 1). It covers the transition from the Upper Rhine Graben to the Middle Rhine Valley region. The central region covering the sample locations comprises the northern part of the wine-growing and loess-covered region 'Rhine-Hesse'. The surrounding area comprises a second wine-growing and loess-covered region, the 'Rhinegau', and parts of the Taunus mountains in Hesse, and of the Hunsrück mountains in Rhineland-Palatinate. The loess regions, with an average elevation of 200 m above sea level, are located on the leeward sides of the surrounding mountains, which have elevations up to 700 m in the study area. They are relatively warm and dry, with mean annual precipitations of 500–550 mm. In contrast, precipitation in the mountainous regions rises up to 850 mm. The soils found in this area are therefore diverse, ranging from Luvisols in the loess area to Cambisols and Podzols in the mountainous regions, which are dominated by quartzite. The loess, which was deposited in the last glacial period in the Pleistocene epoch (Würm glaciation), originates from the surrounding riverbeds (Figure 1), especially from the Rhine-Main lowlands (Schönhals, 1996). Because of the strong influence of loess, this region is considered as an optimal test case for a contextual mapping approach, as topsoil silt content is a function of regional loess translocation under periglacial conditions. The loess distribution is mainly driven by local climate characteristics of the wind and precipitation patterns, which are influenced by the geomorphic settings in the study area at a regional scale. Additionally, subsequent Holocene erosion and accumulation processes are of importance, which are again functions of geomorphic settings and precipitation patterns, as well as land use. The latter is not considered in this study.

We mapped topsoil silt content of the 0–10 cm-depth interval on the basis of 342 samples. Samples were collected in the 1990s

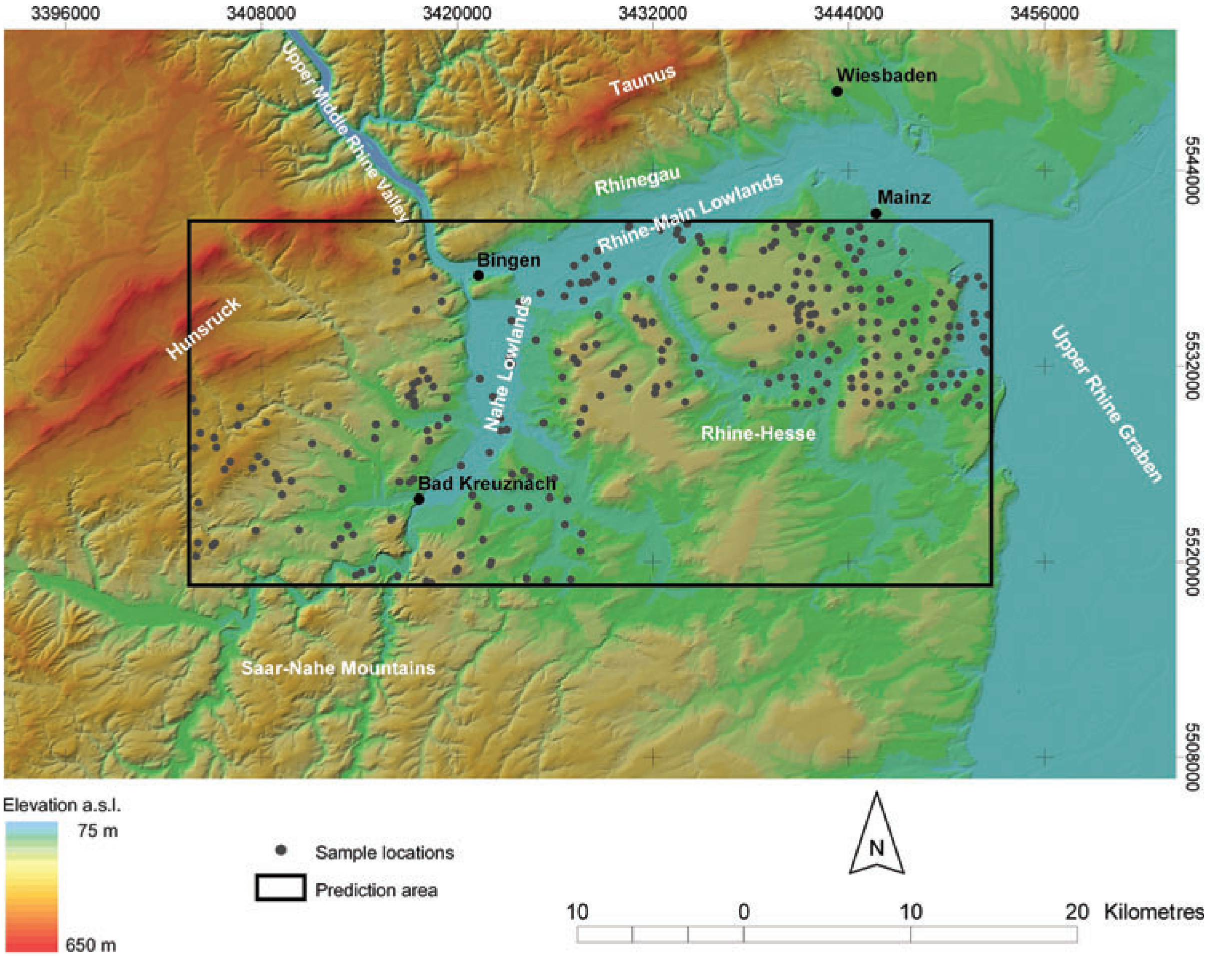

Figure 1 Study area (Rhineland-Palatinate/Hesse, Germany) with landscape units. The core rectangle shows the final prediction area; the surrounding area represents the buffer of 600 pixels (12 000 m) required for ConMap predictions.

as part of soil surveys at a scale of 1:25 000. Information about sampling schemes, positional accuracies and measurement errors is not available. The spatial distribution of topsoil silt content shows the smallest values (<5%) in the Rhine-Main lowlands and the largest values (>80%) in the eastern part of Rhine-Hesse (Figure 2).

