Statistische Verfahren:
Projekt 4 - Nahinfrarotspektroskopie I

Tobias Giesemann
Ferdinand Rewicki
Moritz Preuß

1. April 2019

Zusammenfassung

In der vorliegenden Arbeit wird eine Methode zur Ableitung einer Kalibierfunktion zur Bestimmung des Stickstoffgehalts in Bodenproben vorgestellt. Basierend auf den zugrunde liegenden physikalischen und chemischen Eigenschaften der Nahinfrarotspektroskopie wird ein lineares Modell formuliert. Hierfür wurden potenzielle Prediktoren entsprechend ihrer Variabilität für das Maximalmodell ausgewählt und dann mit Hilfe von Mallow's Cp ein Idealmodell ermittelt. Darüber hinaus wurde der Einfluss des Stichprobenumfangs auf den durch den minimalen Cp-Wert geschätzten Vorhersagefehler untersucht.

Inhaltsverzeichnis

1 Einleitung **1**

2 Hintergrund **1**
2.1 Gebundener Stickstoff 1
2.2 Nahinfrarotspektroskopie 2

3 Methodik **2**
3.1 Datensatz 2
3.2 Statistisches Modell 2
3.3 Modellwahl im Falle der NIR-Spektroskopie 2
3.4 Modellselektion 3
3.5 Validierung des Modells 4
3.6 Theoretische Grundlagen der Simulation 4

4 Implementierung **4**
4.1 Modellwahl 4
4.2 Simulation 5

5 Ergebnisse und Diskussion **5**
5.1 Modellwahl 5
5.2 Simulation 6

6 Fazit **6**

A Einflussgrößen und Parameter des Modells **i**

B R Source Code **i**

Statistische Verfahren:
Projekt 4.1 - Nahinfrarotspektroskopie I

Tobias Giesemann
tobias.giesemann@uni-jena.de

Ferdinand Rewicki
ferdinand.rewicki@uni-jena.de

Moritz Preuß
moritz.preuss@uni-jena.de

Zusammenfassung

In der vorliegenden Arbeit wird eine Methode zur Ableitung einer Kalibierfunktion zur Bestimmung des Stickstoffgehalts in Bodenproben vorgestellt. Basierend auf den zugrunde liegenden physikalischen und chemischen Eigenschaften der Nahinfrarotspektroskopie wird ein lineares Modell formuliert. Hierfür wurden potenzielle Prediktoren entsprechend ihrer Variabilität für das Maximalmodell ausgewählt und dann mit Hilfe von Mallow's Cp ein Idealmodell ermittelt. Darüber hinaus wurde der Einfluss des Stichprobenumfangs auf den durch den minimalen Cp-Wert geschätzten Vorhersagefehler untersucht.

1 Einleitung

Die Zusammensetzung des Boden ist ein wesentlicher Faktor für einen ertragreichen und nachhaltigen Anbau von Pflanzen in der Landwirtschaft. Wichtige Parameter hierfür sind die Anteile des im Boden gebundenen Stickstoffs (N) und organischen Kohlenstoffs (SOC), durch deren Messung Erkenntnisse über die Fruchtbarkeit des Bodens und die Auswirkungen der Bodennutzung gewonnen werden können.[PDD$^+$13] Die Messung der genannten Parameter ist durch sogenannte Fraktionierung von Bodenproben möglich. Etablierte Messverfahren sind jedoch kostenintensiv, nicht vollständig standardisiert, und weisen eine schlechte Reproduzierbarkeit in verschiedenen Laboren auf.[PDD$^+$13] Aus diesem Grund sind Verfahren notwendig, welche eine zuverlässige und effiziente Ermittlung der Menge des organischen Kohlenstoffs und gebundenen Stickstoffs im Boden zulassen. Ein seit den sechziger Jahren vielfach eingesetztes Messverfahren, ist die sogenannten Nahinfrarotspektroskopie.[AHJ10] Dieses erlaubt die Schätzung der nur aufwendig bestimmbaren Einflussparametern auf den Anteil von N (bzw. SOC) in Bodenproben mit Hilfe der leicht messbaren Nahinfrarotspektren. Die Bestimmung und Validierung eines hierfür notwendigen statistischen Modells für den Stickstoffgehalt in der Bodenprobe, ist Gegenstand dieser Arbeit. Darüber hinaus widmet sich diese Arbeit einer statistischen Simulationsaufgabe. Dabei soll untersucht werden, inwiefern sich der über Mallow's C_p-Kriterium geschätzte erwartete Prognosefehler (siehe Kapitel 3) in Abhängigkeit der Stichprobengröße verändert.

2 Hintergrund

2.1 Gebundener Stickstoff

In der Natur geschieht eine fortwährender Austausch von Stickstoff zwischen Lebewesen, Boden und Atmosphäre. Ein Großteil des vorhandenen Stickstoffs liegt in gebundener Form im Erdboden vor und ist ein unentbehrlicher Nährstoff für Pflanzen und Lebewesen. Er wird von Pflanzen beim Wachstum aus der Erde aufgenommen und beim Absterben wieder freigesetzt. Zur Steigerung der Ernte ist eine Anreicherung des Bodens mit Nitrat durch der Einsatz von Düngemitteln daher gängige Praxis in der Landwirtschaft.[Umw17] Dies kann zu Problemen führen, wenn es zu einer Übersättigung des Bodens mit Stickstoff kommt. Durch Ausschwemmen des in Form von Nitrat (NO_3^-) und Ammonium (NH_4^+) gebundenen Stickstoffs kann dieser ins Grundwasser gelangen und eine Gefahr für die Umwelt darstellen. Sowohl für den Erfolg der Landwirtschaft als auch für den Schutz der Umwelt ist daher eine zuverlässige und effiziente Ermittlung des Nitratgehalts im Bodens von entscheidender Bedeutung. Der Messung des Stickstoffs liegen folgende chemische Zusammenhänge zu Grunde:

Die Stoffmengenkonzentration von Stickstoff $c_{(N)}$ lässt sich berechnen durch

$$c^{(\mathrm{N})} := \frac{n^{(\mathrm{N})}}{V}$$

wobei V das Volumen der Lösung und $n_{(N)}$ die enthaltenen Stoffmenge von Stickstoff ist.

