1. Motivation

In the lecture in the slides and exercises you will learn the methods and basic knowledge of cluster analysis. This example should show you where the problems lie in practice and make you aware of the following points:

- cluster analysis should be applied in an exploratory way
- that it is important to transform the data
- that it is often important to standardise the data
- that it is crucial which variables are chosen
- that common measures of quality should be used with caution
- that the choice of the number of clusters is non-trivial
- that the choice of method must be as ideal as possible.
- Only if you have done pretty much everything right, you will get good results.

Since the data contains geographical information, we also want to use it to illustrate the results in maps.

2. kola data

The Kola region is on the Barents Sea and includes parts of Norway, Finland and Russia. Industry is mainly located in Russia. The following figure shows a map of this region.

Pure nature, e.g. in Finland:

versus industry (e.g. around Monchegorsk and Nikel):

The Nornikel company is the largest nickel refinery in the world, based in Nikel and Zapolyarny, see https://www.nornickel.com/business/assets/kola-division-russia/ .

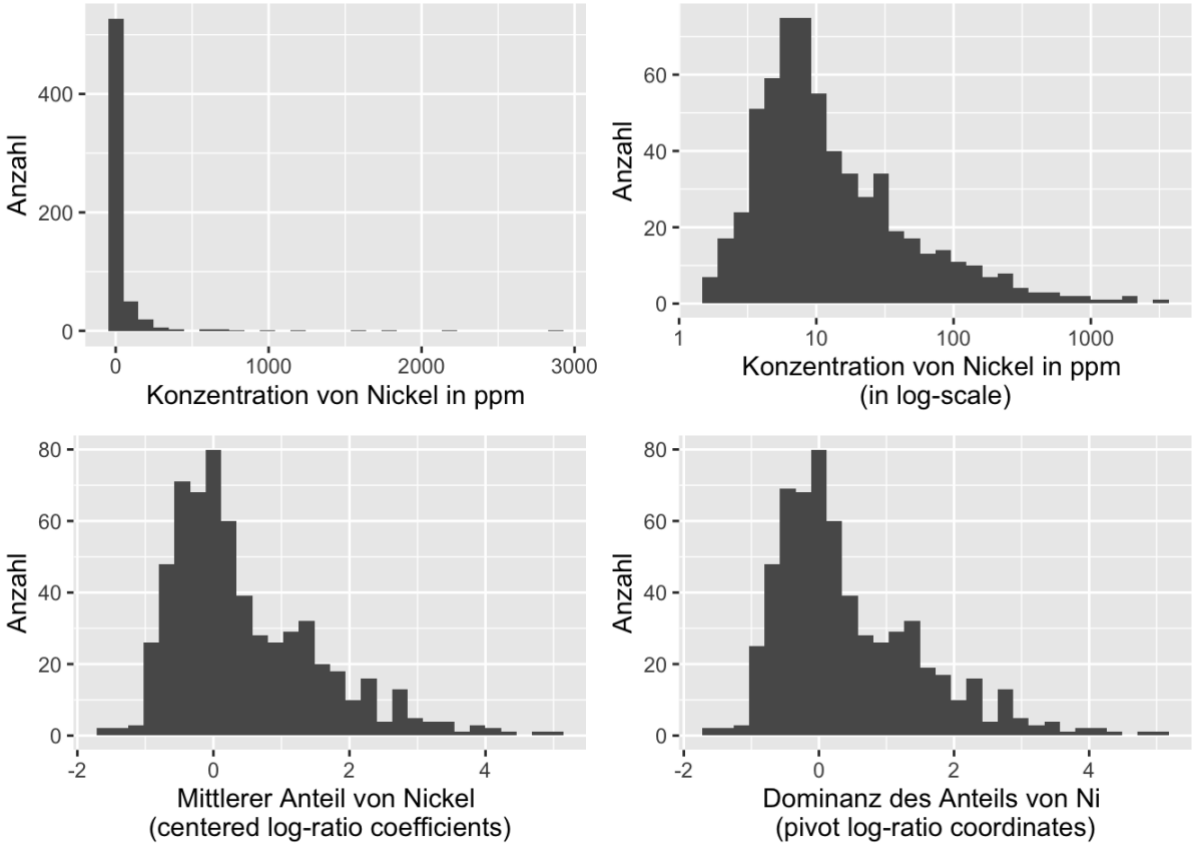
Some compare this area around Nikel, Zapolyarny and Monchegorsk to Tolkien's Mordor from Lord of the Rings. https://www.cryopolitics.com/2018/07/18/monchegorsk-eco-friendly-inferno/

Well, what can we do to prove to Russia how much they are destroying the environment in this region? Because insight is probably not available otherwise: https://www.highnorthnews.com/en/there-are-no-environmental-problems-nikel-says-putins-special-advisor

Sampling
Samples were taken from over 600 sites in five different layers.

