# 数据挖掘课程作业之海藻数据的分析

计算机学院-2120151056-于畅泳

### 一. 问题描述

某些高浓度的有害藻类对河流生态环境的破坏是一个严重的问题。它们不仅破坏河流的生物,也破坏水质。能够监测并在早期对海藻的繁殖进行预测对提高河流质量是很有必要的。

针对这一问题的预测目标,在大约一年的时间内,在不同时间内 收集了欧洲多条河流的水样。对于每个水样,测定了它们的不同化学 性质以及7种有害藻类的存在频率。在水样收集过程中,也记录了一 些其他特性,如收集的季节、河流大小和水流速度。

### 二.数据说明

有200个水样,每条记录是同一条河流在该年的同一个季节的三个月内收集的水样的平均值。

每条记录由 11 个变量构成, 3 个是标称变量, 分别描述水样收集的季节, 河流大小和河水速度, 剩下的 8 个变量是水样的化学参数:

最大 pH 值 (mxPH)

最小含氧量(mn02)

平均氯化物含量(C1)

平均硝酸盐含量(NO3)

平均氨含量(NH4)

平均正磷酸盐含量(oP04)

平均磷酸盐含量(P04)

平均叶绿素含量(Chla)

a1-a7 为 7 种不同有害藻类在相应水样中的频率数目。

### 三. 分析报告及程序

实验环境: Microsoft Windows 10 家庭中文版 (64 位)
Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz(2592 MHz)
16.00 GB 内存

实验软件: RGui (3.3.0)

实验内容:

### 0. 安装 R 添加包

> install.packages('DMwR')

此添加包中包含实验所需数据,名为'algae',当出现:

下载的二进制程序包在

C:\Users\...(程序包安装路径)里时,则表示安装成功。

## 1. 数据摘要

- > library(DMwR)
- > head(algae)

使用 library 命令载入 DMwR 包,展示数据 algae 的前 6 行,如下图:

```
      > library(DMwR)

      载入需要的程辑包: lattice
      载入需要的程辑包: grid

      > head(algae)
      season size
      speed mxPH mn02
      Cl
      NO3
      NH4
      oPO4
      PO4 Chla
      a1
      a2
      a3
      a4
      a5
      a6
      a7

      1 winter small medium 8.00
      9.8
      60.800
      6.238
      578.000
      105.000
      170.000
      50.0
      0.0
      0.0
      0.0
      0.0
      0.0
      0.0
      34.2
      8.3
      0.0

      2 spring small medium 8.05
      8.0
      577.750
      1.288
      370.000
      428.750
      558.750
      1.3
      1.4
      7.6
      4.8
      1.9
      6.7
      0.0
      2.1

      3 autumn small medium 8.10
      11.4
      40.020
      5.33
      346.667
      125.667
      187.057
      15.6
      3.3
      53.6
      1.9
      0.0
      0.0
      0.0
      9.2

      4 spring small medium 8.07
      4.8
      77.364
      2.302
      98.182
      61.182
      138.700
      1.4
      3.1
      41.0
      18.9
      0.0
      1.4
      0.0
      1.4

      5 autumn small medium 8.06
      9.0
      55.350
      10.416
      233.700
      58.
```

> summary(algae)

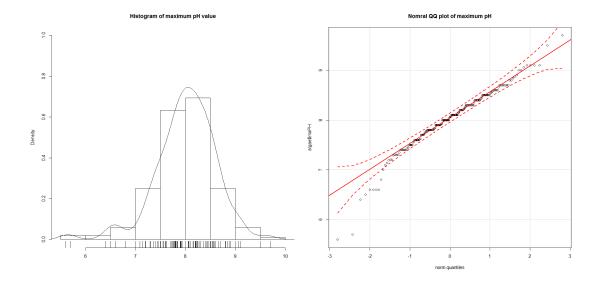
输出数据摘要。对于标称属性,给出每个可能取值的频数。对于数值属性,给出最大、最小、均值、中位数、四分位数及缺失值的个数。

```
> summary(algae)
   season
                 size
                                          mxPH
                                     Min. :5.600
1st Qu.:7.700
                                                      Min. : 1.500
1st Qu.: 7.725
                                                                       Min. : 0.222
1st Qu.: 10.981
                         high :84
low :33
                                                                                          1st Ou.: 1.296
spring:53
            medium:84
                                                                                                           1st Ou.:
 summer:45
             small :71
                         medium:83
                                     Median :8.060
                                                      Median : 9.800
                                                                        Median : 32.730
                                                                                          Median : 2.675
                                                                                                           Median :
                                                                                                                      103.17
                                                               9.118
winter:62
                                            :8.012
                                                      Mean
                                                                                43.636
                                                                                          Mean
                                     Mean
                                                                        Mean
                                                                                                           Mean
                                     3rd Qu.:8.400
                                                      3rd Qu.:10.800
                                                                        3rd Qu.: 57.824
                                                                                          3rd Qu.: 4.446
                                                                                                           3rd Qu.:
                                     Max.
                                                                                          Max. :45.650
                                                      Max. :13.400
                                                                              :391.500
                                                                                                           NA's
                                     NA's
                                                      NA's
                                                                               :10
                                                                                          NA's
                                                                                                 :2
Min. : 0.000
1st Qu.: 0.000
Median : 0.000
                        : 0.000
                                          : 0.000
Min.
                  Min.
                                                     Min.
                  1st Qu.: 0.000
Median : 1.900
                                   1st Qu.: 0.000
Median : 0.000
                                                     1st Qu.: 0.000
Median : 1.000
                           5.064
7.500
        . 1.992
                  Mean
                                   Mean
                                          : 5.964
                                                     Mean
                                    3rd Qu.:
       :44.600
                  Max.
                         :44.400
                                   Max.
                                          :77.600
       oPO4
: 1.00
                           PO4
                                                Chla
                     Min. : 1.00
                                          Min. : 0.200
                                                                                           : 0.000
 Min.
                                                                Min.
                                                                        : 0.00
                                                                                   Min.
                                                                                                         Min.
                     1st Ou.: 41.38
                                                                1st Ou.: 1.50
                                                                                    1st Ou.: 0.000
 1st Ou.: 15.70
                                          1st Ou.: 2.000
                                                                                                         1st Ou.: 0.000
                     Median :103.29
                                          Median : 5.475
                                                                Median: 6.95
 Median : 40.15
                                                                                    Median : 3.000
                                                                                                         Median : 1.550
        : 73.59
                     Mean :137.88
                                                  : 13.971
                                                                        :16.92
                                                                                    Mean
                                                                                            : 7.458
 Mean
                                          Mean
                                                                Mean
                                                                                                         Mean
                                                                                                                    4.309
 3rd Qu.: 99.33
                     3rd Qu.:213.75
                                          3rd Qu.: 18.308
                                                                3rd Qu.:24.80
                                                                                    3rd Qu.:11.375
                                                                                                         3rd Qu.:
        :564.60
                     Max.
                             :771.60
                                          Max.
                                                   :110.456
                                                                Max.
                                                                         :89.80
                                                                                    Max.
                              :2
```

