

DOI:10.3876/j.issn.1000-1980.2017.03.005

基于影响因子数据年际分类的 太湖典型口门流量估算方法

孙 前^{1,2}, 陈 方¹, 刘金涛², 吴国群²

(1. 太湖流域水文水资源监测中心, 江苏 无锡 214024; 2. 河海大学水文水资源学院, 江苏 南京 210098)

摘要: 基于太湖典型口门瓜泾口站 1966—2014 年(1989—2005 年缺)的流量、水位、降水及周围站点水位等数据,采用多元线性回归分析法构建太湖流量估算模型。采用聚类分析方法,对流量的影响因子年际序列矩阵进行相似分类,找出与目标年(待估算流量的年份)相似的年份,然后依据相似年份数据率定回归模型参数。通过与基于常系列数据建立的回归模型进行比较发现,基于相似年份数据建立的回归模型的估算精度更高。

关键词: 太湖瓜泾口站;聚类分析;流量估算;多元线性回归

中图分类号: P333.1 **文献标志码:** A **文章编号:** 1000-1980(2017)03-0218-06

Method for estimation of discharge at typical entrance of Taihu Lake based on annual classification of impact factor data

SUN Qian^{1,2}, CHEN Fang¹, LIU Jintao², WU Guoqun²

(1. Monitoring Center of Hydrology and Water Resources of Taihu Basin Authority, Wuxi 214024, China;

2. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: Based on discharge, water level, and precipitation data at the Guajingkou Station and water level data at surrounding stations from 1966 to 2014 (with data missing from 1989 to 2005), a model for estimation of discharge in Taihu Lake was established with multiple linear regression. The annual sequence matrixes of the impact factors of the discharge were classified using clustering analysis according to the similarity, the years similar to the objective years were determined, and, finally, the regression model parameters were calibrated based on the data from similar years. Compared with the regression model based on the common series data, the regression model based on the data from similar years has a higher accuracy.

Key words: Guajingkou Station in Taihu Lake; clustering analysis; discharge estimation; multiple linear regression

太湖流域重要河湖口门取水总量的估算,对开展太湖取水总量控制和计划用水工作,细化流域水量分配,以及实现依法治水、依法管水等具有重要意义。估算水量的传统一潮推流法,使用广泛,能够达到精度要求,但不适合所有情形的口门,特别是对于闸门变动较大的水闸,每年需重新校测,测验任务繁重,目前还没有较好的解决方法^[1]。国内学者在水量估算方面做了众多研究,如万晓凌等^[2]改进的水位差关系法,较为简易可行,但对测次及使用条件有一定要求。左一鸣等^[3]、朱漫莉等^[4]采用动力学方法进行流域水量平衡过程模拟,此类研究能有效提高计算速度,缩短计算时间,但建模过程较为复杂,工作量大,且计算时间较长^[3]。此外,遥感与同位素技术也被应用于太湖河网区的水量估算。如陆宝宏等^[5]利用水量平衡与同位素质量守恒原理构建了河网地区水量分配模型,可以用来近似估算河网区不同地点的受水量;张国庆等^[6]结合湖面高程与面积变化对区域气候变化的响应,利用遥感卫星数据快速有效估算湖泊水量平衡。然而,水

收稿日期: 2016-05-17

基金项目: 国家“十三五”重点研发计划(2016YFC0401501)

作者简介: 孙前(1991—),男,河南郑州人,硕士研究生,主要从事流域水文模拟研究。E-mail: qhsqsc2006@163.com

样同位素测试和遥感信息提取对人员和设备的要求较高,限制此类方法的推广。

传统的基于数理统计的水文学方法,根据收集的历年数据,利用多元线性回归推求估算公式,特点是方法简单,精度较高^[7],可方便地进行流量估算。且太湖局采用由基点站和巡测断面所组成的水文驻测与巡测相结合的水文观测方式,可为此方法提供大量数据^[8]。但是,由于环湖口门水文条件可能随时间发生改变^[9],数据系列会存在不一致现象。因此,在建立回归模型的过程中,需要对数据系列进行分类,以消除这种影响。为此,笔者通过引入聚类分析方法,对瓜泾口站流量影响因子的时间序列数据^[10-11]进行分类,进而建立流量的估算回归模型,以期提高估算精度。

