Bodenatmung im Nationalpark Hainich

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

Sichtling C Loos D Jajali R

7. September 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 **3**

1.1 Bodenatmung als geophysikalischer Prozess 3

1.2 Statistische Grundlagen 3

1.2.1 Fehlermaße 3

1.2.2 Korrelation 3

1.2.3 Informationskriterien 4

1.2.4 Test auf Vorliegen einer Normalverteilung 4

1.2.5 F-Test in geschachtelten Modellen 4

2 Erstellung eines Modells zur Bodenatmung **6**

2.1 Korrelationsanalyse 6

2.2 Transformationen 6

2.3 Variabelnselektion 6

2.4 Ergebnis 6

2.5 Umsetzung mit R 6

3 Simulation **8**

3.1 Vorgehensweise 8

3.2 Auswertung 8

3.3 Diskussion 10

3.3.1 Mögliche Gründe 10

3.3.2 Auswirkungen der Fehlentscheidungen 11

1 Einleitung

1.1 Bodenatmung als geophysikalischer Prozess

[1]

1.2 Statistische Grundlagen

1.2.1 Fehlermaße

RSS (*Residual sum of Suqres*). SPSE ist der summierte quadratische Fehler zwischen Modellvorhersage und Messwert auf unabhängigen Testdaten:

$$\widehat{SPSE} = \sum_{i=1}^{n} (y_i - \vec{\beta} * x_i^T)^2 \tag{1}$$

Das Modell wurde vorher mit anderen Trainingsdaten erstellt.

1.2.2 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:

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

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:

$$r_s = \frac{\sum_i (rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_i (rg(x_i) - \overline{rg}_x)^2} \sqrt{\sum_i (rg(y_i) - \overline{rg}_y)^2}} \tag{3}$$

$$= \frac{\frac{1}{n} \sum_i (rg(x_i) rg(y_i)) - \overline{rg_x rg_y}}{s_{rg_x} s_{rg_y}} \tag{4}$$

$$= \frac{Cov(rg_x, rg_y)}{s_{rg_x} s_{rg_y}} \tag{5}$$

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.

1.2.3 Informationskriterien

BIC/AIC

1.2.4 Test auf Vorliegen einer Normalverteilung

Shapiro

1.2.5 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 unter Annahme der Nullhypothese 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 ($H_0 : \beta_i = 0$). Die F-Statistik wird gebildet durch:

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}} \tag{6}$$

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. Ist der F-Wert je nach Signifikanzniveau hinreichend groß, kann die Nullhypothese abgelehnt werden: Das erweiterte Modell beschreibt die Daten nun signifikant genauer. Der Fehler RSS wird so stark verringert, dass es auch eine Erhöhung der Freiheitsgrade rechtfertigt.

Die R-Funktion `anova(model1, model2)` führt derartige Tests durch. Ist der p-Value (in R `F Pr(>F)`) kleiner als das Signifikanzniveau, so wird im Rahmen der Variabelnselektion das erweiterte Modell favorisiert.

