**Case study: Tree – based methods**

# Predicting Algae Blooms

(Source: Torgo L. Data Mining with R, 2nd Ed., Chapman & Hall/CRC Press.)

## Background

High concentrations of certain harmful algae in rivers constitute a serious ecological problem with a strong impact not only on river lifeforms and environment, but also on the quality of water provided to consumers. Being able to monitor and perform an early forecast of algae blooms is essential to improving the quality of rivers.

## Case study (Business Understanding Phase)

With the goal of addressing this prediction problem, several water samples were collected in different European rivers at different times during a period of approximately 1 year. For each water sample, different chemical properties were measured as well as the frequency of occurrence of seven harmful algae. Some other characteristics of the water collection process were also stored, such as the season of the year, the river size, and the river speed.

One of the main motivations behind this application lies in the fact that chemical monitoring is cheap and easily automated, while the biological analysis of the samples to identify the algae that are present in the water involves microscopic examination, requires trained manpower, and is therefore both expensive and slow. As such, obtaining models that are able to accurately *predict* the algae frequencies based on chemical properties would facilitate the creation of cheap and automated systems for monitoring harmful algae blooms.

[Another possible objective of this study could be to provide a better understanding of the factors influencing the algae frequencies. In that case, we would want to understand how these frequencies are related to certain chemical attributes of water samples as well as other characteristics of the samples (like season of the year, type of river, etc.).]

## The data (Data Understanding Phase)

There are three datasets for this problem.

1. The first dataset (Analysis.txt) consists of data for n = 200 observations (water samples). To be more precise, each observation in the available datasets is in effect an aggregation of several water samples collected from the same river over a period of 3 months, during the same season of the year.
   Each observation contains information on **11 variables (IVs)**.
   - 3 of these variables are nominal (categorical) and describe:
     - Season of the year when the water samples to be aggregated were collected
     - Size of the river in question

    - Speed of the river in question
- The 8 remaining variables are values (numerical) of different chemical parameters measured in the water samples forming the aggregation, namely:
    - Maximum pH value
    - Minimum value of $O_2$ (oxygen)
    - Mean value of Cl (chloride)
    - Mean value of $NO^{-}_{3}$ (nitrates)
    - Mean value of $NH^{+}_{4}$ (ammonium)
    - Mean of $PO^{3-}$ (orthophosphate) $_4$
    - Mean of total $PO_4$ (phosphate)
    - Mean of chlorophyll

The rest of the variables in the dataset (**a1- a7**) are the frequencies of different algae **(DVs).**

2. The second dataset (Eval.txt) contains information on n=140 extra observations, which can be regarded as a kind of test. It uses the same basic structure as the training data set, but it does not include information concerning the seven harmful algae frequencies (the target). This information is given separately in the file Sols.txt.

The main goal of our study is to correctly predict the frequencies of the seven algae for these 140 water samples (test sample). This means that we are facing a predictive data mining task.

Data access:

- You may import the data files mentioned above and available on Blackboard.
- The data is also available in the package DMwR2. If we load "dplyr" package before loading the data, we get a data frame table object (a tibble) instead of the standard data frame. A tibble is a modern way to work with our data comparatively with data frames.

```
library(dplyr)
library(DMwR2)
data(algae, package="DMwR2")
```

## Requirements:

Similar to the previous Case study, the four main operational tasks are:

Task 1: Understand and import data properly
Task 2: Inspect your data and do the required variable adaptations and transformations
Task 3: Build one or several predictive tree-based model(s) and evaluate their performance.
Task 4: Reflect on implications and recommendations