Zusammenfassung der Arbeiten im Berichtszeitraum 7.11.2018 bis 31.12.2019

SOIL-DE – Entwicklung von Indikatoren zur Bewertung der Ertragsfähigkeit, Nutzungsintensität und Vulnerabilität landwirtschaftlich genutzter Böden in Deutschland

Arbeitspaket AP 5: Analyse der Einflusses von Spektralinformationen auf die maßstabsspezifische Prognosegüte der Bodenparameter Humusgehalt und Bodenart

Dr. Markus Möller

Halle, 28.2.2020
TERRASYS geodatenanalyse (`https:\terrasys.github.io`) im Auftrag des Deutschen Zentrums für Luft- und Raumfahrt (DLR)

Inhaltsverzeichnis

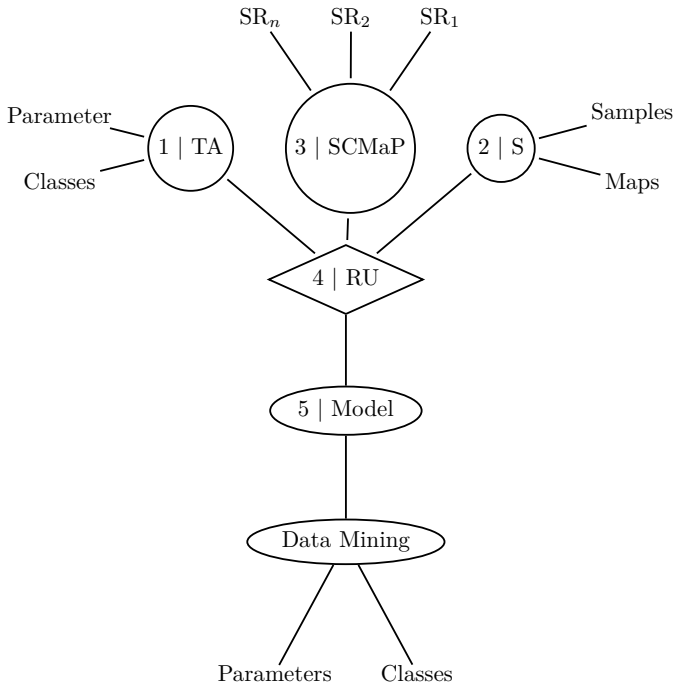

Abbildung 1: Fließschema zur maßstabsspezifischen Prognose von Bodeneigenschaften. SCMaP – Soil Composite Mapping Processor | S – Bodendaten | TA – Reliefanalyse | RU – Bezugseinheit | SR – multi-temporale Spektralreflektanzen.

1 Bodenprognose (Meilensteine 5.1. und 5.2)

Die Arbeiten im Berichtszeitraum fokussierten auf die Entwicklung einer Prozesskette, die eine vergleichende Bewertung des Einflusses erklärender Variablen auf die Prognose von Bodenparametern und -klassen erlaubt. Die Prozesskette ist in Abbildung 1 dargestellt und läasst sich in die folgenden Schritte gliedern:

1. Reliefinformationen (Parameter oder Klassen) ergeben sich aus der Anwendung von Reliefanalysealgorithmen (TA) auf digitale Höhenmodelle.

2. Als bodenkundliche Datengrundlagen (S) dienen Aufschlüsse (Samples) oder Attribute bodenkundlicher Kartenwerke (Maps), die wiederum unterschiedliche Maßstäbe repräsentieren.

3. Multi-temporale Reflektanzen (SR) sind das Ergebnis des Algorithmus “Soil Composite Mapping Processor” (Arbeitspaket AP 4).

4. Die Boden-, Relief- und Spektralinformationen werden mit Bezugseinheiten gekoppelt.

Tabelle 1: Reliefattribute

Abkürzung	Bedeutung	Quelle
FILL	Digitales Höhenmodell (= *Digital Elevation Model*) mit gefüllten Senken	(Planchon & Darboux, 2001)
SLP	Hangneigung	Zevenbergen & Thorne (1987)
CA	Fließakkumulation	Quinn et al. (1991)
VDC	Vertical Distance above Channel network	
TCI	Terrain Classification Index (TCI)	Bock et al. (2007)
TWI	Topographic Wetness Index	Beven & Kirkby (1979)
MBI	Mass Balance Index	Möller et al. (2008, 2012); Möller & Volk (2015)
MRVBF	Multiresolution Index of Valley Bottom Flatness	Gallant & Dowling (2003)
TOP	Topographic (positive) Openness	Yokoyama et al. (2002)
TON	Topographic (negative) Openness	Yokoyama et al. (2002)
NH	Relative Hangposition	Boehner & Selige (2006)
TPI	Topographic Position Index	Guisan et al. (1999)

