编号：A0022

基于 CNN 的赣江水质时空规律分析与预测

论文题目：基于 CNN 的赣江水质时空规律分析与预测

参赛学校：江西财经大学

参赛成员（作者）：李霖、王琨、刘强

指导老师：刘庆　谭祥勇

目录

表格清单

插图清单

摘要

水是人类社会赖以生存和发展的资源，水质污染不仅会直接影响到人类生活用水的安全，同时也对社会的可持续发展产生了严重的影响。科学准确的水质预测模型能够为环境保护部门提供预警信息，及时控制有污染风险的流域，有助于我们在水环境保护中占据主动地位。

在以往水质预测研究中,典型时间序列模型已得到了广泛的应用，但大多数研究都具有将时空规律分开进行研究、影响因素考虑片面化、预测精度较低等问题。因此本文提出了由 CNN 卷积神经网络驱动的领域预测模型，它能够同时考虑水质的时间和空间变化规律，有效捕捉水质数据中复杂的非线性结构，提高水质预测精度。

本文通过收集 2018 年至 2021 年赣江流域的 31 个地表水质监测站点的实时水质数据，建立了基于 CNN 的水质预测模型，并用其拟合赣江流域水质的时空变化规律，预测未来赣江流域水质情况。过程及结果如下：

首先利用 Python 进行数据预处理，由于各指标之间存在量纲差异，再根据《地表水环境质量标准》可知水质类别与各指标之间存在复杂的非线性关系，因此我们将五项水质指标——pH、CODMn、溶解氧、氨氮、总磷，利用 PCA 得出综合指标。并且，根据时间序列的图形化观察与参数检验，验证得出综合指标与水质类别的变化规律有相似性，这为我们的预测变量的选择提供了理论基础。

之后，我们采用 ARIMA 模型对水质类别进行预测，预测结果仅能解释原变量的 56.6%，拟合效果一般；而仅考虑空间规律搭建的一维卷积神经网络，所得的拟合图像显示其拟合效果不佳。

因此，我们搭建了二维卷积神经网络的预测模型，同时考虑水质的时空变化规律，将各站点数据按照上中下游的空间顺序排列后，选取连续六天内相邻的 13 个站点的水质综合指标作为观测样本，第七个站点第七天的水质类别作为预测值，形成一个 13×6 的子矩阵，利用 Python 批量转化为灰度图像后，将其输入到 CNN 模型中按照水质类别进行图像分类，最终实现对水质的预测。

经过实证分析，本文所建立的二维 CNN 水质预测模型，在水质预测的多分类问题上能够达到 83.3%的预测精度，相较于 ARIMA 模型和一维卷积神经网络而言，效果更好，适用于复杂水域水质的时空规律预测，能够在实际应用中帮助监测人员预知流域的污染风险，为临近站点发送预警信号；对于实际值与预测值相差较大的流域，能够让工作人员有针对性地进行人工排查，提高了线下的管控效率和资源配置效率。

关键词：水质预测、CNN 卷积神经网络、灰度图识别、PCA

一、绪论

（一）研究背景及意义

地表水是指存在于地壳表面的水，是河流、沼泽、冰川、湖泊四种水体的总称。它是人类生活、工业农业用水的重要来源之一，因此地表水环境质量的好坏与人们的生活质量高低和社会发展快慢密切相关。

近年来，随着我国经济的快速发展，城市数量和规模的迅速扩大，社会对水的需求不断增大，因此地表水环境污染的问题倍受关注。根据中商产业研究院整理的数据显示，2018 年中国城市污水排放量约为 521.12 亿立方米，比 2017 年增加了 5.83%。根据 2020 年生态环境部通报的全国地表水质数据显示，全国主要江河地表水质类别为Ⅳ类~劣Ⅴ类的断面比例占 16.6%，主要污染指标为化学需氧量、氨氮和 CODMn。因此加强水质管理，实现水资源与社会可持续发展是一项不可忽视的艰巨任务。

为贯彻《环境保护法》和《水污染防治法》，我国以《地表水环境质量标准》作为国家水环境质量标准，加强地表水质管理。该标准通过设定各基本指标的标准限值，根据各指标从属类别的占比进行水质分类评价。而在水质分类过程中容易带有主观评价，且各指标与水质类别之间不存在明显的线性关系。因此需要通过数学方法分析，基于各单项监测指标建立一个客观的水质综合评价指标，解决水质分类在建模中的实用性问题。

现如今，我国的环境监测事业不断发展，并已实现了从传统监测到数据化管控的历史转型，各种监测数据呈指数增长。截至 2021 年，国家地表水环境质量监测网共设置了 3641 个地表水考核断面。基于当下有利的数据环境，实现科学准确的水质时空预测，在实际应用中能够帮助监测人员预知流域的污染风险，为临近站点发送预警信号；对于预测值与实际值差异明显的站点，能够让工作人员有针对性地进行人工排查，提高了线下的管控效率和资源配置效率。

在此背景下，本文基于 2018 年至 2021 年江西省位于赣江流域的 31 个地表水质监测站点提供的五项实时监测指标——pH、CODMn、溶解氧、氨氮、总磷，利用主成分分析方法得出一个水质综合评价指标，并以此了建立 CNN 卷积神经网络水质预测模型，拟合赣江流域水质的时空变化规律，预测未来水质的变化和发展趋势，为促进地表水环境保护和监督助力。

（二）研究现状

近年来，生态环境部门以及众多学者基于现阶段公开数据对水质预测模型进行了大量的研究，但大多模型都面临将时空规律分开进行研究、影响因素考虑片

面化、预测精度较低等问题。经查阅相关研究资料后，我们发现如今的水质预测研究模型主要分为机理性水质预测模型和非机理性水质预测模型。后者主要为黑箱模型法,通常其输入变量为水质历史监测数据,而输出变量为待预测水质指标,从而建模预测。该方法因不关注水环境的内在机理，采用的数学方法较为成熟,预测的效果好，被广泛运用于水质的预测。常用的非机理性水质预测模型有时间序列预测模型、灰色系统预测模型、神经网络预测模型等。

1. 时间序列预测模型

时间序列是指将某一统计指标的观测值按照时间顺序排列而成的数列。水质指标随时间变化所得的数据可以看作时间序列,根据大量的历史监测数据建立时间序列模型，从而实现水质的预测。

唐宗鑫等[7]2002 年将自回归模型引入环境水文地质学领域,研究闽江水质 pH 值与浑浊度,能较好的运用于水质变化的中长期预测。吴涛等[8]2006 年以三峡库区水质指标的水期数据为研究对象,利用 Holt-Winters 乘法预测模型进行水质预测。周志青等[14]2017 年采用建立了 ARIMA 和 RBF-NN 的组合模型，分别考虑水质指标受内、外生变量的影响，对 TN、TP 和 CODMn 三项指标进行预测。罗学科等[5]2020 年提出 ARIMA-SVR 的组合方法，其在水质预测中的应用结果显示该方法具有较高的预测精度，效果优于单一模型。时间序列模型的具体分类见表 1：

表 1 时间序列模型分类表

分类	时间序列趋势	时间序列周期	平稳时间序列	其他时间序列
具体分类	滑动平均	小波分析	ARMA	季节-周期组合
	指数平滑	时间序列周期方差	ARIMA	多变量时间序列
	线性回归	季节性水平	ARFIMA	CAR
	多次滑动平均	季节性交替趋势		门限自回归
	多次指数平滑	季节叠加趋势		均值生成函数预测

2. 灰色系统预测模型

灰色系统理论主要是通过在少数据、贫信息的生成与开发中提取出有价值的信息，从而进一步实现对系统的运行行为和演化规律的描述和监控。自 1982 年以来，该理论在农业、工业、气象等领域得到了成功地应用。

