海藻数据的分析

姓名：于文楠

学号：2120151057

学院：计算机学院

邮箱：ywnbit@163.com

一、数据摘要

读取文件，同时设定表列名，设定缺失数据的字符串为 XXXXXXX。其中第一个参数指向了待分析数据文件的路径位置。

使用 R 代码：

```
mydata <- read.table('Analysis.txt',
                     header=F,
                     dec='.',
                     col.names=c('season','size','speed','mxPH','mnO2','Cl',
                     'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',
                     'a5','a6','a7'),
                     na.strings=c('XXXXXXX'))
```

用 Summary 函数分析数据摘要：

```
summary(mydata)
```

得到如下信息，最小值、前四分位，中位数，平均值，后四分位，最大值以及缺失数量的统计信息，如图：

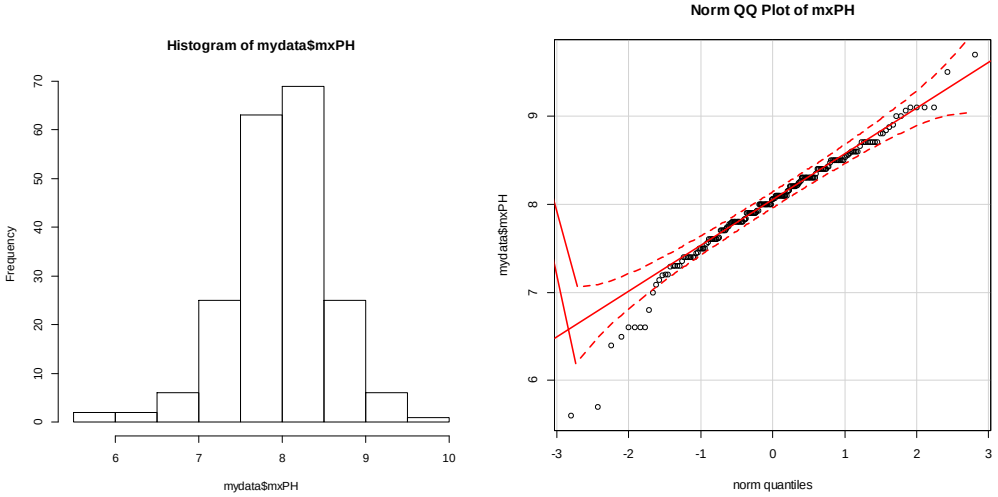
```
   season       size       speed        mxPH           mnO2             Cl              NO3
autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500   Min.   :  0.222   Min.   : 0.050
spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296
summer:45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

二、数据可视化

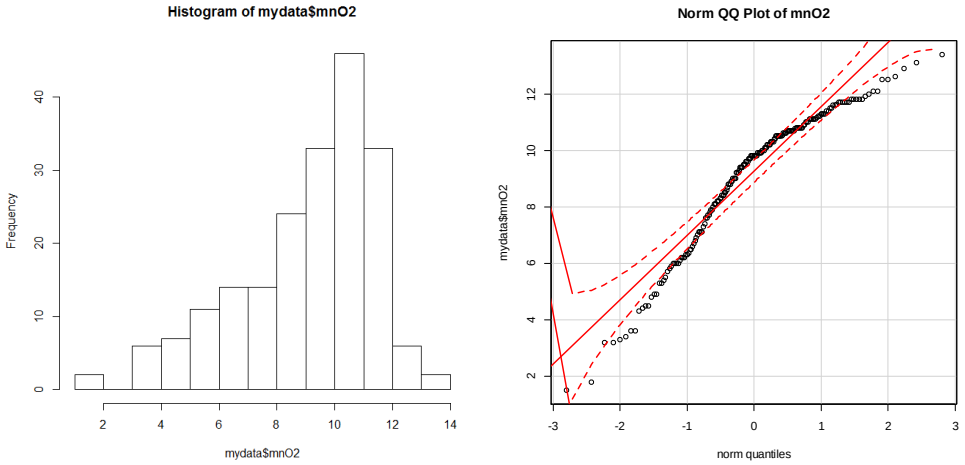
(1) 对数值属性，绘制直方图与 QQ 图检验其正态分布

绘制其直方图与 QQ 图。绘制出的直方图纵轴是其频数 ,横轴是其分布区间。QQ 图中，红色实线为其 QQ 线，虚线为 95%置信度的置信区间。

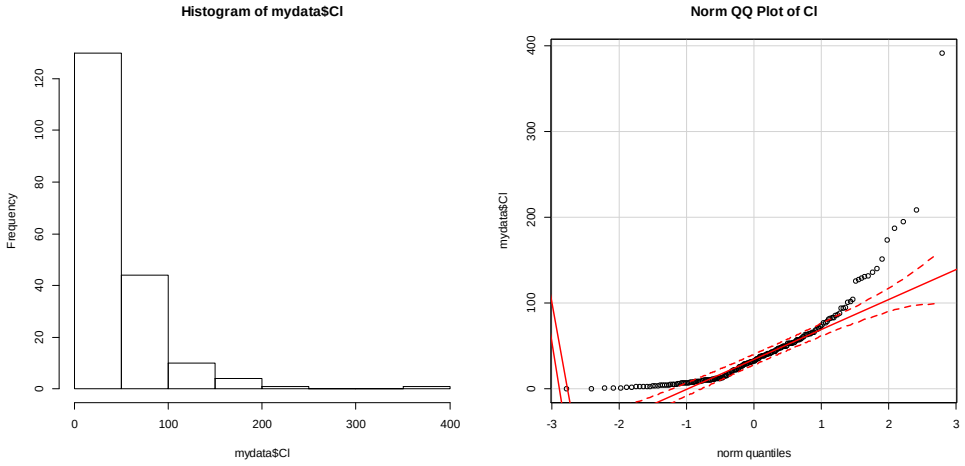
```
hist(mydata$mxPH)

