Erhöhung der Biodiversität von Graslandbeständen mittels p-Wert-korrigierter Assoziationsregeln

Jens Harbers [1]

Abstract: In dieser Ausarbeitung wird gezeigt, wie Pflanzenarten zur Erhöhung der Biodiversität mithilfe der Assoziationsanalyse identifiziert werden. Basierend auf einem frei zugänglichen Datensatz einer Vegetationserhebung wurden eine Assoziationsanalyse durchgeführt und die mit dem Apriorialgorithmus erstellten Regeln mittels dem Chi-Quadrat-Test auf Signifikanz ($p < 0,05$) überprüft. Anschließend wurden die p-Werte nach verschiedenen Methoden adjustiert, um insignifikante Regeln aus der Regelsatztabelle auszuschließen. Je nach Korrekturmethode konnte untermauert werden, dass der Datenumfang der Simulationsstudie zur Erstellung von signifikanten Mustern nicht ausreicht. Somit muss eine Simulationsstudie mit mehr Parzellen angefertigt werden, wie die p-Wert-Korrektur zeigte. Bei unterbleibender Korrektur der p-Werte waren etwa 18 % aller Regeln nach der Filterung relevant, während bei einer p-Wert-Korrektur hingegen keine der erkannten 11 768 312 Regeln statistisch signifikant waren.

Keywords: R, Patternmining, p-Wert-Korrektur, Grasland, Assoziationsregeln

1 Einleitung

Der fortschreitende Verlust an artenreichem Grasland stellt die Landwirtschaft sowie den Naturschutz vor Herausforderungen. Auf artenarmen Flächen muss daher die Biodiversität wieder erhöht werden, um einerseits die Stabilität eines Bestandes zu fördern und andererseits Lebensraum für Insekten und Vögel zu schaffen. Dies setzt resiliente Grünlandbestände voraus, sollen die Bestände langfristig erhalten bleiben. Zudem soll die Landwirtschaft selbst nicht zu stark beeinträchtigt werden, damit keine Betriebe in existenzbedrohende Situationen gelangen, deren Flächen nach Betriebsaufgabe ggf. nicht weiter als Grünland bewirtschaftet werden. Daher muss eine Möglichkeit gefunden werden, passende Arten zu identifizieren, die auf einem Standort ausgebracht werden können, und diese möglichst sicher im Bestand zu etablieren. Eine Methodik stellt hier die Assoziationsanalyse dar, die seit langem in der Mustererkennung eingesetzt wird. Das Ziel ist es, aus einer Transaktionsliste von Items passende Muster (Regeln) zu finden, um relevante Informationen aufzudecken. Die Regeln können in der Ausgangsmenge nur Spezies enthalten, die auf einer Parzelle gefunden wurden, und müssen auf neue, nicht auf der Parzelle befindliche Spezies hinweisen. In der Landwirtschaft stellt dabei die Validierung passender Spezies ein Problem dar, da das Austesten der Saatgutmischungen aus den Regeln erheblichen Aufwand darstellt. Der Stand der Literatur ist es, die Regeln

[1] Landwirtschaftsverlag Münster, Data Analyst, Hülsebrockstraße 2-8, 48165 Münster, jens.harbers@lv.de
https://orcid.org/0000-0001-6634-623X

nach Support, Konfidenz und dem Lift absteigend zu sortieren und auszugeben. Hingegen wird in diesem Beitrag eine andere Vorgehensweise genutzt, da hier stochastische Elemente berücksichtigt werden. Die Forschungsfrage besteht darin, aus der Vielzahl an Assoziationsregeln die geeignetsten zur Biodiversitätserhöhung ausfindig zu machen und so mit den vielversprechendsten Regeln zu arbeiten.

2 Material und Methoden

Dieser Abschnitt beschreibt die Datenverarbeitung und den Ablauf der Berechnungen. Es wird zudem folgendes Szenario angenommen: Eine Parzelle wird aufgrund einer Begehung als ökologisch verbesserungswürdig befunden. Anschließend wurden in direkter Umgebung weitere Parzellen in Hinblick auf vorkommende Pflanzenarten untersucht, um daraus Assoziationen abzuleiten. Im Folgenden wird die Parzelle A10_16, deren Biodiversität mit neuen Pflanzenspezies verbessert werden soll, aus dem ‚schedenveg'-Datensatz herangezogen, der im R-Paket ‚goeveg' [FJ21] liegt. Alle Daten wurden in R 3.4.3 [R 21] verarbeitet. Die Vorverarbeitung wurde mit den Grundfunktionen in R vorgenommen, während die Assoziationsregeln im ‚arules'-Paket mittels der Apriori-Funktion [MBK05] erstellt wurden.

