Predictive soil mapping as a means to aggregate and improve existing soil databases using classification trees and knowledge integration

Jan Willer[A], Rainer Baritz[B], Einar Eberhardt[B] and Reinhold Jahn[A]

[A]Martin Luther University of Halle-Wittenberg, Halle-Wittenberg, Germany, Email Jan.Willer@bgr.de
[B] Federal Institute for Geosciences and Natural Resources, Hannover, Germany, Email Rainer.Baritz@bgr.de

Abstract

A methodology using digital soil mapping has been developed to improve the reconnaissance mapping 1:200.000. It combines soil data from different sources in order to develop a seamless, nation-wide, consistent soil geometric and semantic database. The use of legacy data involves some challenges coming from differing datasets, soil descriptions, mapping strategies and data gaps which are typical for mapping campaigns which last over few decades. These problems have to be solved by applying and developing extensive semantic harmonization and quality control procedures.

Digital soil mapping techniques are considered to be a powerful tool to harmonise data from different data sources, filling gaps in existing soil maps and cross-validation. The model and methodical pathway presented here has been developed as a hybrid approach, combining (a) classification tree analysis of existing soil maps and auxiliary data (elevation models, geological maps, climatic data, etc.) as regionalisation method for discrete soil classes, and (b) the integration of knowledge about regional soil forming processes through expert-based rules. As the final product, predictive conceptual soil maps are generated. The methodology can be used to harmonize soil data from different sources and to speed up the mapping process in areas with a lack of soil mapping data.

Key Words

Digital soil mapping, classification tree, Soil data harmonizing, Project SIAM.

Introduction

The project SIAM (**S**oil **I**nference **a**nd **M**apping Project) aims at putting together a set of digital soil mapping techniques that allow an integration of soil and auxiliary data from different sources and scale to form a consistent level of soil information at a fixed target scale (here: 1:200.000-1:250.000, Germany). The resulting methodology is then applied to gap filling, quality checking and harmonizing existing reconnaissance mapping sheets throughout the study area (Germany, stratified into soilscapes). Even though the national soil project builds on a harmonized assessment schemes, different data sources were used in different parts of the country. In some areas, digital high-resolution mapping data are only quite patchy and not area covering. These differences and data gaps print through into the individual sheets 1:200.000 in different ways. Since qualitative experiences and legacy data sources are to be utilized as much as possible, a hybrid approach was selected. If successfully calibrated, the system is also able to predict conceptual maps in areas not covered with such legacy data.

Methods

The approach presented here combines the analysis of existing soil maps with classification trees and predictive modelling with knowledge integration. Classification trees are an established method in digital soil mapping approaches (Behrens and Scholten 2007), and in comparison with other data mining techniques, the derived models permit a better interpretability by soil scientists. As a pilot area, the map sheet Cologne with a total area of 6400 ^{2} was used (Figure 1). Inside this test sheet, a training area for calibration was selected based on the quality of 1:50,000 mapping and data density (Figure 2), which is representative for most parts of the Rhenish Slate Mountains. Classification trees were used to extract a concept model from the pilot area soil map. The rules of the classification tree were put into the software SolimSolution (Zuh *et al.* 2004) for recording and modifying the model by knowledge integration and producing predictive soil maps (Figure 3).

