Data mining: data analysis seaweed

Schools: School of Computer Science
Student ID: 2820150087 Name: Elfatih abdalla Hassan

The statistical tasks using Matlab complete, including data summaries, data visualization and data processing missing Three tasks.

First, data summary:

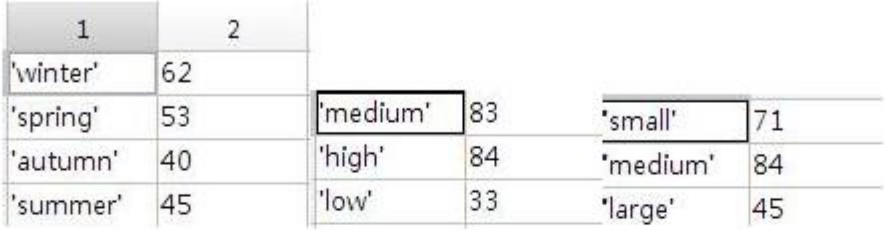
(1) to read the file, set the missing data string XXXXXXX.

```
% Read in data from a file
[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');
% Construct cellular matrix according to data backup
% Of nominal properties
biao = [season ,size,speed];
% Value of property
shu =
[mxPH,mnO2,Cl,NO3,NH4,oPO4,PO4,Chla,a1,a2,a3,a4,a5,a6,a7];
Of all of data
all =[biao,shu];
```

(2) using the nominal property tabulate function analysis data summary:

```
% Nominal frequency statistical properties
freSeason=tabulate(all(:,1));
freSize=tabulate(all(:,2));
freSpeed=tabulate(all(:,3));
```

We can see the data on the nominal frequency statistics, shown in Figure 1:



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

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

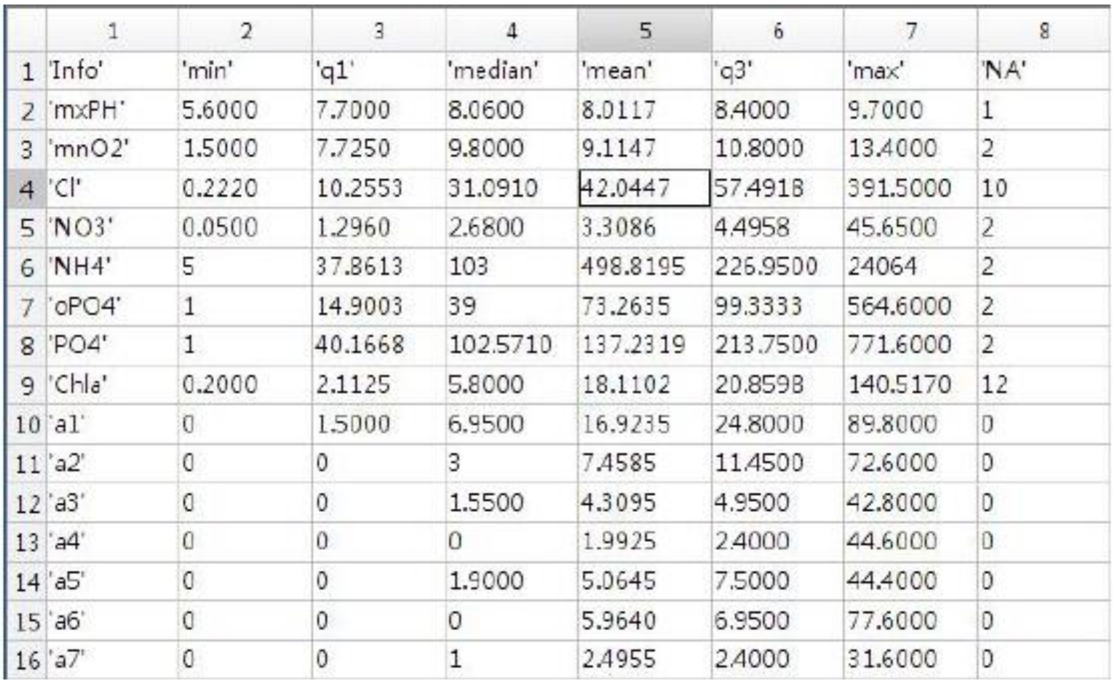
'small'	71
'medium'	84
'large'	45

Nominal frequency attribute information 1 seaweed data

(3) using the min, max, median, mean, prctile function analysis nominal attribute data summary:

```
% Minimum statistical value of the property, the former quartile, median, average, after quartile, max.
    maxsh=max(sh);
    minsh=min(sh);
    meansh = mean(sh);
    mediansh=median(sh);
    q1sh=prctile(sh,25);
    q3sh=prctile(sh,75);
```