李如忠等[4]2002 年依据灰色系统理论,构造了一个由 6 个 GM(1,1)模型组成的灰色动态模型群,并运用该模型群对淮河干流枯水期氨氮浓度变化趋势进行了预测分析。于慧等[12]2014 年将模糊集合理论与马尔科夫链理论引入灰色 GM(1,1)

预测模型,并应用其对海河三岔口断面的 3 项指标——DO、CODMn 和 NH3-N, 2012-2016 年的浓度变化趋势进行预测，结果显示其预测精度较高。

3. 神经网络预测模型

人工神经网络本质上是模仿人脑神经系统, 通过动力学行为和网络变换进行分布式并行信息处理的算法模型。现如今神经网络已经成为了水质预测中的主流模型。

莫慧芳等[6]2004 年建立了东江水质预测的 BP 神经网络模型,并给出了仿真结果，验证了 BP 神经网络模型可以很好地对水质进行预测。刘东君等[3]2011 年基于灰色预测和神经网络的组合模型，对北京密云水库水中的 DO 值进行预测,并与单纯灰色和单纯神经网络模型比较, 结果表明组合模型的预测值相对误差更小, 精度更高。王晓峰等[9]2020 年基于支持向量分类与 GRU 神经网络联合的水质预测方法，得到的联合预测模型的预测精度更准确,模型效果更好。

二、模型构建思路与创新

（一）CNN 模型构建思路

本文主要基于地表水质的五项监测指标数据对赣江地表水水质的时空变化规律进行分析与预测，模型构建流程图如图 1 所示：

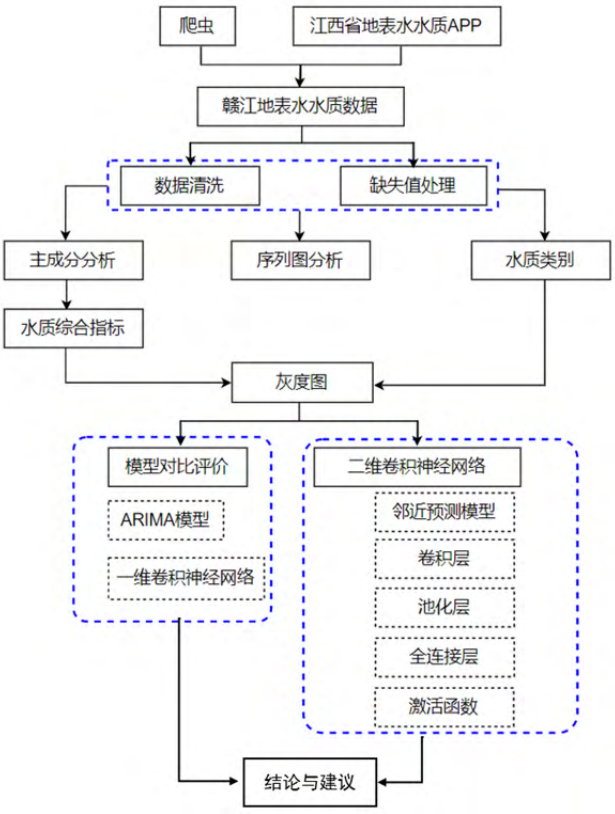

图 1 CNN 模型构建流程图

（二）水质的分析与预测模型创新

CNN 卷积神经网络能够高效地从图像数据中提取特征信息并进行分类，具有极强的图像识别能力，主要应用在自然语言处理、图像处理与分类、视频预测等领域。

1. 在建立预测模型之前，将各站点的五项水质指标（pH、溶解氧、CODMn、氨氮、总磷）通过主成分分析得出一项水质综合指标，并应用该指标预测各站点未来的水质类别，以消除各指标量纲差异的影响，并解决水质指标与水质类别间存在非线性的问题。
2. 将各站点数据按照上中下游的空间顺序排列后的时间序列数据集，切分成若干个邻域水质矩阵，再利用 Python 将各矩阵转化为二维灰度图像输入到模型中，根据预测的水质类别进行图像分类。
3. 在非机理性水质预测模型中，首次引入二维卷积神经网络 CNN 预测模型，同时考虑水质的时间和空间变化规律并对水质类别进行预测。将水质预测建模问题转化成一个计算机视觉任务，充分利用了 CNN 优良的特征提取功能以及其在时间序列预测中的噪声容忍度极佳的优点。

三、数据描述及数据预处理

（一）研究区域概况

作为长江的主要支流之一，赣江是江西省水运的大动脉、最大的河流，亦是远景规划中赣粤运河的组成部分。此外，赣江于 2003 年覆盖的江西总人口数最多，约达 1931.2 万人。因此，对赣江水质时空分布规律的分析了解与预测在经济和社会发展中具有重要的意义。

赣江流域虽然森林覆盖率达 63.6%，水质达标率为 78.7%，总体情况良好；但在生态环境方面其仍然存在不可忽视的问题，主要表现在：森林的质量与结构仍存在问题，防护功能差；水土流失仍较严重；其中赣州段的水质污染较重，一般为Ⅳ类水质，其中下游区支流的污染最为严重，多次有 V 类或劣 V 类水质的出现。

为了更好地识别赣江流域的水质时空分布规律，本文将赣江流域内选择的 31 个水质监测站点按照上中下游的空间顺序进行排列。其中上游地区站点有：大余城郊、龙山口、梓坑、梅江江口、峡山、上犹江江口、新庙前、市自来水厂；中游地区站点有：遂川江江口、通津、蜀水河口、孤江江口、禾水河口、乌江江口、金滩；下游地区站点有：良田村、高安市青州村、棚下（杨村）、湖心岛、浮桥、孔目江江口、罗坊、大洋洲、肖江江口、丰城小港口、周坊、生米、滁槎、

大港、昌邑、吴城赣江。各观测站点的空间分布如图 2 所示：

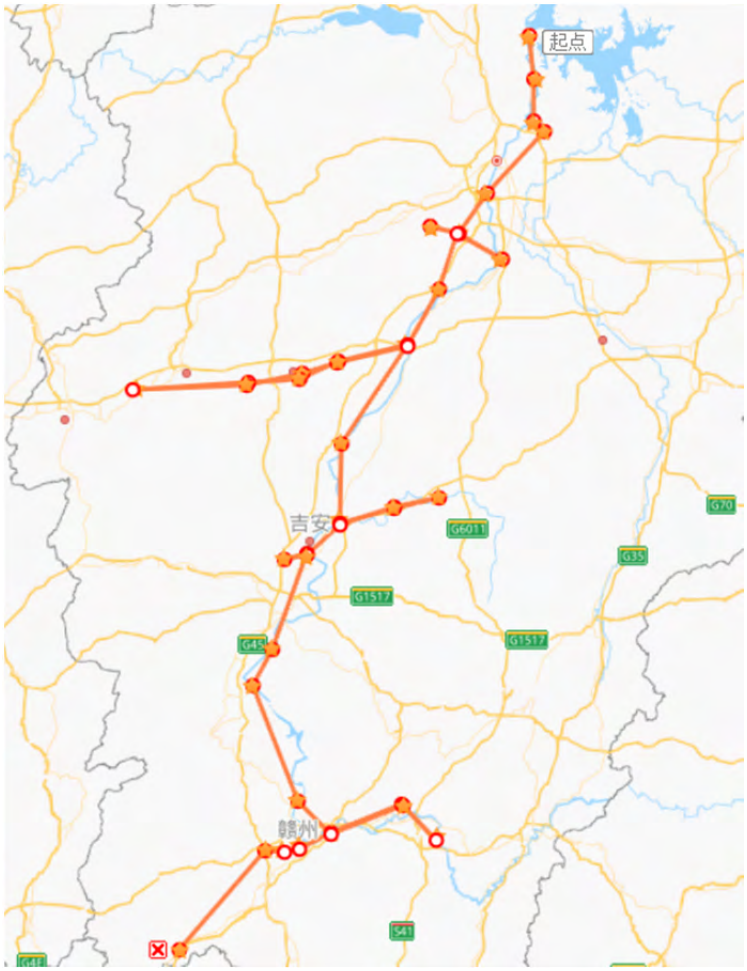

图 2 观测站点分布图

（二）变量描述

地表水环境监测数据指标将选取 pH、CODMn、溶解氧、氨氮以及总磷这 5 项指标对赣江地表水环境质量进行相关研究。赣江各流域的水质目标级别属Ⅲ类，该类别主要适用于生活饮用水的地表水源地二级保护区、渔业水域及游泳区。而数据中显示的水质类别为实时监测的水质实际所处类别。具体描述以及有效值范围的如表 2：

