**数据探索性分析与预处理-海藻数据的分析**

计算机科学与技术 李凯霞 2120151003 硕士

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season，河流大小size和河水速度speed，剩下的8个变量是水样的化学参数，其中缺失值用XXXXXXX代替：

* 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。数据样本如下图:

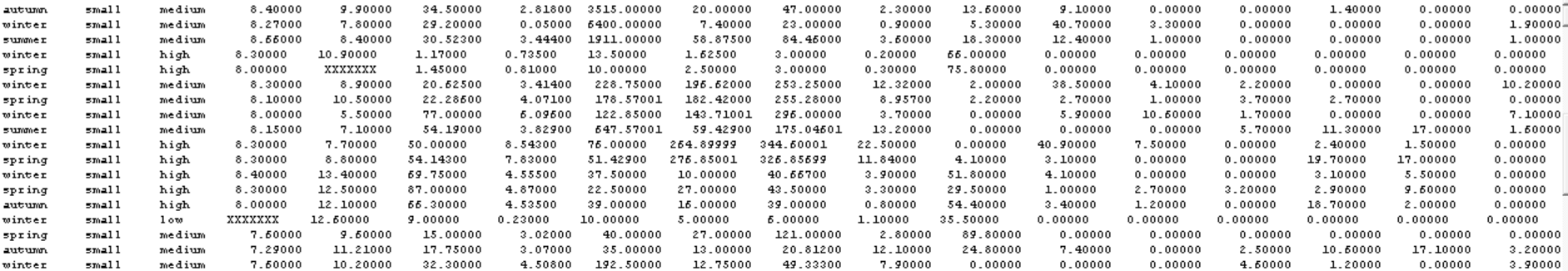


图1 数据样本示例

1. **数据分析**

**1、数据摘要**

对标称属性，给出每个可能取值的频数，数值属性，给出最大Max、最小Min、均值Mean、中位数Median、四分位数1st Qu/3rd Qu及缺失值NA的个数。本实验数据摘要结果保存在..\result\summary.txt中，如下图:

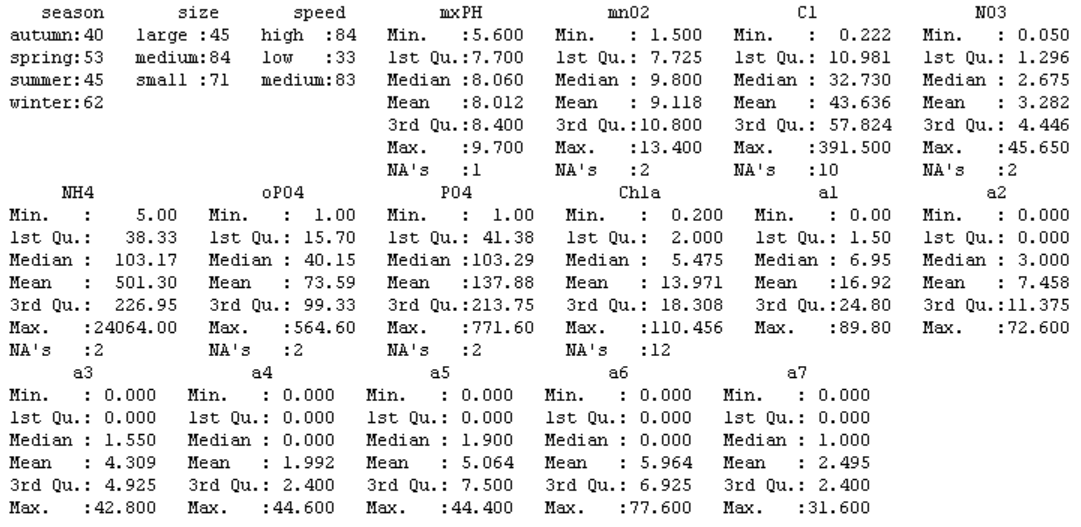


图2 数据摘要运行结果

**2、数据的可视化**

针对数值属性，绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。Q-Q图绘制变量值和正态分布的理论分位数（实线）的散点图，同时给出正态分布的95%置信区间的带状图（虚线）。

