Modellierung der naturräumlichen Einheiten Baden-Württembergs mit Boosted Regression Trees

Geraldine Quénéhervé[1], Markus Alle[2], Andreas Schwab[2], Hans-Joachim Rosner[1]

[1]Fachbereich Geographie, Universität Tübingen · geraldine.queneherve@uni-tuebingen.de
[2]Arbeitsgruppe Geographie, Pädagogische Hochschule Weingarten

Zusammenfassung: In Landschaftsplanung und -ökologie spielen Naturräumliche Ordnungen als Planungsgrundlage eine wichtige Rolle. Dabei kommen in Baden-Württemberg nach wie vor visuell erfasste und händisch klassifizierte Gliederungen zum Einsatz. Diese Arbeit verfolgt einen statistischen Ansatz, um mittels Regressionsanalysen ein objektiveres Klassifikationsverfahren zu erproben. In die Modellierung fließen die Vegetation, Geologie, Landbedeckung, Höhendaten sowie deren Derivate als erklärende Variablen ein. Die naturräumliche Gliederung von MEYNEN & SCHMITHÜSEN (1953-1962) dient als Zielvariable. Von 17 Eingangsvariablen erweisen sich sechs als signifikant. Vor allem die Geologie, Vegetation und Höhendaten besitzen entscheidenden Einfluss auf die automatisierte Ableitung der naturräumlichen Gliederung.

Schlüsselwörter: Geostatistik, Modellierung, Naturraum, Boosted Regression Trees

Abstract: *Within landscape planning and ecology, landscape structures play an important role in decision making. In Baden-Württemberg, the visually interpreted and hand drawn classification is still in use. This study uses the geostatistical approach of Boosted Regression Trees to test the classification in a more objective way. The model uses environmental variables such as potential natural vegetation, geology, land cover, digital elevation models and its derivatives. The target variable is the landscape structure of MEYNEN & SCHMITHÜSEN (1953-1962). Of 17 environmental variables, six are proved to be significant. Especially geology, vegetation and the topography play an important role in automatically drawing landscape structures.*

Keywords: *Geostatistical analysis, modelling, landscape structure, Boosted Regression Trees*

1 Motivation und Stand der Forschung

Für die Landschaftsplanung und -ökologie sowie für den Naturschutz stellen Raumgliederungen eine wichtige Planungsgrundlage dar, da sie die landschaftliche Komplexität reduzieren und Ordnung in die natürliche Mannigfaltigkeit bringen (MANNSFELD 2005). In Baden-Württemberg kommt hauptsächlich das deutschlandweite landschaftliche Gliederungssystem von MEYNEN & SCHMITHÜSEN (1953-1962) zum Einsatz (z. B. LANDSCHAFTSPARK OBERSCHWABEN-BODENSEE 2003). Hierbei handelt es sich um einen sogenannten Top-down-Ansatz, der visuell erfasste landschaftliche Großstrukturen auf Basis von topographischen Karten schrittweise in immer kleinere Teilräume gliedert (vgl. DONGUS 1991). Die Abgrenzung erfolgte analog und nicht automatisiert auf Basis von Expertenwissen. Mittels einem hierarchischen Verfahren wurden mehrere Ordnungsstufen ausgewiesen, deren dritte Stufe sich auf der mesoskaligen Ebene befindet und Grundlage der hier vorliegenden Untersuchung ist. Dabei ist Baden-Württemberg als Teil der deutschlandweiten naturräumlichen Gliederung in 13 Naturräume unterteilt (Abb. 1a).

Die Subjektivität und die regionalen wie zeitlichen Differenzen dieses Verfahrens werden kritisch betrachtet (vgl. z. B. BURAK 2005). Die jüngere Landschaftsökologie verfolgt einen

AGIT – Journal für Angewandte Geoinformatik, 2-2016, S. 204-212. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-622-2, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537622030.

gegensätzlichen Ansatz, welcher großmaßstäbig erhobene Raumeinheiten zu höherrangigen aggregiert (vgl. SYRBE 1999). Bei einem solchen Verfahren spricht man nach RICHTER (1967) von einer naturräumlichen Ordnung. Auch aktuelle Arbeiten, wie die Habitat- und Ökosystemklassifikation von KUTTNER et al. (2015), arbeiten nach diesem Prinzip.