Für eine gegebene Stoffmengenkonzentration c_0 und Stoffmenge n_o einer Probe, definieren wir den Stoffmengen-

anteil $y_{(\mathrm{N})}$ von Stickstoff als,

$$y^{(\mathrm{N})} \coloneqq \frac{c^{(\mathrm{N})}}{c_0} = \frac{n^{(\mathrm{N})}}{n_0}$$

2.2 Nahinfrarotspektroskopie

Bei der Nahinfrarotspektroskopie kommen elektromagnetische Wellen im Bereich zwischen 120 THz und 400 THz bzw. 2.500 nm und 750 nm zum Einsatz.[AHJ10] Das Messverfahren nutzt die Tatsache, dass bei der Bestrahlung einer Probe Teile des Lichts reflektiert, hindurch gelassen oder absorbiert werden. Von besonderem Interesse ist hierbei die Reflexion des Lichts, welche sich in die zwei Komponenten „Spiegelreflexion“ und „diffuse Reflexion“ unterscheiden lässt. Aus den Teilen des Lichts, welche diffus reflektiert werden, können aufgrund der größeren Eindringtiefe Informationen über die Beschaffenheit der Probe gewonnen werden.[AHJ10] Für eine Wellenlänge λ ist das relative Reflexionsvermögen $\delta(\lambda)$ definiert als:

$$\delta\colon (0,\infty) \to (0,\infty), \qquad \delta(\lambda) \coloneqq \frac{P_{\mathrm{r}}(\lambda)}{P_s}$$

wobei $P_{\mathrm{r}}(\lambda)$ die Reflexion einer Probe und P_0 die Reflexion eines Materials mit einem Reflexionsanteil nahe 100% ist.

Weiterhin lässt sich unter Anwendung des Beer'schen Gesetzes ein approximativer Zusammenhang zwischen dem relativen Reflexionsvermögen und der Stoffmenge $c_{(\mathrm{N})}$ des Stickstoffs herstellen. In eine Probe mit n verschiedenen Stoffen sei c_i die Stoffmengenkonzentration des $i$ten in der Probe enthaltenen Stoffes. Dann sei $\varepsilon_i(\lambda)$ ein Koeffizient mit $i \in \mathbb{N}, i \leq n$ sodass

$$-\log \delta(\lambda) = -\log \frac{P_{\mathrm{r}}(\lambda)}{P_s} = \sum_{i=1}^{n} \varepsilon_i(\lambda) c_i$$

3 Methodik

3.1 Datensatz

Der für die Modellwahl verwendete Datensatz beinhaltet die logarithmierten relativen Reflexionswerte $\delta(\lambda)$ bei Wellenlängen zwischen 1400 nm und 2672 nm in einem Abstand von 4 nm sowie die Stoffmengeanteile $y^{(\mathrm{N})}$, $y^{(\mathrm{SOC})}$ und entsprechenden pH-Werte von insgesamt 533 Proben. Für diese Arbeit sind lediglich die gegeben Reflexionswerte und der Stoffmengenanteil des Stickstoffes von Interesse.

Informationen über die chemische Zusammensetzung der Probe in den Reflexionswerten im Nahinfrarotbereich sind stark überlagert.[AHJ10] Aus diesem Grund ist eine sorgfältige Auswahl relevanter Wellenlängen von besonderer Bedeutung für die Erstellung eines zuverlässigen Modells.

3.2 Statistisches Modell

Sei $n \in \mathbb{N}$ die Größe des Datensatzes und $d \in \mathbb{N}$ mit $d < n$ die Anzahl der Wellenlängen im Datensatz. Entsprechend Abschnitt 2.2 definieren wir dann die Einflussgröße x_{ik} für die $i$te Probe und $k$te Wellenlänge als

$$x_{ik} \coloneqq -\log \delta_i(\lambda_k)$$

für jedes $i, k \in \mathbb{N}, i \leq n, k \leq d$.

Der Stoffmengenanteil des Stickstoffs $y^{(\mathrm{N})}$ stellt eine Ausprägung der Zielgröße unseres späteren Modells dar. Wir definieren hierfür den Vektor der Stoffmengenanteile y_i der $i$ten Probe als den n-dimensionalen Vektor

$$y \coloneqq \left(y_i^{(\mathrm{N})}\right)$$

Nachdem wir sowohl die Einflussgrößen als auch die Zielgröße für das lineare Modell definiert haben, lassen sich diese nun in Zusammenhang bringen. Es ist valide anzunehmen, dass sich die Zielgröße durch einen Linearkombination der Einflussgrößen beschreiben lässt. Hierfür definieren wir zunächst Y als einen zufälligen Vektor von y mit

$$\mathbb{E}\, Y \coloneqq \beta_0 + \sum_{j=1}^{k} x_{ik}\beta_j$$

Zudem ist es notwendig eine Variable ε einzuführen welchen den Zufall der Messungen beschreibt. In Matrixschreibweise lässt sich dies durch die Designmatrix $\mathbb{X} \in \mathbb{R}^{n\times(d+1)}$, dem Parametervektor $\beta \in \mathbb{R}^{d+1}$ und dem stochastisch verteilten Parameter ε wie folgt darstellen.

$$Y = \mathbb{X}\beta + \varepsilon$$

Für den Zufallsparameter ε wird weiterhin gefordert, dass

$$\mathbb{E}\,\varepsilon = 0, \qquad \operatorname{cov}\varepsilon = \sigma^2 \mathrm{I}$$

wobei $\sigma^2 \in (0,\infty)$. Weiterhin soll angenommen werden, dass ε normalverteilt ist mit

