China's Main River Water Quality Structure Analysis

Ying-ying Zhang[1], Ting Zheng[2]

[1)] Department of Statistics and Actuarial Science, Chongqing University, Chongqing, China
[2)] Department of Statistics and Actuarial Science, Chongqing University, Chongqing, China

***Abstract*—This article analyses five river basins' water qualities: the Song Hai Liao river basin, the Huaihe river basin, the Yellow river basin, the Yangtze river basin, and the pearl river basin. We consider four water pollution indicators: PH, DO, CODMn, and NH3-N. First, the result of single factor analysis of variance indicates that the four water pollution indicators among the river basins all have significant differences. Then, the result of the multi-class distance classification method indicates that under the condition of the water quality is divided into 6 classes, 4 indicators have the correct rate of 63.64%, and under the condition of the water quality is divided into 5 types, 4 indicators have the correct rate of 78.79%. After that, by drawing a pie chart of the river basins' water quality distribution, we obtain the ranking of the river basins pollution levels: the Yellow River basin > the Huaihe river basin > the Song Hai Liao river basin > the pearl river basin > the Yangtze river basin. Finally, by collecting 2004, 2007, 2010, 2013 four years data, we analyze the changes of the river basins' water quality structure from year to year, and give some suggestions.**

***Keywords*—river basin, water quality structure, single factor analysis of variance, Kruskal-Wallis rank sum test, multiple t test, multi-class distance classification method**

我国主要流域水质结构分析

张应应　郑婷

重庆大学，重庆，中国

摘　要　本文分析 5 个流域：松海辽流域、淮河流域、黄河流域、长江流域、珠江流域的水质情况，考虑 4 个水质污染指标：PH、DO、CODMn、NH3-N。首先，单因素方差分析结果显示各流域间的 4 个水质污染指标都存在显著差异。其次，多分类距离判别法结果显示在水质分为 6 类的情况下 4 个指标能代表 63.64%的正确率，在水质分为 5 类的情况下 4 个指标能代表 78.79%的正确率。再次，通过画出各流域的水质分布饼图，得出各流域污染程度排序为：黄河流域＞淮河流域＞松海辽流域＞珠江流域＞长江流域。最后，通过收集 2004，2007，2010，2013 这四年的数据分析各流域各年水质结构的变化，并给出相关建议。

关键词　流域，水质结构，单因素方差分析，Kruskal-Wallis 秩和检验法，多重 t 检验法，多分类距离判别法

1. 引言

我国大小河川总长 42 万公里，湖泊 7.56 万平方公里，占国土总面积的 0.8%，水资源总量 28000 亿立方米，人均 2300 立方米，只占世界人均拥有量的 1/4，居 121 位，为 13 个贫水国之一[1]。目前中国 640 个城市有 300 多个缺水，2.32 亿人年均用水量严重不足。如此珍贵的水资源，目前却被严重污染，我国污水的处理能力只占 20%左右，各大城市地下水不同程度受到污染。我国七大水系：长江，珠江，松花江，黄河，淮河，海河，辽河，其中有一半河段受到污染，目前的污染状况已不可忽视。

本文数据来自中华人民共和国环境保护部数据中心，选取的是 2013 年 5 月 14 日发布的全国主要流域重点断面 2013 年第 19 周水质状况数据。网址：http://datacenter.mep.gov.cn/report/getCountGraph.do?type=runQianWater

2. 指标分析

本文中涉及到的水质污染指标主要有 4 个：PH、DO、CODMn、NH3-N。依据地表水水域环境功能和保护目标，五类水的说明如下：

Ⅰ类：主要适用于源头水、国家自然保护区；

Ⅱ类：主要适用于集中式生活饮用水地表水源地一级保护区、珍稀水生生物栖息地、鱼虾类产卵场、仔稚幼鱼的索饵汤等；

Ⅲ类：主要适用于集中式生活饮用水地表水源地二级保护区、鱼虾类越冬场、洄游通道、水产养殖区等渔业水域及游泳区；

Ⅳ类：主要适用于一般工业用水区及人体非直接接触的娱乐用水区；

Ⅴ类：主要适用于农业用水区及一般景观要求水域。

首先，为了有一个直观的了解，运用 R 软件给出各类污染指标的线图，见图 1 和图 2。

从图 1 中可以看出，各流域各监测点 PH 值大致都在 6—9 之间。综看各个流域，大多监测点的 PH 值都在 7 以上，而长江流域的江西南昌滁槎监测点的 PH 值最低，为 6.31。适宜农作物生长的 PH 值范围为 6.0～7.5[2]，按这个标准来看，整个黄河流域没有一个达标，其 PH 值都在 7.5 以上，而其他流域的大多数监测点的水质也都已经偏碱性。

