Task chapter 9: Clustering of Thunderstorm Data

Background

Conventional wisdom has it that thunderstorms in winter are so much rarer than in summer because conditions that cause thunderstorms in summer, e.g. substantial amounts of CAPE, only rarely occur in winter. The statistics research group set out to test whether this is actually true. We realized that "thunderstorm", which is defined as having at least one lightning flash, need not be synonymous with "strong convection" since maybe lightning could also take place without strong convection. To reduce the problem to its bare essentials, a fairly small (two thirds the size of Austria) and flat region in northern Germany was selected. Lightning flashes recorded by a lightning locations system (LLS) and ERA5 reanalysis data and variables derived from it that are relevant for the electrification of clouds are used for the period 2010-2019.

Data

Data were used at the resolution of ERA5, i.e. 0.25 degrees horizontally and 1 hour in time. When at least one flash occurred in such a "cell-hour", a thunderstorm was classified. Over the whole period and region, 1576 such cell-hours occurred in winter (DJF). To find out whether different processes are responsible for lightning and whether these processes have a seasonal dependence, four scenarios are considered: winter and no lightning (wnol), winter and lightning (wl), summer and no lightning (snol), and summer and lightning (sl). 1576 cell-hours were cleverly sampled for the remaining three scenarios to yield four samples of exactly same size. The data are contained in the variable `dat` in the file `lightningPCAclustering.rds`. Explanations of the variables in `dat` are given in `lightningPCAclustering_variables_explanations.csv`.

Tasks

Scaling

It is less error-prone for both the scaling and the k-means clustering to assign the ERA5 variables to a separate data frame `era5` since the "scenario" column must be excluded from these operations (which can of course be achieved simply by subsetting with `dat[,-1]`).

Before you cluster, scale the data to the mean and standard deviation of the scenarios without lightning, i.e. wnol and snol, since no lightning is the dominant mode of the atmosphere. Use `scale()` to do that:

`ms <- scale(dat[dat$scenario %in% c("snol", "wnol"),-1])`. Note that the `-1` excludes (minus sign!) the first column (== scenario) from the scaling since only numerical values can be scaled. This `scale` creates the attributes "scaled:center" and "scaled:scale" containing mean and standard deviation, respectively. And then apply the scaling to the whole data set `era5`, again using `scale` with the options

`center = attr(ms, "scaled:center"), scale = attr(ms, "scaled:scale")`

Finally convert the result back to a data frame using `as.data.frame()` and assign to `era5`, which should still have 35 columns of ERA5-derived and scaled variables.

Clustering

Use k-means to assign the data to 5 different clusters. Do not forget to use `set.seed()` beforehand to make your results reproducible. Cluster on `era5` and assign the result to `clK` with

`clK <- kmeans(...)` Use 150 (or a value close to it) for `nstart` and `algorithm = "MacQueen"` as clustering algorithm.

Compute frequencies of clusters in each of the 4 scenarios

First assign a name to *each cluster*, e.g.,

`dat$cluster_new[clK$cluster == 1] <- "wind_field"`. Names (from 1 to 5) should be "wind_field", "mass_field", "average", "cloud_physics_mass", "cloud_physics_wind". While these names seem to come out of the blue at this stage they will become clearer after the PCA (the second task for this week) has been performed.

Compute the frequencies of cases in each cluster using `table()` on `dat$scenario` and `dat$cluster_new` and assign to `tab`. What do you notice about the distribution of the clusters over the scenarios?

Instead of absolute frequencies you can also compute relative frequencies using `proportions()` on `tab`. You need to specify whether you are computing relative frequencies with respect to rows (`margin = 1`) or columns (`margin = 2`). Assign the relative frequencies to `relFreq`.

Plot distribution of the clusters over the scenarios

First, merely for the purpose of easier interpretation of the results, re-order your frequency table in `tab` so that the sequence of rows is "wl", "wnol", "snol" and "sl"; and the sequence of columns is "cloud_physics_wind", "wind_field", "average", "mass_field" and "cloud_physics_mass".

Use `mosaicplot()` on `tab` to plot the distribution; as option use `off = c(15,0)`, which joins the mosaic pieces and eliminates the default space betweent them. For the option `color` use the following colors

```
cols_mosaic <- rev(hcl(c(10, 10, 80, 250, 250), c(70, 60, 50, 60, 70), c(20, 60, 90, 60,
20)))
```

This plot is now easier to digest (for most people) than the numbers in the tables that you computed previously. The five different clusters have been named after the variables that most strongly deviate from "average", i.e. non-lightning, conditions. Note that "mass field" contains variables related to temperature and pressure (and humidity). Try to interpret the mosaic plot in light of this information and with your expert knowledge in meteorology.