In dieser Studie werden für die Ausweisung der naturräumlichen Einheiten Regressionsbäume (*Boosted Regression Trees*) verwendet. Dieses Verfahren kann kategoriale (d. h. nominale und ordinale) Daten zusammen mit metrischen (intervall- und rationalskalierte) Daten ohne Skalentransformation verarbeiten (SCHRÖDER & SCHMIDT 2000). Es arbeitet mit einer klassifizierten Zielvariable [hier die naturräumliche Gliederung nach MEYNEN & SCHMITHÜSEN (1953-1962)], um eine homogene Raumgliederung zu erzeugen. Ziel des Verfahrens ist es, Objekte anhand der Ähnlichkeit ihrer Merkmalsausprägung in möglichst wenige homogene und klar unterscheidbare Klassen zu gliedern. Erste Berechnungen mittels eines geostatistischen Verfahrens führten SCHRÖDER & SCHMIDT (2000) auf deutschlandweiter Ebene in 2 × 2 km Auflösung durch (Abb. 1b).

Ziel dieser Arbeit ist, die Naturräume Baden-Württembergs auf Basis der Naturräume von MEYNEN & SCHMITHÜSEN als Zielklassen in einem überwachten, statistischen Ansatz auf einer kleinräumigen Skala zu modellieren.

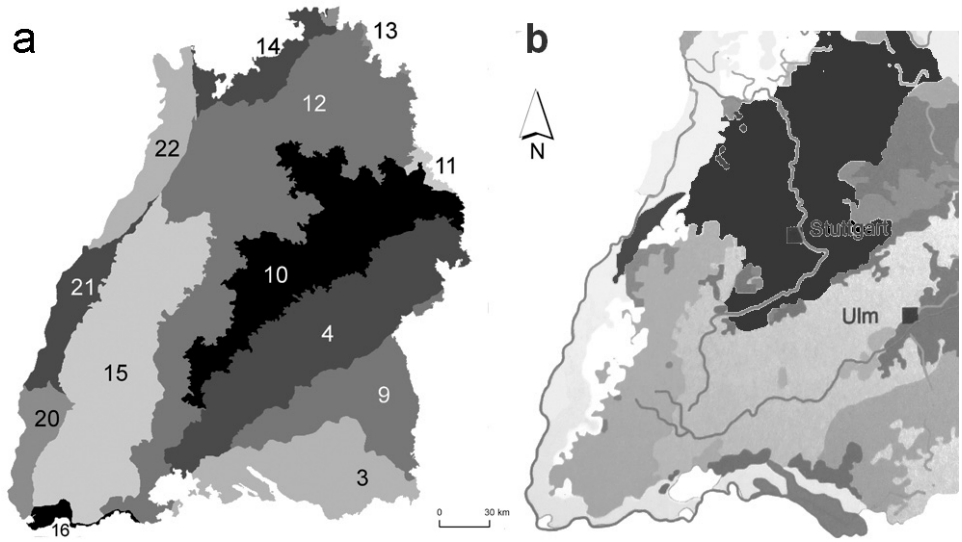

Abb. 1: Darstellung der ausgewiesenen Naturräume:
a) Einteilung Baden-Württembergs in die 3. Ordnung (13 Naturräume) nach MEYNEN & SCHMITHÜSEN (1953-62); b) Standortökologische Raumgliederung (Ausschnitt Baden-Württemberg) nach SCHRÖDER & SCHMIDT (2000)

