对海藻数据的分析

崔绿叶 2120150981

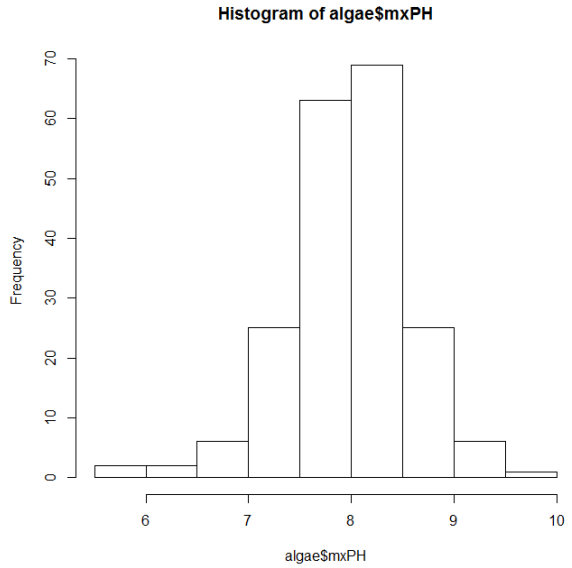
1、 数据摘要

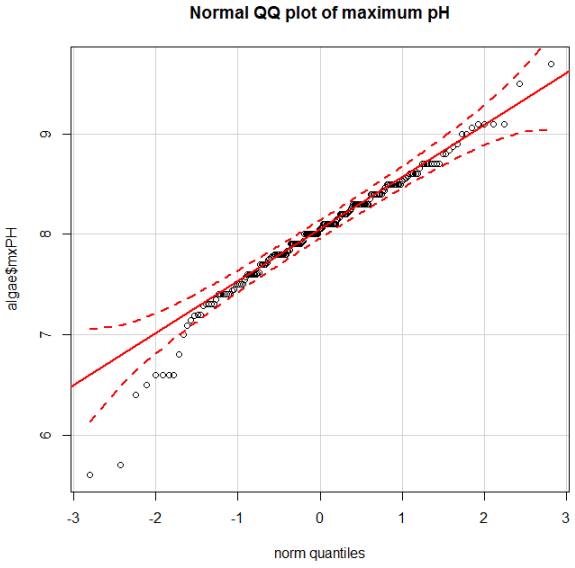
获取数据的描述性统计摘要是获取数据统计特性的一个重要方法。我们通过R语言的Summary函数对数据摘要进行分析， 这个函数给出了数据的统计特征概括，对于标称变量，他给出了每个取值变量的频数。对于数值型变量，R为我们提供了四分之一位数、中位数、均值、四分之三位数以及极值等一系列信息。所得结果如下：

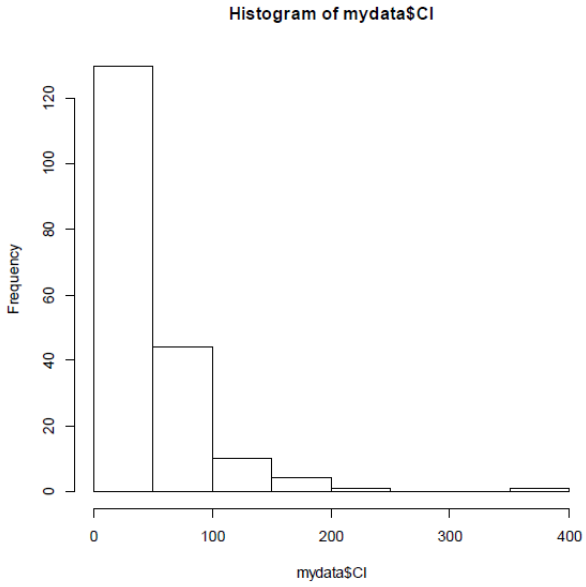
```
   season       size       speed        mxPH            mnO2             CL        
autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500   Min.   :  0.222  
spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725   1st Qu.: 10.981  
summer:45   small :71   medium:83   Median :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  
```

