

基于主成分分析法与人工神经网络耦合模型的水质评价

朱永军, 吴 琼, 湛忠宇

(江苏省水文水资源勘测局南京分局, 江苏 南京 210008)

摘要:为了解决水质评价中评价指标权重难以合理确定、评价模型过于复杂、评价结果不合理等问题,将改进的主成分分析降维能力与人工神经网络自学习能力相结合,提出 PCA-BP 神经网络水质评价模型。实例分析表明,PCA-BP 神经网络在避免了传统的单因子评价法评价结果过于悲观、神经网络法模型复杂的同时,能够确定主要污染物,所得评价结果的合理性、准确性均能够得到保证。

关键词:主成分分析; 人工神经网络; 水质评价; 水环境改善; 六合区

中图分类号:TV21

文献标识码:B

文章编号:1007-7839(2021)08-0048-07

Study on water quality evaluation of Luhe District based on principal component analysis and artificial neural network coupling model

ZHU Yongjun, WU Qiong, ZHAN Zhongyu

(Nanjing Hydrology and Water Resources Survey Bureau of Jiangsu Province, Nanjing 210008, China)

Abstract: In order to solve the problems in water quality evaluation, such as difficulty in determining the weight of evaluation indexes, too complicated evaluation model and unreasonable evaluation results, the improved dimension-reduction ability of principal component analysis was combined with the self-learning ability of artificial neural network, and the PCA-BP neural network model for water quality evaluation was proposed. The case analysis showed that PCA-BP neural network could avoid the pessimistic evaluation result of the traditional single factor evaluation method, and the model of the neural network method was complex. At the same time, it could determine the main pollutants, and the rationality and accuracy of the evaluation results could be guaranteed.

Key words: principal component analysis; artificial neural network; water quality evaluation; water environment improvement; Luhe District

1 研究背景

合理对水体质量进行分析评价能够为水质治理提供科学的方向,同时也是区域改善水环境的基础,目前,存在多种对水质进行评价的方法,传统的单因子评价法^[1]以最差的水质指标所处等级作为评价结果,极易受到极端指标的影响,不能反映出

真实情况;灰色理论^[2],模糊数学法^[3]、层次分析法^[4]在确定指标权重时往往忽略了因子之间的相互影响且主观性较强;传统主成分分析法^[5-6]采用标准差对数据进行标准化,使得同类指标之间的方差为零,消除了指标之间的差异;传统的神经网络法^[7]将所有的监测指标作为输入数据,增加了模型复杂程度,效率低下。本文针对水质评价中评价指

收稿日期:2020-12-31

作者简介:朱永军(1970—),男,工程师,研究方向为水资源与水环境管理。E-mail:1139094860@qq.com

标权重难以合理确定,评价模型过于复杂、评价结果不合理等问题,依据南京市六合区的水质监测资料,提出主成分分析与神经网络相结合的方法,以期合理进行水质评价提供一种新思路。

2 研究方法

2.1 改进的主成分分析

主成分分析^[8]的主要思想是利用正交变换对原始数据进行降维处理,找出一组线性无关的主成分,以此代表原始数据的大部分信息,一般分为以下几个步骤。

1)对原始数据进行标准化处理,消除量纲不同带来的影响。本文采用均值化方法对原始数据进行标准化,在保留同类变量间的差异信息的同时,消除量纲的干扰,计算公式为

$$ZX_{ij} = X_{ij} / \text{Mean}X_i \quad (1)$$

式中, ZX_{ij} 为第*i*个指标的第*j*个数值标准化后的结果; X_{ij} 为第*i*个指标的第*j*个数据的原始值, $\text{Mean}X_i$ 为第*i*个样本的平均值。

由于水质指标中大部分为逆向指标,正向指标即数值越大表明水质越好的指标,采用下式进行标准化处理:

$$ZX_{ij} = (\text{Max}X_i - X_{ij}) / \text{Mean}(\text{Max}X_i - X_{ij}) \quad (2)$$

式中, $\text{Max}X_i$ 为第*i*个指标的最大值。

2)计算标准化后 ZX_{ij} 的相关系数矩阵 R ;

3)计算 R 的特征值 λ_i 和特征向量并将特征向量按照从大到小进行排列;