2 Methodik

2.1 Modellvariablen

In das statistische Modell der *Boosted Regression Trees* fließen für Baden-Württemberg vier thematische Datensätze mit Raumbezug ein. Die Zielvariable ist die naturräumliche Gliederung nach MEYNEN & SCHMITHÜSEN (in LUBW 2015). Folgende Variablen werden als erklärende Variablen verwendet: (1) Die potenzielle natürliche Vegetation (LUBW 2015) also diejenige Vegetation, die unter den gegenwärtigen klimatischen, orographischen und pedologischen Randbedingungen unter Ausschluss menschlicher Einflüsse zu erwarten wäre (TÜXEN 1956). (2) Die Geologische Karte der Bundesrepublik Deutschland 1:1.000.000 (GK1000) (BGR 2015) wird mit den quartären Einheiten des Alpenvorlands genetisch, die älteren Sedimentgesteine nach der Stratigraphie beschrieben. (3) Die Landnutzungsklassifikation erfolgt auf der Basis des CORINE Datensatzes (UBA 2004).

Zusätzliche Bestandteile der statistischen Analyse ist das frei verfügbare SRTM90 Höhenmodell (DLR 2012) und dessen Derivate. Das Höhenmodell wurde mit einem Tiefpass-Filter vorprozessiert um Artefakte zu minimieren. Dieser Filter berechnet den Mittelwert für ein 3x3 Kernel, sodass die Extremwerte in diesem gemittelt werden (LEE 1980). So werden die lokalen Variationen geglättet und das Rauschen minimiert. Anschließend wurde es hydrologisch unter Verwendung des Algorithmus von PLANCHON & DARBOUX (2001) korrigiert. Die abgeleiteten Datensätze reflektieren die Topographie, Erosions- und Depositionsprozesse sowie allgemeine Unterschiede in den Landschaftsformen. Alle Erklärungsvariablen sind in Tabelle 1 aufgeführt.

Tabelle 1: Erklärungsvariablen für die *Boosted Regression Tree* Analyse

Name	**Literaturangabe**
Potenziell natürliche Vegetation (PNV)	LUBW 2015
Geologie	BGR 2015
Landnutzung (CORINE)	UBA 2004
Höhenmodell (nachfolgend Ableitungen daraus)	DLR 2012
Hangneigung	TRAVIS et al. 1975
Exposition	TRAVIS et al. 1975
Longitudinal Curvature	ZEVENBERGEN & THORNE 1987
Cross-Sectional Curvature	ZEVENBERGEN & THORNE 1987
Convergence Index	CONRAD 2002
LS Faktor	MOORE et al. 1991
Relative Slope Position	CONRAD 2002
Valley Depth	BOEHNER & CONRAD 2008
Vertical Distance to Channel Network	CONRAD 2002
Channel Network Base Level	CONRAD 2002
Topographic Wetness Index	BEVEN & KIRKBY 1979
Catchment Area	FREEMAN 1991
Terrain Ruggedness Index	RILEY et al. 1999

Zur Optimierung der Rechenabläufe wurden alle Erklärungsvariablen auf eine Auflösung von 250 m × 250 m transformiert. Die Gesamtfläche des betrachteten Raums beträgt 571.705 Pixel bzw. 35.732 km^2. Die so erhaltene Tabelle mit den Einzelwerten der Erklärungsvariablen geht in die weitere Berechnung ein. Die Zielvariable wurde dieser Tabelle hinzugefügt.

2.2 Data Mining

Die Analyse wurde mit einem *Boosted Regression Tree* (BRT) Ansatz (ELITH et al. 2008) in Salford Systems (TreeNet™) durchgeführt. Regressionsanalysen werden verwendet, um von einem kleineren Modelldatensatz auf den gesamten betrachteten Raum zu schließen. BRTs kombinieren klassische Regressionsbäume (vgl. Abb. 2) mit einem *Boosting* Ansatz, welcher verschiedene Klassen optimal trennt, sowie einem *Bagging* Algorithmus (*bootstrap aggregating*), welcher die Modelgüte verbessert. Ausgangspunkt ist ein Datensatz, welcher die Zielklasse, in der vorliegenden Studie die naturräumliche Gliederung nach MEYNEN & SCHMITHÜSEN, abbildet und dann an jeden Pixel die Werte der Erklärungsvariablen anhängt (vgl. Tabelle 1).