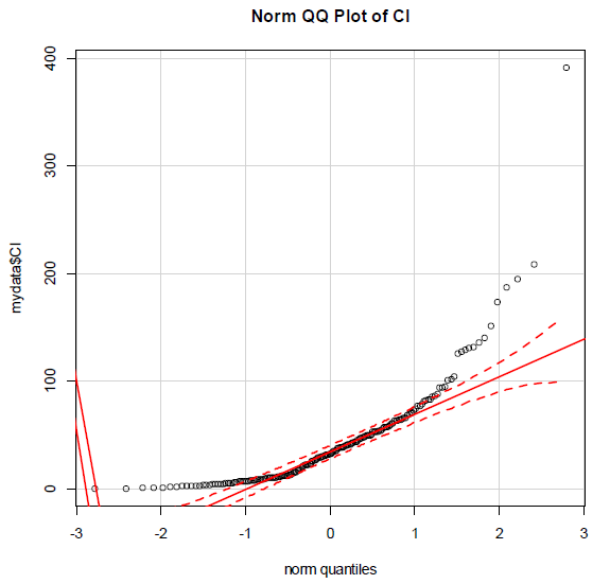
2、 数据可视化

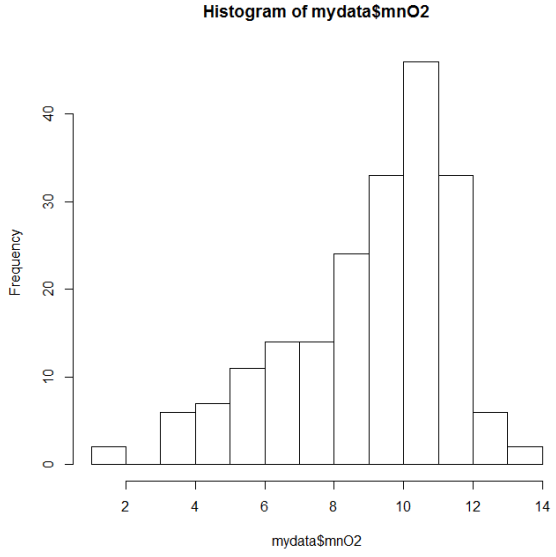
（1）每个数值属性（mxPH、Cl、mnO2、NO3、NH4、oPO4、PO4、Chla）的直方图和 QQ 图如下:

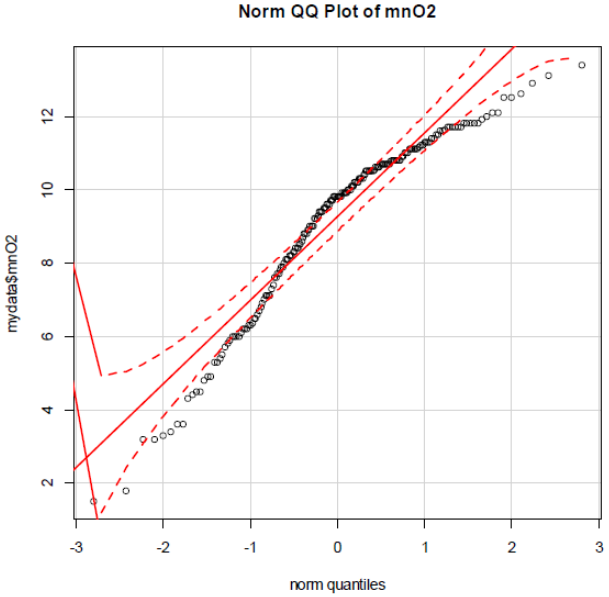

Histogram of algae$mxPH
Frequency
0
10
20
30
40
50
60
70
6
7
8
9
10
algae$mxPH

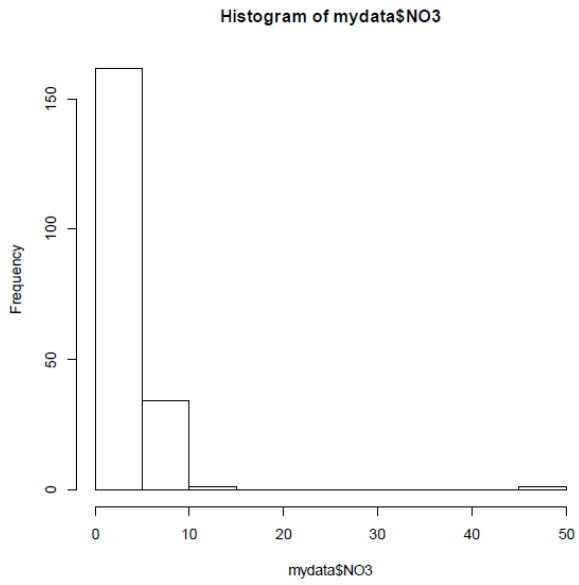

Normal QQ plot of maximum pH
algae$mxPH
6
7
8
9
-3
-2
-1
0
1
2
3
norm quantiles

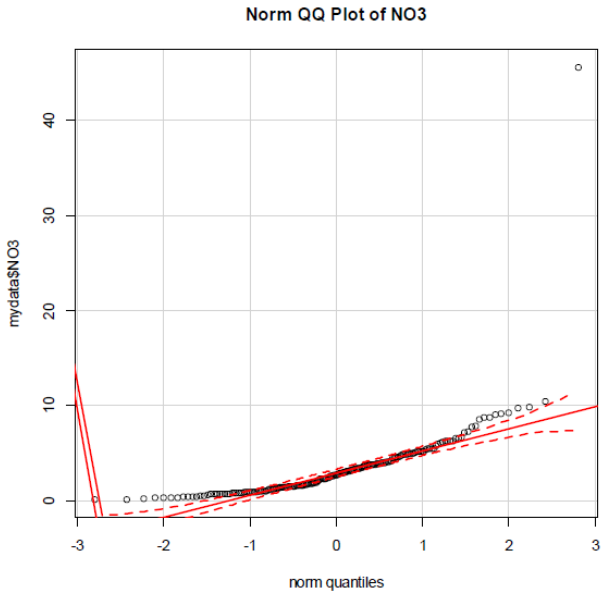

Histogram of mydata$Cl
Frequency
0
20
40
60
80
100
120
0
100
200
300
400
mydata$Cl