4)计算累计方差贡献率确定主成分个数。以前*n*个特征值的和占总特征值的百分比作为累计方差贡献*K*,一般取 $K \geq 85\%$;

5)计算主成分 F_i 的值及主成分综合得分,得分越高说明水质越差,其中

$$F = \sum_{1 \leq i \leq n} \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_n} F_i \quad (3)$$

2.2 BP 人工神经网络

BP神经网络^[9-10]是一种多层前馈神经网络,依靠大量神经元的联系,形成一个非线性的动态系统。BP神经网络一般由输入层、隐含层和输出层三部分组成,其中隐含层可以有一个或多个(图1)。在网络的运行过程中输入的数据由前向后传播,每一层的神经元输出结果只对与其直接相连的下一

层神经元有影响,同一层的神经元直接互不连接,互不干扰。

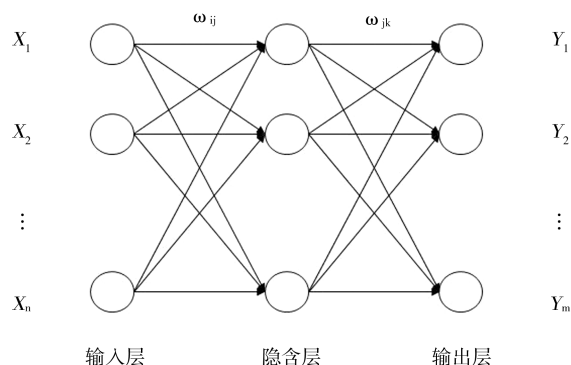


图1 BP神经网络结构图

构建BP神经网络一般有如下几个步骤:

1)对输入数据集 $\{X_i\}$ ($1 < i < n$)和目标数据集 $\{Y_j\}$ ($i < j < m$)进行标准化处理,消除量纲的干扰;

2)确定隐含层神经元数*N*,按照经验 $N = \sqrt{n+m} + p$,其中*p*介于1~10之间;

3)构建神经网络,设定各层传递函数、最大允许步长、模型精度等;

4)当模型满足设定精度时,保存模型,进行水质评价。

3 计算与分析

3.1 主成分分析

六合区位于南京市北部,长江左岸,是国家重要的现代工业基地,滁河由西到东贯穿境内。本文选取2019年南京市六合区23个监测断面年平均水质数据进行实例分析。由于监测数据中金属离子及石油类、挥发酚等按《地表水质量评价标准》(GB3838—2002)评价均为I类,故在主成分分析中不对其进行分析。筛选后主要的评价因子有pH(X_1)、DO(X_2)、 $\text{NH}_3\text{-N}$ (X_3)、 COD_{Mn} (X_4)、 COD_5 (X_5)、F(X_6)、COD(X_7)、TP(X_8)、TN(X_9)共9类。

(1)对数据进行均值化处理,计算相关系数矩阵,由表1可知大部分监测指标相关系数均大于0.3,说明各指标之间存在信息的重叠,因此适用主成分分析对原始数据进行降维处理。

(2)计算特征值和主成分贡献率,得到每个主成分所对应的解释方差、特征值和累计方差贡献率,由表2可知前3个主成分累计反映了原始参数信息的88.469%,可将原来11个影响指标减少为3个,进而大大降低因子的维数。

由主成分荷载矩阵可以看出, F_1 中DO、

表 1 相关系数矩阵

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1.000	-0.059	0.043	0.241	0.187	0.214	0.159	0.080	-0.048
X2	-0.059	1.000	-0.462	-0.646	-0.684	-0.589	-0.670	-0.459	-0.402
X3	0.043	-0.462	1.000	0.528	0.582	0.588	0.481	0.900	0.864
X4	0.241	-0.646	0.528	1.000	0.953	0.849	0.888	0.476	0.324
X5	0.187	-0.684	0.582	0.953	1.000	0.819	0.825	0.542	0.445
X6	0.214	-0.589	0.588	0.849	0.819	1.000	0.925	0.452	0.487
X7	0.159	-0.670	0.481	0.888	0.825	0.925	1.000	0.369	0.313
X8	0.080	-0.459	0.900	0.476	0.542	0.452	0.369	1.000	0.831
X9	-0.048	-0.402	0.864	0.324	0.445	0.487	0.313	0.831	1.000

