数据挖掘大作业一：海藻数据的分析

学院：计算机学院　　学号：2120151036　　姓名：王丹

本次统计任务使用 Matlab 完成，其中包括了数据摘要、数据可视化以及缺失数据处理三项任务。

一、　数据摘要

(1) 读取文件，设定缺失数据的字符串为 XXXXXXX。

```
%从文件中读入数据
[season,size,speed,mxPH,mnO2,Cl,NO3,NH4,oPO4,PO4,Chla,a1,a2,a
3,a4,a5,a6,a7] =
textread('./Analysis.txt','%s%s%s%s%s%s%s%s%s%s%s%s%s%s%s%s%s%s%s
%s');
%根据数据构建元胞矩阵备用
%标称属性
biao = [season ,size,speed];
%数值属性
shu =
[mxPH,mnO2,Cl,NO3,NH4,oPO4,PO4,Chla,a1,a2,a3,a4,a5,a6,a7];
%全部数据
all =[biao,shu];
```

(2) 使用 tabulate 函数分析标称属性的数据摘要：

```
%统计标称属性的频数
freSeason=tabulate(all(:,1));
freSize=tabulate(all(:,2));
freSpeed=tabulate(all(:,3));
```

可以看到对标称数据的频数统计，如图 1 所示：

1	2
'winter'	62
'spring'	53
'autumn'	40
'summer'	45

'medium'	83
'high'	84
'low'	33

'small'	71
'medium'	84
'large'	45

图 1 海藻数据的标称属性频率信息

(3) 使用 min，max，median，mean，prctile 函数分析标称属性的数据摘要：

```
%统计数值属性的最小值、前四分位，中位数，平均值，后四分位，最大值
maxsh=max(sh);
minsh=min(sh);
meansh = mean(sh);
mediansh=median(sh);
q1sh=prctile(sh,25);
q3sh=prctile(sh,75);
```

数值型数据的最小值、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 2 所示：

	1	2	3	4	5	6	7	8
1	'Info'	'min'	'q1'	'median'	'mean'	'q3'	'max'	'NA'
2	'mxPH'	5.6000	7.7000	8.0600	8.0117	8.4000	9.7000	1
3	'mnO2'	1.5000	7.7250	9.8000	9.1147	10.8000	13.4000	2
4	'Cl'	0.2220	10.2553	31.0910	42.0447	57.4918	391.5000	10
5	'NO3'	0.0500	1.2960	2.6800	3.3086	4.4958	45.6500	2
6	'NH4'	5	37.8613	103	498.8195	226.9500	24064	2
7	'oPO4'	1	14.9003	39	73.2635	99.3333	564.6000	2
8	'PO4'	1	40.1668	102.5710	137.2319	213.7500	771.6000	2
9	'Chla'	0.2000	2.1125	5.8000	18.1102	20.8598	140.5170	12
10	'a1'	0	1.5000	6.9500	16.9235	24.8000	89.8000	0
11	'a2'	0	0	3	7.4585	11.4500	72.6000	0
12	'a3'	0	0	1.5500	4.3095	4.9500	42.8000	0
13	'a4'	0	0	0	1.9925	2.4000	44.6000	0
14	'a5'	0	0	1.9000	5.0645	7.5000	44.4000	0
15	'a6'	0	0	0	5.9640	6.9500	77.6000	0
16	'a7'	0	0	1	2.4955	2.4000	31.6000	0

图 2 海藻数据的数值属性摘要信息

由摘要信息可以看出，每个季节采集的样本数量相当，其中，大部分采自高速和中等流速的河流。缺失数据累计有 33 个，主要分布在 Cl 与 Chla，其中 Cl 缺失数为 10，Chla 缺失最多，为 12 个。

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

程序以 mxPH 为例，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。

```
%绘制直方图
  hist(sh(:,i));
  xlabel(name{i});
  ylabel('Value');
%绘制 QQ 图
  qqplot(sh(:,i));
  xlabel(name{i});
  ylabel('Value');
```