Even though all samples were located in Rhineland-Palatinate, the DEM used in this study also covers parts of Hesse. Therefore, we merged the two DEMs of Rhineland-Palatinate and Hesse. As the resolutions were different (10 m in Rhineland-Palatinate and 20 m in Hesse), we re-sampled the 10-m DEM to a resolution of 20 m before merging. The vertical accuracy for both DEMs is ± 0.5 m.

Contextual elevation mapping; ConMap

The approach. ConMap is based on the differences in elevation from each pixel in a circular neighbourhood to the centre pixel. The data are stored in a table in X, Y, Z_1, Z_2, $\ldots Z_n$ format, where X and Y are the geographical coordinates and $Z_1 - Z_n$ are the columns for the elevation differences. The differences are calculated for every pixel over the study area and are directly used as features (independent variables; predictors) without any further mathematical or statistical processing in the learning approach to discern the relationships between these differences and the soil property in question through the sample points.

In contrast to common digital soil mapping studies with the ConMap approach, we did not use standard terrain attributes as features in the regression approach, but more simple topographical indicators (the elevation differences computed above). The use of such simple indicators is the major difference and the major advantage of the method compared with common terrain attributes because no complex mathematical functions are used and no decision has to be made as to which specific algorithm to use for computing the terrain attribute.

Another advantage is the ability to capture contextual information by simply extending the neighbourhood size in which elevation differences are calculated. The neighbourhood used can be extended to any size to include regional surface shapes without increasing mathematical complexity. The size is only restricted to

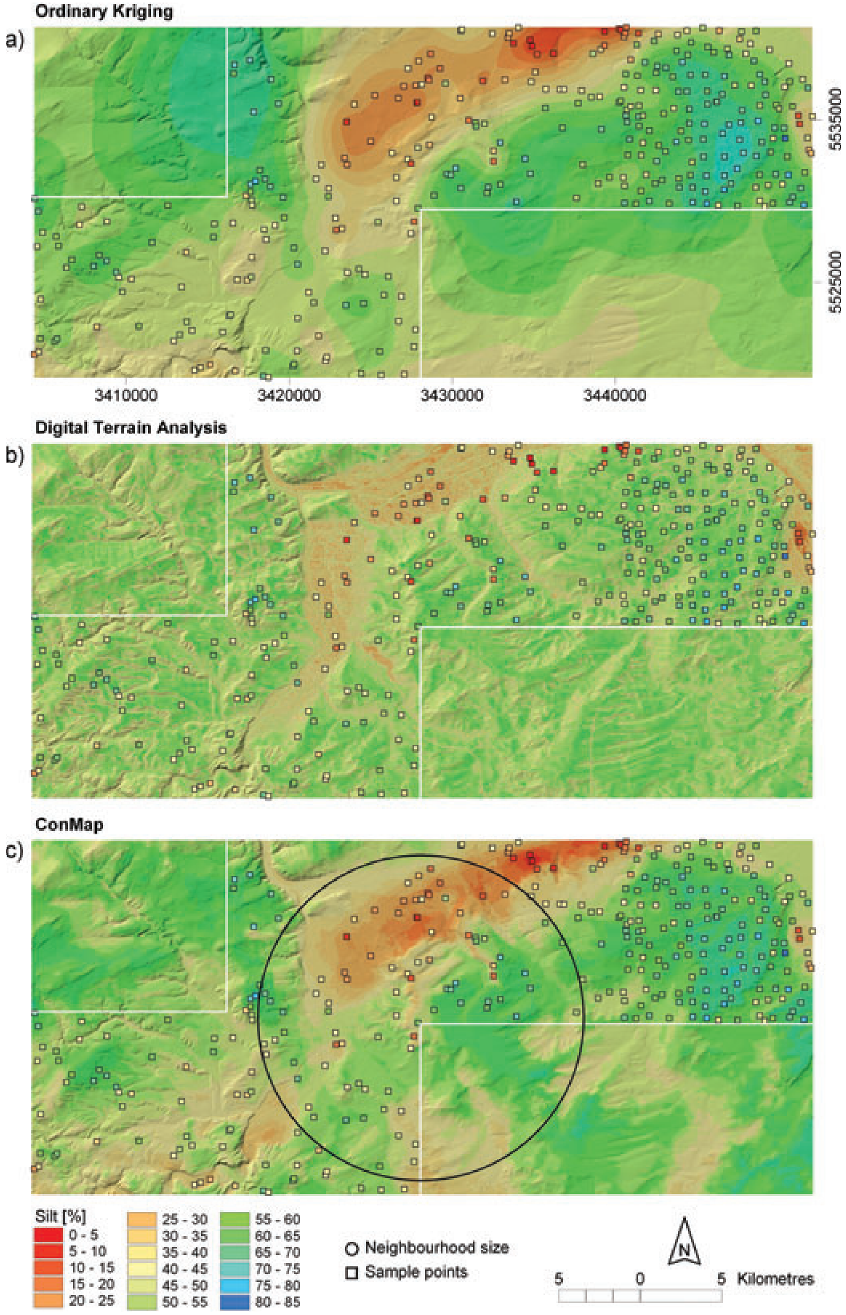

Figure 2 Interpolation and prediction results of topsoil silt content; (a) ordinary kriging; (b) digital terrain analysis; (c) ConMap.

the extent of the DEM available. As all neighbourhood sizes are included in one training dataset, ConMap comprises local surface functions as well as spatial arrangements of broader geomorphologic entities in larger spatial contexts.