DO 是溶解氧，表述水体中的含氧量。根据标准限值表，溶解氧越大，水质越好，总体来看，珠江流域和淮河流域的溶解氧偏低，黄河流域和松海辽流域的溶解氧较好。当灌溉水中缺氧时，易使土壤处于还原状态而产生有机酸和硫化氢，抑制作物根系对氧气的吸收，造成作物减产或农产品质量下降。我国地表水环境质量标准规定农业用水区的溶解氧大于 2mg/l。按这个标准，各流域的水在溶解氧方面还是符合农业用水标准的。

CODMn 是高锰酸盐指数，表征有机物量的多少，氧化剂一般是高锰酸钾。其值越高，水体污染越严重。从图 2 中可以看出，整个长江流域和珠江流域的 CODMn 都是比较低的，在 5 以下。而松海辽流域和淮河流域则有多个监测点超过 III 类限值，黄河流域有一个监测点超过 IV 类限值。

NH3-N 称为氨氮，表述水体收受有机物污染的结果。一般污染物为含 N、P 较高的化肥、农药，还有生活垃圾。这样的情况被称为水体富营养化，易产生水华现象（水藻大量繁殖，水体含氧量降低，造成鱼虾等水生物死亡），太湖蓝藻事件就是由氨氮超标造成。从图上可以看出，除黄河和淮河的某些监测点氨氮数异常高外，其他水域的氨氮数一般都没有超过 IV 类限值。

3. 方差分析

我们感兴趣的是各流域间的 PH 值，DO 值，CODMn 值和 NH3-N 值是否有差异，即分析各流域间的同一污染指标值的均值是否相同。这里由于涉及到的变量不止一个，我们采用方差分析法，以流域作为影响因素，分析各污染指标是否有显著差异。下面给出各污染指标的箱线图，如图 3。其中，1~5 分别表示松海辽流域、淮河流域、黄河流域、长江流域、珠江流域。

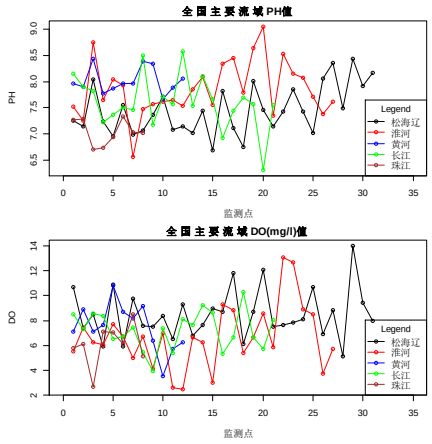

图 1. 主要流域 PH, DO 指标值

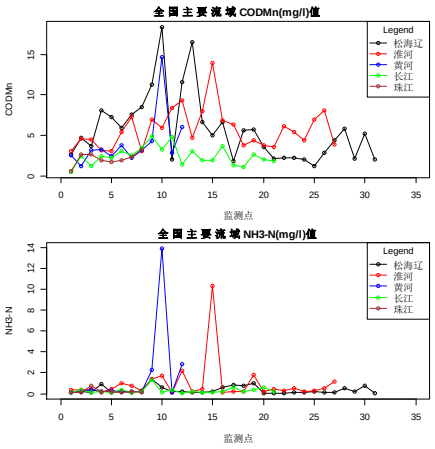

图 2. 主要流域 CODMn, NH3-N 指标值

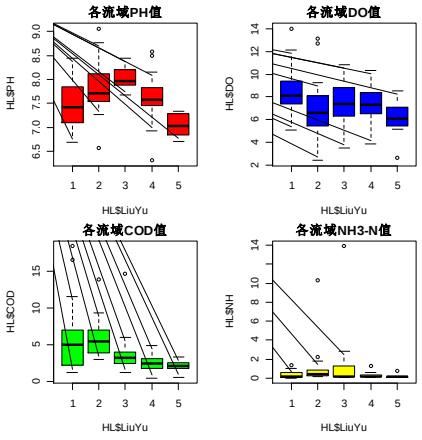

图 3. 各污染指标箱线图

首先，有必要检验一下我们的数据是否满足方差分析的前提条件[3]：独立正态性和方差齐性。

对我们的四个指标的数据进行正态性检验，使用shapiro.test()函数。总结如下：对 PH 值和 DO 值，各个流域均通过正态性检验；对 CODMn 值，仅长江流域和珠江流域的 CODMn 值通过正态性检验；对 NH3-N 值，各个流域均没有通过正态性检验。

