„Data-Mining zur Bestimmung von Makronährstoffen (P) auf Basis kleinräumig erhobener Variablen“

Michael Marz, Peter Wagner, Thomas Chudy

Professur für Landwirtschaftliche Betriebslehre
Martin-Luther-Universität Halle-Wittenberg
Karl-Freiherr-von-Fritsch-Straße 4
06120 Halle (Saale)
michael.marz@landw.uni-halle.de

Abstract: Ausgehend vom kleinräumig erhobenen pH-Werten als Variable, sollen geeignete Methoden und Modelle des Data-Mining eruiert, angewandt und ggf. (weiter-)entwickelt werden, um die Makronährstoffe Phosphor und Kalium zu schätzen. Im Fokus dieses Beitrages stehen die Eingrenzung numerischer Modelle und Inputparameter, welche für weitere Schritte die bestmöglichen Ergebnisse zur Abschätzung von Phosphorgehalten im Boden versprechen. Die Untersuchungen zeigen, dass der Klassifizierungs- und Regressionsbaumknoten und das künstliche Neuronale Netz die geeignetsten Modelle sind, um die P-Gehalte im Boden anhand des pH-Wertes, der Bodenart (0-25cm) und des feldfruchtspezifischen Phosphorentzugs zu ermitteln.

1 Einleitung

Die kleinräumige Ermittlung von Phosphor (P)- und Kaliumgehalten (K) im Boden ist für eine effiziente Düngung von Schlägen relevant, um zum einen optimale Erträge erzielen und zum anderen möglichst ressourcenschonend wirtschaften zu können. Dauerversuche auf einem Testschlag verdeutlichen eine lokale Varianz zwischen 1,7 bis 25mg P/100g bzw. 4,5 bis 34,7mg K/100g Bodenprobe. Eine üblicherweise gleichmäßige Düngung orientiert sich nicht am realen Nährstoffbedarf: Das Ertragspotential wird nicht ausgeschöpft bzw. Dünger wird verschwendet und belastet die Umwelt. Gängige Methoden zur Bestimmung eines kleinräumigen Düngebedarfs dieser Makronährstoffe stützen sich auf eine Probennahme vor Ort und eine anschließende Laboranalyse [TH12]. Diese arbeits- und kostenintensiven Verfahren könnten dahingehend sowohl in wirtschaftlicher Hinsicht als auch in der räumlichen Auflösung verbessert werden, indem die Analyse sensorgestützt vor Ort zum Zeitpunkt der Befahrung bzw. Begehung erfolgt. Praxisrelevante Lösungsansätze zur vor-Ort-Analytik von P und K, wie es bei anderen für den Ackerbau relevanten Parametern möglich ist, existieren jedoch noch nicht. Die vorgestellte Arbeit gliedert sich ein in ein Projekt zur Entwicklung eines robusten Messsystems zur vor-Ort-Analytik des pH-Werts direkt bei der Bodenprobenahme im teilflächenbezogenen Ackerbau. Ausgehend vom kleinräumig erhobenen pH-Wert als Variable sowie weiteren parametrisierten statischen Größen und dynamischen Prozessen, sollen

geeignete Methoden und Modelle des Data-Mining eruiert, angewandt und ggf. (weiter-) entwickelt werden, um die Makronährstoffe P/K zu schätzen. Im Fokus dieses Beitrages zum Zeitpunkt der initialen Projektphase stehen die Eingrenzung numerischer Modelle und Inputparameter, welche für weitere Schritte die bestmöglichen Ergebnisse zur Abschätzung von Phosphorgehalten im Boden versprechen.

2 Methoden

Der hier vorgestellte Arbeitsablauf (Abbildung 1) der initialen Projektphase beschreibt die Schritte einer Vorauswahl relevanter Eingangsparameter sowie numerischer Modelle bei der eine bestmögliche Abschätzung von Phosphorgehalten im Boden zu erwarten ist. Mit den gewonnenen Erkenntnissen wird eine Grundlage für eine umfangreichere sowie detailliertere Untersuchung geschaffen.