library(car)

qqPlot(mydata$mxPH,main='Norm QQ Plot of mxPH')
```



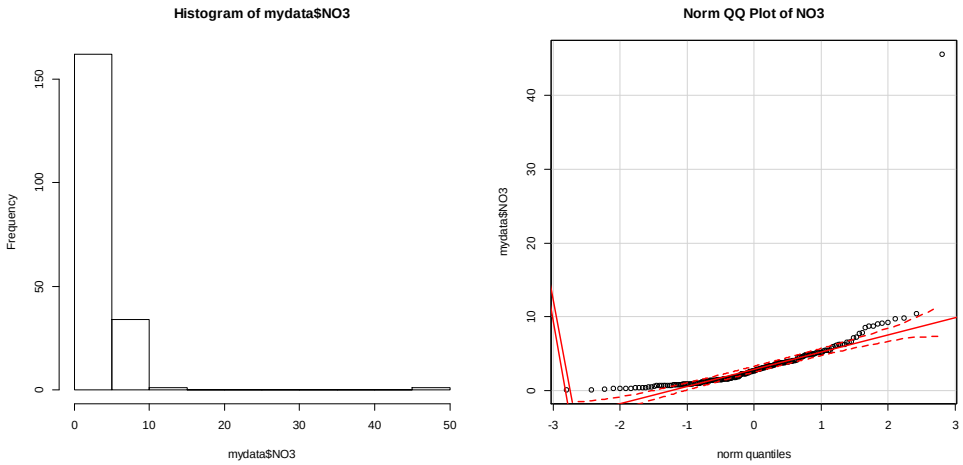
mxPH 的直方图与 QQ 图



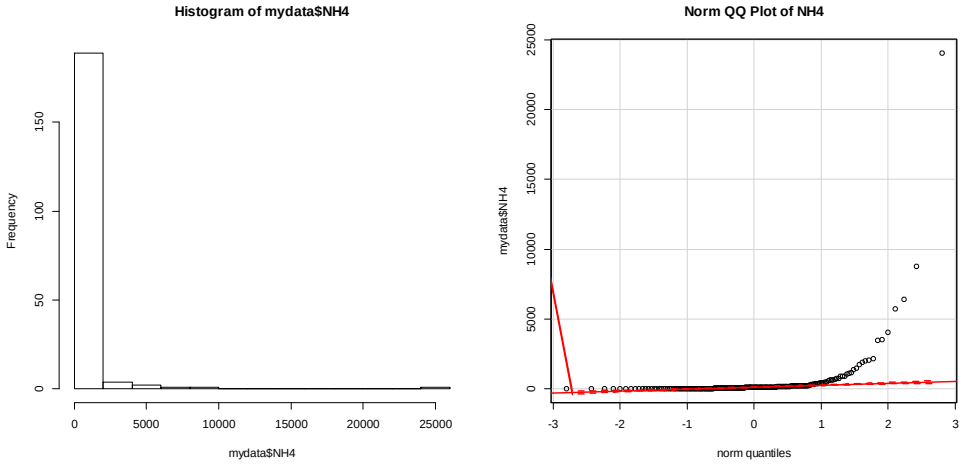
mnO2 的直方图与 QQ 图



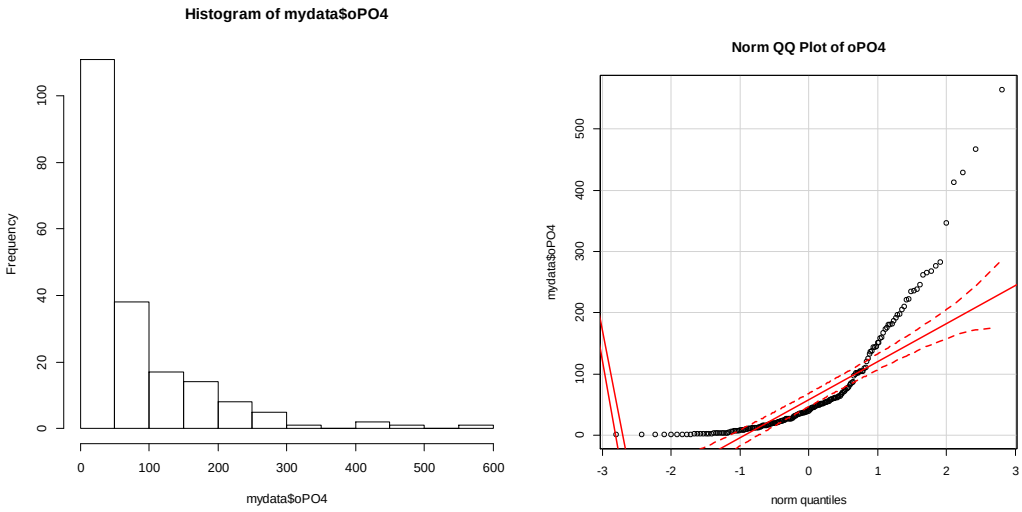
Cl 的直方图与 QQ 图