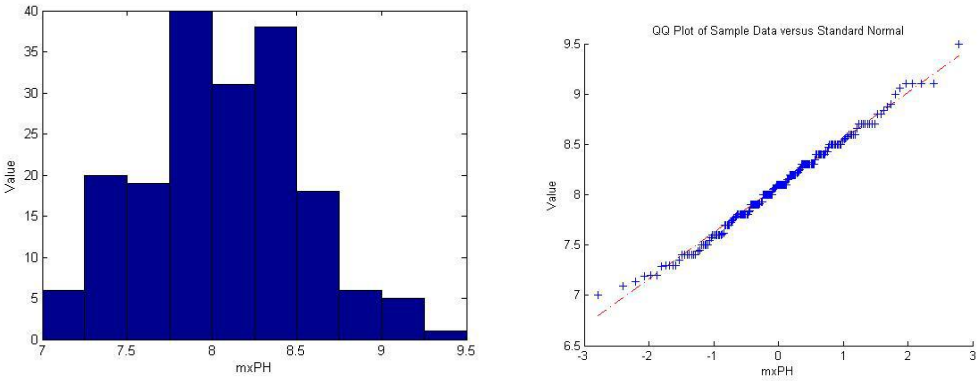


图 3 mxPH 的直方图与 QQ 图

由图 3 可看出，mxPH 的柱状图近似属于正态分布，其 QQ 图中，大部分点位于 QQ 线附近，所以可以认为，mxPH 属于正态分布。

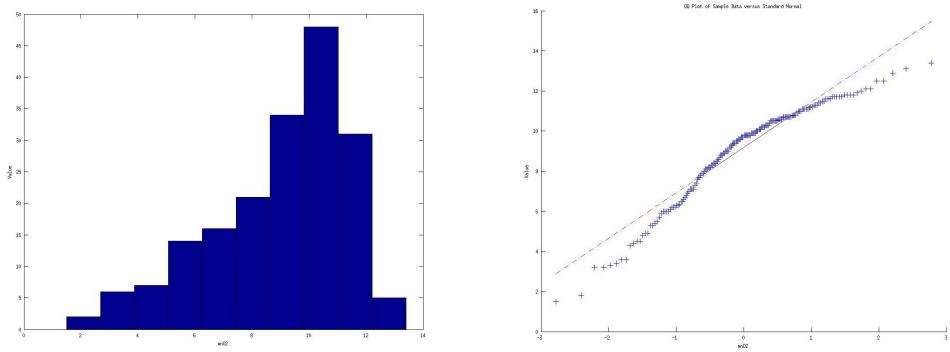


图 4 mnO2 的直方图与 QQ 图

由图 4 可看出，mnO2 的柱状图近似属于负偏态分布，其 QQ 图中，大量数据点已经在偏离 QQ 线，所以可以认为，mnO2 不属于正态分布。

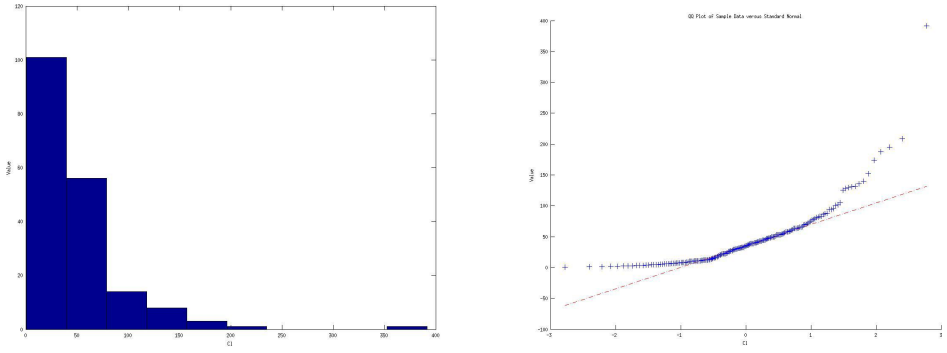


图 5 Cl 的直方图与 QQ 图

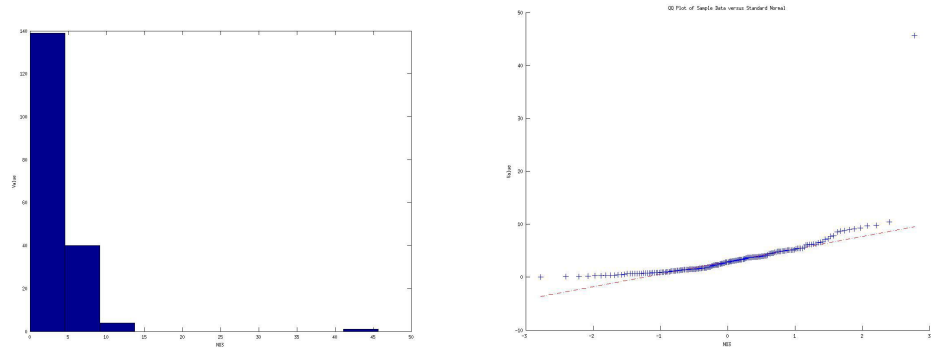


图 6 NO3 的直方图与 QQ 图

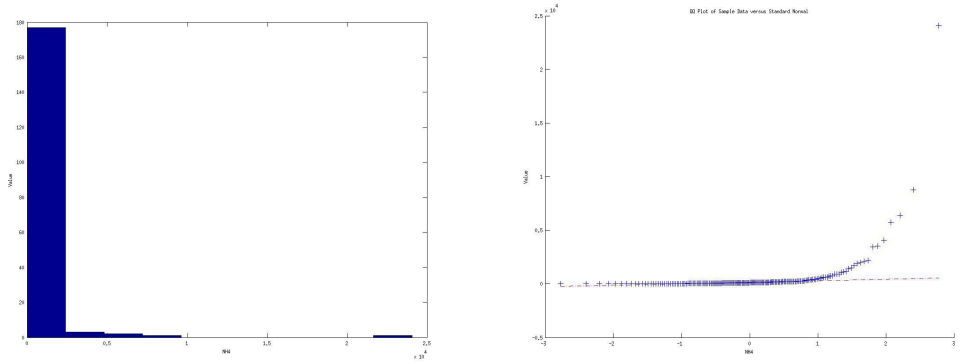


图 7 NH4 的直方图与 QQ 图

