

基于相关分析和回归模型的引江济太水量水质关系分析

朱跃龙¹, 朱佳丽¹, 李士进¹, 吴浩云², 陆铭锋², 陈 方²

(1 河海大学 计算机与信息学院, 江苏 南京 210098; 2 水利部太湖流域管理局, 上海 200434)

摘 要: 在分析引江过程中水质水量关系时, 先采用相关分析理论判断水质与水量是否具有相关性, 以及两者的相关程度, 再根据相关分析结果, 建立初始水质情况与达标时所需引水量之间的关系模型, 并对模型进行检验. 由于望亭立交闸下的水质会受到望虞河干流各站点水质的影响, 因此在考察水质水量的关系时, 初始水质情况由各站点的初始水质共同决定, 达标时所需的引水量由各站点水质均达标时流经望亭立交闸下的累计引水量决定. 在对水质水量建立回归模型时, 考虑到季节变化的因素以及异常值的影响, 根据相关分析结果分别建立不同情况下的回归模型, 并通过对不同模型进行比较来确定最终模型, 从而保证水质水量关系模型的正确性.

关键词: 引江济太; 水质水量分析; 相关分析; 回归模型; 望虞河流域

中图分类号: TV214

文献标识码: A

文章编号: 1000-7180(2012)03-0047-04

Analysis of Water Quality and Quantity in Yangtze River-Taihu Lake Water Transfer Project Based on Correlation Analysis and Regression Models

ZHU Yue-long¹, ZHU Jia-li¹, LI Shi-jin¹, WU Hao-yun²,
LU Ming-feng², CHEN Fang²

(1 School of Computer & Information Engineering, Hohai University, Nanjing 210098, China;

2 Taihu Basin Administration of the Ministry of Water Resources, Shanghai 200434, China)

Abstract: In analysis of how the water quality changes with the change of water quantity, firstly we adopt correlation analysis to determine the relevance between water quality and water quantity, and then build the relational model of the initial water quality and the ultimately required amount of water, finally we test the model. In the process of engineering experiment, the highlight is that how much water is required in Changshu water control to make the water under the Wangting interchange gates reach the standard. For the water quality under Wangting interchange gates is influenced by the water along the Wangyu River, we suppose that the initial water quality is measured by the initial water in all sites, and that when the water in each site along the Wangyu River reaches the standard, the total amount of water flow through the Wangting interchange gates is the required amount, so that we can guarantee the data used to modeling is reliable. In the process of modeling, considering the influence of seasonal changes and outliers, we respectively build different models based on correlation analysis, and then by comparison, we determine the final model, which has guaranteed the accuracy of our approach.

Key words: Yangtze River-Taihu Lake Water Transfer Project; analysis of water quality and quantity; correlation analysis; regression model; Wangyu River

1 引言

近年来,随着太湖流域经济的高速发展以及乡镇企业的繁荣,大量废污水未经处理直接排入江河湖泊,使河流湖泊水质受到严重的污染.因此,文中提出了“引江济太”工程体系.通过调引长江水入太湖及流域河网来提高水体自净能力,以缓解流域水恶化的趋势,从而发挥了防洪除涝的巨大作用,并且给太湖流域带来了显著的社会、经济和环境效益^[1].

在“引江济太”工程中,长江水从常熟水利枢纽引入,流经望虞河流域,最后经望亭水利枢纽到达太湖.在此过程中,只有保证望亭水利枢纽的水质达标,引入的水才能达到净化太湖的效果^[2].国内外专家学者提出了很多水质预测的方法.颜剑波^[3]等采用多元回归分析的方法用于黄河水质预测.李亦芳^[4]等人提出结合人工神经网络和回归分析的方法进行水质预测,通过回归分析模型得到的预测区间来控制人工神经网络中的异常值,从而提高预测精度.徐业平^[5]提出根据空间信息统计学的基本原理,建立河流水质模拟、预测的方法,进行量质联合评价.

根据项目研究的主要任务,选择高锰酸盐指数(CODMn)和氨氮(NH₃-N)为目标水质指标.通过分析上游各站点这些指标随引水量的变化过程,以及上游站点浓度的变化对下游站点的影响,建立望虞河沿线水质与水量之间的关系,从而预测望亭立交闸下水质达标所需要的引水量.研究过程中,通过分析不同情况下相关系数的变化,来判断异常值的存在和季节的影响,并通过检验回归模型来进一步保证水质水量关系模型的精确度.

2 相关性分析

2.1 相关分析理论

相关分析(correlation analysis)是研究两个或两个以上变量之间相关程度大小以及用一定函数来表达其相互关系的方法.一般可以借助相关系数、相关表与相关图来进行相关分析^[6].

