

对海藻数据的分析

海藻数据的分析

1. 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。

2. 数据说明

下载数据: [地址](#)

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(mnO2)
- 平均氯化物含量(Cl)
- 平均硝酸盐含量(NO3)
- 平均氨含量(NH4)
- 平均正磷酸盐含量(oPO4)
- 平均磷酸盐含量(PO4)
- 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。

3. 数据分析要求

3.1 数据可视化和摘要

数据摘要

- 对标称属性，给出每个可能取值的频数，
- 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。

数据的可视化

针对数值属性，

- 绘制直方图，如 mxPH，用 qq 图检验其分布是否为正态分布。
- 绘制盒图，对离群值进行识别

对 7 种海藻，分别绘制其数量与标称变量，如 size 的条件盒图

3.2 数据缺失的处理

分别使用下列四种策略对缺失值进行处理：

- 将缺失部分剔除
- 用最高频率值来填补缺失值
- 通过属性的相关关系来填补缺失值
- 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

4. 提交内容

- 分析报告
- 分析程序
- 预处理后的数据集

1. 数据摘要

获取数据的描述性统计摘要是获取数据统计特性的一个重要方法。我们通过 R 语言的 `Summary` 函数对数据摘要进行分析，这个函数

给出了数据的统计特征概括，对于标称变量，他给出了每个取值变量的频数。对于数值型变量，R 为我们提供了四分之一位数、中位数、均值、四分之三位数以及极值等一系列信息。所得结果如下：

```

