MSc. Research Methods – Statistikteil
Lösungstext

- Übung 3.1: Multiple Regression -

Methoden

Es wurden Pflanzenartenzahlen von Steppenrasen in der Ukraine auf 199 10 m² grossen Probeflächen erhoben und zu diesen 23 Umweltvariablen erhoben. Da jeweils ein Messwert fehlte, wurden die Bodenvariablen Sand-, Schluff- und Tongehalt von der weiteren statistischen Analyse ausgeschlossen. Mittels Pearson's Korrelationskoeffizient wurde auf Abhängigkeiten zwischen den Prädiktorvariablen getestet und aus Paaren mit einer Beziehung mit $|r| > 0.6$ nur jeweils eine Variable beibehalten. Entsprechend wurden die hoch mit Jahresmitteltemperatur korrelierten Werte Meereshöhe (negativ), Temperaturamplitude (positiv) und Niederschlag (negativ) ausgeschlossen, ferner Leitfähigkeit (positiv mit pH) und Stickstoffgehalt (positiv mit organischem Kohlenstoffgehalt). Damit umfasste das globale lineare Modell (Funktion lm in R) die in Tab. 1 aufgeführten Variablen. Quadratische Terme oder Interaktionen zwischen Variablen wurden nicht berücksichtigt. Auf ein glm mit Poisson-Regression wurde verzichtet, da eine visuelle Inspektion eines Boxplots der Artenzahlen keine relevanten Abweichungen von einer Normalverteilung ergab.

Die Modellauswahl fand mittels Multimodel Inference (dredge-Funktion im MuMIn package in R) statt. Die Modellgüte wurde mittels AICc beurteilt. Zur Beurteilung der Bedeutung von Variablen wurden Importance Values (Summe der Akaike weights in allen Modellen, die die betreffende Variable beinhalten) ausgerechnet. Die Richtung der Beziehung wurde aus der Parameterschätzung im globalen Modell bestimmt. Schliesslich wurde ein gemitteltes Modell (aller möglichen Modelle, gewichtet nach deren Akaike weights) erstellt, um die Effektgrössen zu bestimmen. Die Adäquanz des gewählten Modells wurde in Residualplots visuell inspiziert und ergab keine nennenswerten Verletzungen der Voraussetzungen eines parametrischen Modells.

Ergebnisse

Das beste Modell nach AICc beinhaltete die Variablen CaCO3, CN-Verhältnis, Beweidungsintensität, Heat load-Index und Streuauflage. Da sich die nächstbesten Modelle aber um weniger als ΔAICc = 2 unterscheiden, wurden für die Gesamtbeurteilung die Importance values sowie die Koeffizienten des über alle 8192 betrachteten Modelle nach Akaike weights gemittelten average models herangezogen (Tab. 1). Mit einem Importance value von nahezu 100% war der Heat load-Index die wichtigste Einflussgrösse (negativ). Vier weitere Umweltvariablen waren ebenfalls in mehr als 50% der statistisch relevanten Modelle enthalten (in dieser Reihenfolge):

Tab. 1. Ergebnisse der Multimodel Inference. Die Variablen sind nach absteigenden Importance values sortiert. Die Parameterschätzung bezieht sich nur auf die Modelle, welche die entsprechende Variable beinhalten und ist nach Akaike weights gewichtet. Unter „Hochkorrelierte Variablen" sind jene aufgeführt, die wegen ihres engen Zusammenhangs mit der genannten Variablen nicht ins globale Modell aufgenommen wurden.

Variable (Einheit)	Hochkorrelierte Variablen (Richtung des Zusammenhangs)	Importance value	Parameter-schätzung
Achsenabschnitt		–	37.6
Heat-load-Index (-)		1.00	-12.9
Steuauflage (%)		0.92	-0.1
$CaCO_3$ (%)		0.82	+0.2
CN-Verhältnis (-)		0.73	-0.6
Beweidungsintensität (0 = keine … 3 = hoch)		0.68	+1.3
Deckung Steine und Felsen (%)		0.43	-0.1
Jahresmitteltemperatur (°C x* 10)	Temperaturamplitude (+), Jahresniederschlag (-), Meereshöhe (-)	0.39	+0.2
Mikrorelief (cm)		0.33	+0.1
Deckung Kies (%)		0.31	-0.1
Deckung Feinboden (%)		0.31	-0.1
Organischer Kohlenstoff (%)	Stickstoffgehalt (+)	0.30	+0.3
pH	Leitfähigkeit (+)	0.26	+0.2
Hangneigung (°)		0.26	+/-0.0