Multivariate Statistics: Exercise

Eduard Szöcs (szoecs@uni-landau.de)

January 25, 2016

1 Part I

The red squirrel (*Sciurus vulgaris*) is an endangered species in Scotland. To setup a conservation plan, it is important to know the effect of forest composition on red squirrel abundance.
In this exercise we will use data of [1].
Some informations about the study: 52 forest plots (`Id`) were setup and the number of trees (`Ntrees`), tree height (`TreeHeight`), diameter of trees at breast height (`DBH`) and canopy cover (`CanopyCover`) in these plots was recorded. Feeding remains of cones were counted (`SqCones`) and serve as an index for red squirrel abundance.

A research question was: *What habitat variables (no. trees, diameter and height of trees, canopy cover) influence squirrel abundance (as measured by the number of stripped cones observed)?*

You can download the data from here:
`https://raw.githubusercontent.com/EDiLD/permanova_lecture/master/exercises/exercise1/data/RedSquirrels.txt`

1.1 Tasks

1. Read the datafile `RedSquirrels.txt` into R. Your table should look like:

```
## 'data.frame': 52 obs. of  6 variables:
##  $ Id         : Factor w/ 52 levels "Abern1","Abern10",..: 1 12 23 27 28 29 30 31 32 2 ...
##  $ SqCones    : int  61 4 15 9 42 4 12 27 0 4 ...
##  $ Ntrees     : int  32 4 34 22 22 21 19 15 12 9 ...
##  $ DBH        : num  0.23 0.27 0.17 0.23 0.18 0.23 0.22 0.26 0.23 0.12 ...
##  $ TreeHeight : num  20.4 15.2 16 22.4 19.4 ...
##  $ CanopyCover: num  91.3 61.5 91.4 92 93.2 93.5 88.5 88 89.8 73.3 ...
```

2. Fit an appropriate model to the number of Cones explained by the habitat variables. (This includes model checking and possible removal of variables).
3. What variables influence the Squirrel abundance and how?
4. Write the resulting model equation in your script.

2 Part II

In this exercise we will analyse a dataset from Australia collected by R. Schäfer and colleagues [2]. The focus of this study was to analyse the effects of pesticide toxicity and salinisation on macroinvertebrate communities [3]. However, also other variables have been collected to check for their influence.
Macroinvertebrates, pesticides and other environmental variables were sampled at 24 sites situated in a 120km radius around Melbourne on three sampling occasions. These sites covered a gradient of both, pesticide exposure and salinisation. Pesticide toxicity was expressed in terms of Toxic Units (TU) with respect to *Daphnia magna* and salinity in terms of electrical conductivity ($\mu S/cm$ at 25°C).
There are two data files:

envdata.csv 22 measured environmental variables at the sites. Some of the variables have already been transformed. See Table 1 for details. The file is available for download here: `https://raw.githubusercontent.com/EDiLD/permanova_lecture/master/exercises/exercise3/data/envdata.csv`

abudata.csv Counts of 75 taxa collected during the study - mostly on family level. The file is available for download here: `https://raw.githubusercontent.com/EDiLD/permanova_lecture/master/exercises/exercise3/data/abudata.csv`

The first three columns are the same in both files: `Site`, `Month` and `Site_Month` are ID variables - they code uniquely each sample. These should not be included in your analysis, but are useful to join both tables.

2.1 Tasks

1. Read both datafiles `RedSquirrels.txt` into R.
2. We want to know which variables are correlated. Conduct a PCA on the environmental data set, **excluding** the two variables of interest `log_Cond` and `log_maxTU` (we are interested only in the variation of the remaining variables).
3. We want to reduce the correlated dataset. To how many axes can this be reduced by PCA?

Table 1: Overview of environmental variables.

Column	Variable	Unit	Transformation
`T`	Temperature	°C	-
`pH`	pH	-	-
`oxygen`	Dissolved oxygen	% sat.	-
`Depth`	stream depth	m	-
`max_width`	maximum stream width	m	-
`min_width`	minimum stream width	m	-
`rif_prec`	Pool	%	-
`pool_perc`	Riffle	%	-
`Bedrock`	Bedrock	%	-
`Boulder`	Boulder (>25.6 cm)	%	-
`Cobble`	Cobble (6.4 - 26.5 cm)	%	-
`Pebble`	Pebble (1.6 - 6.4 cm)	%	-
`Gravel`	Gravel (0.2 - 1.6 cm)	%	-
`Sand`	Sand (0.06 - 0.2 cm)	%	-
`Clay.silt`	Clay (<0.06 cm)	%	-
`log_Cond`	Conductivity	uS / cm	log10
`log_Nh4`	Ammonia	mg / L	log10
`log_NO2`	Nitrite	mg / L	log10
`log_NO3`	Nitrate	mg / L	log10
`log_PO4`	Phosphate	mg / L	log10
`log_Turb`	Turbidity	NTU	log10
`log_maxTU`	Maximum TU D. magna	-	log10

4. What proportion of variance can be explained by the first two axes?
5. Which of the following statements is correct?
   (a) Temperature (`T`) and `pH` are correlated.
   (b) `Clay.silt` and `log_Turb` are negatively correlated.
   (c) The proportion of riffles (`rif_perc`) and pools (`pool_perc`) are negatively correlated.
   (d) `Clay.silt` and Temperature (`T`) are not correlated.
   (e) Sand and Gravel are not correlated.

   One or more statements are correct.
6. Which of the following statements is correct?
   (a) The first axis could be interpreted as a gradient in hydrological conditions.
   (b) The first and second axis are not correlated.
   (c) The second axis is a gradient of chemical conditions.
   (d) `Clay.silt` is the most important variable for the first axis.

   One or more statements are correct.
7. Fit a RDA model, explaining the macroinvertebrate communities with the variables conductivity (`log_Cond`), pesticide load (`log_maxTU`), as well as the first two PCA axes from task 2 as predictors (to avoid problems with collinearity)!
8. Which of the following statements is correct?
   (a) Conductivity and PC1 are the two most influential predictors.
   (b) Lymnaeidea are mainly found at sites with high electrical conductivity.
   (c) The abundance of Simulidae is affected by hydrology.
   (d) The sample in row 64 has higher salinity then the sample in row 22.
   (e) Physidae can tolerate a relatively high amount of toxicity.

   One or more statements are correct.
9. What proportion of total variance can be explained by the predictor variables?
10. Which axes show a statistically significant amount of variation?
11. Which predictors explain most of the variation in the data?

References

[1] Silvia Flaherty, G Patenaude, A Close, and PWW Lurz. The impact of forest stand structure on red squirrel habitat use. *Forestry*, 85(3):437–444, 2012.

[2] Ralf B. Schäfer, Mirco Bundschuh, Duncan A. Rouch, Eduard Szöcs, Peter C. von der Ohe, Vincent Pettigrove, Ralf Schulz, Dayanthi Nugegoda, and Ben J. Kefford. Effects of pesticide toxicity, salinity and other environmental variables on selected ecosystem functions in streams and the relevance for ecosystem services. *Science of the Total Environment*, 415(1):69–78, 2012.

[3] E. Szöcs, B. J. Kefford, and Ralf B. Schäfer. Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates? *Science of the Total Environment*, 437(1):121–126, 2012.