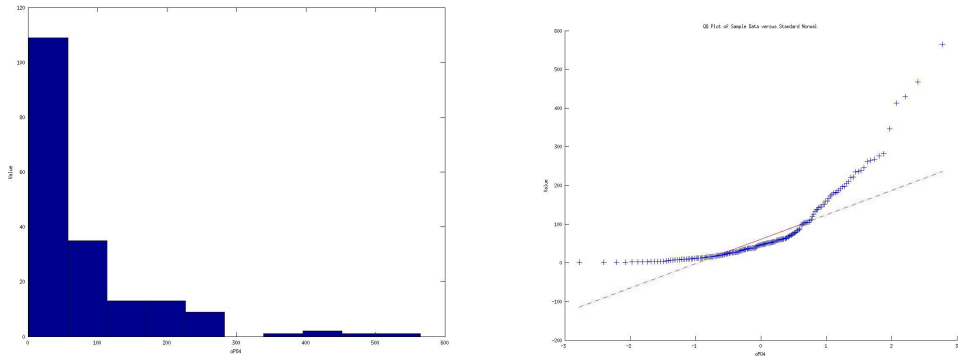


图 8 oPO4 的直方图与 QQ 图

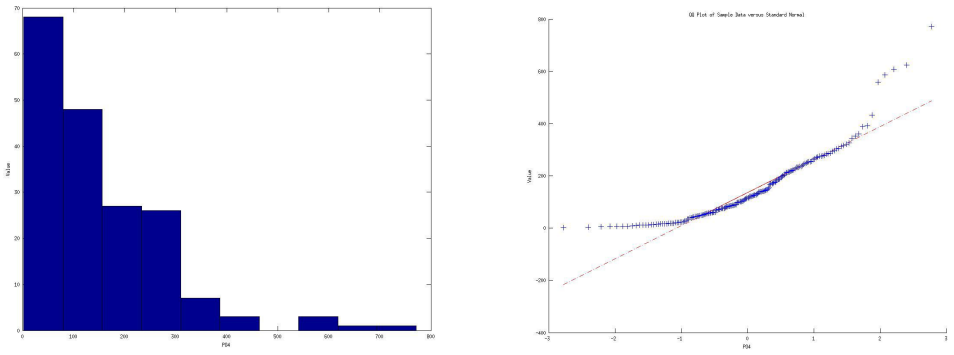


图 9 PO4 的直方图与 QQ 图

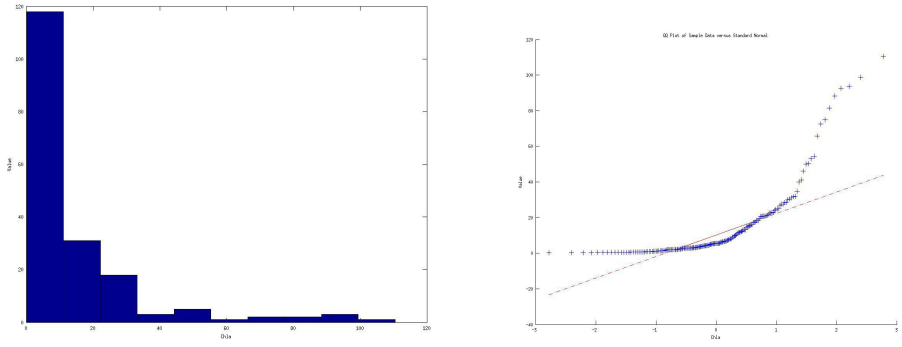


图 10 Chla 的直方图与 QQ 图

接下来对 Cl,NO3,NH4,oPO4,PO4,Chla 的直方图与 QQ 图进行了分析，其 QQ 图中，大量数据点远离 QQ 线，因此均不属于正态分布。

(2) 绘制盒图，识别离群点

此处以 mxPH 为例，绘制其盒图的命令如下：

```
%绘制盒图
  boxplot(sh(:,1));
  ylabel(name{i})
```

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。具体图参看图 12-19。

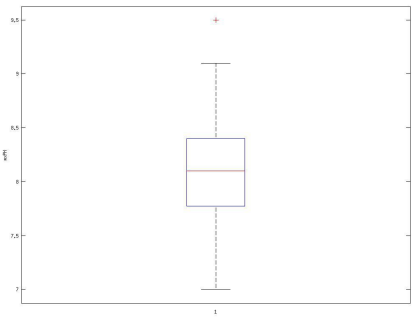