在本实验中，通过直方图观察mxPH非常符合正态分布，通过Q-Q图检验，其数据点大多数都在95%置信区间内。通过直方图观察NH4不符合正态分布，通过Q-Q图检验，其数据点大多数都在95%置信区间外。同样观察得，mnO2比较符合正态分布，其余参数均不符合。mxPH、NH4的直方图及Q-Q图如下。

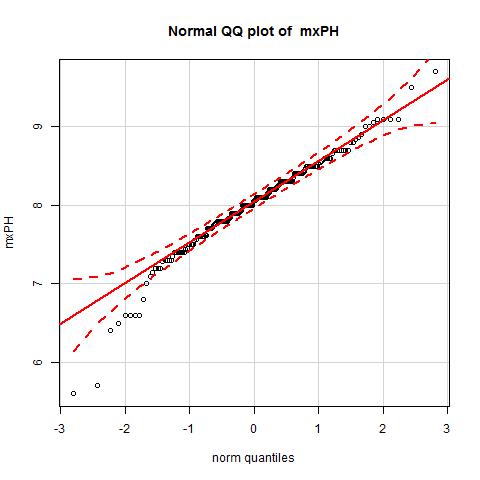
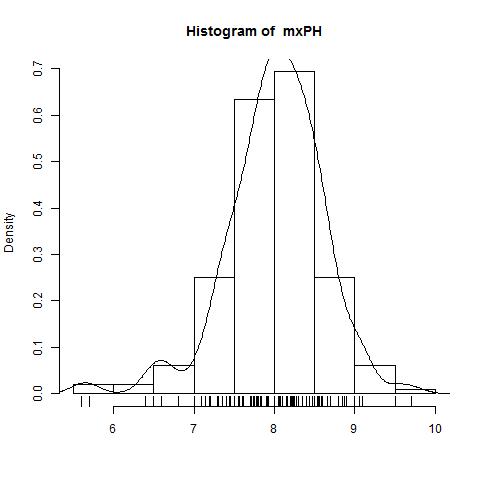


图3 变量mxPH的直方图和Q-Q图

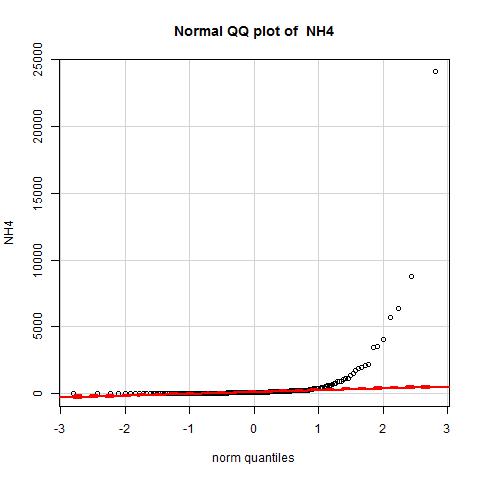
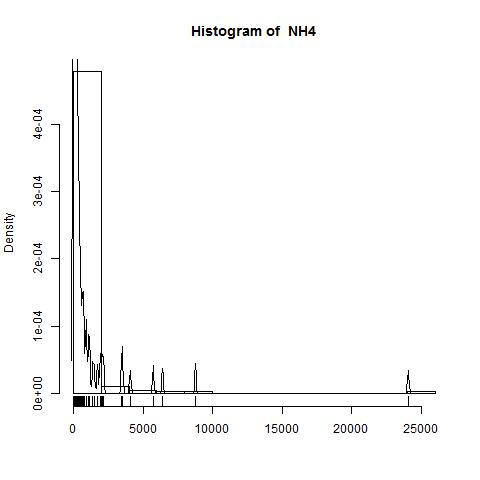


图4 变量NH4的直方图和Q-Q图

绘制盒图，对离群值进行识别。盒图的边界代表变量的第一个四分位数和第三个四分位数，而框内的水平线是变量的中位数。设r是变量的四分位距，盒图上方的小横线是小于或等于第三个四分位数加1.5\*r的最大的观测值，而盒图下方的小横线是大于或等于第一个四分位数减去1.5\*r的最小观测值。盒图上方小横线上面或者下方小横线下面的数据点通常认为是离群点。mnO2的盒图如下。