1 资料来源与研究方法

1.1 资料来源

根据太湖流域水系特点及行政分区,将其分为8个四级水资源区(图1),其中与环太湖出入湖水量最为密切的有5个区,分别为湖西区、浙西区、阳澄淀泖区、武澄锡虞区和杭嘉湖区^[12]。

依据太湖流域水文监测站点的分布情况以及计算方法对资料系列的要求,以位于太湖东侧、资料最全面的瓜泾口站为研究对象,建立模型估算1966—1988年、2006—2014年(共32 a)中任一目标年的流量。建模时以同期的瓜泾口站水位、降水^[13]、蒸发及瓜泾口下游与其具有较密切水力联系的水位站陈墓站、商榻站^[14]的水位作为可能的影响因子,同时各数据均采用月平均值。

1.2 研究方法

首先,收集瓜泾口站及邻近站点的相关数据,基于SPSS软件,采用单相关系数法(Pearson法)遴选与流量相关的影响因子;然后,确定目标年,并分别采用常规方法与分类方法构建流量估算公式。常规方法采用目标年之前所有年份的数据,利用多元线性回归分析方法建立月平均流量与所遴选的影响因子之间的关系模型,并通过最小二乘法确定该模型各项的系数。分类方法通过引入聚类分析,对瓜泾口站影响因子的多年数据进行分类,从而选出与目标年数据相似的年份,依据相似年份的数据进行多元线性回归分析,构建流量估算模型。

如果以年为单位,以目标预测年的流量为对象,则某些年份的输入矩阵可能会与其他年份的相似。分类方法在建立回归模型前,对数据矩阵进行分类,得出与目标年数据相似的年份,然后用该年份的影响因子矩阵率定模型参数,估算精度会显著提高。

根据遴选的影响因子(自变量 $x_1, x_2, \dots, x_m$)对流量(因变量 y)建立的多元线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon \quad (1)$$

式中: ε ——相互独立且服从 $N(0, \sigma^2)$ 的随机变量; m ——自变量个数; $\beta_0 \sim \beta_m$ ——系数。假设共有 n 组实测数据,可以根据实测值采用最小二乘法对参数 $\beta_0 \sim \beta_m$ 做出估计。

1.3 结果评价

选用年总流量累计误差 δ (式(2))及确定性系数 R^2 (式(3))对预测结果进行评价。

$$\delta = \frac{\sum \Delta}{\sum Q_{\text{实}}} \times 100\% \quad (2)$$

$$R^2 = \frac{\sum (Q_{\text{估}} - Q_{\text{实}})^2}{\sum (Q_{\text{实}} - \bar{Q})^2} \quad (3)$$

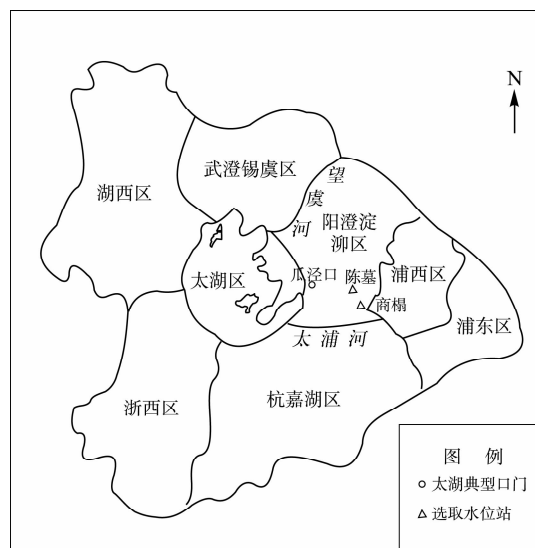


图1 太湖流域水资源分区与瓜泾口站及周围站点分布
Fig.1 Water resources division and locations of Guajingkou Station and surrounding stations in Taihu Lake Basin

式中: $\Sigma \Delta$ ——年流量估算偏差的年累计值; $\Sigma Q_{\text{实}}$ ——实测年总流量; $Q_{\text{实}}$ ——实测流量; $Q_{\text{估}}$ ——估算流量; $\bar{Q}$ ——平均实测流量。