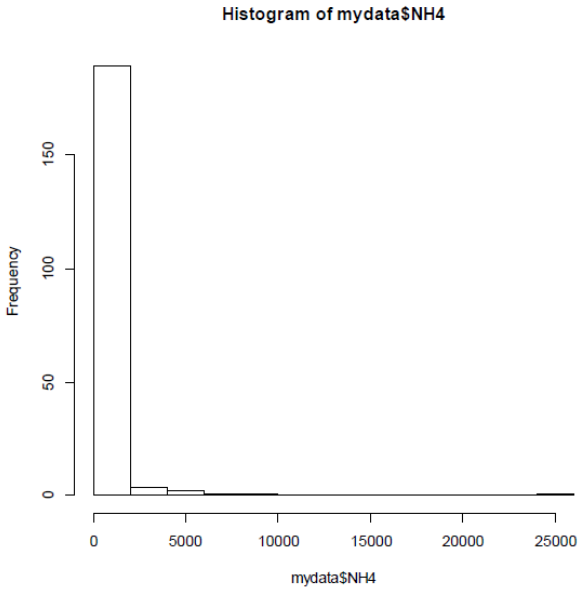

Norm QQ Plot of Cl
mydata$Cl
0
100
200
300
400
-3
-2
-1
0
1
2
3
norm quantiles

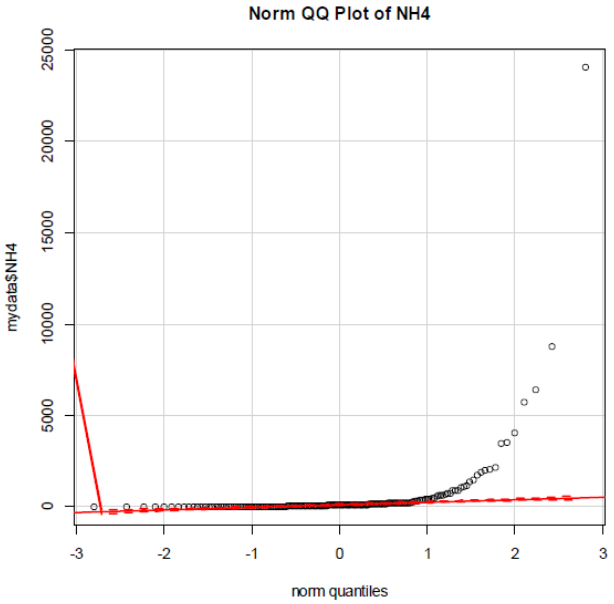

Histogram of mydata$mnO2
Frequency
0
10
20
30
40
2
4
6
8
10
12
14
mydata$mnO2

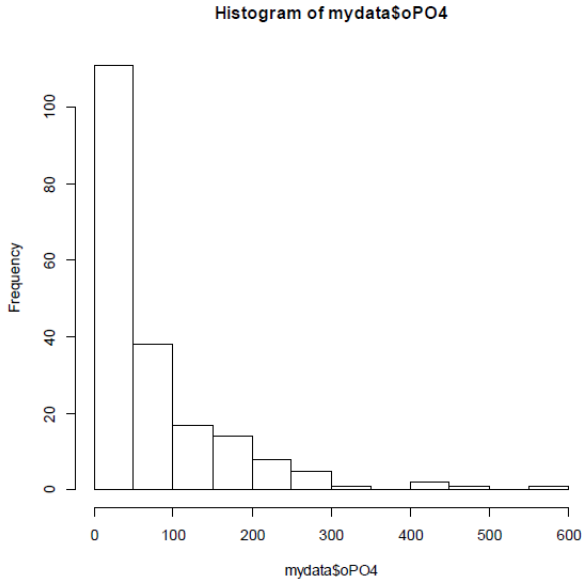

Norm QQ Plot of mnO2
mydata$mnO2
2
4
6
8
10
12
-3
-2
-1
0
1
2
3
norm quantiles

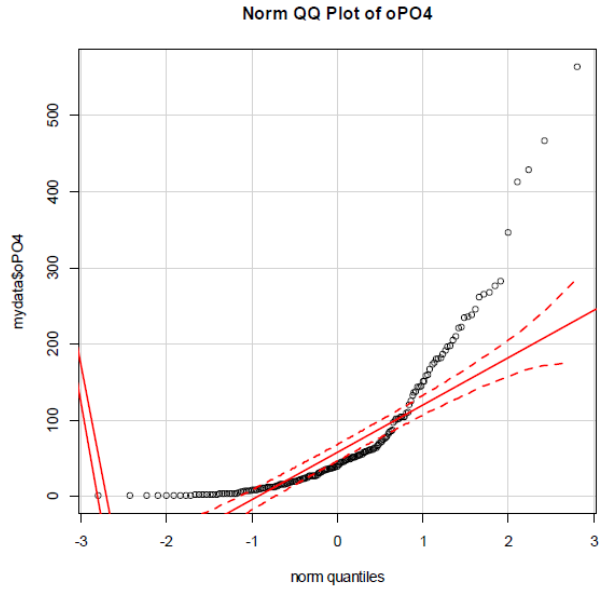

Histogram of mydata$NO3
Frequency
0
50
100
150
0
10
20
30
40
50
mydata$NO3

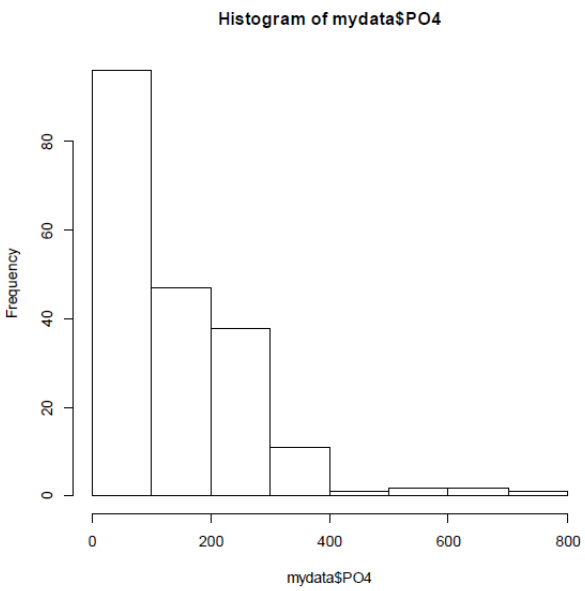

Norm QQ Plot of NO3
mydata$NO3
0
10
20
30
40
-3
-2
-1
0
1
2
3
norm quantiles

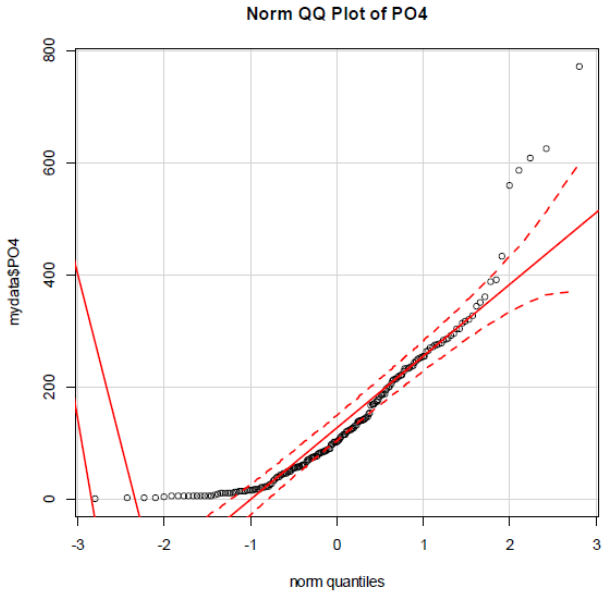

Histogram of mydata$NH4
Frequency
0
50
100
150
0
5000
10000
15000
20000
25000
mydata$NH4