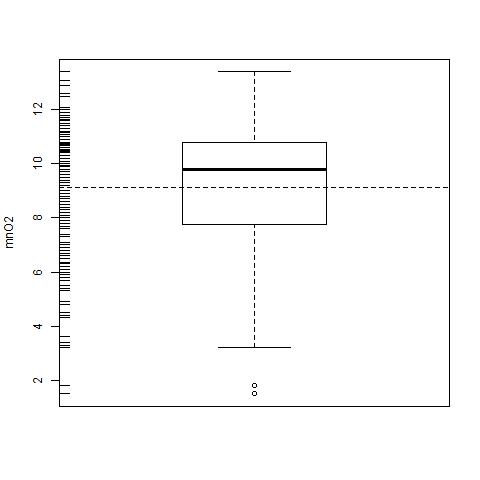


图5 变量mnO2的盒图

对7种海藻，分别绘制其数量与标称变量，如size的条件盒图。条件绘图是依赖于某个特定因子的图形表示。因子是一个取值为有限集合的标称变量。例如，对于标称变量size的某个特定值的样本子集，可以研究标称变量size如何影响变量a1值的分布。海藻a1基于size的条件盒图如下。

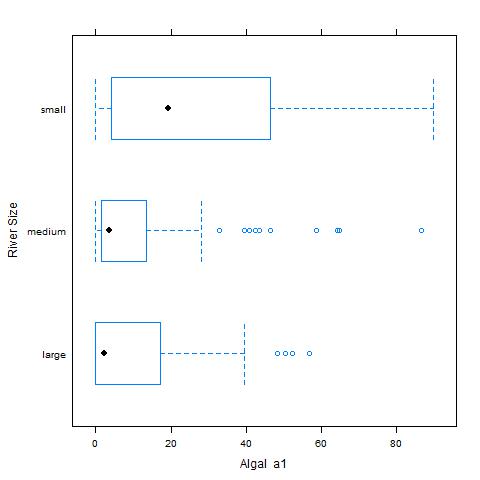


图6 海藻a1基于size的条件盒图

为了便于对比，本实验原始数据输出为..\result\Analysis\_Orginal.csv文件中。可视化结果在..\result\picture\Orginal中

**3、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:

1. **将缺失部分剔除**

剔除缺失值，优点是实现简单，当缺失记录所占比例在可用数据集中非常小的时候，可以选择该方法。在实际应用中，可以首先选择剔除缺失值较多的样本，再采用其他方式填补剩余的缺失值。

在本实验数据中，还有缺失项的样本共有16个，剔除后，剩余184个。输出结果在..\result\Analysis\_Delete.csv文件中，可视化结果在..\result\picture\Delete中。

1. **用最高频率值来填补缺失值**

采用众数填补缺失值，优点简单速度快，但可能存在较大的数据偏差。

最高频率填充输出结果在..\result\Analysis\_Frequency.csv文件中，可视化结果在..\result\picture\Frequency中。

1. **通过属性的相关关系来填补缺失值**

在本实验中，200个样本值中，16个样本有缺失项，8个数值属性都有缺失值。观察数据后发现样本62和199含有太多的缺失值，为不具备参考的样本，剔除后，NH4和NO3就没有缺失值了。剔除后有缺失值的属性为mxPH、mn02、Cl、PO4、Chla，属性NH4、NO3、oPO4无确实值。通过探寻变量之间的相关关系，使用变量的相关关系填补缺失值。在本实验计算所得变量之间的相关值矩阵结果保存在..\result\CorrelationMatrix.txt中，如下图:

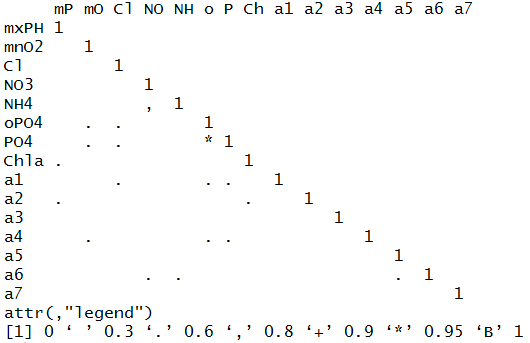


图7 变量之间的相关值矩阵

通过变量之间的相关值矩阵，发现变量PO4和oPO4的相关性>0.9，非常相关；NH4和NO3的相关性>0.6，比较相关(剔除62、199后不存在缺失项，不用填补)；oPO4和mnO2、oPO4和Cl、Chla和mxPH的相关性>0.3，一般相关。本实验中相关系数的计算结果保存在..\result\CorrelationCoefficient.txt中。