下面进行方差齐性检验，使用 bartlett.test()函数。省略程序，结果显示：在 0.02 的显著性水平下，只有 PH 和 DO 两个指标通过方差齐性检验。CODMn 和 NH3-N 的显著性 P 值分别为 1.187e-08，< 2.2e-16，远远没有通过检验。综合上面的正态性检验结果，只有 PH 和 DO 两个指标适合做方差分析。

3.1 单因素方差分析

下面对 PH 和 DO 两个指标做单因素方差分析，这里取流域为影响因素，使用 aov()函数，程序及输出省略。结果显示：P 值均小于 0.05，即认为各流域间的 PH 值和 DO 值皆有显著差异。

3.2 Kruskal-Wallis 秩和检验

对于没有通过齐性检验的 CODMn 与 NH3-N 指标，这里采用 Kruskal-Wallis 秩和检验法。秩统计量的分布与总体的分布无关，可以摆脱对总体分布假设的束缚，这里采用秩和检验法是恰当的，使用 kruskal.test()函数，程序及输出省略。结果显示：P 值均小于 0.05，即各流域间 CODMn 与 NH3-N 值是存在显著差异的。

3.3 均值的多重比较

前面得出各流域间的 PH，DO，CODMn 与 NH3-N 皆有显著差异。但这样的结论太过笼统，我们希望得出 5 个流域间均值两两比较的结果。这里用多重 t 检验法，使用 pairwise.t.test()函数。省略程序结果，分析如下：

对 PH 值来说，在 0.05 的显著性水平下，松海辽流域(1)和长江流域(4)、淮河流域(2)和黄河流域(3)、淮河流域(2)和长江流域(4)没有显著差异。其余两两比较的结果被认为差异显著。

对 DO 值来说，认为松海辽流域(1)和淮河流域(2)、松海辽流域(1)和长江流域(4)、松海辽流域(1)和珠江流域(5)有显著差异，其余皆没有显著性差异。结合之前的线图和箱线图也可以看出，松海辽流域的 DO 值普遍较其他流域好。

对 CODMn 值来说，认为松海辽流域(1)和长江流域(4)，松海辽流域(1)和珠江流域(5)，淮河流域(2)和长江流域(4)，淮河流域(2)和珠江流域(5)有显著差异，其余皆没有显著性差异。

对 NH3-N 值来说，认为松海辽流域(1)和黄河流域(3)，黄河流域(3)和长江流域(4)有显著差异，其余皆没有显著性差异。

4. 判别分析

前面分析了各指标之间存在显著差异，下面我们将视角转移到这些指标到底在多大程度上能反应该监测点水质的情况。这里我们采用多分类距离判别的方法[3,4]将结果与实际的水质进行比较，用（1－误判率）来衡量这些指标的代表程度。为了使结果更一目了然，这里对原始数据按水质结构进行重新排列，得到该 99 个监测点中，I 类水质到劣 V 类水质的监测点个数分别为 7，36，25，17，6，8 个。程序及输出省略。结果显示误判个数为 36 个，误判率 36.36%，这样的误判率还是比较大的，即实际的水质模型算出来的水质结构与仅通过这 4 个指标判别出来的水质结果存在一些差异，但这 4 个指标依然能代表 63.64%的正确率。

分析上述结果，将第 2 类判成第 3 类水质的可能性比较大，这里不妨将 2，3 类水质合并为一类，对我们总体上把握各流域的水质结构是没有太大影响的。省略程序及结果，这里误判个数为 21 个，误判率 21.21%，正确率 78.79%，这样的结果是有所改善的。即我们仅采用这 4 个污染指标还是能够对总体的水质有很好的把握。

5. 水质结构分析

根据 2013 年 5 月 14 日发布的数据，可得到 2013 年各流域的水质分布饼图，图形省略。由饼图可知，总体来说，各流域污染程度排序为：黄河流域＞淮河流域＞松海辽流域＞珠江流域＞长江流域

为了对各水域的污染状况有一个纵向的了解，这里还收集了 2010 年，2007 年，2004 年，这三年第 19 周的数据。将各流域各年水质结构变化整理成表 1。

首先来看松海辽流域，I 类水质所占比例逐年减少，劣 V 类水质所占比例也逐年减少，II，III，IV 类水质所占比例趋于增加。说明松海辽流域在治理劣 V 类水质上还是有一些成效的，但同时也应重视 I 类水质资源的保护。