表 2 水质指标描述与有效值范围

水质指标	有效值	释义
pH	0~14	反映水质的酸碱程度，水质过酸或过碱对生态环境以及人体健康都会造成不良影响。
CODMn	0~22.5mg/L	高锰酸盐指数，通过高锰酸钾的氧化反应测定水中有机物和无机物的量，是反映地表水环境污染程度的一项重要指标。
溶解氧	0~12mg/L	溶解在水中的氧，用于衡量水体自净能力的一项指标。溶解氧被消耗后，其恢复到初始状态所需的时间越长，则说明水体的自净能力越弱，即污染严重。
氨氮	0~3mg/L	以游离氨（NH3）和铵离子（NH4+）形式在水中存在的氮，可导致水体富营养化，是水体中的主要耗氧污染物。
总磷	0~0.6mg/L	水中各种形态磷的总和，过量的磷主要会导致水体污秽异臭，使湖泊富营养化和海湾出现赤潮。
水质类别	Ⅰ~劣Ⅴ(Ⅵ)类	各站点当天地表水环境质量所处类别，分为 6 个等级

针对以上五项指标的地表水水质国家标准如表 3 所示：

表 3 地表水水质国家标准表

水质指标		Ⅰ类	Ⅱ类	Ⅲ类	Ⅳ类	Ⅴ类
pH（无量纲）				6-9		
CODMn（mg/L）	≤	2	4	6	10	15
溶解氧（mg/L）	≥	7.5	6	5	3	2
氨氮（mg/L）	≤	0.15	0.5	1.0	1.5	2.0
总磷（mg/L）	≤	0.02	0.1	0.2	0.3	0.4

考虑到水质监测系统在中间环节中可能会出现问题, 导致监测数据库中包含了一些异常值。并且将江西省地表水水质 APP 提供的水质类别数据中划分的 6 个等级与地表水水质国家标准划分的 5 个等级相结合考虑, 为不同的指标设定了不同的有效值范围（除 pH 值外，将Ⅴ类水质标准上浮 50%的区间定为Ⅵ类），超过该范围值的按照异常值排除，后续视同缺失值处理。

（三）数据来源

数据文件夹中的数据为从由江西省生态环境厅提供的江西省地表水水质 APP 上利用 Fiddler 与 Python 爬虫所得数据。数据包中 CNNP1821.xlsx 为主成分分析后所得的综合指标数据集, CNNQ1821.xlsx 为 APP 中实际水质类别数据集。

安装 APP 的二维码以及爬虫代码见数据包。

（四）数据预处理

1.数据的获取与整理

为了使水质评价的预测更具现实意义，在爬取数据时选择以天为单位进行爬取，共爬取了 2018 年 7 月 2 日——2021 年 4 月 29 日各数据指标。利用 Python 和 Excel 表格操作对数据作预处理，删除无用列以及无用站点数据，并将赣江流域各站点整合在一起。

2. 缺失值处理

利用 Python 将各站点数据按照完整的时序排列后，结合 R 语言对数据集进行多重插补法，以下结果展示均以上犹江江口为例，其他站点与该站点结果相似：

① 对缺失值的分布进行可视化处理，结果如图 3 所示：

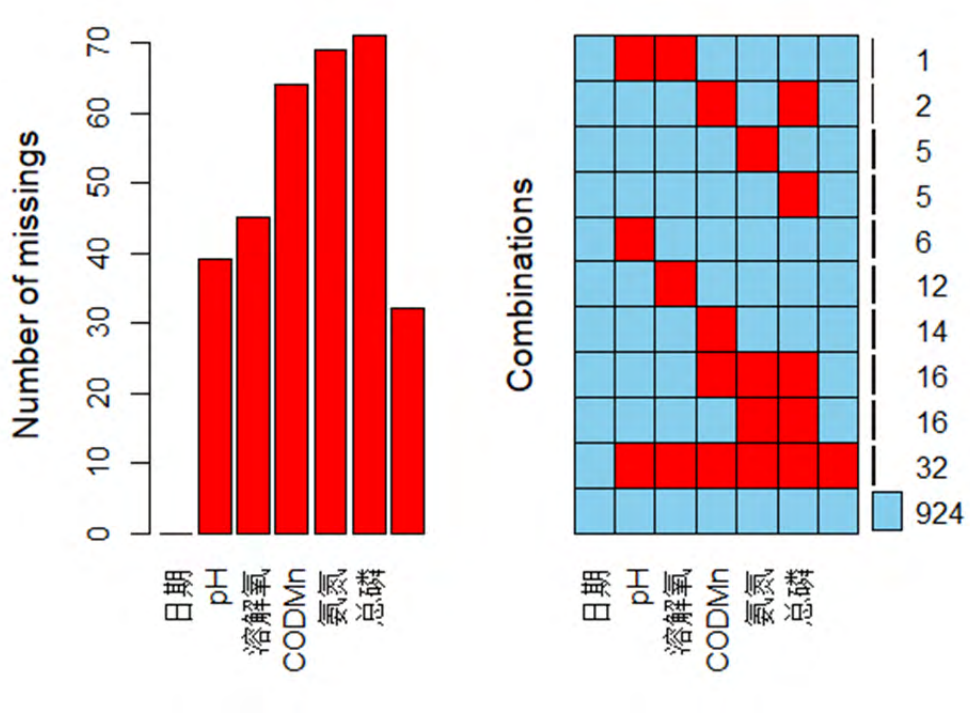

图 3 数据缺失值的分布图

② 利用相关性分析确定缺失数据的类型，结果如表 4 所示：

表 4 缺失数据的相关性矩阵

	pH	溶解氧	CODMn	氨氮	总磷	水质类别
pH	1	0.779	0.623	0.598	0.589	0.903
溶解氧	0.779	1	0.575	0.551	0.542	0.838
CODMn	0.623	0.575	1	0.703	0.724	0.696
氨氮	0.598	0.551	0.703	1	0.908	0.668
总磷	0.589	0.542	0.724	0.908	1	0.658
水质类别	0.903	0.838	0.696	0.668	0.658	1

由表 4 可知一起缺失的相对可能性较小的指标是溶解氧和总磷（$r=0.542$），

而其余的指标缺失可能性相对较高，r 在 0.55~0.90 之间浮动，因此可以确定缺失数据的类型为 MAR 型，即为随机缺失，适用 MICE 包进行多重插补。

③ 本文以线性回归模型作为多重插补模型，利用 MICE 包中随机森林的方法对各站点检测指标中的缺失值进行非参数插补，生成了 5 个插补数据集。

④ 利用 Pool 函数对 5 个回归模型汇总进行 F 检验，检验整个插补方法是否合格，结果如表 5 所示：

表 5 插补后的 F 检验结果

术语	估计值	标准误	统计值	自由度	P 值
(截距项)	0.008	0.008	0.003	2.765	0.010
CODMn	0.010	0.002	6.549	107.112	0.000
氨氮	0.171	0.015	11.130	8.232	0.000