season      size      speed      mxPH      mnO2      CL
autumn:40   large :45   high :84   Min.   :5.600   Min.   : 1.500   Min.   : 0.222
spring:53   medium:84   low  :33   1st Qu.:7.700   1st Qu.: 7.725   1st Qu.: 10.981
summer:45   small :71   medium:83   Median :8.060   Median : 9.800   Median : 32.730
winter:62                                Mean   :8.012   Mean   : 9.118   Mean   : 43.636
                                3rd Qu.:8.400   3rd Qu.:10.800   3rd Qu.: 57.824
                                Max.    :9.700   Max.    :13.400   Max.    :391.500
                                NA's    :1      NA's    :2      NA's    :10

      NO3      NH4      oPO4      PO4      Chla
Min.   : 0.050   Min.   : 5.00   Min.   : 1.00   Min.   : 1.00   Min.   : 0.200
1st Qu.: 1.296   1st Qu.: 38.33   1st Qu.: 15.70   1st Qu.: 41.38   1st Qu.: 2.000
Median : 2.675   Median : 103.17   Median : 40.15   Median :103.29   Median : 5.475
Mean   : 3.282   Mean   : 501.30   Mean   : 73.59   Mean   :137.88   Mean   : 13.971
3rd Qu.: 4.446   3rd Qu.: 226.95   3rd Qu.: 99.33   3rd Qu.:213.75   3rd Qu.: 18.308
Max.   :45.650   Max.   :24064.00   Max.   :564.60   Max.   :771.60   Max.   :110.456
NA's   :2        NA's   :2        NA's   :2        NA's   :2        NA's   :12

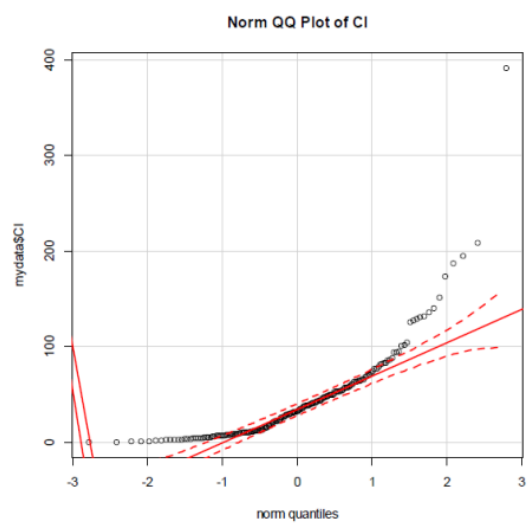
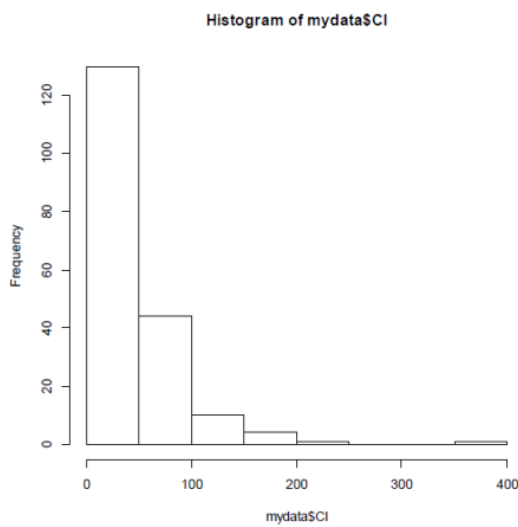
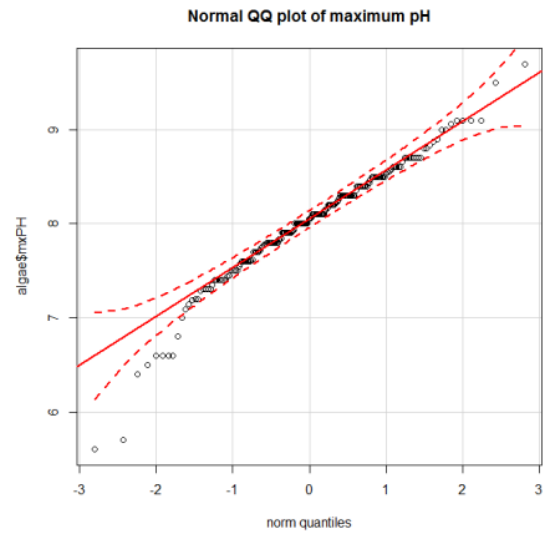
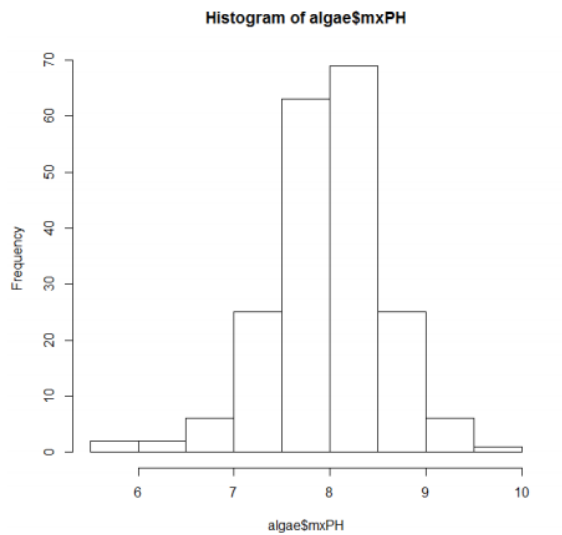
      a1      a2      a3      a4      a5
Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 1.50   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median : 6.95   Median : 3.000   Median : 1.550   Median : 0.000   Median : 1.900