表 2 特征值及累计方差贡献率

主成分	特征值	初始特征值方差 百分比/%	累计贡献率/ %
1	5.403	60.030	60.030
2	1.617	17.964	77.994
3	0.943	10.475	88.469
4	0.469	5.212	93.681
5	0.289	3.214	96.894
6	0.149	1.656	98.551
7	0.074	0.823	99.374
8	0.035	0.386	99.759
9	0.022	0.241	100.000

表 3 主成分荷载矩阵

X	F1	F2	F3
X1	0.185	-0.367	0.902
X2	-0.748	0.135	0.225
X3	0.805	0.522	0.107
X4	0.889	-0.369	-0.039
X5	0.909	-0.246	-0.063
X6	0.889	-0.259	-0.036
X7	0.856	-0.405	-0.158
X8	0.746	0.572	0.176
X9	0.682	0.661	0.063

COD_{Mn}、COD₅、F、COD 对其影响程度较大,可认为 F1 在一定程度上表示了水体的有机污染;F2 中 NH₃-N、TP、TN 对其影响程度较大,可认为 F2 在一定程度上表示了水体的无机污染;F3 中 pH 对其影响程度较大,故可认为 F3 在一定程度上表示了水体的酸碱性。

(3)根据主成分荷载矩阵计算各主成分的对应指标的得分系数,由此计算 F1、F2、F3 的值并根据公式(4)计算综合得分 F,主成分得分越大说明水质越差,如表 4 所示。其中按照《地表水环境质量标准》各类水质的标准值计算得到 I 类水质主成分

综合得分的 -2.891、II 类为 -2.008、III 类为 -0.706、IV 类为 1.788、V 类 3.817。

$$F = 0.108ZX_1 - 0.178ZX_2 + 0.360ZX_3 + 0.211ZX_4 + 0.229ZX_5 + 0.224ZX_6 + 0.184ZX_7 + 0.426ZX_8 + 0.486ZX_9 \quad (4)$$

由表 4 可知仅有 3 个断面的评价结果与单因子评价法保持了相同,这是因为这些断面中大多数水质指标都处单因子评价法的评价等级。其余 20 个断面评价结果提升了 1~3 个等级,这是因为这些断面水质指标仅有少数处于单因子评价法的评价等级。主成分分析法的综合考虑了所有评价指标,避

表 4 主成分分析评价结果

监测断面	F1	F2	F3	F	排序	单因子法	PCA
沙洲渡口	-3.833	0.830	1.240	-2.343	3	V	II
远古水业	-4.438	2.053	-3.146	-3.031	1	V	I
南厂码头	-4.017	0.964	1.753	-2.393	2	V	II
扬子 8 号码头	-3.788	1.091	1.344	-2.261	4	V	II
滁河宁连公路大桥	0.879	-0.321	0.106	0.559	15	劣 V	IV
龙津桥	2.541	0.797	-0.577	1.787	20	劣 V	IV
六合大桥	1.286	0.697	-0.633	0.914	18	劣 V	IV
六合铁路桥	2.122	1.366	0.302	1.683	19	劣 V	IV
红山窑闸	-0.128	0.122	-0.583	-0.130	12	劣 V	IV
陈摆江渡口	-1.233	-0.003	0.120	-0.825	7	劣 V	III
独山渡口	0.122	-1.654	0.133	-0.158	10	IV	IV
竹镇南桥	-0.288	-1.923	-0.471	-0.542	8	IV	IV
同心桥	-1.361	-1.904	0.175	-1.199	5	IV	III
方州桥	3.126	1.960	1.259	2.558	23	劣 V	V
八百河桥	0.037	-1.871	-0.227	-0.287	9	IV	IV
东沟大桥	1.419	-0.606	0.367	0.909	17	劣 V	IV
马汊河大桥	0.726	0.565	-0.230	0.555	14	劣 V	IV
岳子河闸	0.248	-0.777	-0.609	-0.016	13	V	IV
三汊湾	0.138	-1.485	-0.087	-0.144	11	V	IV
友谊桥	1.296	-0.939	0.308	0.768	16	V	IV
灵钢河桥	3.464	1.661	-0.524	2.553	22	劣 V	V
安桥	3.034	0.282	0.433	2.149	21	劣 V	V
划子口闸	-1.352	-0.903	-0.452	-1.105	6	V	III