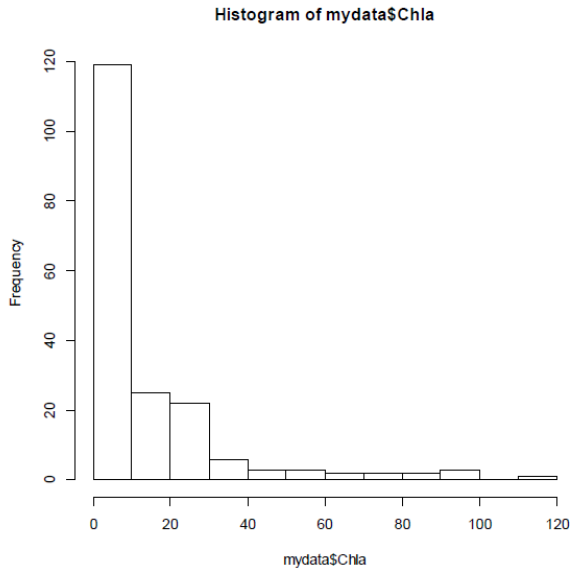

Norm QQ Plot of NH4
mydata$NH4
0
5000
10000
15000
20000
25000
-3
-2
-1
0
1
2
3
norm quantiles

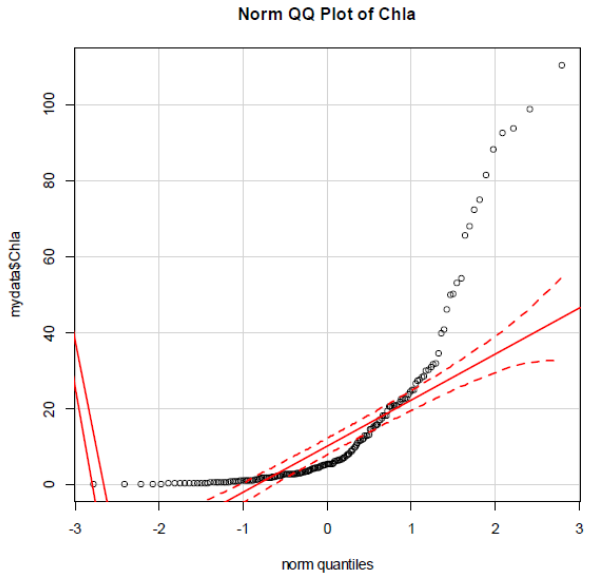

Histogram of mydata$oPO4
Frequency
0
20
40
60
80
100
0
100
200
300
400
500
600
mydata$oPO4


Norm QQ Plot of oPO4
mydata$oPO4
0
100
200
300
400
500
-3
-2
-1
0
1
2
3
norm quantiles