## 2. 数据的可视化

- > library(car)
- > par(mfrow=c(1,2))
- > hist(algae\$mxPH, prob=T, xlab='', main='Histogram of maximum pH value', ylim=0:1)
- > lines(density(algae\$mxPH, na.rm=T))
- > rug(jitter(algae\$mxPH))
- > qqPlot(algae\$mxPH, main='Nomral QQ plot of maximum pH')
- > par(mfrow=c(1,1))

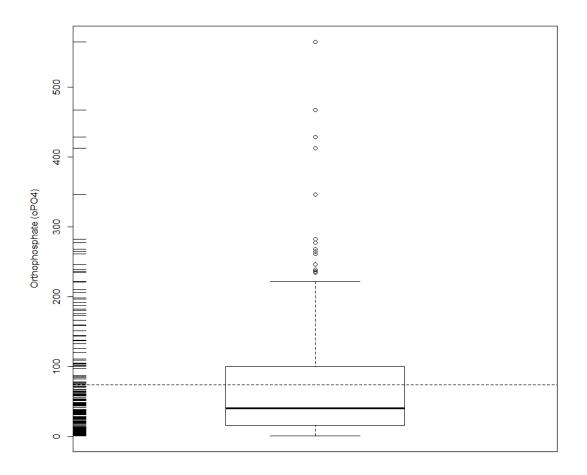
针对数值属性,绘制 mxPH 属性的直方图,用 Q-Q 图检验其分布是否为正态分布。



上图右侧是通过 qqP1ot 得到的 Q-Q 图,它绘制变量值和正态分布的理论分位数的散点图。同时,它给出正态分布的 95%置信区间的带状图(虚线)。从图中可知,变量有几个小的值明显在 95%置信区间之外,它们不服从正态分布。

- > boxplot(algae\$oP04, ylab = 'Orthophosphate (oP04)')
- > rug(jitter(algae\$oP04), side = 2)
- > abline(h = mean(algae\$oPO4, na.rm = T), 1ty = 2)

绘制 oP04 属性的箱图,图中可以看出,oP04 的分布集中在较小的观测值周围,因此分布正偏。大部分水样的 oP04 值比较低,但也有几个水样的观测值比较高,甚至特别高。



针对7种海藻,下面分别绘制其数量与标称变量 size 的条件箱图

- > bwplot(size  $^{\sim}$  a1, data=algae, ylab='River Size',xlab='Algae A1')
- > bwplot(size  $^{\sim}$  a2, data=algae, ylab='River Size',xlab='Algae A2')
- > bwplot(size ~ a3, data=algae, ylab='River Size',xlab='Algae A3')
  - > bwplot(size  $^{\sim}$  a4, data=algae, ylab='River

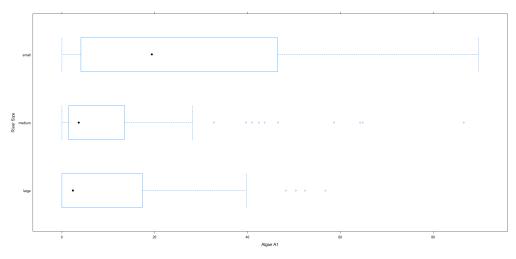
Size', xlab='Algae A4')

> bwplot(size  $^{\sim}$  a5, data=algae, ylab='River Size',xlab='Algae A5')