2 应用实例

2.1 水文影响因子遴选

由于瓜泾口站、陈墓站、商榻站数据较全,因此为了保证数据系列的一致性,选取瓜泾口站水位 Z_g 、瓜泾口站降水 P_g 、陈墓站水位 Z_{cm} 、商榻站水位 Z_{st} 、瓜泾口站与陈墓站水位差 $(Z_g - Z_{cm})$ 、瓜泾口站与商榻站水位差 $(Z_g - Z_{st})$ 6 个因子进行影响因子分析。所选站点数据均采用月平均数据,水位数据均为黄海基面。

采用 Pearson 法对多年以及单个年份的影响因子数据进行相关性分析,结果见表 1。由表 1 可知, Z_g 、 P_g 、 $Z_g - Z_{cm}$ 、 $Z_g - Z_{st}$ 与流量 Q_g 的相关系数较高, t 检验的显著性概率均小于 0.05,表明瓜泾口站 Q_g 与这些影响因子始终呈显著相关。而 Z_{cm} 、 Z_{st} 与 Q_g 的相关性较差,未通过显著性检验。由表 1 可知,显著影响因子没有因年份的不同发生改变,故最终选定的影响因子为: Z_g 、 P_g 、 $Z_g - Z_{cm}$ 、 $Z_g - Z_{st}$ 。

表 1 各影响因子与流量 Q_g 的相关分析结果
Table 1 Results of analysis of correlation between each impact factor and discharge Q_g

水文影响因子	Pearson 相关系数				显著性概率			
	多年	1996 年	...	2014 年	多年	1996 年	...	2014 年
Z_g	0.64	0.51	...	0.69	0.00	0.00	...	0.00
P_g	0.29	0.32	...	0.37	0.04	0.02	...	0.03
Z_{cm}	0.19	0.15	...	0.09	0.47	0.63	...	0.51
Z_{st}	0.11	0.15	...	0.14	0.52	0.63	...	0.67
$Z_g - Z_{cm}$	0.57	0.91	...	0.93	0.00	0.00	...	0.00
$Z_g - Z_{st}$	0.43	0.82	...	0.81	0.01	0.00	...	0.00

2.2 聚类分析

因各影响因素的单位不同,因此需要对指标数据进行标准化,本文采用 Z-score 标准化方法,也称为标准差标准化。

由聚类树状图(图 2)可知,瓜泾口站 1966—1988 年数据与 2006—2014 年数据被分成 2 类。这是由于 2002 年后,“引江济太”工程以及瓜泾口的疏浚工程的实施,使得瓜泾口站的数据情况发生了质的变化^[13],所以 1966—1988 年的数据无法用于 2002 年后流量情况的估算。因此,分别在 2 类数据中进行分析,并各选取一个目标年进行流量的推求。为使目标年所具有的历史数据足够长,且保障在数据系列中目标年具有一个最为相似的年份,故以年份相对靠后以及在图 2 中位置相对靠上的 1988 年和 2011 年为例进行聚类分析。根据图 2 及聚类分析中的相似性水平 λ 值,同一组中相对位置靠上的 2 样本,相似度更高,因此与 1988 年数据最接近的年份为 1985 年,而与 2011 年数据最接近的年份为 2008 年。

2.3 流量计算

2.3.1 1988 年流量计算

由图 2 可知,与 1988 年最接近的是 1985 年的数据。所以用 1985 年数据做多元线性回归,得出流量计算公式:

$$Q_g = -66.388 + 157.801Z_g + 0.008P_g - 129.398(Z_g - Z_{cm}) - 18.981(Z_g - Z_{st}) \tag{4}$$

同时采用瓜泾口站 1966—1987 年历史数据推求 1988 年流量,得到常规方法的流量估算公式:

$$Q_g = -61.322 + 25.202Z_g - 0.023P_g + 2.038(Z_g - Z_{cm}) - 1.017(Z_g - Z_{st}) \tag{5}$$