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}\right)$$

sodass sich für das Gesamtmodell gilt

$$Y \sim \mathcal{N}\left(\mathbb{X}\beta, \sigma^2 \mathrm{I}\right)$$

3.3 Modellwahl im Falle der NIR-Spektroskopie

Seien $Y \in \mathbb{R}^{\mathrm{n}}$ die Zielgröße in einem statistischen Modellwahlverfahren und $\mathbb{X} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$ eine Designmatrix. Zur Wahl einer geeigneten Menge von k Einflussparametern auf die Zielgröße y_i wird klassischerweise die Modellwahl über eine hierarchische Aufstellung von linearen Modellen erreicht. Beginnend mit dem minimalen Modell $E(Y_i) = \beta_0$ werden nach und nach neue potenzielle Einflussparameter x_{ik}

in das Modell hinzugefügt. Zu jedem dieser neuen x_{ik} wird dann eine Teststatistik aufgestellt, die darauf hinweist, ob der gewählte Parameter wichtig ist ist oder nicht. Dabei ist die Nullhypothese, dass x_{ik} keinen Einfluss auf die Zielgröße hat: $H_0 : \beta_k = 0$ und wird abgelehnt, falls $H_1 : \beta_k \neq 0$ zutrifft. Dies wird über die T-Teststatistik erreicht, wobei für den Fall, dass H_0 richtig ist, gilt:

$$\frac{\hat{\beta_k}}{\sqrt{\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}_{kk}}} \sim t_{n-(k+1)}.$$

Dieses Verfahren ist vor allem dann besonders gut geeignet, wenn man bereits theoretisch fundierte Annahmen über die Einflussgrößen machen kann [Sch19]. Mit diesem Modellwahlverfahren ergeben sich hier allerdings einige Schwierigkeiten: Durch die große Anzahl potenzieller Einflussgrößen im Form von Reflexionswerten kann a priori nur schwer eine inhaltliche Deutung vorgenommen werden. Daher ist Identifizierung weniger wichtiger Einflussgrößen und somit eine effektive hierarchische Modellwahl nicht möglich. Demnach muss in dieser Arbeit die Anzahl der möglichen Einflussgrößen stark erhöht werden und hier bekommen wir ein Problem mit der T-Teststatistik. Es ließen sich sehr viele unterschiedliche Kombinationen von Einflussgrößen aufstellen und in eine hierarchische Form bringen. Doch da wir bei der T-Teststatistik ein *zufälliges* Intervall konstruieren, gegen das unsere Hypothese getestet wird, steigt die Wahrscheinlichkeit, bei oft wiederholten Tests fälschlicherweise die Nullhypothese abzulehnen mit Anzahl der Versuchen. Ein automatisiertes Modellwahlverfahren, das viele unterschiedliche Modelle vergleichen kann, ist also mittels des T-Tests nicht zu erreichen [Sch19]. Stattdessen bietet sich eine Modellwahl basierend auf dem erwarteten Prognosefehler (ßum of prediction squared error", SPSE) an:

$$SPSE \coloneqq \mathbb{E}(\sum_{i=1}^{n}(Y_{i+n} - \hat{Y_i}^{(M)})^2)$$

Hierbei sind die Werte in Y_{i+n} neue Beobachtungen zum Erwartungsvektor x_i und $\hat{Y_i}^{(\mathrm{M})} = x_i^{(\mathrm{M})}\hat{\beta_i}^{(\mathrm{M})}$ die Prognosewerte aus dem zu testenden Modell M. Der Prognosefehler lässt sich in 3 Terme zerlegen: Einen irreduzierbaren Prognosefehler, der unabhängig von dem momentan betrachteten Modell ist, einen Biasterm, der die Abweichung des aktuellen Modells M vom Prognosemodell als Summe der quadrierten Prognose-Verzerrungen anzeigt und einen Varianzterm, der die Ungenauigkeiten widerspiegelt, die sich aus der Schätzung von $p = (|M| + 1)$ unbekannten Parametern ergibt:

$$SPSE^{(\mathrm{M})} = n\sigma^2 + p\sigma^2 + (bias^{(\mathrm{M})})^2$$

Der SPSE lässt sich über unterschiedliche Wege schätzen:
(1) mithilfe neuer Beobachtungen,
(2) (wiederholter) Zerlegung der Ursprungsdaten in Test- und Trainingsdaten (Kreuzvalidierung) oder
(3) mittels Schätzung basierend auf der Residuenquadratsumme (residual squared sum", RSS), hier im Vergleich zu o.g. SPSE:

$$RSS^{(M)} \coloneqq \sum_{i=1}^{n}(Y_i - \hat{Y_i}^{(M)})^2$$

$$SPSE^{(M)} \coloneqq \sum_{i=1}^{n}\mathbb{E}(Y_{i+n} - \hat{Y_i}^{(M)})^2$$

Es kann gezeigt werden, dass RSS den Wert von SPSE systematisch unterschätzt, dies jedoch durch die Verwendung der geschätzten Varianz des maximalen Modells behoben werden kann.[Sch19]:

$$SPSE^{(\mathrm{M})} \coloneqq RSS^{(\mathrm{M})} + 2\tilde{\sigma}^2_{full}(k+1) \tag{1}$$

Die Minimierung des SPSE entspricht der Minimierung des Mallow's Cp- Kriteriums, das für die folgenden Analysen getestet werden soll. Dabei gilt:

$$C_p^{(\mathrm{M})} = \frac{1}{\sigma^2_{\mathrm{full}}}\sum_{i=1}^{\mathrm{n}}(y_i - \hat{y_i}^{(\mathrm{M})})^2 - n + 2(k+1)$$

