

Statistische Verfahren SS 2016

Projekt 4 - Nahinfrarotspektroskopie

Problemstellung:

Herleitung von Kalibrierungsfunktionen zur Schätzung des Bodenkohlenstoffgehalts mit Hilfe der Nahinfrarotspektroskopie

Datensatz: (NIR.csv, Dr. A. Don, Heinrich von Thünen-Institut Braunschweig)

Der Datensatz enthält Daten zum Bodenkohlenstoffgehalt (SOC), Bodennickstoffgehalt (N) und dem pH-Wert sowie die dazugehörigen Nahinfrarotspektren, d.h. die gemessenen Reflexionswerte bei verschiedenen Wellenlängen aus dem Nahinfrarotbereich, für 534 Bodenproben. Diese Daten sollen zur Herleitung von Prognosefunktionen verwendet werden, die die Schätzung der nur aufwändig bestimmbarer Bodenparameter SOC, N und pH aus den viel leichter messbaren Nahinfrarotspektren gestattet.

Aufgaben:

- Bestimmen Sie mit Hilfe eines geeigneten Modellselektionsverfahrens Vorhersagemodelle für den Kohlenstoffgehalt, den Stickstoffgehalt und den pH-Wert.

Simulationsaufgabe:

- Die große Anzahl von Wellenlängen führt zu einer sehr großen Anzahl von Kandidatenmodellen. Bei der Modellwahl basierend auf Mallows C_p -Kriterium ist daher zu erwarten, dass der tatsächliche erwartete Vorhersagefehler SPSE unterschätzt wird.
Wählen Sie eines der im ersten Teil hergeleiteten Prognosemodelle als Grundlage zur Simulation von Pseudobeobachtungen. Entscheiden Sie sich für jeden simulierten Pseudo-Datensatz basierend auf Mallows C_p -Kriterium für ein bestes Modell. Vergleichen Sie den aus dem minimalen C_p -Wert geschätzten Vorhersagefehler mit dem tatsächlichen, theoretisch erwarteten Vorhersagefehler SPSE.
Untersuchen Sie dabei den Einfluss des Stichprobenumfangs (indem Sie aus dem Originaldatensatz zufällig Spektren für die Designmatrix der Simulation auswählen) sowie den Einfluss der Anzahl der potentiellen Prädiktoren (indem Sie die Anzahl der verwendeten Wellenlängen in der Designmatrix reduzieren)