免了某一评价因子将其他因子的信息完全覆盖,让一些处于“劣势”的指标得到了反映,评价结果与单因子评价法在整体上的趋势是相同的,因此具有一定的合理性。但是可能存在过于乐观的评价结果,比如远古水业断面,除 DO、COD₅为 I 类外,其余各指标均在 III 类和 IV 类之间,因此主成分分析法将其评价为 I 类过于乐观。

(5)通过主成分分析对原本的 9 个评价指标进行降维,得出的 3 个主成分能够反映原指标 88.469% 的信息,大大简化了信息处理的维度;由主成分的综合表达式可以看出 NH₃-N、TP、TN 相比其他指标在权重方面占有绝对的优势,因此认为 NH₃-N、TP、TN 是六合区的主要污染物,这与南京市水资源公报里的分析结果同样是吻合的,因此后续

搭建神经网络模型,以此 3 项指标作为模型的输入数据。

3.2 构建 BP 神经网络模型

(1) 生成样本数据

样本的数量和差异性对神经网络的模拟精度有重要影响,为获得足够多的样本,将各项指标测定国标法中最低检出浓度作为 I 类下限值,结合《地表水环境质量标准》(GB3838—2002)可以得到各级水质的上下限值,在各类水质等级之间进行随机插值(比如当 N、TP、TN 分别处于(0.15,0.5]、(0.02,0.1]、(0.2,0.5]之间时,该水体水质肯定属于 II 类),考虑到研究区域水质多为劣 V 类,因此共设置六个水质等级(I~劣 V 类),每两级之间随机插值生成 450 个样本,共 2 700 个样本。

表 5 地表水环境质量标准各项指标限值 单位:mg/L

指标	I 类下限	I 类	II 类	III 类	IV 类	V 类	劣 V 类
NH ₃ -N	0.01	0.15	0.5	1	1.5	2.0	>2.0
TP	0.01	0.02	0.1	0.2	0.3	0.4	>0.4
TN	0.05	0.2	0.5	1	1.5	2	>2.0

(2) 样本数据预处理

在 MATLAB 中将样本矩阵 P 的每一个元素归一化到 $[-1,1]$,样本集作为输入样本时是一个 $5 \times 2\,700$ 的矩阵,其中每一列代表一个样本,共 2 700 个样本。

(3) 确定目标矩阵

输出层共有 6 种水质类别,因此输出层选用 6 个神经元。用 6×1 的矩阵表示每个输出类别,其中 $(1,0,0,0,0,0)^T$ 表示 I 类水质、 $(0,1,0,0,0,0)^T$ 表示 II 类水质、 $(0,0,1,0,0,0)^T$ 表示 III 类水质、 $(0,0,0,1,0,0)^T$ 表示 IV 类水质、 $(0,0,0,0,1,0)^T$ 表示 V 类水质、 $(0,0,0,0,0,1)^T$ 表示劣 V 类水质。每一个输入样本对应一个输出矩阵,因此目标集 T 为一个 $3 \times 2\,700$ 的矩阵。

(4) 创建神经网络

在 MATLAB 中输入样本集 $[P,T]$,将样本中的 75% 用于训练网络,10% 用于验证,15% 用于测试;隐含层神经元个数按照经验公式取值在 $[4,13]$ 之间,先选取 4 个神经元进行训练,然后依次增加神经元的个数直到 15,依据神经元个数和均方误差(图 2)及神经元个数和训练步长(图 3)的关系,确定隐含层神经元个数为 11。

此时模型运行了 76 步,在第 70 步时达到了最佳表现,如图 4 所示。通过输入层、隐含层、输出层神经元个数的确定,最终的神经网络采用 3-11-6 的 3 层网络结构。