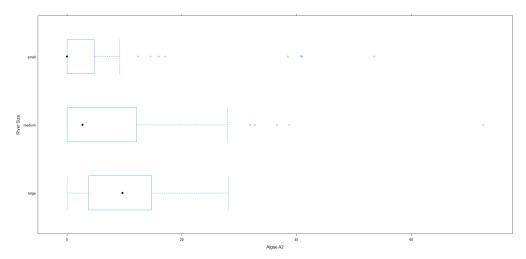
> bwplot(size ~ a6, data=algae, ylab='River Size',xlab='Algae A6')

> bwplot(size  $^{\sim}$  a7, data=algae, ylab='River Size',xlab='Algae A7')

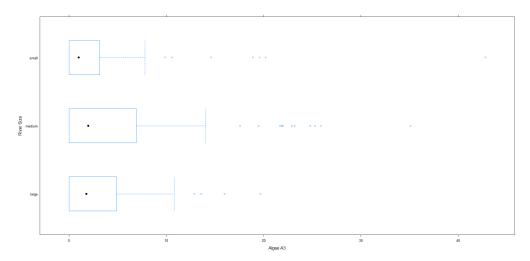
条件箱图 (a1, a2, a3, a4, a5, a6, a7) 分别如下:



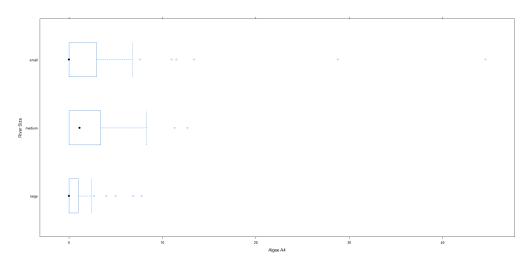
在规模较小的河流中,海藻 a1 的频率较多。



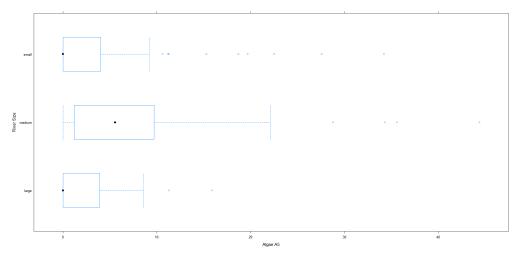
在规模较小和中等的河流中,海藻 a2 的频率较多。



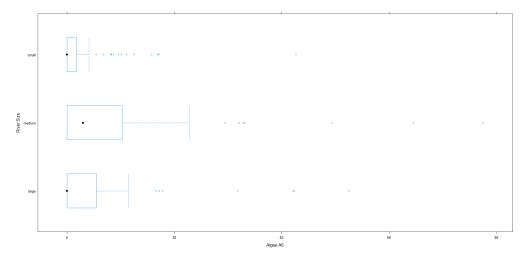
在规模中等的河流中,海藻 a3 的频率较多。



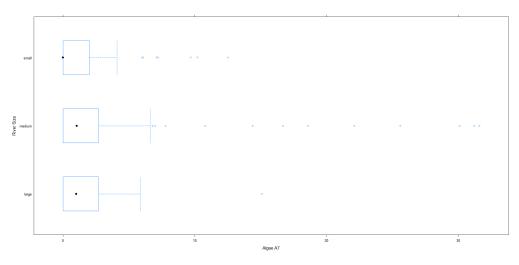
在规模较小和中等的河流中,海藻 a4 的频率较多。



在规模中等的河流中,海藻 a5 的频率较多。



在规模中等的河流中,海藻 a6 的频率较多。



在三种规模的河流中,海藻 a7 的频率相差不大。

# 3. 数据缺失的处理

> algae[!complete.cases(algae),]

