Tagungsbeitrag zu:
Jahrestagung der DBG, Kom. V

Titel der Tagung:
Böden - eine endliche Ressource

Veranstalter:
DBG, September 2009, Bonn

Berichte der DBG
(nicht begutachtete online Publikation)
http://www.dbges.com

Projekt SIAM - Entwicklung eines Boden-Landschaftsmodells zur Datenharmonisierung und Qualitätssicherung für Bodenübersichtskarten

J. Willer[*], *R. Baritz*[+], *E. Eberhardt*[+], *G. Milbert*[x], *R. Jahn*[*]

Schlüsselwörter: *Digitale Bodenkartierung, Classification Tree, Datenharmonisierung*

Zusammenfassung

Für die Bodenübersichtskarte von Deutschland 1 : 200.000 (BÜK 200) werden Bodendaten aus den Bundesländern, die über mehrere Jahrzehnte erhoben wurden, zu einer blattschnittfreien, einheitlichen Übersichtskarte und Bodendatenbank zusammengeführt. Dafür muss historisch bedingt auf eine sehr heterogene Datenbasis zurückgegriffen werden. Das macht umfangreiche Qualitätskontrollen und Harmonierungsschritte erforderlich, wobei für jedes Blatt aufgrund der unterschiedlichen Datengrundlagen neue Anforderungen auftreten. Im von der BGR initiierten Projekt SIAM (**S**oil **I**nference **a**nd **M**apping Project) werden Methoden aus dem Bereich der digitalen Bodenkartierung entwickelt und getestet, die als Werkzeuge für die Qualitätssicherung und Datenharmonisierung eingesetzt werden können. Das hier entwickelte Verfahren zum Aufbau eines Boden-Landschaftsmodells ermöglicht es, an Referenzblättern kalibrierte Modelle zur Bodenformenvergesellschaftung abzuleiten, um Bodenkarten unterschiedlicher Maßstäbe und Qualität nach objektiveren Kriterien zu vereinheitlichen. Die Prognosekarten zeichnen generelle Verteilungsmuster nach, solange sich Beziehungen zwischen kartierten Bodeneinheiten und dem Relief bzw. Ausgangsgestein finden lassen. Generalisierungsnotwendigkeiten lassen Übereinstimmungen der Prognose mit größermaßstäbigen Vergleichskarten von ca. 50 % angemessen erscheinen.

[*] Martin-Luther-Universität Halle-Wittenberg
Weidenplan 14, 06108 Halle (Saale)
[+]Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), Stilleweg 2, 30655 Hannover
[x] Geologischer Dienst Nordrhein-Westfalen
De-Greiff-Straße 195, 47803 Krefeld
Korrespondenz: jan.willer@bgr.de

Pilotgebiet Blatt Köln

In Kooperation mit dem Geologischen Dienst NRW wird das Boden-Landschaftsmodell auf dem BÜK 200-Blatt Köln (Abb. 1) entwickelt. An dem etwa 7650 km^2 großen Gebiet haben das Rheinische Schiefergebirge und die Niederrheinische Bucht große Anteile. Zurzeit wird das Modell für das Bergland entwickelt.

Abb. 1: Lage des Pilotgebiet Blatt Köln

Methode

Die hier vorgestellte Methode kombiniert eine Classification Tree-Analyse als klassisches Datamining-Verfahren (BEHRENS & SCHOLTEN 2007) mit einem wissensbasierten Ansatz für die Prognose von Bodenvergesellschaftungen

im Übersichtsmaßstab. Die Classification Tree-Analyse wird verwendet, um Beziehungen zwischen Verbreitungsmustern von Bodengesellschaften und davon unabhängigen Umweltvariablen (hier Reliefparameter und Geologie) zu analysieren. Grundlage sind u. a. existierende Bodenkarten. Dabei kann das in älteren Kartenwerken meist nur implizit enthaltene Kartierwissen in Form von Regelsätzen extrahiert werden, die in ähnlichen Landschaften zur Bodenprognose genutzt werden können. Durch die Weiterverarbeitung dieser Regeln in der Software SolimSolutions (ZHU et. al. 2001), ist es möglich, neuere Erkenntnisse zu integrieren und die zuvor erstellten Regeln fortzuschreiben. Zur Entwicklung des Modells wurde ein Lerngebiet innerhalb des Pilotgebiets gewählt, das für große Teile des Rheinischen Schiefergebirges repräsentativ ist (Abb. 2). Zu Validierung wurde ein unabhängiger Datensatz aus einem angrenzenden Gebiet gewählt.