2.1 Definition relevanter Begriffe und Maße

Nachfolgend werden relevante Maße der Assoziationsregeln definiert.

Das Itemset bezeichnet die Menge aller Spezies auf einer Parzelle.

Die Hypothese (X) einer Assoziationsregel wird Vorläufer bzw. Antecedent genannt.

Der Konsequent (Y) beschreibt das Itemset des Folgegliedes einer Assoziationsregel.

Der Support kennzeichnet die relative Häufigkeit, wie oft eine Regel zutrifft.

Die bedingte Wahrscheinlichkeit, die angibt, wie oft Y zutrifft, wenn X eingetreten ist, wird als Konfidenz bezeichnet.

Der in R ausgegebene p-Wert beschreibt die Grenzwahrscheinlichkeit, an der die Nullhypothese des Chi-Quadrat-Tests auf Unabhängigkeit von X und Y verworfen wird und dient als Ausgangswert für die p-Wert-Korrekturmethoden.

2.2 Datenverarbeitung

Der verwendete Datensatz ‚schedenveg' zeigt 28 verschiedene Parzellen an zwei Standorten, auf denen insgesamt 155 verschiedene Pflanzenspezies aufgenommen wurden. Zuerst wurde der Datensatz in ein Transaktionsdatenformat gebracht, in dem jede Parzelle eine Zeile darstellt und jede Spezies eine Spalte. Der Mindestsupport betrug 0,12.

Um Regeln auszuschließen, die kein hohes Vertrauensmaß enthalten, wurde vor der Berechnung außerdem eine Mindestkonfidenz von 0,05 gewählt. Der p-Wert wurde im Vorhinein auf 0,05 festgelegt.

Es wurden folgende Datenverarbeitungsschritte ausgeführt:

1. Umformen der Daten in ein Transaktionsformat, sodass die Parzellen und Spezies in zwei Spalten vorliegen.
2. Bestimmung der Itemsets der Vorläufer und Nachfolger.
3. Filterung von Regeln, damit nur Regeln, deren Vorläufer nur auf existierende Spezies im Bestand verweisen, enthalten sind.
4. Reduktion des Regeldatensatzes auf Nachfolger, deren Items ausschließlich auf neue, nicht im Feld vorhandene, Spezies hinweisen.
5. Ausführen des Chi-Quadrat-Tests und Ausgabe des p-Wertes.
6. Anwendung der p-Wert-Korrekturmethoden von Holm [Ho79], Benjamini-Yekutieli [BY01] und der Falscherkennungsrate (FDR) nach Benjamini-Hochberg [BH95]. Diese ist notwendig, um die Alphafehlerkumulierung zu adjustieren, die bei der mehrfachen Anwendung des Chi-Quadrat-Tests auf denselben Datenbestand entsteht.
7. Entfernung von insignifikanten Regeln ($p \geq 0{,}05$) nach p-Wertadjustierung und Regeln ohne p-Wert.
8. Sortieren der Liste nach Signifikanz (aufsteigend), Konfidenz (absteigend) und Support (absteigend).
9. Entnahme der ersten Regel je Nachfolger und Speicherung in einem neuen Objekt.
10. Sortieren des neuen Objektes nach Konfidenz (absteigend) und Support (absteigend) sowie Ausgabe einer Tabelle mit den zehn passendsten Spezies (siehe Tab. 1).

3 Ergebnisse

3.1 p-Wert-adjustierte Regeln

Abhängig von der Methodik einer p-Wert-Korrektur zeigte sich, dass je nach Stärke der Korrektur viele Regeln als insignifikant gelten. Insbesondere bei der Korrektur nach Holm ist erkennbar, dass alle p-Werte nach einer Korrektur als nichtsignifikant gelten und somit die Datengrundlage für das Pattern-Mining unzureichend ist. Diese zeigen durchwegs den Wert 1 in Abbildung 1. Bei der unterbliebenen p-Wert-Korrektur zeigten sich 80 % aller Regeln in Abbildung 1 als insignifikant. Je nach der Entscheidungsschwelle (dem kritischen p-Wert) sinkt die Anzahl signifikanter Regeln beim Anstieg des globalen

Signifikanzniveaus. Zudem ist erkennbar, dass alle Regeln je nach Korrekturmethode als insignifikant zu allen betrachteten Signifikanzniveaus (0,05;0,01;0,001) ausfielen. Damit stellen sich bei Anwendung einer der gewählten Korrekturmethoden keine vom Zufall unterscheidbaren Regeln dar. Abbildung 1 zeigt, dass bei Anwendung einer Korrekturmethode keine signifikanten Regeln bei einem Niveau von $p = 0{,}05$ mehr verbleiben. Ohne p-Wert-Korrektur blieben 8 855 signifikante Regeln übrig ($p < 0{,}05$). Gemessen am Ausgangsregeldatensatz von 11 768 312 Regeln wurden 99,924 % aller Regeln entfernt.