有缺失数据的部分如下:

```
> algae[!complete.cases(algae),]
    season
             size speed mxPH mnO2
                                          C1
                                              NO3 NH4
                                                           oPO4
                                                                     PO4
                                                                           Chla
                                                                                                 a4 a5 a6 a7
0.0 2.1 1.4 2.1
                     high 6.80 11.1 9.000 0.630
                                                          4.000
                                                                           2.70 30.3
                                                                                       1.9 0.0
    autumn
             small
                                                   20
                                                                      NA
                                                                                       0.0 0.0
                                                                                                 0.0 0.0
             small
                     high 8.00
                                   NA 1.450 0.810
                                                          2.500
                                                                           0.30
                                                                                75.8
    spring
    winter
             small
                      low
                             NA 12.6 9.000 0.230
                                                          5.000
                                                                   6.000
                                                                           1.10 35.5
                                                                                       0.0 0.0
                                                                                                 0.0 0.0 0.0 0.0
                     high 6.60 10.8
55
    winter
             small
                                          NA 3.245
                                                     10
                                                          1.000
                                                                   6.500
                                                                             NA 24.3
                                                                                       0.0 0.0
                                                                                                 0.0 0.0 0.0 0.0
                                                                             NA 82.7
             small medium 5.60 11.8
                                          NA 2.220
                                                                                       0.0 0.0
                                                                                                 0.0 0.0 0.0 0.0
                                                          1.000
                                                                   1.000
    spring
             small medium 5.70 10.8
                                          NA 2.550
                                                          1.000
                                                                   4.000
                                                                             NA 16.8
                                                                                       4.6 3.9 11.5 0.0 0.0 0.0
                                                                             NA 46.8
NA 46.9
                                                                                       0.0 0.0 28.8 0.0 0.0 0.0 0.0 0.0 0.0 13.4 0.0 0.0 0.0
    spring
             small
                     high 6.60
                                 9.5
                                          NA 1.320
                                                     20
                                                          1.000
                                                                   6.000
59
    summer
             small
                     high 6.60 10.8
                                          NA 2.640
                                                     10
                                                          2.000
                                                                  11.000
             small medium 6.60 11.3
                                          NA 4.170
                                                          1.000
                                                                   6.000
                                                                             NA 47.1
                                                                                       0.0 0.0
    autumn
    spring
             small medium 6.50 10.4
                                          NA 5.970
                                                     10
                                                          2.000
                                                                  14.000
                                                                             NA 66.9
                                                                                       0.0 0.0
                                                                                                 0.0 0.0 0.0 0.0
    summer
             small medium 6.40
                                   NA
                                          NA
                                                NA
                                                     NA
                                                             NA
                                                                  14.000
                                                                             NA 19.4
                                                                                       0.0 0.0
                                                                                                2.0 0.0 3.9 1.7
                     high 7.83 11.7
                                       4.083 1.328
                                                          3.333
    autumn
             small
                                                     18
                                                                   6.667
                                                                             NA 14.4
                                                                                       0.0 0.0
                                                                                                 0.0 0.0 0.0 0.0
116 winter medium
                     high 9.70 10.8 0.222 0.406
                                                     10
                                                         22.444
                                                                  10.111
                                                                             NA 41.0
                                                                                       1.5 0.0
                                                                                                0.0 0.0 0.0 0.0
                     low 9.00 5.8 NA 0.900 142
high 8.00 10.9 9.055 0.825 40
                                         NA 0.900 142 102.000 186.000 68.05 1.7 20.6 1.5
055 0.825 40 21.083 56.091 NA 16.8 19.6 4.0
161 spring
            large
                                                                                                2.2 0.0 0.0 0.0
                                                                  56.091
                                                                                                 0.0 0.0 0.0 0.0
184 winter
            large
            large medium 8.00
> nrow(algae[!complete.cases(algae),])
[1] 16
```

> nrow(algae[!complete.cases(algae),])

[1] 16

缺失行数为16行

(1) 将缺失部分剔除

> algae <-na.omit (algae)

直接将有缺失的部分剔除,如下图:

```
> algae(-na.omit(algae)
> algae[!complete.cases(algae),]
[1] season size speed mxPH mnO2 Cl NO3 NH4 oPO4 PO4 Chla a1 a2 a3 a4 a5 a6 a'
<0 行> (或0-长度的row.names)
```

(2) 用最高频率值来填补缺失值

使用平均值填补数据,如使用平均值数值填补 mxPH:

使用中位数填补数据,如使用中位数数值填补 Ch1a:

```
> data(algae)
  algae[!complete.cases(algae),]
                                    Cl
                                         NO3 NH4
                                                    oPO4
                                                             PO4
                                                                 Chla
                                                                         a1
                                                                             a2 a3
                                                                                      a4 a5 a6 a7
            size speed mxPH mnO2
   season
           small high 6.80 11.1 9.000 0.630 20
                                                                  2.70 30.3 1.9 0.0 0.0 2.1 1.4 2.1
                                                   4.000
                   high 8.00 NA 1.450 0.810 10
                                                                             0.0 0.0
   spring
           small
                                                   2,500
                                                           3.000
                                                                  0.30 75.8
                                                                                     0.0 0.0 0.0 0.0
                         NA 12.6 9.000 0.230 10
   winter
           small
                    low
                                                   5.000
                                                           6.000
                                                                  1.10 35.5 0.0 0.0
                                                                                     0.0 0.0 0.0 0.0
                  high 6.60 10.8
   winter
           small
                                    NA 3.245 10
                                                   1.000
                                                           6.500
                                                                    NA 24.3
                                                                             0.0 0.0
                                                                                     0.0 0.0 0.0 0.0
           small medium 5.60 11.8
                                    NA 2.220
                                                   1.000
                                                           1.000
                                                                             0.0 0.0
                                                                                     0.0 0.0 0.0 0.0
   spring
           small medium 5.70 10.8
                                    NA 2.550 10
                                                   1.000
                                                           4.000
                                                                    NA 16.8
                                                                             4.6 3.9 11.5 0.0 0.0 0.0
   autumn
58
   spring
           small high 6.60 9.5
                                    NA 1.320 20
                                                   1.000
                                                           6.000
                                                                    NA 46.8 0.0 0.0 28.8 0.0 0.0 0.0
                  high 6.60 10.8
                                                   2.000 11.000
   summer
           small
                                    NA 2.640 10
                                                                    NA 46.9 0.0 0.0 13.4 0.0 0.0 0.0
   autumn small medium 6.60 11.3
                                    NA 4.170
                                             10
                                                  1.000
                                                           6.000
                                                                    NA 47.1
                                                                             0.0 0.0 0.0 0.0 1.2 0.0
   spring
           small medium 6.50 10.4
                                    NA 5.970 10
                                                  2.000
                                                          14.000
                                                                    NA 66.9
           small medium 6.40
                             NA
                                    NA NA NA
                                                     NA
                                                          14.000
                                                                    NA 19.4
                                                                             0.0 0.0
                                                                                     2.0 0.0 3.9 1.7
   summer
63
   autumn small high 7.83 11.7 4.083 1.328
                                              18
                                                   3.333
                                                           6.667
                                                                    NA 14.4 0.0 0.0
                                                                                     0.0 0.0 0.0 0.0
116 winter medium high 9.70 10.8 0.222 0.406
                                             10
                                                 22.444
                                                          10.111
                                                                    NA 41.0
                                                                            1.5 0.0
                                                                                     0.0 0.0 0.0 0.0
                                    NA 0.900 142 102.000 186.000 68.05 1.7 20.6 1.5
161 spring large
                   low 9.00 5.8
                                                                                     2.2 0.0 0.0 0.0
                 high 8.00 10.9 9.055 0.825 40 21.083
                                                                    NA 16.8 19.6 4.0
           large
                                                          56.091
                                                                                     0.0 0.0 0.0 0.0
199 winter large medium 8.00 7.6
                                   NA
                                         NA NA
                                                      NA
                                                             NA
                                                                    NA 0.0 12.5 3.7 1.0 0.0 0.0 4.9
> algae[is.na(algae$Chla), 'Chla']<-median(algae$Chla, na.rm = T)</pre>
> algae[c(28,38,48,55,56,57,58,59,60,61,62,63,116,161,184,199),]
                                    Cl NO3 NH4
   season size speed mxPH mnO2
                                                   oP04
                                                                   Chla
                                                                         a1
                                                                              a2 a3
                                                                                       a4 a5 a6 a7
                                                             NA
                  high 6.80 11.1 9.000 0.630 20
                                                                  2.700 30.3 1.9 0.0 0.0 2.1 1.4 2.1
   autumn
           small
                                                   4.000
   spring
           small high 8.00 NA 1.450 0.810 10
                                                   2.500
                                                           3.000
                                                                  0.300 75.8
                                                                             0.0 0.0
                                                                                      0.0 0.0 0.0 0.0
                   low NA 12.6 9.000 0.230 10
                                                           6.000
48
   winter
           small
                                                   5.000
                                                                  1.100 35.5
                                                                              0.0 0.0 0.0 0.0 0.0 0.0
                  high 6.60 10.8
                                    NA 3.245 10
                                                   1.000
                                                           6.500
                                                                  5.475 24.3
           small
                                                                              0.0 0.0 0.0 0.0 0.0 0.0
   winter
           small medium 5.60 11.8
                                    NA 2.220
                                                           1.000
                                                                  5.475 82.7
                                                                              0.0 0.0 0.0 0.0 0.0 0.0
   spring
                                                   1.000
   autumn
           small medium 5.70 10.8
                                    NA 2.550 10
                                                   1.000
                                                           4.000
                                                                  5.475 16.8
                                                                              4.6 3.9 11.5 0.0 0.0 0.0
   spring
           small high 6.60 9.5
                                    NA 1.320 20
                                                   1.000
                                                           6.000
                                                                  5.475 46.8
                                                                              0.0 0.0 28.8 0.0 0.0 0.0
                   high 6.60 10.8
                                                                  5.475 46.9
                                    NA 2.640 10
                                                  2.000
                                                         11.000
                                                                              0.0 0.0 13.4 0.0 0.0 0.0
   summer
           small
   autumn
           small medium 6.60 11.3
                                     NA 4.170
                                                   1.000
                                                           6.000
                                                                  5.475 47.1
                                                                              0.0 0.0
           small medium 6.50 10.4
                                     NA 5.970 10
                                                  2.000
                                                          14.000
                                                                  5.475 66.9
                                                                              0.0 0.0 0.0 0.0 0.0 0.0
                                                          14.000
   summer
           small medium 6.40
                             NA
                                    NΔ
                                         NA NA
                                                     NΔ
                                                                  5.475 19.4
                                                                              0.0 0.0 2.0 0.0 3.9 1.7
   autumn small high 7.83 11.7 4.083 1.328 18 winter medium high 9.70 10.8 0.222 0.406 10
                                                   3.333
                                                           6.667
                                                                  5.475 14.4 0.0 0.0 0.0 0.0 0.0 0.0
63
                                             10
                                                 22.444
                                                          10.111
                                                                  5.475 41.0
116 winter medium
                                                                             1.5 0.0 0.0 0.0 0.0 0.0
161 spring large
                    low 9.00
                             5.8
                                    NA 0.900 142 102.000 186.000 68.050 1.7 20.6 1.5 2.2 0.0 0.0 0.0
                 low 9.00 5.8 NA U.900 172 102.000 101111
high 8.00 10.9 9.055 0.825 40 21.083 56.091
                                                                 5.475 16.8 19.6 4.0 0.0 0.0 0.0 0.0
184 winter
           large
199 winter large medium 8.00 7.6
                                    NA
                                         NA NA
                                                      NA
                                                             NA 5.475 0.0 12.5 3.7 1.0 0.0 0.0 4.9
```