5. Bei der eigentlichen maßstabsspezifischen Prognose werden Verfahren des maschinellen Lernens angewendet.

Alle Schritte der Prozesskette innerhalb der Programmumgebung **R** in Form von Funktionen umgesetzt (R Core Team, 2018). Die Funktionen sind auf der Plattform GitHub[1] abgelegt.

1.1 Reliefanalyse

Zwischen Relief- und Bodeneigenschaften bestehen enge Beziehungen (Ad-hoc AG Boden, 2005). Mit der flächenhaften Verfügbarkeit von digitalen Reliefdaten sind Reliefableitungen und -klassifikationen von besonderer Bedeutung für die digitale Prognose von Bodenklassen und -eigenschaften (Minasny & McBratney, 2016; Arrouays et al., 2020).

Um der Maßstabsabhängigkeit von Bodeneigenschaften gerecht werden zu können, zielt die Funktion `fTerrA()` zielt auf die Ableitung von bodenkundlich relevanten Re-

[1] `https://github.com/terrasys/ScaleP`

Tabelle 2: `fTerrA`: Parameter und Ergebnisdaten.

Parameter/Ergebnisse	**Bedeutung**
`DEM.DIR`	Verzeichnis mit DEM
`DEM`	DEM-Name ohne Dateiformat
`DEM.FRM`	DEM-Dateiformat
`OUT.DIR`	Ausgabeverzeichnis
`EPSG`	DEM-Projektion entsprechend `https://spatialreference.org`
`AGGREGATE`	Faktor zur Erhöhung der Rasterzellengröße
`TA`	Präfix der resultierenden Reliefattributedateien
`[TCI | MBI | NH | TPI | MRVBF] = TRUE`	Auswahl zu berechnender Reliefattribute
`VECTOR = TRUE`	Vektorisierung des aggregierten DEM-Datensates
`names_TerrainAttributes.txt`	Bedeutung der Reliefattributdateien
`[DEM]_AGGREGATE[AGG]_[TA][...].sgrd`	Namenskonvention der Reliefattributdateien (vgl. Tab. 1)
`[DEM]_AGGREGATE[AGG]_[TA].shp`	vektorisiertes aggregiertes Höhenmodell
`[DEM]_AGGREGATE[AGG]_ [channel network][...].sgrd`	verschiedene Aggregierungsniveaus des Tiefenliniennetzes (channel network)

liefattributen, die verschiedene Aggregations- bzw. Maßstabsniveaus des Reliefs repräsentieren (Tab. 1). Das betrifft die Reliefattribute NH und TPI, für die Varianten mit verschiedenen "Moving Window"-Größen abgeleitet worden sind. Die Berechnungsvarianten der Reliefattribute VDC und TCI basieren auf verschiedenen Aggregationsniveaus des Tiefenliniennetzes. Die MBI-Versionen sind schließlich Ausdruck von Varianten der Differenzierbarkeit von dominanten und subdominanten Reliefformen.

`fTerrA()` ist in erster Linie eine Sammlung von Funktionen des **R**-Paketes `RSAGA` (Brenning et al., 2018), worüber Funktionalitäten der Reliefanalysesoftware SAGA-GIS angesprochen werden können (Conrad et al., 2015). Tabelle 2 fasst die Parameter der Funktion `fTerrA()` zusammen, deren Aufruf dem Namen des digitalen Höhenmodells, dessen Dateiformat und Projektion sowie die Angabe eines Aggregierungsfaktor erfordert, mit dem wiederum die Rasterzzellengröße verändert werden kann. Zusätzlich sind die Ergebnisdateien aufgelistet, die beim Ausführen der Funktion entstehen. Dazu

Tabelle 3: `fZonaSt`: Parameter und Ergebnisdaten.