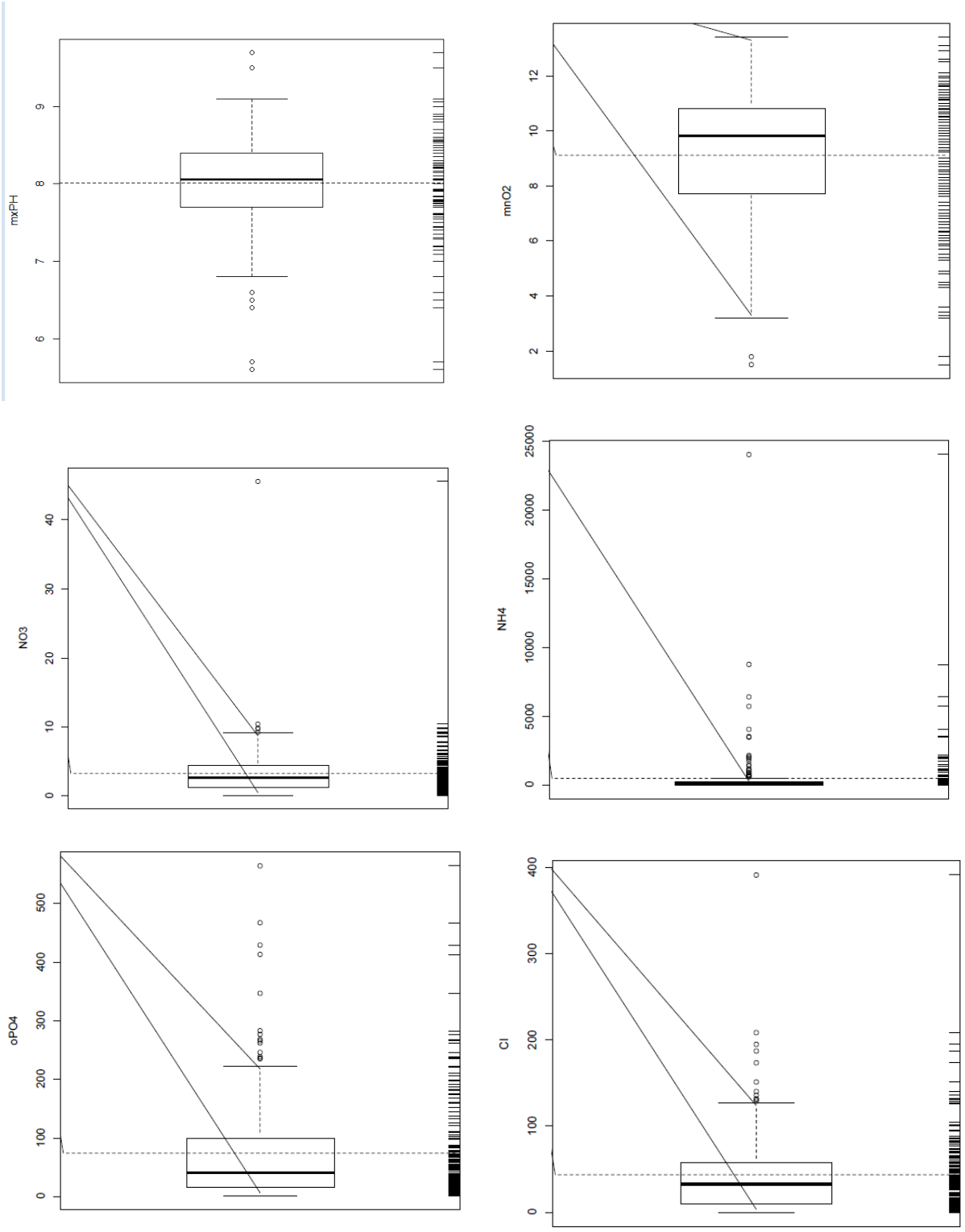


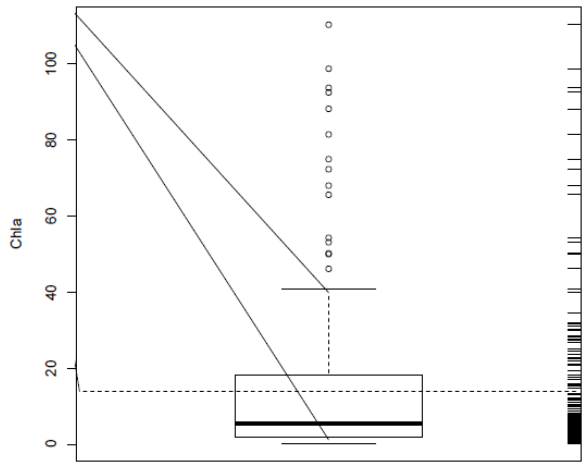


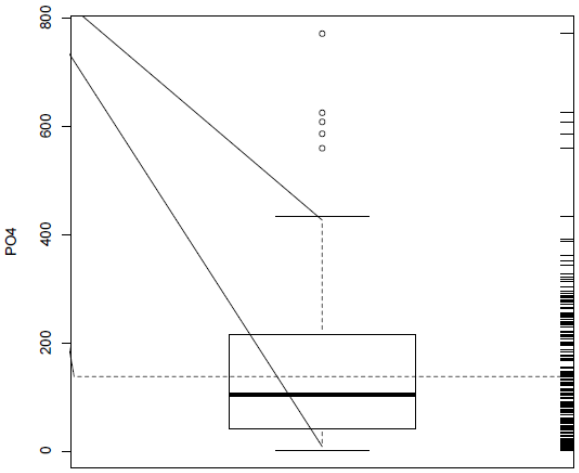
以上绘制的直方图横轴是其分布区间，纵轴是其频数；QQ图中，红色实线为其QQ线，虚线为95%置信度的置信区间。结果表明直方图显示变量mxPH的分布非常接近正态分布，它的值大多数都集中在变量的均值附近；QQ图绘制了变量值与正态分布的理论分位数的散点图，同时他给出正态分布的95%的置信区间的带状图，除去有几个小的值明显在95%置信区间之外，基本服从正态分布。由数据图表可以看出，mxPH2基本符合正态分布，其他属性大部分数据点在QQ图

