Statistische Verfahren:
Projekt 4 - Nahinfrarotspektroskopie I

Tobias Giesemann
Ferdinand Rewicki
Moritz Preuß

March 28, 2019

Abstract

We describe and assess ...

Contents

Statistische Verfahren:
Projekt 4 - Nahinfrarotspektroskopie I

Tobias Giesemann
tobias.giesemann@uni-jena.de

Ferdinand Rewicki
...@gmail.com

Moritz Preuß
moleo.preuss@gmail.com

Abstract

We describe and assess ...

1 Einleitung

Obtaining ...

2 Background

2.1 Soil Parameters

Let A be any substance in a given soil sample dissolved in a solution of volume $V \in (0, \infty)$ and let $n_A \in (0, \infty)$ be the amount of A in the sample. Then the molar concentration c_A of A is given by

$$c_A := \frac{n_A}{V}$$

Now let c_0 be the molar concentration of this whole sample and n_0 the amount of the whole sample....

2.2 Near Infrared Spectroscopy

NIRS uses electromagnetic waves [**?** , 246],

The reflectance

$$\varrho\colon (0,\infty) \to (0,\infty), \qquad \varrho(\lambda) := \frac{P_\mathrm{r}(\lambda)}{P_0}$$

of a

3 Methodology

3.1 Multivariate lineare Regression

Ein lineares Modell (LM) zur Bestimmung der Abhängigkeit einer Zielgröße von mehreren Einflussvariablen wird multivariate lineare Regression genannt [Wei05]. Ein typisches LM hat die Form

$$E(Y|\mathbb{X}) = \beta_0 + \beta_1\mathbb{X}_1 + \beta_2\mathbb{X}_2 + ... + \beta_k\mathbb{X}_k, \qquad (1)$$

wobei $Y \in \mathbb{R}^\mathrm{n}$ die Zielgröße und $X \in \mathbb{R}^{\mathrm{n}\times(\mathrm{k}+1)}$ die Matrix von d Einflussparametern und n Beobachtungen (Designmatrix) sind. Y wird als normalverteilter Zufallsparameter mit $Y \sim N(X\beta, \sigma^2)$ angenommen. Lineare Modelle sind nun eine Methode, die Einflussstärke und Richtung der Variablen in der Designmatrix auf die Zufallsgröße Y mithilfe der Maximum-Likelihood Methode zu bestimmen:

$$\hat{\beta}(Y) = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y \qquad (2)$$

Des Weiteren ergibt sich ein Schätzer für die Varianz σ^2 des Modells:

$$\hat{\sigma^2}(Y) = \frac{1}{n-(k+1)}||Y - \mathbb{X}\hat{\beta}(Y)||^2 \qquad (3)$$

Für eine Realisierung von $y := (y_i) \in \mathbb{R}^\mathrm{n}$ von Y definieren wir [Sch19]:

$$\hat{y} := \mathbb{X}\hat{\beta}(y) = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y \qquad (4)$$

$$\tilde{\sigma^2} := \hat{\sigma}^2(y) \qquad (5)$$

3.2 Modellwahl im Falle der NIR-Spektroskopie

Seien $Y \in \mathbb{R}^\mathrm{n}$ die Zielgröße in einem statistischen Modellwahlverfahrens 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} 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:

$$\frac{\hat{\beta_k}}{\sqrt{\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}_{kk}}} \sim t_{n-(k+1)}. \qquad (6)$$

Dieses Verfahren ist vor allem dann besonders gut geeignet, wenn man bereits theoretisch fundierte Annahmen über die

