# Distributional Modelling in R

## 07 - Trees and Forests - Exercises

In this example, we will analyze the frequency of motorcycle rides at Sonnenberg in the Harz region of Germany. The data frame can be download with the following R code

```
R> download_data <- function(data = "prepared_Sonnenberg.rda") {
+    file <- paste0("https://nikum.org/dmr/Data/", data)
+    tdir <- tempfile()
+    dir.create(tdir)
+    download.file(file, file.path(tdir, data))
+    load(file.path(tdir, data), envir = .GlobalEnv)
+  }
R> download_data("prepared_Sonnenberg.rda")
R> library("zoo")
R> x <- zoo(data$bikes, data$date)
R> plot(x, xlab = "Time", ylab = "#Bikes")
```

The objective of this analysis is to develop a comprehensive probabilistic model that provides the best explanation for the number of `bikes` observed. A first try dummy model is also contained in the `.rda` (object `model`).
The data consists of the following variables:

| Variable | Description |
|---|---|
| `date` | The date. |
| `trucks` | Number of trucks. |
| `cars` | Number of cars. |
| `bikes` | Number of motorbikes. |
| `others` | Number of other vehicles. |
| `wday` | The day of the week. |
| `weekend` | Weekend or weekday? |
| `yday` | The day of the year. |
| `sinyday` | A sine transformation of `yday`. |
| `cosyday` | A cosine transformation of `yday`. |
| `tlmin` | Minimum temperature in °C. |
| `tlmax` | Maximum temperature in °C. |
| `tl` | Mean temperature in °C. |
| `rh` | Mean relative humidity in percent. |
| `td` | Average dew point temperature °C |
| `cloudiness` | Average cloud cover in percent. |
| `rain` | Amount of precipitation in mm (snow and rain). |
| `sunshine` | Sunshine duration in minutes. |
| `wind` | Mean wind speed in m/s |
| `windmax` | Maximum wind speed in m/s |

1. To start, estimate a classical GAMLSS model using the **gamlss2** package and try to find a suitable model given the covariates. Split the data into a training and a test set, using years 2021 and 2022 as the training set and 2023 as the test set. Identify which covariates appear to have the strongest effects on the number of bikes.

2. In the next step, estimate a distributional tree model (package **disttree**) using all covariates. Plot the estimated tree and compare the results with those obtained from the classical GAMLSS model.

3. Finally, estimate a distributional forest model. Assess whether this increases predictive performance by examining the out-of-sample log-likelihood score for the year 2023.

4. Using your most accurate model, calculate the probability of observing more than 500 and 1000 bikes for each day of the year and visualize the results.