图 11 mxPH 盒图

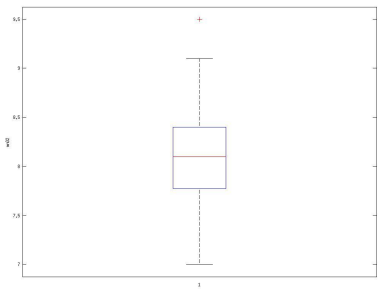


图 12 mnO2 盒图

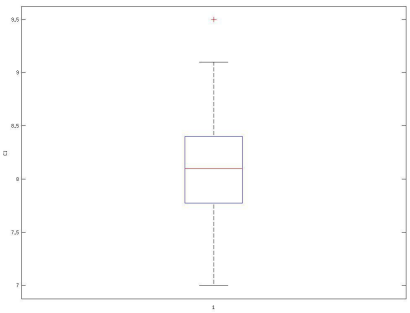


图 13 Cl 盒图

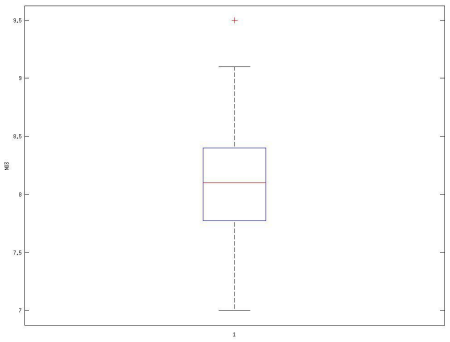


图 14 NO3 盒图

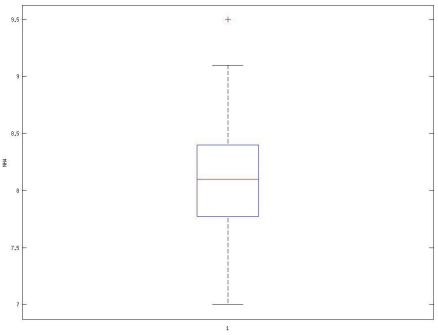


图 15 NH4 盒图

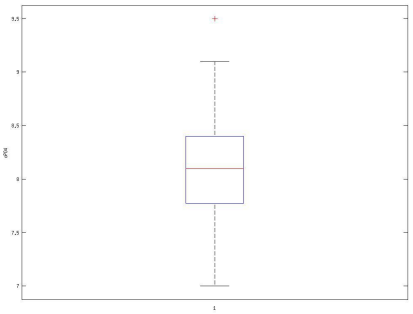


图 16 oPO4 盒图

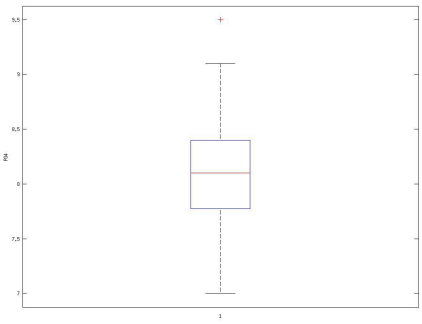


图 17 PO4 盒图

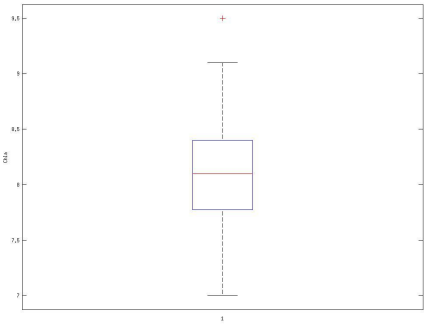


图 18 Chla 盒图

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。