The minimum numerical data, the first quartile, median, average, after quartile, and the maximum number of missing statistics, shown in Figure 2:



	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

numeric attribute data in Fig2 summary information seaweed As can be seen from the summary information, a considerable number of samples collected each season, in which most of the stream taken from the high-speed and medium Speed river. There are a total of 33 missing data, mainly in the Cl and Chla, wherein Cl missing number is 10, Chla missing Most, 12.

Second, data visualization

(1) on the value of property, draw a histogram and test its normal QQ plot Procedures to mxPH, for example, draws its histogram QQ FIG. Draw a histogram and the vertical axis is its frequency, its horizontal axis Distribution range. QQ plot, the solid red line for QQ line.

```
%Draw histograms
  hist(sh(:,i));
  xlabel(name{i});
  ylabel('Value');
  % QQ plot
  qqplot(sh(:,i));
  xlabel(name{i});
  ylabel('Value');
```

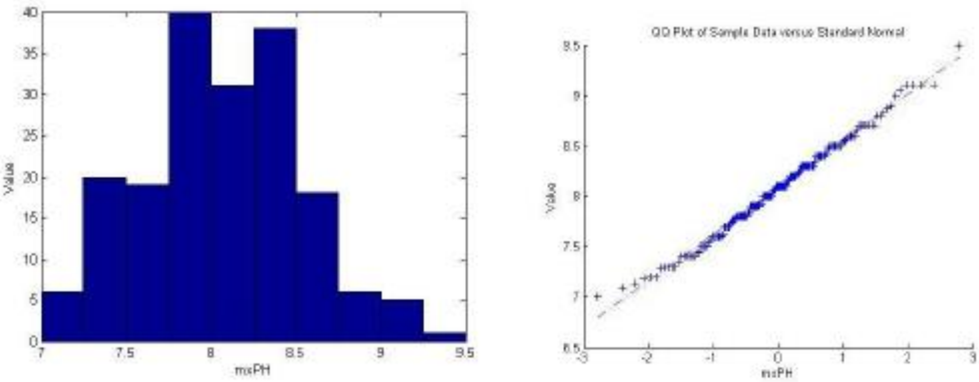


Figure 3 mxPH histograms and QQ plot As can be seen from Figure 3, mxPH histogram approximately normally distributed, the QQ plot, most point in the QQ line Nearby, it is considered that, mxPH normally distributed.

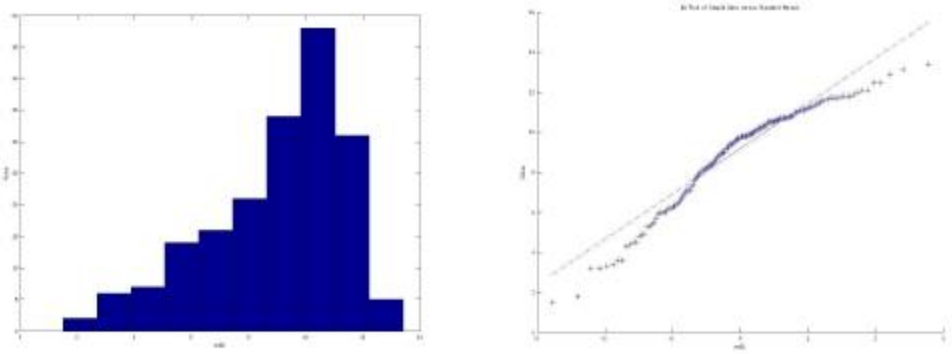


Figure 4 mnO2 histograms and QQ plot As can be seen from Figure 4, mnO2 histogram approximation are in negative skewness, which QQ plot, has a large number of data points Deviate from the QQ line, so it can be considered, mnO2 does not belong to the normal distribution.