Mean   :16.92   Mean   : 7.458   Mean   : 4.309   Mean   : 1.992   Mean   : 5.064
3rd Qu.:24.80   3rd Qu.:11.375   3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500
Max.   :89.80   Max.   :72.600   Max.   :42.800   Max.   :44.600   Max.   :44.400

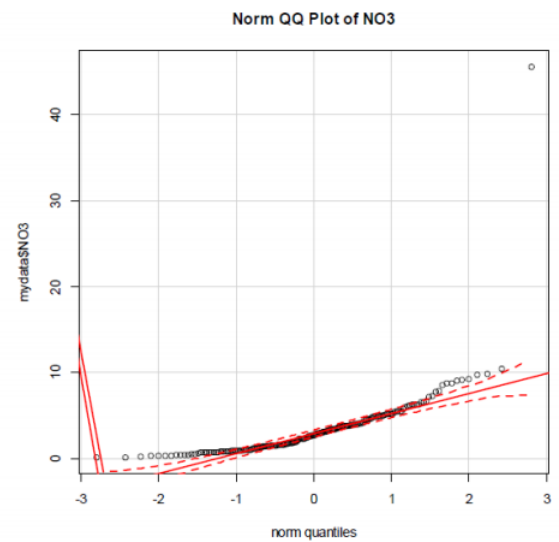
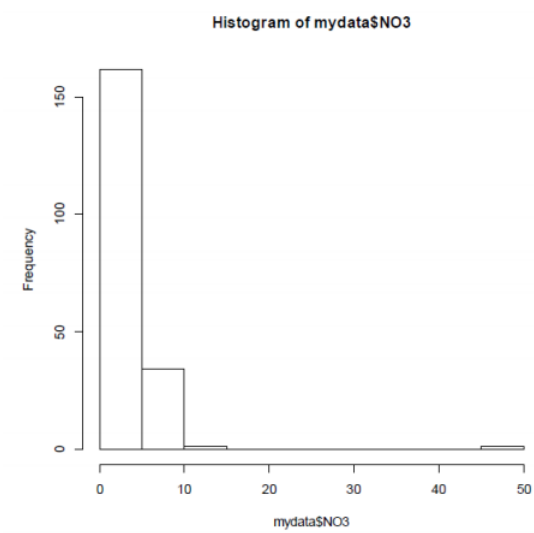
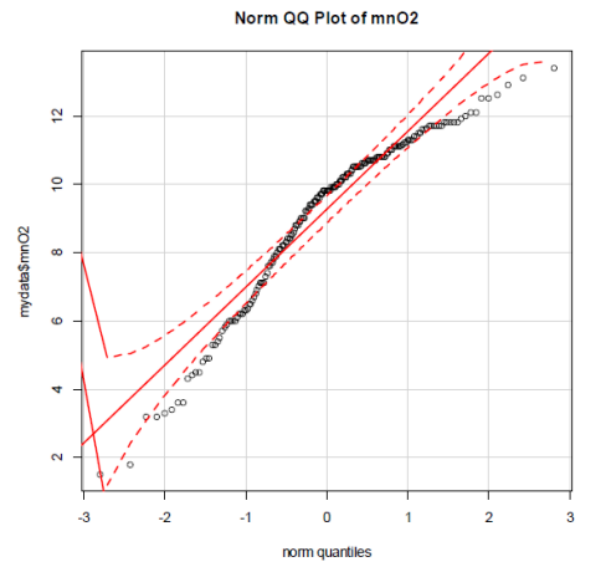
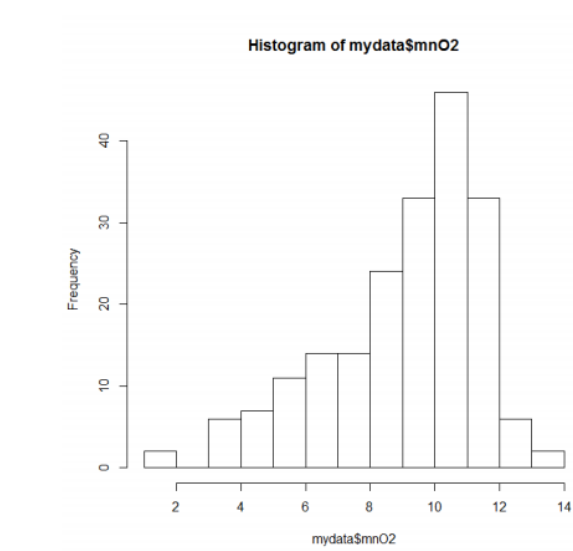
      a6      a7
Min.   : 0.000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.000
Median : 0.000   Median : 1.000
Mean   : 5.964   Mean   : 2.495
3rd Qu.: 6.925   3rd Qu.: 2.400
Max.   :77.600   Max.   :31.600

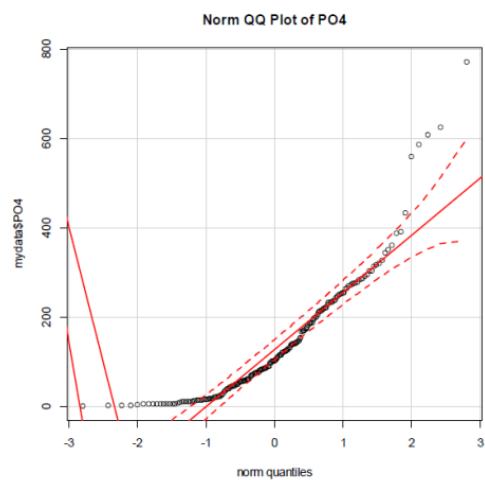
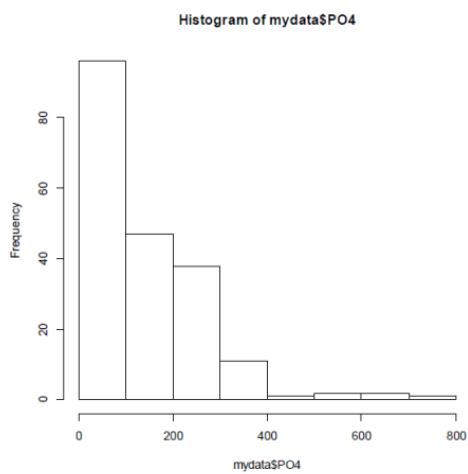
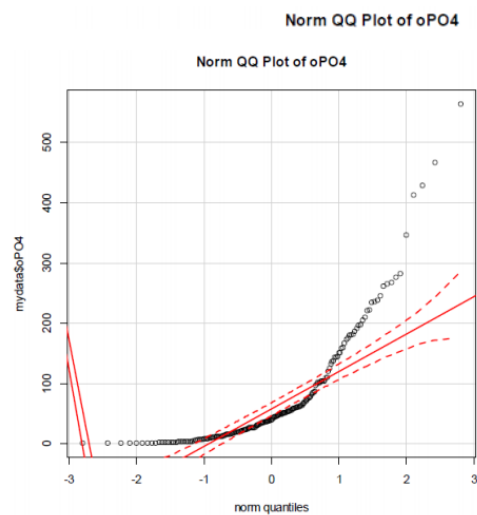
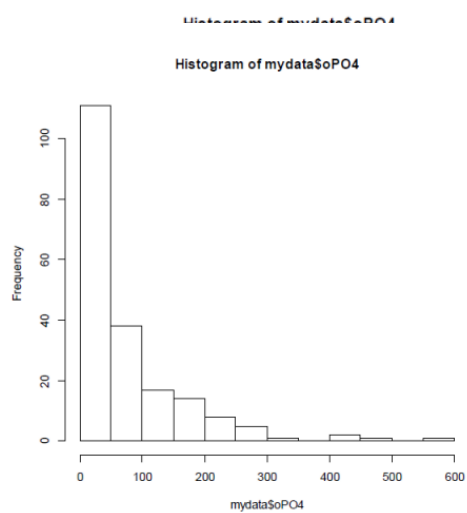
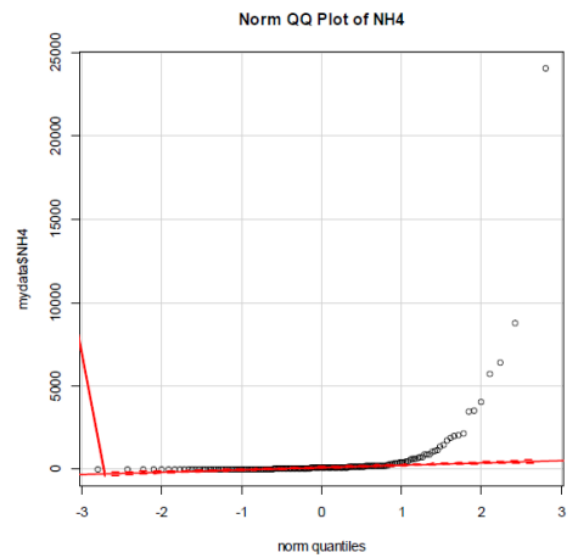
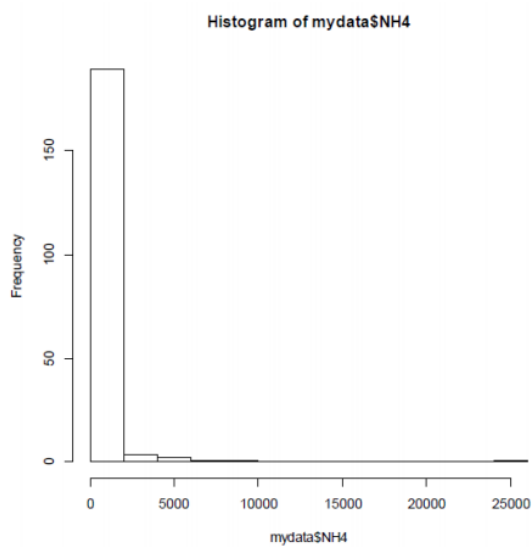
```