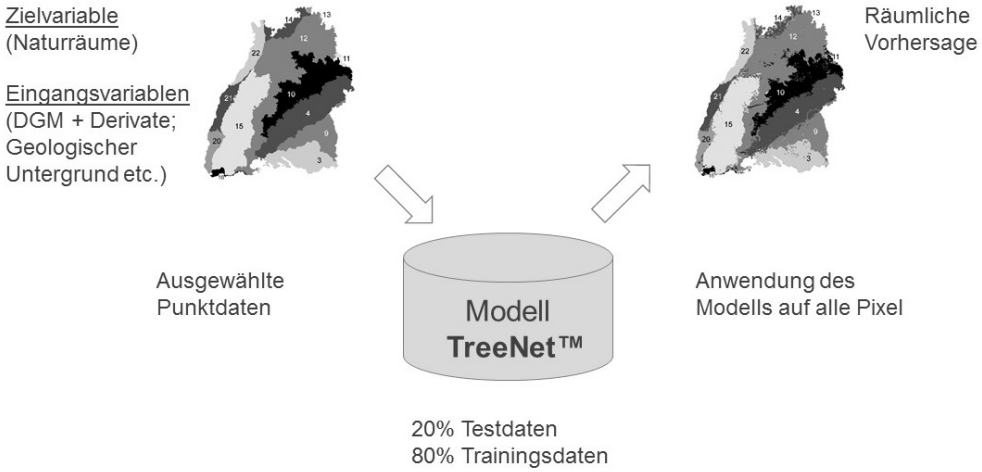

Abb. 2: Aufbau des räumlichen Data-Mining-Ansatzes

Im Entscheidungsbaum werden anschließend die einzelnen Klassen der Zielvariable mit der besten Kombination von Erklärungsvariablen beschrieben. Somit können auch unbekannte Klassen der Zielvariablen durch die im Modell vorgegebene Kombination von Erklärungsvariablen erschlossen werden (vgl. Abb. 3). Die Analyse läuft daher als Klassifikationsmodell.

Die *Boosting* Funktion erzeugt jeden Baum auf Grundlage des originalen Datensatzes, aber jeder neue Baum basiert auf Informationen aus den zuvor erstellten Bäumen. *Bagging* bildet wiederholt Entscheidungsbäume aus dem Trainingsdatensatz um das Vorhersagemodell zu bilden und mittelt dann diese Ergebnisse. Damit wird die Varianz reduziert und die Vorhersagegüte erhöht.

Im Modell gehen die einzelnen Erklärungsvariablen unterschiedlich stark in die zu erklärende Klasse ein (*variable importance*). Diese Gewichtung kann über alle Klassen insgesamt angegeben werden. Dabei wird die Änderung der Vorhersagegüte berechnet, wenn jede Verbindung zwischen der Erklärungs- und der Zielvariable gelöscht wird. Wenn sich in der Modellgüte große Änderungen ergeben, ist die Erklärungsvariable wichtig für das Modell.

