海藻数据的分析报告

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

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

3.1 数据可视化和摘要

首先使用将数据转换成 arff 格式进行计算。该格式概览如下：

```
@relation Analysis

@attribute season {spring, summer, autumn, winter}
@attribute size {small, medium, large}
@attribute speed {low, medium, high}
@attribute mxPH real
@attribute mnO2 real
@attribute CL real
@attribute NO3 real
@attribute NH4 real
@attribute oPO4 real
@attribute PO4 real
@attribute Chla real
@attribute a1 real
@attribute a2 real
@attribute a3 real
@attribute a4 real
@attribute a5 real
@attribute a6 real
@attribute a7 real
```

数据摘要

- 对标称属性，给出每个可能取值的频数:

spring	summer	autumn	winter
53	45	40	62

small	medium	large
71	84	45

low	medium	high
33	83	84

- 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:

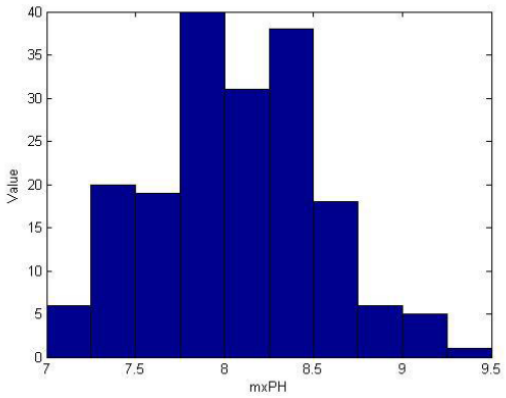
```
              mxPH        mnO2          Cl         NO3          NH4  \
count  199.000000  198.000000  190.000000  198.000000   198.000000
mean     8.011734    9.117778   43.636279    3.282389   501.295828
std      0.598305    2.391253   46.831312    3.776474  1962.545467
min      5.600000    1.500000    0.222000    0.050000     5.000000
25%      7.700000    7.725000   10.981250    1.296000    38.333250
50%      8.060000    9.800000   32.730000    2.675000   103.166500
75%      8.400000   10.800000   57.823500    4.446250   226.950000
max      9.700000   13.400000  391.500000   45.650000  24064.000000

              oPO4         PO4        Chia          a1          a2          a3  \
count  198.000000  198.000000  188.000000  200.000000  200.000000  200.000000
mean    73.590596  137.882101   13.971197   16.923500    7.458500    4.309500
std     91.136436  128.993740   20.495920   21.348376   11.028202    6.948537
min      1.000000    1.000000    0.200000    0.000000    0.000000    0.000000
25%     15.700000   41.375250    2.000000    1.500000    0.000000    0.000000
50%     40.150000  103.285500    5.475000    6.950000    3.000000    1.550000
75%     99.333250  213.750000   18.307500   24.800000   11.375000    4.925000
max    564.600000  771.600000  110.456000   89.800000   72.600000   42.800000

               a4          a5          a6          a7
count  200.000000  200.000000  200.00000  200.000000
mean     1.992500    5.064500    5.96400    2.495500
std      4.417404    7.491401   11.66071    5.158564
min      0.000000    0.000000    0.00000    0.000000
25%      0.000000    0.000000    0.00000    0.000000
50%      0.000000    1.900000    0.00000    1.000000
75%      2.400000    7.500000    6.92500    2.400000
max     44.600000   44.400000   77.60000   31.600000
<class 'pandas.core.frame.DataFrame'>
```

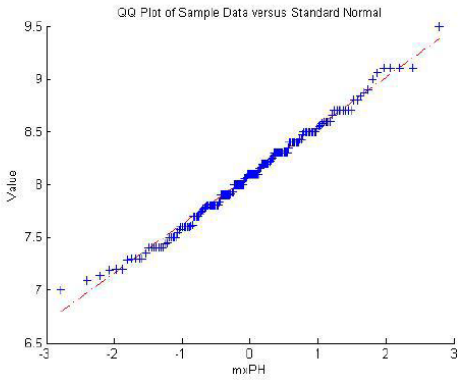
数据的可视化

针对数值属性：

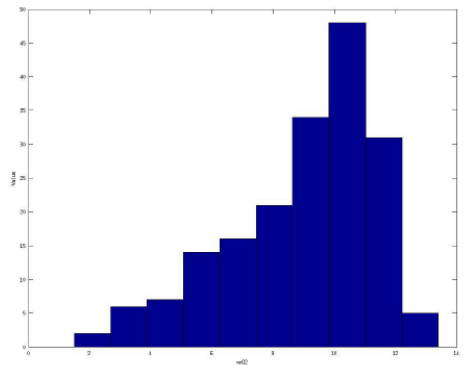
绘制直方图，并用 qq 图检验其分布是否为正态分布：
可以判断仅 mxPH 为正态分布。