### 使用数据的中心趋势来填补数据:

```
algae[!complete.cases(algae),]
    season
             size speed mxPH mnO2
                                      Cl
                                           NO3 NH4
                                                       oPO4
                                                                     Chla
28
    autumn
            small
                    high 6.80 11.1 9.000 0.630
                                                      4.000
                                                                     2.70 30.3
                                                                                1.9 0.0
                                                                                         0.0 2.1 1.4 2.1
                                                20
                                                                 NA
38
    spring
            small
                    high 8.00
                               NA 1.450 0.810
                                                10
                                                      2.500
                                                              3.000
                                                                     0.30 75.8
                                                                                0.0 0.0
                                                                                         0.0 0.0 0.0 0.0
                          NA 12.6 9.000 0.230
48
                                                10
                                                      5.000
                                                              6.000
                                                                     1.10 35.5
                                                                                0.0 0.0
                                                                                         0.0 0.0 0.0 0.0
   winter
            small
                     low
            small
                    high 6.60 10.8
                                      NA 3.245
                                                10
                                                      1.000
                                                              6.500
                                                                       NA 24.3
                                                                                0.0 0.0
                                                                                         0.0
    winter
            small medium 5.60 11.8
                                                      1.000
                                                              1.000
                                                                       NA 82.7
                                                                                0.0 0.0 0.0 0.0 0.0 0.0
    spring
                                      NA 2.220
    autumn
57
            small medium 5.70 10.8
                                      NA 2.550
                                                10
                                                      1,000
                                                              4.000
                                                                       NA 16.8
                                                                                4.6 3.9 11.5 0.0 0.0 0.0
                                                      1.000
58
   spring
           small
                   high 6.60 9.5
                                      NA 1.320
                                                20
                                                              6.000
                                                                       NA 46.8
                                                                                0.0 0.0 28.8 0.0 0.0 0.0
                    high 6.60 10.8
                                                            11.000
59
   summer
           small
                                      NA 2.640
                                                10
                                                      2.000
                                                                       NA 46.9
                                                                                0.0 0.0 13.4 0.0 0.0 0.0
                                                                       NA 47.1
            small medium 6.60 11.3
                                      NA 4.170
                                                      1.000
                                                              6.000
                                                                                0.0 0.0
    spring
            small medium 6.50 10.4
                                      NA 5.970
                                                10
                                                     2.000
                                                            14.000
                                                                       NA 66.9
                                                                                0.0 0.0 0.0 0.0 0.0 0.0
61
62
   summer
           small medium 6.40
                               NA
                                      NA
                                           NA
                                                NA
                                                        NA
                                                            14.000
                                                                       NA 19.4
                                                                                0.0 0.0 2.0 0.0 3.9 1.7
                                         1.328
                                                      3.333
                    high 7.83 11.7
                                   4.083
                                                18
                                                              6.667
                                                                       NA 14.4
                                                                                0.0 0.0
                                                                                         0.0 0.0 0.0 0.0
63
   autumn
           small
                    high 9.70 10.8 0.222
116 winter medium
                                         0.406
                                                10
                                                    22.444
                                                             10.111
                                                                       NA 41.0
                                                                               1.5 0.0
                                                                                         0.0 0.0 0.0 0.0
161 spring
           large
                     low 9.00 5.8
                                      NA 0.900 142 102.000 186.000 68.05
                                                                          1.7 20.6 1.5
                                                                                         2.2 0.0 0.0 0.0
184 winter
           large
                   high 8.00 10.9 9.055 0.825
                                                40
                                                    21.083
                                                            56.091
                                                                       NA 16.8 19.6 4.0 0.0 0.0 0.0 0.0
199 winter
           large medium 8.00 7.6
                                      NA
                                            NA
                                                NA
                                                        NA
                                                                 NA
                                                                       NA 0.0 12.5 3.7
                                                                                         1.0 0.0 0.0 4.9
> algae<-algae[-manyNAs(algae),]
> algae<-centralImputation(algae)
> algae[c(28,38,48,55,56,57,58,59,60,61,62,115,160,183),]
                                                                       Chla
    season
            size speed mxPH mnO2
                                       Cl
                                           NO3 NH4
                                                       oPO4
                                                                 PO4
                                                                             a1
                                                                                   a2 a3
                                                                                            a4 a5 a6 a7
                   high 6.80 11.1 9.000 0.630 20
                                                       4.000 104.000
                                                                      2.700 30.3
                                                                                  1.9 0.0 0.0 2.1 1.4 2.1
   autumn
           small
   spring
           small
                    high 8.00
                               9.8
                                    1.450 0.810
                                                                      0.300 75.8
    winter
            small
                     low 8.06 12.6 9.000 0.230
                                                 10
                                                       5.000
                                                               6.000
                                                                      1.100 35.5
                                                                                  0.0 0.0
                                                                                           0.0 0.0 0.0 0.0
48
55
   winter
            small
                    high 6.60 10.8 32.730 3.245
                                                 10
                                                       1.000
                                                               6.500
                                                                      5.475 24.3
                                                                                  0.0 0.0
                                                                                           0.0 0.0 0.0 0.0
                                                       1.000
56
    spring
            small medium 5.60 11.8 32.730 2.220
                                                  5
                                                               1.000
                                                                      5.475 82.7
                                                                                  0.0 0.0 0.0 0.0 0.0 0.0
    autumn
            small medium 5.70 10.8 32.730 2.550
                                                  10
                                                       1.000
                                                               4.000
                                                                      5.475 16.8
                                                                                  4.6 3.9 11.5 0.0 0.0 0.0
    spring
            small
                    high 6.60 9.5 32.730 1.320
                                                 20
                                                       1.000
                                                               6.000
                                                                      5.475 46.8
                                                                                  0.0 0.0 28.8 0.0 0.0 0.0
59
    summer
            small
                    high 6.60 10.8 32.730 2.640
                                                 10
                                                      2.000
                                                             11.000
                                                                      5.475 46.9
                                                                                  0.0 0.0 13.4 0.0 0.0 0.0
            small medium 6.60 11.3 32.730 4.170
60
   autumn
                                                 10
                                                      1.000
                                                              6.000
                                                                      5.475 47.1
                                                                                  0.0 0.0 0.0 0.0 1.2 0.0
            small medium 6.50 10.4 32.730 5.970
                                                              14.000
                                                                      5.475 66.9
61
                                                 10
                                                      2.000
                                                                                  0.0 0.0
                                                                                           0.0 0.0 0.0 0.0
   spring
                                   4.083 1.328
                                                      3.333
                                                                      5.475 14.4
            small
                    high 7.83 11.7
                                                 18
                                                               6.667
                                                                                  0.0 0.0
116 winter medium
                    high 9.70 10.8 0.222 0.406
                                                 10 22.444
                                                              10.111
                                                                      5.475 41.0
                                                                                 1.5 0.0 0.0 0.0 0.0 0.0
                     low 9.00 5.8 32.730 0.900 142 102.000 186.000 68.050 1.7 20.6 1.5 2.2 0.0 0.0 0.0
161 spring
           large
                    high 8.00 10.9 9.055 0.825
                                                             56.091 5.475 16.8 19.6 4.0 0.0 0.0 0.0 0.0
184 winter
                                                 40
                                                     21.083
           large
```

