湖泊表面水温预测与可视化方法研究*

杨 昆[1,2],喻臻钰[1,2],罗 毅[1,2],商春雪[3],杨 扬[1,2]
(1. 云南师范大学 信息学院 昆明 650500;
2. 西部资源环境地理信息技术教育部工程研究中心 昆明 650500;3. 云南师范大学教务处 昆明 650500)

摘 要:湖泊表面水温是水生态环境的重要因子,直接影响流域生态系统及生物多样性。准确获取湖泊表面水温、预测表面水温时空变化过程是控制和改善流域水生态环境的基础,同时也是预防和治理蓝藻水华爆发的关键。为此,以滇池为研究区,以2005~2016 年 10 个水质监测站点的 54 个水质数据(水温、叶绿素 a、pH、高锰酸盐指数、溶解氧等)为基础数据集,将支持向量回归(SVR)、主成分分析法(PCA)及反向传播人工神经网络(BPANN)3 种算法相结合,组成混合预测模型,并将克里金插值法与地理信息系统相结合,实现滇池水温 12 年来历史变化过程的情景再现及未来 5 年变化趋势的情景模拟。研究结果表明,模型的平均相对误差为 0.5%,均方根误差为 1.452 3,R^2 = 0.904 9,具有误差低、泛化高的综合预测性能;空间可视化分析结果表明,2005~2020 年水温高于 20℃的区域呈现北向南扩散趋势,蓝藻水华爆发可能性由局部性变为全面性,这与昆明市快速城镇化发展及全球气候变暖密切相关。

关键词: 湖泊表面水温;支持向量回归;主成分分析;反向传播人工神经网络;蓝藻水华
中图分类号: X82 TH89 **文献标识码**: A **国家标准学科分类代码**: 610.3040

Lake surface water temperature prediction and visualization

Yang Kun[1,2], Yu Zhenyu[1,2], Luo Yi[1,2], Shang Chunxue[3], Yang Yang[1,2]
(*1. School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China*;
2. The Engineering Research Center of GIS Technology in Western China of Ministry of Education of China, Kunming 650500, China;
3. Teaching Affairs Department, Yunnan Normal University, Kunming 650500, China)

Abstract: Lake surface water temperature (LSWT) is an important factor in aquatic environment, which directly affects watershed ecosystem and biodiversity. Precise LSWT measurement and prediction is essential to control and improve the aquatic ecological environment of the river basin, and is also the key to prevent and control the outbreak of cyanobacteria bloom. Focusing on Dianchi Lake, 54 water quality parameters (LSWT, chlorophyll a, pH, permanganate index, dissolved oxygen, etc.) of 10 water quality monitoring sites from 2005 to 2016 are used as the data set. A hybrid forecasting model is presented composed of ε-support vector regression (ε-SVR), principal component analysis (PCA) and back propagation artificial neural network (BPANN). Moreover, Kriging method is combined with geographic information system (GIS) to realize the scene reproduction of the historical changes of the Dianchi lake LSWT and water quality in the past 12 years and the trend simulation of the next 5 years. Results show that the average relative error of the model is 0.5%, the mean square error is 1.452 3, R^2 is 0.904 9. Spatial visualization results indicate that the region with LSWT over 20 ℃ diffuses obviously from north to south. The outbreak of cyanobacteria bloom changes from locally to globally, which is related to the expansion of Kunming urbanization and meteorological environment.

Keywords: lake surface water temperature; support vector regression (SVR); principal component analysis (PCA); back propagation artificial neural network (BPANN); cyanobacteria

收稿日期:2017-05 Received Date: 2017-05
*基金项目:国家自然科学基金(41761084)、国家 863 计划(2012AA121402)、教育部博士点专项基金(20115303110002)、云南省自然科学基金青年(2016FD020)项目资助

0 引　　言

湖泊表面水温(lake surface water temperature, LSWT)—湖泊 0 ~ 1 m 的水温,是湖泊生态环境监测与评价研究领域中重要的物理参数。湖泊表面水温变化直接影响湖泊物种结构和分布[1]。2015 年《Nature》杂志在 NEWS 板块及其子刊《Scientific Data》报道了近 30 年来全球大部分湖泊表面水温快速上升这一研究发现[2-3],湖泊表面水温升高对湖泊生态系统的影响巨大,动态监测湖泊表面水温变化、模拟和预测湖泊表面水温变化时空过程是目前国内外学者研究的热点问题。从数据获取难易程度考虑,与其他水质参数相比,湖泊表面水温更容易观测,能够从较大的空间尺度和较长的时间尺度开展研究工作;从参数敏感程度考虑,当区域环境(包括气候环境及地貌环境)发生变化,湖泊表面水温的响应最为敏感迅速。

滇池是云南省最大的高原湖泊,也是我国第六大淡水湖。近 30 年来,滇池水质由 20 世纪 60 年代的 II 类恶化为劣 V 类,成为我国污染最严重的湖泊之一[4]。虽然各级政府部门投入大量物力与和人力,对滇池流域生态环境进行长期的监测与治理[5]。但在全球气候变暖(自然环境的改变)、城镇化快速发展(人文环境的改变)双重作用下,滇池水体富营养化日趋严重,蓝藻水华频繁爆发,目前滇池仍是我国污染最严重的湖泊之一[6-8]。导致滇池流域保护与治理成效不明显的主要原因是目前对城市型湖泊水污染的形成机理认识不够充分,忽略了湖泊水体表面水温的变化对滇池生态系统平衡造成的影响。有必要从微观尺度监测湖泊表面水温变化机理,从宏观尺度分析湖泊表面水温变化时空过程,进而从源头控制和改善湖泊水生态环境,监测和预警蓝藻水华爆发等水环境突发事件。

有关湖泊表面水温变化对生态环境方面的研究,国外,Brookes J. D. 等人[9]研究结果表明,蓝藻水华爆发主要是由于湖泊营养盐浓度及湖泊表面水温两个参数共同作用的结果。国内,朱根海等人[10]研究结果表明,近 40 年来,南极气温升高了 0.6℃,同时伴随藻类丰度迅速升高,两者具有密切的关系,调查结果表明,水温与营养盐是影响湖泊藻类生长繁殖的主要因子。上述研究结果表明,湖泊表面水温直接影响流域气、地、水间相互作用过程中物质能量的交换,是湖泊生态环境中最重要的指标,连续动态监测湖泊表面水温、模拟湖泊表面水温历史变化变化过程、预测湖泊表面水温未来变化趋势是控制和改善湖泊水生态环境的基础。