图 3 给出了常规方法与分类方法推求的 1988 年流量过程线,从图 3 可以看出,分类方法推求的结果较好,不仅趋势一致,且系统误差也大大降低,仅个别月份出现较大误差,结果较为理想。

2.3.2 2011 年流量计算

由图 2 可知,与 2011 年最接近的是 2008 年的数据,所以用 2008 年数据做多元线性回归,得出的 2011 年瓜泾口站流量计算公式为

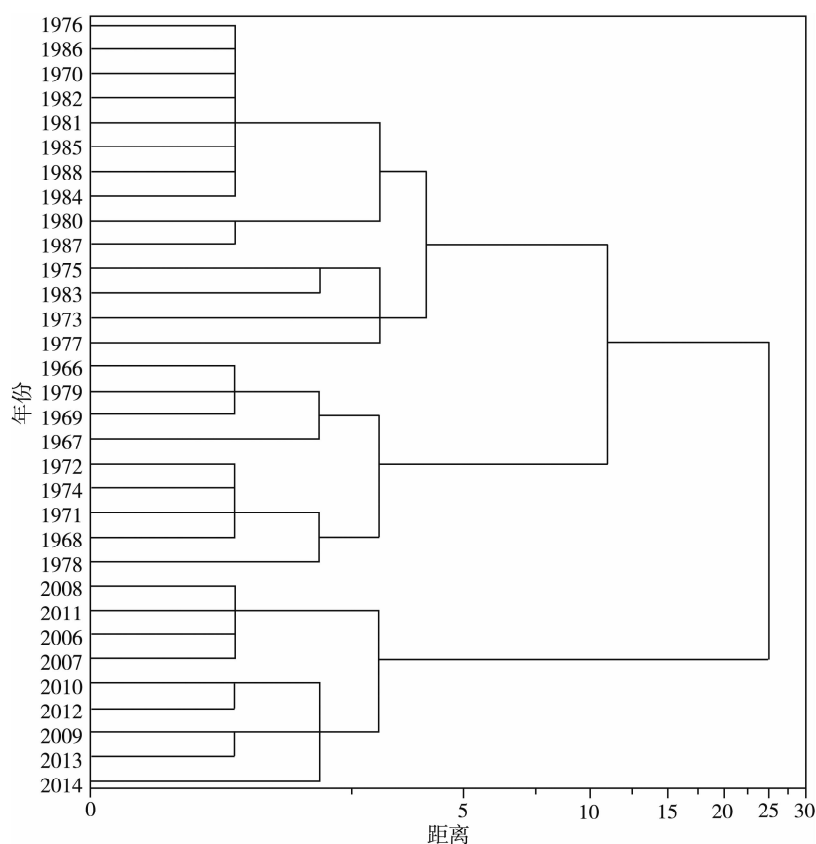


图2 瓜泾口站1966—1988年、2006—2014年数据聚类结果

Fig. 2 Data hierarchical diagram during periods from 1966 to 1988 and from 2006 to 2014 at Guajingkou Station

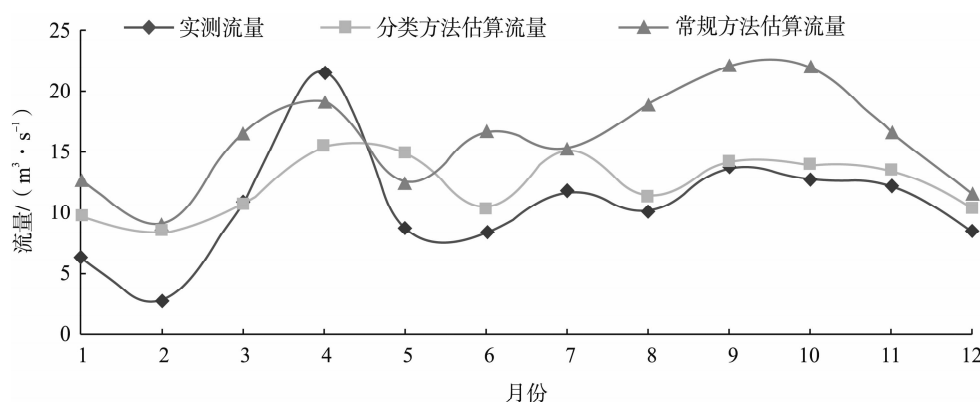


图3 常规方法与分类方法推求的1988年流量过程线

Fig. 3 Discharge hydrographs in 1988 obtained from conventional method and classification method

$$Q_g = -168.108 + 229.435Z_g - 0.009P_g + 1.646(Z_g - Z_{cm}) + 199.527(Z_g - Z_{st}) \quad (6)$$

