海藻数据的分析报告

姓名：宁小东

学号：2120151024

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。

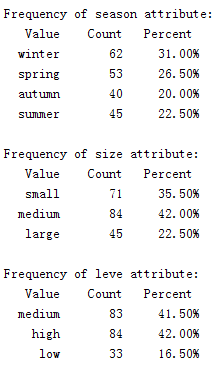
1. 数据分析

分析程序的具体说明请参考readme。程序的编写语言是Matlab，编写平台是Matlab R2014b，共包含5个文件。

* 1. 数据摘要

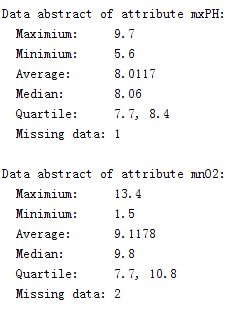
利用分析程序对数据进行了摘要以及可视化处理，数据摘要结果如下：

首先为标称属性的频数统计结果：



3个属性分别为season、size和level（即季节、河流大小和流速）。程序统计了各种属性下不同取值的频数（count）和所占比例即频率（percent）。例如season属性下的autumn取值频数为40。

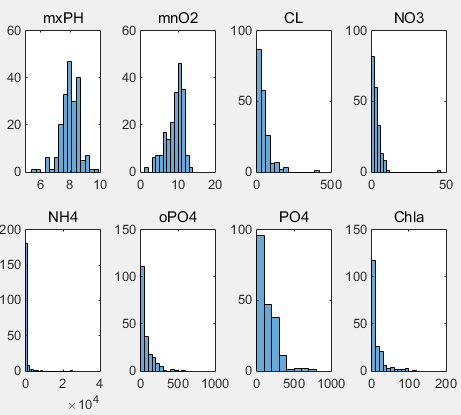
对于数值属性，程序也进行了各种统计，例如对如下的属性的数据摘要：



如图为数据摘要的部分（仅有mxPH和mnO2两属性），其余摘要与上图同理，包含最大、最小、均值、中位数、两个四分位数以及缺失数据的个数。

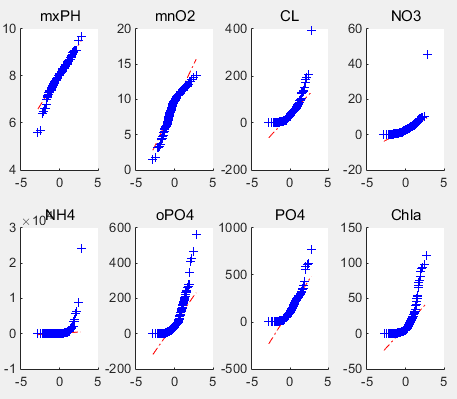
* 1. 数据可视化

对数值属性数据进行可视化处理，首先显示各个数值属性的直方图：



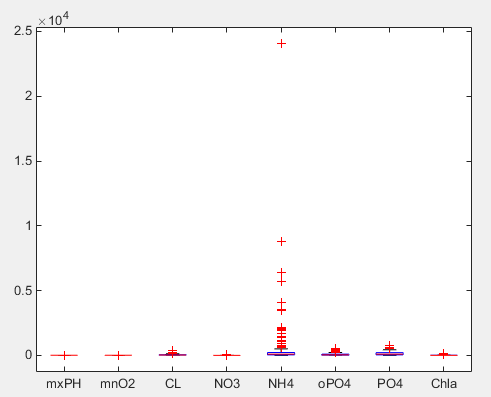
能看出mxPH属性相对近似于正态分布，mnO2属性则偏向于负轴，其余属性值大部分集中在值较低的区间，而NH4属性极值差距非常大。