Abbreviated insight

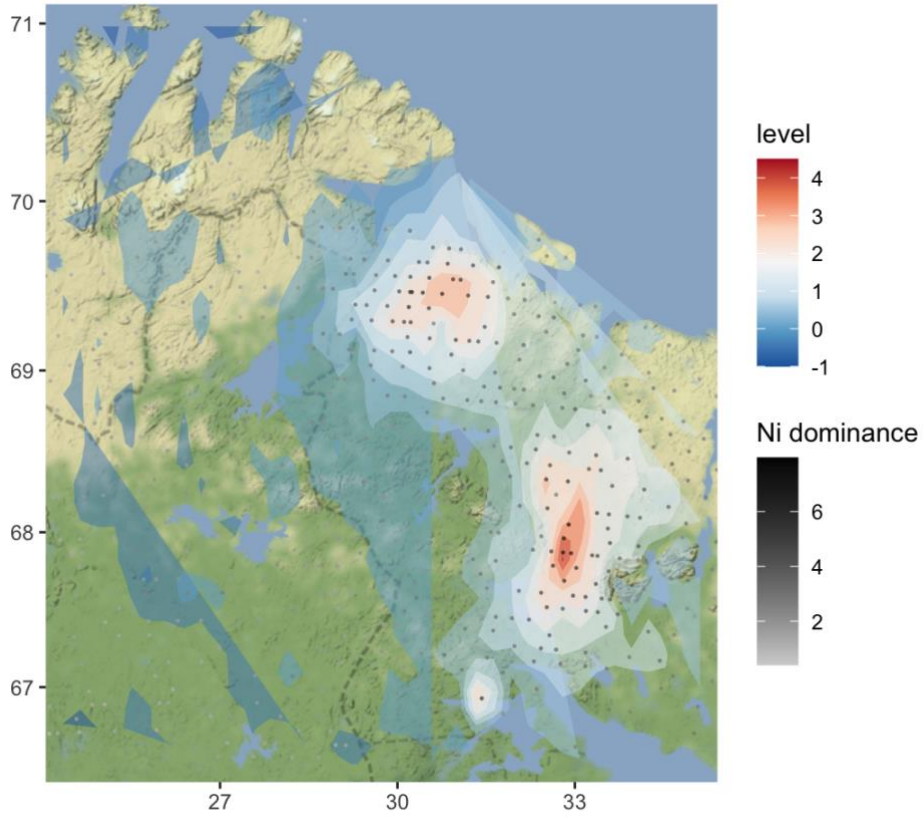
Univariate distributions can be viewed well, e.g. for Ni (nickel)



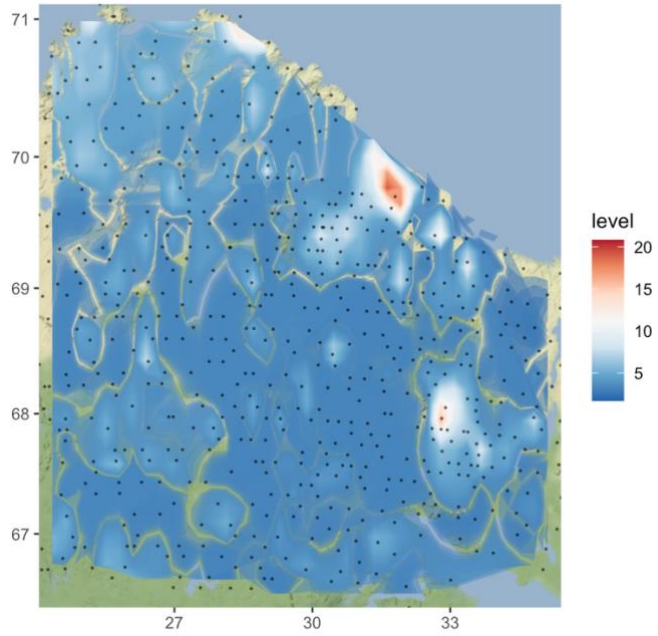
And on the map: (we have not bothered with the legend here, after all, we do not want to carry out a univariate evaluation). We see increased amounts of nickel especially around Monchegorsk and Nikel.

The amounts of Ni are especially high in the soils around Nikel, Zapolyarni and Monchegorsk compared to the amounts of other variables.

Geochemists usually make such maps of each chemical element and then (somewhat desperately) compare about 80 maps with each other. In other words, they need your help to look at the data multivariately.



A multivariate outlier detection shows that especially in these three areas unusually high values were measured.



We want to analyse more details with a cluster analysis.