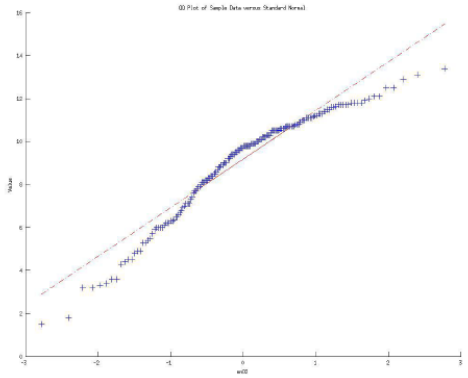
@attribute mxPH real：



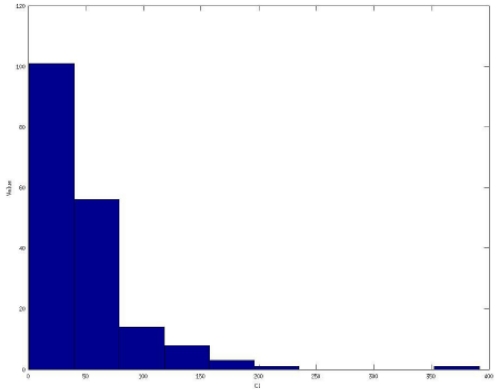


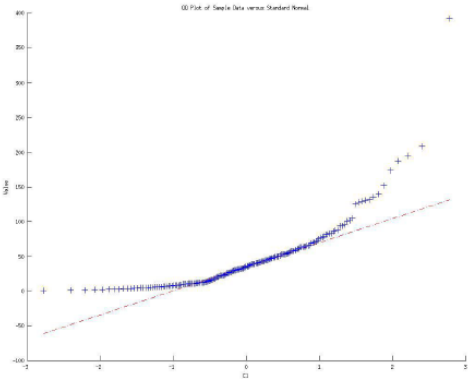
@attribute mnO2 real：



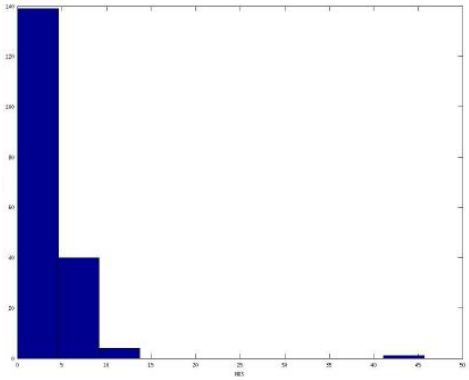