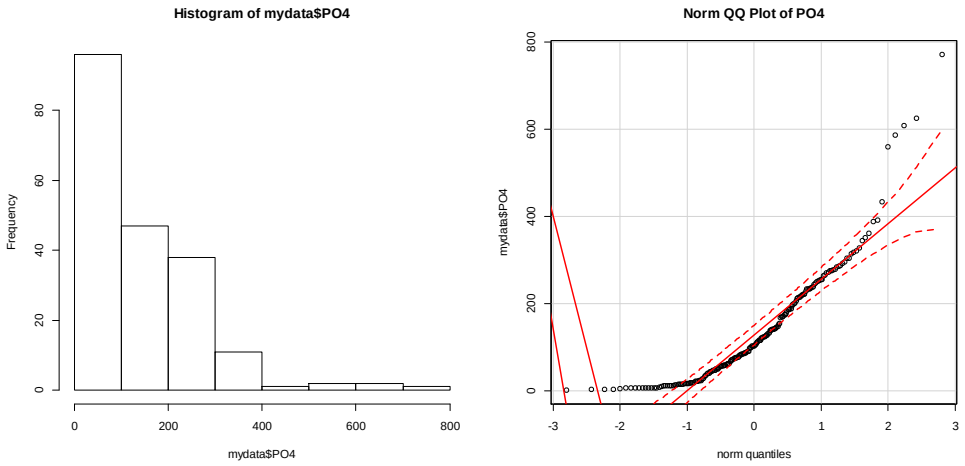
NO3 的直方图与 QQ 图



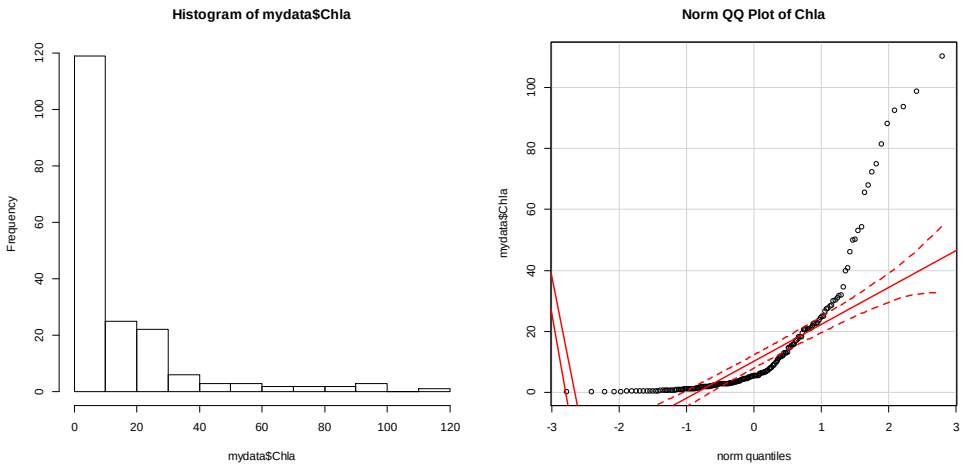
NH4 的直方图与 QQ 图



oPO4 的直方图与 QQ 图



PO4 的直方图与 QQ 图



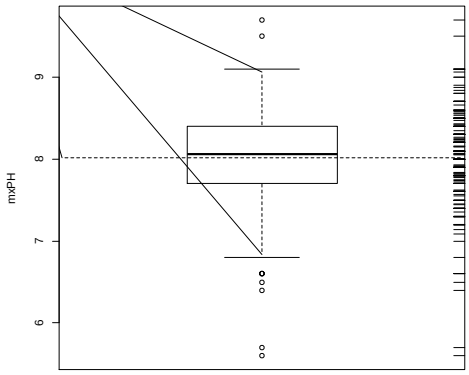
Chla 的直方图与 QQ 图

(2) 绘制盒图，识别离群点

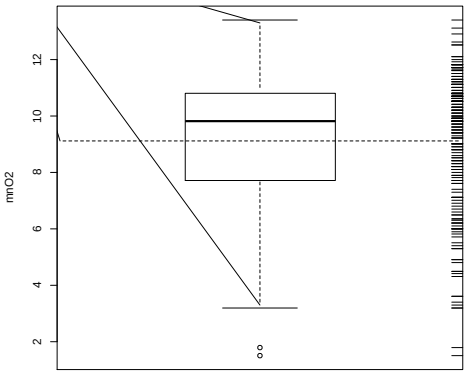
执行如下 R 代码：