3.4 Modellselektion

Sei Λ das Spektrum der erhobenen Wellenlängen im vorliegenden Datensatz. Das erste Ziel dieser Arbeit ist es, diejenigen Wellenlängen $\lambda \in \Lambda$ herauszufinden, deren logarithmierte Reflektionswerte x_j durch den vom Stickstoffgehalt der Bodenproben beeinflusst werden. Es wurden mehrere Selektionsverfahren verglichen, die auf zwei Annahmen basieren:
(1) Die erste Ableitung x'_j der Reflexionswerte zeigt an die *Veränderungen* der x_j über das gesamte Spektrum an. Da bei der großen Anzahl an potenziellen Einflussgrößen eine Vorauswahl schwierig ist, wurden also diejenigen mit auffälligem mathematischen Verhalten ausgewählt: (a) diejenigen Reflexionen, deren 1. Ableitung über einem Schwellwert τ_1 lag oder (b) diejenigen, deren Werte unterhalb eines Schwellwertes lagen.
(2) Darüber hinaus wurden diejenigen Wellenlängen ausgewählt, die eine hohe Variablilität $var(x'_j) > \tau_2$ aufweisen. Zur Berechnung der Variablilität wurden wiederum unterschiedliche Verfahren angewandt. In Modell (a) wurde die Differenz zwischen dem minimalen und dem maximalen Ableitungswert normaiert auf den Mittelwert, während (b) und (c) ohne Normierung arbeiten. In (b) wird lediglich

zusätzlich der Betrag verwendet:

(a) $var(x'_j) = \frac{|max(x'_j) - min(x'_j)|}{mean(x'_j)}$
(b) $var(x'_j) = |max(x'_j) - min(x'_j)|$
(c) $var(x'_j) = max(x'_j) - min(x'_j)$

Der Vergleich zwischen den Modellen ergab, dass der SPSE am kleinsten wird, wenn wir diejenigen Reflexionen verwenden, deren erste Ableitung und Variablilität über τ_1, liegt. (2b), oder:

$$X_{select} = \{x'_j | max(x'_j) - min(x'_j)| > \tau_1\}$$

Die so ausgewählten Wellenlängen wurden als Maximalmodell in der Berechnung von Mallow's C_p gegeben. Da immer noch sehr viele Variablen im Modell sind, wurde die performantere Rückwärtsselektion für die konkrete Berechnung verwendet.

3.5 Validierung des Modells

Ein Parameter zur Messung der globalen Anpassungsgüte einer Regression ist R^2, definiert als [Lan07]

$$R^{2^{(M)}} \coloneqq \frac{\sum_{i=1}^{n} \left(\hat{y}_i^{(M)} - \overline{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}$$

Dieser beschreibt den Anteil der durch die Regression erklärten Varianz in den Daten und nimmt Werte zwischen 0 und 1 an. Während ein R^2 von 1 für einen perfekten linearen Zusammenhang steht bedeutet ein Wert von 0, dass kein linearer Zusammenhang vorliegt. Für das gewählte Modell ergibt sich ein Wert von $R^2 = 0.82$. Zudem ist es sinnvoll neben dem Parameter R^2 ein Korrelationsdiagramm zwischen der durch das Modell geschätzten Ausprägung $\hat{y}_i$ und dem wahren Wert y_i zu betrachten. Dies ist in Abbildung 2 dargestellt.

3.6 Theoretische Grundlagen der Simulation

Wie in Abschnitt 3.3 beschrieben gibt der SPSE den erwarteten Prognosefehler an. Es ist offensichtlich, dass dieser Wert möglichst gering sein sollte, um ein verlässliche Vorhersagen zu neuen Daten zu generieren.

Im Allgemeine kann der SPSE mit Hilfe neuer Beobachtungen geschätzt werden[Sch19]. Da diese nicht zur Verfügung stehen werden stattdessen neuen Pseudobeobachtungen der Zielgröße mit Hilfe des gewählten Modells generiert. Der aus der Pseudobeobachtungen geschätzte Prognosefehler kann der mit dem wahren wahren Prognosefehler des Modellsverglichen werden. Bei gegebenem Modells M wird der wahre Prognosefehler durch

$$SPSE^{(\mathrm{M})} \coloneqq n\sigma^2 + (|M| + 1)\,\sigma^2$$

berechnet.

Der neue Vektor der Pseudobeobachtungen ist gegeben durch,

$$\widetilde{Y} \coloneqq \widetilde{y}_i + \varepsilon$$

wobei für den stochastisch verteilen Parameter ε gilt

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}\right)$$

und als Varianz die Varianz des Modells angenommen wird.

$$\sigma^2 \coloneqq \left(\tilde{\sigma}^2\right)^{(M)}$$

Aus den generierten Pseudodaten wird ein neues bestes Modell $\widetilde{M}$ auf basierend auf dem kleinsten Cp-Wert ausgewählt, für dass gilt.

$$\widetilde{Y} \sim \mathcal{N}\left(\mathbb{X}^{(\widetilde{\mathrm{M}})}\widetilde{\beta}, \sigma^2 I\right)$$

Hierfür wird die Varianz als unbekannt angenommen. Neben dem Vergleich des wahren und geschätzten Prognosefehlers bei gleicher Stichprobengröße soll zudem überprüft werden welcher Einfluss die Änderung der Stichprobengröße auf den Wert des geschätzten SPSE hat. Hierzu werden die verschiedenen Stichprobengrößen von $n = (150, 200, 250, 300, 350, 400, 450, 500, 533)$ gewählt und zufällig aus dem Datensatz ausgewählt. Um die Verzerrung der Ergebnisse durch zufällige Ausreißer zu vermeiden werden die Simulationen mehrfach durchgeführt.

