Bodenatmung im Nationalpark Hainich

Erstellung eines statistischen Modells und Diskussion von Fehlentscheidungen im Rahmen der Variabelnselektion

Sichtling C Loos D Jajali R

15. August 2016

PROJEKTARBEIT

im Rahmen des Moduls *Statistische Verfahren*
an der Friedrich-Schiller-Universität Jena

Die Freisetzung von CO_2 in die Atmosphäre ist ein fundamentaler Bestandteil von Modellen zur Berechnung klimatischer Phänomene. Die *Bodenatmung* ist hierbei der dominierende Prozess auf Landgebieten. Zur Modellierung dieses Prozesses wurde im Nationalpark Hainich Messdaten im Einfluss von ca. 30 Variabeln erhoben. Daraus wird nun ein statistisches Modell erstellt. Bei der Variabelnselektion können statistische Unsicherheiten zu Fehlern führen. Dies wird im folgendem diskutiert.

Inhaltsverzeichnis

1 Einleitung

1.1 Bodenatmung als geophysikalischer Prozess

[1]

1.2 Statistische Grundlagen

1.2.1 Korrelation

Unter *Korrelation* versteht man die „Wechselbeziehung“ zweier Zufallsgrößen. Dieser Zusammenhang kann entweder *linear* (Korrelationskoeffizient nach Pearson) oder lediglich *monoton* sein. Ist eine Korrelation nicht gegeben, so scheint die Zufallsgröße als ungeeigneter Prädiktor für die jeweils andere Variabel. Kausale Beziehungen erfordern Korrelation.
Der *lineare, empirische Korrelationskoeffizient nach Person* zwischen den Variabeln X und Y wird definiert durch:

$$Kor_e(x, y) := \varrho_e(x, y) := r_{xy} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

mit den empirischen Mittelwert $\bar{x}$ aus n Messwerten von X. Möchte man lediglich ein *lineares* Modell erstellen, so ist die Pearson-Korrelation zu wählen.
Eine allgemeinere Beschreibung der Korrelation ist die nach *Spearsman*, welche auf Ranglisten beruht:

$$\begin{aligned}
r_s &= \frac{\sum_i (rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_i (rg(x_i) - \overline{rg}_x)^2}\sqrt{\sum_i (rg(y_i) - \overline{rg}_y)^2}} && (2)\\
&= \frac{\frac{1}{n}\sum_i (rg(x_i) rg(y_i)) - \overline{rg_x rg_y}}{s_{rg_x} s_{rg_y}} && (3)\\
&= \frac{Cov(rg_x, rg_y)}{s_{rg_x} s_{rg_y}} && (4)
\end{aligned}$$

Somit lassen sich auch nicht-lineare Korrelationen in Datensätzen erkennen. Insbesondere im Betrachtung des exponentiellem Verhaltens in Abhängigkeit von der Temperatur ist diese Art der Korrelation wichtig. Zur Berechnung des Anteils an der *erklärten* Varianz in linearen Modellen allerdings kann diese Variante nicht verwendet werden.

1.2.2 F-Test in geschachtelten Modellen

Der *F-Test* überprüft, ob zwei verschachtelte Modelle mit den Featuremengen $M_1 \subseteq M_2$ sich signifikant unterscheiden. Hierbei wird auf den gleichen Testdaten evaluiert. Die Statistik ist F-verteilt und abhängig von den Freiheitsgraden (Anzahl der Features im Modell). Es wird die Nullhypothese überprüft, ob die hinzugefügten Features des erweiterten Modells statistisch irrelevant sind ($\beta_i = 0$). Die F-Statistik wird gebildet durch:

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}} \tag{5}$$

mit den quadrierten Residuen RSS, den Anzahl an Freiheitsgraden p_i in den Modellen M_1 und das um weitere Feature erweiterte Modell M_2.

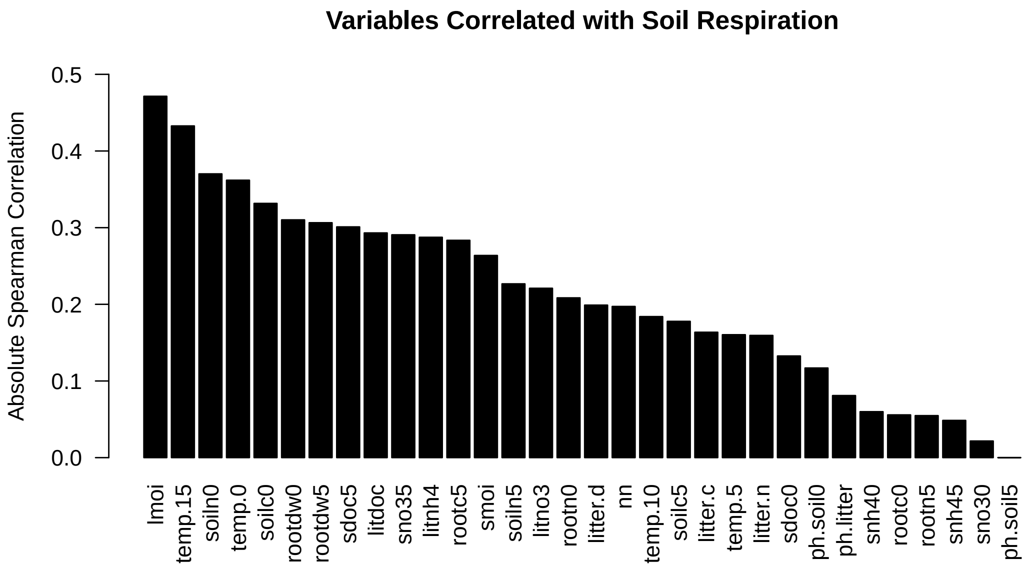

Abbildung 1: **Spearsman-Korrelation** der Einflussgrößen mit der Bodenatmung.

2 Erstellung eines Modells zur Bodenatmung

2.1 Korrelationsanalyse

2.2 Transformationen

$exp(Tmp)$ anstatt Tmp?

2.3 Variabelnselektion

2.4 Ergebnis

Qualität des gewählten Modells

2.5 Umsetzung mit R

```
## corelation
#correlation from soil.res with all others
hainich.r <- abs(cor(hainich, method = "pearson"))["soil.res",-1]
hainich.r.ordered <- hainich.r[order(hainich.r, decreasing = T)]
barplot(hainich.r.ordered, las = 2, ylim = c(0,0.5), col = "black"
    ,
```

```
ylab = "Absolute Pearson Correlation", main = "Variables Correlated with Soil Respiration")

## scatter plot
pairs(~ rootdw0 + smoi + nn,
data = hainich, main = "Scatterplot Matrix")
```

Literatur

[1] N. Jr. My article, 2006.