Die Basis für alle Versuche der initialen Projektphase ist ein 64 Hektar großer Schlag bei Görzig in Sachsen-Anhalt. Für dieses Feld existieren im Rahmen des Monitorings jährliche Messungen der Parameter pflanzenverfügbarer Phosphor (P), Kalium (K), Magnesium (Mg) und pH an 46 (bis 2010) bzw. 45 (ab 2011) Probenahmepunkten. Für die gleichen Messgrößen stehen Daten von kleinräumigen Messungen mit 1079 (2007), 314 (2011) und 508 (August 2011) Probenahmen zur Verfügung. Die verwendeten Eingangsdaten werden in Primär- und Sekundärparameter gegliedert. Die Primärparameter P und pH werden als projektrelevante Größen immer berücksichtigt. Als Sekundärparameter werden Eingangsgrößen klassifiziert, die laut Literatur eine besondere Gewichtung bei der Verfügbarkeit von Phosphor aufweisen, jedoch bei der Modellierung wechselnd inkludiert werden. Diese umfassen Bodencharakteristika (Bodenart- und typ) aller Horizonte als Mischwert bzw. nur die des Oberbodens (0 – 25cm), scheinbare elektrische Leitfähigkeit (ECa) sowie natürliche Bedingungen die in der allgemeinen Bodenabtragsgleichung (ABAG) zusammengefasst bzw. auch in den einzelnen Faktoren (Regenerosivitätsfaktor - R, Bodenerodierbarkeitsfaktor - K, Hanglängen- und Hangneigungsfaktor - LS) ausgedrückt werden. Die Ergebnisfindung orientiert sich an der (räumlichen) Datenstruktur der Punkteingangsdaten der Primärparameter. Darüber hinaus unterscheiden sich die gewählten Sekundärparameter in ihrer räumlichen Auflösung und Verortung der Ausgangsinformationen. In der Konsequenz werden für eine Kongruenz zum Primärinput die Werte der Sekundärparameter durch Spline-Interpolation räumlich angeglichen. Die numerische Modellierung hat die Erklärung der Zielgröße P anhand verschiedener Eingangsgrößen bzw. Inputparameter zum Ziel. Eine zu Beginn ergebnisoffene Auswahl verschiedener numerischer Modelle soll dabei helfen, den besten Lösungsansatz zu identifizieren. Nach ihrer Eignung getestet werden künstliche neuronale Netze (KNN) [Ro93], Klassifikations- und Regressionsmethoden der Support Vector Machines (SVM) sowie k-Nearest-Neighbor (k-NN), einfache (lineare) Regressionsanalysen sowie darüber hinaus entscheidungsbaumbasierende Algorithmen Chi-square Automatic Interaction Detectors (CHAID) und ein Klassifizierungs- und Regressionsbaumknoten (C&R - Classification & Regression) [Lo11]. Die Ergebnisgüte wird anhand des Bestimmtheitsmaßes zwischen der Zielgröße P und des modellierten Parameters $P bestimmt.

Beginnend (**Fehler! Verweisquelle konnte nicht gefunden werden.**) mit einer (1) explorativen Gesamtdatenanalyse werden je Datensatz alle Primär- und Sekundärparameter bei einer numerischen Modellierung von P-Gehalten berücksichtigt. Im Vordergrund steht die Selektion der Parameter hinzu einem Minimalinput für geeignete Modelle. Im Anschluss (2) erfolgt eine detaillierte kleinräumige Evaluierung mit dem Ziel der Auswahl geeigneter Modelle unter einer möglichst praxisnahen Ausgangssituation. Diesbezüglich werden weitere, jedoch in der Datenbasis noch nicht jährlich verfügbare Parameter der ECa sowie des feldfruchtspezifischen P-Entzugs (fPe) in numerische Modelle überführt. Darüber hinaus erfolgt die Evaluierung unter dem Gesichtspunkt der Modellstabilität. Nach der Feststellung, welche Inputparameter und numerische Modelle bestmögliche Ergebnisse erreichen können, werden (3) exemplarische Evaluierungen zur Abschätzungsgenauigkeit durchgeführt. Die gewählten Modelle werden zunächst mit Werten eines Ausgangsjahres trainiert. Im Anschluss erfolgt eine Abschätzung der Ziel- bzw. Kontrollgröße P mit Daten eines Folgejahres. Mit einer Korrelation zwischen realen und modellierten Zielgrößen kann eine erste Aussage über die Modellgüte getroffen werden.

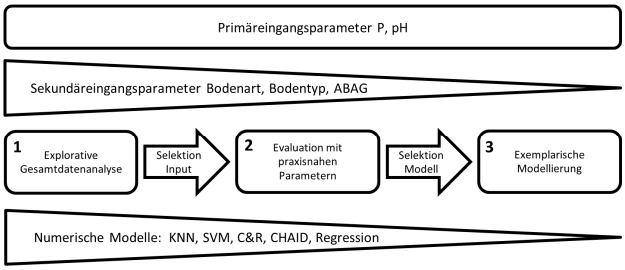

Abbildung 1: Arbeitsablauf – Selektive Wahl von Inputparametern und numerischen Modellen