Parameter/Ergebnisse	**Bedeutung**
`TA.DIR`	Verzeichnis mit Reliefattributen [`*.sgrd`]
`TA`	Präfix der Reliefattribute und Spaltennamen der Ergebnisdatei
`POLYGON.DIR`	Verzeichnis mit Polygonvektordatensatz (Bezugseinheiten)
`POLYGON.SHP`	Name des Polygonvektordatensatzes [`*.shp`]
`OUT.DIR`	Ausgabeverzeichnis
`[DEM]_AGGREGATE[AGG]_[TA].shp`	Polygonvektordatensatz mit Reliefattributen
`[DEM]_AGGREGATE[AGG]_[TA]_CorMatr.csv`	Korrelationsmatrix aller Reliefattribute

gehören in erster Linie die Reliefattribute, die im SAGA GIS-Rasterformat `*.sgrd` vorgehalten werden. Die darüber hinaus abgeleiteten `*.shp`-Dateien repräsentieren das vektorisierte aggregierte Höhenmodell (Polygonvektordatensatz) sowie verschiedene Aggregierungsniveaus des Tiefenliniennetzen (channel network; Linienvektordatensatz).

1.2 Zonale Statistik

Die Funktion `fZonaSt()` beinhaltet einen zonalen Statistikalgorithmus (Tab. 3), der Bestandteil der **R**-Paketes `RSAGA` (Brenning et al., 2018) ist und die Verknüpfung von Rasterdaten und Bezugseinheiten erlaubt. Letztere repräsentieren bodenkundlich relevante Polygone, die beispielsweise aus der Verschneidung von verschiedenen Bodeneingangsdaten, aus der Segmentierung von Reliefattributen oder der Vektorisierung von Rasterzellen resultieren. Die Rasterdaten sind das Ergebnis der Funktionen `fTerrA()` (Kap. 1.1) und `fSCMaP()` (Kap. ???).

1.3 Modellierung

Die Prognose der Zielklassen und numerischen Parameter basiert auf Algorithmen des maschinellen Lernens (= *Data Mining*), die im **R**-Paket `caret` implementiert sind (Kuhn, 2008; Kuhn & Johnson, 2013). Die eigentliche Prognose wird mit dem im bodenkundlichen Kontext weit verbreitete und etablierte Entscheidungsbaumalgorithmus "Random Forest" durchgeführt (Behrens et al., 2018; Taghizadeh-Mehrjardi et al., 2020).

Tabelle 4: `ClasP`: Parameter und Ergebnisdaten.

Parameter/Ergebnisse	**Bedeutung**
`POLYGON.DIR`	Verzeichnis mit Polygonvektordatensatzes (Bezugseinheiten)
`POLYGON.SHP`	Vollständiger Name und Pfad des Polygonvektordatensatzes im `*.shp`-Format
`TRAIN.DIR`	Verzeichnis mit Trainingsdatensatz
`TRAIN.SHP`	Trainingsdatensatz im `*.shp`-Format
`OUT.DIR`	Ausgabeverzeichnis
`T.CLASS`	Spaltenname der Zielkategorie
`M.TRAIN`	Methode des maschinellen Lernens (`https://topepo.github.io/caret/train-models-by-tag.html`)
`PART`	Verhältnis ($\in [0,1]$) zwischen Trainungs- und Testdatensatz
`PF.TA`	`TRAIN.SHP`-Spaltennamen mit Reliefattributen
`UP.TRAIN=TRUE`	Ausgleich unbalanzieter Zielklassen im Trainingsdatensatz
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_part[PART].shp`	Trainingsdatensatz im `*.shp`-Format zur Modellbildung
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_part[PART].shp`	Testdatensatz im `*.shp`-Format zur Validierung
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_BP.shp`	Prozentuale Anteile der Zielklassen im Trainings- und Testdatensatz
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_VarImp.csv`	Bedeutung (Variable Importance) der erklärenden Variablen für die Prognose
`[TRAIN.SHP]_[POLYGON.SHP]_[T.CLASS]_MODEL-[M.TRAIN]_part[PART].shp`	Prognoseergebnis mit den Spalten [T.CLASS_SIM] und [T.CLASS_PRB]
`[TRAIN.SHP]_[POLYGON.SHP]_[T.CLASS]_MODEL-[M.TRAIN]_part[PART]_CV.txt`	Gesamtgenauigkeit basierend auf Kreuzvalidierung
`[TRAIN.SHP]_[POLYGON.SHP]_[T.CLASS]_MODEL-[M.TRAIN]_part[PART]_CM.csv`	Konfusionsmatrix basierend auf der Prognose des Testdatesatzes
`[TRAIN.SHP]_[POLYGON.SHP]_[T.CLASS]_MODEL-[M.TRAIN]_part[PART]_AM.csv`	Genauigkeitsmaße basierend auf der Konfusionsmatrix
`[TRAIN.SHP]_[POLYGON.SHP]_[T.CLASS]_MODEL-[M.TRAIN]_part[PART]_BP.pdf`	Prozentuale Anteile der Zielklassen im Prognoseergebnis

Tabelle 5: `NumP`: Parameter und Ergebnisdaten.