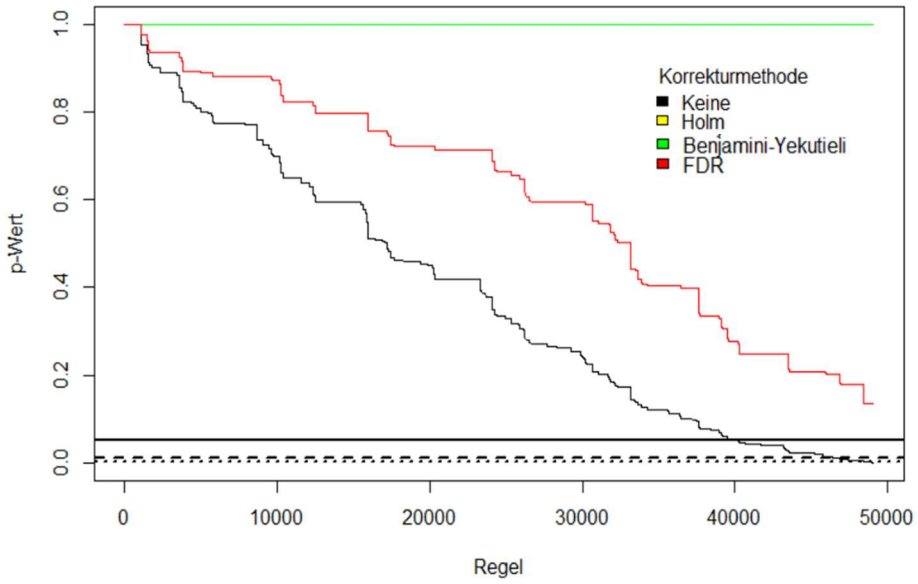

Abb. 1: p-Werte von 49 138 Regeln, die nach verschiedenen p-Wert-Korrekturmethoden adjustiert wurden. Es wurden nur Regeln berücksichtigt, deren Vorläufer auf einer Parzelle vorkommen und deren Nachfolger nicht in der Ausgangsmenge enthalten sind. Waagerechte Linien kennzeichnen Signifikanzschwellen von 0,05 (durchgezogen) sowie 0,01 (gestrichelt) und 0,001 (punktiert). Die FDR kennzeichnet die Falscherkennungsrate. Die Linie der Benjamini-Yekutieli-Korrektur überdeckt die p-Werte der Korrekturmethode nach Holm. Sie zeigen für alle Regeln einen korrigierten p-Wert von 1.

3.2 Die Top 10 der Spezies zur Bestandsverbesserung

In Tabelle 1 wurden Assoziationsregeln für zehn Spezies ausgegeben, die nach den Sortierkriterien oben standen und sich für die Ansaat auf dem Plot A10_16 eignen. Es zeigte sich, dass die Regeln in Tabelle 1 einen signifikanten p-Wert ($p < 0{,}05$) aufwiesen, wobei der Support eine Bandbreite zwischen 0,179 und 0,750 hatte.

Vorläufer	Nachfolger	Support	Konfidenz	p-Wert
{DacGlom}	{ArrElat}	0,750	0,875	0,000183
{GalAlbu,LeuIrcu, TarRude,TriPrat}	{ConArve}	0,179	0,833	0,000197
{GalAlbu,HolLana, LolPere,TarRude}	{RhiMino}	0,179	0,833	0,00245
{HolLana}	{FesRubr}	0,500	0,824	0,000805
{DacGlom,LeuIrcu}	{TriDubi}	0,321	0,818	0,00021
{MedLupu}	{PriVeri}	0,429	0,800	0,0248
{HolLana,TriPrat}	{AntOdor}	0,286	0,800	0,000267
{HolLana,TriPrat}	{HelPube}	0,286	0,800	0,00101
{DacGlom,PlaLanc}	{RumAcet}	0,643	0,783	0,000933
{GalAlbu,LeuIrcu}	{VicAngu}	0,321	0,750	0,000805