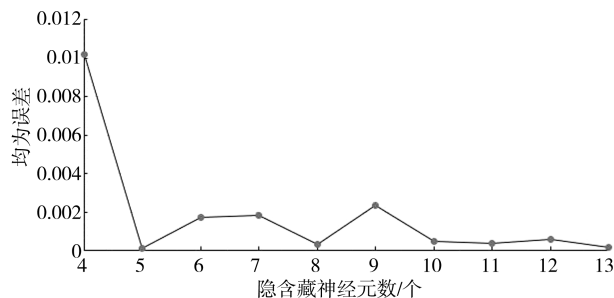


图 2 隐含层神经元数与均方误差的关系

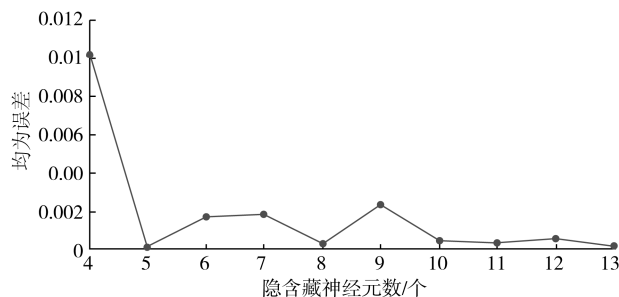


图 3 隐含层神经元数与步长的关系

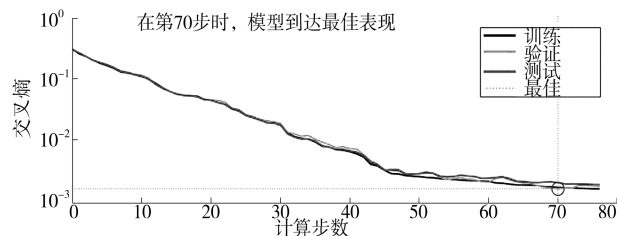


图 4 $n=11$ 时训练表现

(5) 进行水质评价

由表 6 可以看出,PCA-BP 神经网络法的评价结果与单因子评价法和主成分分析法的评价结果在整体趋势上同样是一致的。相比单因子评价法,龙津桥、六合大桥、方州桥、马汊河大桥、灵钢河桥、安桥 6 个断面神经网络的评价结果与其完全一致。远古水业、南厂码头、扬子 8 号码头等断面评价结果上升了 2 个等级;滁河宁连公路大桥、六合铁路桥、友谊桥等断面 1 个等级。劣 V 类断面占比由 47.8% 下降为 26.1%;V 类断面占比由 34.8% 下降为 13.0%;IV 类断面减少 100%;新增 III 类及 III 类以上断面 7 个。与主成分分析法相比,陈摆江渡口、岳子河闸、划子口闸断面与其评价结果完全一致,对于远古水业断面,PCA-BP 神经网络法对上文的主成分分析法的结果进行了一定程度上的“纠正”,避免了评价结果的过分乐观。

表 6 神经网络评价结果

监测断面	I 类	II 类	III 类	IV 类	V 类	劣 V 类	PCA – BP	PCA	单因子
沙洲渡口	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	III	II	V
远古水业	0.0000	0.0002	0.9998	0.0000	0.0000	0.0000	III	I	V
南厂码头	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	III	II	V
扬子 8 号码头	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	III	II	V
滁河宁连公路大桥	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	V	IV	劣 V
龙津桥	0.0000	0.0000	0.0000	0.0000	0.0009	0.9991	劣 V	IV	劣 V
六合大桥	0.0000	0.0000	0.0000	0.0000	0.0381	0.9619	劣 V	IV	劣 V
六合铁路桥	0.0000	0.0000	0.0000	0.0001	0.9999	0.0000	V	IV	劣 V
红山窑闸	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	IV	IV	劣 V
陈摆江渡口	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	III	III	劣 V
独山渡口	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	II	IV	IV
竹镇南桥	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	II	IV	IV
同心桥	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	I	III	IV
方州桥	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	劣 V	V	劣 V
八百河桥	0.0000	0.9837	0.0163	0.0000	0.0000	0.0000	II	IV	IV
东沟大桥	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	V	IV	劣 V
马汊河大桥	0.0000	0.0000	0.0000	0.0000	0.3715	0.6285	劣 V	IV	劣 V
岳子河闸	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	IV	IV	V
三汊湾	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	III	IV	V
友谊桥	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	IV	IV	V
灵钢河桥	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	劣 V	V	劣 V
安桥	0.0000	0.0000	0.0000	0.0000	0.0006	0.9994	劣 V	V	劣 V
划子口闸	0.0000	0.0003	0.9997	0.0000	0.0000	0.0000	III	III	V

4 结 论

本文针对六合区 23 个断面水质监测数据, 首先利用改进的主成分分析法对各断面水质进行评价, 并由此确定主要污染物, 实现评价因子降维的目的。然后结合 BP 神经网络, 构建 PCA – BP 神经网络评价模型, 同时利用主成分分析法的评价结果