3. Auxiliary functions

Later you may need these two auxiliary functions:

- findCenter()
- plotmeans()

```
findCenter <- function(x, clustering){
  agg <- aggregate(x, list(clustering), mean)[, -c(1)]
  return(agg)
}


plotmeans <- function(x, cl = cl1){
  cen1 <- findCenter(x[, 3:ncol(x)], cl)
  df <- t(cen1)
  colnames(df) <- paste0("cluster", 1:ncol(df))
  df <- data.frame(df)
  df$variable <- rownames(df)
  df <- reshape2::melt(df)
  colnames(df) <- c("variable", "cluster", "mean")
  df$id <- as.integer(as.factor(df$variable))
  library(ggplot2)
  ggplot(df, aes(x = id, y = mean, label = variable)) +
    geom_text() +
    facet_wrap(~cluster) + theme_bw() + xlab("") +
    theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank()) +
    ylab("cluster means")
}
```

4. Task definition

1. Produce a well-interpretable result by making a good choice of standardisation x transformation x variable selection x clustering method. Important: Justify your choice! To achieve a good result, it is necessary to try out a lot, but also be familiar with the cluster methods.
2. Also visualise the results in maps and compare the cluster means for each cluster (plotmeans).
3. Also evaluate the quality of the clustering. Comment on your assessment.

Hints:

- It helps if you think about which variables stand for pollution (e.g. Ni, but there is more of it), which stand for the influence of the sea (e.g. Na, but there is more of it), etc., i.e. make yourself smart about the measured chemical elements.
- Analysis from the C horizon, for example, may be different than from the O horizon, as the (older) rock structure is important in the former, and human influence in the latter.
- This time chatGPT will not be of much help, as the problems and possibilities concerning cluster analysis are not really discussed.
- Due to the complexity of the task, they are not expected to find a very good cluster result by an ideal choice of parameters and methods. There are also points for a suboptimal choice that can be earned by describing the results well.

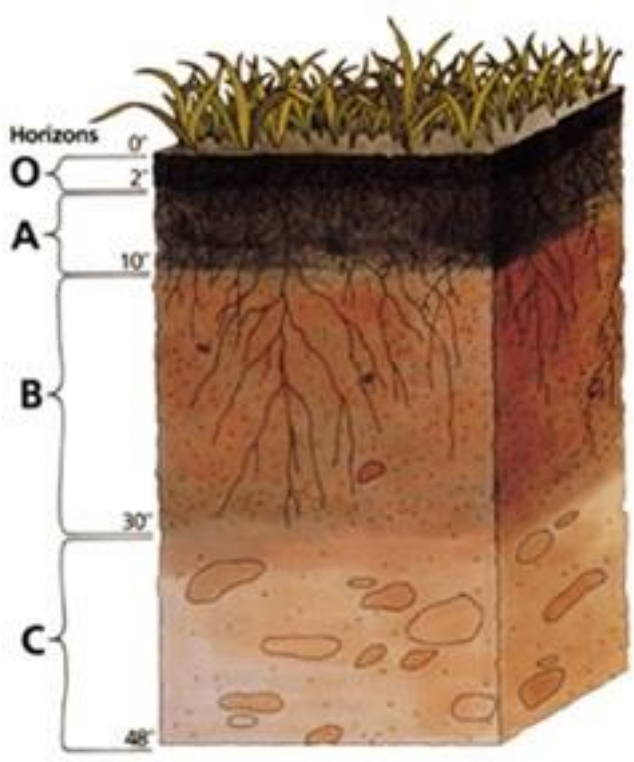
```
# ?ohorizon
data("ohorizon", package = "StatDA")
vars <- c('XCOO','YCOO','Ag','Al','As','Au','B','Ba','Be',
          'Bi','Br','C','Ca','Cd','Cl','Co','Cr','Cu','Fe',
          'H','Hg','K','La','Mg','Mn','Mo','N','Na','Ni',
          'P','Pb','Pd','Rb','S','Sb','Sc','Se','Si','Sr',
          'Th','Tl','U','V','Y','Zn')
x <- ohorizon[, vars]
coord <- x[, 1:2] # Koordinaten
x <- x[, 3:ncol(x)] # chem. Konzentrationen
dim(x)
## [1] 617  43
dim(na.omit(x))
## [1] 611  43
```

5. data selection

Depending on the matriculation number, you use a different data set.

The following data sets are available:

bhorizon: B-horizon of the Kola Data
chorizon: C-horizon of the Kola Data
humus: Humus Layer (O-horizon) of the Kola Data
moss: Moss Layer of the Kola Data



Concentrations of about 80 chemical elements were measured.

Select your data set by inserting your matriculation number (instead of 123456). For example, for 123456 this would be:

```
Matrikelnummer <- 123456
set.seed(123)
s <- sample(c("bhorizon", "chorizon", "humus", "moss"), 1)
s
## [1] "humus"
```

Work with your randomly drawn data set. The data sets are available in the package mvoutlier. For the matriculation number 123456 this would be:

```
# load data
library(mvoutlier)
assign(s, get(s))
```

My Matriculation number: 20490660

6. delivery format

We want to use modern data science tools. Therefore, a submission in MS-Word, MS-Excel and similar is not allowed.

Please submit the following files on Moodle:

an R-Mardown (Rmd) and a
a complicated HTML or PDF file.
In addition to your interpretations, please make your (commented) R-code visible as well as the generated graphics.