Tab. 1: Top-10-Regelsätze für die Ansaat neuer Spezies auf dem Plot A10_16, geordnet nach Konfidenz (absteigend) und Support (absteigend). ArrElat: Arrhenatheretalia elatioris; AntOdor: Anthoxanthum odoratum; ConArve: Convolvulus arvensis; DacGlom: Dactylis glomerata; FesRubr: Festuca rubra; GalAlbu: Galium album; HelPube: Helictotrichon pubescens; HolLana: Holcus lanatus; LeuIrcu: Leucanthemum ircutianum; LolPere: Lolium perenne; MedLupu: Medicago lupulina; PriVeri: Primula veris; RhiMino: Rhinanthus minor; RumAcet: Rumex acetosa; TarRude: Taraxacum sect. Ruderali; TriDubi: Trifolium dubium; TriPrat: Trifolium pratensis; VicAngu: Vicia angustifolia

4 Diskussion und Ausblick

Dieses Vorgehen zeigt, wie einzelne Spezies als neue Etablierungskandidaten identifiziert werden können, allerdings hat die Methodik Anpassungspotential. In der Ökologie ist bei der Aufwertung der Flächen mittels Saatgutmischungen gängig, jedoch bietet das R-Paket ‚arules' keine Möglichkeit für Saatgutmischungen, daher sei auf das Python-Modul ‚mlxtend' [Se18] verwiesen, womit Nachfolger mit mehreren Items ausgeben werden. Zudem ist zu beachten, dass es für das Ranking der Regeln auch nullinvariante Maße gibt, die entgegen dem Chi-Quadrat-Test nicht auf Nulltransaktionen reagieren und in Betracht gezogen werden sollen. Nullinvariante Maße kennzeichnen solche, die sich nicht ändern, wenn die Anzahl der Regeln, die weder den betrachteten Vorläufer noch den betrachteten Nachfolger beinhalten, sich ändert. Beispiele für Maße sind Cosine, Kulczynski und Max Confidence, wie in [WCH10] dargestellt. Diese sind die am besten geeigneten Maße, um Regeln auch aus nichtbalancierten Datensätzen zu sortieren [KBH11].

Ferner kommen biotische und abiotische Faktoren hinzu, die für die Bestimmung der optimalen Spezies als Etablierungskandidat relevant sind. Die Trieb- und Konkurrenzkraft einer Spezies stellt eine essentielle Kenngröße zur Etablierungswahrscheinlichkeit dar und

ist wiederum von vielen Umweltfaktoren abhängig [AYQ20]. Auch abiotische Faktoren wie die Frosthärte und Sonneneinstrahlung werden durch die Assoziations-analyse nur unzureichend gewürdigt, da hier keine Vorgruppierung der Standorte erfolgte und zudem eine Analyse mit Assoziationsregeln konstant bleibende Umweltfaktoren annimmt, was beim Klimawandel nicht haltbar ist. In einer Folgeuntersuchung sollen die Standorte vor der Assoziationsanalyse vorgruppiert werden und Assoziationsregeln je Standort berechnet werden.

Danksagung: Ich bedanke mich herzlich bei der Landwirtschaftsverlag GmbH für die Bereitstellung der notwendigen Ressourcen für das Projekt.

Literaturverzeichnis

[AYQ20] Ahmed, H. A.; Yu-Xin, T.; Qi-Chang, Y.: Optimal control of environmental conditions affecting lettuce plant growth in a controlled environment with artificial lighting: A review. South African Journal of Botany 2/130, S. 75-89, 2020.

[BH95] Benjamini, Y.; Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 1/57, S. 289–300, 1995.

[BY01] Benjamini, Y.; Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 4/29, S. 1165–1188, 2001.

[FJ21] Friedemann Goral; Jenny Schellenberg: goeveg: Functions for Community Data and Ordinations, 2021.

[Ho79] Holm, S.: A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 2/6, S. 65–70, 1979.

[KBH11] Kim, S.; Barsky, M.; Han, J.: Efficient Mining of Top Correlated Patterns Based on Null-Invariant Measures. In (Gunopulos, D. et al. Hrsg.): Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, Berlin, Heidelberg, S. 177-192, 2011.

[MBK05] Michael Hahsler; Bettina Gruen; Kurt Hornik: arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. Journal of Statistical Software 15/14, S. 1-25, 2005.

[R 21] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[Se18] Sebastian Raschka: MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. The Journal of Open Source Software 24/3, 2018.

[WCH10] Wu, T.; Chen, Y.; Han, J.: Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining and Knowledge Discovery 3/21, S. 371-397, 2010.