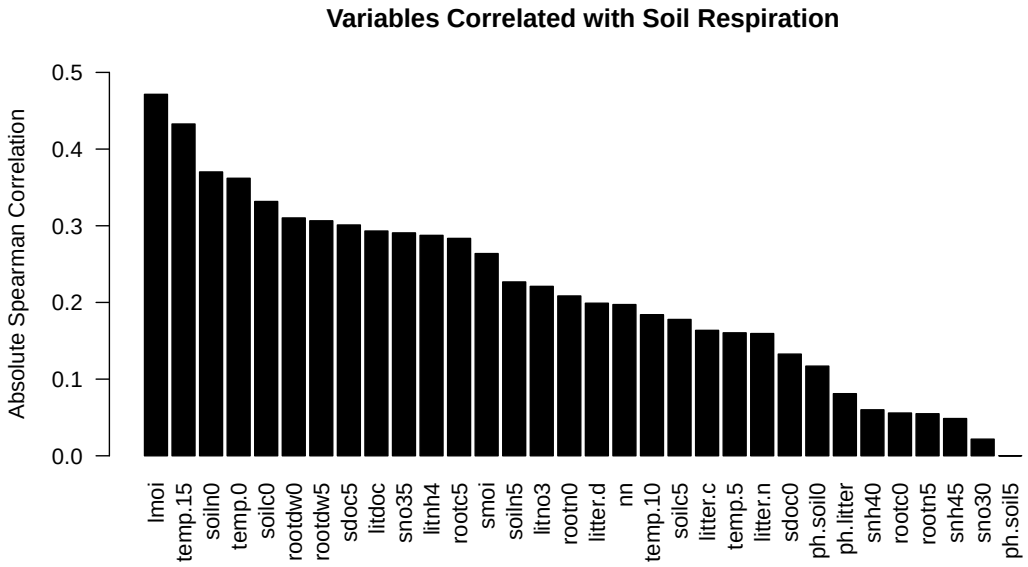

Abbildung 1: **Spearsman-Korrelation** der Einflussgrößen mit der Bodenatmung.

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

```
1 ## corelation
2 #correlation from soil.res with all others
3 hainich.r <- abs(cor(hainich, method = "pearson"))["soil.res",-1]
4 hainich.r.ordered <- hainich.r[order(hainich.r, decreasing = T)]
5 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")
```

3 Simulation

In Kapitel 2 wurde folgendes Modell ausgewählt:

$$soil.res = \vec{\beta} * (1, lmoi, temp15, smoi)^T \tag{7}$$

3.1 Vorgehensweise

Die Daten wurden in 80% zum Training und 20% zum Testen aufgeteilt. Anschließend wurde das "wahre"Modell um jeweils ein weiteres Feature erweitert. Das erweiterte Modell $soil.res = \vec{\beta} * (1, lmoi, temp15, smoi, x)^T$ wurde mittels ANOVA mit dem "wahren"Modell verglichen. In R berechnete die Funktion `anova(realModel, nxtModel )$F[2]` den jeweiligen F-Wert. Wichtig hierbei war, dass beide Modelle auf den selben Trainingsdaten erstellt wurden und es sich um *verschachtelte* Featuremengen handelte. Die Partitionierung der Daten erfolgte sowohl mittels Kreuzvalidierung als auch mit einem Monte-Carlo-Ansatz. Bei der 4-fachen Kreuzvalidierung wurden die Daten in vier möglichst gleich großen Partitionen aufgeteilt. Eine Partition wurde zum Testen des Reproduktionsfehlers verwendet, die Anderen zum Training. Beim Monte-Carlo-Ansatz wurden 100 mal zufällig 20% der Daten zum Testen ausgewählt. Dadurch kann die Wahrscheinlichkeit der Fehlentscheidungen im Bezug auf die Nullhypothese besser geschätzt werden. Die Ermittlung des Reproduktionsfehlers $\widehat{SPSE}$ erfolgte mittels unabhängigen Testdaten. Um zu überprüfen, ob sich das Modell überhaupt verbessert hat, wurde die Differenz $\Delta\widehat{SPSE}$ im Bezug auf das Ausgangsmodell berechnet.

3.2 Auswertung

In Abbildung 2 sind die F-Werte und Fehler bei der Reproduktion von unabhängigen Testdaten gezeigt. Im Rahmen des Variablenselektionsverfahrens wird jeweils das Feature zum Modell hinzugenommen, welches um den größten F-Wert verfügt. Im diesem Falle wäre das *litter.d*. Betrachtet man allerdings den erwarteten Reproduktionsfehler $\widehat{SPSE}$, so führt das Hinzunehmen der Variable *ph.soil*5 zum Modell mit minimalen Fehlern. Hier ist der F-Wert aber um $\approx 3,5$ geringer. Die Nullhypothese wird hier fälschlicherweise abgelehnt und stattdessen *litter.d* zum Modell hinzugenommen. Auch kann es vorkommen, dass eine Zufallsvariable überwiegend Rauschen beinhaltet und sich das Modell dann durch hinzufügen ebendieser Variable lediglich verschlechtert (Vgl. Abb. 4)
Nach der Monte-Carlo-Simulation war das Modell mit dem geringsten $\widehat{SPSE}$ stets das Modell mit den 9. höchsten F-Wert. Während der 200 Simulationen verblieb der Wert unverändert.