3 Ergebnisse

In der explorativen Gesamtanalyse wurden verschiedene Datensätze (nach Jahr) des Monitoring und der Großuntersuchungen mit den zugehörigen Primär- sowie variierende Sekundärparameter einbezogen. Es hat sich gezeigt, dass für die Abschätzung des P-Gehaltes unter Einbezug des pH-Wertes hohe Bestimmtheitsmaße erreicht werden können, wenn als Sekundärparameter mindestens Bodencharakteristika des Oberbodens (0 bis 25cm) berücksichtigt werden. Dies inkludiert die Bodenart als nominale Kategorie sowie die prozentuale Korngrößenverteilung (Ton, Schluff, Sand). Das im Vergleich höchste Bestimmtheitsmaß für einen Datensatz unterscheidet sich nicht nennenswert und steht in keinem erkennbaren Zusammenhang zu weiteren Sekundärparametern. Bei der Gegenüberstellung der Datensätze des Monitoring beträgt das jeweils höchste Bestimmtheitsmaß $r^2 \geq 0{,}901$ (CHAID). Im Vergleich erreichen die numerischen Modelle mit den Datensätzen der Großuntersuchungen Bestimmtheitsmaße von $r^2 \geq 0{,}755$ (k-NN). Die im Anschluss erfolgte detaillierte kleinräumige Evaluierung identifiziert bestmöglich geeignete Modelle unter Einbezug des zuvor ermittelten Minimalinputs und zusätzlich praxisrelevanter Parameter der ECa und des fPe. Werden bei den Monitoring-

datensätzen die ECa und der fPe in der Analyse berücksichtigt, so erreicht der C&R-Baum ein Bestimmtheitsmaß von 0,976 und das KNN als zweitbestes Modell 0,962. Ein Ausschluss der ECa führt zu keiner signifikanten Veränderung. Das KNN erreicht beim Datensatz der Großuntersuchungen mit Abstand das beste Ergebnis ($r^2 = 0{,}835$). An zweiter Stelle ist CHAID ($r^2 = 0{,}709$) zu nennen. Ohne Berücksichtigung der ECa sinkt das Bestimmtheitsmaß des KNN auf 0,792. In beiden Fällen erhöht die Inkludierung des fPe das Bestimmtheitsmaß. Dies ist plausibel, da die P-Gehalte im Boden u.a. von der Entnahme durch die Feldfrüchte beeinflusst werden. Die Untersuchungen ergeben, dass der C&R-Baum und KNN die geeignetsten Modelle sind, um die P-Gehalte im Boden anhand des pH-Wertes, der Bodenart und des fPe zu ermitteln. Darauf aufbauend wurden zwei Tests zwischen unterschiedlichen Jahren durchgeführt: Die mit den Eingangsdaten des Monitoring aus dem Jahr 2008 trainierten Modelle erreichen für das Evaluierungsjahr 2010 die Bestimmtheitsmaße 0,717 (KNN) sowie 0,855 (C&R). Die zweite Evaluierung basiert auf zwei zeitlich aufeinanderfolgenden Datensätzen zwischen 2011 und 2012. Hierbei bildet das zuvor bessere Modell C&R-Baum die P-Gehalte wesentlich schlechter ab ($r^2 = 0{,}233$). Im Vergleich ist das KNN das bessere Modell ($r^2 = 0{,}698$).

4 Diskussion und Ausblick

Der Vergleich der Bestimmtheitsmaße aus den vorangegangenen Untersuchungen zwischen den Datensätzen der Monitoringpunkte und Großuntersuchungen deutet darauf hin, dass eine Maßstababhängigkeit zwischen Eingangsdaten (P-Gehalt, pH) in Bezug auf die Sekundärdaten (Bodenart) besteht. Eine Interpolation von Punktdaten, um eine räumliche Anpassung zu erreichen, wirkt sich scheinbar nicht positiv auf die Ergebnisse aus. Die Untersuchungen zeigten, dass der Klassifizierungs- und Regressionsbaumknoten und das künstliche Neuronale Netz die geeignetsten Modelle sind, um die P-Gehalte im Boden anhand des pH-Wertes, der Bodenart (0-25cm) und des feldfruchtspezifischen Phosphorentzugs zu ermitteln. In zukünftigen Projektschritten ist die Betrachtung räumlicher Zusammenhänge und eine Untersuchung der Struktur der Eingangsdaten besonders wichtig, um Fehlerquellen der Modellierung bestmöglich ausschließen zu können. Ein genereller Ausschluss von Fehlerquellen ist nicht möglich, da bereits die Datensammlung (Feldbeprobung, Laboranalyse) eine gewisse Ergebnisstreuung aufzeigt. Des Weiteren ist im Rahmen der Raum- und Datenanalyse eine konditionsgesteuerte Wahl von numerischen Modellen denkbar.

Literaturverzeichnis

[Lo11] Loh, W.-Y.: Classification and regression trees. In Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1; S. 14–23.

[Ro93] Rojas, R.: Theorie der neuronalen Netze. Eine systematische Einführung. Springer-Verlag, Berlin, New York, ©1993.

[TH12] Thun, R.; Hoffmann, G.: Die Untersuchung von Böden. VDLUFA-Verlag, Darmstadt, 2012.