Einflussgrößen machen kann. Mit diesem Modellwahlverfahren ergeben sich hier allerdings einige Schwierigkeiten, wobei die für unseren Fall besonders schwerwiegenden herausgehoben werden: In dieser Arbeit haben wir es mit einer großen Anzahl potenzieller Einflussvariablen auf dem Nah-Infrarotspektrum zu tun. A priori kann schwer eine inhaltliche Deutung vorgenommen werden, die gewisse Wellenlängen bevorzugt. Daher ist eine hierarchische Modellwahl mit wenigen, theoretisch begründeten Einflussvariablen 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, wird die Wahrscheinlichkeit bei oft wiederholten Tests fälschlicherweise die Nullhypothese abzulehnen mit Anzahl der Versuchen immer größer. An ein automatisiertes Modellwahlverfahren, das in dieser Arbeit von Vorteil ist, ist also mittels des T-Tests nicht zu erreichen [Sch19]. Stattdessen bietet sich eine Modellwahl basierend auf dem erwarteten Prognosefehler ("sum of prediction squared error", SPSE) an:

$$SPSE := E(\sum_{i=1}^{n}(Y_{i+n} - x_i^{(M)}\hat{\beta}_i{}^{((M)})^2) \tag{7}$$

Hierbei sind die Werte in Y_{i+n} neue Beobachtungen zum Erwartungsvektor x_i und $x_i^{(M)}\hat{\beta}_i{}^{((M)}$ ist sind 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:

$$SPSE^{(M)} = n\sigma_{full}^2 + p\sigma_{full}^2 + (bias^{(M)})^2 \tag{8}$$

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

$$RSS^{(M)} := \sum_{i=1}^{n} E(Y_i - \hat{Y}_i{}^{(M)})^2 \tag{9}$$

$$SPSE^{(M)} := \sum_{i=1}^{n} E(Y_{i+n} - \hat{Y}_i{}^{(M)})^2 \tag{10}$$

Es kann gezeigt werden, dass RSS den Wert von SPSE systematisch unterschätzt, dass diese Unterschätzung jedoch behoben werden kann, indem für alle Modelle die Varianzschätzung aus dem maximalen Modell verwendet wird[Sch19]:

$$SPSE^{(M)} := RSS^{(M)} + 2\tilde{\sigma}_{full}^2(k+1) \tag{11}$$

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

$$Cp^{(M)} = \frac{1}{\sigma_{full}^2}\sum_{i=1}^{n}(y_i - \hat{y_i}{}^{(M)})^2 - n + 2(k+1) \tag{12}$$

3.3 Modellselektion

Das erste Ziel dieser Arbeit ist es, diejenigen Nahinfrarot-Wellenlängen herauszufinden, die einen Einfluss auf den Bodenstickstoffgehalt haben. Für diese Aufgabe nahmen wir zunächst die erste Ableitung der Designmatrix, um eine bessere Einsicht in die Daten zu bekommen(siehe **??**). Für unsere Analyse wurden diejenigen Wellenlängen in das maximale Modell aufgenommen, deren durchschnittliche Ableitungswerte über dem mit der roten Linie angezeigten Schwellwert lagen. Über dieses maximale Modell wurde nun mittels Mallows Cp ein bestes Modell(3.2) berechnet.

3.4 title

4 Implementation

4.1 Choosing a Neighbour

We stated in section **??** that we want to select a "good" model for the prediction.

4.2 Additional Functions

All other functions were defined following a standard scheme. It follows from **??** that

$$\mathrm{cost}(M) := C_{\mathrm{p}}^{(M)}$$

4.3 Preprocessing

Implementing the algorithm described in **??** takes a sizeable toll on computing power. The most expensive calculations are performed in the computation of the residual sum of squares

