Statistik-Projekt: Bodenatmung

Wir gehen wie folgt vor:

Erstellung des Modells

- Der Datensatz enthält sehr viele Features im Vergleich zur Anzahl an Messungen. Der Suchraum der Variabelnselektion wäre viel zu groß.
- **Vorauswahl.** Es werden nur stark korrelierende Variablen in Betracht gezogen
- Einflüsse wie z.B. die der *Temperatur* sind nicht linear. **Link-Funktion.** Ins lineare Modell wird z.B. die Transformierte $log(Temp)$ genommen.
- **Kopplungen.** Zur Vereinfachung werden lediglich die Kopplungen zwischen $Temp_i$ und $Temp_j$ sowie zwischen $Temp$ und $moisture$ betrachtet.
- **Modellqualität.** Genommen wird das *kleinste* Modell, welches einen $SPSE < 0.05 * E(soil.res)$ hat. Hierbei wird das Modell auf *Trainingsdaten* ($\approx 80\%$) der Daten gelernt und auf Testdatensatz miteks $SPSE$ evaluiert. Diese Untermengen des Datensatzen bilden eine *Partition*.
- Um Overfitting entgegenzutreten: **Variabelnselektion.** Mit Hilfe des R-Pakets `leaps` wird das Modell mit dem geringsten BIC ausgewählt, welches das Kriterium der Modellqualiät erfüllt.

Simulation

- Der Datensatz ist zu klein, sodass es keinen Sinn ergibt, alle Features ins Modell aufzunehmen. *Sparse linear models* sind gefragt.
- viele Variabeln des Datensatzes sind *statistisch abhänig*. Dadurch sind die Maxima der F-Statistiken nicht mehr F-verteilt. So kann der Fehler entstehen, dass eine Variabel *fälschlicherweise* doch zum Modell hinzugenommen wird, obwohl es objektive Kriteria gegenüber $(SPSE, BIC, ...)$ das Modell *nicht* besser amcht
- **Simulation.**
- Angenommen, Modell

$$E(soil.res) = \beta_0 + smoi * \beta_1 + temp10 * \beta_2$$

  sei gegeben. Dieses "wahre"Modell ist das Ergebnis des vorherigen Prozesses.
- Nun wird im Rahmen der *forward selection* geprüft, ob es Sinn ergibt, die zusätzliche Variable $rootdw$ hinzuzunehmen. Derartige Verfahren verwenden die *F-Statistik* als Prüfgröße:
- Sei Modell 1 das "wahre" Ausgangsmodell und Modell 2 das um $rootdw$ erweiteret Modell von 1. RSS ist der summierte, quadratische Fehler im

Bezug auf die Prädikation des Modells einer Zeile der *disjunkten* Test-Daten. Ferner sei p_i die Anzahl an Features des Modells i. Dann ist:

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}}$$

- Für jeden Test-Datensatz gibt es somit pro Schritt der Selektion und pro zusätzlicher Variable einen F-Wert. Die *Verteilung* dieser F-Statistiken für die unterschiedlichen Test-Messungen wird nun betrachtet.
- Ferner wird nicht nur der F-Wert, sondern auch der von stat. Abhänigkeiten *unabhänige* Wert $SPSE$ betrachtet.
- Die *beste Auswahl* ist die mit der stärksten Abnahme des unabhänigen Kriteriums $SPSE$. Ausgewählt wird allerdings ausslichlich nach maximalem F-Wert. Demnach wird immer bei $argmax(F) \neq argmax(-\Delta SPSE)$ ein Fehler begangen. Der relative Anteil der Fehlentscheidungen in der Test-Simulation ergibt eine Schätzung dafür, wie häufig die Nullhypothese $H_0 : \beta_i = 0$ fälschlicherweise entschieden wird.