Tagungsbeitrag zu: Jahrestagung der DBG, Kommission II
Titel der Tagung: Böden – eine endliche Ressource
Veranstalter: DBG, September 2009, Bonn
Berichte der DBG (nicht begutachtete online Publikation)
http://www.dbges.de

Einsatz eines Genetischen Algorithmus zur verbesserten PLS-Kalibrierung von Kennwerten der Bodenfruchtbarkeit (C_{org}, C_{hwl}) aus Spektroradiometerdaten

Michael Vohland[1], Christoph Emmerling[2]

Einleitung

Die Reflexionssignale von Böden im sichtbaren und nahen Infrarot-Wellenlängenbereich (VIS/NIR) sind immer „Mischsignaturen“, d.h. sie sind Ausdruck der sich in komplexer Art und Weise überlagernden Spektralsignaturen der jeweils vorhandenen Bodenkonstituenten. Spektroradiometerdaten liefern einerseits eine Fülle spektraler Information, die für quantitative Ansätze ausgenutzt werden kann; andererseits kann die Güte der Kalibrierung durch kollineare und eventuell verrauschte Spektralvariablen beeinträchtigt sein. Unter solchen statistischen Randbedingungen hat sich in der Chemometrie die Partial Least Squares Regression (PLS) als Standardwerkzeug etabliert. Untersucht wurde in der vorliegenden Studie, ob und wie stark die PLS-Kalibriergüte durch Variablenselektion mittels eines Genetischen Algorithmus (GA) verbessert werden kann. Diese Betrachtung erfolgte für ein heterogenes Kollektiv von spektroradiometrisch vermessenen Bodenproben am Beispiel der Zielgrößen C_{org} und C_{hwl}.

Material

Untersucht wurden Böden, die aufgrund ihres Ausgangssubstrates, der vorherrschenden Bodenart und der Bodennutzung deutliche Unterschiede in ihrem Gehalt an organischer Bodensubstanz (OBS) aufwiesen. Hierzu wurden in jeweils vierfacher Wiederholung Mischproben aus dem Oberboden eines jeweils landwirtschaftlich genutzten Kolluvisol-Tschernosems aus Rheinhessen, einer Sand-Braunerde aus Rostock, eines Auenbodens (allochthone Braunerde) bei Trier sowie eines Podsols aus Lias-Sandstein unter Kiefern beprobt. Dieses Probenkollektiv wurde durch Proben aus Bv- und Cv- Horizonten der Braunerden, sowie Ae-, Bh- und Cv-Horizonten der Podsole ebenfalls in vierfacher Wiederholung ergänzt. Die Untersuchung umfasste somit insgesamt 30 Proben (n = 30).

Die Mischproben wurden homogenisiert, < 2 mm gesiebt und anschließend luftgetrocknet gemörsert. Die Analyse von C- und N- Gehalten erfolgte durch trockene Veraschung im Sauerstoffstrom (EuroEA, Fa. HEKAtech). Karbonathaltige Proben wurden zuvor in einer 0,22 N HCl-Lösung mittels Ultra Turrax im Suspensionsverfahren vorbehandelt. Die Messung erfolgte am Shimadzu TOC-V- Analysator. Zur Bestimmung des heißwasserlöslichen C (C_{hwl}) wurden 10g Bodenprobe mit 50ml H_2O_{dest} mittels Kjeldatherm (Fa. Gerhardt) aufgeschlossen und anschließend ebenfalls am Elementaranalysator (Shimadzu TOC-V) analysiert.

Für die spektrometrische Analyse wurde luftgetrocknetes und gemörsertes Probenmaterial verwendet. Als Spektroradiometer wurde ein FieldSpec II Pro FR-Instrument (ASD) eingesetzt, das den VIS/NIR- Spektralbereich von 0,35-2,5 µm hochauflösend abdeckt. Die Messgeometrie (Nadirsicht des Sensors, 30° Beleuchtungs-Zenitwinkel) wurde für alle Proben konstant gehalten. Durch Referenzmessungen über einem Spectralon©-Panel wurden die absoluten bidirektionellen Reflexionswerte für jede Bodenprobe ermittelt.