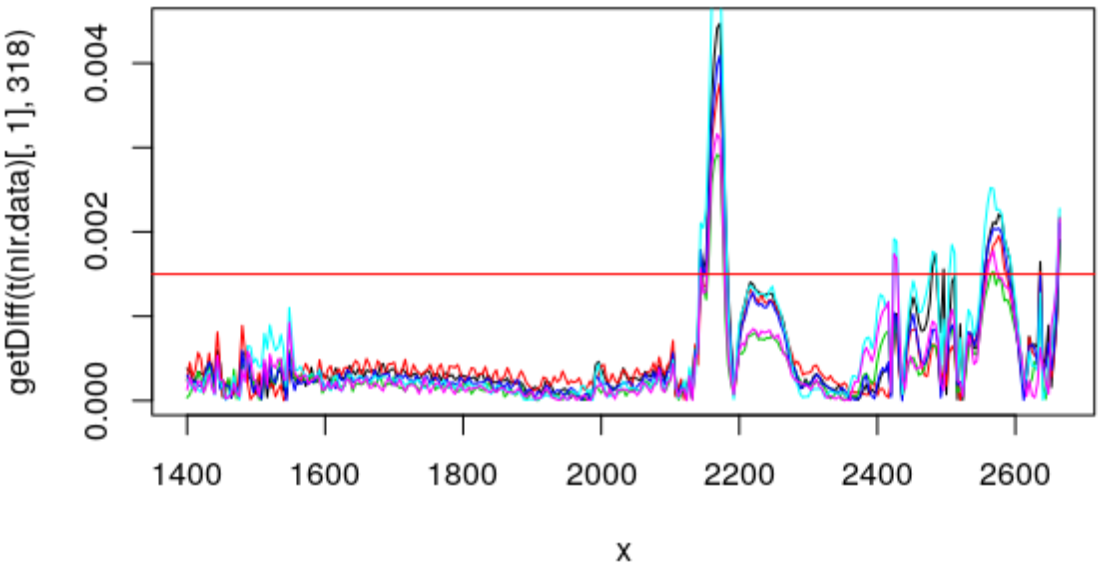
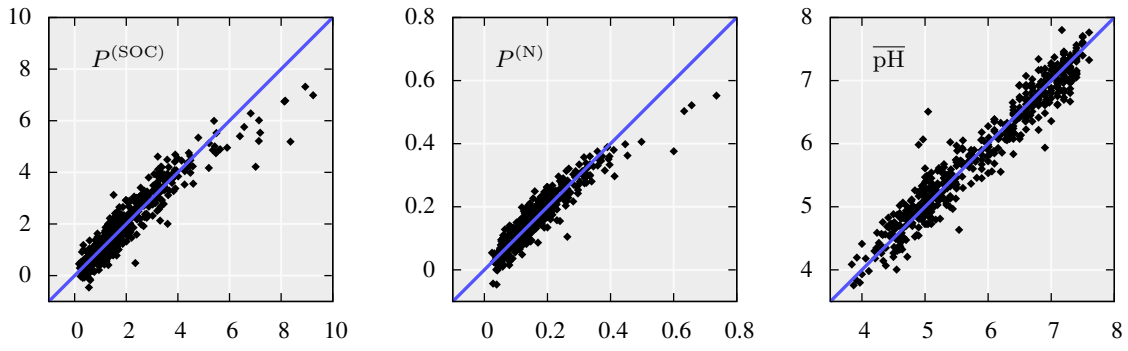


Figure 1: Ableitungen der Wellenlängen von fünf zufälligen Spektren



(a) $\hat{p}^{(\mathrm{SOC})} \sim p^{(\mathrm{SOC})}$ $(R^2)^{(\mathrm{SOC})} = 0.871$

(b) $\hat{p}^{(\mathrm{N})} \sim p^{(\mathrm{N})}$ $(R^2)^{(\mathrm{N})} = 0.861$

(c) $\widehat{\mathrm{pH}} \sim \mathrm{pH}$ $(R^2)^{(\mathrm{pH})} = 0.940$

Figure 2: Correlation diagrams plotting $\hat{y}$ on y and the blue line representing the id

5 Calibration

5.1 Model Selection

SANN returned as minimum values for the respective models

$$C_{\mathrm{p}}^{(\mathrm{SOC})} = -17.46$$
$$C_{\mathrm{p}}^{(\mathrm{N})} = -28.51$$
$$C_{\mathrm{p}}^{(\mathrm{pH})} = -57.76$$