简单相关系数反映两个变量之间线性相关密切程度和相关方向的统计测定,它是其他相关系数形成的基础.两个随机变量 X 和 Y 的简单相关系数的计算公式为:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}} \quad (1)$$

或化简为:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}} \quad (2)$$

相关系数具有以下性质:

(1) 相关系数的取值范围在 -1 和 +1 之间,即: $-1 \leq r \leq 1$;

(2) 若计算结果 r 为正,则表明两变量为正相关;若 r 为负,则表明两变量为负相关;

(3) 相关系数 r 的数值越接近于 1(-1 或 +1),表示相关性越强;越接近于 0,表示相关性越弱.如果 $r=1$ 或 -1 ,则表示两个变量完全直线性相关.如果 $r=0$,则表示两个变量完全不相关(不是线性相关);

2.2 水质水量相关性分析

由于上游的引水要经过一定的延时才能对下游站点产生影响,因此在统计水质达标所需的引水量时,需要考虑到站点之间的延时.这里仅以 CODMn 有关的水质水量相关性分析过程进行说明.在分析水质水量关系的过程中,由于望虞河沿线各站点水质的初始情况反映了各站点的初始污染状况,污染越严重就需要引入更多的水来缓解,因此需要结合各站点的初始水质来分析达标时水质与所需引水量之间的关系.其中,初始水质用引水前望虞河各站点 CODMn 的初始浓度来表示,达标时所需的引水量用望虞河各站点以及望亭立交闸下水质均达标时流经望亭水利枢纽的累计引水量来表示.

本文借助 SAS 软件^[7]计算水质水量的相关性,分别计算了各站点 CODMn 初始浓度之和与达标时所需引水量之间的相关系数、各站点 CODMn 最大初始浓度与达标时所需引水量之间的相关系数.在计算相关系数过程中,先把历年数据作为一个整体计算出总的相关系数;考虑季节因素对水质变化的影响,又把历年数据分成春夏季和秋冬季两部分,分别计算相关系数;同时,在分析过程中发现 2003 年的数据可能有异常,于是又在去除 2003 年数据的情况下,对相关系数进行了重新计算,并对不同情况下的相关系数进行比较.具体见表 1.

从表 1 可以看出,在去除 2003 年数据的情况下,初始浓度之和与引水量之间的相关系数、最大初始浓度与引水量之间的相关系数都增大了很多,这说明 2003 年的数据很有可能是异常的;当数据根据

季节进行分类之后,相关系数也均增大了很多,说明把水质水量数据按照季节分类也是合理的.

表 1 CODMn 对应的相关系数

变量	总相关性	春夏季相关性
初始浓度之和— 引水量	0.016	0.69
初始浓度之和— 引水量(03 年除外)	0.41	0.81
最大初始浓度— 引水量	0.054	0.19
最大初始浓度— 引水量(03 年除外)	0.16	0.58

3 基于回归分析确定水质水量关系

3.1 多元回归模型

回归分析(regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法.回归分析是通过规定因变量和自变量,建立回归模型,并根据实测数据来求解模型的各个参数,然后评价回归模型是否能够很好地拟合实测数据;如果拟合误差较小,则可以根据自变量作进一步预测.

多元线性回归数学模型:假定因变量 y 与自变量 $x_1, x_2, \dots, x_m$ 间存在线性关系,多元线性回归数学模型为:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_m x_{mj} + \epsilon_j$$
$$j = 1, 2, \dots, n \tag{3}$$

式中, $x_1, x_2, \dots, x_m$ 为可以观测的一般变量(或为可以观测的随机变量); y 为可以观测的随机变量,随 $x_1, x_2, \dots, x_m$ 而变,受试验误差影响; ϵ_j 为相互独立且都服从 $N(0, \sigma^2)$ 的随机变量;可以根据实际观测值对 $\beta_0, \beta_1, \dots, \beta_m$ 以及方差 σ^2 作出估计.

设 y 对 $x_1, x_2, \dots, x_m$ 的 m 元线性回归方程为:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m \tag{4}$$

其中, $b_0, b_1, \dots, b_m$ 为 $\beta_0, \beta_1, \dots, \beta_m$ 的最小二乘估计值.即 $b_0, b_1, \dots, b_m$ 应使实际观测值 y 与回归估计值 $\hat{y}$ 的偏差平方和最小.

3.2 水质水量关系回归模型的建立

由水质水量的相关分析可知,达标时所需的引水量与初始浓度之和、最大初始浓度均有一定的正相关关系,因此采用二元回归分析来获得累计引水

量、最大初始浓度、初始浓度之和这三者之间的关系 $y = a * x_1 + b * x_2 + c$,其中因变量 y 为达标时所需的累计引水量,自变量 x_1 = 最大初始浓度,自变量 x_2 = (初始浓度之和 - 最大初始浓度).同时,由于春夏季和秋冬季水质达标时所需要的累计引水量不同,且 2003 年的数据可能是异常的,这些都会影响到回归模型的精确度,因此这里将根据不同情况分别采用 SAS 工具进行多元回归分析,建立水质达标时的总引水量与最大初始浓度、初始浓度之和的回归模型.由于春夏季节的数据包含了 2003 年的数据,因此需要分两种情况讨论,而秋冬季节的数据不包括 2003 年的数据,则不需要分情况说明.

