15-Multivariate Analysis of Lake Data

2026-01-27

1 Introduction

This exercise explores basic multivariate principles using a lake data set from the German Umweltbundesamt (Umweltbundesamt, 2021).

Goal: You will perform an exploratory data analysis, followed by ordination (PCA, NMDS) and cluster analysis to identify patterns in lake characteristics.

2 Data Preparation

2.1 Data set and terms of use

The lake data set originates from the public data repository of the German Umweltbundesamt (Umweltbundesamt, 2021). The data set provided can be used freely according to the terms and conditions published at the UBA web site, that refer to § 12a EGovG with respect of the data, and to the Creative Commons CC-BY ND International License 4.0 with respect to other objects directly created by UBA.

The document and codes provided here can be shared according to more permissive CC BY 4.0.

2.2 Loading the Data

The following code loads the data and prepares the column names. Ensure the Excel file is in your working directory or adjust the path.

```
library("readxl")
library("vegan")
lakes <- as.data.frame(read_excel("3_tab_kenndaten-ausgew-seen-d_2021-04-08.xlsx",
                                  sheet="Tabelle1", skip=3))
names(lakes) <- c("name", "state", "drainage", "population", "altitude",
                  "z_mean", "z_max", "t_ret", "volume", "area", "shore_length",
                  "shore_devel", "drain_ratio", "wfd_type")

## create abbreviated lake identifiers
rownames(lakes) <- paste0(1:nrow(lakes), substr(lakes$name, 1, 4))
```

2.3 Variable Selection and Cleaning

We need to filter the dataset to include only numerical physical variables and remove incomplete rows.

Task

1. Create a subset of the data named `dat`.
2. Select the following columns: `drainage`, `population`, `altitude`, `z_mean`, `z_max`, `t_ret`, `volume`, `area`, `shore_length`, `shore_devel`, `drain_ratio`.
3. Remove all rows containing missing values (`NA`).

3 Data Inspection and Transformation

Before running multivariate statistics, we must understand the distribution of our data.

Task

1. Create boxplots of the `dat` dataframe.
2. Apply `scale()` to the data inside the boxplot function to make variables comparable.
3. Create aadditional boxplots using square-root and log transformed data .

Note: The `altitude` variable contains some negative values (below sea level). Replace any values < 0 with 0 before transforming.

Question: Why is it necessary to scale (z-transform) the data before visualizing it in a single boxplot?

Question: Compare the raw (scaled) boxplots with the square-root and log-transformed boxplots. Which transformation appears most suitable for this dataset and why?

4 Principal Components Analysis (PCA)

We will now reduce the dimensionality of the data using PCA.

Task

1. Perform a PCA on the **transformed** data using `prcomp()`. *Remember to scale the data within the function.*
2. Print the summary of the PCA object.
3. Create a biplot of the result.

Question: Look at the `summary()` output. How much of the total variance is explained by the first two principal components (PC1 and PC2) combined?

Question: Interpret the biplot.

- Which variables are strongly correlated with PC1 or PC2?
- Which variables are correlated with each other?
- Which variables are orthogonal?
- Which lakes are characterized by which variables?
- Are there clusters?

5 Nonmetric Multidimensional Scaling (NMDS)

PCA assumes linear relationships. Lets try NMDS, which is based on rank orders of distances.

Task

1. Use the `metaMDS` function from the **vegan** package on the transformed data.
2. Set `distance = "euclid"` (since these are physical variables, not species counts).
3. Plot the result using `type="text"` to see the lake names.

Question: Compare the NMDS plot to the PCA biplot. Do you see similar patterns in how the lakes are grouped?

6 Cluster Analysis

Finally, we will classify the lakes into groups using hierarchical clustering.

Task

1. Compute a distance matrix of the scaled, transformed data.
2. Run a hierarchical cluster analysis (`hclust`) using "Complete Linkage", Ward's method (`method="ward.D2"`), and "Single Linkage".
3. Plot the dendrograms.

Question: Which aggregation method is best suited for this data set?

7 Combining Clustering and Ordination

To visualize the clusters better, we will project them onto the NMDS plot.

1. Recreate the cluster object with the most suitable agglomeration scheme from the previous task.
2. Plot the dendrogram.
3. Cut the dendrogram into, for example, 5 groups using `cutree()`.
4. Plot the NMDS ordination again (empty plot).
5. Add text labels to the NMDS plot, using the cluster groups to color the text.

Question: Based on the colored NMDS plot, do the clusters correspond well to the geometric distances in the NMDS plot?

8 Discussion

Discuss the results of the applied methods regarding the characterization of the lake data set and the technical advantages and disadvantages of the applied methods.

Read essential parts of the the vegan documentation.

Information about the lakes can be found in Nixdorf et al. (2004). In addition to this, look for more recent information.

References

Nixdorf, B., Hemm, M., Hoffmann, A., & Richter, P. (2004). *Dokumentation von zustand und entwicklung der wichtigsten seen deutschlands* (Texte No. 05/04). Umweltbundesamt (UBA). https://www.umweltbundesamt.de/publikationen/dokumentation-von-zustand-entwicklung-wichtigsten

Umweltbundesamt. (2021). *Kenndaten ausgewählter Seen Deutschlands*. https://www.umweltbundesamt.de/daten/wasser/zustand-der-seen#okologischer-zustand-der-seen