4 Implementierung

4.1 Modellwahl

Wie in Abschnitt 3.4 beschrieben, benötigen wir zur Auswahl der relevanten Prädiktoren die erste Ableitung der Reflektionswerte. Die Methode *getSlope* berechnet diese als Differenz benachbarter Messwerte. Daraus ergibt sich in Methode *criterionSlopeDist* die Variabilität je Wellenlänge als Differenz des größten und des kleinsten Wertes. Alle Wellenlängen mit einer Variablität größer einem vorgegebenen Schwellwert (hier: 0.001) werden in Methode *selectFeatures* für das Maximalmodell ausgewählt. Alle ausgewählen Variablen gehen linear in das Maximalmodell ein.

Mittels der im Paket leaps bereitgestellten Methode regsubsets wird aus dem 149 Prädiktoren umfassenden Maximalmodell das Modell als Bestes bestimmt, welches den kleinsten C_p Wert aufweist. Die Variablen sowie die geschätzen Parameter des ausgewählen Modells sind in Tabelle 2 dargestellt.

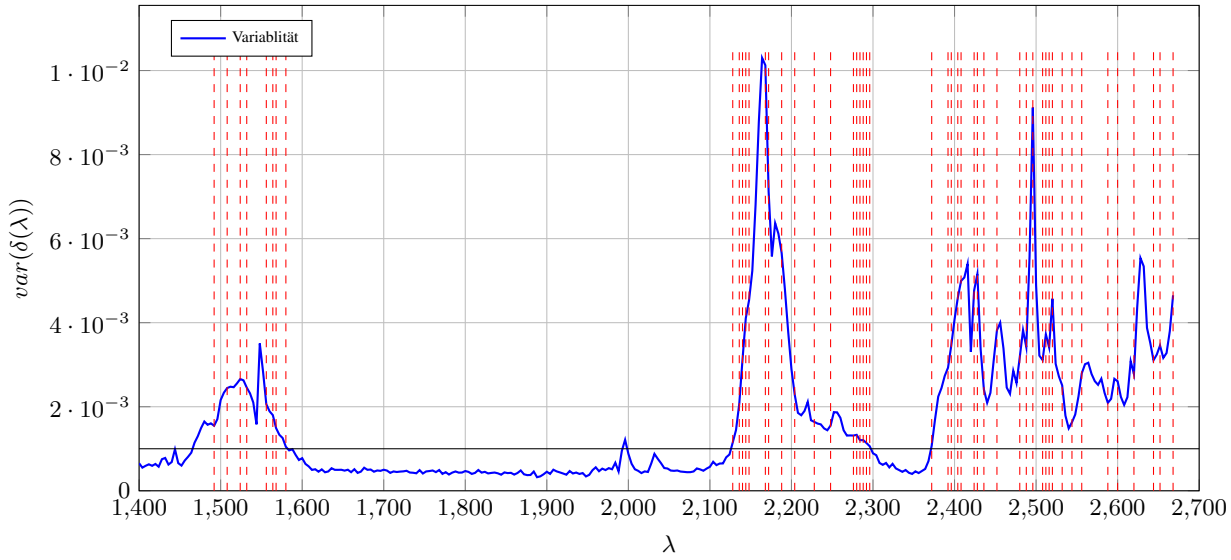

Abbildung 1: Variablität der Reflektionswerte. Die horizontale Linie markiert den Schwellwert von 0.001, die vertikalen Linien die im optimalen Modell enthaltenen Features.

4.2 Simulation

Das so ausgewählte, optimale Modell wird nun verwendet, um Pseudobeobachtungswerte zu simulieren. Die Simulation erfolgt in mehreren Runden und für verschiedene Stichprobengrößen von 150, 200, 250, 300, 350, 400, 450 und 500 zufällig ausgewählten sowie für die gesamten 533 Spektren des vorliegenden Datensatzes. In Methode *simulateOnDatSubset* erfolgt zunächst die zufällige Auswahl der übergebenen Anzahl an Sprektren. Anschließend wird das unter Abschnitt 4.1 bestimmte, optimale Modell verwendet, um neue Stockstoffwerte zu erzeugen, wobei, wie in Abschnitt 3.6 beschrieben, eine Normalverteilung der Zufallsgröße angenommen wird. In jedem Simulationsdurchlauf wird der Ergebnisvektor als arithmetisches Mittel aus 1000 Durchgängen erzeugt. Mit den so erzeugten Pseudobeobachtungen, sowie den originalen Reflektionswerten wird nun ein zweites mal die Methode *regsubsets* aus dem *leaps* Paket aufgerufen, um für den neuen Datensatz das beste Modell mittels Mallows' C_p-Kriterium zu bestimmen. Für jedes beste Modell wird die Modellgröße sowie der C_p-Wert erfasst.
Im Anschluss an die Somulation erfolgt die Berechnung des erwarteten Prognosefehler, welche in Methode *calculateTrueSpse* nach Formel (1) erfolgt. Die Schätzung des erwarteten Prognosefehlers aus den CP Werten der Simulation erfolgt in Methode *calculateEstimatedSpse* nach der Formel

$$\hat{\mathrm{SPSE}}^{(\mathrm{M})} := (C_p^{(\mathrm{M})} + (|M| + 1))\tilde{\sigma}^2_{full}$$

5 Ergebnisse und Diskussion

5.1 Modellwahl

Unter Anwendung der in Abschnitt 3.4 bzw 4.1 vorgestellten Methode wurde aus einem 149 Prädiktoren umfassenden Maximalmodell das in Tabelle 2 angegebene, 50 Prädiktoren große Modell anhand des kleinsten C_p-Wertes von ≈ 3.92 als optimales Modell ausgewählt.