可见各项回归系数在 $p<0.01$ 的水平上很显著，因此可认为插补后各检验数据集合格，并且选择其中一组数据集作为本文的研究数据集。

3.插补后监测数据的基本信息

为了检查插补后数据集的大致分布，利用 SPSS 中的描述统计对各监测指标数据进行分析：

表 6 监测数据的基本信息

	N	最小值	最大值	均值	标准偏差
pH	1033	4.00	9.00	7.04	0.63
溶解氧	1033	4.60	12.00	8.35	1.39
CODMn	1033	0.20	5.60	1.48	0.70
氨氮	1033	0.02	1.18	0.10	0.12
总磷	1033	0.005	0.46	0.04	0.04
水质类别	1033	1.00	6.00	1.97	0.63
有效个案数(成列)	1033				

由表 6 可知，处理后的数据集没有缺失数据以及异常值出现，从而开始进一步的分析。

4.各指标时间序列的图形化观察

根据各指标与水质分类的序列图进行趋势观察分析，从整体上了解各指标与水质类别的变化趋势以及相关性。

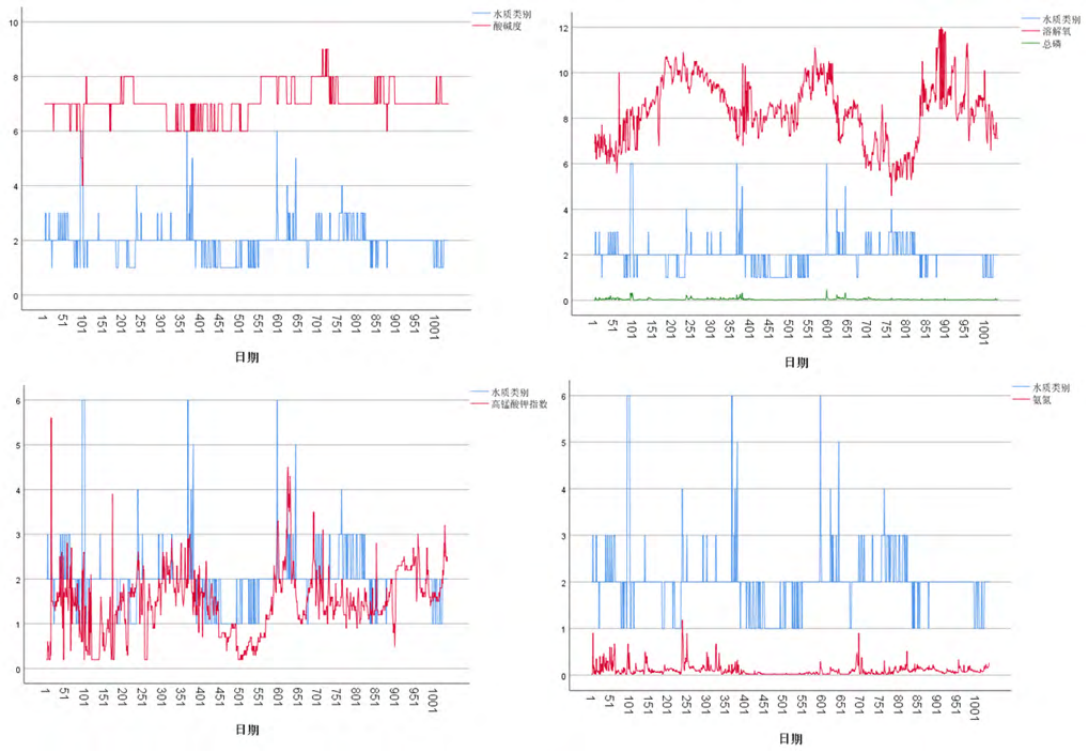

图 4 各指标与水质类别的序列图

由图 4 可知，尽管各指标与水质类别的变化趋势存在一定的相似性，但由于各指标量纲的不同并且水质类别为多分类数据，为后续的水质预测分析加大了难度，因此本文选择对各指标进行主成分分析以得出综合指标为后续水质预测准备。

四、五项水质指标的主成分分析

主成分分析法（PCA）是一种利用正交变换，用一组线性无关且较少的变量替代一组可能存在相关性的变量的统计方法；线性无关的变量称为主成分。

（一）主成分分析法的基本原理

统计分析中，多个变量之间往往存在一定程度上的相关性，从而增加了分析的难度。因此在主成分分析过程中，首先对原始数据进行标准化，消除因量纲不同可能带来的影响；之后人们通过线性组合的方式，将相关的变量转换为不相关的变量，且用较少的主成分就能得到大部分信息。主成分中所得的特征根即为各变量对应的方差，其表示了新指标的变异性，即所含信息量的大小，并根据其大小排序将各主成分排序得第一主成分、第二主成分等。

（二）主成分分析法的计算步骤

1. 原始指标数据的标准化采集 p 维随机向量$\boldsymbol{X}=\left(X_1,X_2,\ldots,X_p\right)^T$,n 个样品$\boldsymbol{x_i}=\left(x_{i1},x_{i2},\ldots,x_{ip}\right)^T$，构造样本阵，对样本阵元进行如下标准化变换：

($i = 1，2，\cdots，n$；$n>p$)，($j = 1，2，\cdots，p$)

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

其中$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$，$s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}$，得标准化阵 Z。

2. 对标准化阵 Z 求相关系数矩阵

$$\boldsymbol{R} = (r_{ij})_{p\times p} = \frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n-1}$$

其中，$r_{ij} = \frac{\sum_{a=1}^{n} z_{ai}\cdot z_{aj}}{n-1}$，（$i，j=1，2，\cdots，p$）

解样本相关矩阵$\mathbf{R}$的特征方程$|\boldsymbol{R} - \lambda \boldsymbol{I_P}| = 0$得 p 个特征根, 确定主成分。

本文的判断依据为特征根是否大于 1，从而确定 m 值。对每个λ_j，($j=1，2，\cdots，m$)，解方程$R\boldsymbol{b} = \lambda_j \boldsymbol{b}$，得单位特征向量$\boldsymbol{b_j}$。

3. 将标准化后的指标变量转换为主成分($j=1，2，\cdots，m$)

$$U_{ij} = \boldsymbol{z_i^T b_j}$$

$\boldsymbol{U_1}$称为第一主成分，$\boldsymbol{U_2}$称为第二主成分，…，$\boldsymbol{U_m}$成为第 m 主成分。