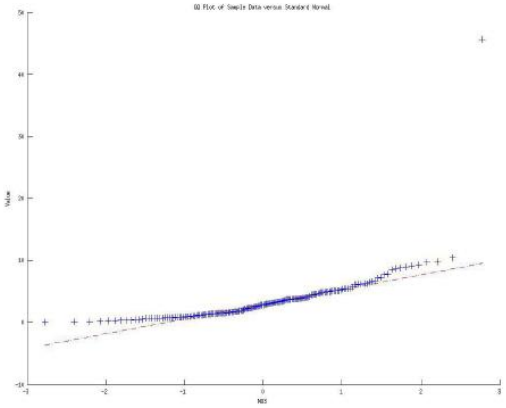
@attribute CL real：



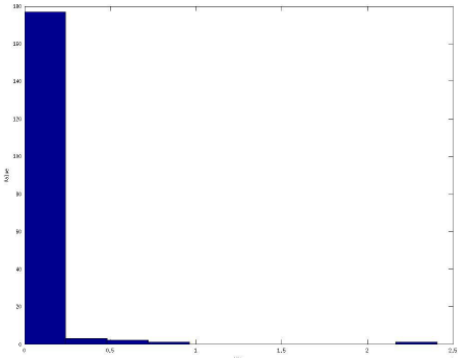


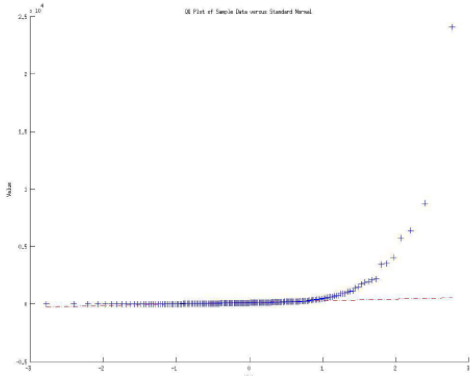
@attribute NO3 real：



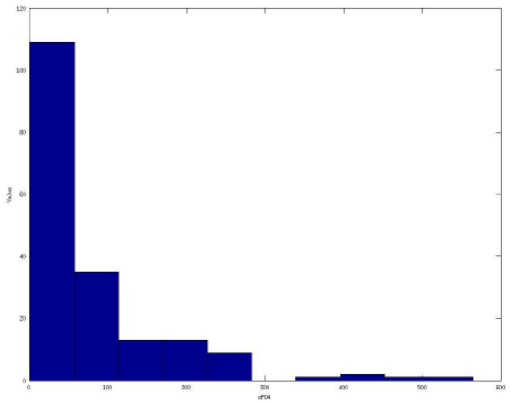