再看淮河流域，该流域对劣 V 类水质的治理可以说是卓有成效，但该流域近 10 年来所有的监测点都没有 I 类水质，说明整体上的水质还是欠佳的。

表 1. 各流域各年水质结构变化表

松海辽	2004	2007	2010	2013
Ⅰ	12.5%	9.5%	9.1%	6.5%
Ⅱ	31.3%	23.8%	22.7%	32.3%
Ⅲ	18.8%	9.5%	40.9%	25.8%
Ⅳ	0.0%	23.8%	13.6%	22.6%
Ⅴ	12.5%	14.3%	0.0%	6.5%
劣Ⅴ	25.0%	19.0%	13.6%	6.5%
淮河	2004	2007	2010	2013
Ⅰ	0.0%	0.0%	0.0%	0.0%
Ⅱ	12.5%	0.0%	18.5%	11.1%
Ⅲ	37.5%	62.5%	37.0%	33.3%
Ⅳ	25.0%	50.0%	22.2%	33.3%
Ⅴ	12.5%	0.0%	7.4%	11.1%
劣Ⅴ	62.5%	37.5%	14.8%	11.1%
黄河	2004	2007	2010	2013
Ⅰ	0.0%	11.1%	11.1%	0.0%
Ⅱ	25.0%	44.4%	44.4%	66.7%
Ⅲ	25.0%	22.2%	22.2%	8.3%
Ⅳ	25.0%	0.0%	0.0%	0.0%
Ⅴ	12.5%	0.0%	0.0%	0.0%
劣Ⅴ	12.5%	22.2%	22.2%	25.0%
长江	2004	2007	2010	2013
Ⅰ	5.9%	11.1%	26.3%	23.8%
Ⅱ	58.8%	72.2%	63.2%	47.6%
Ⅲ	11.8%	5.6%	5.3%	23.8%
Ⅳ	5.9%	11.1%	5.3%	4.8%
Ⅴ	17.6%	0.0%	0.0%	0.0%
劣Ⅴ	0.0%	0.0%	0.0%	0.0%
珠江	2004	2007	2010	2013
Ⅰ	14.3%	0.0%	0.0%	0.0%
Ⅱ	71.4%	62.5%	75.0%	62.5%
Ⅲ	0.0%	25.0%	12.5%	25.0%
Ⅳ	0.0%	0.0%	0.0%	0.0%
Ⅴ	0.0%	12.5%	12.5%	12.5%
劣Ⅴ	14.3%	0.0%	0.0%	0.0%

黄河流域，I 类水质比例减少，劣 V 类水质比例逐年增多，虽然第 II 类水质逐年增多，但总体上，该流域还是应该注意劣 V 类水质的治理。

长江流域，近 10 年都没有劣 V 类水质，且 I 类水质比例几乎逐年增加，比例向前三类水质靠拢，说明整个长江流域的水质很好。不存在较大的污染问题。

珠江流域，水质绝大部分集中在 II，III 类，整体水质还是不错的。近年来也不存在劣 V 类水质，但没有 I 类水质，终归有些不足。

6. 结论

本文分析 5 个流域：松海辽流域、淮河流域、黄河流域、长江流域、珠江流域的水质情况，考虑 4 个水质污染指标：PH、DO、CODMn、NH3-N。首先画出 4 个指标的线图对 5 个流域的污染情况有了直观的了解。然后画出 5 个流域间的 4 个指标值的箱线图，可以大致看出其差异。为了做单因素方差分析，先对数据进行正态性检验和方差齐性检验。发现只有 PH 和 DO 两个指标可以进行单因素方差分析，结果显示各流域间的 PH 值和 DO 值皆有显著差异。对于没有通过方差齐性检验的 CODMn 与 NH3-N 指标，采用 Kruskal-Wallis 秩和检验法，结果显示各流域间的 CODMn 与 NH3-N 值也是存在显著差异。然后采用多重 t 检验法对 5 个流域间均值做两两比较。为了检测 4 个指标反应监测点水质的情况，我们采用多分类距离判别法将分类结果与实际的水质分类结果进行比较，发现在水质分为 6 类的情况下 4 个指标能代表 63.64%的正确率，在水质分为 5 类的情况下 4 个指标能代表 78.79%的正确率。通过画出各流域的水质分布饼图，得出各流域污染程度排序为：黄河流域＞淮河流域＞松海辽流域＞珠江流域＞长江流域。通过收集 2004，2007，2010，2013 这四年的数据分析各流域各年水质结构的变化，并给出相关建议。

参考文献(References)

[1] J.R. Li. *Technical Report of the National Water Environment Information Database*. Beijing: Remote Sensing Technology Application Center of the Ministry of Water Resources, 2001.

[2] L.H. Dong, *Study on Growth Model of Water Quality in the River*. Tianjin University, 2010.

[3] Y. Xue and L.P. Chen, *Statistical Modeling and R Software*. Beijing: Tsinghua University Press, 2009.

[4] X.M. Wang, *Applied Multivariate Analysis*. Third Edition. Shanghai: Shanghai University of Finance and Economics Press, 2009.