To determine the optimal size of the moving window (the neighbourhood), which is analogous to the optimal integration of different scales, and to analyse the change in prediction accuracy in relation to different scales, we used a stepwise approach over different circular neighbourhood sizes. Because of the size of the study area, we compared neighbourhood radii of $ra = 2$, 5, 10, 20, 40, 60, …, 600 pixels, which corresponds to a maximum diameter of 24 km based on a resolution of 20 m. Using large neighbourhoods has two major limitations: (i) when the neighbourhood is too large it will run into the boundary of DEM, and (ii) the number of features constructed, which is equivalent to the number of pixels in the circular neighbourhood, can be very large. For example, a radius of 100 pixels would result in 31 416 features. The problem of multicollinearity can occur and the principle of parsimony has to be considered.

Reduction of the contextual spatial feature space. Spatial density functions can be applied to reduce the large feature space,

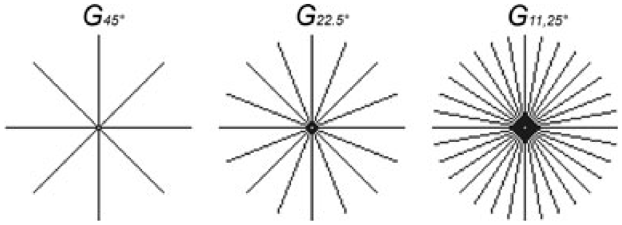

Figure 3 Spatial contextual moving window kernels for different spatial density sampling functions of G_{45°, $G_{22.5^\circ}$ and $G_{11.25^\circ}$ for a radius of 30 pixels.

which would emerge if all pixels in large neighbourhoods were taken into account. This can be achieved if pixels more remote from the centre pixel are only accounted for sparsely. Many spatial sampling approaches to solve this task seem reasonable. In this study we tested three geometrical approaches (G) where only those pixels were sampled that form rays at certain angles, originating from the window centre in a star-shaped manner. We chose 45°, 22.5° and 11.25° intervals (Figure 3). Hence, the amount of features (d) per kernel is a function of the radius (ra) expressed as the number of pixels (Equations 1–3):

$$G_{45^\circ}: \quad d_{45} \approx 4ra + 4\frac{ra}{\sqrt{2}}, \tag{1}$$

$$G_{22.5^\circ}: \quad d_{22.5} \approx 4ra + 4\frac{ra}{\sqrt{2}} + 8\frac{ra}{1.0824}, \tag{2}$$

$$G_{11.25^\circ}: \quad d_{11.25} \approx 4ra + 4\frac{ra}{\sqrt{2}} + 8\frac{ra}{1.2027} + 8\frac{ra}{1.0824} + 8\frac{ra}{1.0196}. \tag{3}$$

Because of the fixed pixel size and partial overlaps of neighbouring rays near the centre pixel, these formulae must be regarded as approximations. Compared with a neighbourhood where all pixels are included this leads to a reduction (*red*) of the feature space (%) according to:

$$G_{45^\circ}: \quad red_{45} \approx \left(1 - \frac{d_{45}}{\pi^* ra^2}\right)^* 100, \tag{4}$$

$$G_{22.5^\circ}: \quad red_{22.5} \approx \left(1 - \frac{d_{22.5}}{\pi^* ra^2}\right)^* 100, \tag{5}$$

$$G_{11.25^\circ}: \quad red_{11.25} \approx \left(1 - \frac{d_{11.25}}{\pi^* ra^2}\right)^* 100. \tag{6}$$

For a radius of 100 pixels, the G_{45° neighbourhood consists of 683 features, which is a reduction of 97.8% compared with the 31 416 features of the full neighbourhood.

Spatial prediction and analysis using the random forests approach. In recent years, techniques for combining multiple predictions, known as ensemble approaches, have become very popular, as ensemble predictions are generally more accurate than individual predictions (Maclin & Opitz, 1997). Ensembles are aggregations of multiple predictions often based on changes in the training dataset resulting from re-sampling. Two intuitive, simple-to-implement, yet powerful approaches are bagging (Breiman, 1996) and the random subspace method (Ho, 1998). Bagging is an instance (or sample)-based approach where multiple predictions on individual bootstrap replicates (random sampling with replacement; Efron & Tibshirani, 1993) of the original dataset are averaged. The random subspace method is a feature-based approach where averaging is performed over multiple randomized feature subsets.

The random forests approach (Breiman, 2001) is a combination of both techniques where multiple classification or regression trees (Breiman *et al*., 1984) are generated. For regression the final prediction is the average of the suite of individual tree outputs (Breiman, 2001). The random forests model has been applied in digital soil mapping by Grimm *et al*. (2008) and various other environmental mapping applications (e.g. Cutler *et al*., 2007; Peters *et al*., 2007). We used the 'Random Forest' package (Liaw & Wiener, 2009) for the R statistical language (R Development Core Team, 2009).

The random forests approach provides options to analyse feature importance (Breiman, 2001). Therefore, each feature is randomly permuted at each split and the rate of change of the out-of-bag (OOB) data left out in each bootstrap sample when generating a single tree in the forest) error compared with the original feature is used as an indicator for its importance (Breiman, 2001; Grimm *et al*., 2008). As this measure is tested against the independent OOB datasets it does not over-fit (Prasad *et al*., 2006) and it is unbiased towards different state spaces of the predictors if continuous variables are used solely (Strobl *et al*., 2007), as in the present study.

Optimization. To optimize the random forests model, we systematically tested different parameter settings in a grid- or hyper-learning approach (Schmidt *et al*., 2008). The parameter that we optimized was 'mtry', which specifies the number of variables selected randomly at each split in a tree. If all variables are tested for each split, random forests is similar to an ensemble of bagged trees, as no random feature subsets are used.