有关湖泊水环境预测方法方面的研究,国外,Dai C 等人[11]提出一种基于支持向量回归(support vector regression, SVR)和遗传算法的模型,用于识别水传递之间的功能关系和湖泊营养状况,并应用于牛兰江-滇池水运工程的流域内水转运管理;Zakhem B. A. 等人[12]以 83 个地下水样品为数据,采用主成分分析方法,将 14 个变量缩减为 4 个主成分,并以此证明了多变量统计分析在地下水水质变化研究中的有效性。国内,张淑清等人[13]基于 PCA 法提出处理多天气因素的基于 L-M 优化算法的 BP 神经网络(levenberg marguardt back propagation, LMBP)电力负荷预测模型,采用 LMBP 进行预测分析,有效地提高了网络的收敛速度和泛化能力;赵超等人[14]针对污水处理过程建模中样本数据可能存在的测量误差对模型性能的影响,提出一种自适应加权最小二乘支持向量机(adaptive weighted least squares-support vector machine, AWLS-SVM)回归的软测量建模方法,并应用此方法建立污水处理过程出水水质关键参数的软测量模型,获得了较好的效果;赵彦涛等人[15]针对输入变量选择易受时延影响的问题,提出一种基于互信息和最小二乘支持向量机(mutual information least square support vector machine, MI – LSSVM)的软测量建模方法;张宝印等人[16]针对检测数据少、效率低等问题,提出一种基于主成分分析和改进的轮换对称分块支持向量机的损伤识别算法,同时引入遗传算法获得更好的分类性能,结合主成分分析提取主特征向量降低向量维度和缩短训练时间,最终实现了可靠的算法;李小珉等人[17]在 BP 神经网络学习收敛速度慢、易陷入局部极小的缺陷上,加入具有良好搜索全局最优解能力的遗传算法,改进后的方法具有较好的非线性拟合能力和更高的预测准确性;丁硕等人[18]提出了一种将灰关联分析与 BP 神经网络相结合的预测模型,并通过对单一模型的对比分析得到组合模型具有更高的精度与更精简的结构。

本文在前期研究成果的基础上[19-21],提出将支持向量回归 ε-SVR、主成分分析及反向人工神经网络(back propagation artificial neural network, BPANN)相结合,组成混合水温预测模型;利用滇池 10 个水质监测站点(白鱼口、草海中心、滇池南、断桥、观音山东、观音山西、观音山中、海口西、晖湾中、罗家营)2005 ~ 2016 年的历史实测数据及研究团队实测到的水质数据作为建模数据集和验证集,对滇池流域 2005 ~ 2020 年的历史水温变化过程和未来水温变化趋势进行预测分析,并利用地理空间分析手段实现滇池水环境水温变化时空过程模拟。

1 研究区域与数据源

滇池位于云贵高原中部,呈南北向分布,处于亚热带

高原西南季风气候区,气候变化主要受西南季风和热带大陆气团交替控制[22],湖面海拔约 1 886 m,面积约 330 km^2,平均水深约为 5 m,属于半封闭型湖泊,仅有西南部的海口为出水口[23],对滇池流域的气候调节具有重要作用。滇池是流域灌溉、调畜、受纳的主要水体,也是城市发展的载体。同时,随着国家一带一路战略的提出和实施,昆明市成为了面向南亚、东南亚重要的辐射中心,控制和改善滇池生态环境是推进“一带一路”战略的环境基础。为此选择昆明滇池作为研究区具有区域特色(高原城市型湖泊)和时代特色(一带一路战略提出)。本研究所使用的历史数据来源于云南省环境科学研究院,包括白鱼口、草海中心、滇池南、断桥、观音山东、观音山西、观音山中、海口西、晖湾中、罗家营等 10 个水质监测站点的 2005 年 1 月 1 日 ~2016 年 12 月 30 日的水温、叶绿素 a 浓度、pH 值等 54 个参数的日监测数据,以及研究团队利用设计的水质监测节点实测的水质数据。研究区及监测站点地理位置分布如图 1 所示。

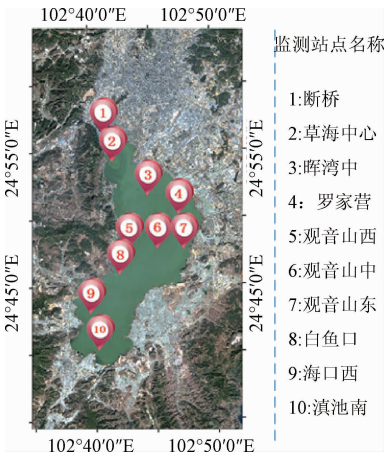


图 1 研究对象监测点分布

Fig. 1 Monitoring sites distribution map

2 预测模型构建

2.1 ε-SVR

SVR 建立于数理统计学基础,基于结构风险最小优化原则,使得神经网络应用中较复杂的结构选择问题转换为相对较为容易的核函数选择问题。SVR 将原凸二次优化问题转为有更简单变量约束的对偶凸二次优化问题,确保找到全局最优解的基础上,较好地解决数据量小、维度高、非线性的问题,具有很好的推广能力,同时巧妙解决了维数灾难问题。ε-SVR 在 SVR 的基础上加入了不敏感损失函数 ε,将 SVM 推广到非线性系统的回归估计,展现了极好的学习能力,其建模过程如下。

已知训练集:

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i)\} \in (x \times y) \tag{1}$$

式中:$x_i \in R^n$, $y_i \in R^n$, $i = 1,2,\cdots,n$。首先选择恰当的 $K(x, x')$ 核及待定系数。常见的核函数有多项式(Polynomial)核函数为 $K(x, y) = [\varepsilon(x \cdot y) + C]^d$ ($d = 1,2,\cdots$);径向基(radial basic function, RBF)核函数为 $[K(x, y) = \exp(-\varepsilon \|x - y\|^2)]$;Sigmoid 核函数为 $K(x, y) = \tanh\varepsilon(x \cdot y) + C$,式中 d、ε、C 为待定系数;然后最优化问题可通过下式求得:

$$\min \frac{1}{2}\sum_{i,j=1}^{n}(a_i^* - a_i)(a_j^* - a_j)K(x_i \times x_j) + \varepsilon\sum_{i=1}^{n}(a_i^* - a_i) - \sum_{i=1}^{n}y_i(a_i^* - a_i) \tag{2}$$

式中:$\sum_{i=1}^{n}(a_i^* - a_i) = 0$, $0 \leqslant a_i, a_i^* \leqslant c/n$, $i = 1,2,\cdots,n$。从而求得支持向量:

$$W_0 = \sum_{i=1}^{n}x_i(a_i^* - a_i) \tag{3}$$

在求取有约束条件的优化问题时,拉格朗日乘子法可以求取等式约束最优值。若含有不等式约束,KKT(Karush-Kuhn-Tucker)条件可以求取最优值。根据 KKT 条件,有 $a_i^* \times a_i = 0$ 成立,即仅有支持向量对应的拉格朗日乘子不为 0。因此,能够仅用训练样本的少数支持向量实现函数估计。