Die relativ stark verrauschten Spektralbereiche (0,35-0,39 bzw. > 2,4 µm) wurden entfernt. Durch ein Resampling auf 10 nm Auflösung erfolgte eine weitere Reduzie-

Universität Trier, Fachbereich VI (Geographie/Geowissenschaften), 54286 Trier
[1] Fernerkundung und Geoinformationsverarbeitung, [2] Bodenkunde
vohland@uni-trier.de

rung auf insgesamt 201 Spektralvariablen (0,4; 0,41; 0,42 … 2,4 µm). Um Streueffekte und Variationen der Gesamtalbedo (bedingt z.B. durch bereits geringfügige Variationen der Messgeometrie) zu korrigieren, wurde eine Standardisierung der einzelnen Spektren mit der „Standard Normal Variate“ Transformation durchgeführt.

Der Genetische Algorithmus

Abb. 1 dokumentiert den implementierten Genetischen Algorithmus (GA). Als Fitnesskriterium diente das erzielte kreuzvalidierte Bestimmtheitsmaß der PLS-Regression (Leave-One-Out (LOO) Kreuzvalidierung) zur Schätzung der Zielvariablen C_{org} und C_{hwl}.

Durchgeführt wurden relativ kurze Läufe (200 Evaluationen pro Lauf), um zufällige Korrelationen („overfitting“) zu vermeiden. 100 Läufe bildeten einen Zyklus, nach 10 Zyklen entschied die Selektionshäufigkeit einer Variable im gesamten Prozess über ihre Verwendung im finalen PLS-Regressionsmodell.

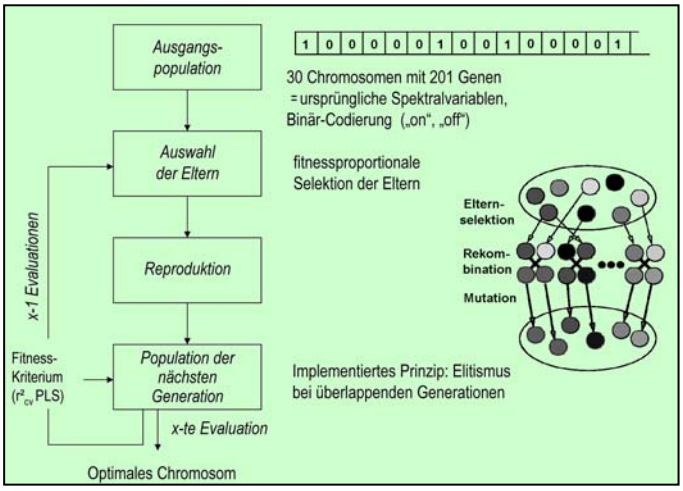

Abb. 1 Funktionsweise des GA zur Selektion von Spektralvariablen als Prädiktorvariablen im Rahmen der GA-PLS.

Ergebnisse

Zunächst wurden auf Basis aller 201 Spektralvariablen PLS-Regressionsgleichungen aufgestellt. Die Anzahl von Faktoren, die jeweils in die Modelle einbezogen wurden, wurde auf Basis einer LOO-Kreuzvalidierung ermittelt (C_{org}: 8 latente Variablen, C_{hwl}: 6 latente Variablen).

In der eigentlichen Kalibration konnten auf dieser Basis hohe Gütemaße erzielt werden. Für C_{org} ergab sich ein r^2 von 0,87, der root mean squared error (RMSE) betrug 0,26 % (relativer RMSE (rRMSE) = 0,22) und für den RDP (ratio of standard deviation (der gemessenen Werte) to RMSE of prediction) wurde ein Wert von 2,82 erzielt. In der internen Validierung (Kreuzvalidierung) erwiesen sich diese Ergebnisse aber als wenig stabil, das kreuzvalidierte Bestimmtheitsmaß betrug zum Beispiel nur noch 0,57 (Abb. 2).