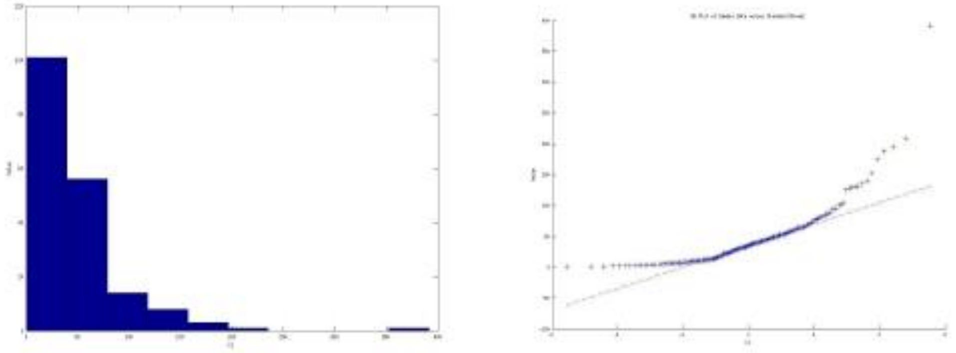
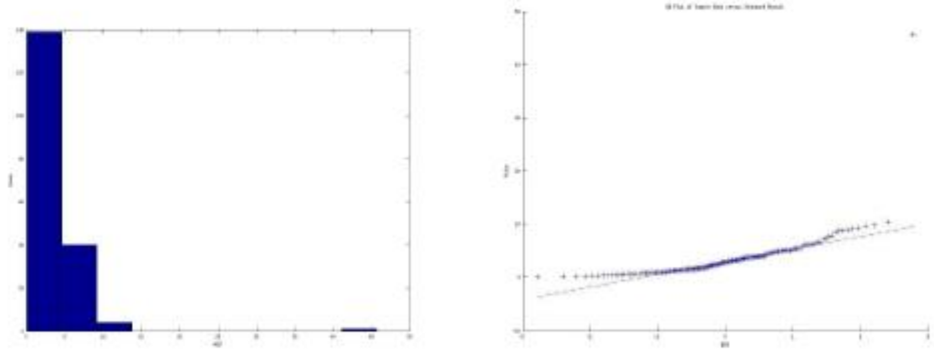
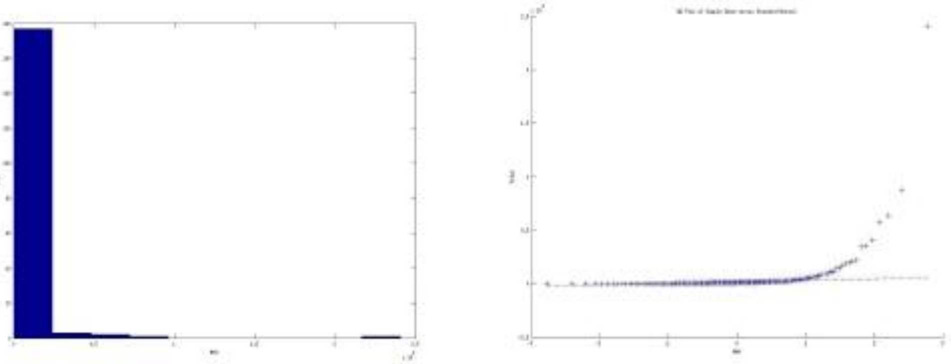