采用瓜泾口站2006—2010年历史数据推求2011年流量,得出常规方法的流量估算公式:

$$Q_g = -186.485 + 200.760Z_g - 0.024P_g - 6.376(Z_g - Z_{cm}) - 149.617(Z_g - Z_{st}) \quad (7)$$

图4给出了常规方法与分类方法推求的2011年流量过程线。由图4可知,使用与2011年最为接近的2008年推求出的结果与常规方法推求的结果均较好,趋势基本一致。

2.3.3 误差分析

由图3~4、表2可得,1988年确定性系数由常规方法的0.58上升到分类方法的0.77,年流量相对误差相应地由50.5%下降到15.7%,各月的总流量相对误差也有大幅下降,因此,分类方法相较于常规方法,无

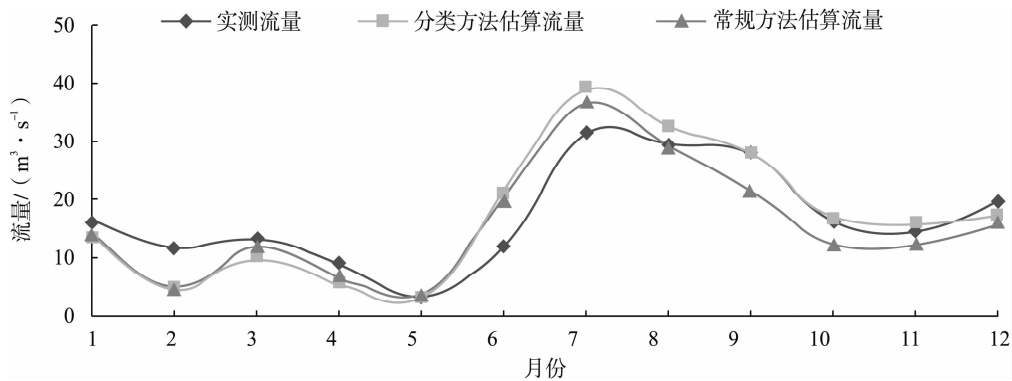


图4 常规与分类方法推求的2011年流量过程线
Fig.4 Discharge hydrographs in 2011 obtained from conventional method and classification method

论是流量过程还是流量的预测精度均有较大幅度的提升。分类方法需要长系列年的数据予以支撑,数据年份较多的时候,更容易找到相似程度较高的年份,推求出的效果也会更佳。例如2011年,常规方法与聚类分析的方法推求出的结果差别不大,这正是由于缺少与2011年数据相似程度较高的年份来率定模型参数,在这种情况下,可以分别计算和比较2种方法的确定性系数及相对误差,选择预测精度相对较高的方法。

需要说明的是,2种方法对不同目标年的预测效果存在不同程度的差异,主要是由于瓜泾口站每年的水文条件不尽相同,收集的影响因子数据可能不够全面^[15],或者存在较大或较小的“特殊值”,文中所建立的多元线性回归模型仍属于统计学范畴,容易受到特殊值的扰动,影响预测效果,同时对于波峰、波谷数据的预测误差也相对较大^[16]。

表2 常规方法与分类方法估算的流量相对误差
Table 2 Relative errors of discharge estimated with conventional method and classification method