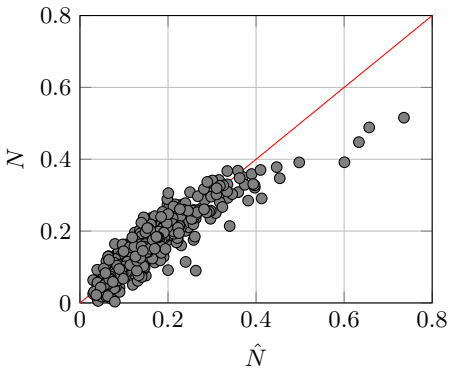

Abbildung 2: Korrelationsplot

Der Wert des Bestimmtheitsmaßes $R^2 = 0.82$ zeigt, dass das gewählte Modell die Varianz in den Daten gut erklären kann. Dies spiegelt auch das in Abbildung 2 abgebildete Korrelationsdiagramm wieder, in welchem der wahre Wert auf der x-Achse und der geschätzte Stickstoffmengenanteil auf der y-Achse aufgetragen sind. Abweichungen sind lediglich im oberen Bereich ab 0.3 zu beobachten, wo der Stoffmen-

n	$\hat{SPSE}$
150	0.4964226
200	0.4312206
250	0.521319
300	0.6097658
350	0.7058959
400	0.7947748
450	0.8807918
500	0.9750933
533	1.028092

Tabelle 1: Über Mallow's C_p geschätzte SPSE Werte mit zugehörigen n

genanteil des Stickstoffs durch das Modell unterschätzt wird. Dies lässt dich vermutlich durch die geringe Datendichte in diesen Bereichen erklären. Insgesamt kann man sagen, dass die bisherigen Untersuchungen auf ausreichend gute Prognosen durch das Modell hindeuten.

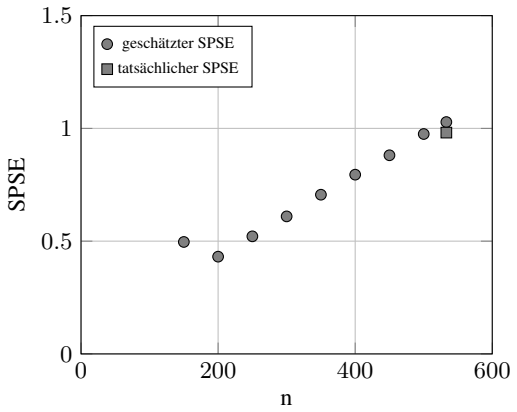

Abbildung 3: geschätzer sowie tatsächlicher SPSE in Abhängigkeit der Stichprobengröße

5.2 Simulation

Die vorliegende Arbeit sollte den Einfluss des Stichprobenumfangs auf die SPSE-Schätzung mithilfe Mallow's C_p untersuchen. Die hier angegebenen Werte spiegeln den Mittelwert aus 10 Durchläufen wieder.

Wie in Abbildung 3 zu erkennen ist steigt der geschätzte SPSE-Wert außer bei $n < 200$ monoton und nahezu linear mit zunehmender Stichprobengröße. Dies lässt auf einen starken linearen Einfluss von n auf den SPSE schließen. In der Berechnung des SPSE ist n sowohl als Faktor im unkontrollierten Prognosefehler als auch im Biasterm vorhanden, was den Einfluss auf das Ergebnis bereits vermuten ließ. Der tatsächliche SPSE von 0,981 wird bei der Schätzung über Mallow's C_p mit derselben Stichprobengröße (533) um knapp 5% überschätzt und liegt bei 1,028. Daraus lässt sich schließen, dass der Biasterm durch die zufällige Simulation wohl leicht überschätzt wurde. Wir können also schlussfolgern, dass SPSE Werte, die auf verschiedenen Stichprobengrößen beruhen, normiert werden müssen, um miteinander vergleichbar zu sein.

6 Fazit

In der vorliegenden Arbeit wurde mit Hilfe von Mallow's C_{p}-Kriterium eine Prognosefunktion für Stoffmengenanteile von Stickstoff durch die Messung von Reflexionswerte verschiedener Wellenlängen im Nahinfrarotbereich hergeleitet. Hierbei wurde zunächst eine Vorauswahl relevanter Wellenlängen nach verschiedenen Kriterien getroffen und jenes Kriterium ausgewählt welches, bei weiterer Featureselektion zu dem kleinsten $SPSE$ in Kombination mit einem minimalen C_p des Modells führte. Es zeigte sich, dass die Variabilität der differenzierten Reflexionswerte ein geeignetes Kriterium der Selektion darstellt.

Um die tatsächliche Eignung des gewählten Modells zu überprüfen wurde zunächst das Bestimmtheitsmaß R^2 untersucht. Zur Evaluation der Schätzung des erwarteten Prognosefehlers $SPSE$ wurde darüber hinaus Pseudobeobachtungen durch das gewählte Modell generiert. Die Schätzwerte des erwarteten Prognosefehlers verdeutlichen den starken Einfluss des Stichprobenumfangs auf dessen Genauigkeit. Die Ergebnisse der Untersuchungen zeigen außerdem die Notwendigkeit einer sorgfältigen Auswahl der Einflussgrößen für das Maximalmodell. In diesem Zusammenhang könnten in Zukunft auch andere Vorgehensweisen für die Selektion der Einflussgrößen getestet werden. Denkbar wäre beispielsweise eine Formulierung des Problems als Minimierungsproblem.[Men16]

Insgesamt lässt sich zusammenfassen, dass mit dem gewählten Vorgehen eine geeignete Prognosefunktion bzw. Kalibrierung gefunden wurde.

Literatur

[AHJ10] Lidia Esteve Agelet and Charles R Hurburgh Jr. A tutorial on near infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry*, 40(4):246–260, 2010.

[Lan07] Stefan Lang. Regression, modelle, methoden und anwendungen, 2007.

[Men16] Pawellek M Menzel, K. Statistical methods: Prediction of soil parameters through near infrared spectroscopy. 2016.