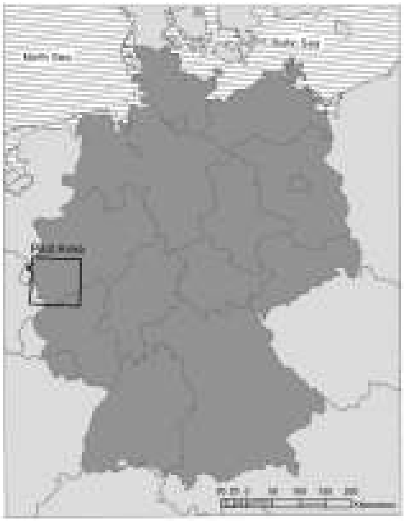

Figure 1. Location of the Pilot Area

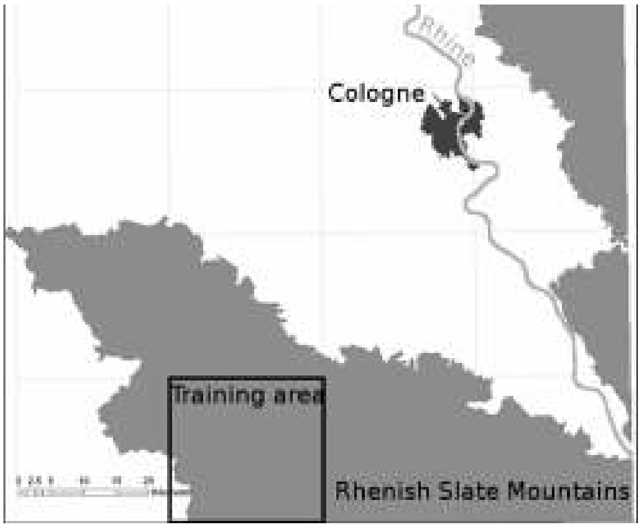

Figure2: Location of the training area within the Pilot area

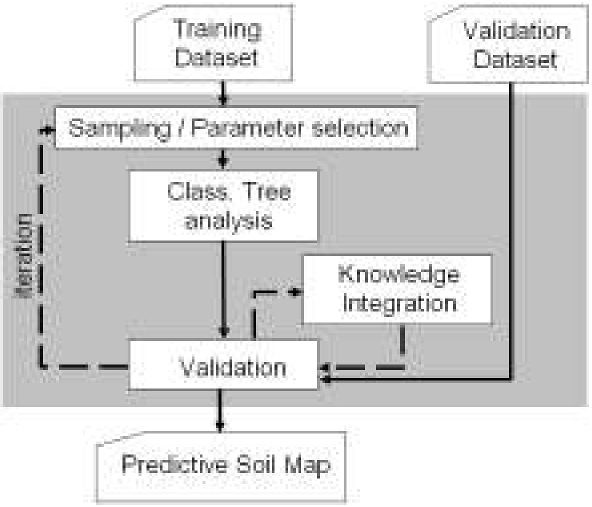

Figure 3. Overview of method

Data selection and preparation

Selection of the data is a crucial step within the model development. Accuracy of the results depends on the accuracy of the input data, but as well on the parameter selection appropriate for the target scale. At a scale of 1:200k, environmental factors relevant to describe the general spatial patterns of soil associations differ from larger scales. As we use data from various scales and sources, reduction of noise originating from previous generalisation steps is essential. Soil information was derived from the soil map 1:50,000 of North Rhine-Westphalia. The legend units of this soil map describe discrete associations of soil units that combine genetic soil types with parent material classes. For the target scale, these map units need to be aggregated and spatially generalised to form even more complex units.

Classification Tree Analysis

For the Classification Tree Analysis, we used the Software GUIDE (Loh 2008, 2009). It has some advantages over the better known CART algorithm (Breimann *et al.* 1983), as it allows unbiased split selection at each node and also interaction detection within the nodes. For the split selection we used the simple linear model. More complex models using the kernel or nearest-neighbour method did not significantly improve the results while making the tree more complex and less interpretable.

Data selection included four main steps:

(1)	Soil scapes: delineation of mountainous regions with similar geological environment	Cluster analysis for grid datasets (SAGA Software package) has combined the following data: – standard deviation of altitude (250 m radius) – relative altitude (50 km radius) – geological units (geolog. map 1:100k)
(2)	Apriori definition of target map units	For orientation, map units of the soil map 1:50,000 were aggregated to a target scale of 1:200k by the geological service. The aggregation was based on soil morphogenetic types, depth to bedrock, texture, base saturation and parent material. The 34 Units of the map 1:50,000in the test area were aggregated into 14 units suitable for the 1:200k map.
(3)	Development of ancillary data base	– Terrain analysis using a local DEM 25 m (SAGA GIS, Köthe 2006) – Geological map 1:100k Climate data showed no significant correlation with the soil distribution, but tended to produce artificial over-fitting in the classification tree analysis. The same held true for land use data derived from satellite images.
(4)	Noise reduction	a: DEM filtering: – Gaussian filter to achieve a moderate smoothing of the DGM – multidirectional filter (Selige *et al.* 2006) b: Buffer functions to settle misfits between the soil map and geologic map polygons with their respective topographies and degree of generalisation c: Identification of outlier soil polygons following the approach of (Qi 2006).

Knowledge Integration

The rules obtained during the classification tree analysis transferred to the inference modelling software SolimSolution (Zhu *et al.* 2001) in order to explicitly record the classification rules. In a subsequent step, the model was further modified, based on regional expert knowledge or by means of other statistical approaches, because fuzzy membership functions can deliberately be changed by the user. For each pixel of the predictive map, the rules applied and the decisive parameter for its assignment to a soil class can be extracted. Classification misfits can be traced back to the deciding rule, even if highly complex classification rules are used. In this way, it is possible to integrate expert knowledge or adapt it with complementary datasets. In the current state of the model, we use a knowledge-based rule for one soil association (dominated by Haplic Luvisols, see Figure 4) that is underrepresented in the training area.

Validation

As a validation, we compared the model results with existing soil maps in the training area and in a separate validation area (neighbouring sheet 1:50,000) with similar environmental conditions (Figure 4). User's, producer's and overall accuracy were calculated following Congalton and Green (1999). Validation with auxiliary data is planned but not concluded yet.

Results and Conclusions

We were able to predict all 14 apriori map units (Figure 4) that were defined in the training area with an overall accuracy of 0.53 in the training area and 0.51 in the validation area, as compared to the existing soil map 1:50,000. The model developed in the training area shows visually plausible results for a large area of the Rhineland Slate Mountains with similar geological setting. Hence, our approach provided an independent tool for comparing soil maps from different origin and scale. Due to the data selection and the parameter settings, a reasonable generalization for our target scale of 1:200k of the soil map has been achieved. For certain areas, a lack of input data, e.g. incomplete data on the parent material or historical land use, produced hardly assessable inaccuracies. As a conclusion, the comparative low congruence with the existing soil map is not only a result of model inaccuracy, but can also be attributed to the required level of generalization.