接着选取不同核函数 $K(x_i, x_j)$ 作为内积回旋。对于非线性回归问题,可通过非线性变换将输入向量映射到高维特征空间,转化为类似的线性回归问题解决。为避免高维特征空间的“维数灾难”问题,采用 Hilbert 空间中内积的回旋形式,用输入空间的一个核函数等效高维特征空间的内积形式。

最后,构造决策函数:

$$f(x) = \sum_{i=1}^{n}\beta_i K(x_i, x) + \bar{b} \tag{4}$$

$$\beta_i = \bar{a}_i^* - \bar{a}_i \quad i = 1,2,\cdots,n \tag{5}$$

式中:$\bar{b}$ 以式(6)、(7)计算,$\bar{a}_j \in (0, c/n)$,$\bar{a}_k^* \in (0, c/n)$,若选取 $\bar{a}_j$,则:

$$\bar{b} = y_i - \sum_{i=1}^{n}\beta_i K(x_i, x_j) + \varepsilon \tag{6}$$

若选取 $\bar{a}_k^*$,则:

$$\bar{b} = y_k - \sum_{i=1}^{n}\beta_i K(x_i, x_j) - \varepsilon \tag{7}$$

2.2 PCA

PCA 法通过降维技术将原变量中多个具有一定相关性的指标进行一系列线性组合约化为少数几个综合指标,并使新变量在彼此不相关的前提下尽可能多地反映原变量的信息,被广泛用于指标合成。PCA 数据信息主要反映于数据变量的方差,方差越大,包含信息越多,以

累计方差贡献率衡量。通常一种好的指标合成技术,应尽可能少地丢失原始信息,因此,最终主成分个数的选择,将取决于其对原始指标变量的解释程度。第一主成分是数据变异最大的方向,只取第一主成分是一种极端的强行丢弃降维方法,其前提是第一主成分的方差贡献率足够大,PCA 具体步骤如下。

首先建立自相关矩阵 $\boldsymbol{R}$,并计算其特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$ 与特征向量 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m$,即:

$$\boldsymbol{R} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}^{*}}{(N-1)} \tag{8}$$

式中:$\boldsymbol{X}^{*}$ 为归一化处理后的数据矩阵。

然后确定主成分个数,方差贡献率 η_i 和累计方差贡献率 $\eta_{\Sigma}(p)$ 分别为:

$$\eta_i = \frac{100\%\,\lambda_i}{\sum_{i}^{m}\lambda_i} \tag{9}$$

$$\eta_{\Sigma}(p) = \sum_{i}^{p}\eta_i \tag{10}$$

通常累计方差贡献率大于 75% ~95% 时,对应的前 p 个特征值大于 1 的主成分便包含 m 个原始变量所能提供的绝大部分信息,主成分个数就是 p 个。那么 p 个主成分对应的特征向量为:

$$\boldsymbol{U}_{m\times p} = [u_1, u_2, \cdots, u_p] \tag{11}$$

则 n 个样本的 p 个主成分构成的矩阵为:

$$\boldsymbol{Z}_{N\times p} = \boldsymbol{X}^{*}_{N\times m}\boldsymbol{U}_{m\times p} \tag{12}$$

2.3　BPANN

BPANN 人工神经网络即反向传播人工神经网络,由一个含有 N 个节点的数据输入层 s,一个含有 H 个节点的隐含层 h 和仅有 1 个节点的输出层 r 3 层网络组成,各相邻两层之间单向传播,其学习规则采用梯度下降法,并通过阈值的判断反向传播而不断调整网络权值,使得整个网络的误差平方和最小,BPANN 的建模过程如下。

首先从输入节点输入样本的 N 个特征值,并确定激活函数向前传播。模型隐含层神经元采用 tansig 激活函数,输出层神经元采用线性函数激活函数。

隐含层节点的输出为:

$$h(t,j) = \int\sum_{k=1}^{N} x(t,k)w(t,j,k) + \theta(t,j) \tag{13}$$

输出层节点的输出为:

$$y(t) = \int\sum_{t=1}^{N} h(t,j)w_h(t,j) + \theta(t) \tag{14}$$

式中:$w(t,j,k)$ 为输入层 k 节点对隐含层 j 节点的连接权值,$w(t,j)$ 为隐含层 j 节点对输出节点的连接权值,$x(t,k)$ 为输入的第 t 个样本的第 k 个特征值,$\theta(t,j)$ 和 $\theta(t)$ 分别为隐含层和输出层的阈值。

然后将式(13)、(14)中表示网络各层间连接权值 w、w_h 和阈值 θ 取(-1,1)的随机量作为初始值,作为输入样本进行学习。每次学习完成,比较实际输出值与期望值的误差,若误差小于指定精度,则学习结束并输出此时的最佳结果,否则将误差信号沿原连接路径进行反向传播,并逐步调整各层的连接权值和阈值,直到误差小于指定精度、训练次数达到指定次数或训练时间达到上限时结束。

2.4　ε-SVR-PCA-BPANN 组合预测模型构建

综上所述,ε-SVR 不过分依赖样本集的数量,且学习与泛化能力较优于 BPANN,但其需要计算和存储核函数矩阵,当样本集较大时使得计算复杂度大幅增加;BPANN 收敛速度快,结构简单,具有全局逼近能力,不存在局部最小问题,但学习方法采用经验最小化原则,带有很大的经验成分,可能出现过学习问题。为解决上述问题,并兼顾预测精度和计算复杂度,加入 PCA 方法考虑各特征变量与整体变量的相关性因素,以实现多变量较高精度预测为目的,综合以上 3 种方法的优势,构建基于 ε-SVR-PCA-BPANN 的组合预测模型,实现水温的模拟预测,模型构建原理如图 2 所示。

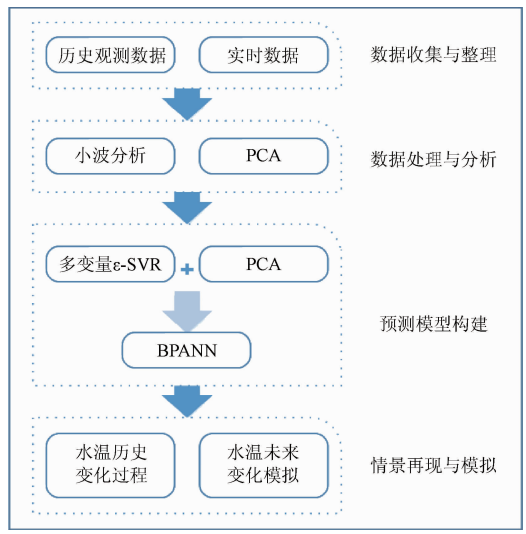


图 2　组合预测模型原理

Fig. 2　Schematic diagram of the combined forecasting model

基于 ε-SVR-PCA-BPANN 的组合预测模型构建过程如下:

1)数据收集与整理

(1)非零缺失值处理。当样本的数据量较大时,由于传感器故障及数据传输的问题,缺失值难以避免,如果缺失值在 10% 范围内,可以通过随机插值、均值填充、临近数据填充等方法进行补偿。由于原始水质数据为日数据,其样本量较大,为了避免缺失值对预测结果的影响,采用线性插值法对缺失值进行处理。

(2)剔除异常值。在统计检验方法中,由于传感器

工作过程中难免出现因外部环境因素的干扰导致数据异常,故有必要对样本空间的异常值进行剔除。剔除异常值较为常用的方法有 Grubbs 检验、Dixon 检验和 t-检验。其中,t-检验的效果最佳,本文选取此方法进行异常值的剔除。

(3)原始数据标准化处理。为消除量纲不同、数值差异过大而带来的影响,对原变量作标准化处理。假设有 m 个指标 x_1, x_2,…,x_m 分别表示 N 个对象的特性,用 $N\times m$ 矩阵表示,即:

$$\boldsymbol{X}_{N\times m}\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} \end{bmatrix} \tag{15}$$

进行中心标准化处理生成标准矩阵 $\boldsymbol{Y}$,即:

$$x_{ij}^{*} = \frac{x_{ij}-\bar{x}_j}{s_j} \tag{16}$$

式中:$i=1,2,\cdots,N$,$j=1,2,\cdots,m$。$\bar{x}_j$、s_j 分别为指标变量 x_j 的均值和方差。

2)数据处理与分析

(1)相关性分析与降维。由于样本数据达到了 54 维,若直接将数据作为模型输入,将大大降低模型预测效果,且难以在有效时间内训练得到满意结果,因此有必要对样本数据进行降维处理。目前常用的降维方法有缺失值比率、随机森林、主成分分析等。其中,缺失值比率方法是将数据变量缺失值大于某个阈值的变量去除,可简便、快捷地直接删除某些缺失信息较多的变量;随机森林方法由于不能给出一个连续性的输出,更适用于分类问题;PCA 能够在缺失信息很少的前提下,将多个指标正交变换为几个综合指标(主成分),相对其余方法具有更高准确性。

综合几种方法的适用性及优缺点,本研究采用缺失值比率和 PCA 结合的方法,避免了因缺失信息过多而无法进行 PCA 分析的不足(此时的缺失值包括非零缺失值与 0)。通过缺失值分析,淘汰缺失值比率达到 10% 的变量 20 个,剩下的变量重新生成为新的样本进行 PCA 处理得到 18 个变量,再进行相关性分析与显著性检验,最终确定 15 个变量作为模型的输入,按月取均值处理后生成样本数据集。

(2)小波分析。由于各种气象因子、水文过程以及生态系统与大气之间的物质交换过程都可以看作是随时间有周期性变化的信号,因此小波分析方法同样适用于湖泊水环境领域,从而对各种湖泊水环境变化过程复杂的时间格局进行分析。

可以将具有等时间步长 δ_t 的离散时间系列 x_n($n=1,\cdots,N$)的连续小波变换定义为小波函数 ψ_0 尺度化以及转换下的 x_n 的卷积:

$$W_N^X(s) = \sqrt{\frac{\delta t}{s}}\sum_{n=0}^{N-1} x_n \cdot \varphi^{*}\left[\frac{(n'-n)\delta t}{s}\right] \tag{17}$$

式中:* 表示共轭复数,N 是时间系列的总数据个数,$(\delta_{t/s})1/2$ 是一个用于小波函数标准化的因子从而使得小波函数在每个小波尺度 s 上具有单位能量。

Morlet 小波不但具有非正交性而且还是由 Gaussian 调节的指数复值小波。

$$\varphi_0(t) = \pi^{-1/4}\mathrm{e}^{\mathrm{i}\omega_0 t}\mathrm{e}^{-t^2/2} \tag{18}$$

式中:t 为时间,ω_0 是无量纲频率。当 $\omega_0=6$,小波尺度 s 与傅里叶周期基本相等($\lambda = 1.03$ s),所以尺度项与周期项可以相互替代。由此可见,Morlet 小波在时间与频率的局部化之间有着很好的平衡。此外,Morlet 小波中还包含着更多的振动信息,小波功率可以将正、负峰值包含在一个宽峰之中。

$|W_n^X(s)|^2$ 定义为小波功率谱,该功率谱表达了时间系列在给定小波尺度和时间域内的波动量级。由于采用的 Morlet 母小波为复值小波,因此 $W_n^X(s)^2$ 也为复数,其复值部分可以解释为局部相位。将小波功率谱在某一周期上进行时间平均,可以得到小波全谱:

$$W^2(s) = \frac{1}{N}\sum_{n=0}^{N-1}|W_n(s)|^2 \tag{19}$$

小波全谱能够表明时间系列真实功率谱的无偏、一致估计。由于小波全谱可以显示出背景谱量度,所以局部小波谱的峰值可以得到验证。通过小波全谱图中可以清晰的辨别时间系列的周期波动特征及其强度。

3)基于 ε-SVR 分段训练样本

将预处理后样本数据集 S 的各个变量进行分段处理,并将其作为 ε-SVR 模型的输入变量,设置核函数类型为 RBF 径向基函数,标准正态分布的双侧检验概率为 0.001,进行 5 倍交叉验证,从而寻找并获取最佳 c、g 参数,然后进行 ε-SVR 分段训练得到样本集 $S1$。

4)基于 PCA 计算变量权重

计算预处理后样本数据集 S 的相关系数矩阵、特征根、相应的标准特征向量以及贡献率,并依据累计贡献率提取主成分,综合考虑第一主成分与最终预测变量抽取比例较高的成分,选取此成分的提取系数作为计算权重 W。

5)基于 ε-SVR 与 PCA 的样本数据优化

将 ε-SVR 与 PCA 处理后的数据集进行组合得到数据集 $S2$。

$$S2 = \sum_{i=1}^{n} S1(i)\times W(i) \tag{20}$$

式中:n 为变量总数,$S1(i)$ 为步骤 2)处理后的数据集中的第 i 个变量,W 为步骤 3)处理后的数据集中的第 i 个权重。

6)基于 BPANN 的预测结果输出

将 $S2$ 的 n 个变量以 n 行矩阵的形式作为 BPANN 模

型输入变量,设置模型最大训练次数为 1 000 次,训练要求精度为 0.01,学习率为 0.01,输出变量为水温,隐含层为 3,隐含节点为 5,并以此模型进行预测得到最佳组合模型的预测结果。