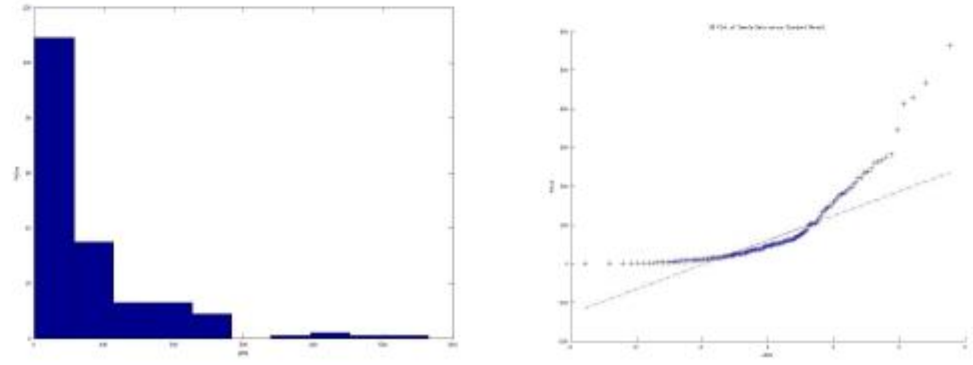
Figure 5 Cl histograms and QQ plot



Histogram and QQ of FIG. 6 NO3



Histogram and QQ of FIG. 7 NH4



Histogram and QQ of FIG. 8 oPO4

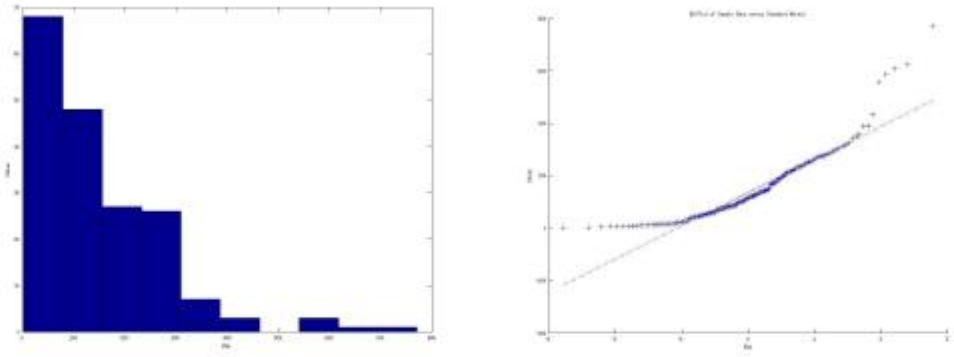


Figure 9 PO4 histograms and QQ plot

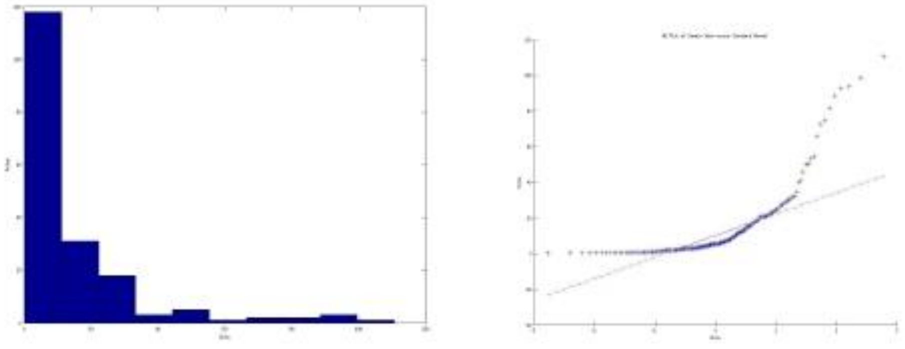


Figure 10 Chla histograms and QQ plot

Next to Cl, NO3, NH4, oPO4, PO4, Chla histograms and QQ plot was analyzed, its QQ plot, a large number of data points away from the QQ line, so not normally distributed.

(2) Draw box diagram, identify outliers

MxPH here for example, draws its box diagram command as follows:

```
% Figure Drawing Box
  boxplot(sh(:,1));
  ylabel(name{i})
```

Rug function draws each point on the vertical axis of the projected circumstances, abline mean data is plotted in the figure in dashed lines in a way. Each attribute box diagram, you can analyze the quantity and distribution of the outliers. Figure specific reference to Figure 12-19.