Validation

Accuracy measures. Begleiter & El-Yaniv (2008) argue that in order to obtain an honest assessment of the total system's performance, both the estimation of mtry and the accuracy estimation of the supervised training should be embedded within a cross-validation procedure (Kohavi, 1995). Therefore, we used a 10-fold cross-validation in this study to avoid over-fitting and to handle the lack of parsimony (McBratney *et al*., 2003) arising from the large number of features generated using ConMap. The root mean square error (RMSE) was used to determine the optimal model parameters for ConMap:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2}, \tag{7}$$

where X_i are the observed values and Y_i are the predicted values. We also report the corresponding coefficient of determination (R^2):

$$R^2 = \frac{\left(\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})\right)^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2 \sum_{i=1}^{N} (Y_i - \overline{Y})^2}. \tag{8}$$

As Random Forest relies on multiple random components in model building and cross-validation is also based on randomized data partitioning, we report average prediction accuracies of three independent model runs and the corresponding standard deviations.

Comparison with standard and multi-scale terrain analysis. We compared the ConMap approach with standard, as well as with the multi-scale terrain analysis approaches introduced by Behrens *et al.* (2009), and using the following terrain attributes, which are described in detail in Behrens *et al.* (2009): elevation, steepest slope, deviation from bearing (0°, 45°, 90° and 135°), mean curvature, minimum curvature, maximum curvature, relative profile curvature, relative horizontal curvature, topographic roughness, contributing area, compound topographic index, relative hillslope position, local elevation and distance to channels. The same prediction and validation approaches as used for ConMap were applied. In contrast to ConMap, we carefully pre-processed the DEM in terms of filtering noise and filling sinks and pits.

In the comparison with the multi-scale approach introduced by Behrens *et al.* (2009), calculation times limit the maximum neighbourhood size that can be tested. For example, calculating the average for a circular neighbourhood with a diameter of 24 km for one single terrain attribute in the study area takes approximately 70 hours on a 3 GHz PC using ArcView GIS. Compared with this, generating both the training and the prediction datasets for all neighbourhood sizes for ConMap takes approximately 40 hours.

Even though this is based on a multi-threaded version of ConMap using four CPU cores instead of the single core filtering approach, filtering is computationally much more demanding. As different radii need to be calculated for all terrain attributes, the filter approach is not reasonably applicable for such large neighbourhoods. Because of these computational limitations we used only the following filter sizes for the multi-scale approach: 140 × 140, 300 × 300, 460 × 460 and 620 × 620 m. This size is also comparable to the average hillslope length in the area, as recommended by Gallant & Hutchinson (2008) for contextual terrain attributes, as well as to the filter sizes tested in the study by Behrens *et al.* (2009). However, the maximum extent is not comparable to the ConMap approach.

Comparison with spatial interpolation using kriging

Because of the visible spatial trend in the topsoil silt-distribution in the sample dataset (Figure 2a) we compared ConMap with ordinary kriging (Goovaerts, 1997) as an additional candidate prediction method. Ordinary kriging is one of the most frequently used geostatistical estimators (Siska *et al.*, 2005) that yields accurate results (Moyeed & Papritz, 2002). We used the gstat library (Pebesma, 2004) for the R statistical language (R Development Core Team, 2009). As we did for ConMap, we used 10-fold cross-validation to estimate kriging interpolation performance.

Results and discussion

ConMap

Because of the large number of features for high-density neighbourhoods (as with $G_{22.5^\circ}$ and $G_{11.25^\circ}$, see Figure 3) and computational limitations of the software used, predictions were only possible for all densities with neighbourhood diameters below 7.2 km. This corresponds to a maximum of 5743 features for the $G_{11.25^\circ}$ feature density and a radius of 180 pixels. The differences in prediction accuracy regarding the densities of the different neighbourhoods were marginal. The averaged *RMSE* values over all neighbourhoods from ra = 2 up to ra = 180 pixels were 14.3 (G_{45°), 14.5 ($G_{22.5^\circ}$) and 14.5 ($G_{11.25^\circ}$). As there were no significant differences in the *RMSE* values, we used the G_{45° neighbourhood for further analysis. This follows the principle of parsimony as the G_{45° neighbourhood comprised the smallest number of features. Furthermore, it also allows for the largest neighbourhoods.

Figure 4 shows the prediction results (*RMSE*) for the G_{45° approach for all neighbourhood sizes tested in this study. It can be seen that with larger neighbourhoods much better prediction results were achieved than with small neighbourhoods. Prediction accuracies obtained with ConMap start with an *RMSE* of 15.9 ($R^2 = 0.16$) for a radius of 40 m. The best prediction results with an *RMSE* of 11.2 ($R^2 = 0.61$) were achieved for diameters greater than 20 km. The increase in prediction accuracy is non-linear and not uniformly continuous, as shown by an additional increase in prediction accuracy between neighbourhood diameters of 15 and 20 km. This additional increase can be interpreted as another important scale or surface property emerging in this spatial neighbourhood range. Hence, important geomorphic units driving loess distribution can be accounted for. Figure 2(c) shows the spatial predictions using a neighbourhood diameter of 20 km. Figure 5 shows the corresponding scatterplot of observed against predicted values.

The standard deviation of the variance explained across the three independent model runs and all neighbourhood sizes tested was 2% on average. This comprises the random effects of the random forests modelling approach (bagging and the random subspace method) as well as of the 10-fold cross-validation procedure. Therefore, the influence of the random effects can be regarded as being marginal.