3　结果与分析

3.1　算法实现与测试

以滇池 2005 年 1 月 1 日 ~2016 年 12 月 31 日 10 个监测站点 54 个水质因子的日监测数据为原始数据,在数据预处理中,使用线性插值法替换非零缺失值、t-检验排除异常值,然后对数据归一化处理,最后使用缺失值比率和 PCA 进行降维,从而获得 15 个变量、144 组数据作为模型输入样本集 S。其中,抽取 116 组数据(约占总样本数据的 80.56%)为训练集,28 组数据(约占总样本数据的 19.44%)为测试集。降维之后的 15 个变量的前 3 个主成分分析如图 3 所示。

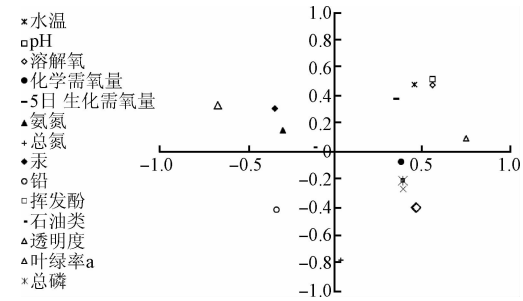
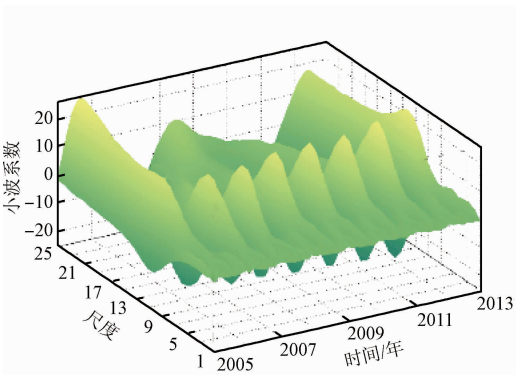


图 3　主成分分析结果

Fig. 3　Results of Principal component analysis

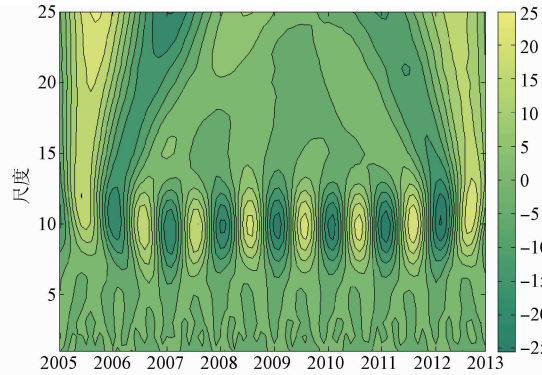
特征向量大于 1 的一共有 5 个成分,前 3 个主成分累积贡献率达到 41.28% ,其中第一主成分贡献率达到了 17.73% 。在第一主成分中,提取比例较大的是叶绿素 a 浓度、水温、pH、透明度,总氮含量抽取比例较小。