4. 对 m 个主成分进行综合评价, 将选取得特征根归一化, 即有（$i = 1,2,\cdots，m$）

$$w_i = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

记$\boldsymbol{W} = (w_1, w_2, \ldots, w_m)^T$,因此构造综合评价函数为

$$F_i = \boldsymbol{U_i W}$$

对 m 个主成分进行加权求和, 即得各样本的最终评价值F_i, W 为各权数, 即为再选取的主成分中的每个主成分方差贡献率。

（三）主成分分析结果的可视化与分析

利用 SPSS 将由主成分分析所得的水质综合指标与平台提供的水质类别指标进行时间序列图以及互相关图分析，观察与检验其变化趋势的相似性与相关性：

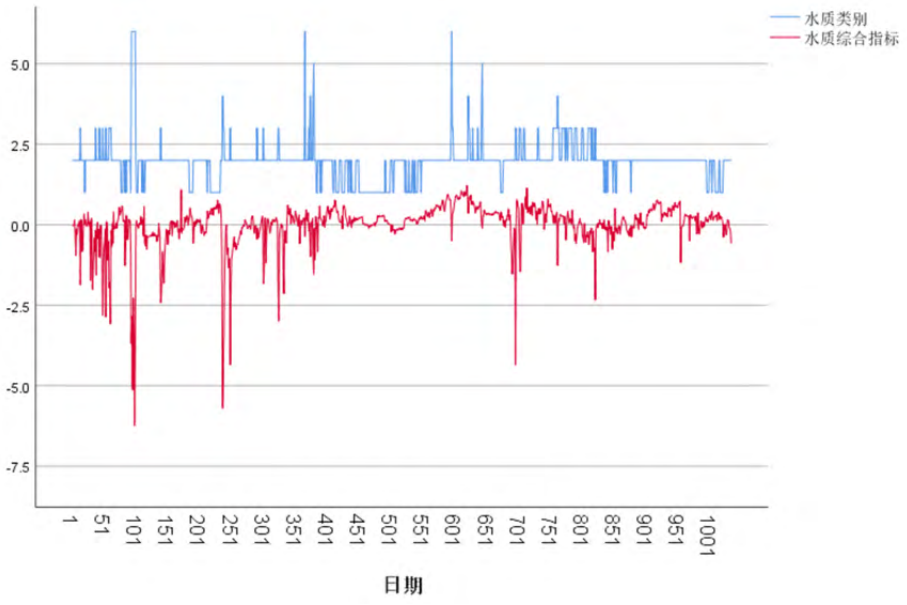

图 5 水质两项指标的序列图

由图 5 可知二者的变化趋势相似，主要相对于横轴，呈对称分布。

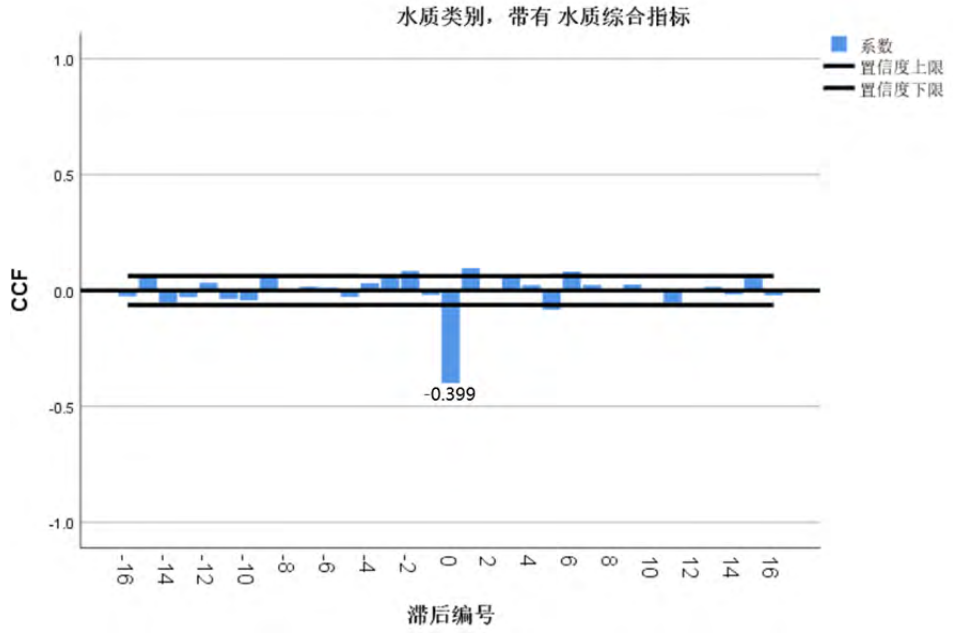

图 6 水质两项指标的互相关图

由图 6 的可知，观测指标与实际指标的时间序列之间在时间滞后，即滞后编号为 0 时，互相关函数值为-0.399，明显处于随机区间以外，从而认为二者有显著的相关性。因此我们认为由主成分分析所得的综合指标在一定程度上可以作为预测水质类别的观测数据。

五、CNN 卷积神经网络模型构建与评价

卷积神经网络（Convolutional Neural Networks，CNN）是一类具有深度结构利用卷积计算的前馈神经网络，是深度学习领域的代表算法之一；为近年来深度学习的迅速发展做出了巨大的贡献。CNN 主要包括三个维度的卷积神经网络，其中一维的 CNN 主要用于序列类数据的处理，二维的 CNN 主要用于图像处理与分类，三维的 CNN 主要用于医学图像或是视频预测等领域。

（一）卷积神经网络建模思路

CNN 凭借其优秀的计算能力，在图像识别和分类等技术领域占据主要地位，近年来在自然语言处理，时间序列建模，异常检测等领域得到了非常成功的应用。

卷积神经网络结构可以是无穷多个的。如图 7 所示为一种典型的卷积神经网络，由卷积层，池化层和全连接层构成。

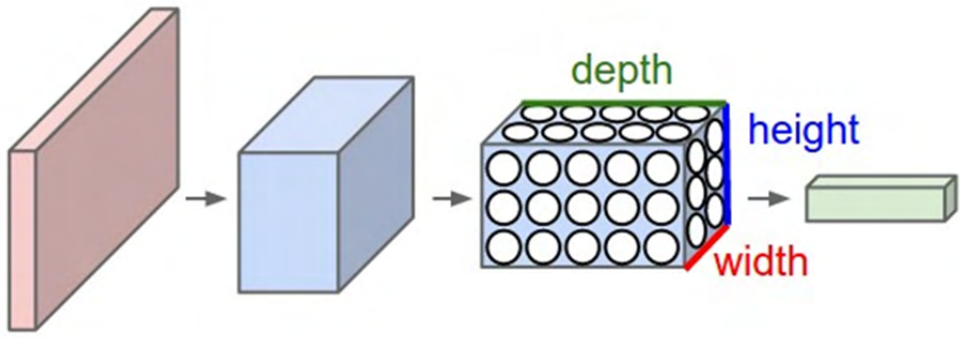

图 7 典型的卷积神经网络结构

本文根据经缺失值处理、主成分分析后得出水质综合指标，构建出邻域水质矩阵，将矩阵转化为二维灰度图像并输入到 CNN 模型中，经过卷积层、池化层、全连接层的处理后，根据水质类别进行图像分类，具体流程如图 8 所示。