Parameter/Ergebnisse	**Bedeutung**
`POLYGON.DIR`	Verzeichnis mit Polygonvektordatensatzes (Bezugseinheiten)
`POLYGON.SHP`	Vollständiger Name und Pfad des Polygonvektordatensatzes im `*.shp`-Format
`TRAIN.DIR`	Verzeichnis mit Trainingsdatensatz
`TRAIN.SHP`	Trainingsdatensatz im `*.shp`-Format
`OUT.DIR`	Ausgabeverzeichnis
`T.NUM`	Spaltenname des numerischen Zielparameters
`M.TRAIN`	Methode des maschinellen Lernens (`https://topepo.github.io/caret/train-models-by-tag.html`)
`PART`	Verhältnis ($\in [0,1]$) zwischen Trainungs- und Testdatensatz
`PF.TA`	`TRAIN.SHP`-Spaltennamen mit Reliefattributen
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_part[PART].shp`	Trainingsdatensatz im `*.shp`-Format zur Modellbildung
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_part[PART].shp`	Testdatensatz im `*.shp`-Format zur Validierung
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_DP.shp`	Dichtefunktionen (Density plot) des Trainings- und Testdatensatzes
`[TRAIN.SHP]_[T.CLASS]_model-[M.TRAIN]_VarImp.csv`	Bedeutung (Variable Importance) der erklärenden Variablen für die Prognose
`[TRAIN.SHP]_[POLYGON.SHP]_[T.NUM]_MODEL-[M.TRAIN]_part[PART].shp`	Prognoseergebnis
`[TRAIN.SHP]_[POLYGON.SHP]_[T.NUM]_MODEL-[M.TRAIN]_part[PART]_CV.txt`	Gesamtgenauigkeit basierend auf Kreuzvalidierung
`[TRAIN.SHP]_[POLYGON.SHP]_[T.NUM]_MODEL-[M.TRAIN]_part[PART]_ACCtrain.csv`	Genauigkeitsmaße basierend auf der Prognose des Trainingsdatensatzes
`[TRAIN.SHP]_[POLYGON.SHP]_[T.NUM]_MODEL-[M.TRAIN]_part[PART]_ACCtest.csv`	Genauigkeitsmaße basierend auf der Prognose des Testdatensatzes

Das Verfahren teilt unter optionaler Berücksichtigung von thematischen Klassen den n-dimensionalen Merkmalsraum von erklärenden Variablen solange, bis der höchste statistische Zusammenhang bei minimaler Varianz erreicht wird (Breiman, 2001; Liaw & Wiener, 2002).

Zur Validierung der Klassifikationsergebnisse wird der Gesamtdatensatz unter Berücksichtigung der Zielklassen- bzw. Zielparameterverteilung in einen Trainings- und Testdatensatz von 75 % bzw. 25 % geteilt. Die Modellbildung basiert auf dem Trainingsdatensatz. Auf dessen Grundlage wird auch eine Kreuzvalidierung durchgeführt wird, aus der sich die Gesamtgenauigkeit des Modells ergibt. Der Testdatensatz dient der unabhängigen Validierung (Khaledian & Miller, 2020), auf den das trainierte Modell angewendet wird. Zur Bewertung der numersichen Parameter dienen das Bestimmtheitsmaß R^2 und das Streuungsmaß $RMSE$ (= *Root Mean Square Error*). Die Beurteilung der prognostizierten Klassen wird mithilfe der klassenbezogenen F1-Werte und der Gesamtgenauigkeit vorgenommen. Der F1-Wert stellt das gewichtete harmonische Mittel aus Genauigkeit ("precision") und Trefferquote ("recall") einer jeden Klasse dar (Manning et al., 2008): "precision" beschreibt das Verhältnis richtig klassifizierter Fälle zur Gesamtzahl der simulierten Fälle einer Klasse. "recall" kennzeichnet das Verhältnis richtig klassifizierter Fälle zur Gesamtzahl der tatsächlichen Fälle einer Klasse. Die Gesamtgenauigkeit berechnet sich wiederum aus dem Verhältnis aller korrekten Treffer und der Gesamtzahl der Fälle berechnet (Stehmann, 1997).

