# Chapter 02 - Predicting Algae Blooms

O objetivo deste estudo é realizar a predição de expansão de algas em rios da Europa.

Foram coletadas diversas amostras de água em rios da Europa, durante aproximadamente um ano. Foram medidas diferentes propriedades químicas e a frequêncida de ocorrência de 7 tipos prejudiciais de algas. Outros dados foram, ainda, incluídos, como estação do ano, dimensão e velocidade do rio.

Como a coleta e análise de propriedades químicas do rio pode ser feita de forma automática, rápida e barata e a análise biológica para identificar algas envolve pessoal altamente treinando, além de ser cara e lenta, a obtenção de modelos de predição do crescimento das algas a partir de parâmetros químicos é bastante oportuna.

## Os dados

Existem 2 datasets para esta análise. O primeiro possui 200 amostras com as seguintes variáveis:

* Estação do ano;
* Tamanho do rio;
* Velocidade do rio;
* 8 parâmetros químicos:
  + Valor máximo do pH;
  + Valor mínimo de Oxigênio (O2);
  + Valor máximo de Cloro (Cl);
  + Valor médio de Nitratos (NO-3);
  + Valor médio de Amônia (NH+4);
  + Média de Ortofosfato (PO3-4);
  + Média de Fosfaro (PO4); e,
  + Média de Clorofila.
* Associado a estes dados há a frequência de 7 algas prejudiciais.

O segundo dataset possui 140 amostras com os dados acima, porém sem as frequências dos 7 tipos de algas prejudiciais. Nosso objetivo é fornecer um modelo que permita prever estes valores.

## Instalando o pacote do livro

Para a primeira execução do código, retire o “#” da linha abaixo:

# install.packages('DMwR')

## Carregando e verificando os dados

library(DMwR)

## Warning: package 'DMwR' was built under R version 3.6.3

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

str(algae)

## 'data.frame': 200 obs. of 18 variables:  
## $ season: Factor w/ 4 levels "autumn","spring",..: 4 2 1 2 1 4 3 1 4 4 ...  
## $ size : Factor w/ 3 levels "large","medium",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ speed : Factor w/ 3 levels "high","low","medium": 3 3 3 3 3 1 1 1 3 1 ...  
## $ mxPH : num 8 8.35 8.1 8.07 8.06 8.25 8.15 8.05 8.7 7.93 ...  
## $ mnO2 : num 9.8 8 11.4 4.8 9 13.1 10.3 10.6 3.4 9.9 ...  
## $ Cl : num 60.8 57.8 40 77.4 55.4 ...  
## $ NO3 : num 6.24 1.29 5.33 2.3 10.42 ...  
## $ NH4 : num 578 370 346.7 98.2 233.7 ...  
## $ oPO4 : num 105 428.8 125.7 61.2 58.2 ...  
## $ PO4 : num 170 558.8 187.1 138.7 97.6 ...  
## $ Chla : num 50 1.3 15.6 1.4 10.5 ...  
## $ a1 : num 0 1.4 3.3 3.1 9.2 15.1 2.4 18.2 25.4 17 ...  
## $ a2 : num 0 7.6 53.6 41 2.9 14.6 1.2 1.6 5.4 0 ...  
## $ a3 : num 0 4.8 1.9 18.9 7.5 1.4 3.2 0 2.5 0 ...  
## $ a4 : num 0 1.9 0 0 0 0 3.9 0 0 2.9 ...  
## $ a5 : num 34.2 6.7 0 1.4 7.5 22.5 5.8 5.5 0 0 ...  
## $ a6 : num 8.3 0 0 0 4.1 12.6 6.8 8.7 0 0 ...  
## $ a7 : num 0 2.1 9.7 1.4 1 2.9 0 0 0 1.7 ...

summary(algae)

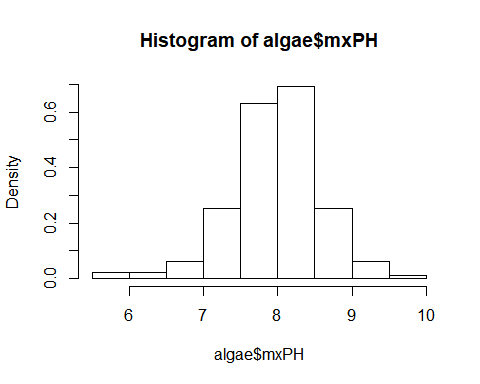
## season size speed mxPH mnO2   
## autumn:40 large :45 high :84 Min. :5.600 Min. : 1.500   
## spring:53 medium:84 low :33 1st Qu.:7.700 1st Qu.: 7.725   
## summer:45 small :71 medium:83 Median :8.060 Median : 9.800   
## winter:62 Mean :8.012 Mean : 9.118   
## 3rd Qu.:8.400 3rd Qu.:10.800   
## Max. :9.700 Max. :13.400   
## NA's :1 NA's :2   
## Cl NO3 NH4 oPO4   
## Min. : 0.222 Min. : 0.050 Min. : 5.00 Min. : 1.00   
## 1st Qu.: 10.981 1st Qu.: 1.296 1st Qu.: 38.33 1st Qu.: 15.70   
## Median : 32.730 Median : 2.675 Median : 103.17 Median : 40.15   
## Mean : 43.636 Mean : 3.282 Mean : 501.30 Mean : 73.59   
## 3rd Qu.: 57.824 3rd Qu.: 4.446 3rd Qu.: 226.95 3rd Qu.: 99.33   
## Max. :391.500 Max. :45.650 Max. :24064.00 Max. :564.60   
## NA's :10 NA's :2 NA's :2 NA's :2   
## PO4 Chla a1 a2   
## Min. : 1.00 Min. : 0.200 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 41.38 1st Qu.: 2.000 1st Qu.: 1.50 1st Qu.: 0.000   
## Median :103.29 Median : 5.475 Median : 6.95 Median : 3.000   
## Mean :137.88 Mean : 13.971 Mean :16.92 Mean : 7.458   
## 3rd Qu.:213.75 3rd Qu.: 18.308 3rd Qu.:24.80 3rd Qu.:11.375   
## Max. :771.60 Max. :110.456 Max. :89.80 Max. :72.600   
## NA's :2 NA's :12   
## a3 a4 a5 a6   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 1.550 Median : 0.000 Median : 1.900 Median : 0.000   
## Mean : 4.309 Mean : 1.992 Mean : 5.064 Mean : 5.964   
## 3rd Qu.: 4.925 3rd Qu.: 2.400 3rd Qu.: 7.500 3rd Qu.: 6.925   
## Max. :42.800 Max. :44.600 Max. :44.400 Max. :77.600   
##   
## a7   
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 1.000   
## Mean : 2.495   
## 3rd Qu.: 2.400   
## Max. :31.600   
##