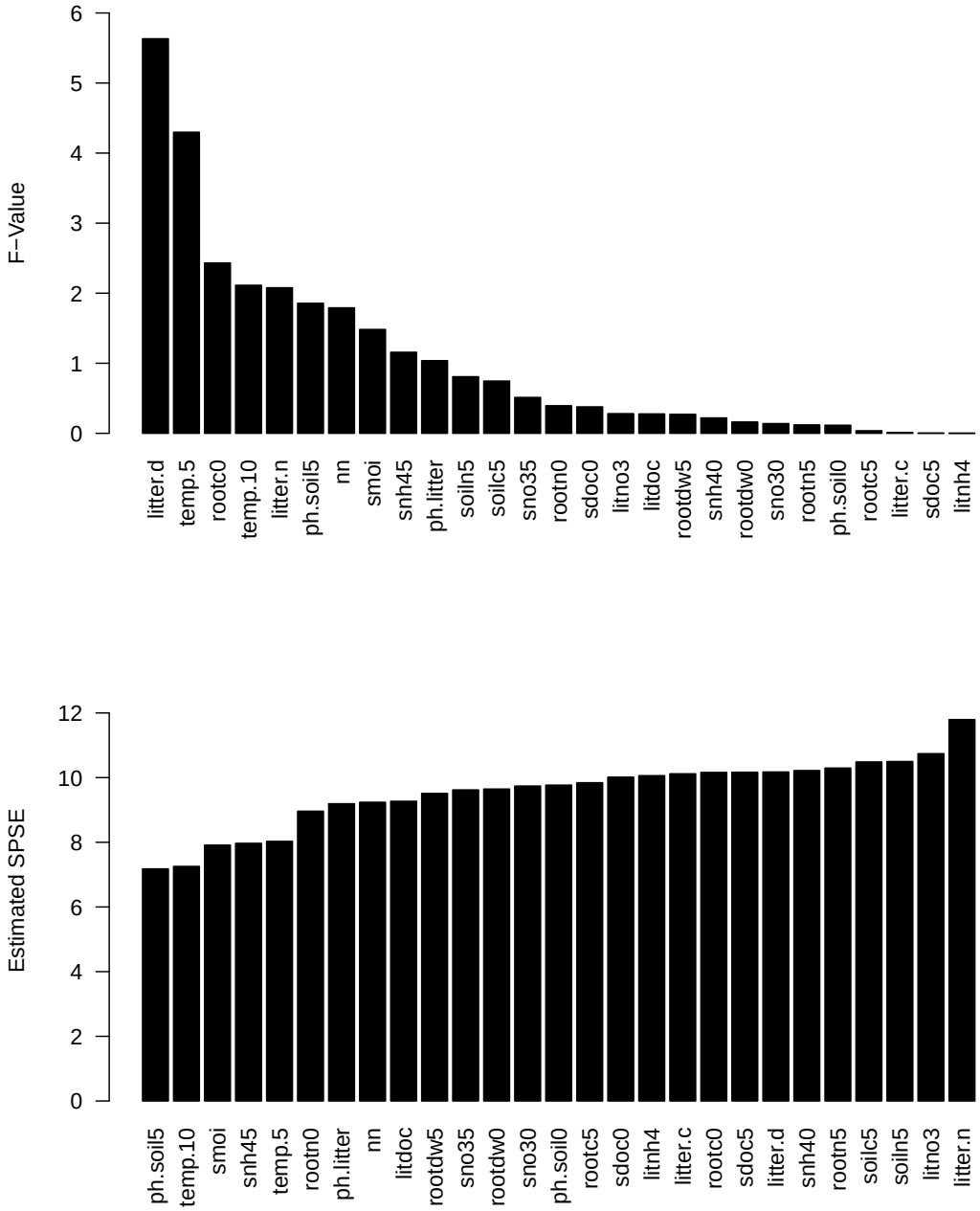

Abbildung 2: **Beispiel einer Simulation.** Das in Kapitel 2 erstellte "wahre"Modell wurde um jeweils ein Feature erweitert. Dargestellt ist der Fehler **SPSE** und der zugehörige **F-Wert** in Abhänigkeit von der gewählten zusätzlichen Variable in der ersten Runde der Kreuzvalidierung. Im Verlaufe der Variablenselektion wird das Modell mit dem höchsten F-Wert gewählt. Dies ist allerdings selten das Modell mit dem geringsten Fehler.

Unter der Nullhypothese (Erweitertes Modell ist nicht besser als das "wahre"Modell)

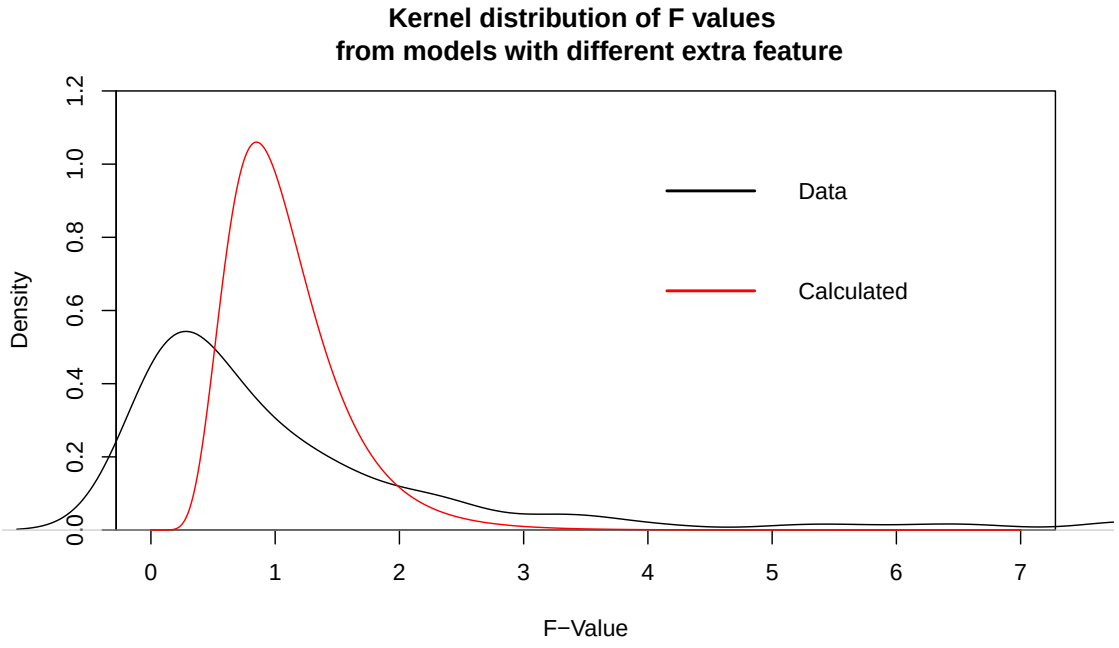

Abbildung 3: **Approximierte Dichtefunktionen der F-Werte.** Gezeigt ist die Kerneldichtefunktion der F-Werte der Simulationen und die berechnete F-Dichtefunktion mit den Freiheitsgraden der ANOVA-Tabelle unter Berücksichtigung der Nullhypothese. Der Unterschied der beiden Funktionen deutet auf fehlerhafte F-tests hin.

ist die F-Statistik eine F-verteilte Zufallsvariabel. Die Dichtefunktion der Daten aus der Kreuzüberprüfung (approx. durch Kernel-Desity) ist nach Abb. 3 mitunter stark abweichend. Die Simulation möchte demnach das wahre Modell erweitern (H_0 wird abgelehnt). Dies führt zu Overfitting, da das Modell bereits als wahr angenommen wurde und demnach keine weiteren Features hinzugefügt werden sollten.