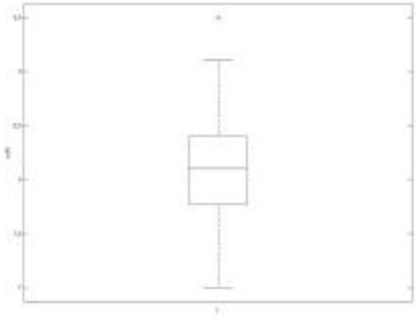


Figure 11 mxPH cartridge box diagram

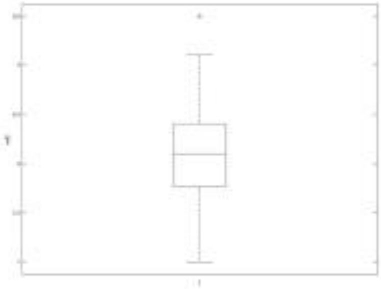


Figure 12 mnO2

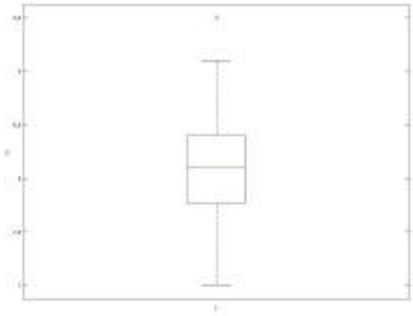


Figure 13 Cl cartridge box diagram

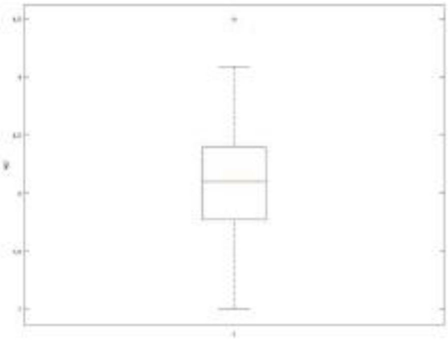


Figure 14 NO3

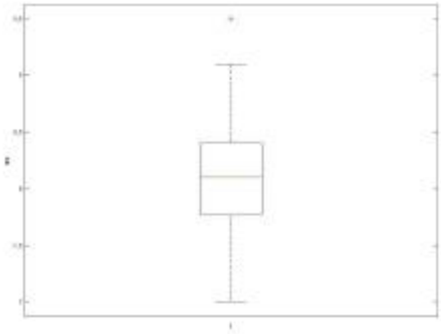


Figure 15 NH4 cartridge box diagram

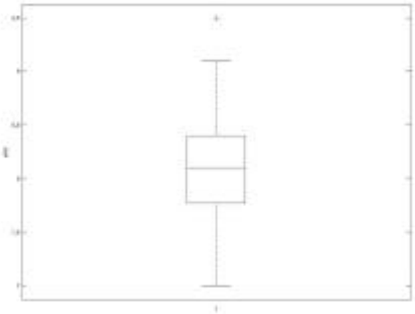
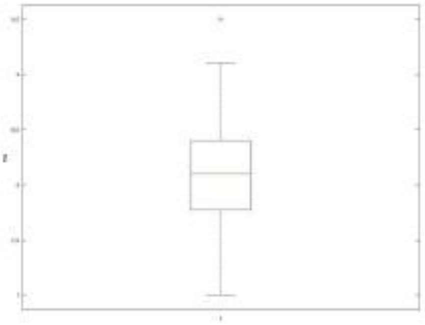


Figure 16 oPO4



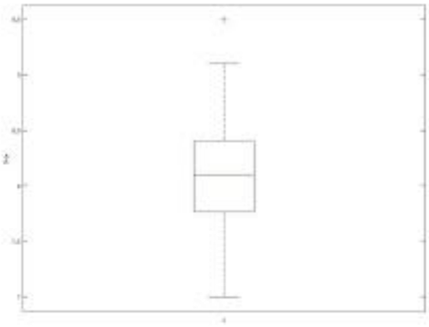


Figure 17 PO4 cartridge box diagram

Figure 18 Chla

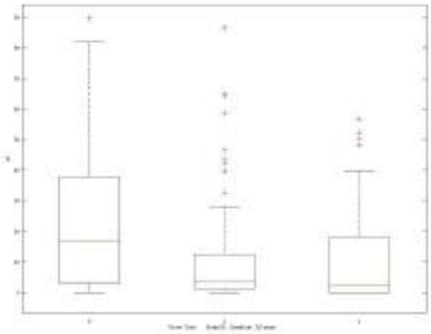
As can be seen from the figure, mnO2 more evenly distributed, and NO3 and NH4 in both a higher value outliers can be It can provide noise data or specific examples.

(3) on Seven seaweed, we draw the number and size of river conditions box diagram

Here to a1 seaweed, for example, draws its size and river conditions box diagram, the command is as follows:

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

This figure reflects a1 seaweed box showing the shape in different sizes river conditions. In turn draw a1-a7 algae conditions FIG box, as shown in Figure 19- 25.