(3) 对七种海藻，绘制其数量与河流大小的条件盒图

此处以 a1 海藻为例，绘制其与河流大小的条件盒图，命令如下：

```
boxplot(a1,G);
xlabel('River Size');
ylabel('a1');
```

此图可以反映 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 19-图 25 所示。

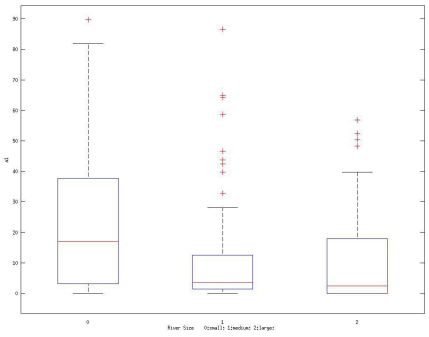


图 19 a1 海藻与河流大小的条件盒图

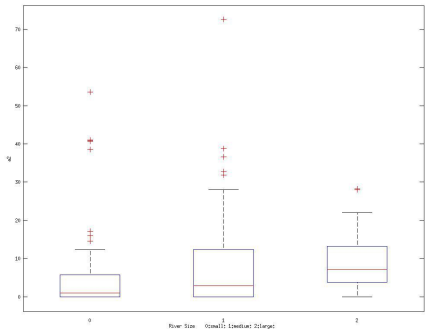
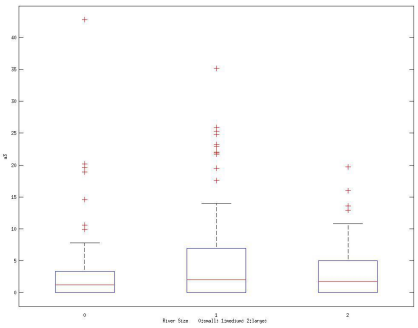


图 20 a2 海藻与河流大小的条件盒图



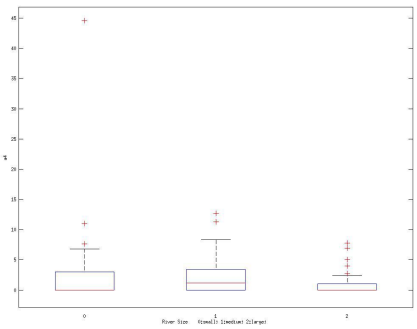
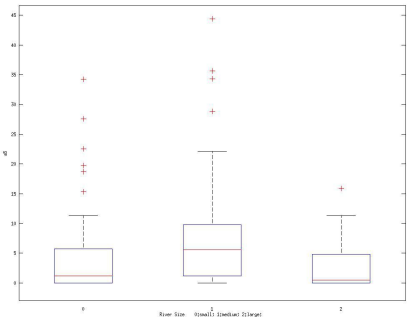