3.3 Diskussion

3.3.1 Mögliche Gründe

Dies kann mehrere Gründe haben: Gemessene Features müssen nicht zwangsläufig von der Bodenatmung statistisch abhängig sein. Je geringer die Korrelation dieser beiden Zufallsvariablen, desto höher ist die Wahrscheinlichkeit, dass es sich nur um Rauschen handelt. In diesem Fall sollte diese Variabel nicht zum Modell hinzugefügt werden.

Zufälligerweise kann dies aber auch zu hohen F-Werten führen, welches zu fehlerhaften Entscheidungen bei den Hypothesentests führt. Ferner fordert der F-Test normalverteilte Variablen. Folgt eine Messgröße nicht dieser Verteilung, ist der Test nicht aussagekräftig. Auch eine nicht repräsentative Stichprobe fälscht das Ergebnis. Uns lagen lediglich 38 Observationen vor; dies könnte zu gering sein. Eine fehlerhafte Datenerhebung führt auch zu verfälschten Ergebnissen. Letztendlich kann der F-Test auch nur zufällig richtig sein. Ein geringer p-Value schließt keine Fehlentscheidungen aus; sie werden lediglich unwahrscheinlicher.

3.3.2 Auswirkungen der Fehlentscheidungen

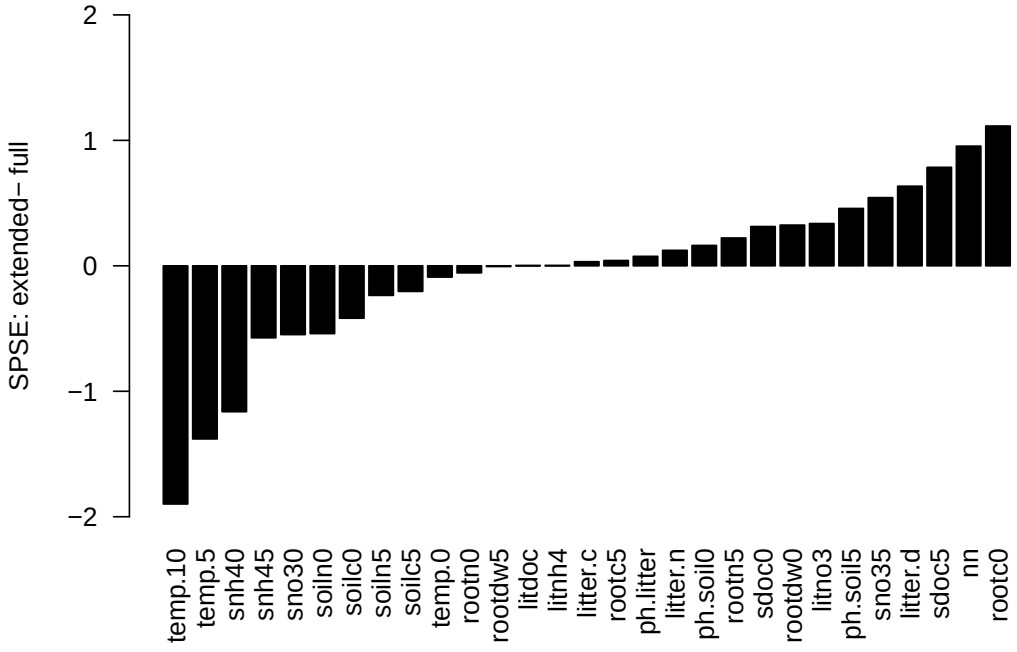

Abbildung 4: **Verbesserung der Modelle.** Dargestellt ist die Differenz $\Delta\widehat{SPSE}$ aus dem „wahren“ und dem erweiterten Modell. Positive Werte deuten darauf hin, dass die hinzugefügte Variable dem Modell lediglich Rauschen beifügt.

Literatur

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