```
boxplot(mydata$mxPH,ylab='mxPH')
rug(mydata$mxPH,side=4)
abline(h=mean(mydata$mxPH,na.rm=T),lty=2)
```

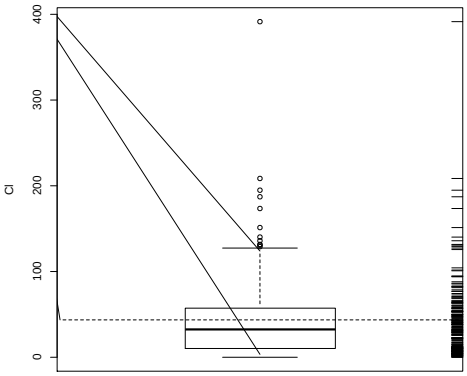
Rug 函数绘制了每个点在纵轴上的投影情况，abline 则绘制了数据的均值，在图中以虚线的方式呈现。由各属性的盒图，可以分析出离群点的数量以及分布情况。



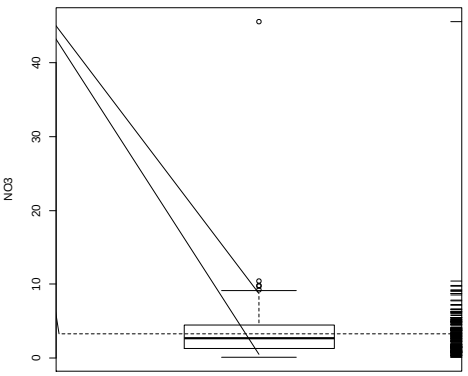
mxPH 盒图



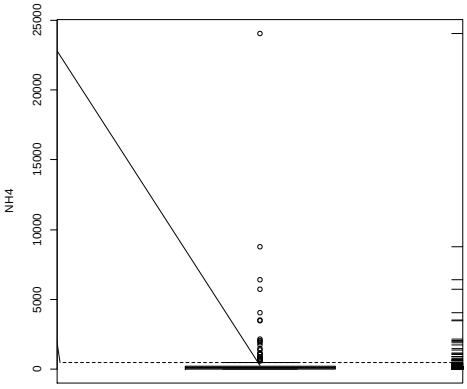
mnO2 盒图



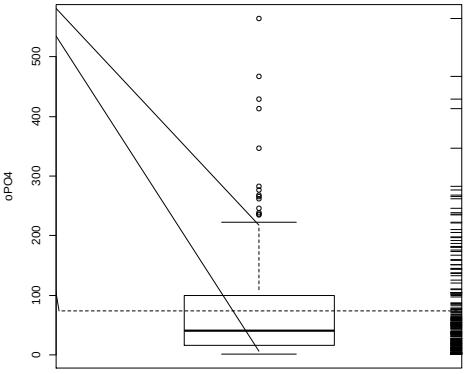
Cl 盒图



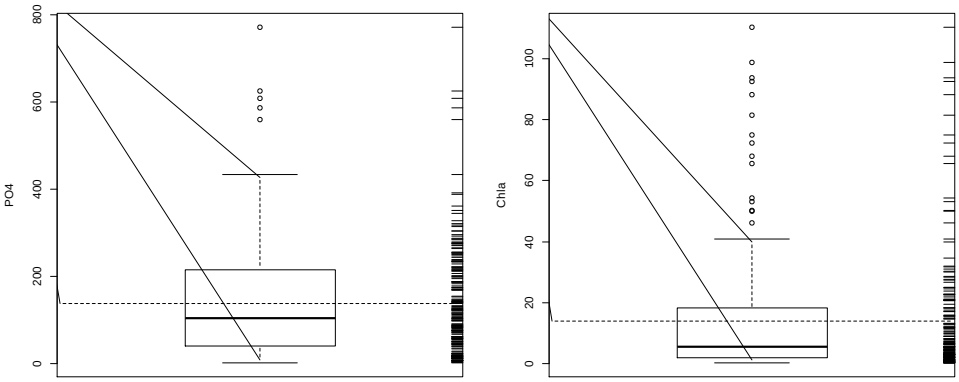
NO3 盒图



NH4



oPO4



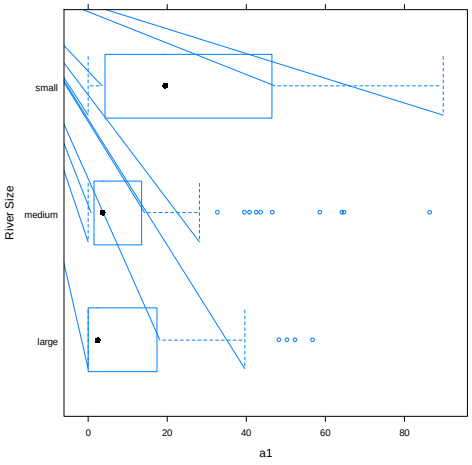
PO4 盒图　　Chla 盒图

(3) 对七种海藻，绘制其数量与河流大小的条件盒图

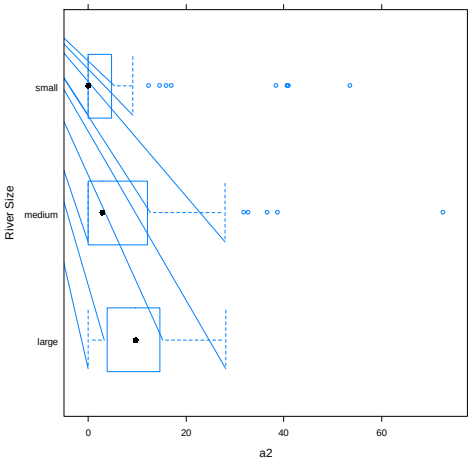
使用 R 代码，绘制其与河流大小的条件盒图，命令如下：