图 21 a3 海藻与河流大小的条件盒图　　图 22 a4 海藻与河流大小的条件盒图



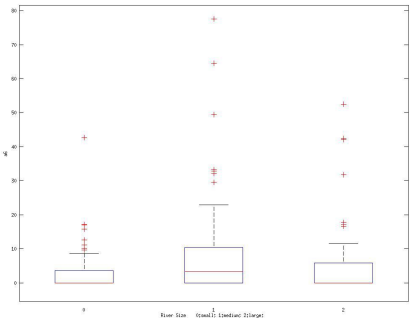


图 23 a5 海藻与河流大小的条件盒图　　图 24 a6 海藻与河流大小的条件盒图

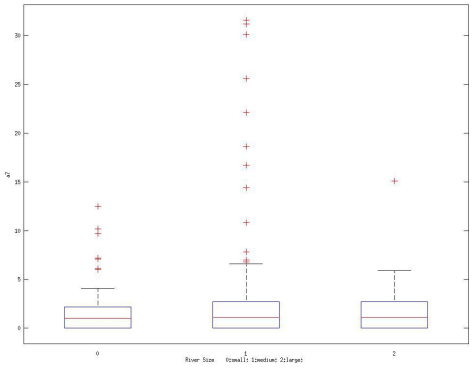
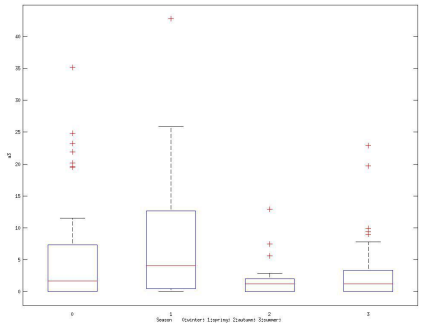


图 25 a7 海藻与河流大小的条件盒图

由这些图可分析出，在小型河流中，a1 有更高的频数，而 a3,a5,a6 在中型河流中更多一些。



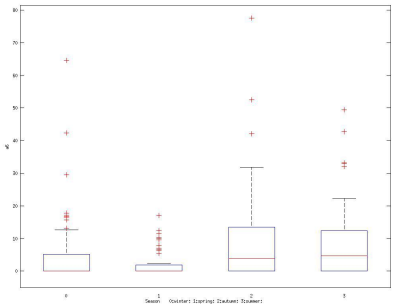
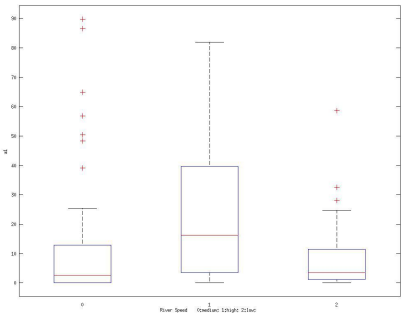


图 26 a3 海藻与季节的条件盒图　　图 27 a6 海藻与季节的条件盒图

绘制海藻与季节的条件盒图，其中 a3 与 a6 海藻表现对季节较为敏感，a3 海藻在春季与冬季较多，而 a6 海藻在秋季与夏季处于较高水平。



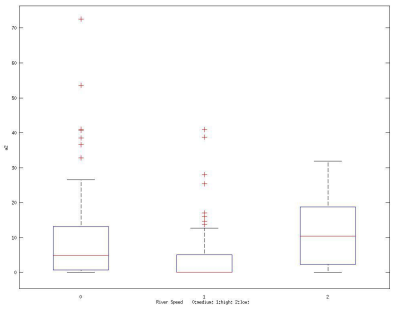


图 28 a1 海藻与流速的条件盒图　　图 29 a2 海藻与流速大小的条件盒图

绘制海藻与河流流速的条件盒图，其中 a1 与 a2 海藻表现对河流的流速表现敏感，a1 在较高流速中，表现出较多的数量，而 a2 海藻在高流速的河流中数量较少。

三、　数据缺失的处理

(1) 将缺失部分剔除

剔除缺失数据命令如下：

```
%剔出缺失数据
all1 =all;   %原始数据集
nm=all(d,:);   %存在缺失数据的数据集
all1(d,:)=[];    %已删除缺失数据后的数据集
```

总共剔除了 16 条数据。

(2) 用最高频数值来填补缺失值

```
dim=numel(all)/length(all);
freout=cell(dim,1);
for ic = 1:dim%列数
    for ir = 1:length(all)%行数
        temp(ir) = length(find(strcmp(all(:,ic),all(ir,ic))));
    end
    [~, id] = max(temp,[],1);
    freout(ic) = all(id(1,1),ic);
end
all2 = all;
for ir = 1:length(A)
    position =A(ir,:);
    all2{position(1),position(2)}=freout(position(2),1);      %用最高频率值填补缺失值后的数据集
end
```

使用该方法，对数值数据，采用最高频率值填充。

(3) 通过属性的相关关系来填补缺失值

使用 corrcoef 函数察看两个变量的相关性

```
relationship=corrcoef(shu1);
[~,index]=sort(relationship,2);
index=index(:,14);
```