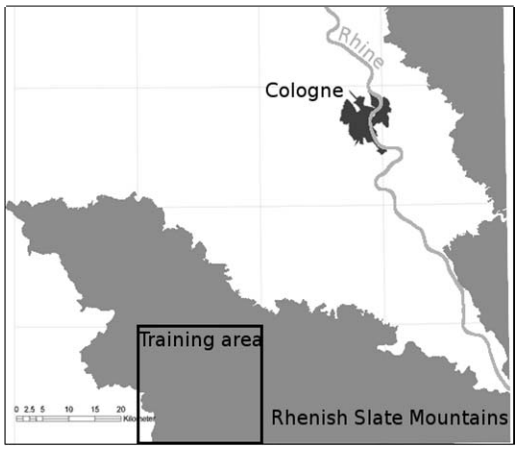

Abb.2: Lage des Lerngebiet innerhalb des Pilotgebiet Blatt Köln

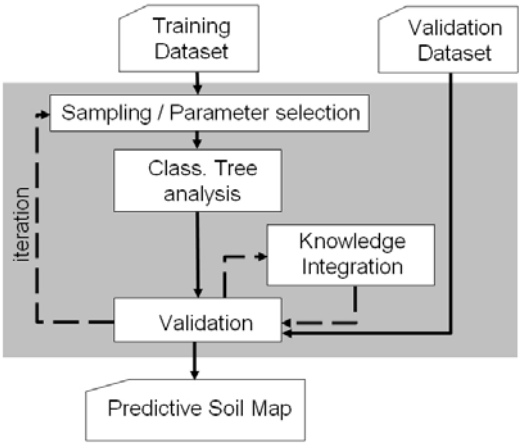

Abb. 3: Methodenüberblick

Abb. 3. gibt einen schematischen Überblick über die Methode.

Datenauswahl und Vorbehandlung

Da die Eingangsdaten in unterschiedlichen Maßstäben und mit unterschiedlicher Genauigkeit vorlagen, waren einige Schritte zur Minimierung des dadurch verursachten „Rauschens“ in den Daten nötig. Ausgewählt wurden nur Daten, die deutschlandweit in vergleichbarer Qualität zur Verfügung stehen. In der BÜK 200 werden nicht einzelne Bodenformen dargestellt, sondern komplexe Bodenformenassoziationen. Die Bodeninformationen aus dem Lerndatensatz mussten daher entsprechenden Bodenformenassoziationen zugeordnet werden. Für die statistische Analyse war es nötig, die Datensätze nach unterschiedlichen Landschaftsräumen zu stratifizieren. Hierfür wurden geomorphographische Kriterien herangezogen. Die Datenvorbereitung umfasste die folgen vier Schritte:

(1) *Stratifizierung der Datensätze für die statistische Analyse*
Clusteranalyse für Rasterdaten (SAGA GIS, KÖTHE 2006)) mit den folgenden Daten:
- Standardabweichung der Höhe (250 m Radius)
- Relative Höhe (50 km Radius)
- Aggregierte Einheiten der Geologischen Karte 1 : 200K

Unterteilung in drei Unterdatensätze, die getrennt behandelt werden

(2) *A priori-Definition der Zielklassen (Bodenformenassoziationen)*
Die Einheiten der Bodenkarte 1 : 50.000 wurden durch den Geologischen Dienst NRW für 1 : 200.000 aggregiert. Die Aggregierung basierte auf dem Bodentyp, der Tiefe zum Festgestein, Bodenart, Basensättigung und Ausgangsgestein. Die 34 Kartiereinheiten der Bodenkarte 1 : 50.000 im Lerngebiet wurden dabei zu 14 Kartiereinheiten zusammengefasst.