采用 Morlet 小波变换进行周期性分析,分析结果如图 4 所示。X、Y、Z 轴分别表示时间(年份)、尺度和小波系数。



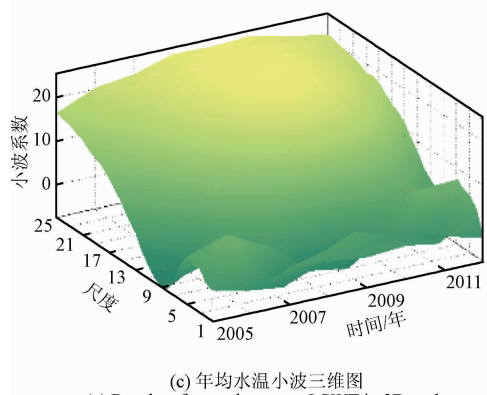
(a) 月均水温小波三维分析结果

(a) Results of month average LSWT in 3D scale average



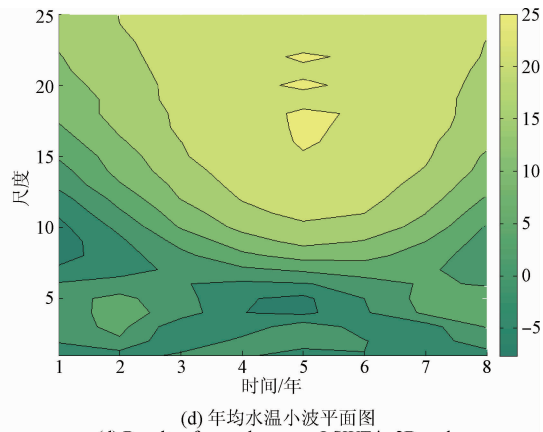
(b) 月均水温小波平面图

(b) Results of month average LSWT in 2D scale



(c) 年均水温小波三维图

(c) Results of annual average LSWT in 3D scale



(d) 年均水温小波平面图

(d) Results of annual average LSWT in 2D scale

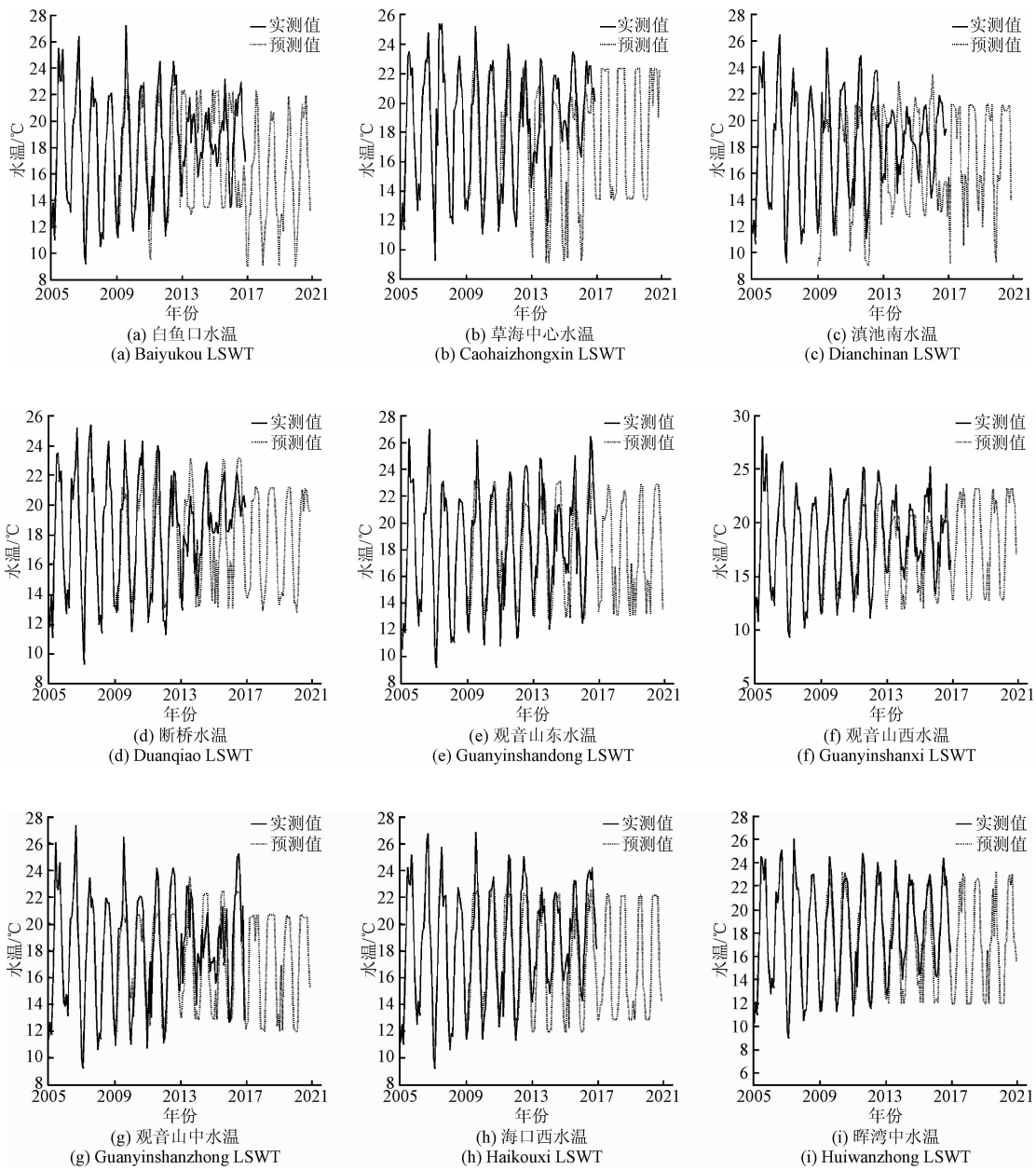
图 4　Morlet 小波分析结果

Fig. 4　Morlet wavelet analysis results

以月为单位的小波分析中,表现出 6 ~ 14 年尺度的中低频周期变化规律的水温演变过程,且在此尺度上出现了冷-暖交替的现象;以年为单位的小波分析中,存在着 2 ~ 5 年、16 ~ 18 年、20 ~ 21 年和 22 ~ 23 年尺度的周

期变化规律,在 2 ~ 5 年尺度上出现了冷-暖-冷的震荡,且此周期变化在整个分析阶段表现极其稳定,在其余尺度上出现的震荡不稳定。且 16 ~ 18 年、20 ~ 21 年和 22 ~ 23 年尺度的能量最强,但周期变化具有局部性(2008 年中旬 ~ 2009 年中旬),虽然 2 ~ 5 年尺度上能量较弱,但周期分布比较明显,几乎贯穿整个研究时域。水温演变时间序列中存在 3 个较为明显的峰值,依次对应 4 年和 17 年,最大峰值为 17 年,表明此时间尺度的周期震荡最强,为年均水温变化的第一主周期,4 年为第二主周期,2006 年中旬为第一主周期与第二主周期的交界点,即产生突变,这两个周期的波动控制着整个时域内年均水温的变化特征。

利用 ε-SVR 分别对样本集 S 的各子样本序列进行训练得到样本集 $S1$,同时利用 PCA 计算权重 W,然后将 $S1$ 与 W 组合得到的样本集 $S2$ 作为 BPANN 的输入样本集,预测结果如图 5(a) ~ (j)所示。通过多种模型的大量组合实验,本研究提出的 ε-SVR-PCA-BPANN 组合模型平均相对误差最低($SA = 0.5\%$),通过了 $\alpha = 0.001$的假设性检验,预测结果显著($P = 2^{-16} < 0.001$),各模型预测结果的均方根误差及相对误差如图 5(k) ~ (l)所示。各类模型的预测误差统计如表 1 所示。



(a) 白鱼口水温
(a) Baiyukou LSWT

(b) 草海中心水温
(b) Caohaizhongxin LSWT

(c) 滇池南水温
(c) Dianchinan LSWT

(d) 断桥水温
(d) Duanqiao LSWT

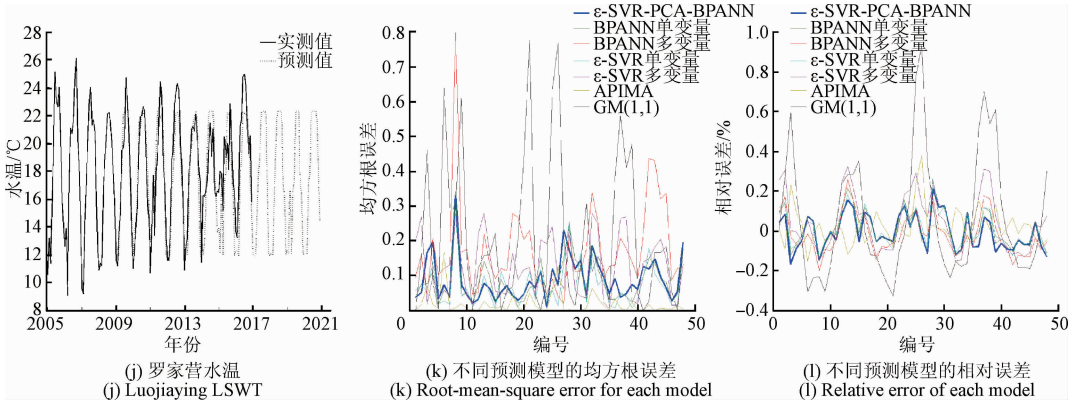
(e) 观音山东水温
(e) Guanyinshandong LSWT

(f) 观音山西水温
(f) Guanyinshanxi LSWT

(g) 观音山中水温
(g) Guanyinshanzhong LSWT

(h) 海口西水温
(h) Haikouxi LSWT

(i) 晖湾中水温
(i) Huiwanzhong LSWT



(j) 罗家营水温
(j) Luojiaying LSWT

(k) 不同预测模型的均方根误差
(k) Root-mean-square error for each model

(l) 不同预测模型的相对误差
(l) Relative error of each model

图 5　2005 ~ 2020 年滇池各监测站点水温随时间变化及不同预测模型的误差

Fig. 5　Monitoring station chlorophyll a Changing characteristic from 2005 to 2020 and different prediction model error