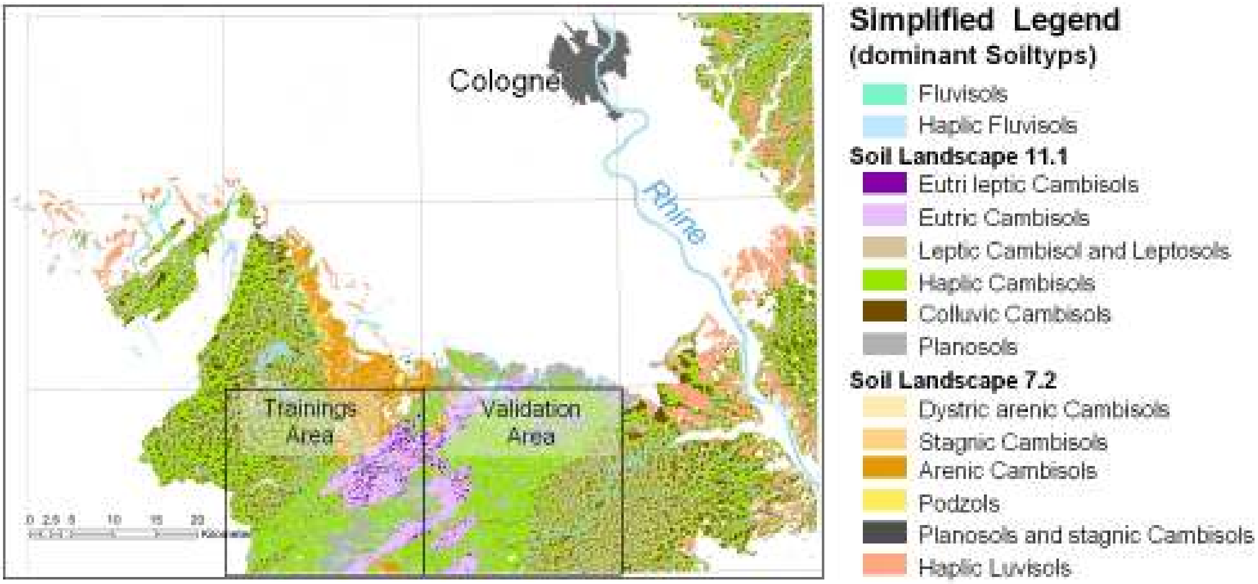

Figure 4. Predictive soil map for the Rhenish Slate Mountains

Outlook

For further improvement of the model, more accurate parent material information shall be obtained by a combined approach of relief analysis, parent material information from the soil map and spatial distance to the respective source rock units of the geological map. Knowledge integration should be improved, in a way that allows uncertainty assessment, as it is considered to be an important advantage over models using only geostatistical methods. Quantifying the effect of the generalisation on the model accuracy shall be done by a systematic comparison with unpruned trees highly adapted to the training area.

References

Behrens T, Scholten, T (2007) A comparison of data-mining techniques in predictive soil mapping. In 'Digital soil mapping. An introductory perspective'. (Eds P Lagacherie, AB McBratney, M Voltz) pp. 353-364. vol 31. (Elsevier).

Breiman L, Friedman JH, Olshen RA, Stone CJ (1983) CART: Classification and Regression Trees. Wadsworth: Belmont, CA.

Congalton RG, Green K (1999) Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, pp. 45 – 47. (Boca Raton).

Köthe R, Bock M (2006) Development and use in practice of SAGA modules for high quality analysis of geodata. *Göttinger Geographische Abhandlungen*, **115**, 85-96.

Loh W (2008) Classification and Regression Tree Methods. In 'Encyclopedia of Statistics in Quality and Reliability' (Eds Ruggeri, Kenett, Faltin) pp. 315-323. (Wiley).

Loh, W (2009) Improving the precision of classification trees. *Annals of Applied Statistics* **3**, 13 pp.

Lee J (1980) Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**, 165-168.

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.

Selige T, Böhner J, Bock M (2006) Processing of SRTM X-SAR Data to correct interferometric elevation models for land surface process applications. *Göttinger Geographische Abhandlungen*, **115**, 97-104.

QI F (2004) Knowledge Discovery from Area-Class Resource Maps: Data Preprocessing for Noise Reduction. *Transactions in GIS* **8**, 297-308.

Zhu AX, Hudson B, Burt V, Lubich K, Simonson D (2001) Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. *Soil sci. Soc. Am. J.* **65**, 1463-1472.

Zhu A, Moore AC, Smith MP, Liu J, Burt J, Qi F, Simonson D, Hempel J, Lubich K (2004) Advances In Information Technology For Soil Surveys: The Solim Effort. In 'Innovative techniques in soil survey: “Developing the foundation for a new generation of soil resource inventories and their utilization” (Eds Eswaran H, Vijarnsorn P, Vearasilp T, Padmanabhan E) pp. 25-42.