Figures 3a, 3b and 3c in appendix A show

5.2 Goodness of Fit

6 Simulation - Schätzgenauigkeit von Mallow's Cp

Die hier vorgestellte Untersuchung soll zeigen, ob und inwiefern sich Mallow‘s Cp als Schätzwert für den SPSE an diesen annähert, wenn sich der Umfang der Stichprobe verändert. Genauer gesagt ziehen wir ein Sample von oben erstelltem Modell, zu dem die exakte Berechnung des SPSE für alle von uns gewählten Stichprobengrößen (z.B. n = [5, 10, 25, 50, 100, 200, 500]) möglichist. Der ermittelte SPSE Wert dient uns im späteren Vergleich als Ground Truth. Für jeden Stichprobenumfang ermitteln wir zudem den Wert von Mallow‘s Cp, der ein Schätzwert des ursprünglichen SPSE darstellt. Ermittelt werden soll dabei, inwiefern sich der Umfang der Stichprobe auf die Abweichung des Schätzers von der Ground Truth (SPSE) auswirkt.

7 Conclusion

Using Mallow's C_{p} criterion, we calibrated three predictive models for the soil parameters $p^{(\mathrm{SOC})}$, $p^{(\mathrm{N})}$ and pH. To construct ...

References

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

[Wei05] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

A Prediction Parameters and Models

Table 1: Estimated model parameters of $P^{(\mathrm{SOC})}$ on selected model

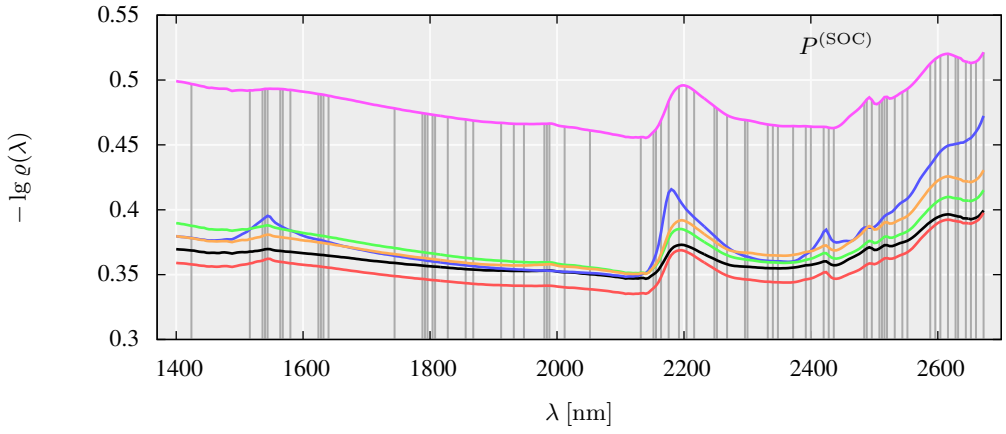
λ_i [nm]	$\beta_i^{(\mathrm{SOC})}$	λ_i [nm]	$\beta_i^{(\mathrm{SOC})}$	λ_i [nm]	$\beta_i^{(\mathrm{SOC})}$	λ_i [nm]	$\beta_i^{(\mathrm{SOC})}$
—	-1.47103	1808	1991.63	2204	-2319.71	2496	1956.13
1424	-811.326	1828	1568.91	2216	1075.21	2508	-5057.56

Table 2: Estimated model parameters of $\overline{\mathrm{pH}}$ on selected model