### (3) 通过属性的相关关系来填补缺失值

```
> symnum(cor(algae[, 4:18], use='complete.obs'))
```

获得变量之间的相关值矩阵,如下图:

```
> symnum(cor(algae[,4:18],use='complete.obs'))
     mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH 1
mn<sub>02</sub>
         1
C1
             1
NO3
                1
NH4
                    1
oPO4
                        1
PO4
                          1
Chla
                            1
a1
                                1
a2
                                    1
a3
                                       1
a4
                                           1
a5
                                               1
                                                  1
a6
a7
                                                      1
attr(,"legend")
[1] 0 \ ' 0.3 \.'
                     0.6
                          ١,,,
                               0.8 '+' 0.9 '*' 0.95 'B' 1
```

由图可知, P04与 Opo4之间相关值很高。

> lm(PO4~oPO4, data=algae)

获得两个变量之间的线性关系,并通过线性关系 P04 = 42.897 + 1.293 \* 0po4 来填补数据:

```
> data(algae)
> algae[28,]
  season size speed mxPH mnO2 Cl NO3 NH4 oPO4 PO4 Chla
                                                         a1 a2 a3 a4 a5 a6 a7
28 autumn small high 6.8 11.1 9 0.63 20 4 NA 2.7 30.3 1.9 0 0 2.1 1.4 2.1
> algae<-algae[-manyNAs(algae),]
> lm(PO4~oPO4.data=algae)
lm(formula = PO4 ~ oPO4, data = algae)
Coefficients:
                  oPO4
(Intercept)
    42.897
> algae[28,'PO4']<-42.897+1.293*algae[28,'oPO4']
> algae[28,]
  season size speed mxPH mnO2 Cl NO3 NH4 oPO4
                                                PO4 Chla a1 a2 a3 a4 a5 a6 a7
                                            4 48.069 2.7 30.3 1.9 0 0 2.1 1.4 2.1
28 autumn small high 6.8 11.1 9 0.63 20
```