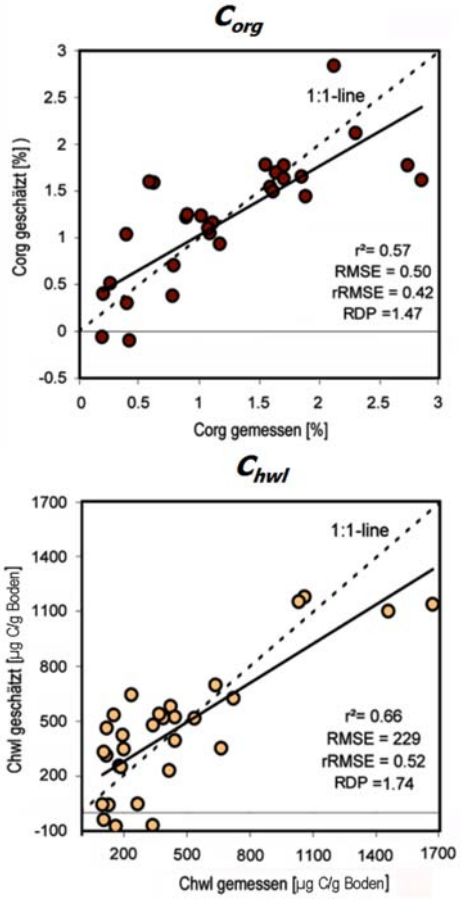

Abb. 2 PLS-Kalibrationsgüte (kreuzvalidiert) bei Verwendung aller 201 Spektralvariablen

Nachfolgend wurde der Genetische Algorithmus in der implementierten Form genutzt, um aus den 201 Spektralvariablen die bestgeeigneten Variablen auszuwählen. Auf Basis der Selektionshäufigkeiten der einzelnen Spektralvariablen für das jeweils fitteste Chromosom (in 10mal 100 Läufen) (Abb. 3) wurden für C_{org} 53, für C_{hwl} 43 Spektralvariablen ausgewählt und in die PLS-Analyse einbezogen.

Die zur Kalibrierung von C_{org} selektierten Spektralvariablen lagen hauptsächlich im

sichtbaren Bereich (520-660 nm) und weiterhin in den nIR-Bereichen 1770-1950 nm, 2010-2110 nm und 2200-2400 nm. Für C_{hwl} wurden nur Spektralvariablen jenseits des sichtbaren Bereichs ausgewählt (760-900 nm, 1000-1060 nm, 1240-1310 nm und 2150-2350 nm).

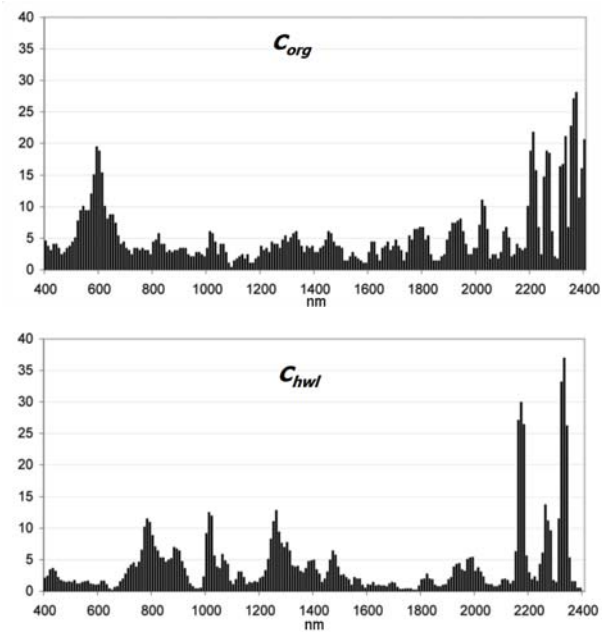

Abb. 3 Durchschnittliche Selektionshäufigkeit der Spektralvariablen bei Anwendung des GA.