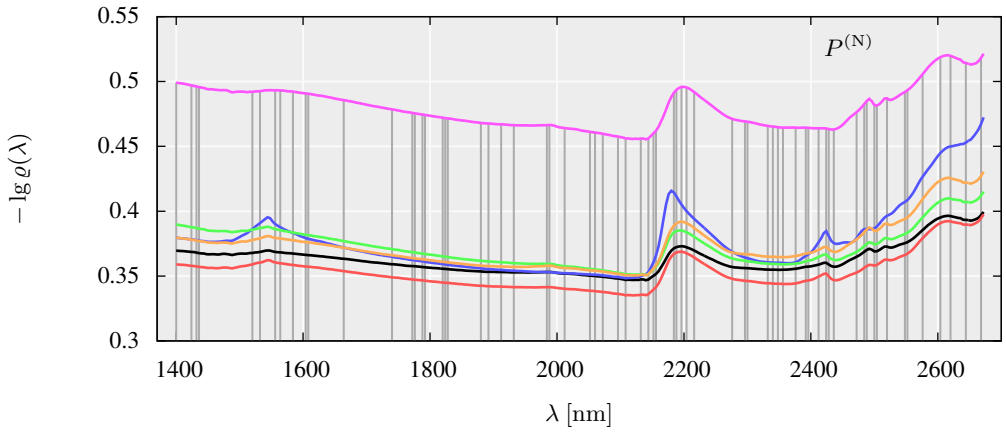
λ_i [nm]	$\beta_i^{(\mathrm{pH})}$	λ_i [nm]	$\beta_i^{(\mathrm{pH})}$	λ_i [nm]	$\beta_i^{(\mathrm{pH})}$	λ_i [nm]	$\beta_i^{(\mathrm{pH})}$
—	5.57628	1864	-508.298	2220	699.185	2460	545.634
1436	135.244	1896	-623.655	2224	-659.264	2464	-519.611

Table 3: Estimated model parameters of $P^{(\mathrm{N})}$ on selected model

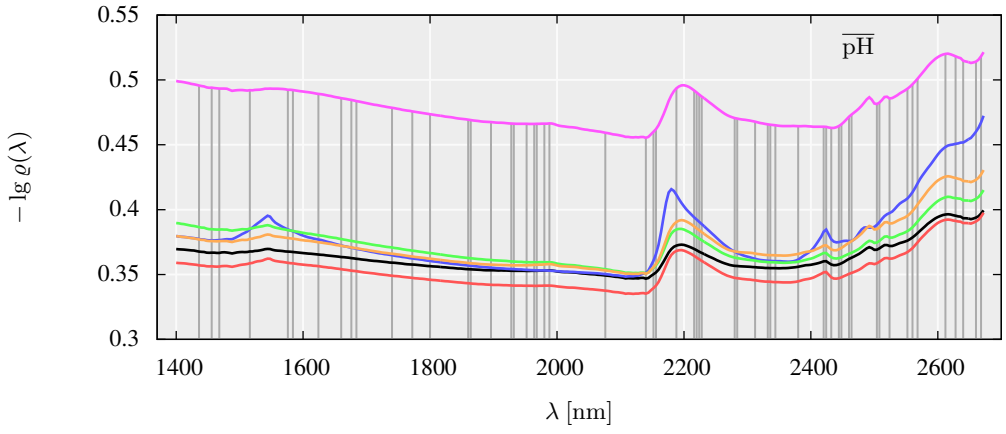
λ_i [nm]	$\beta_i^{(\mathrm{N})}$	λ_i [nm]	$\beta_i^{(\mathrm{N})}$	λ_i [nm]	$\beta_i^{(\mathrm{N})}$	λ_i [nm]	$\beta_i^{(\mathrm{N})}$
—	-0.0287506	1820	169.949	2156	95.2657	2428	116.231
1400	48.2214	1824	-272.304	2184	-99.54	2436	-60.6976



(a) $p^{(\mathrm{SOC})}$ with $\mathrm{p}^{(\mathrm{SOC})} = 72$



(b) $p^{(\mathrm{N})}$ with $\mathrm{p}^{(\mathrm{N})} = 69$



(c) pH with $\mathrm{p}^{(\mathrm{pH})} = 58$

Figure 3: Displaying the spectra from figure **??** with wavelength included in the selected models for each response highlighted by vertical grey lines

B R Source Code

Statutory Declaration

We herewith declare that ...

Jena, 25th of August 2016

Kazimir Menzel Markus Pawellek