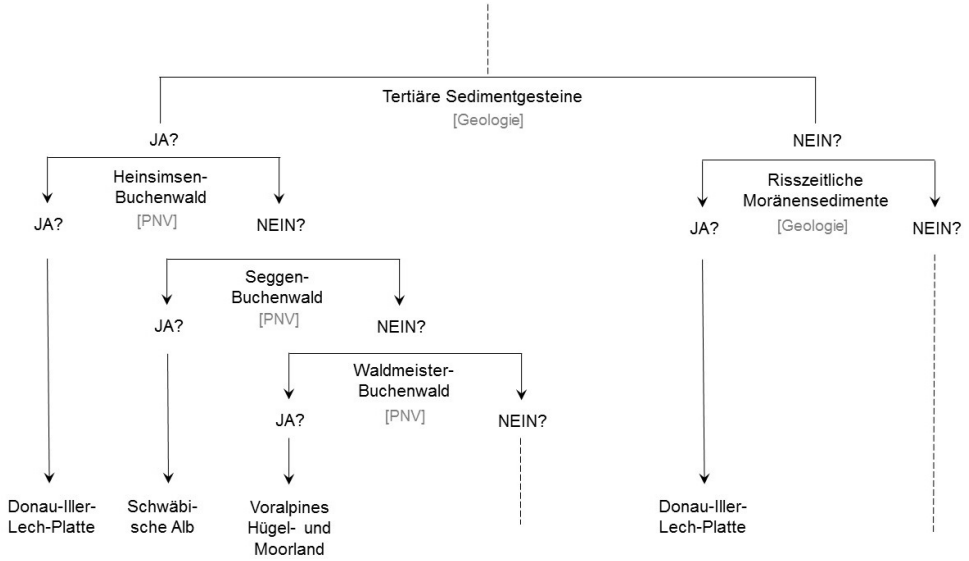

Abb. 3: Beispiel aus einem Entscheidungsbaum (vereinfachte Ansicht)

Dem Ausgangsdatensatz wurde eine geschichtete Stichprobe an Punktdaten (1 % der Gesamtfläche, d. h. 5.717 Punkte) mit einem Abstand von 100 m zu den Naturraumgrenzen der Zielvariable entnommen. Die Punktdaten wurden mittels eines *Random Point* Verfahrens anteilig auf die Flächen der jeweiligen Klassen in der naturräumlichen Gliederung erzeugt. Diesen Punktdaten wurden anschließend die lokalen Werte der 17 Erklärungsvariablen angefügt. Sie bilden gemeinsam den vollständigen Datensatz für die statistische Analyse. Von den 5.717 Punkten wurden automatisiert 4.573 als Trainingsdaten ausgewählt, um die verbliebenen 20 % der Punktdaten zu testen.

3 Ergebnisse

Das Model mit allen 17 Variablen berechnete eine Vorhersagegenauigkeit (*prediction success*) von 76,70 % (Testdatensatz). Laut *variable importance* handelt es sich bei den folgende sechs Erklärungsvariablen als die wichtigsten: 1) Geologie, 2) Channel Network Base Level, 3) Valley Depth, 4) Höhenmodell, 5) Potenzielle natürliche Vegetation und 6) Vertical Distance to Channel Network. Diese sechs Variablen wurden anhand ihrer relativen Erklärungskraft, größer wie 15 % ausgewählt.

Tabelle 2: Modellgüte der einzelnen Naturräume für den Testdatensatz (n = 1.181)

Naturraum	n	% richtig klassifiziert
03 Voralpines Hügel- und Moorland	80	93,75
04 Donau-Iller-Lech-Platte	102	94,12
09 Schwäbische Alb	168	91,07
10 Schwäbisches Keuper-Lias-Land	147	86,39
11 Fränkisches Keuper-Lias-Land	8	62,50
12 Neckar- und Tauber-Gäuplatten	295	87,12
13 Mainfränkische Platten	3	66,67
14 Odenwald	40	97,50
15 Schwarzwald	189	92,59
16 Hochrheingebiet	7	100,00
20 Südliches Oberrhein-Tiefland	36	83,33
21 Mittleres Oberrhein-Tiefland	48	95,83
22 Nördliches Oberrhein-Tiefland	58	93,10

Die Modellierung mit den oben genannten sechs Variablen führt zu einer Erklärung von 81,82 % (Testdatensatz) und wird in dieser Studie als Ergebnis vorgestellt (vgl. Tabelle 2).

4 Diskussion

Es wird vermutet, dass die geringere Vorhersagegenauigkeit des Models mit 17 Variablen auf eine Überanpassung (*overfitting*) der Variablen zurückzuführen ist. In Klassen mit geringerem Stichprobenumfang ist die Fehlklassifizierung beim Ansatz mit 17 Variablen deutlich höher als bei dem mit sechs Variablen.

Die Naturräume „Mittleres Oberrhein-Tiefland“ und „Nördliches Oberrhein-Tiefland“ werden vom Modell sehr gut, das heißt homogen, abgebildet. Sie bilden geologisch wie auch morphologisch klar abgrenzbaren Strukturen. Im Modell des Testdatensatzes wurde die Klasse „21“ (n = 48) zwei Mal anderen Klassen zugeordnet. Elemente der Klasse „22“ (n = 58) wurden viermal falsch zugeordnet.