The feature importance for the final ConMap model (Figure 2c) using the G_{45° neighbourhood is shown in Figure 6 and Figure 7. On average, across all rays of the G_{45° neighbourhood, four zones

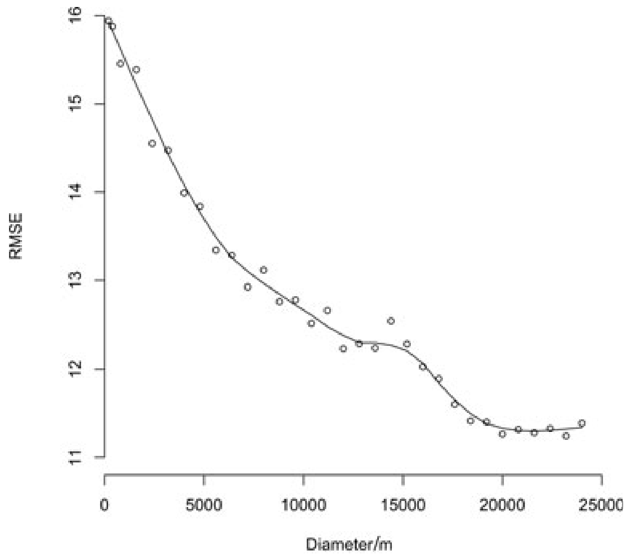

Figure 4 Neighbourhood size plotted against random forests prediction results for topsoil silt content (*RMSE*). The line shows averaged results using a lowess smoother.

can be identified as important drivers or scales for topsoil silt content distribution with radii ranges of approximately 200–750, 2500–4200, 6250–7000 and 8750–10 000 m (Figure 7). The first zone is related to the site and the catena scale, whereas the others are measures of regional spatial context as demanded by MacMillan (2004). The spatial anisotropy in the feature importance reflects the importance of accounting for directional components in terrain-based digital soil mapping. The latter is most important, as all other approaches discussed in the present study rely only on non-directional information. Further studies are needed to further investigate the relationship between anisotropy and pedogenesis.

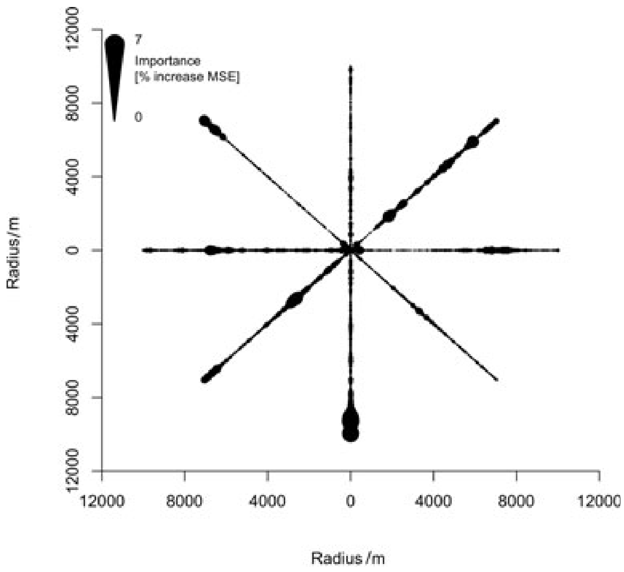

Figure 6 Feature importance of the final ConMap prediction model (Figure 2c) using a radius of 500 pixels and the G_{45° neighbourhood kernel. The larger the dots the greater the importance of each feature compared with randomly permutated version of the features.

Comparison with standard and multi-scale terrain analysis

Standard terrain analysis returned an *RMSE* of 16.1 ($R^2 = 0.15$), the largest value for *RMSE* of all approaches tested in this study. The difference in prediction results can be analysed when the statistical distributions of the prediction results are compared across all approaches. Figure 5 shows that the range of the predicted topsoil silt contents was much smaller compared with the sample set as well as with the ConMap and kriging predictions. The multi-scale approach comprising neighbourhood diameters up to 620 m increased prediction accuracy to an *RMSE* of 15.5

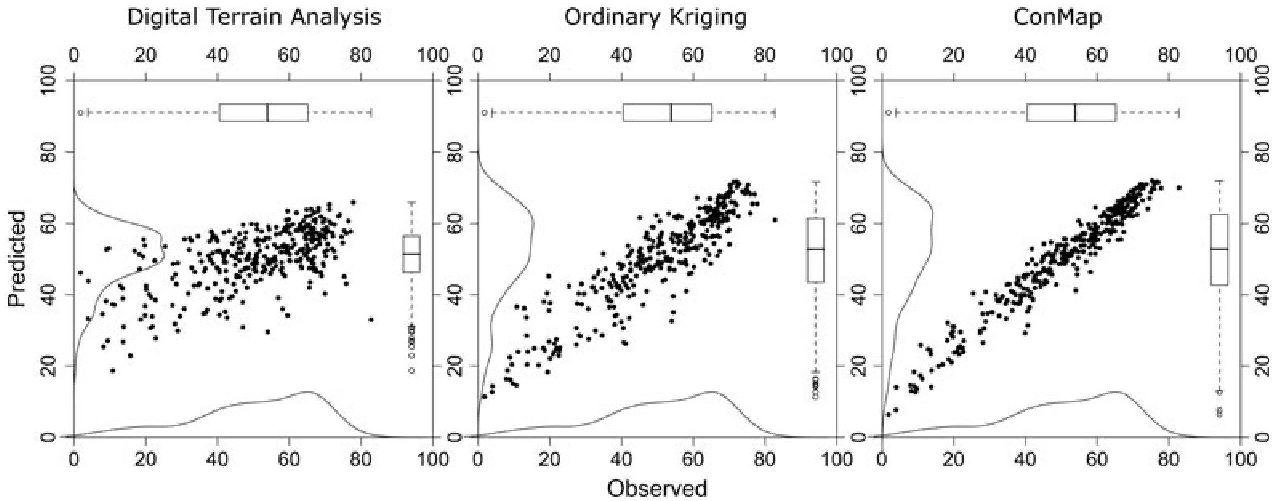

Figure 5 Observed plotted against predicted scatterplots overlaid with boxplots and kernel density plots of topsoil silt content for standard terrain analysis, ordinary kriging and ConMap predictions.