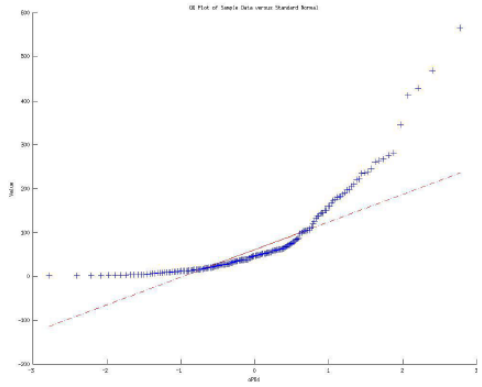
@attribute NH4 real：



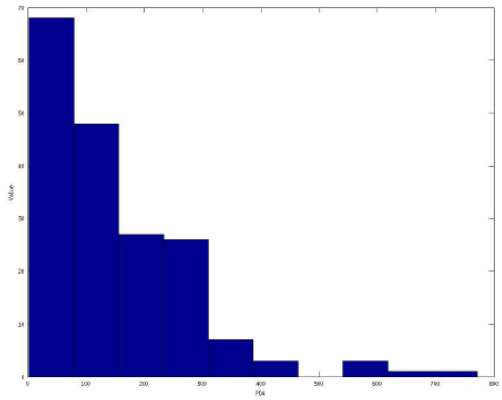


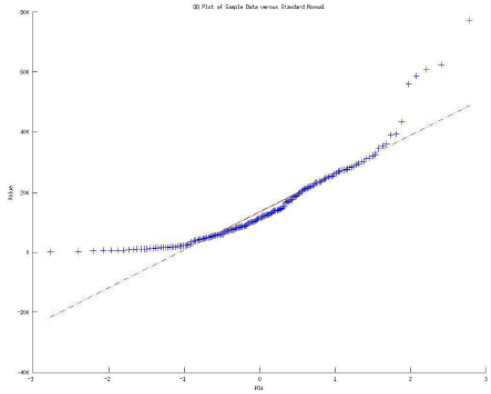
@attribute oPO4 real：



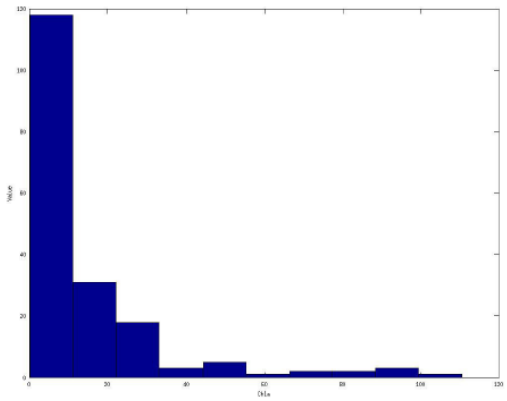