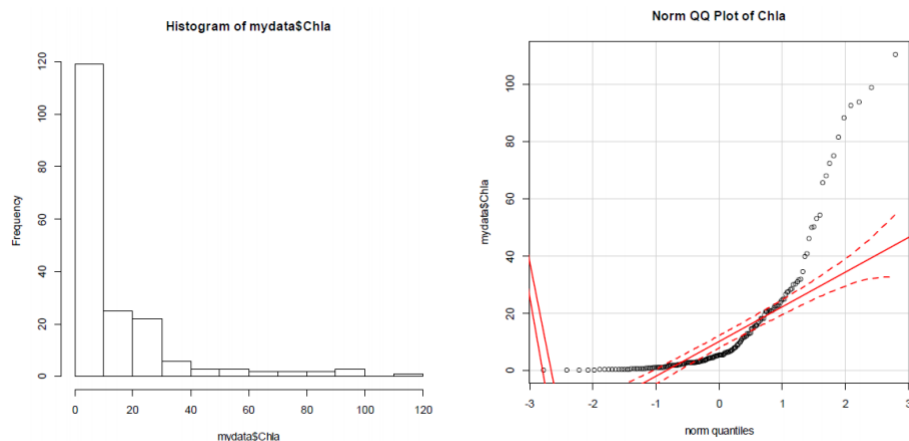
2、数据可视化

(1) 每个数值属性 (mxPH、 Cl、 mnO2、 NO3、 NH4、
oPO4、
PO4、 Chla) 的直方图和 QQ 图如下:



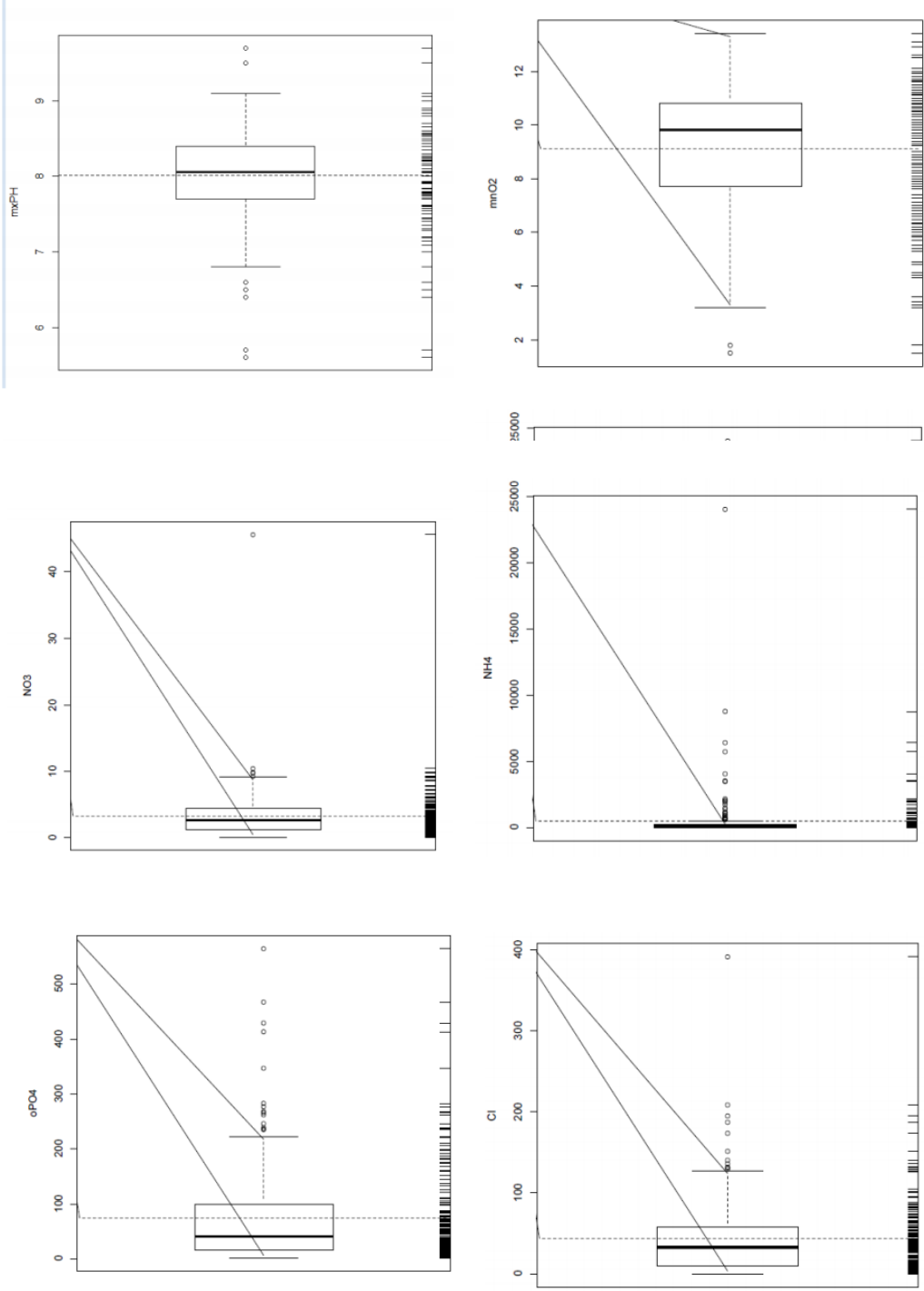


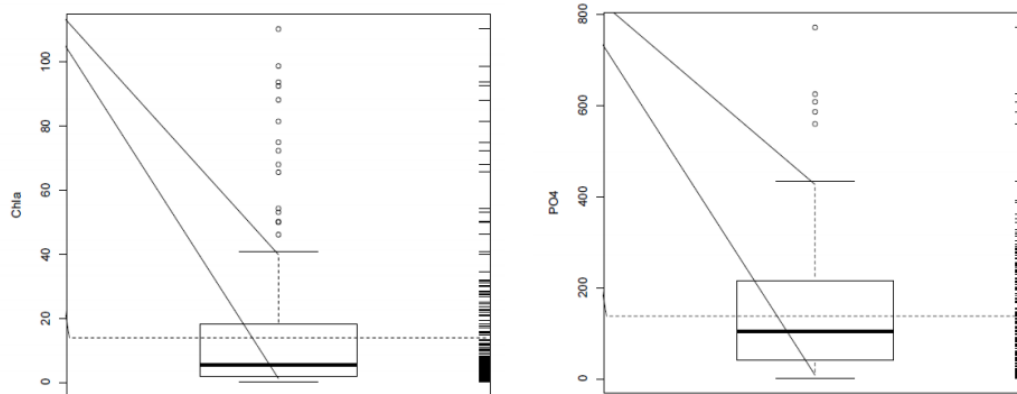




以上绘制的直方图横轴是其分布区间，纵轴是其频数； QQ 图中，红色实线为其 QQ 线， 虚线为 95%置信度的置信区间。 结果表明直方图显示变量 mxPH 的分布非常接近正态分布，它的值大多数都集中在变量的均值附近； QQ 图绘制了变量值与正态分布的理论分位数的散点图，同时他给出正态分布的 95%的置信区间的带状图， 除去有几个小的值明显在 95%置信区间之外，基本服从正态分布。 由数据图表中离开了 95%的置信区间，所以不满足正态分布。