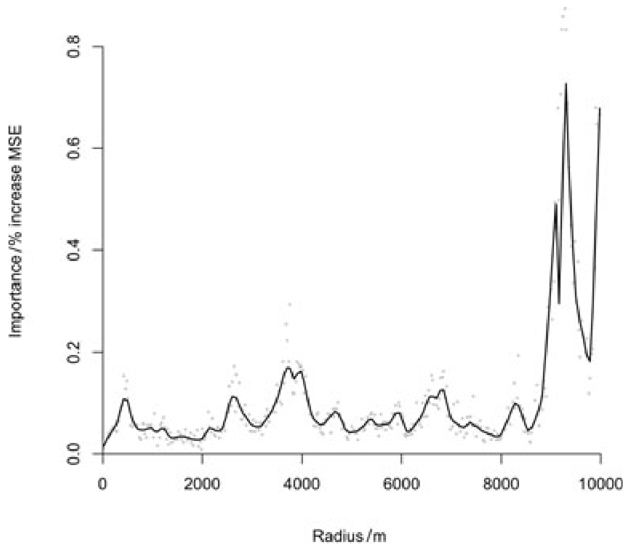

Figure 7 Averaged feature importance over all rays of the G_{45° neighbourhood kernel with a radius of 10 000 pixels used for the final ConMap prediction model (Figure 2c). The dots show the averaged values, the smoothed line the corresponding lowess curve.

($R^2 = 0.21$). ConMap provided similar results at the same neighbourhood diameter.

Comparison with spatial interpolation using kriging

The experimental and the theoretical semi-variogram of the topsoil silt content (Figure 8) had a pronounced spatial trend. Figure 2(a) shows the corresponding kriging interpolation results using gstat (Pebesma, 2004) and using a spherical semi-variogram model with a range of 15 556 m, a nugget variance of 87 and a sill value of 397. The cross-validation error was 11.7 (*RMSE*) with a corresponding R^2 value of 0.55. The standard deviation across the three independent model runs was 1% in terms of variance explained. The statistical distribution of topsoil silt-content was comparable to the ConMap predictions. However, the range of the predicted values was smaller (Figure 5). Up to neighbourhood sizes of 15 000 m in diameter there was a relationship between spatial correlation as shown by the variogram (Figure 8) and the prediction results for the G_{45° approach across all neighbourhood sizes (Figure 4). However, there was an additional increase in prediction accuracy in larger neighbourhoods up to a maximum neighbourhood diameter of 20 000 m, which is not discerned from the variogram.

The lack of auto-correlation of silt content between pixels >15 000 m apart, but a presence of correlation with terrain, may result from the random forests method, which is non-linear and non-parametric and can cope with complex multivariate interactions of predictors (Strobl *et al*., 2007, 2008). Further studies are needed to examine the general relation between the variogram and prediction results across all neighbourhood sizes as well as the effect of performance increase above the range of the variogram.

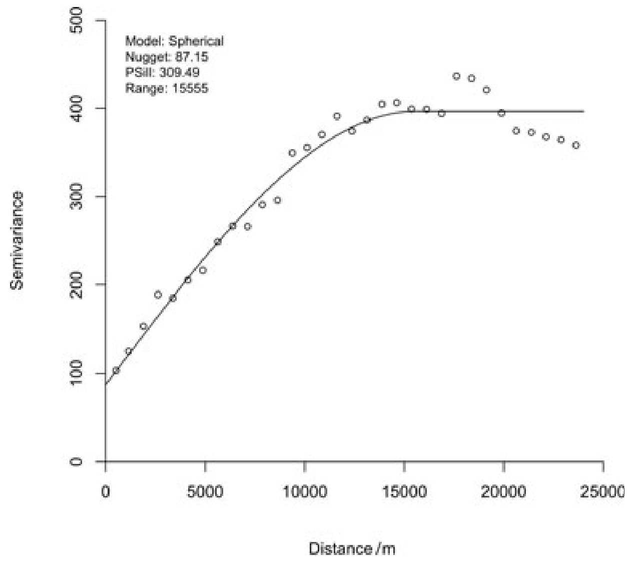

Figure 8 Experimental and theoretical variogram for topsoil silt content.

Integration of scales

One major advantage of ConMap for digital soil mapping applications is the fact that ConMap accounts for multi-scale variation over large neighbourhoods. With regard to the concept of spatial hierarchy of land units for soil and land resource surveys as introduced by Gallant *et al*. (2008), ConMap allows the integration of impacts across multiple scales from the site to broad physiography in one approach, and increases the accuracy of prediction. Generally, (digital) soil mapping requires that the process scales match the observation scales. This ‘concept of emergent properties’ (Gallant *et al*., 2008) is implicit in ConMap, as multiple terrain is featured at different scales and is adaptively chosen in a single regression approach. Thus, ConMap integrates across scales, but does not down- or up-scale terrain features. The relevant process scales are expressed and can be analysed using feature selection approaches, even though further research on this topic is still needed.

Technical aspects

Because it is based on elevation differences to the centre pixel only, ConMap is a rather simplistic approach in terms of terrain model complexity as compared with some common terrain analysis approaches (Shary *et al*., 2002). However, it generates many more features. As Occam’s razor only applies for models with similar prediction accuracy, ConMap should be preferred over common-terrain analysis approaches. Nevertheless, regression approaches resistant to over-fitting and multicollinearity, such as random forests, partial least squares regression or Support Vector Machines, have to be used.

The two major limitations of the ConMap approach are that the optimum neighbourhood size cannot be determined *a priori* and that, with regard to common pedogenetic concepts, ConMap cannot be used to identify relevant features such as slope even though ConMap allows us to analyse the influence of scale, which might give an insight into the pedogenetic processes beyond the catena scale.

Performance

The most important result of this study is that prediction accuracy for the topsoil silt content increased remarkably with larger neighbourhood sizes. ConMap out-performs standard digital terrain analysis by a factor of four in terms of the variance explained. With multi-scale digital terrain analysis as tested in the present study, ConMap has similar prediction accuracies at the same, yet relatively small, neighbourhood range. However, because of computational costs, the multi-scale approach cannot be reasonably extended to the same neighbourhood sizes as used with ConMap. In this respect other approaches that take account of scale, such as wavelet transforms, should be tested in further comparative studies.