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：

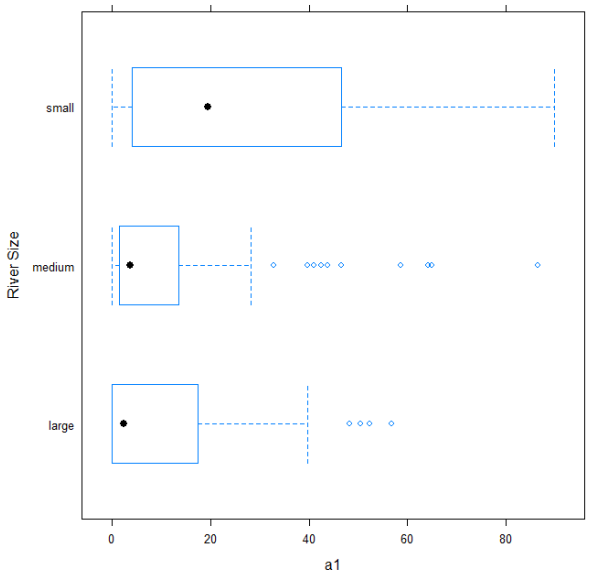


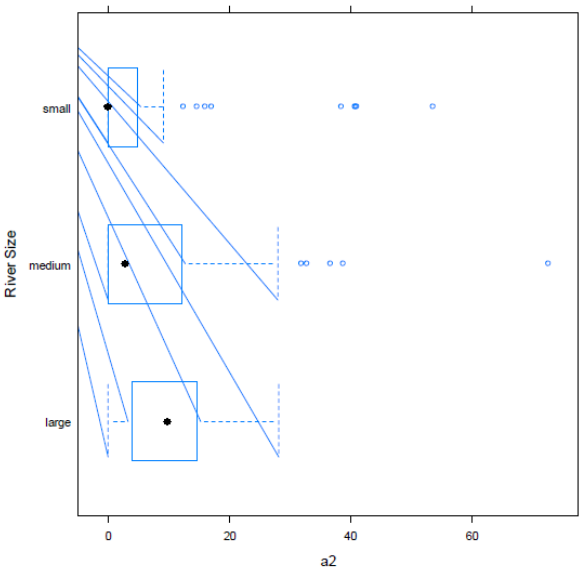


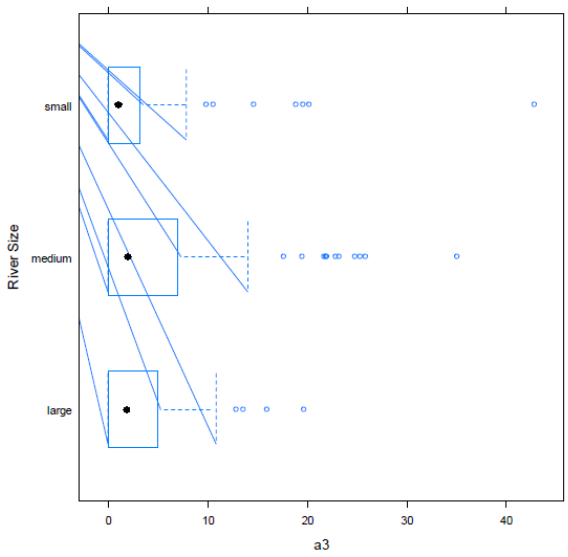


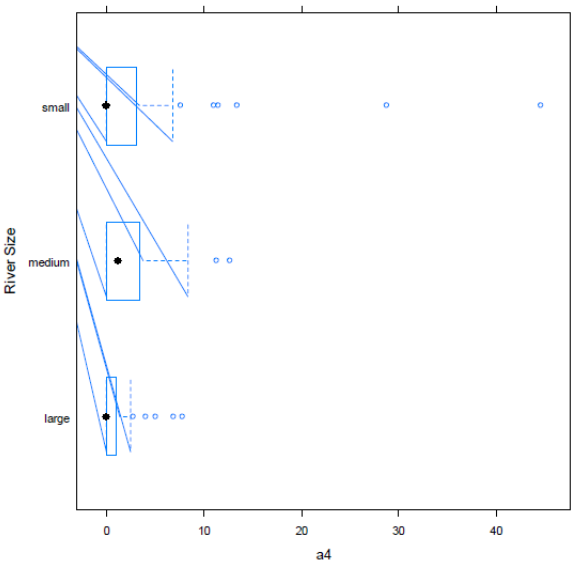
结果表明，较小的离群点多于较大的离群点导致平均值较中位线稍小，总体来说，离群点较少。