In den „Neckar- und Tauber-Gäuplatten“ finden sich vereinzelt Einheiten aus der Klasse „Schwäbisches Keuper-Lias Land“ auf. Im Modell wurden 19 Stichprobenpunkte der Klasse „12“ (n = 295) der Klasse „10“ zugeordnet. Im Gegensatz dazu wurde die Klasse „10“ (n = 147) 19-mal der Klasse „12“ zugewiesen. Die Variable mit der höchsten Erklärungskraft ist für beide Klassen die Geologie. In Abbildung 4 (Box A) finden sich an dieser Stelle die geologischen Schichteinheiten des „Höheren Mittel- und Oberkeupers“, welcher ansonsten in Baden-Württemberg direkt an den „Unterjura“ angrenzt. Daher wurde dieser Bereich in den Naturraum „10“ fehlklassifiziert. Ein möglicher Lösungsansatz hierfür wäre eine höhere Gewichtung anderer Inputvariablen.

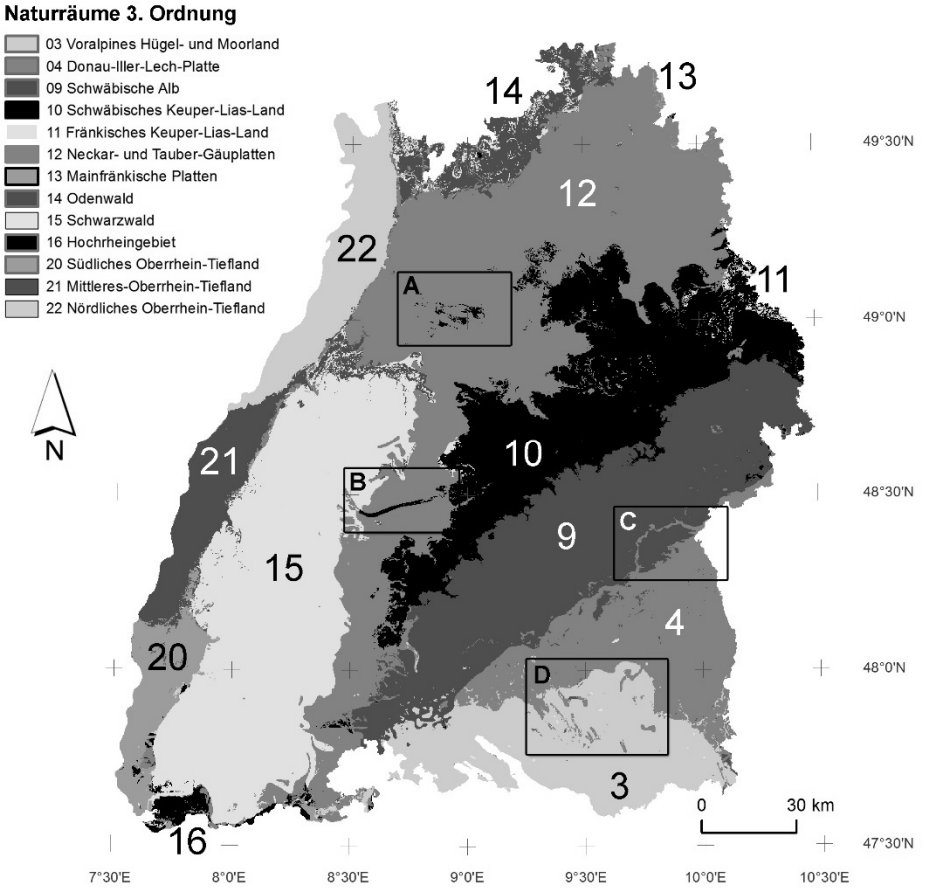

Abb. 4: Vorhersage der Naturräumlichen Ordnung Baden-Württembergs in der Rasterauflösung 250 m

In Box D findet sich die Einheit des „Voralpinen Hügel- und Moorlands“ der Klasse „Donau-Iller-Lech-Platten“. Hier beeinflusst die Morphologie in Form von Talstrukturen das Modellergebnis.