Because of the clear spatial trend in topsoil silt-content (Figure 2a and Figure 8), which is a reflection of the local loess translocation in the study area, ordinary kriging also had much better prediction results than standard or multi-scale terrain analysis. The statistical distributions obtained from kriging and ConMap were comparable. However, the variance explained by ConMap was 6% better than that by kriging.

ConMap and the scorpan model

The previous discussion leads to two major findings of this study: (i) as ConMap is based on elevation differences only, the trend observed in the data must be induced by the geomorphic settings in the study area driving local climate pattern, and (ii) ConMap can account indirectly for spatial trends. Hence, for further methodological and conceptual analysis, interpretation and evaluation of the ConMap approach, the inclusion of the spatial domain in prediction approaches has to be discussed.

In this context, the ConMap approach is a typical example for the scorpan paradigm introduced for quantitative empirical digital soil-mapping (McBratney *et al*., 2003). Under the terms of this modelling approach, a soil property or a soil class (S_a *or* S_c) is modelled as a function of other soil properties (s), climate (c), organisms (o), relief (r), parent material (p), age (a) and space or spatial position (n). Because it is an important factor for quantitative empirical modelling, McBratney *et al*. (2003) added the factor (n) to their scorpan extension of Jenny's (1941) famous state factor concept. There are two main concepts directly designed to operate in the spatial domain, those of kriging (Matheron, 1960) and geographically weighted regression (Brunsdon *et al*., 1996). To integrate the factor n into regression or supervised classification predictions, two approaches, direct (n_d) and indirect (n_i), seem reasonable. The most straightforward n_d approach is to add the X and Y coordinates as additional features and is therefore based on absolute positional information. Hence, it describes the entire dataset and can be used to map spatial trends directly as in trend surface analysis. With the multi-scale terrain analysis approaches introduced by Moran & Bui (2002), Wood (1996), Behrens *et al*. (2009), and Zhu *et al*. (2008), a spatial component is included by means of the size of the neighbourhood window. Therefore, these approaches integrate space, but do this indirectly (n_i). In this respect, ConMap can immediately account for space in varying neighbourhoods. Additionally, it comprises information on the relative position within this window space, which can be analysed using measures of feature importance.

The present study gives an example where the range of spatial auto-correlation is of comparable importance to an increase in prediction accuracy based on contextual terrain information (Figures 5 and 8). Thus, the spatial trend found in the topsoil silt dataset can be assumed to have resulted from differences in relief. Compared with ordinary kriging, which is based solely on spatial statistical assumptions (Odeh *et al*., 1994), ConMap is based on the common pedogenetic model of state factors (Jenny, 1941) and integrates space indirectly (n_i) in relief (r). Thus, r and n are not independent factors in the ConMap approach.

The relationship between the range of the variogram and the neighbourhood size needs to be further investigated on other datasets to see if it is possible to generalize rules. The study shows that even though there were some similarities in the spatial range, prediction accuracy can still increase in neighbourhoods above the range of the variogram using ConMap. Comparisons with other approaches that directly (n_d) or indirectly (n_i) integrate the spatial domain, such as regression kriging (Odeh *et al*., 1994), geographically weighted regression (Brunsdon *et al*., 1996) and those based on wavelet analysis, are needed.

Conclusions

This study introduces a new approach for terrain-based digital soil mapping named ConMap using contextual elevation differences across multiple neighbourhoods instead of common terrain attributes such as slope. It out-performs standard as well as another multi-scale digital terrain analysis approach and can model spatial trends indirectly. ConMap also returns slightly better results than ordinary kriging in this case study, even though there is a pronounced spatial trend in topsoil silt distribution.

The scorpan concept allows quantitative empirical modelling, and the present study shows an example where multiple scales should be considered as part of space. The discussion on the approach used to incorporate the spatial domain in machine learning-based digital soil mapping approaches is therefore most interesting. In this respect, further studies are required to reveal whether there is some general relationship between the range of the variogram and the neighbourhood size. Additional work is also necessary for interpreting the directional component in the importance of the elevation differences across different scales.

We conclude that the approach introduced here shows a great potential for future digital soil mapping studies, where spatial contextual information is important, as well as for analysing spatial soil variation in general. Further studies within other landscapes and on other mapping problems are needed.

Acknowledgements

We are grateful to the Federal Geological Surveys of Rhineland-Palatinate and Hesse for providing the data.

References

Begleiter, R. & El-Yaniv, R. 2008. *A Generic Tool for Performance Evaluation of Supervised Learning Algorithms*. Technical Report CS-2008-01. URL http://www.cs.technion.ac.il/∼ronbeg/gcv/DATAS/CS-2008-01.pdf [accessed on 3 March 2009]. Computer Science Department, Technion, Israel.

Behrens, T. 2003. *Digitale relief analyse als basis von boden-landschafts-modellen – am beispiel der modellierung periglaziärer lagen im ostharz.* Doctoral dissertation, Justus-Liebig-Universität, Gießen, Germany.

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

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

Behrens, T., Zhu, A.-X., Schmidt, K. & Scholten, T. 2009. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, DOI: 10.1016/j.geoderma.2009.07.010.

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

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

Breiman, L., Friedman, J., Olshen, R. & Stone, C. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Brunsdon, C., Fotheringham, A.S. & Charlton, M.E. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281–298.

Cutler, R.D., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. *et al*. 2007. Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Efron, B. & Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, Dordrecht.

Evans, I.S. 1980. An integrated system of terrain analysis and slope mapping. *Zeitschrift für Geomorphologie*, **36**, 274–295.