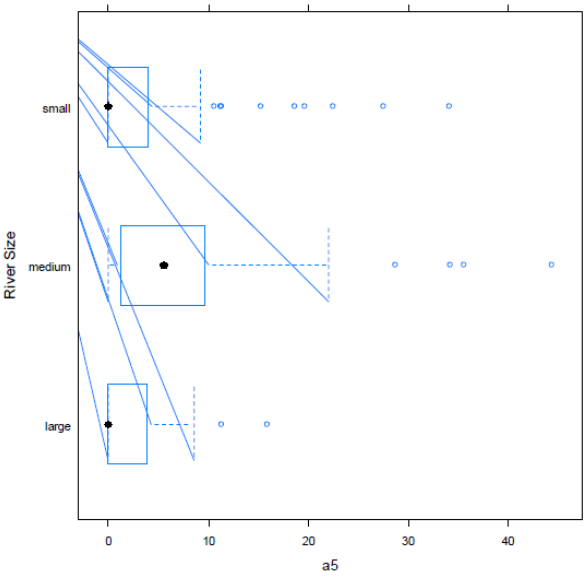
（3）7 种海藻与河流大小的条件盒图：

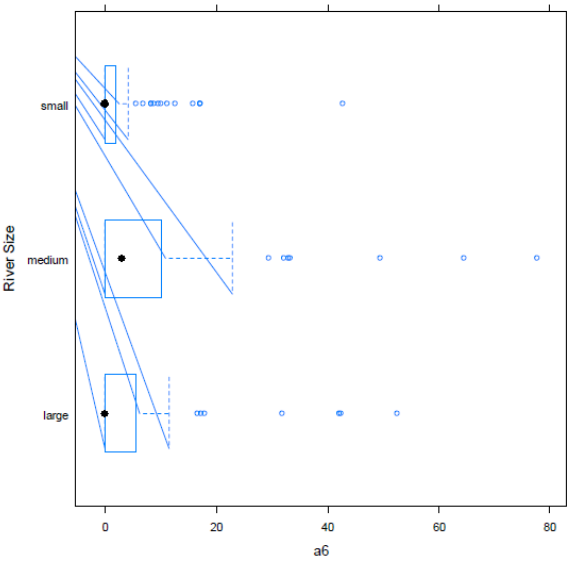


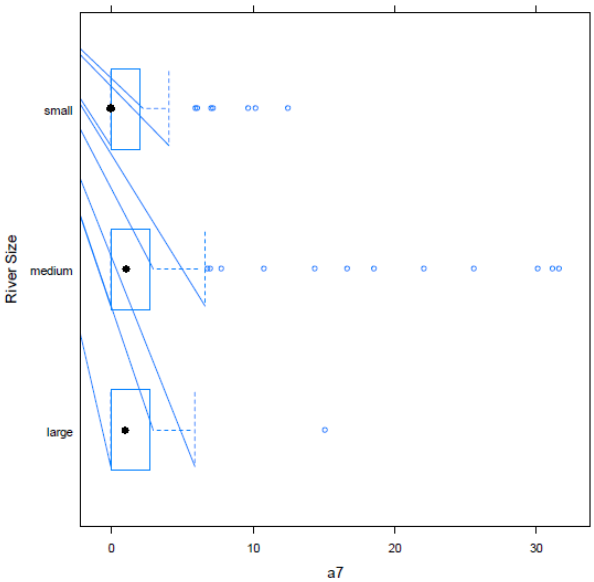



River Size
small
medium
large
0
10
20
30
40
a3


River Size
small
medium
large
0
10
20
30
40
a4


River Size
small
medium
large
0
10
20
30
40
a5


River Size
small
medium
large
0
20
40
60
80
a6


River Size
small
medium
large
0
10
20
30
a7

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/OmitedData.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 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
```

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.txt',
col.names = F,row.names = F, quote = F)
```