表 1　各类算法的误差统计

Table 1　Statistical error data of various algorism

指标	组合方法	BPANN 单变量	ε-SVR 单变量	BPANN 多变量	ε-SVR 多变量	ARIMA	GM(1,1)
均方根误差	1.452 3	2.600 5	2.443 6	2.004 2	1.852 0	2.037 1	4.366 8
平均相对误差	0.005 0	0.015 7	0.046 8	0.012 4	0.011 0	0.027 9	0.067 9
确定系数	0.904 9	0.816 1	0.879 4	0.888 3	0.899 7	0.844 9	0.143 9

3.2　预测结果可视化分析

克里金(Kriging)法也称空间局部估计或空间局部插值,该方法在考虑了样点的形状、大小和空间相互位置关系、待估样点相互空间位置关系以及变异函数提供的结构信息之后,对该待估样点值进行线性无偏最优估计。通过 Kriging 法插值,将协方差函数和变异函数描述的空间相关性或空间异质性在二维平面上以某种格局的形式表现,不但对分析滇池流域水温的区域化现象具有重要意义,而且对定量研究其空间分布格局的特点和格局的空间构型提供了有效支持。湖泊表面水温与叶绿素 a 浓度复合时空分布如图 6 所示。

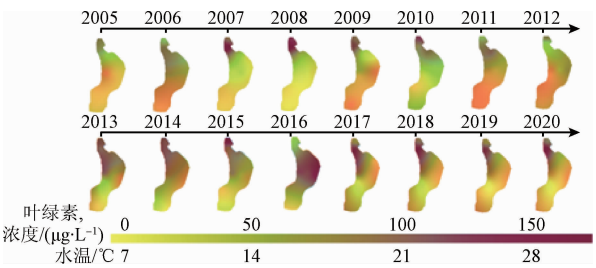


图 6　2005 ~ 2020 年滇池水温与叶绿素 a 浓度的空间分布

Fig. 6　Spatial distribution of LSWT and Chlorophyll a concentration in Dianchi Lake from 2005 to 2020

以白鱼口监测点为例,年平均水温为 17.87℃(接近于蓝藻爆发的适宜温度),2006 年 9 月达到了最高值 26.4℃,2007 年 2 月为最低值 9.2℃。分析结果表明,每年 6 月 ~ 9 月为高温区,10 月 ~ 次年 2 月为低温区,叶绿素 a 浓度存在一致的明显波动周期,每年叶绿素 a 浓度的最高值均在高温区间内,2005 年叶绿素 a 浓度的波动最小,但在 7 月 ~ 次年 2 月之间仍存在一个波动周期。其余月份的波动相对较小,可将其视为一个波动周期,即每年 3 月 ~ 6 月。随着年份的增加,周期波动的规律越来越明显。

蓝藻生长的最佳温度为 30 ~ 35℃,水温为 26℃时,最适宜于蓝藻的聚集、上浮而形成水华;当水温低于 20℃时,水生高等植物抑藻能力较强,当水温高于 20℃时,其抑藻能力急剧下降。滇池流域 2005 ~ 2016 年均水温为 17.87℃,7 月 ~ 10 月平均水温为 22.35℃,11 月 ~ 次年 2 月的平均水温为 13.87℃。分析结果表明 7 月 ~ 10 月内的蓝藻生长温度最为适宜,为蓝藻爆发提供了有利条件。综上所述,以蓝藻水华形成的“4 阶段理论”,结合滇池流域气候因素与其他各监测点的实测数据进一步分析表明,可将滇池蓝藻水华生长规律归纳为 3 月 ~ 6 月为复苏阶段,7 月 ~ 10 月为生长与上浮聚集阶段,11 月 ~ 次年 2 月为衰亡与休眠阶段。

地理空间分析结果表明 2005 ~ 2016 年,水温高于 20℃的区域主要位于滇池北部的草海和滇池的西南部和东部。2016 年之后,叶绿素 a 浓度高于 100 μg/L 的分布区域向西南方向移动,且 2015 ~ 2020 年叶绿素 a 浓度高于 100 μg/L 的区域覆盖范围高达 30% 。

4 结 论

本文提出将 ε-SVR、BPANN 及 PCA 3 种方法进行组合,构建水温长期预测模型,并利用滇池 10 个监测站点 2005 ~ 2016 年来获取的 54 个水质因子作为样本空间,实现对滇池 2017 ~ 2020 年以来 10 个监测站点水温预测,同时利用地理空间分析手段实现了滇池 2005 ~ 2020 年水温时空变化的情景模拟与预测。

研究结果表明,模型的平均相对误差为 0.5% ,均方根误差为 1.452 3,R^2 = 0.904 9,具有误差低、泛化高的综合预测性能;滇池水体表面水温总体呈上升趋势的同时表现出稳定的周期性变化;滇池蓝藻水华生长规律归纳为 3 月 ~ 6 月为复苏阶段,7 月 ~ 10 月为生长与上浮聚集阶段,11 月 ~ 次年 2 月为衰亡与休眠阶段;由于城镇化快速发展和气候变化等一系列自然与人文因素的相互作用和影响下,滇池水质仍在 V 类和劣 V 类之间波动,合理优化和调控城市发展,控制滇池表面水温升高趋势是控制和改善滇池水质的关键。

参考文献

[1] SCHMID M, HUNZIKER S, WÜEST A. Lake surface temperatures in a changing climate: A global sensitivity analysis[J]. Climatic Change, 2014, 124(1-2):1-15.

[2] KINTISCH E. Earth's lakes are warming faster than its air[J]. Science, 2015, 350(6267): 1449.

[3] SHARMA S, GRAY D K, READ J S, et al. A global database of lake surface temperatures collected by in situ and satellite methods from 1985 - 2009[J]. Scientific data, 2015(2):150008.

[4] TIAN C Y, LIU Z D, ZHANG Y H, et al. Hydrothermal liquefaction of harvested high-ash low-lipid algal biomass from Dianchi Lake: Effects of operational parameters and relations of products[J]. Bioresource Technology, 2015, 184(5):336-343.

[5] LI Z J, ZHENG Y X, ZHANG D W, et al. Impacts of 20-year socio-economic development on aquatic environment of lake Dianchi basin[J]. Journal of Lake Sciences, 2012, 24(6):875-882.

[6] GAO W, HOWARTH R W, SWANEY D P, et al. Enhanced N input to lake Dianchi basin from 1980 to 2010: Drivers and consequences[J]. Science of the Total Environment, 2015, 505(10):376-384.