[PDD+13] C Poeplau, A Don, M Dondini, J Leifeld, R Nemo, J Schumacher, N Senapati, and M Wiesmei-

er. Reproducibility of a soil organic carbon fractionation method to derive rothc carbon pools. *European Journal of Soil Science*, 64(6):735–746, 2013.

[Sch19] Jens Schumacher. Skript zur vorlesung statistische verfahren im wintersemester 2018/19, 2019.

[Umw17] Umweltbundesamt. Stickstoff, 2017.

A Einflussgrößen und Parameter des Modells

x_i	β_i	x_i	β_i	x_i	β_i	x_i	β_i	x_i	β_i
nm1492	-33.036	nm2144	308.476	nm2288	364.667	nm2452	34.662	nm2588	83.166
nm1508	-41.247	nm2148	-205.012	nm2292	-288.261	nm2480	90.572	nm2600	-101.046
nm1524	59.924	nm2168	91.553	nm2296	182.079	nm2488	-110.665	nm2620	63.035
nm1532	46.614	nm2172	103.26	nm2372	43.232	nm2496	132.982	nm2644	-81.213
nm1556	48.434	nm2188	60.423	nm2392	-118.104	nm2508	-372.868	nm2652	-54.94
nm1564	-99.414	nm2204	-117.815	nm2396	147.093	nm2512	403.274	nm2668	100.931
nm1568	62.567	nm2228	69.33	nm2404	166.541	nm2516	-318.052		
nm1580	-43.808	nm2248	91.69	nm2408	-209.319	nm2520	209.089		
nm2128	-105.035	nm2276	-321.278	nm2424	-106.745	nm2532	51.908		
nm2136	157.86	nm2280	328.749	nm2428	215.418	nm2544	-202.54		

Tabelle 2: Enthaltene Features und geschätzte Parameter β des ausgewählten Modells