再观察属性的QQ图：

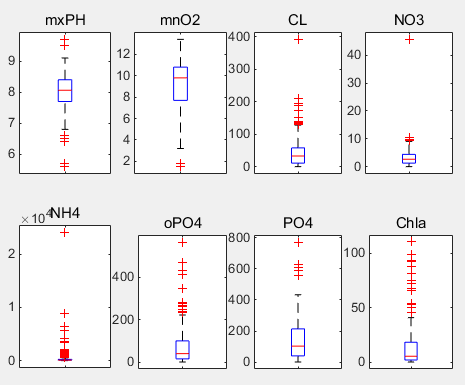


能够看出仅前两个数据部分拟合正态分布，其余数据差距较大。

8个数据的盒图为：

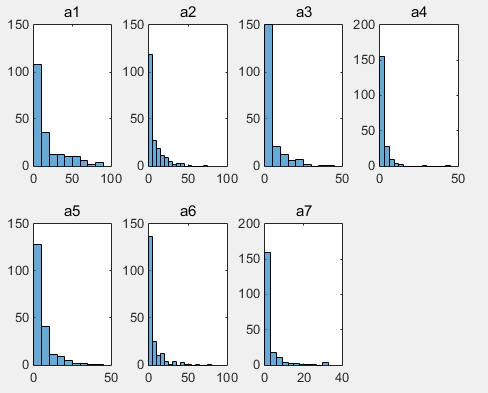


其中位于盒中的红线为中位数，盒顶端和低端为四分位数，盒两端的红线为最大、最小值，加号为离群点。其中本实验NH4属性由于极值过大，产生了较偏的离群点，导致整幅图效果下降。可以单独观察其盒图为：

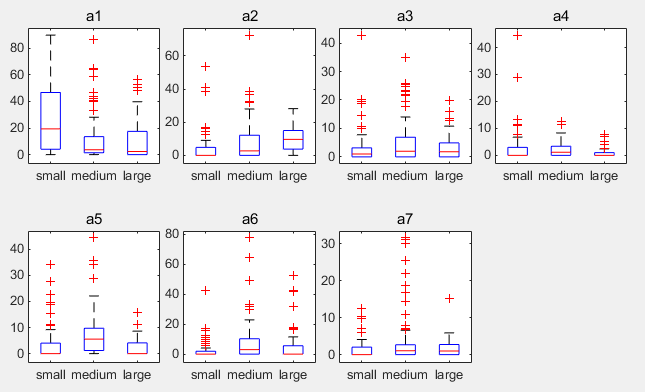


能看出CL，NH4，Chla属性都有较多的离群点。

海藻的数量用如下直方图统计：



海藻分布大致集中在低区间。其相对于size的条件盒图为：



能够看出各个海藻在不同河流规模下有不同的分布，例如a1在小河流中数量均值、四分位数、极大值都明显高于中等规模以及大河流。

* 1. 数据集预处理

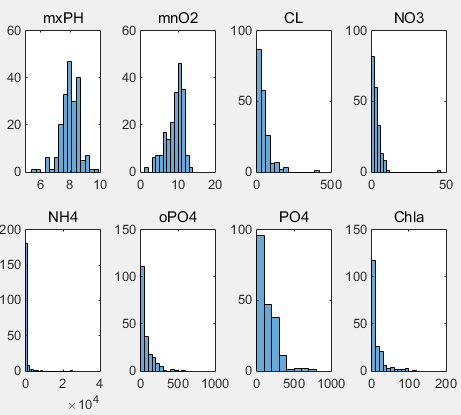
对数据进行3种（剔除属于缺省处理方式）处理方式的结果如3个txt文件所示。分别使用了最大值、按属性填补以及按样本相似性填补。三种方式的做法为：

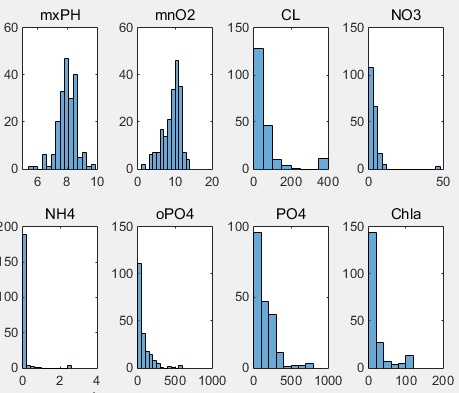
使用最大值：对于某一属性下的缺失，使用该属性的极大值代替缺失的数据；

按属性填补：计算两个属性的相关性，相关性越大表明可以根据另一属性推断缺失属性的值。通过另一属性的回归分析，计算当前的缺失值；

按相似性填补：计算两个样本的相似程度，越相似证明越可以使用该样本推断当前含缺失值的样本。

预处理后的数据集也可以进行可视化，例如将极值填补后的结果与原直方图相比：





上图为处理前，下图处理后，能看出NO3，NH4，Chla有明显的变化。

其余可视化属性与此同理。