对 PCA – BP 神经网络评价模型的评价结果从侧面进行验证, 就评价效果而言, PCA – BP 神经网络评价模型既避免了单因子评价法中某一评价因子将其他因子的信息完全覆盖的弊端, 同时也避免了主成分分析法评价结果的过分乐观, 因此 PCA – BP 神经网络评价模型评价结果更为客观真实。总体来说, PCA – BP 神经网络评价模型在解决了水质评价

中评价指标权重难以合理确定,评价模型过于复杂、评价结果不合理等问题的基础上,为六合区的水资源保护与治理工作提供了参考。

参考文献:

- [1] 李博川. 不同水质评价方法在河流水质评价中的应用比较[J]. 区域治理, 2019(28):69-71.
- [2] 王平, 王云峰. 综合权重的灰色关联分析法在河流水质评价中的应用[J]. 水资源保护, 2013, 29(5):52-54,64.
- [3] 蒋宝林. 模糊数学在句容河水质评价中的应用[J]. 黑龙江环境通报, 2019, 43(3):60-63.
- [4] 闫荣荣. 基于 AHP 的地表水环境评价分析[J]. 太原师范学院学报(自然科学版), 2019, 18(2):89-91.
- [5] 林卉, 李楠, 黄伯当, 等. 基于主成分分析的南流江

水质评价[J]. 广东化工, 2020, 47(4):144-146, 148.

- [6] 周及, 关卫省, 付林涛. 基于多元统计的西安市河流水质评价及污染源解析[J]. 水资源保护, 2020, 36(2):79-84.
- [7] 曹阳阳. 基于 RBF 神经网络的燕山南麓水库群水质评价[J]. 水资源开发与管理, 2019(2):38-41.
- [8] 周星宇, 黄晓荣, 赵洪彬. 基于主成分分析法的河流水文改变指标优选[J]. 人民长江, 2020, 51(6):101-106.
- [9] 舒服华. 基于 BP 神经网络预测我国进口石材值[J]. 石材, 2019(12):33-36, 62.
- [10] 张轩, 张行南, 江唯佳, 等. 秦淮河流域东山站水位预报研究[J]. 水资源保护, 2020, 36(2):41-46.

(上接第 3 页)

和自媒体等新技术、新形式的运用,满足社会、公众对生产、生活中相关水利知识的迫切需求;完善科普工作体制机制,增强各级主管部门之间的联系,通过争取财政及社会各界资金的支持,探究水利科普激励方法,加强对水利科普场馆建设、运行管理的督促和协助,努力拓宽水利科普经费来源,引导和鼓励相关企业、社会团体等投入水利科普事业。

7 加强对外合作,探寻多元渠道

完善水利外事管理制度,支持科研人员更广泛深入地参与国际科技交流与合作。积极拓宽对外交流合作渠道,强化与国际水利科技创新管理机构、高校和研究机构的交流与合作。联合国际合作伙伴单位开展研究,分享技术进步成果,进一步提升对外合作的能力和水平。支持和鼓励具有自主知识产权的水利技术、产品和标准走出去,增强国际竞争力,全面提升中国水利在国际舞台的影响力和话语权。

8 结 语

科技创新永无止境,江苏作为发展先行地区和

水利大省,加快制定省“十四五”水利科技创新发展规划,可为社会各界及全体水利工作人员提供有力指导,保障经济社会可持续发展,在全国范围内树立水利科技创新模范起到推动作用。

参考文献:

- [1] 两院院士大会中国科协第十次全国代表大会在京召开[J]. 党建, 2021(6):4-7.
- [2] 董明锐. 科技创新支撑现代水利[N]. 中国水利报, 2018-01-18.
- [3] 周惠娟. 水利科技项目创新成果与管理[J]. 水科学与工程技术, 2019(2):90-92.
- [4] 王琳琳. 三“跑”并存:跟跑并跑领跑[J]. 科技中国, 2015(2):10-11.
- [5] 江芳. 多方发力巧解水利建设资金之“渴”[J]. 中国水利, 2015(24):218-221.
- [6] 尹克剑. 工程质量检测单位建立质量、环境和职业健康安全管理体系的思考[J]. 水利发展研究, 2018, 18(6):63-65.