Dzień 3 - Testy statystyczne i regresja - zadania

Patryk Czortek, Marcin K. Dyderski

3 kwietnia 2019

Zadania do wykonania

Badano skład gatunkowy roślinności runa parków miejskich na Pomorzu (plik `fito.parki.csv`) w zależności od: udziału chwastów polnych (`arch_rat`), ergazjofitów (uciekinierów z miejsc uprawy i lokalnie zadomowionych; `erga_rat`), gatunków leśnych (`fore_rat`), obcych (`keno_rat`) oraz procentowego udziału typów pokrycia terenu w sąsiedztwie parków: wód stojących (`Water_0100`), terenów zurbanizowanych (`Settl_1000`), gęstości sieci rzecznej (`RivDens_1000`), pól uprawnych (`Agri_1000`) oraz lasów (`Forest_1000`). Dane te zawarto w pliku `cechen.parki.csv`. Linki: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/fito.parki.csv] oraz [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/cechen.parki.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
fito<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/fito.parki.csv',
               sep=';')
cechen<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/cechen.parki.csv',
               sep=';')
```

a) Na podstawie długości gradientu ocenić, która analiza (RDA czy CCA) będzie najodpowiedniejsza do analizy powyższych danych
b) Przeprowadzić odpowiednią analizę, a wyniki zobrazować w postaci wykresu (bez lub przy użyciu pakietu `ggplot2`). Które zmienne najbardziej wpływają na skład gatunkowy roślinności runa parków miejskich?
c) Z wykresu widać, że niektóre zmienne są ze sobą silnie skorelowane, co sugeruje, że w analizie powinniśmy użyć mniejszą liczbę predyktorów, nie uwzględniając tych o najwyższych wartościach VIF. Przy użyciu funkcji `vif.cca()` wskażcie, które predyktory mają największe wartości VIF (powyżej 10).
d) Stwórz model uwzględniający tylko predyktory o najniższych wartościach VIF. Wykonaj wykres i analizę PERMANOVA. Które predyktory istotnie wpływają na skład gatunkowy runa?
e) Dokonaj krokowej selekcji zmiennych modelu z podpunktu (d). Uwzględniając kryterium AIC wskaż zmienną/zmienne, która/które powinny być uwzględnione w modelu finalnym (istotnie wpływające na skład gatunkowy runa). Czy w tym przypadku ordynacja bezpośrednia jest odpowiednią metodą do analizy powyższych danych? Czy może wystarczy zwykła regresja?

Propozycje do pracy z własnym zbiorem danych

2. Jakiego typu orynację można wykorzystać w Twoich danych? Czy może być przydatna do testowania hipotez? Czy jako metoda eksploracyjna, pokazująca jak zmienia się kompozycja gatunkowa w gradientach? Jeśli nie badasz zbiorowisk/zgrupowań zastanów się nad wykorzystaniem PCA/CA jako metody eksploracyjnej analizy danych zamiast macierzy korelacji. Wykonaj prostą ordynacje na danych wycentrowanych lub standaryzowanych.