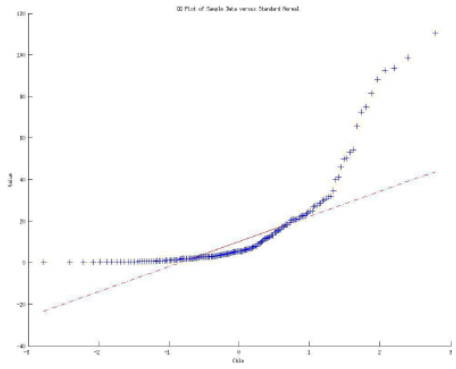
@attribute PO4 real：



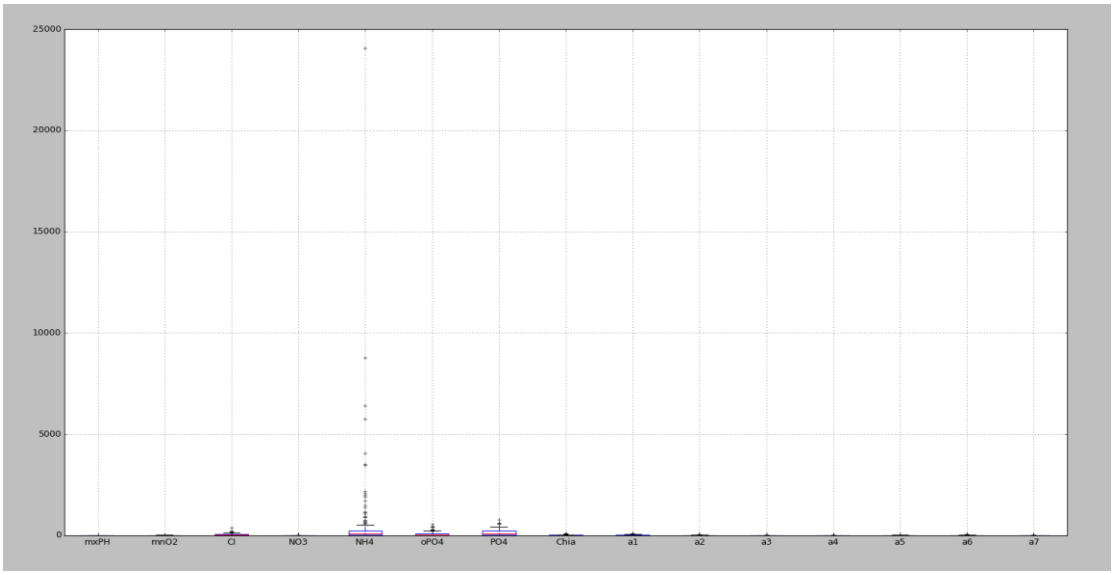


@attribute Chla real：

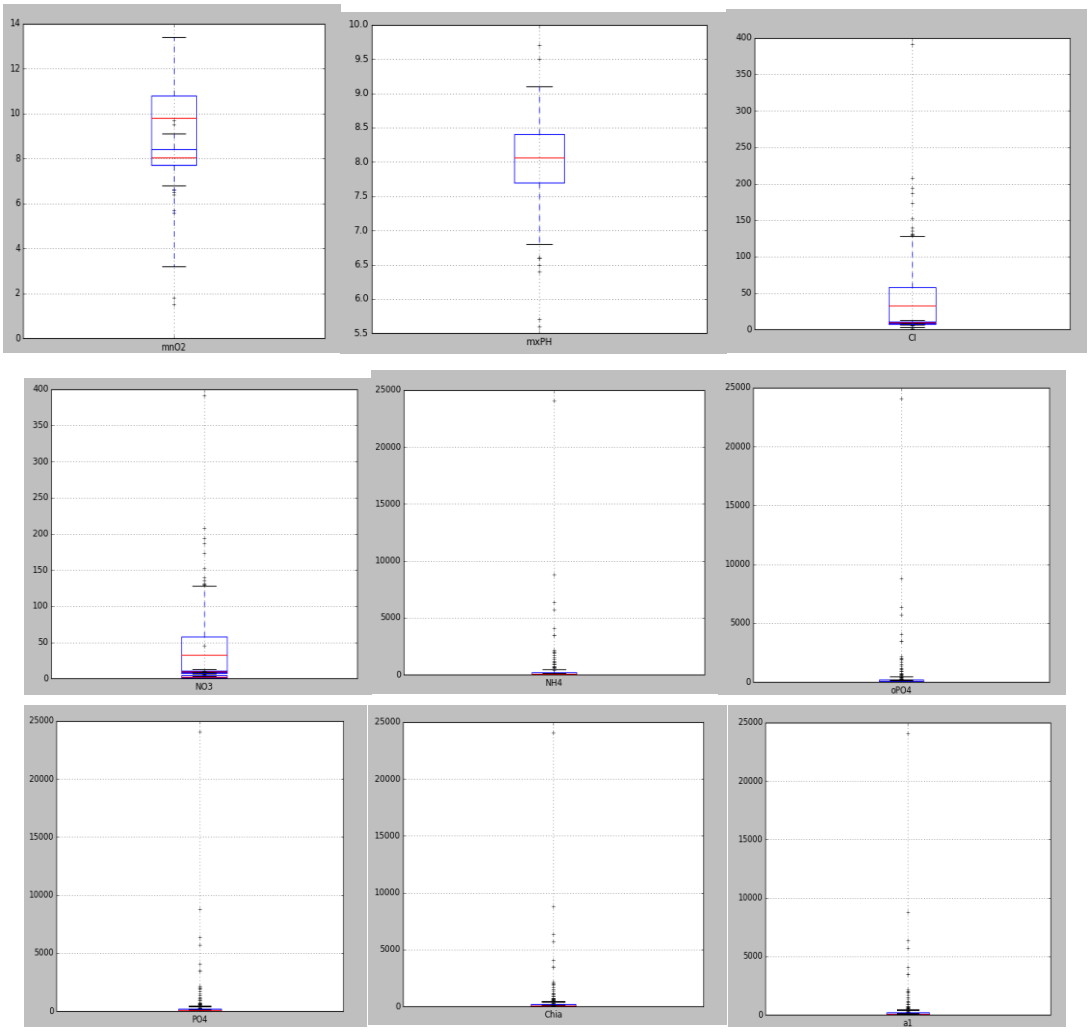


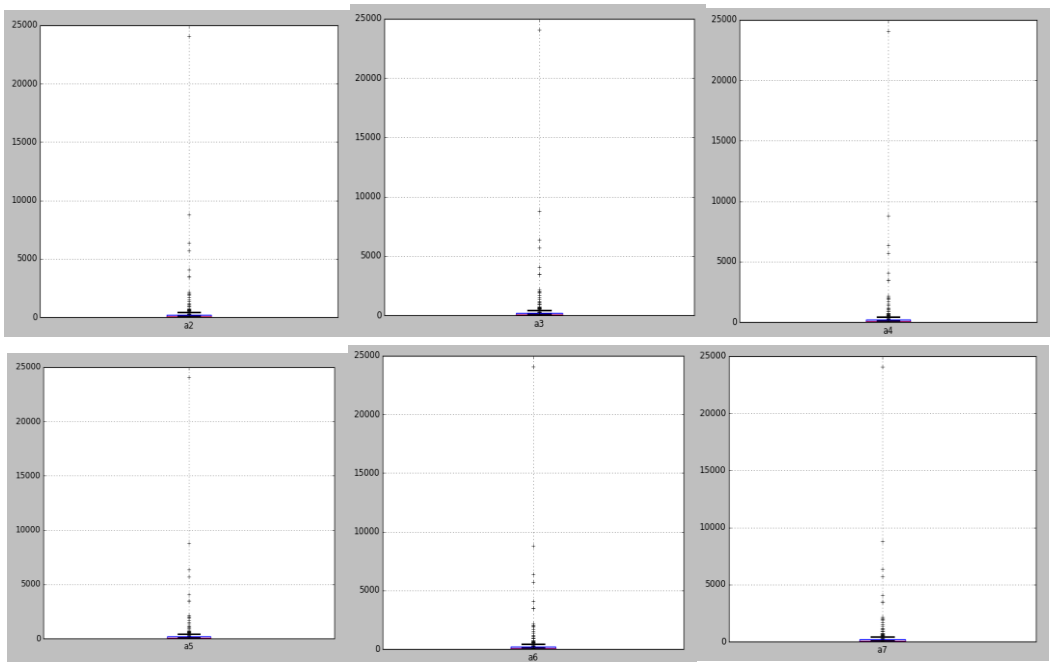