(3) *Aufbau der Datenbasis*

- DGM mit 25 m Rasterweite, Ableitung verschiedener Reliefparameter in SAGA GIS. Im Modell wurden die Parameter Saga-Wetness-Index, Höhe über Tiefenlinie, Oberflächenwölbung (horizontal, längs und quer), Länge des Oberflächenabflusses, Hangneigung und -richtung verwendet
- Geologische Karte 1 : 100.000

Nicht verwendet:

- Klimadaten
- Corine-Landnutzungsdaten

(beide Datensätze zeigten keine signifikante Korrelation mit der Bodenverbreitung im Testgebiet und bewirkten in der Classification Tree-Analyse eine Überanpassung (Overfitting), an das Lerngebiet

(4) *Fehlerbehandlung in den Daten*

a: DGM-Filterung:

- Gauss-Filter um eine schwache Glättung des DGM zu erreichen
- Multidirectional Lee-Filter (SELIGE et. al. 2006), stärkere, aber hangparallele Glättung.

b: Entfernung von Pufferzonen an den Grenzen der Bodenpolygone, um den Effekt verschiedener Topographien in den Datensätzen abzumildern. Entfernung von widersprüchlichen Kombinationen aus dem Lerndatensatz nach inhaltlichen Kriterien.

c: Identifikation von Ausreißer-Polygonen nach QI (2004).

Classification Tree-Analyse

Die Classification Tree-Analyse ist eine etablierte Methode in der digitalen Bodenkartierung (BEHRENS & SCHOLTEN 2007). Im Gegensatz zu ähnlichen Verfahren wie neuronalen Netzen oder random forests sind die Ergebnisse besser bodenkundlich interpretierbar. Für die Analyse wurde die Software GUIDE (LOH 2008, 2009) verwendet. Die Optimierung der Pruning-Einstellungen, die zur Vermeidung einer Überanpassung an das Lerngebiet wichtig ist und zu einer Generalisierung der Resultate beiträgt, erfolgte in Anlehnung an SCHMITT et. al. (2008) in einem iterativen Verfahren.

Integration von Wissen

Das Ergebnis der Classification Tree-Analyse wird in die Software SolimSolution importiert, wobei die Klassifizierungs-Regeln in Zugehörigkeitsfunktionen überführt werden. Die Regeln können danach sofort in ein Prognosemodell umgesetzt werden. Dabei ist es möglich, für jedes Pixel der Prognosekarte die zugrundeliegende Regel abzufragen. Zusätzlich ist eine schrittweise Optimierung des Modells möglich, indem Bereiche mit unplausiblem Prognoseergebnis identifiziert und die zugrundeliegenden Regeln unter Berücksichtigung von Expertenwissen und/oder zusätzlicher statistischer Auswertungen angepasst werden können.
Dieser Schritt ermöglicht es, das Modell so weiter zu entwickeln, dass es über den Wissensstand der Trainingskarte hinausgeht. Für die laufende Modellentwicklung für das Blatt Köln wurde zum Beispiel für die Verbreitung der Parabraunerden, die im Lerndatensatz unterrepräsentiert sind, eine wissensbasierte Regel ergänzt.

Validierung

Zur Einschätzung der Prognosequalität wurde das Modell mit der existierenden Bodenkarte 1 : 50.000 im Lern- und in einem Validierungsgebiet (Abb. 4) verglichen. Berechnet wurden die User's, producer's und overall accuracy nach CONGALTON und GREEN (1999). Weitere Validierungsschritte mit zusätzlichen Daten und unter Berücksichtigung des Zielmaßstabes sind geplant.

Ergebnisse und Diskussion

Das Prognosemodell (Abb. 4) konnte alle vierzehn im Lerngebiet vorkommenden Kartiereinheiten (Abb. 4) mit einer durchschnittlichen Übereinstimmung (overall accuracy) von 0,53 im Lerngebiet und 0,51 im Validierungsgebiet in ihrer Verbreitung nachvollziehen, bezogen auf die Bodenkarte 1 : 50.000. Das im Lerngebiet entwickelte Modell zeigt visuell plausible Er