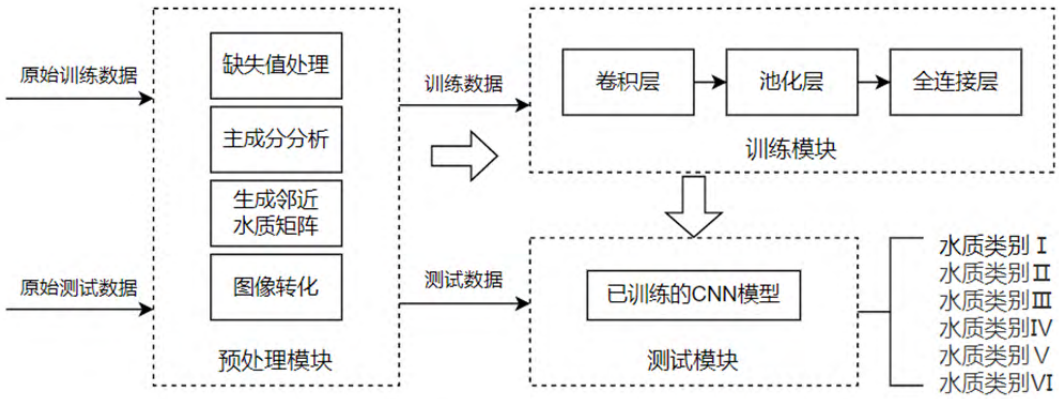

图 8 模型构建流程图

本文训练模块所采用的卷积神经网络结构如图 9 所示，由 3 层卷积层、2 层

池化层、1 层平滑层和 2 层全连接层组成。图像在 CNN 卷积结构中通过特征提取、反馈调整、类别判断，最终实现图像分类。

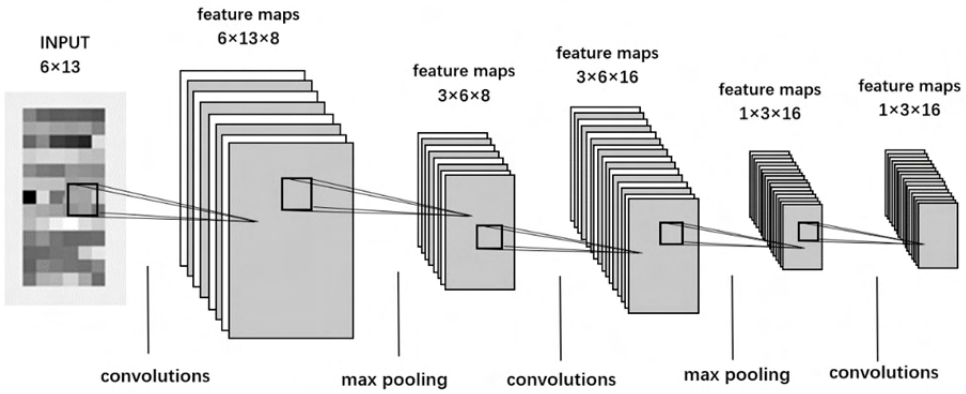

图 9 本文的卷积神经网络结构

（二）邻近预测模型

在对各监测站点进行水质类别的预测时，为了充分考虑附近地区水质情况所带来的影响以及 CNN 在图像识别中的优秀性能，本文在时间序列的基础上，建立了由卷积神经网络(CNN)驱动的相邻预测模型。通过构建邻近预测框架，生成了赣江流域水质数据的灰度图像，再结合二维卷积神经网络(CNN)，以构建水质类别预测模型。

1. 邻域水质矩阵

本文将水质监测站点的历史数据通过主成分分析得出水质综合指标（q），并将各站点的综合水质指标按照时间上的先后顺序、空间上河流流经顺序进行整理排序，构建出邻域水质矩阵$\varepsilon_{q_{x,t}}(x_1, x_2, s)$。

$$\varepsilon_{q_{x,t}}(x_1, x_2, s) = \begin{bmatrix} q_{x-x_1,t-s}, & \cdots & q_{x-x_1,t-2}, & q_{x-6,t-1} \\ \vdots & & \vdots & \vdots \\ q_{x-1,t-s}, & \cdots & q_{x-1,t-2}, & q_{x-1,t-1} \\ q_{x,t-s}, & \cdots & q_{x,t-2}, & q_{x,t-1} \\ q_{x+1,t-s}, & \cdots & q_{x+1,t-2}, & q_{x+1,t-1} \\ \vdots & & \vdots & \vdots \\ q_{x+x_2,t-s}, & \cdots & q_{x+x_2,t-2}, & q_{x+x_2,t-1} \end{bmatrix}$$

x:赣江流域内 31 个水质监测站点，并按照上游至下游的顺序将各站点进行编号（1≤x≤31）

$q_{x,t}$：x 站点在时间 t 时的水质综合指标

s：滞后天数

$\varepsilon_{q_{x,t}}(x_1, x_2, s)$：x 监测站点及其上游$x_1$个站点、下游$x_2$个站点滞后 s 天的水质情况

本文选取$x_1 = x_2 = 6$，$s = 6$，即通过 x 监测站点及其上下游 6 个站点前六天的水质情况，来预测 x 站点第七天的水质类别，相邻水质矩阵如表 7 所示。

表 7 邻域水质矩阵

	Day1	……	Day5	Day6	Day7
站点 1	$q_{x-6,t-6}$	……	$q_{x-6,t-2}$	$q_{x-6,t-1}$	
……	……	……	……	……	
站点 6	$q_{x-1,t-6}$	……	$q_{x-1,t-2}$	$q_{x-1,t-1}$	
站点 7	$q_{x,t-6}$	……	$q_{x,t-2}$	$q_{x,t-1}$	L
站点 8	$q_{x+1,t-6}$	……	$q_{x+1,t-2}$	$q_{x+1,t-1}$	
……	……	……	……	……	
站点 13	$q_{x+6,t-6}$	……	$q_{x+6,t-2}$	$q_{x+6,t-1}$	

2. 二维灰度图像

由于主成分分析得出的水质综合指标与水质类别有着显著的相关性, 因此由本文构建的邻域水质矩阵$\varepsilon_{q_{x,t}}(x_1, x_2, s)$在理论上能够良好的预测 x 站点在时间 t 时的水质类别。该矩阵可以转化为二维灰度图像，输入到 CNN 模型中进行水质类别图像分类，其中水质类别 I 至VI的二维灰度图像如图 10 所示。

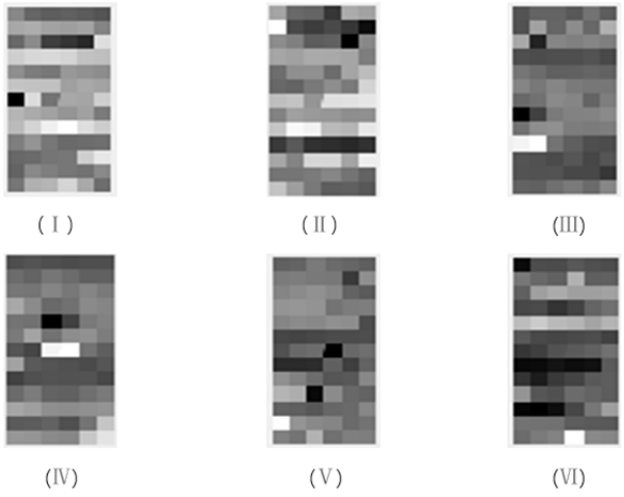

图 10 各水质类别的二维灰度图举例

将水质类别的预测问题构造为一个计算机视觉任务，充分利用 CNN 善于提取图像特征并进行分类的性质，能够充分利用水质数据的邻域效应，从而提高模型预测精度。

（三）CNN 建模过程与实证分析

本文搭建了同时考虑水质的时空变化规律的二维卷积神经网络的预测模型，建模过程与实证分析结果如下:

1. 卷积神经网络结构

（1）卷积层

卷积层是卷积神经网络中最核心的一层，它承担了卷积神经网络中大部分繁重的计算。在本文中，卷积层读取一个 6×13×1 的输入层，设置过滤器(Filter)数为 8，感受野为 3×3，通过卷积核与输入层之间的点乘运算进行图像局部特征的提取。其中卷积层结构如图 11 所示：

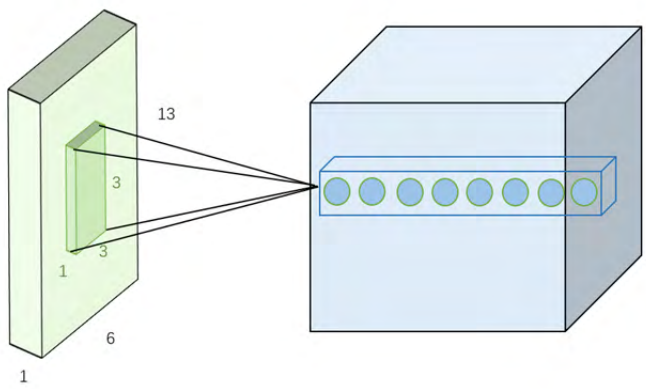

图 11 卷积层结构

第一层卷积设置了八个卷积核，将其拼接得到一个 6×13×8 的输出，从而提取了图像的八种特征。卷积层的运算公式如下，q_{ij} 表示邻域水质矩阵的第 i 行第 j 列元素值，x_{ij} 表示卷积核中第 i 行第 j 列的元素值，b 为偏差项。

$$f(x) = \sum q_{ij}\, x_{ij} + b$$

（2）池化层

在连续的卷积层之间插入池化层，能够减少网络中参数矩阵的尺寸，为最后的全连接层减少参数数量与计算量，从而达到减小过拟合发生概率的效果。池化层在每个输入通道上独立操作，利用 MAX 调整其空间大小 ;本文使用大小为 2x2 的过滤器的池化层，在输入的每个通道上以 2 的宽度和高度下采样，如图 12 所示。