Die einzelnen Schritte der Klassifikationsprozedur sind in den Funktionen `fClasP()` und `fNumP()` zusammengefasst. In den Tabellen 4 und 5 sind die Parameter und Ergebnisdateien der Funktionen sowie deren Bedeutung aufgelistet. Neben dem Prognoseergebnis werden Dateien mit verschiedenen Genauigkeitsmaßen ausgegeben und visualisiert.

2 Bodenerosionsmodellierung (Meilenstein M5.3)

...

Literatur

Ad-hoc AG Boden (2005). *Bodenkundliche Kartieranleitung (KA 5)*. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. 5. Auflage.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., de Forges, A. C. R., Morgan, C. L., Roudier, P., Poggio, L., & Mulder, V. L. (2020). Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional*, 20, e00255.

Behrens, T., Schmidt, K., MacMillan, R. A., & Viscarra Rossel, R. A. (2018). Multi-scale digital soil mapping with deep learning. *Scientific Reports*, 8(1).

Beven, K. & Kirkby, M. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24(1), 43–69.

Bock, M., Boehner, J., Conrad, O., Koethe, R., & Ringeler, A. (2007). Methods for creating Functional Soil Databases and applying Digital Soil Mapping with SAGA GIS. In T. Hengl, P. Panagos, A. Jones, & G. Toth (Eds.), *Status and prospect of soil information in south-eastern Europe: soil databases, projects and applications*, number EUR 22646 EN in Scientific and Technical Research series (pp. 149–162). Luxemburg: Office for Official Publications of the European Communities.

Boehner, J. & Selige, T. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In J. Boehner, K. McCloy, & J. Strobl (Eds.), *SAGA – Analysis and Modelling Applications*, volume 115 of *Göttinger Geographische Abhandlungen* (pp. 13–28). Göttingen, Germany: University of Göttingen.

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

Brenning, A., Bangs, D., & Becker, M. (2018). *RSAGA: SAGA Geoprocessing and Terrain Analysis*. R package version 1.3.0.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Böhner, J. (2015). System for automated geoscientific analyses (saga) v. 2.1.4. *Geoscientific Model Development*, 8(7), 1991–2007.

Gallant, J. C. & Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research*, 39(12), 1347–1359.

Guisan, A., Weiss, S. B., & Weiss, A. D. (1999). GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, 143(1), 107–122.

Khaledian, Y. & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81, 401 – 418.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. New York, Heidelberg, Dordrecht, London: Springer.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18–22.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Minasny, B. & McBratney, A. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301–311.

Möller, M., Koschitzki, T., Hartmann, K.-J., & Jahn, R. (2012). Plausibility test of conceptual soil maps using relief parameters. *CATENA*, 88(1), 57–67.

Möller, M. & Volk, M. (2015). Effective map scales for soil transport processes and related process domains - statistical and spatial characterization of their scale-specific inaccuracies. *Geoderma*, 247-248, 151–160.

Möller, M., Volk, M., Friedrich, K., & Lymburner, L. (2008). Placing soil-genesis and transport processes into a landscape context: A multiscale terrain-analysis approach. *Journal of Plant Nutrition and Soil Science*, 171(3), 419–430.

Planchon, O. & Darboux, F. (2001). A fast, simple and versatile algorithm to fill the depressions of digital elevation models. *Catena*, 46, 159–176.

Quinn, P., Beven, K., Chevallier, P., & Planchon, O. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes*, 5, 59–79.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stehmann, S. (1997). Selecting and interpretation measures of thematic classification accuracy. *Remote Sensing of Environment*, 62, 77–89.

Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N., & Scholten, T. (2020). Synthetic resampling strategies and machine learning for digital soil mapping in iran. *European Journal of Soil Science*. im Druck.

Yokoyama, R., Shirasawa, M., & Pike, R. (2002). Visualizing topography by openness: A new application of image processing to digital elevation models. *Photogrammetric Engineering & Remote Sensing*, 68(3), 251–266.

Zevenbergen, L. W. & Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12(1).