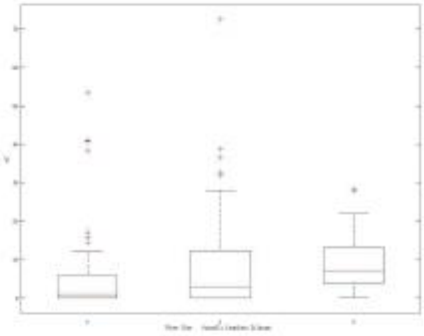
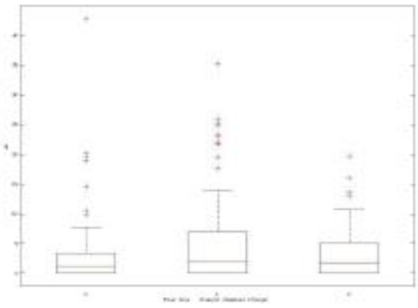


Figure 19 a1 algae and river conditions cartridge size

Figure 20 a2 algae and river conditions sized box diagram



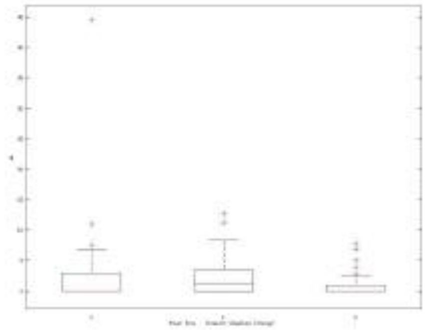
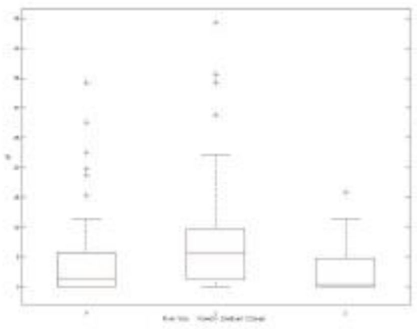


Figure 21 a3 algae and river conditions cartridge size

Figure 22 a4 size of algae and river conditions box diagram



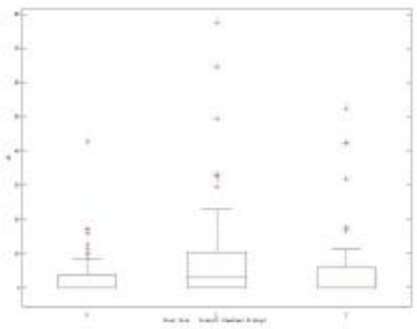


Figure 23 a5 algae and river conditions cassette size

Figure 24 a6 algae and river conditions sized box diagram

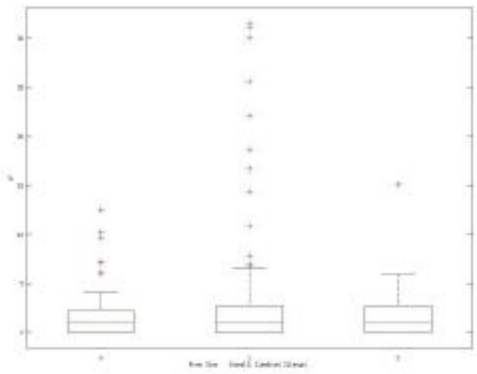
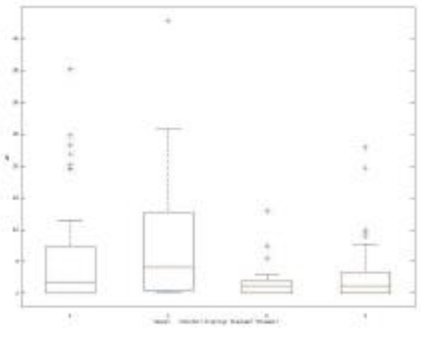


Figure 25 a7 algae and river conditions sized box diagram

From these figures can be analyzed in a small river, a1 higher frequency, and a3, a5, a6 more in the medium rivers some.



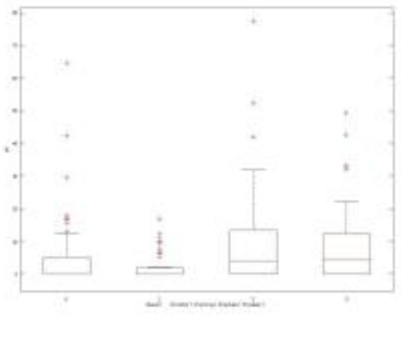
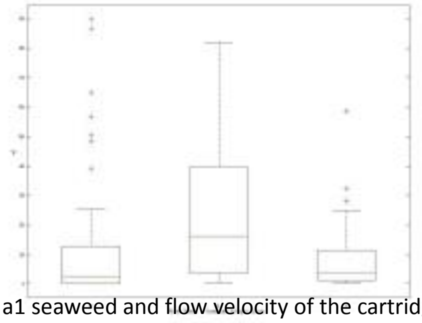


Figure 26 a3 algae and seasonal conditions cartridge

Figure 27 a6 algae and seasonal conditions box diagram

Drawing algae and seasonal conditions box diagram in which the a3 a6 with the performance of the season more sensitive algae, seaweed in the spring a3 And more in winter, while in summer and autumn in seaweed a6 at a high level.