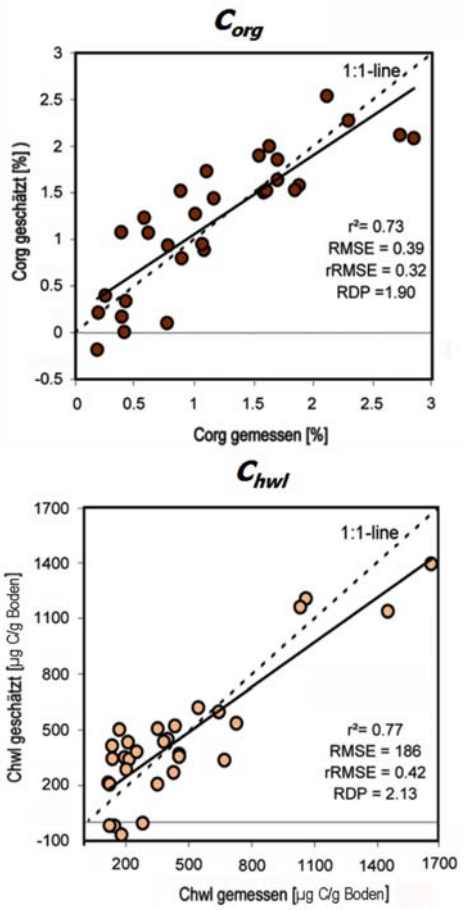

Abb. 4 GA-PLS-Kalibrationsgüte (kreuzvalidiert) bei Verwendung selektierter Variablen.

Durch Verwendung der GA-ausgewählten Spektralvariablen konnte die Güte der PLS-Regressionsmodelle deutlich verbessert werden – sowohl für C_{org} als auch für C_{hwl} (Abb. 4). Dabei wurden im Rahmen der PLS-Faktoranalyse wiederum 8 (C_{org}) bzw. 6 (C_{hwl}) latente Variablen (Faktoren) definiert und im Regressionsmodell verwendet. Für beide Bodengrößen lag die Trendlinie in den Scatterplots jetzt nahe der 1:1-Linie; alle berechneten (kreuzvalidierten) Gütemaße zeigten eine deutlich stabilere Kalibrierung durch GA-PLS anstelle der „einfachen“ PLS ohne „fitnessorientierte“ Selektion von Ausgangsvariablen.

Resümee

Trotz der in zahlreichen Studien dokumentierten Robustheit der PLS gegenüber verrauschten Spektralvariablen konnte die Güte der PLS-Modellierung durch eine gütemaßorientierte Variablenselektion mit einem GA (GA-PLS) deutlich gesteigert werden. Für beide untersuchten Größen erwiesen sich Spektralbereiche als besonders relevant, die wesentlich von mit OBS interagierenden Bodenkonstituenten beeinflusst werden (z.B. Tonminerale, Bodenfeuchte). Die für die Schätzung von C_{org} und C_{hwl} selektierten Spektralvariablen zeigten zwar Überlappungen, wiesen aber auch recht deutliche Unterschiede auf. Dies könnte darauf hindeuten, dass für die leicht umsetzbare C_{hwl}-Fraktion auch spektral gesehen spezifische Mechanismen und Interaktionen wirksam sind, die bei der Schätzung des Summenparameters C_{org} nur von untergeordneter Bedeutung sind. Allerdings liegt hier die Grenze der eingesetzten statistischen Methode, die auf die Verbesserung der Kalibrationsgüte abzielt, nicht aber kausale Zusammenhänge zwischen Spektralvariablen und Bodenkenngrößen aufdecken kann.

Danksagung

Die vorliegende Studie wurde durch den Forschungsfonds der Universität Trier gefördert.