head(algae)

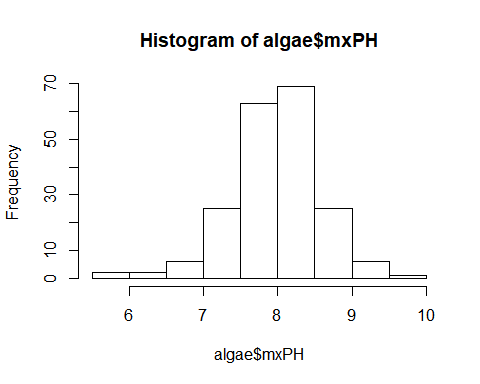
## season size speed mxPH mnO2 Cl NO3 NH4 oPO4 PO4 Chla a1  
## 1 winter small medium 8.00 9.8 60.800 6.238 578.000 105.000 170.000 50.0 0.0  
## 2 spring small medium 8.35 8.0 57.750 1.288 370.000 428.750 558.750 1.3 1.4  
## 3 autumn small medium 8.10 11.4 40.020 5.330 346.667 125.667 187.057 15.6 3.3  
## 4 spring small medium 8.07 4.8 77.364 2.302 98.182 61.182 138.700 1.4 3.1  
## 5 autumn small medium 8.06 9.0 55.350 10.416 233.700 58.222 97.580 10.5 9.2  
## 6 winter small high 8.25 13.1 65.750 9.248 430.000 18.250 56.667 28.4 15.1  
## a2 a3 a4 a5 a6 a7  
## 1 0.0 0.0 0.0 34.2 8.3 0.0  
## 2 7.6 4.8 1.9 6.7 0.0 2.1  
## 3 53.6 1.9 0.0 0.0 0.0 9.7  
## 4 41.0 18.9 0.0 1.4 0.0 1.4  
## 5 2.9 7.5 0.0 7.5 4.1 1.0  
## 6 14.6 1.4 0.0 22.5 12.6 2.9

Vamos obter abaixo um histograma de “mxPH”. O parâmetro “prob = T” nos dá a probabilidade de cada intervalo de valor, caso contrário teríamos a contagem.

hist(algae$mxPH, prob = T)



hist(algae$mxPH)

 O histograma abaixo, com o QQ Plot permite verificar a normalidade dos dados.

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Instale o pacote abaixo, se necessário.

# install.packages('car')

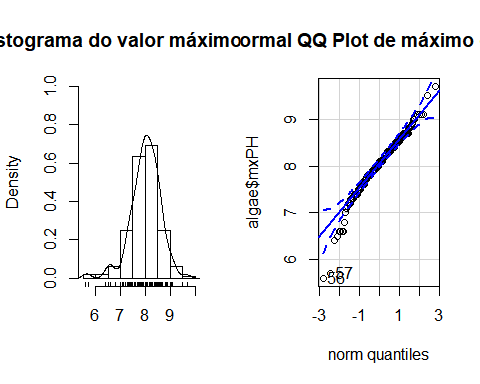
library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.6.3

par(mfrow=c(1,2)) # dividindo o painel do gráfico por 1 linha e 2 colunas  
hist(algae$mxPH, prob = T, xlab = "",  
 main = "Histograma do valor máximo de pH", ylim = 0:1)  
lines(density(algae$mxPH,na.rm = T))  
rug(jitter(algae$mxPH))  
qqPlot(algae$mxPH, main = "Normal QQ Plot de máximo de pH")



## [1] 56 57

par(mfrow=c(1,1))

A parte de baixo do histograna mostra a distribuição dos dados. Ali é possível ver dois valores baixos muito apartados dos outros valores… Possivelmente outliers.

o QQ Plot mostra a distribuição normal (linha contínua azul) e um intervalo de confiança de 95% (linha tracejada azul). Note que alguns valores baixos estão fora do intervalo de confiança da distribuição normal.

Outra possibilidade é criar um Boxplot, que permite uma rápida visualização da distribuição dos dados, a posição da mediana, os quartis e eventuais outliers.

boxplot(algae$oPO4, ylab = "Ortofosfato - oP04")  
rug(jitter(algae$oPO4), side = 2) # "rug" gera os "risquinhos" ao lado do gráfico, mostram o espalhamento dos dados; "jitter" melhora a visualização destes risquinhos, evitando a sobreposição deles; side = 2 é o eixo Y.  
abline(h = mean(algae$oPO4, na.rm = T), lty = 2) # essa linha contendo a média permite a comparação com a mediana, no caso, o descasamento das duas mostra o efeito dos outliers na amostra.