Gallant, J. & Hutchinson, M. 2008. Digital terrain analysis. In: *Guidelines for Surveying Soil and Land Resources* (eds N.J. McKenzie, M.J. Grundy, R. Webster, R. & A.J. Ringrose-Voase), pp. 75–79. CSIRO Publishing, Collingwood, Victoria.

Gallant, J., McKenzie, N. & McBartney, A. 2008. Scale. In: *Guidelines for Surveying Soil and Land Resources* (eds N.J. McKenzie, M.J. Grundy, R. Webster & A.J. Ringrose-Voase), pp. 27–43. CSIRO Publishing, Collingwood, Victoria.

Gerrard, A.J. 1981. *Soils and Landforms. An Integration of Geomorphology and Pedology*. Allen & Unwin, London.

Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

Grimm, R., Behrens, T., Märker, M. & Elsenbeer, A. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forests analysis. *Geoderma*, **146**, 102–113.

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

Ho, T.K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **20**, 832–844.

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

Jenny, H. 1941. *Factors of Soil Formation: A System of Quantitative Pedology*. McGraw-Hill, New York.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1143. Morgan Kaufmann, Montreal, Canada.

Lagacherie, P. 2008. Digital soil mapping: a state of the art. In: *Digital Soil Mapping with Limited Data* (eds A. Hartemink, A. McBratney & M.L. Mendoca-Santos), pp. 3–14. Springer, Dordrecht.

Lark, R.M. 2007. Decomposing digital soil information by spatial scale. In: *Digital Soil Mapping–An Introductory Perspective.* (eds P. Lagacherie, A. McBratney & M. Voltz), pp. 310–326. Developments in Soil Science Series, Elsevier, Amsterdam.

Liaw, A. & Wiener, M. 2009. *Random Forest: Breiman and Cutler's Random Forests for Classification and Regression*. URL http://cran.r-project.org/web/packages/randomForest/index.html [accessed on 3 March 2009].

Maclin, R. & Opitz, D. 1997. An empirical evaluation of bagging and boosting. *Proceedings of the 14th National Conference on Artificial Intelligence, Providence, Rhode Island*, pp. 546–551. AAAI Press, Cambridge, MA.

MacMillan, R. 2004. *Automated Knowledge-based Classification of Landforms, Soils and Ecological Spatial Entities*. URL ftp://ftp-fc.sc.egov.usda.gov/NSSC/NCSS/Conferences/regional/12-BobMacMillian.doc [accessed on 3 March 2009].

Matheron, G. 1960. *Krigeage d'un Panneau Rectangulaire par sa Périphérie*. Note géostatistique, 28, CG. Ecole des Mines de Paris, Paris.

McBratney, A.B., Mendonca-Santos, M.L. & Minasny, B. 2003. On digital soil mapping. *Geoderma*, **117**, 3–52.

Mendonça-Santos, M.L., McBratney, A.B. & Minasny, B. 2007. Soil prediction with spatially decomposed environmental factors. In: *Digital Soil Mapping – An Introductory Perspective.* (eds P. Lagacherie, A. McBratney & M. Voltz), pp. 277–286. Developments in Soil Science Series, Elsevier, Amsterdam.

Moran, C. & Bui, E. 2002. Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science*, **16**, 533–549.

Moyeed, R.A. & Papritz, A. 2002. An empirical comparison of kriging methods for non-linear spatial point prediction. *Mathematical Geology*, **34**, 365–386.

Odeh, I., McBratney, A. & Chittleborough, D. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, **63**, 197–214.

Pebesma, E.J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683–691.

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P. & Huybrechts, W. 2007. Random forests as a tool for predictive ecohydrological modelling. *Ecological Modelling*, **207**, 304–318.

Prasad, A.M., Iverson, L.R. & Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.

R Development Core Team 2009. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria*. URL http://www.R-project.org [accessed on 3 March 2009].

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.

Schönhals, E. 1996. *Ergebisse bodenkundlicher Untersuchungen in der Hessischen Lößprovinz mit Beiträgen zur Genese des Würm-Lösses*. Boden und Landschaft, 8. Justus-Liebig-Universität Gießen, Germany.

Shary, P.A., Sharaya, L.S. & Mitusov, A.V. 2002. Fundamental quantitative methods of land surface analysis. *Geoderma*, **107**, 1–35.

Siska, P.P., Goovaerts, P., Hung, I-K. & Bryant, V.M. 2005. Predicting ordinary kriging errors caused by surface roughness and dissectivity. *Earth Surface Processes & Landforms*, **30**, 601–612.

Smith, M.P., Zhu, A.-X., Burt, J.E. & Stiles, C. 2006. The effects of DEM resolution and neighbourhood size on digital soil survey. *Geoderma*, **137**, 58–69.

Strobl, C., Boulesteix, A.L., Zeileis, A. & Hothorn T. 2007. Bias in random forest variable importance measure: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Strobl, C. Boulesteix, A.L., Kneib, T., Augustin, T. & Zeileis, A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Wood, J. 1996. *The geomorphological characterization of digital elevation models*. Doctoral dissertation, University of Leicester, Leicester, UK.

Zevenbergen, L.W. & Thorne, C.R. 1987. Quantitative analysis of land surface topography. *Earth Surface Processes & Landforms*, **12**, 47–56.

Zhu, A.X. 2008. Spatial scale and neighbourhood size in spatial data processing for modeling nature environment. In: *Representing, Modeling and Visualizing the Natural Environment* (eds N.J. Mount, G.L. Harvey, P. Aplin & G. Priestnall), pp. 47–165. Innovations in GIS, CRC 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, A.X., Burt, J.E., Smith, M., Wang, R.X. & Gao, J. 2008. The impact of neighbourhood size on terrain derivatives and digital soil mapping. In: *Advances in Digital Terrain Analysis* (eds Q. Zhou, B. Lees & G. Tang), pp. 333–348. Springer, New York.