- 绘制盒图，对离群值进行识别：



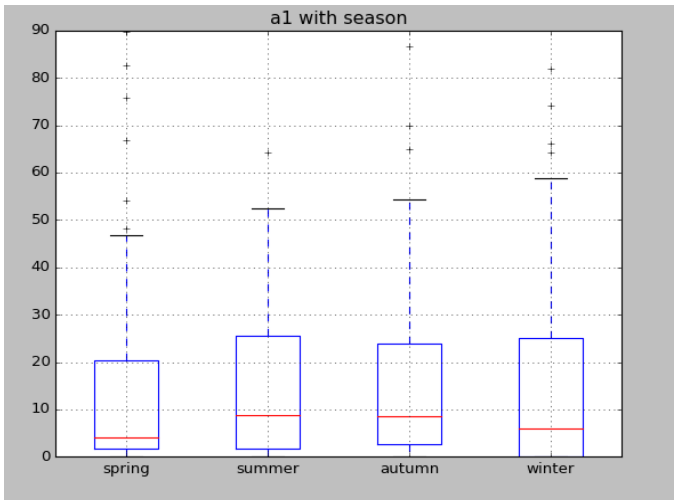
整体图并不能清楚判断离群点，故作每个属性的盒图如下：

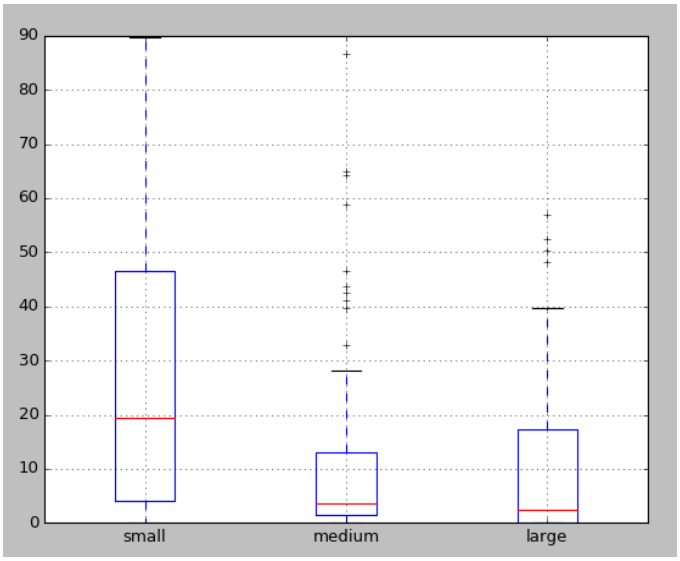


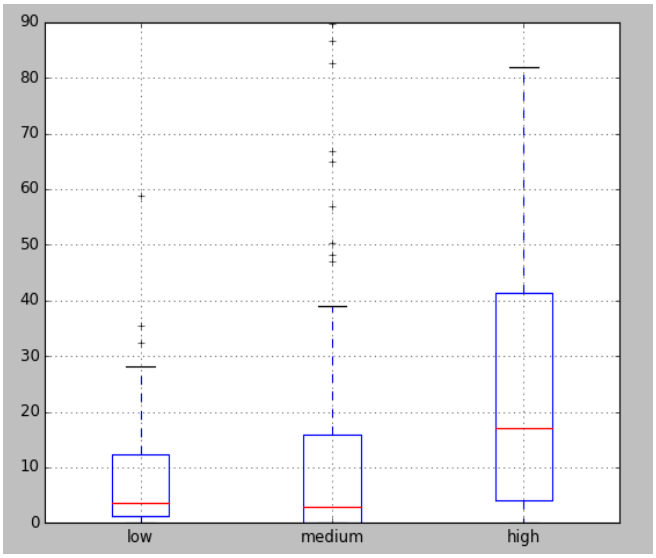