Im Ergebnis des Testdatensatzes wurde in den Klassen mit der geringsten Häufigkeit, den „Mainfränkische Platten“ (n = 3) und „Fränkisches Keuper-Lias-Land“ (n = 8), ein Fehler von 33,3 % bzw. 37,5 % berechnet. Die Klasse “Hochrheingebiet“ mit n = 7 wurde jedoch vollständig korrekt zugeordnet. Es ist zu überlegen, ob diese drei räumlich unterrepräsentierten Klassen ganz ausgeschlossen werden sollten, um das Gesamtergebnis nicht zu verfälschen.

Insgesamt fällt auf, dass die landschaftliche Gliederung nach MEYNEN & SCHMITHÜSEN (1953-1962) stark mit der Geologie korreliert, d. h. die Bearbeiter haben die landschaftlichen Einheiten vor allem auf Grundlage geologischer Karten abgegrenzt. Bei der Regionalisierung ist daher die geologische Struktur ebenfalls der wichtigste Indikator für die Raumgliederung,

wie zwischen Freudenstadt und Nürtingen (Box B in Abb. 4) zu erkennen ist. Dieses sogenannte „Schwäbische Lineament“ bezeichnet eine tektonische Störungszone, in der die geologischen Einheiten „Gipskeuper“ und „Bunte Mergel Schichten“ des „Schwäbischen Keuper-Lias Landes“ aufgeschlossen sind, daher auch die Zuordnung zu dieser Klasse. Im mittelpleistozänen Donautal zwischen Ehingen und Ulm (Box C) liegen die „Jungen Talfüllungen“ der quartären Geologie an der Oberfläche und das Modell weist daher diesen Bereich der Klasse „Donau-Iller-Lech-Platte“ zu.

5 Ausblick

Die Naturräume Baden-Württembergs sollen auf Basis von überwachten und unüberwachten geostatistischen Ansätzen (z. B. Co-Kriging oder Regression-Kriging) modelliert werden. Hierbei kommen Regressions- und Clusteranalysen zum Einsatz. Darüber hinaus soll die graphische Darstellung der Schätzgüte am jeweiligen Pixel in Form einer *error map* erstellt werden. Geplant ist ebenfalls eine Modellierung der Naturräume 4. Ordnung.

Um für die Verwendung in Planungsbehörden ein homogeneres Kartenbild zu erzeugen, wäre der Einsatz digitaler Filter zur Generalisierung der Grenzen denkbar.

Anmerkungen

Wir danken dem Gutachter für seine fruchtbaren Hinweise und Anregungen.

Literatur

BEVEN, K. & KIRKBY, M. J. (1979), A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Journal, 24 (1), 43-69.

BGR – Bundesanstalt für Geowissenschaften und Rohstoffe (2015), Geologische Karte der Bundesrepublik Deutschland 1:1.000.000 (GK1000).

BOEHNER, J. & CONRAD, O. (2008), Terrain Parameters described in the SAGA-GIS Software, v.2.1.0. http://sourceforge.net/projects/saga-gis/files/latest/download?source=files (16.06.2014).

BURAK, A. (2005), Eine prozessorientierte landschaftsökologische Gliederung Deutschlands: Ein konzeptioneller und methodischer Beitrag zur Typisierung von Landschaften in chorischer Dimension. Flensburg (Deutsche Akademie für Landeskunde, Selbstverlag; zugl. Diss. Univ. Bochum 2004).

CONRAD, O. (2002), Terrain Parameters described in the SAGA-GIS Software, v.2.1.0. http://sourceforge.net/projects/saga-gis/files/latest/download?source=files (16.06.2014).

DLR (2012), SRTM X-SAR Digital Elevation Models. http://eoweb.dlr.de:8080/eoweb-ng/licenseAgreements/DLR_SRTM_Readme.pdf (21.10.2015).

DONGUS, H. (1991), Die naturräumlichen Einheiten auf Blatt 187/193 Lindau-Oberstdorf. In: Geographische Landesaufnahme 1:200.000. Naturräumliche Gliederung Deutschlands. Hrsg. v. d. Bundesforschungsanstalt f. Landeskunde und Raumordnung. Bonn.