[7] WANG Y, YANG H, ZHANG J, et al. Characterization of nalkanes and their carbon isotopic composition in sediments from a small catchment of the Dianchi watershed [J]. Chemosphere, 2015, 119(1):1346-1352.

[8] YU Q, CHEN Y, LIU Z, et al. The influence of a eutrophic lake to the river downstream: Spatiotemporal algal composition changes and the driving factors[J]. Water, 2015, 7(5):2184-2201.

[9] BROOKES J D, CAREY C C. Resilience to blooms[J]. Science, 2011, 334(6052): 46-47.

[10] 朱根海, 扈传昱, 何剑锋, 等. 全球气候变化对南极淡水藻类的影响[J]. 中国环境科学, 2010 ,30(3): 400-404.
ZHU G H, HU CH Y, HE J F, et al. Effects of global climate change on freshwater algae in Antarctica [J]. China Environmental Science, 2010,30 (3): 400-404.

[11] DAI C, TAN Q, LU W T, et al. Identification of optimal water transfer schemes for restoration of a eutrophic lake: An integrated simulation-optimization method[J]. Ecological Engineering, 2016, 95(10):409-421.

[12] ZAKHEM B A. Using principal component analysis (PCA) in the investigation of aquifer storage and recovery (ASR) in Damascus Basin (Syria)[J]. Environmental Earth Sciences, 2016, 75(15):1123.

[13] 张淑清, 任爽, 师荣艳,等. 基于多变量气象因子的 LMBP 电力日负荷预测[J]. 仪器仪表学报, 2015, 36(7):1646-1652.
ZHANG SH Q, REN SH, SHI R Y, et al. LMBP Daily load forecasting based on multivariable meteorological factors[J]. Journal of Instrument & Instrumentation, 2015, 36(7): 1646-1652.

[14] 赵超, 戴坤成, 王贵评,等. 基于 AWLS-SVM 的污水处理过程软测量建模[J]. 仪器仪表学报, 2015, 36(8):1792-1800.
ZHAO CH, DAI K CH, WANG G J, et al. Soft-modeling of sewage treatment process based on AWLS-SVM [J]. Journal of Instrument & Instrumentation, 2015, 36 (8):1792-1800.

[15] 赵彦涛, 单泽宇, 常跃进,等. 基于 MI-LSSVM 的水泥生料细度软测量建模[J]. 仪器仪表学报, 2017, 38(2):487-496.

ZHAO Y T, SHAN Z Y, CHANG Y J, et al. Soft measurement modeling of cement raw material fineness based on MI-LSSVM[J]. Chinese Journal of Scientific Instrument, 2017, 38(2):487-496.

[16] 张宝印, 董恩生. 基于 PCA-GA-RSPSVM 的复合材料损伤检测技术研究[J]. 电子测量与仪器学报, 2017, 31(9):1402-1407.
ZHANG B Y, DONG EN SH. Research on damage detection technology of composite materials based on PCA-GA-RSPSVM[J]. Journal of Electronic Measure-ment and Instrument, 2017, 31(9):1402-1407.

[17] 李小珉, 尹明. 基于遗传算法的 BP 神经网络电子系统状态预测方法研究[J]. 电子测量技术, 2016, 39(9):182-186.
LI X M, YIN M. Research on state prediction method of BP neural network electronic system based on genetic algorithm[J]. Electronic Measurement Technology, 2016, 39(9):182-186.

[18] 丁硕, 巫庆辉, 常晓恒,等. 基于灰色 BP 神经网络的实验材料供应预测[J]. 国外电子测量技术, 2016, 35(12):78-82.
DING SH, WU Q H, CHANG X H, et al. Prediction of experimental material supply based on grey BP neural network[J]. Foreign Electronic Measurement Technology, 2016, 35(12):78-82.

[19] LUO Y, YANG K, YU Z, et al. Dynamic monitoring and prediction of Dianchi Lake cyanobacteria outbreaks in the context of rapid urbanization[J]. Environmental Science & Pollution Research, 2017, 24(6):5335-5348.

[20] 杨昆, 陈俊屹, 罗毅,等. 滇池流域不透水表面扩张监测与时空过程分析[J]. 仪器仪表学报, 2016, 37(12):2717-2727.
YANG K, CHEN J Y, LUO Y, et al. Monitoring of surface impervious surface of Dianchi Lake Basin and analysis of spatiotemporal process[J]. Journal of Instrument & Instrumentation, 2016,37(12):2717-2727.

[21] 杨昆, 罗毅, 徐玉妃,等. 基于无线传感器网络与 GIS 的蓝藻水华爆发动态监测与模拟[J]. 农业工程学报, 2016, 32(24):197-205.
YANG K, LUO Y, XU Y F, et al. Dynamic monitoring and simulation of cyanobacteria bloom based on wireless sensor network and GIS [J]. Journal of Agricultural Engineering, 2016, 32(24):197-205.

[22] 何佳, 徐晓梅, 杨艳,等. 滇池水环境综合治理成效与存在问题[J]. 湖泊科学, 2015, 27(2):195-199.
HE J, XU X M, YANG Y, et al. Experimental study on the effect and existing problems of water environment comprehensive management in Dianchi Lake [J]. Chinese Journal of Lake Science, 2015, 27(2):195-199.

[23] 高喆, 曹晓峰, 黄艺,等. 滇池流域水生态功能一二级分区研究[J]. 湖泊科学, 2015, 27(1):175-182.
GAO ZH, CAO X F, HUANG Y, et al. Study on water ecology function of Dianchi lake basin [J]. Chinese Journal of Lake Science and Technology, 2015, 27(1):175-182.

作者简介

杨昆,1998 年于澳大利亚新南威尔士大学获得硕士学位,现为云南师范大学信息学院院长、博士生导师,主要研究方向为地理信息系统、数据融合等。
E-mail:kmdcynu@163.com

Yang Kun received his M. Sc. degree from the University of New South Wales, Australia in 1998. Now he is the dean of the School of Information and a doctoral supervisor in Yunnan Normal University. His main research interests include GIS and data fusion.

罗毅(通讯作者),分别在 2009 年、2012 年和 2014 年于哈尔滨理工大学获得学士学位、硕士学位和博士学位,现为云南师范大学软件工程系主任,主要研究方向为无线传感器网络、地理信息系统。
E-mail:luoyi861030@163.com

Luo Yi(Correspondent author) received his B. Sc., M. Sc. and Ph. D. degrees all from Harbin University of Science and Technology in 2009, 2012 and 2014, respectively. Now he is the director of Software Engineering of Yunnan Normal University. His main research interests include GIS and WSNs.