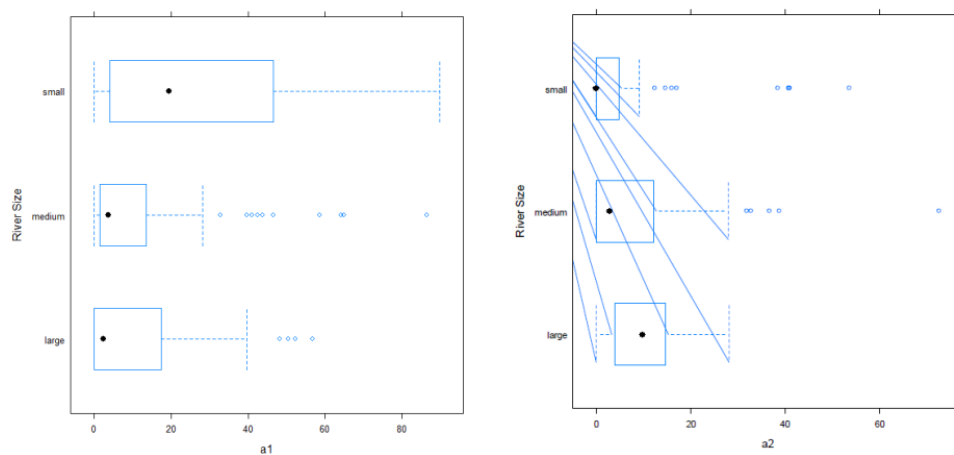
（ 2） 每个数值属性的盒图如下：

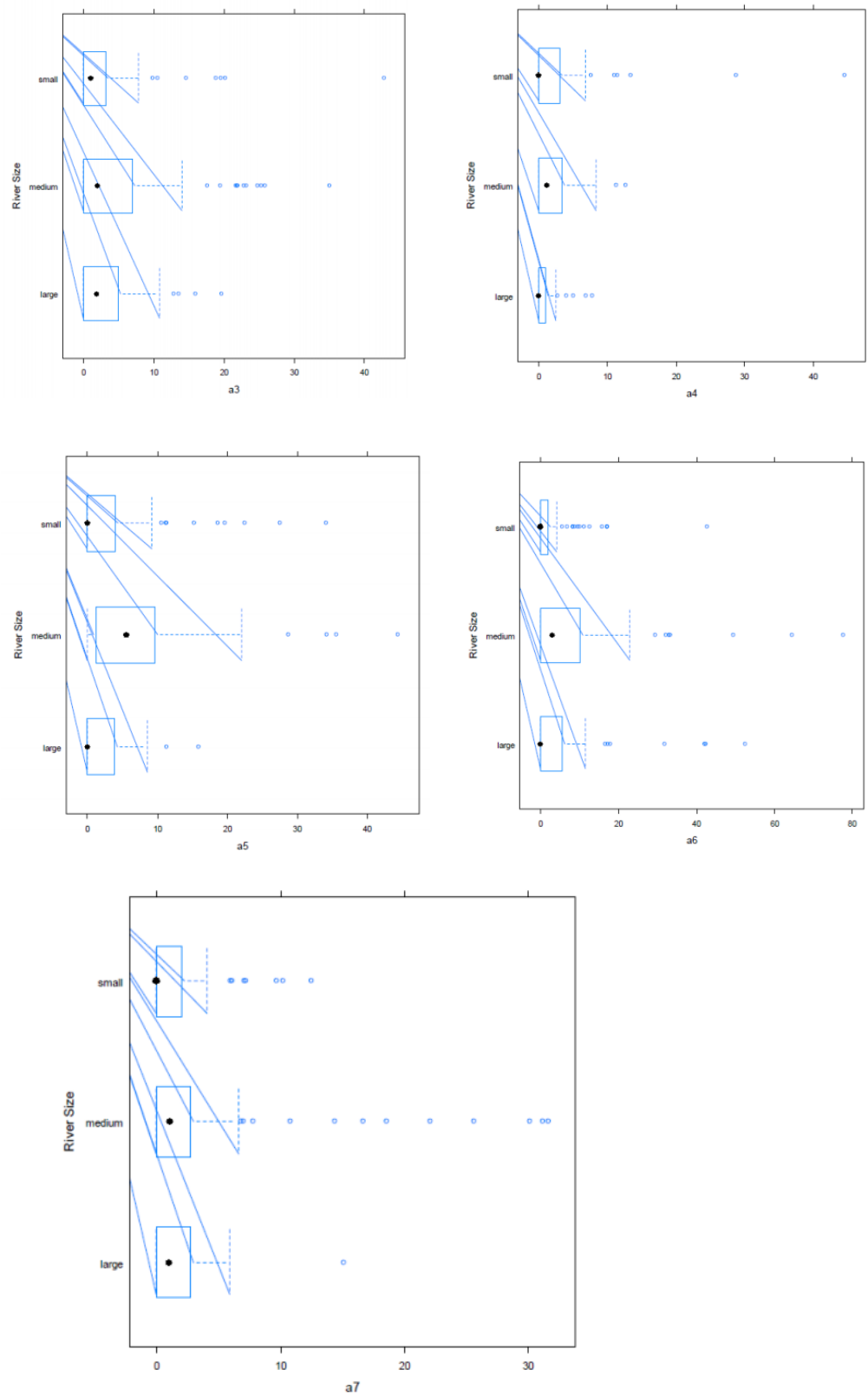




结果表明，较小的离群点多于较大的离群点导致平均值较中位线稍小，
总体来说，离群点较少。

(3) 7 种海藻与河流大小的条件盒图





从图中可知，在规模较小的河流中，海藻 **a1** 的频率最高。

3、 缺失数据的处理：

在许多水样中，一些变量含有缺失值。这种情形在现实问题中非常普遍，这会导致一些不能处理缺失值的分析方法无法应用。

以下通过四种方式处理缺失数据。

（ 1） 剔除缺失部分数据

```
#剔除缺失数据
omitdata = na.omit(algae)
write.table(omitdata, 'E:/study_material/硕-课件/数据挖掘/海藻数据分析/OmittedData.txt',
            col.names = F, row.names = F, quote = F)
```

（ 2） 利用最高频率值来填补缺失值

```
#使用高频数据替换
library(DMwR)
preprocess2 = algae[-manyNAs(algae),]
preprocess2 = centralImputation(preprocess2)
write.table(preprocess2, 'E:/study_material/硕-课件/数据挖掘/海藻数据分析/CentralImputationData.txt',
            col.names = F, row.names = F, quote = F)
```

（ 3） 通过属性的相关关系来填补缺失值

```
#通过变量相关性填补缺失值
symnum(cor(algae[,4:18], use='complete.obs'))
lm(formula=PO4~oPO4, data=algae)
preprocess3 = algae[-manyNAs(algae),]
fillPO4 <- function(oP) {
  if(is.na(oP))
    return(NA)
  else return (42.897 + 1.293 * oP)
}