### (4) 通过数据对象之间的相似性来填补缺失值

我们通过寻找与含有缺失值的数据最相似的 10 个水样,并用它们来填充缺失值。通过 knnImputation()函数用中位数来填补缺失值。

> algae <-knnImputation(algae, k=10, meth='median')

```
> data(algae)
> algae<-algae[-manyNAs(algae),]
  algae[!complete.cases(algae),]
            size speed mxPH mnO2
                                                                             a2 a3
                                     Cl
                                         NO3 NH4
                                                    oPO4
   season
                                                             NA
           small
                   high 6.80 11.1 9.000 0.630
                                                                  2.70 30.3
                                                                            1.9 0.0 0.0 2.1 1.4 2.1
   autumn
                                                   4.000
   spring
           small
                   high 8.00 NA 1.450 0.810
                                              10
                                                   2.500
                                                           3,000
                                                                  0.30 75.8
                                                                            0.0 0.0 0.0 0.0 0.0 0.0
                   low NA 12.6 9.000 0.230
                                                           6.000
                                                   5.000
                                                                  1.10 35.5
                                                                                     0.0 0.0 0.0 0.0
   winter
           small
                                              10
                                                                            0.0 0.0
   winter
                   high 6.60 10.8
                                                   1.000
                                                   1.000
           small medium 5.60 11.8
                                    NA 2.220
                                                           1.000
                                                                   NA 82.7
                                                                            0.0 0.0 0.0 0.0 0.0 0.0
   spring
   autumn
           small medium 5.70 10.8
                                    NA 2.550 10
                                                   1.000
                                                           4,000
                                                                    NA 16.8
                                                                             4.6 3.9 11.5 0.0 0.0 0.0
           small high 6.60 9.5
small high 6.60 10.8
                                                           6.000
   spring
                                    NA 1.320 20
                                                   1.000
                                                                    NA 46.8
                                                                            0.0 0.0 28.8 0.0 0.0 0.0
   summer
                                    NA 2.640 10
                                                          11.000
                                                                    NA 46.9
                                                                            0.0 0.0 13.4 0.0 0.0 0.0
                                                           6.000
           small medium 6.60 11.3
                                     NA 4.170
                                                   1.000
                                                                    NA 47.1
                                                                            0.0 0.0
   autumn
                                              10
   spring small medium 6.50 10.4
                                    NA 5.970 10
                                                   2.000
                                                          14,000
                                                                    NA 66.9
                                                                            0.0 0.0 0.0 0.0 0.0 0.0
           small high 7.83 11.7 4.083 1.328
                                              18
                                                   3.333
                                                           6.667
                                                                    NA 14.4
                                                                            0.0 0.0 0.0 0.0 0.0 0.0
63
   autumn
116 winter medium
                  high 9.70 10.8 0.222 0.406
                                              10 22.444
                                                          10.111
                                                                    NA 41.0
                                                                            1.5 0.0
                    low 9.00 5.8
                                    NA 0.900 142 102.000 186.000 68.05 1.7 20.6 1.5
161 spring large
                  high 8.00 10.9 9.055 0.825 40 21.083
184 winter large
                                                         56.091
                                                                   NA 16.8 19.6 4.0 0.0 0.0 0.0 0.0
> algae<-knnImputation(algae, k=10, meth='median')
> algae[c(28,38,48,55,56,57,58,59,60,61,62,115,160,183),]
                                           NO3 NH4
            size speed mxPH mnO2
                                       C1
                                                       oPO4
                                                               PO4 Chla
                  high 6.80 11.10 9.0000 0.630 20
                                                                     2.70 30.3 1.9 0.0 0.0 2.1 1.4 2.1
   autumn small
                                                     4.000 14.300
                  high 8.00 10.95
                                                                     0.30 75.8 0.0 0.0
                                   1.4500 0.810 10
                                                             3.000
                                                                                        0.0 0.0 0.0 0.0
   spring
           small
                                                      2.500
   winter
           small
                    low 7.75 12.60
                                    9.0000 0.230 10
                                                                     1.10 35.5
                                                              6.500
           small
                  high 6.60 10.80
                                   5.6500 3.245 10
                                                      1.000
                                                                     0.55 24.3
                                                                               0.0 0.0
   winter
                                                                                        0.0 0.0 0.0 0.0
   spring
           small medium 5.60 11.80
                                   7.8450 2.220
                                                      1.000
                                                             1.000
                                                                     0.60 82.7
                                                                               0.0 0.0
                                                                                       0.0 0.0 0.0 0.0
                                   8.8450 2.550 10
   autumn
           small medium 5.70 10.80
                                                      1.000
                                                              4.000
                                                                     0.55 16.8
                                                                               4.6 3.9 11.5 0.0 0.0 0.0
           small high 6.60 9.50 11.4335 1.320 20
                                                     1.000
                                                             6.000
                                                                     1.85 46.8
   spring
                   high 6.60 10.80
                                   7.0000 2.640 10
                                                            11.000
                                                                     0.60 46.9
                                                                               0.0 0.0 13.4 0.0 0.0 0.0
    summer
           small
                                                      2.000
   autumn
           small medium 6.60 11.30
                                   7.8450 4.170 10
                                                      1.000
                                                              6.000
                                                                     0.85 47.1
                                                                               0.0 0.0
                                                                                        0.0 0.0 1.2 0.0
           small medium 6.50 10.40
                                   6.9815 5.970
                                                10
                                                      2.000
                                                            14.000
                                                                     0.95 66.9
                                                                              0.0 0.0
                                                                                       0.0 0.0 0.0 0.0
   spring
   autumn
           small high 7.83 11.70
                                   4.0830 1.328
                                                      3.333
                                                             6.667
                                                                     0.90 14.4 0.0 0.0
116 winter medium high 9.70 10.80 0.2220 0.406 10 22.444
                                                            10.111
                                                                    1.35 41.0 1.5 0.0 0.0 0.0 0.0 0.0
161 spring large
                    low 9.00 5.80 55.5415 0.900 142 102.000 186.000 68.05
                                                                          1.7 20.6 1.5
                                                                                       2.2 0.0 0.0 0.0
184 winter large high 8.00 10.90 9.0550 0.825 40 21.083 56.091 9.50 16.8 19.6 4.0 0.0 0.0 0.0 0.0
```