```
library(lattice)
bwplot(size~a1,data=mydata,ylab='River Size',xlab='a1')
```

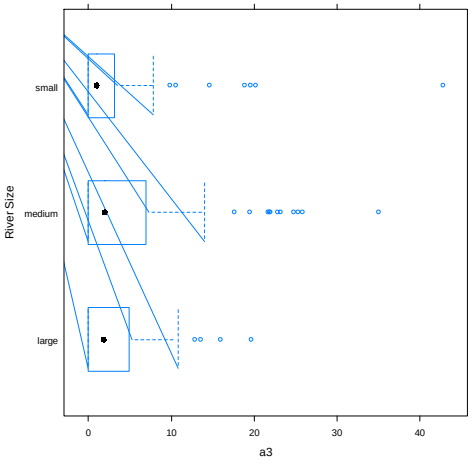
下图反映 a1 海藻在不同河流大小条件下的盒图形状。依次绘制 a1-a7 海藻的条件盒图，如图



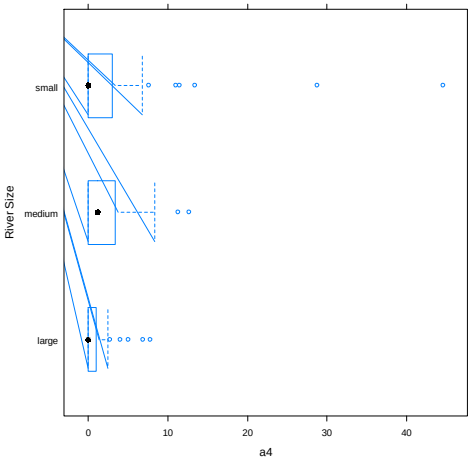
a1 海藻与河流大小的条件盒图



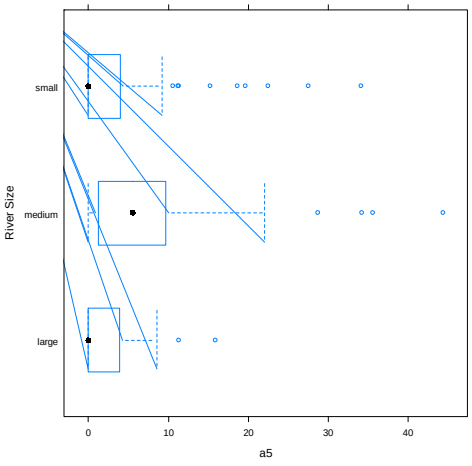
a2 海藻与河流大小的条件盒图



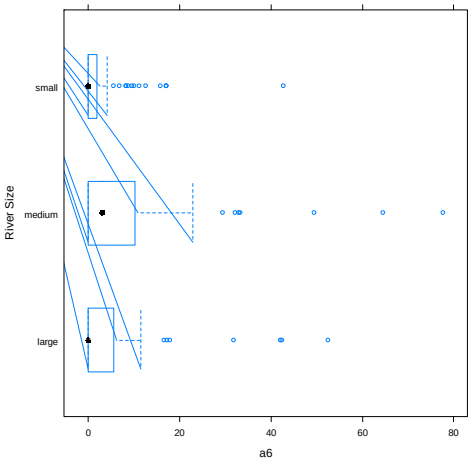
a3 海藻与河流大小的条件盒图



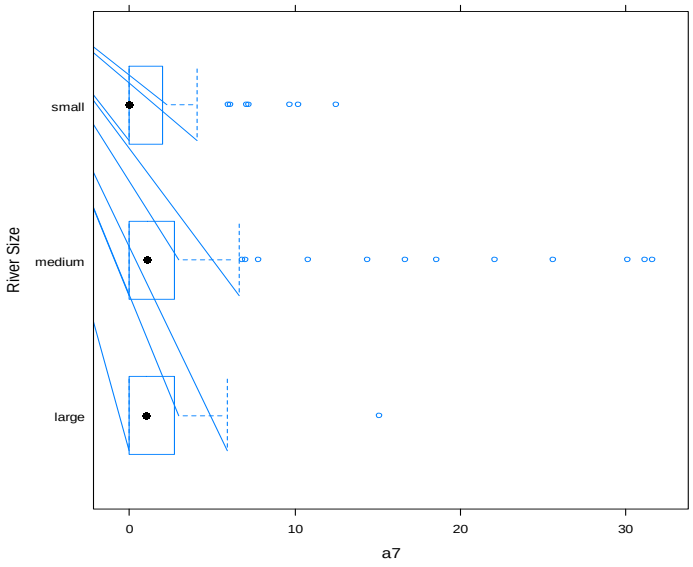
a4 海藻与河流大小的条件盒图



a5 海藻与河流大小的条件盒图



a6 海藻与河流大小的条件盒图



a7 海藻与河流大小的条件盒图

从图中可以看出 a1 有更高的频数，但是 a3,a5,a6 在中型河流中更多一些。

三、数据缺失的处理

(1) 将缺失部分剔除

剔除缺失数据与写入文件的命令如下：

omitdata = na.omit(mydata)　剔出缺失数据

write.table(omitdata,'OmitedData.txt',col.names = F,row.names = F, quote = F) 写入文件

(2) 使用高频数值来填补缺失值

```
library(DMwR)
preprocess2 = mydata[-manyNAs(mydata),]
preprocess2 = centralImputation(Preprocess2)
write.table(preprocess2,'D:/DataMining/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

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

得到属性之间的相关性如下图：

```
      mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH  1
mnO2     1
Cl          1
NO3            1
NH4            ,  1
oPO4     .  .        1
PO4      .  .        * 1
Chla  .                  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
```

从图中可以看出 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。

使用一下代码得到其线性模型：

lm(formula=PO4~oPO4, data=mydata)

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

Coefficients:
(Intercept)         oPO4
     42.897        1.293
```

oPO4 与 PO4 的线性模型分析

得到结果如图 28 所示，表示得到的线性模型为 PO4=42.897 + oPO4* 1.293.

使用线性模型来填充 PO4 与 oPO4 的数据：

```
preprocess3 = mydata[-manyNAs(mydata),]
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)
```

(4) 使用数据对象之间的相似型填补缺失值

```
preprocess4 = knnImputation(mydata,k=10)
write.table(preprocess4,'D:/DataMining/knnImputationData.txt',col.n
ames = F,row.names = F, quote = F)
```