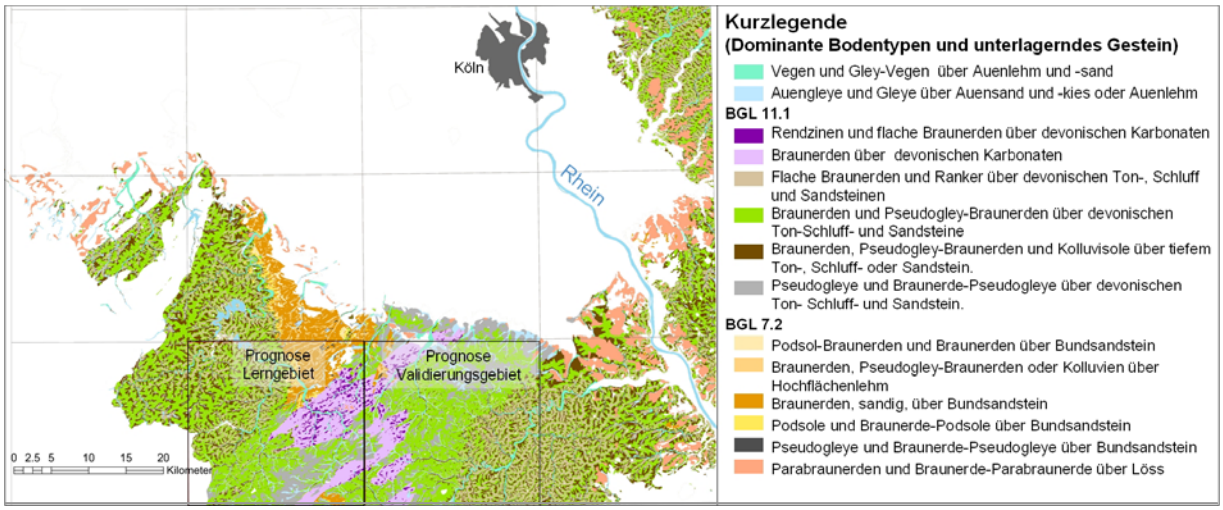

Abb. 4: Prognoseergebnis für das Bergland und Lage des Validierungsgebiet

gebnisse für die Anteile des Rheinischen Schiefergebirges am Arbeitsgebiet mit ähnlichen geologischen Rahmenbedingungen. Damit bietet das Modell ein unabhängiges Werkzeug, um Bodenkarten verschiedener Qualität und Herkunft zu vergleichen. Durch die Datenauswahl und die Parametereinstellungen konnte bereits eine erste Generalisierung für den Maßstab 1 : 200.000 erreicht werden. In einigen Gebieten verursachen fehlenden Eingangsdaten, z. B. unvollständige Informationen zum Ausgangsgestein oder zur historischen Landnutzung, Abweichungen gegenüber der Bodenkarte. Eine Quantifizierung dieser Fehler ist schwer möglich. Zusätzlich ist die vergleichsweise geringe Übereinstimmung mit der Bodenkarte auch auf die beabsichtigte Generalisierung für den Zielmaßstab zurückzuführen.

Ausblick

Für eine Erhöhung der Modellgenauigkeit sollen die Informationen zum Ausgangsgestein aus der abgedeckten geologischen Karte verbessert werden. Hierzu soll die räumliche Entfernung zu den geologischen Einheiten in Abhängigkeit von der Reliefposition berücksichtigt werden. Die Integration von Wissen soll hinsichtlich einer besseren Abschätzung von Unsicherheiten verbessert werden. Der Effekt der durchaus erwünschten Generalisierung durch die Einstellungen bei der Classification Tree-Analyse auf die Modellgenauigkeit soll durch einen Vergleich mit nicht vereinfachten Bäumen und bestmöglicher Anpassung an das Lerngebiet abgeschätzt werden.

Literatur

BEHRENS, T., SCHOLTEN, T. (2007): A comparison of data-mining techniques in predictive soil mapping. In LAGACHERIE, P., MCBRATNEY, A. B., VOLTZ, M. (Eds): Digital soil mapping. An introductory perspective, 353-364. (Elsevier).

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 RUGGERI, KENETT, FALTIN (Eds): Encyclopedia of statistics in quality and reliability, pp. 315-323. (Wiley).

LOH, W (2009): Improving the precision of classification trees. Annals of Applied Statistics, 3.

QI, F. & ZHU, A.X. (2003): Knowledge discovery from soil maps using inductive learning. Int. J. Geographical Information Science, 17, 771-795.

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

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.

ZHU, A.X., 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.