ELITH, J., LEATHWICK, J. R. & HASTIE, T. (2008), A working guide to boosted regression trees. Journal of Animal Ecology, 7 (4), 802-813.

FREEMAN, G. T. (1991), Calculating catchment area with divergent flow based on a regular grid. Computers and Geosciences, 17, 413-22.

KUTTNER, M., ESSL, F., PETERSEIL, J., DULLINGER, S., RABITSCH, W., SCHINDLER, S. & MOSER, D. (2015), A new high-resolution habitat distribution map for Austria, Liechtenstein, southern Germany, South Tyrol and Switzerland. eco.mont (Journal on Protected Mountain Areas Research), 7 (2), 18-29.

LANDSCHAFTSPARK BODENSEE-OBERSCHWABEN (2003), Gesamtdarstellung, erstellt von FUTOUR und HAGE+HOPPENSTEDT PARTNER, München/Rottenburg. http://www.hhp-raumentwicklung.de/materialien/up/LP_BO/Gesamtdarstellung.pdf (01.04.2015).

LEE, J. S. (1980), Digital image enhancement and noise filtering by use of local statistics. IEEE Trans Pattern Anal Mach Intell, 2 (2), 165-168.

LUBW (Landesanstalt für Umwelt, Messungen und Naturschutz BW) (2015). http://udo.lubw.baden-wuerttemberg.de/public/index.xhtml (02.07.2015).

MANNSFELD, K. (2005), Naturräumliche Gliederung Sachsens – Ordnung der Mannigfaltigkeit. In: Landschaftsgliederungen in Sachsen. Mitteilungen des Landesvereins Sächsischer Heimatschutz e. V., Sonderheft 2005, 2-8.

MEYNEN, E. & SCHMITHÜSEN, J. (Hrsg.) (1953-1962), Handbuch der naturräumlichen Gliederung Deutschlands. Bundesanstalt für Landeskunde und des Zentralausschusses für deutsche Landeskunde. Remagen.

MOORE, I. D., GRAYSON, R. B. & LADSON, A. R. (1991), Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes, 5 (1), 3-30.

PLANCHON, O. & DARBOUX, F. (2002), A fast, simple and versatile algorithm to fill the depressions of digital elevation models. CATENA, 46 (2-3), 159-176.

RICHTER, H. (1967), Naturräumliche Ordnung. Wissenschaftliche Abhandlungen der Geographischen Gesellschaft der DDR, 5. Berlin

RILEY, S. J., DEGLORIA, S. D. & ELLIOT, R. (1999), A Terrain Ruggedness Index that quantifies topographic heterogeneity. Intermountain Journal of Sciences, 5 (1-4), 23-27.

SCHRÖDER, W. & SCHMIDT, G. (2000), Raumgliederung für die Ökologische Umweltbeobachtung des Bundes und der Länder. Umweltwissenschaften und Schadstoff-Forschung, 12 (4), 236-243.

SYRBE, R.-U. (1999), Raumgliederungen im mittleren Maßstab. In: ZEPP, H. & MÜLLER, M. (Hrsg.): Landschaftsökologische Erfassungsstandards. Forschungen zur deutschen Landeskunde. Flensburg, 463-489.

TRAVIS, M. R., ELSNER, G. H., IVERSON, W. D. & JOHNSON, C. G. (1975), VIEWIT: Computation of Seen Areas, Slope, and Aspect for Land-Use Planning. Report PSW-11, Berkely, California, USA.

TÜXEN, R. (1956), Die heutige potentielle natürliche Vegetation als Gegenstand der Vegetationskartierung. Angewandte Pflanzensoziologie, 13, 5-42.

UBA – Umweltbundesamt (2004), CORINE Land Cover (CLC2000). http://www.corine.dfd.dlr.de/intro_de.html (07.08.2015).

ZEVENBERGEN, L. W. & THORNE, C. R. (1987), Quantitative analysis of land surface topography. Earth Surface Processes and Landforms, 12 (1), 47-56.