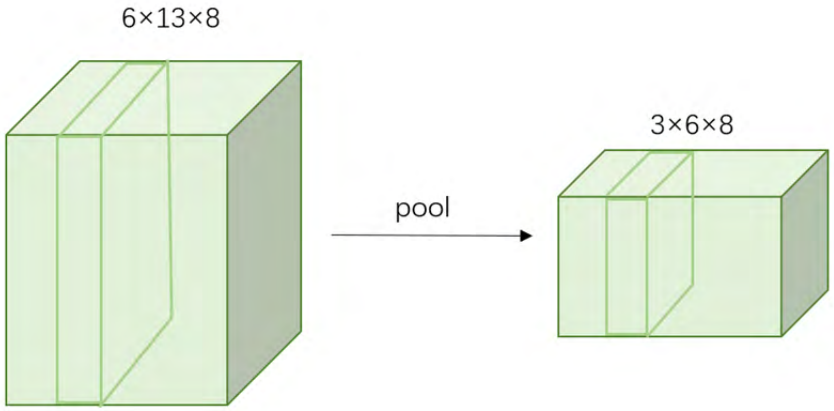

图 12 池化层结构

（3）全连接层

全连接层是将经卷积层和池化层处理后的数据进行非线性组合, 而后将进行降维处理的特征组合输出到输出层。其中, 全连接层运算公式如下：ω为权重系数, x 为全连接层输入，c 为偏置。

$$f(x) = \omega x + c$$

（4）激活函数

激活函数决定了某个特定的神经元是否被激活, 以及这个神经元接受到的特征信息是否有用，在神经网络的建立过程中发挥了重要的作用。它能够对输入信息进行非线性变换，并将非线性变换后的输出信息输入到下一层神经元中。常见的激活函数主要有：sigmoid 函数、tanh 函数、ReLU 函数、Leaky ReLU 函数等。本文采用 sigmoid 函数和 ReLU 函数作为激活函数。

Sigmoid 函数将定义在$(-\infty,+\infty)$的输入映射到(0,1)上，其公式为：

$$f(x) = \frac{1}{1+e^{x}}$$

ReLU 函数在当其输入值为正数时，便不存在梯度消失问题，且此函数的计算速度比其他函数快得多。ReLU 函数的公式为：

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x < 0 \end{cases}$$

2. 实证分析结果

二维卷积神经网络能较好地提取水质在时间和空间上的分布特征, 提高了预测精度。其中 CNN 模型的超参数设置如表 8 所示：

表 8 CNN 模型的超参数设置

模型组件	内核尺寸	过滤器数	输出结构
卷积层 1	3×3	8	6×13×8
最大池化层 1	2×2	NULL	3×6×8
卷积层 2	3×3	16	3×6×16
最大池化层 2	2×2	NULL	1×3×16
卷积层 3	3×3	16	1×3×16
平滑层	NULL	NULL	1×1×48
全连接层 1	NULL	NULL	1×1×32
全连接层 2	NULL	NULL	1×1×1

由图 13 可知，二维卷积神经网络的准确度可以达到 83.3%，同时在迭代过程中损失函数曲线稳定趋于 0。在处理多分类问题上准确率能够达到 83%。

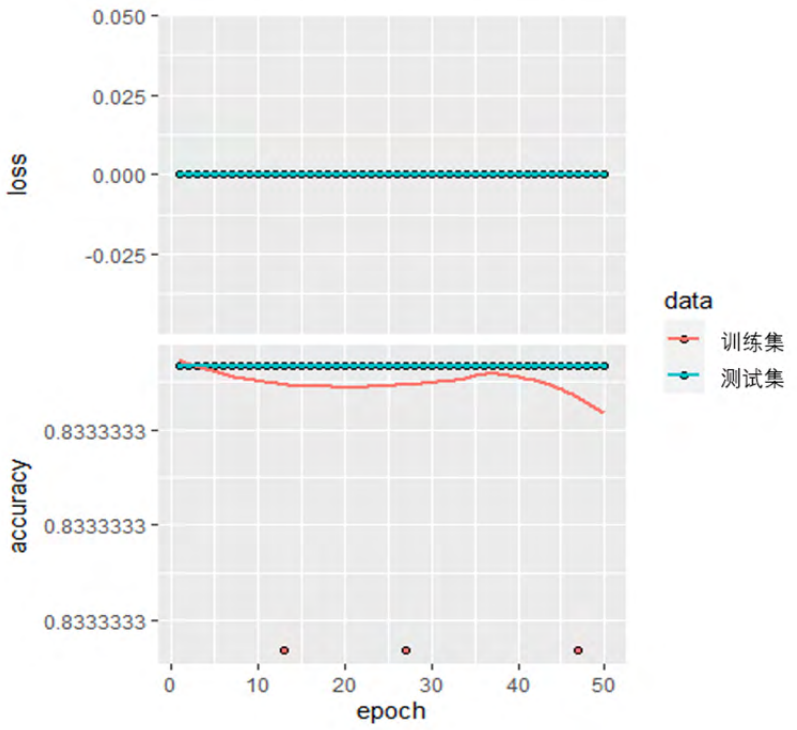

图 13 二维卷积神经网路模型准确率

因此可见本文提出的邻近预测模型能够更好的进行水质预测，且二维卷积神经网络模型具有优良的预测性能，能够为水质监管与治理提供有效的帮助。

（四）模型对比

为了对比不同模型之间对于水质预测问题的性能差异，我们随后建立了传统的时间序列分析中的 ARIMA 模型，与根据水质空间特征建立的一维卷积神经网

络预测模型，建模过程与实证分析结果如下:

1. ARIMA 模型

本节选取赣江流域上犹江江口站点为研究对象，建立 ARIMA 模型并对该站点的水质类别进行预测，模型建立流程如图 14 所示。

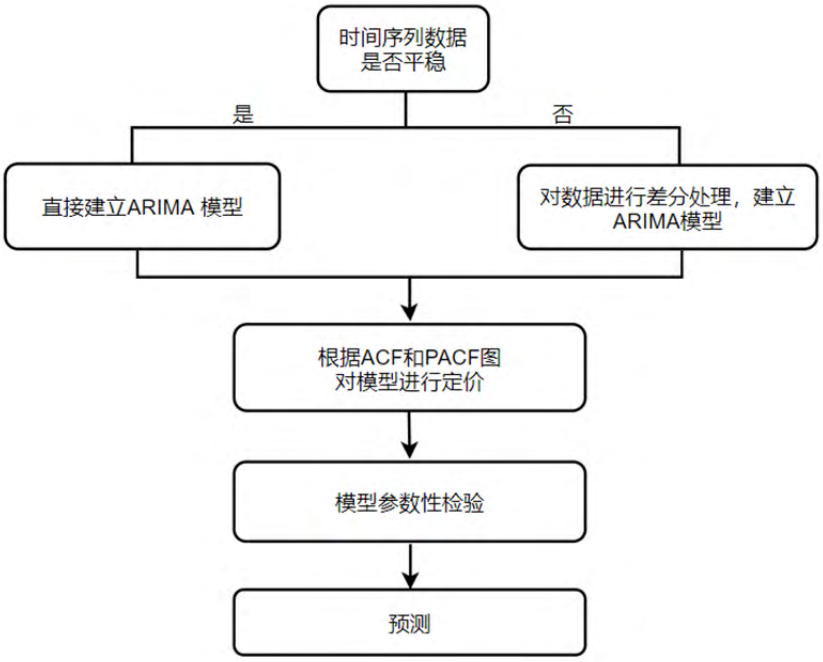

图 14 ARIMA 模型构建流程图

（1）数据平稳性检验

通过观察图 15 可以看出，水质综合指标和水质类别都是非平稳的序列，而 ARIMA 模型主要用于拟合具有平稳性的时间序列，因此对水质综合指标和水质类别进行了一阶差分，此时 ADF 检验结果显示两个序列都是平稳的。