preprocess3[is.na(preprocess3$PO4), 'PO4'] <- sapply(preprocess3[is.na(preprocess3$PO4), 'oPO4'], fillPO4)
write.table(preprocess3, 'E:/study_material/硕-课件/数据挖掘/海藻数据分析/linearDefaultData.txt',
            col.names = F, row.names = F, quote = F)
```

```

> symnum(cor(algae[,4:18],use='complete.obs'))
      mP mO C1 NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH 1
mnO2   1
C1     1
NO3    1
NH4    , 1
oPO4   . . 1
PO4    . . * 1
Ch1a   . 1
a1     . . . 1
a2     . . 1
a3     . 1
a4     . . 1
a5     . 1
a6     . . 1
a7     . 1
attr("legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1

```

由结果可知， oPO4 与 PO4 相关度超过 0.9， 所以可以用这两个属性作

相关分析， 互相填补缺失数据。用如下代码获得其线性模型：

```

> lm(formula=PO4~oPO4, data=algae)

Call:
lm(formula = PO4 ~ oPO4, data = algae)

Coefficients:
(Intercept)          oPO4
      42.897         1.293

```

（ 4） 通过数据对象之间的相似性来填补缺失值

```

#通过数据对象之间的相似型来填补缺失值
preprocess4 = knnImputation(algae,k=10)
write.table(preprocess4,'E:/study_material/硕-课件/数据挖掘/海藻数据分析/knnImputationData.
col.names = F,row.names = F, quote = F)

```