Task chapter 9: Clustering of Thunderstorm Data

David Kurz, Malte Hildebrandt

2023S707737 VU Geostatistik

Universität Innsbruck

June 27, 2023

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.

→ The scaling is applied to the subsetted data, which performs a standardization on the variables. This is done by centering, to have a mean of zero and scaling to have a standard deviation of one. When variables have different units, scaling to have a standard deviation of one is higly recommended.

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.

→ Here the number of subgroups in the data is pre-defined with 5 clusters with similarities in each group. The idea of K-mean clustering is to have a very small within-cluster-variance defined by the Eucledean-distance between each number of observations in k clusters. "nstart" specifies the number of random initializations to try, from which the best clustering solution will be selected based on the lowest total within-cluster sum of squares.

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 ssign 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.

1. The Cluster "average": appears most frequently in the "sl" (878 occurrences) and "snol" (205 occurrences) scenarios. It has a smaller occurrence in the "wl" scenario (10 occurrences) and a very minimal occurrence in the "wnol" scenario (2 occurrences).

Frequencies of cases in each cluster scenario

	average	cld. phys. mass	cld. phys. wind	mass field	wind field
sl	878	24	444	132	98
snol	205	20	43	1213	95
wl	10	300	20	103	1143
wnol	2	37	0	1144	393

2. Cluster "cld. phys. mass": has a relatively low occurrence in all scenarios: 24 occurrences in "sl," 20 occurrences in "snol," 300 occurrences in "wl," and 37 occurrences in "wnol."

3. Cluster "cld. phys. wind": is most prevalent in the "sl" scenario with 444 occurrences. It has a smaller occurrence in the "snol" scenario (43 occurrences) and a negligible occurrence in both the "wl" and "wnol" scenarios (20 occurrences and 0 occurrences, respectively).

4. Cluster "mass field":is predominant in the "snol" scenario with 1213 occurrences. It has a substantial occurrence in the "wl" scenario (103 occurrences) and a significant occurrence in the "wnol" scenario (1144 occurrences). However, it has a much lower occurrence in the "sl" scenario with only 132 occurrences.

5. Cluster "wind field": has a significant occurrence in the "wl" scenario with 1143 occurrences. It also appears frequently in the "wnol" scenario with 393 occurrences. In contrast, it has a much lower occurrence in both the "sl" (98 occurrences) and "snol" (95 occurrences) scenarios.

The cluster occurrences suggest potential associations between certain clusters and specific scenarios, indicating differences in the patterns or characteristics captured by the clustering algorithm.

relative frequencies of cluster scenarios row-wise (rounded)

	average	cld. phys. mass	cld. phys. wind	mass field	wind field
sl	0.5571	0.0152	0.2817	0.0838	0.0621
snol	0.13	0.0127	0.0273	0.7697	0.0603
wl	0.0063	0.1904	0.0127	0.0654	0.7253
wnol	0.0013	0.0235	0	0.7259	0.2494

	average	cld. phys. mass	cld. phys. wind	mass field	wind field
sl	0.8018	0.063	0.8757	0.0509	0.0567
snol	0.1872	0.0525	0.0848	0.468	0.0549
wl	0.0091	0.7874	0.0394	0.0397	0.6611
wnol	0.0018	0.0971	0	0.4413	0.2273

relative frequencies of cluster scenarios column-wise (rounded)

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.

distribution of clusters scenarios
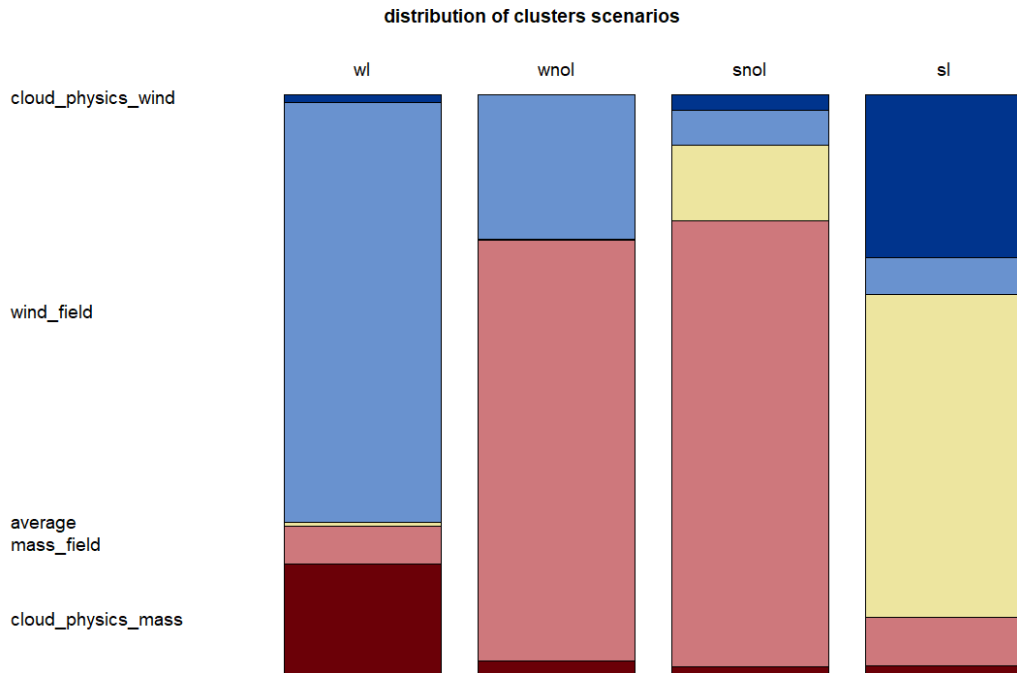
→ In the winter season cloud physics mass and especially the windfield as a scenario is the dominating cluster with its occurances if lightning is observed in winter. So a strong windfield with distinct direction and mass as cloud physics parameter are the driving factors for lightning in winter. Where in contrast the mass field, so when high pressure is dominating inhibits lightning occurances.

→ In summer, the cluster of mass field respectively high pressure is slightly more dominant than in winter. The most dominant lightning cluster in summer is the "average" where none of these 5 pre-defined clusters have its occurance. I think its due to forming lightning in summer either when there are hot temperatures or when the airmasses and so the temperatures change due a coldfront passage. Second frequent cluster is wind as cloud physics which may indicate some coldfront passage or cold sector forming lightning in during high windspeeds.