由图可看出，mnO2 分布较为均匀，而 NO3 与 NH4 中都有个值较高的离群点，可能为噪声数据或者特殊样例。

对 7 种海藻，分别绘制其数量与标称变量，如 size 的条件盒图：
(a1 及 a7 类似，这里给出 a1）



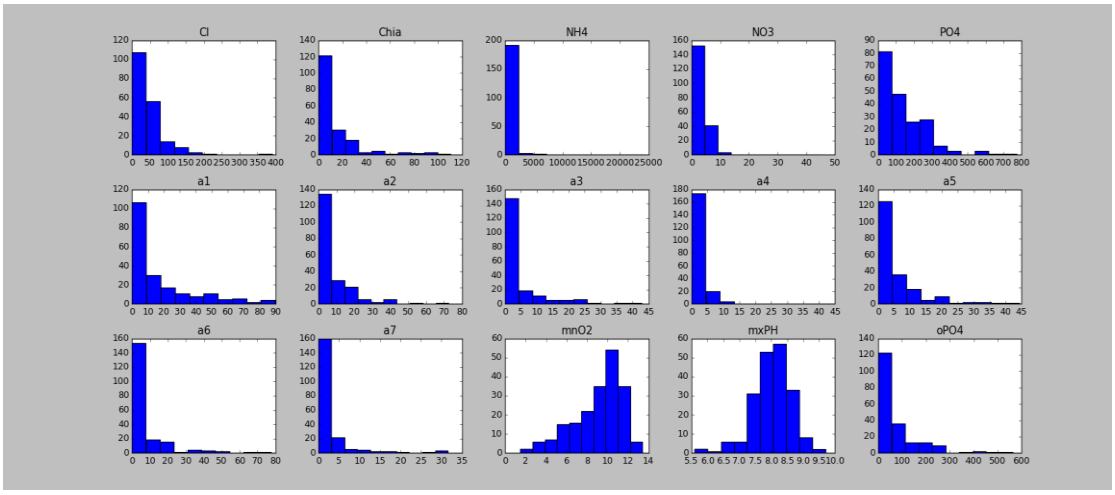




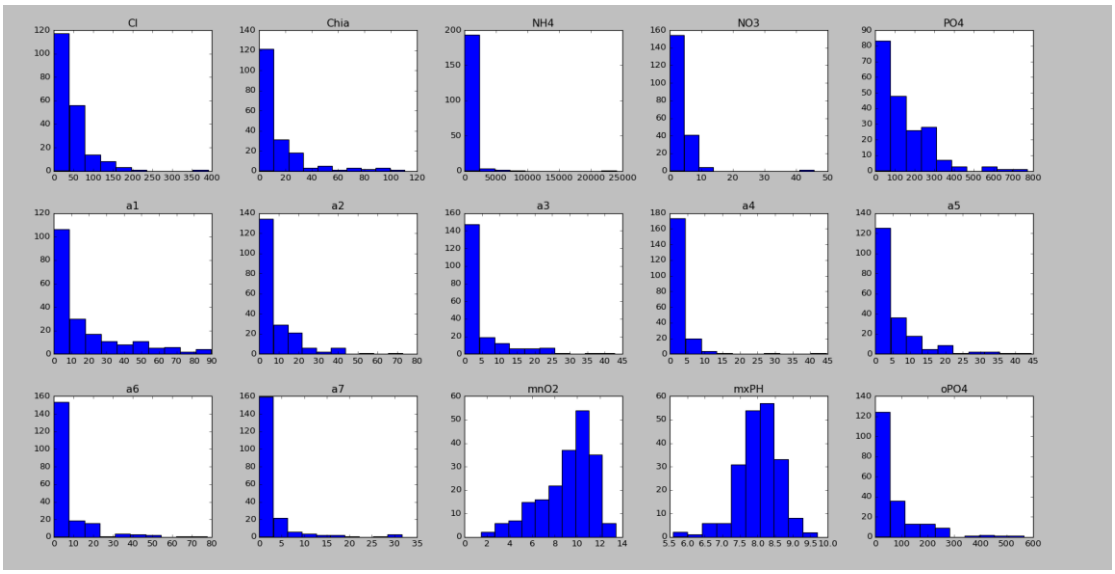
3.2 数据缺失的处理

分别使用下列四种策略对缺失值进行处理:

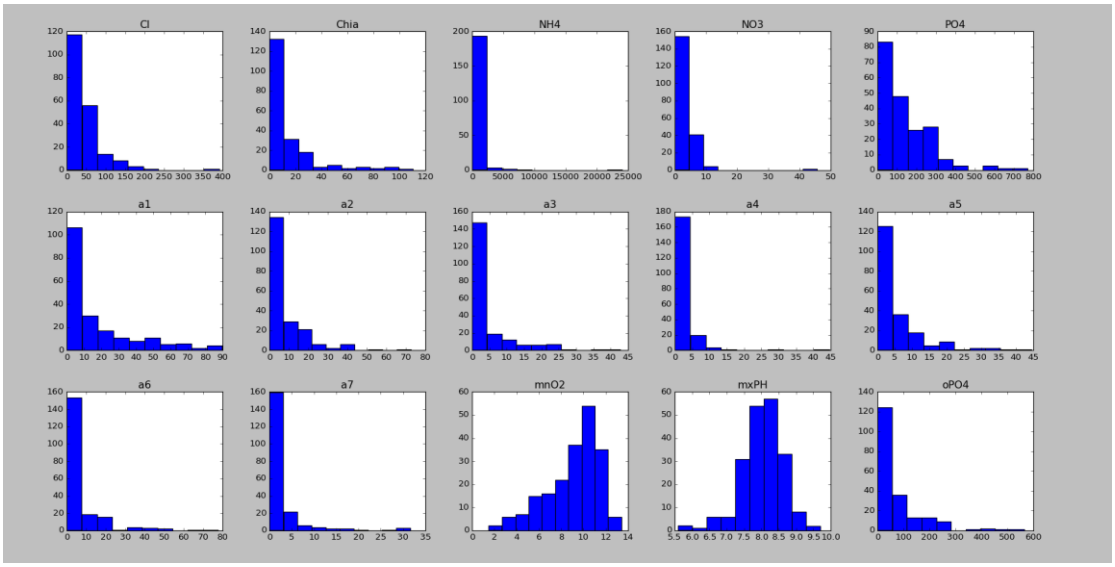
- 将缺失部分剔除



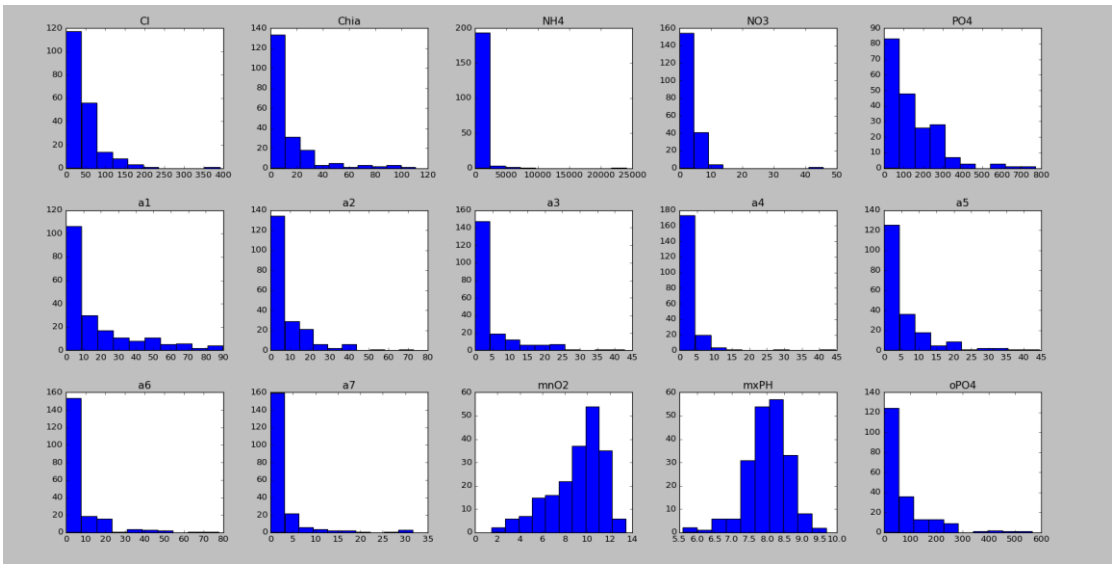
- 用最高频率值来填补缺失值



- 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。