B R Source Code

```
# ##################################################
# File: modelselect.R
# Scope: functions used for initial model selection
# ##################################################

getSlope = function(data, col_from, col_to) {
  slope = matrix(nrow = nrow(data), ncol=col_to-col_from)
  for (i in 1:(col_to-col_from)) {
    slope[,i] = data[,col_from+i] - data[,col_from+i-1]
  }

  return(slope)
}

# find wavelengths with highest variablilty as the distance of max and min slope
criterionSlopeDist = function(data, col_from, col_to) {
  slope = getSlope(data, col_from, col_to)
  min = apply(slope, 2, min)
  max = apply(slope, 2, max)
  dist = max - min

  return (dist)
}

selectFeatures = function(data, feat.all, th, crit) {
  df = data.frame(
    WAVELENGTHS[1:length(WAVELENGTHS)-1],
    colnames(data[, COLUMN_ID_NM_FROM:(COLUMN_ID_NM_TO-1)]),
    crit
  )
  colnames(df) = c("x", "nm", "y")
  feat.select = subset(df, y >= th, select=c("x", "nm", "y"));

  return (feat.select)
}
```

```
# ################################################
# File: simulation.R
# Scope: functions used for simulation
# ################################################
source("helper.R")

getSDs = function(designMatrix, means, data.origin) {
  residuals = as.vector(data.origin$N - means)
  sd = sqrt(as.numeric(t(residuals) %*% residuals / (dim(data.origin)[1] - dim(designMatrix)[2])))

  return(sd)
}

simulate = function(model, data, num, times) {
  simdata = data.frame(matrix(nrow=num, ncol=times));

  coeffs = as.vector(coef(model))
  designmatrix = model.matrix(model, data=data)

  means = designmatrix %*% coeffs
  sds = getSDs(designmatrix, means, data)
  for (i in 1:times) {
    simdata[i] = rnorm(num, mean=means, sd = sds)
  }

  return(rowMeans(simdata))
}

getDataSubset = function(data.origin, numRows) {
  if (numRows < nrow(data.origin)) {
    return(nirs.data[sample(nrow(data.origin), numRows), ])
  }

  return(nirs.data)
}

simulateOnDataSubset = function(model, data, fmForm, fmFeat, numRows, times) {
  simdata = getDataSubset(data, numRows)
  simdata = subset(simdata, select=c(c("SOC", "N", "pH"), as.character(fmFeat[,"nm"])))
  simdata$N = simulate(model, simdata, numRows, times)

  simsets = regsubsets( fmForm, data=simdata, really.big=T, nvmax=nrow(fmFeat)+1, method="backward")
  simOptModelId = which.min(summary(simsets)$cp)
  simOptModelCpValue = min(summary(simsets)$cp)

  simResult = c(simOptModelId, simOptModelCpValue)
  names(simResult) = c("id", "cp")

  return(simResult)
}

calculateR2 = function(model, data, num, times){
  y_exp = simulate(model, data, num, times)
  y_mean = sum(data[,2])/dim(data)[1]
  y_obs = c(data[,2])
  var_exp = y_exp - y_mean
  var_true = y_obs - y_mean
  R2 = (t(var_exp) %*% var_exp) / (t(var_true) %*% var_true)

  return(R2)
}
```

```
# ################################################
# File: helper.R
# Scope: convinience functions
# ################################################

buildFormula = function(features, colName, respVar) {
  featString = paste(features[,colName], collapse="+")
  featString = paste(1, featString, sep="+")
  formulaString = paste(respVar, featString, sep = " ~ ")

  form = as.formula(formulaString)

  return (form)
}

addSelectedFeatureToPlot = function(val, color) {
  abline(v=val, col=color)
}

getModelFromSubsets = function(subsets, modelId, responsevar, data) {
  X <- summary(subsets)$which
  xvars <- dimnames(X)[[2]][-1]
  id <- X[modelId,]
  form <- reformulate(xvars[which(id[-1])], responsevar, id[1])
  lm <- lm(form, data)

  return(lm)
}
```

```
# ################################################
# File: spse.R
# Scope: functions to calculate SPSE
# ################################################

calculateRss = function(model) {
  rss = sum(residuals(model)^2)
}

calcauletSigma2TildeFull = function(fullModel, fullModelSize, sampleSize) {
  sigma2.tilde.full = calculateRss(fullModel) / (sampleSize - fullModelSize)

  return(sigma2.tilde.full)
}

calculateTrueSpse = function(model, modelSize, fullModel, sampleSize, fullModelSize) {
  sigma2.tilde.full = calcauletSigma2TildeFull(fullModel, fullModelSize, sampleSize)
  spse = calculateRss(model) + 2*sigma2.tilde.full*modelSize

  return(spse)
}

calculateEstimatedSpse = function(cp, fullModel, fullModelSize, sampleSize) {
  sigma_2_tilde_full = calcauletSigma2TildeFull(fullModel, fullModelSize, sampleSize)
  spse = cp * sigma_2_tilde_full + sampleSize * sigma_2_tilde_full

  return(spse)
}
```

```
# ################################################
# File: main.R
# Scope: main R Script
# ################################################

# ###############
# DEPENDENCIES
# ###############
require(leaps)
source("helper.R")
source("modelselect.R")
source("simulation.R")
source("spse.R")

# ###############
# CONFIGURATION
# ###############
nirs.data = read.csv("../NIR.csv", sep=";")

set.seed(13)

# general
COLUMN_ID_N = 2
COLUMN_ID_NM_FROM = 4
COLUMN_ID_NM_TO = ncol(nirs.data)
NUM_ROWS = nrow(nirs.data)
WAVELENGTHS = seq(1400, 2672, 4)
RESPONSEVAR = "N"

# modelselection
THRESHOLD = 0.001

# simulation
SAMPLE_SIZES = c(150, 200, 250, 300, 350, 400, 450, 500, 533)
SIM_ITERATIONS = 10


# ###############################
# 1. MODELSELECTION
# ###############################
nirs.criterion.slopedist = criterionSlopeDist(nirs.data, COLUMN_ID_NM_FROM, COLUMN_ID_NM_TO)
selectedCriterion = nirs.criterion.slopedist

# select wavelengths with highes variablilty
plot(WAVELENGTHS[1:length(WAVELENGTHS)-1], selectedCriterion, type = "l", col=1)
abline(h=THRESHOLD, col=2)
fullModel.features = selectFeatures(nirs.data, WAVELENGTHS, THRESHOLD, selectedCriterion)

# build model from selected wavelength
fullModel.formula = buildFormula(fullModel.features, "nm", RESPONSEVAR)

# modelselection based on Mallow's CP from full model
nirs.subsets = regsubsets(
    fullModel.formula, nirs.data, really.big=T, nvmax=nrow(fullModel.features)+1, method="backward"
)
nirs.lm.full = lm(fullModel.formula, data=nirs.data)
nirs.cp = min(summary(nirs.subsets)$cp)
nirs.optModelId = which.min(summary(nirs.subsets)$cp)
nirs.lm.opt = getModelFromSubsets(nirs.subsets, nirs.optModelId, "N", nirs.data)
nirs.spse.true = calculateTrueSpse(nirs.lm.opt, nirs.optModelId, NUM_ROWS)
nirs.spse.true

# plot slope with selected features
plot(WAVELENGTHS[1:length(WAVELENGTHS)-1], selectedCriterion, type = "l", col=1)
selfeat = as.integer(substr(names(nirs.lm.opt$coefficients)[-1],3,6))
sapply(selfeat, addSelectedFeatureToPlot, 3)
```

```
# ###############################
# 2. SIMULATION
# ###############################
R2 = calculateR2(nirs.lm.opt, nirs.data, NUM_ROWS, 1000)
R2

sim.models = data.frame(matrix(nrow=length(SAMPLE_SIZES), ncol=SIM_ITERATIONS))
sim.cp = data.frame(matrix(nrow=length(SAMPLE_SIZES), ncol=SIM_ITERATIONS))
for (run in 1:SIM_ITERATIONS) {
  for (i in 1:length(SAMPLE_SIZES)) {
    simResult = simulateOnDataSubset(
        nirs.lm.opt, nirs.data, fullModel.formula, fullModel.features, SAMPLE_SIZES[i], 1000
    )
    sim.models[i, run] = simResult["id"]
    sim.cp[i, run] = simResult["cp"]
  }
  print(paste(run, "/", SIM_ITERATIONS, sep=""))
}

# calculate SPSEs and compare
sigma2.tilde.full = calcauletSigma2TildeFull(nirs.lm.full, nrow(fullModel.features)+1, NUM_ROWS)

spse_true = calculateTrueSpse(
    nirs.lm.opt, nirs.optModelId, nirs.lm.full, NUM_ROWS, nrow(fullModel.features)+1
)

sim.spse.est = (sim.cp + SAMPLE_SIZES) * sigma2.tilde.full
plot(x=150, y=mean(as.vector(t(sim.spse.est[1,]))), xlim=c(0,550), ylim=c(0,2), pch=16, col=1)
for (i in 2:length(SAMPLE_SIZES)) {
  points(x=SAMPLE_SIZES[i], y=mean(as.vector(t(sim.spse.est[i,]))), pch=16, col=1)
}
points(x=NUM_ROWS, y=spse_true, col=2, pch=16)

# ############################
# ADDITIONAL PLOTS
# ############################
# 1. correlation plot
simN = simulate(nirs.lm.opt, nirs.data, 533, 1000)
plot(nirs.data$N, simN, xlim=c(0,0.8), ylim=c(0,0.8))
abline(a=0,b=1, col=2)
```

Eigenständigkeitserklärung

Hiermit bestätigen wir, dass wir die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet haben. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen) entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

Jena, 01. April 2019

Ferdinand Rewicki Moritz Preuß Tobias Giesemann