(1) 在去除 2003 年数据的情况下,春夏季的回归模型为:

$$Y = 975X_1 + 390X_2 - 9493 \tag{5}$$

(2) 在包含 2003 年数据的情况下,春夏季的回归模型为:

$$Y = 1399X_1 + 1217X_2 - 26646 \tag{6}$$

(3) 秋冬季节的回归模型为:

$$Y = 265X_1 + 233X_2 \tag{7}$$

3.3 模型确定与检验

在模型检验的过程中,分别利用式(5)~(7)求出相应情况下所需引水量的计算值,并把计算值和统计值进行比较,求出两者之间的绝对误差,从而根据误差值来判断模型的优劣,其中绝对误差值越小,说明模型越好,反之则越差.具体见表 2.

表 2 去除 2003 年情况下春夏季所需引水量的
计算值和统计值

年份	统计值(单位: 10000 m ³)	计算值(单位: 10000 m ³)	绝对误差(单位: 10000 m ³)
2007-5-7	4991	5366	375
2004-7-23	5690	5327	363
2005-6-17	3952	3884	68
2006-8-10	3610	3689	79
2003-7-23	10927	6283	4644

从表 2 可以看出,在去除 2003 年数据的情况下,根据春夏季回归模型求出的计算值与统计值之间的误差值均较小,而把 2003 年数据代入模型所求得的绝对误差值很大,具体见表 3.

从表 3 可以看出,在包含 2003 年数据的情况下,根据春夏季回归模型求出的计算值与统计值之间的绝对误差值都很大.

表3 包含2003年情况下春夏季所需引水量的
计算值和统计值

年份	统计值(单位: 10000 m ³)	计算值(单位: 10000 m ³)	绝对误差(单位: 10000 m ³)
2007-5-7	4991	6574	1583
2004-7-23	5690	6123	433
2005-6-17	3952	4578	626
2006-8-10	3610	2327	1283
2003-7-23	10927	9598	1329

根据以上对比分析可知,2003年的数据确实存在异常,在进行回归分析时,去除2003年数据是正确的处理;同时,根据引水季节的不同,把水质水量数据分成春夏季和秋冬季分别建立回归模型也是合理的.因此最终春夏季水质水量的关系为:

$$Y = 975X_1 + 390X_2 - 9493$$

秋冬季水质水量的关系为:

$$Y = 265X_1 + 233X_2$$

4 结束语

本文通过相关分析确定初始水质情况与达标时所需引水量之间的相关性,根据相关系数的变化情况,对异常数据进行了初步判断,并分析了季节变化对水质变化过程的影响,再根据相关分析的有关结论,分别建立不同情况下的水质水量回归模型,并通过回归模型的检验来确定最终的水质水量关系.本文方法取得了较好的结果.

在引水过程中,水质亦可能会受到城市工业污

染、大气降水等情况的影响,使水质的变化过程更为复杂.同时,由于数据的缺失,在某些年份水质随引水量的变化过程不能完全呈现.以上两点都可能会影响到水质水量模型的精度.如果能解决以上问题,则本文水质水量模型的精度将会得到进一步提高.

参考文献:

- [1] 吴浩云.引江济太调水实验关键技术研究 and 应用[J]. 中国水资源,2008(1):6-8.
- [2] 刘宁.对引江济太调水试验工程的初步认识和探讨[J]. 中国水利,2004(4):36-38.
- [3] 严剑波,阮晓红,孙瀚.多元回归分析在黄河水质预测中的应用[J]. 人民黄河,2010,32(3):35-36.
- [4] 李亦芳,程万里,刘建厅.基于人工神经网络和回归分析的水质预测[J]. 郑州大学学报,2008,29(1):106-109.
- [5] 徐业平.河流水质水量综合评价方法研究综述[J]. 水资源保护,2005,21(4):34-36.
- [6] 范金城,梅长林.数据分析[M]. 5版. 北京:科学出版社,2006.
- [7] 汪嘉冈. SAS V8 基础教程[M]. 北京:中国统计出版社,2003.

作者简介:



朱跃龙 男,(1959—),博士,教授.研究方向为智能信息处理.

(上接第46页)

- [4] 李曼.基于语义的Web服务发现研究[D]. 长沙:中南大学信息科学与工程学院,2009.
- [5] 张忠平,田淑霞,刘洪强.一种综合的本体相似度计算方法[J]. 计算机科学,2008(12):142-145.
- [6] Budanitsky A, Hirst G. Evaluating word net-based measures of lexical semantic relatedness[C]// Association for Computational Linguistics. Sweden, 2005.
- [7] W3C. Simple object access protocol [EB/OL]. [2010-07-19]. <http://www.w3.org/TR/soap>.

- [8] Lovasz L, Plummer M. Matching theory [M]. Amsterdam: North-Holland, 1986.

作者简介:



甄景波 男,(1981—),硕士研究生.研究方向为语义网络服务发现.