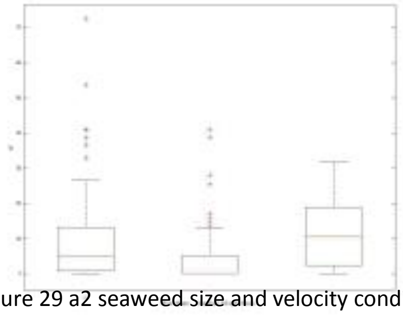


Figure 28 a1 seaweed and flow velocity of the cartridge Figure 29 a2 seaweed size and velocity conditions box diagram

Drawing algae and river flow conditions box diagram, where a1 and a2 seaweed performance of the river flow rate performance sensitive, a1 At higher flow rates, showing large quantities and at high velocity a2 seaweed small number of rivers.

Third, the missing data processing

(1) Excluding the missing parts

Excluding missing data command is as follows:

```
% Excluding missing data
all1 =all;    % Of the original data set
nm=all(d,:);    % There is a missing data set of data
all1(d,:)=[];     % after the missing data has been deleted Dataset
```

Excluding a total of 16 pieces of data.

(2) with the highest frequency value to fill in missing values

```
dim=numel(all)/length(all);
freout=cell(dim,1);
for ic = 1:dim% Number of columns
    for ir = 1:length(all)% Rows
        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);    % With the highest frequency of missing data values after
paddingset
    end
```

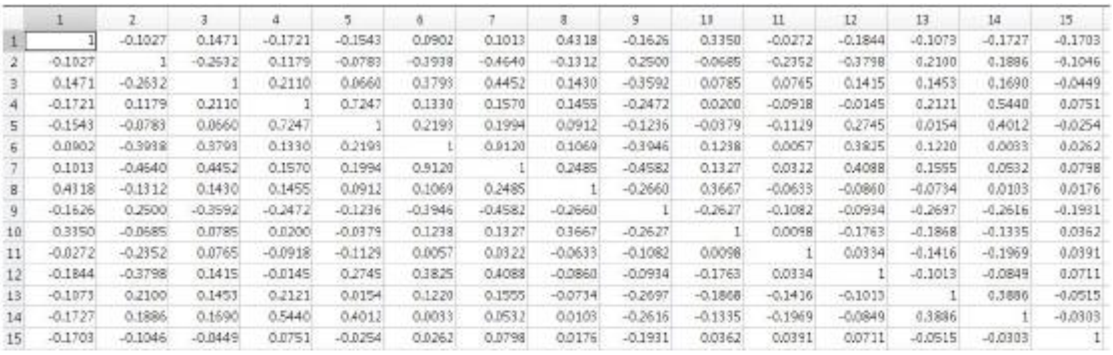
Using this method, the value of the data, with the highest frequency value is filled.

(3) to fill in missing values through correlation properties

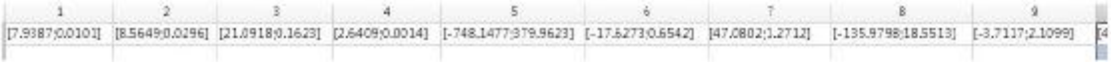
Use corrcoef View function of two variables correlation

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

You can see the results, shown in Figure 30



	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



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]	[4

30 degrees attributes

Correlation analysis with these two most relevant attributes, each to fill in missing data. Obtain its linear model with the following code type:

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

FIG. 31 between the relevant properties of linear model

The results obtained shown in Figure 31.

```
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}));    % Correlation with the attribute data set to fill the missing values after
end
```

(4) to fill the missing values by a similar type of data objects

```
dist =pdist2(shu2,va);
[~,in] = sort(dist);
in=in(1:10,1);
all4{position(1),position(2)}=num2str(mean(shu1(in,position(2)-3))); % Through the data object
Similarity between data sets to fill in missing values after
```