可以看到结果，如图 30 所示

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	-0.1027	0.1471	-0.1721	-0.1543	0.0902	0.1013	0.4318	-0.1626	0.3350	-0.0272	-0.1844	-0.1073	-0.1727	-0.1703
2	-0.1027	1	-0.2632	0.1179	-0.0783	-0.3938	-0.4640	-0.1312	0.2500	-0.0685	-0.2352	-0.3798	0.2100	0.1886	-0.1046
3	0.1471	-0.2632	1	0.2110	0.0660	0.3793	0.4452	0.1430	-0.3592	0.0785	0.0765	0.1415	0.1453	0.1690	-0.0449
4	-0.1721	0.1179	0.2110	1	0.7247	0.1330	0.1570	0.1455	-0.2472	0.0200	-0.0918	-0.0145	0.2121	0.5440	0.0751
5	-0.1543	-0.0783	0.0660	0.7247	1	0.2193	0.1994	0.0912	-0.1236	-0.0379	-0.1129	0.2745	0.0154	0.4012	-0.0254
6	0.0902	-0.3938	0.3793	0.1330	0.2193	1	0.9120	0.1069	-0.3946	0.1238	0.0057	0.3825	0.1220	0.0033	0.0262
7	0.1013	-0.4640	0.4452	0.1570	0.1994	0.9120	1	0.2485	-0.4582	0.1327	0.0322	0.4088	0.1555	0.0532	0.0798
8	0.4318	-0.1312	0.1430	0.1455	0.0912	0.1069	0.2485	1	-0.2660	0.3667	-0.0633	-0.0860	-0.0734	0.0103	0.0176
9	-0.1626	0.2500	-0.3592	-0.2472	-0.1236	-0.3946	-0.4582	-0.2660	1	-0.2627	-0.1082	-0.0934	-0.2697	-0.2616	-0.1931
10	0.3350	-0.0685	0.0785	0.0200	-0.0379	0.1238	0.1327	0.3667	-0.2627	1	0.0098	-0.1763	-0.1868	-0.1335	0.0362
11	-0.0272	-0.2352	0.0765	-0.0918	-0.1129	0.0057	0.0322	-0.0633	-0.1082	0.0098	1	0.0334	-0.1416	-0.1969	0.0391
12	-0.1844	-0.3798	0.1415	-0.0145	0.2745	0.3825	0.4088	-0.0860	-0.0934	-0.1763	0.0334	1	-0.1013	-0.0849	0.0711
13	-0.1073	0.2100	0.1453	0.2121	0.0154	0.1220	0.1555	-0.0734	-0.2697	-0.1868	-0.1416	-0.1013	1	0.3886	-0.0515
14	-0.1727	0.1886	0.1690	0.5440	0.4012	0.0033	0.0532	0.0103	-0.2616	-0.1335	-0.1969	-0.0849	0.3886	1	-0.0303
15	-0.1703	-0.1046	-0.0449	0.0751	-0.0254	0.0262	0.0798	0.0176	-0.1931	0.0362	0.0391	0.0711	-0.0515	-0.0303	1

图 30 属性相关度

用这两个相关性最大的属性作相关分析，互相填补缺失数据。用如下代码获得其线性模型：

```
    c1=shu1(:,ir);
    c2=shu1(:,index(ir));
    c2=[ones(length(c2),1),c2];
    [b,~,~,~,~]=regress(c1,c2);
```

1	2	3	4	5	6	7	8	9
[7.9387;0.0101]	[8.5649;0.0296]	[21.0918;0.1623]	[2.6409;0.0014]	[-748.1477;379.9623]	[-17.6273;0.6542]	[47.0802;1.2712]	[-135.9798;18.5513]	[-3.7117;2.1099]

图 31 相关属性间线性模型参数

得到结果如图 31 所示.

```
all3=all;
for ir = 1:length(A)
    position =A(ir,:);
all3{position(1),position(2)}=num2str(pra{1,position(2)-3}(1,1)+pra{1,position(2)-3}(2,1)*str2n
um(all3{position(1),index(position(2)-3)+4}));      %用属性相关关系填补缺失值后的数据集
end
```

(4) 通过数据对象之间的相似型来填补缺失值

```
    dist =pdist2(shu2,va);
    [~,in] = sort(dist);
    in=in(1:10,1);
    all4{position(1),position(2)}=num2str(mean(shu1(in,position(2)-3))); %通过数据对象之
间的相似性来填补缺失值后的数据集
```