月份	1988 年					2011 年				
	总流量/(m ³ ·s ⁻¹)			相对误差/%		总流量/(m ³ ·s ⁻¹)			相对误差/%	
	实测	常规法	分类法	常规法	分类法	实测	常规法	分类法	常规法	分类法
1	6.3	12.6	9.7	100.9	55.0	16.4	14.1	13.4	-13.9	-18.3
2	2.8	9.1	8.3	221.5	195.6	11.7	4.6	5.0	-60.4	-57.0
3	11.0	16.5	10.8	49.7	-2.1	13.4	12.0	10.0	-10.5	-25.6
4	21.5	19.1	15.4	-11.1	-28.3	9.2	7.1	5.5	-23.0	-39.7
5	8.7	12.4	14.2	42.8	63.2	3.3	3.5	3.1	6.8	-5.6
6	8.4	16.6	10.3	99.0	23.9	12.0	19.9	21.0	66.0	75.4
7	11.8	15.2	15.1	29.0	28.0	31.6	36.6	39.3	15.9	24.2
8	10.1	18.8	11.3	86.4	11.9	29.5	29.0	32.4	-1.7	9.9
9	13.7	22.0	14.2	60.4	3.5	28.0	21.5	27.9	-23.2	-0.4
10	12.7	21.9	14.0	72.5	9.9	16.2	12.3	16.9	-24.2	4.6
11	12.2	16.5	13.4	35.5	10.2	14.7	12.5	16.1	-15.1	9.5
12	8.5	11.5	10.4	35.9	22.7	19.6	15.9	17.3	-18.8	-11.8

3 结 语

建立了基于多元回归方程的太湖典型口门流量估算模型,建模时尝试引入水文时间序列相似分析方法,即以年份为单位,通过聚类分析方法对已收集的历史数据序列进行分类,据此建立回归模型并进行目标年流量估算。对瓜泾口站1966—2014年(缺1989—2005年)资料序列的聚类分析,分别选用常系列数据和最相似年份(1985年和2008年)数据建立回归模型。预测结果显示,基于聚类分析的线性回归方法明显优于常规方法。这表明,在具备足够长的历史资料的前提下,这种改进方法可以有效提高太湖主要口门流量估算的精度。

值得注意的是,分类方法需要一定序列长度的历史数据,通常为20 a以上,所以对目标站点的选取是有

一定要求的。此外,系列中数据相似程度较高的年份,采用该方法的改进效果更为显著(例如对1988年的预测效果提升要好于2011年)。

参考文献:

- [1] 仲兆林. 常州沿江感潮河道水文站水位流量关系综合定线分析[J]. 广东水利水电, 2009(11):33-34. (ZHONG Zhao lin. Comprehensive alignment analysison the stage-discharge relation at hydrological station intided channel along the river in Changzhou[J]. Guangdong Water Resources and Hydropower,2009(11):33-34. (in Chinese))
- [2] 万晓凌, 陆小明, 周毅, 等. 感潮水闸引水量计算方法研究[J]. 长江科学院院报, 2013, 30(4):17-20. (WAN Xiaoling, LU Xiaoming, ZHOU Yi, et al. Method of calculating water diversion amount of tidal sluice[J]. Journal of YangtzeRiver Scientific Research Institute,2013, 30(4):17-20. (in Chinese))
- [3] 左一鸣, 崔广柏. 二维水动力模型的并行计算研究[J]. 水科学进展, 2008, 19(6):846-850. (ZUO Yiming, CUI Guangbai. Parallel computation for 2D flow model [J]. Advances in Water Science,2008, 19(6):846-850. (in Chinese))
- [4] 朱漫莉, 高海鹰, 徐力刚, 等. 基于系统动力学方法的鄱阳湖流域水量平衡过程模拟与分析[J]. 水资源保护, 2015, 31(3):46-52. (ZHU Manli, GAO Haiying, XU Ligang, et al. Simulation and analysis of water balance process in Poyang Lake Basin based on system dynamic approach [J]. Water Resources Protection, 2015,31(3):46-52. (in Chinese))
- [5] 陆宝宏, 孙营营, 马乐军, 等. 利用同位素质量守恒原理估算太湖河网受水量[J]. 河海大学学报(自然科学版), 2009, 37(6):645-649. (LU Baohong, SUN Yingying, MA Lejun, et al. Estimation of water amount in river network of Taihu Lake based on isotopic mass conservation [J]. Journal ofHohaiUniversity(Natural Sciences),2009, 37(6):645-649. (in Chinese))
- [6] 张国庆, XIE Hongjie, 姚檀栋, 等. 基于ICESat 和 Landsat 的中国十大湖泊水量平衡估算[J]. 科学通报, 2013(26):2664-2678. (ZHANG Guoqing, XIE HongJie, YAO Tandong, et al. Water balance estimates of ten greatest lakes in China using ICESat and Landsat data[J]. Chin Sci Bull, 2013(26):2664-2678. (in Chinese))
- [7] 向速林, 刘占孟, 尤本胜. 地下水流量预测的多元线性回归分析模型研究[J]. 水文, 2006, 26(6):36-37. (XIANG Sulin, LIU Zhanmeng, YOU Bensheng. Study on multivariate linear regression analyzing model for groundwater discharge forecasting [J]. Journal of ChinaHydrology,2006, 26(6):36-37. (in Chinese))
- [8] 刘文龙, 寿祥, 张文龙. 太湖进出水量平衡计算与分析[J]. 中国西部科技, 2010, 9(14):34-35. (LIU Wenlong, SHOU Xiang, ZHANG Wenlong. Estamites and analysisofTaihu Lake Water Balance. [J] Science and Technology of West China, 2010,9(14):34-35. (in Chinese))
- [9] 沈国华, 王谦. 太湖进出水量变化对水环境的影响[J]. 水资源保护, 2013, 29(1):46-50. (SHEN Guohua, WANG Qian. Influences of input and output flow on water environment in Taihu Lake [J]. Water Resources Protection,2013, 29(1):46-50. (in Chinese))
- [10] 汤成友, 郭丽娟, 王瑞. 水文时间序列逐步回归随机组合预测模型及其应用[J]. 水利水电技术, 2007, 6(6):1-4. (TANG Chengyou, GUO Lijuan, WANG Rui. Application of prediction model for stochastic combination of stepwise regression of hydrologic time series[J]. Water Resources and Hydropower Engineering,2007, 6(6):1-4. (in Chinese))
- [11] 欧阳如琳, 任立良, 周成虎. 水文时间序列的相似性搜索研究[J]. 河海大学学报(自然科学版), 2010, 3(3):241-245. (OUYANG Rulin, REN Liliang, ZHOU Chenghu. Similarity search in hydrological time series[J]. Journal of Hohai University (Natural Sciences), 2010, 3(3):241-245. (in Chinese))
- [12] 申金玉, 甘升伟, 陈润, 等. 环太湖出入湖水量影响因素分析及对策措施研究[J]. 水资源保护, 2011, 27(6):48-52. (SHEN Jinyu, GAN Shengwei, CHEN Run, et al. Analysis of influencing factors of inflow and outflow around Taihu Lake andcountermeasures[J]. Water Resources Protection, 2011, 27(6):48-52. (in Chinese))
- [13] 申金玉, 石亚东, 甘升伟, 等. 太湖流域湖西区入湖水量变化趋势及成因分析[J]. 水资源保护, 2011, 27(4):48-50. (SHEN Jinyu, SHI Yadong, GAN Shengwei, et al. Changing trend of water entering western area of Taihu Lake Basin and causalanalysis[J]. Water Resources Protection, 2011,27(4):48-50. (in Chinese))
- [14] 戈礼宾, 潘彩英, 张泉荣, 等. 太湖水位资料系列分析与水位特征值计算[J]. 水资源研究, 2008, 29(4):7-8. (GE Libing, PAN Caiying, ZHANG Quanrong, et al. The analysis with the water leveland eigenvalue calculation in Taihu Lake [J]. Water Resources Research, 2008, 29(4):7-8. (in Chinese))
- [15] 李文运, 张伟, 戈建民, 等. 水量平衡分析方法及应用[J]. 水资源保护, 2011, 27(6):83-87. (ZHANG Wenyun, ZHANG Wei, GE Jianmin, et al. Water balance analysis method and its application[J]. Water Resources Protection,2011, 27(6):83-87. (in Chinese))
- [16] 周文斌, 车倩. 多元线性回归法在水文预报中的应用[J]. 山西建筑, 2009, 35(1):359-360. (ZHOU Wenbin, CHE Qian. Use of multi factor line regression method for hydrogicalforecast [J]. Shanxi Architecture, 2009, 35(1):359-360. (in Chinese))