首先利用PO4和oPO4的相关性，填补PO4的缺失值。经过计算，PO4和oPO4的线性相关公式为：PO4=42.897+1.293\*oPO4。

还有mxPH、mn02、Cl、Chla含有缺失值，可以通过探索数值属性与标称变量之间的关系，来填补缺失值。这种分析方法比较繁琐，但是可以应用到少量名义变量的较小数据集的分析中。如mnO2的缺失项为season=spring，size=small，speed=high，通过图示观察，该条件下(图中3行2格)数据主要集中在8-10之间，靠近均值Mean=9.118，所以采用均值填补该缺失项。但是mxPH在这三个条件下的变化不明显。季节、河流大小和速度引起的mnO2、mxPH的变化，如下图：

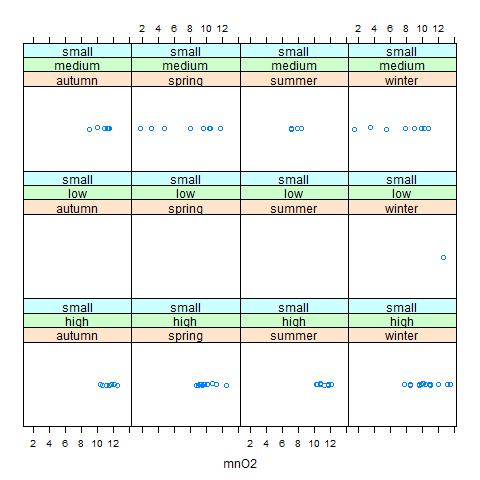


图8 变量mnO2在季节、河流大小和速度条件下的分布

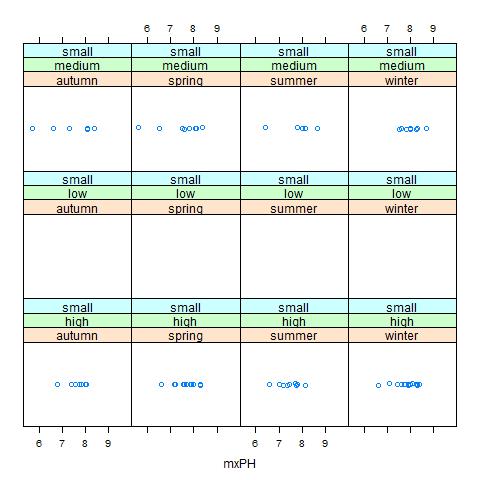


图9变量mxPH在季节、河流大小和速度条件下的分布

同时由于Cl、Chla的缺失项较多，观察结果不明显。本实验仍采用mnO2和oPO4、Cl和oPO4、Chla和mxPH的线性相关性来填补缺失项，由于相关性较小，会存在偏差，这样做的目的是为了对比填充后的结果。经过计算，Chla和mxPH的线性相关公式为:Chla=-139.4+19\*mxPH，mxPH=7.92896+0.01047\*Chla。mnO2和oPO4的线性相关公式为:mnO2= 9.93341-0.01093\*oPO4。Cl和oPO4的线性相关公式:Cl= 28.4771+ 0.1992\*oPO4。

相关性填充输出结果在..\result\Analysis\_Correlation.csv文件中，可视化结果在..\result\picture\Correlation中。

1. **通过数据对象之间的相似性来填补缺失值**

相似性经常由描述观察值的多远度量空间的变量所定义，在本实验中，采用最常用的欧式距离，其公式定义为 。

采用欧氏距离寻找与任何含有确实值的样本最相似的10个样本，并用它们来填补缺失值。采用这些最相似数据的加权平均，权重的大小随着距待填补缺失值的样本的距离增大而减小。本实验中，采用高斯核函数从距离获得权重，设距离为d，则其权重为。

相似性填充输出结果在..\result\Analysis\_Similarity.csv文件中，可视化结果在..\result\picture\Similarity中。