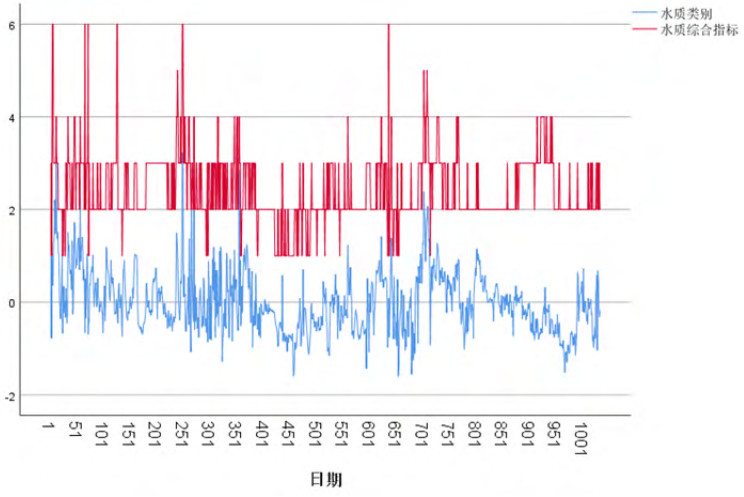

图 15 水质综合指标和水质类别时序图

（2）根据 ACF 图和 PACF 图确定模型参数

由图 16，通过观察水质类别一阶差分后的自相关图和非相关图，并利用 R 语言中的 auto. arima 函数，本文选择 ARIMA（1,1,1）模型。

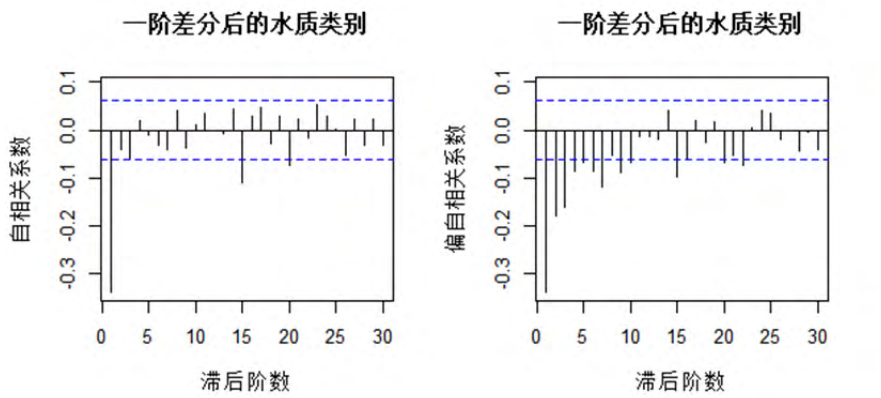

图 16 ACF 图与 PACF 图

（3）模型拟合结果

由表 9 可知，该模型平稳的 R 方为 0.556，表明现有模型仅能够解释原变量的 56.6%，平均绝对百分误差（MAPE）为 13.808＞10，因此模型拟合效果并不理想。

表 9 ARIMA 模型拟合结果

模型	ARIMA (1,1,1)
平稳 R 方	0.556
RMSE	0.423
MAPE	13.808
显著性	0.001

传统的 ARIMA 模型基于历史的水质监测数据进行推演，对异常数据较敏感。该模型仅对时间规律进行探究，并没有考虑到附近地区水质所带来的影响，因此具有一定的局限性。

2. 一维卷积神经网络

一维卷积神经网络可以提取同一时间下站点之间的空间规律，并以此进行回归拟合。本节选取同一时间赣江流域相邻的 15 个站点作为观测样本，搭建了仅考虑空间规律的一维卷积神经网络，对中间站点水质类别进行回归拟合。

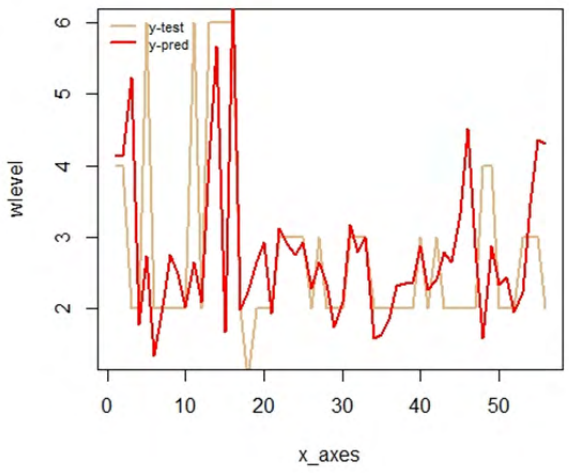

图 17 站点综合指标与水质类别回归

一维卷积神经网络对水质类别的预测结果与真实值的对比见图 17，可以看出模型的拟合效果一般。

表 10 一维卷积神经网络拟合结果

模型	一维卷积神经网络
平稳 R 方	0.1784
RMSE	1.2597
MAPE	0.3092
LOSS	0.1550

为了进一步验证预测结果，我们输出了如表 10 所示的检验统计量。由表 10 可知, 该模型平稳的 R 方为 0.1784, 表明现有模型仅能够解释原变量的 17.84%, LOSS 值为 0.1550>0.01 说明该模型的效果不好。该模型进行水质类别的预测时仅考虑到了站点之间的空间关系，其预测性能也是有限的。

六、结论与展望

（一）结论

1. 基于三种水质预测模型的结论

（1） 相较于 ARIMA 模型和一维卷积神经网络，二维卷积神经网络预测模型具有优良的特征提取功能, 并在时间序列的预测中有优秀的噪声容忍度，能较好地反映水质在时间和空间上的分布特征并进行分类。

（2） 二维卷积神经网络预测模型具有优越的预测性能, 在多分类问题的预

测准确率可以达到 83.3%，这一数据证明了 CNN 卷积神经网络在复杂流域的水质预测中具有十分突出的效果。

2. 基于大数据时代下水质预测的结论

本文所用数据来自由江西省生态环境厅官网提供的江西省地表水水质 APP。近年来为响应国家号召，实现环境保护透明化所公布的数据，尽管数据量达到了上万条，但因为站点水质数据的缺失值以及部分水质类别的数据量较少导致在图像识别的训练集选取中，信息利用率较低。我们认为若充分利用大数据时代数据公开这一特性，获取更多方面的数据，优化模型，对于水环境的预测与保护、环保部门的工作都将带来巨大的帮助。

（二）展望

本文根据赣江流域水质监测数据，建立水质预测模型，对赣江流域污染治理和水质改善具有实际参考价值。但本文仍存在不足之处，有待进一步探讨和改进。

1. 由于各水质指标间的具体关系未知，本文仅初步利用 PCA 得到的水质综合指标可能不是最优主成分结果。查阅资料后发现可利用非线性主成分分析法，即 KPCA，使主成分结果更加客观全面地反映各指标的特征。
2. 由于许多监测站点的数据有较多缺失值，因此本文仅选取了水质数据较为全面的 31 个站点作为研究对象，站点分布较为稀疏。随着水质监测公开数据的发展，将有助于对水质时空分布特征分析的改进。
3. 本文使用了邻近预测模型中的纯模型，从非机理性水质预测方面对赣江流域的水质类别进行预测，因此在模型改进中可以考虑使用混合模型：

$$q_{x,t} = f_1(f_2(\varepsilon_{q_{x,t}}(x1, x2, s),\ w(X_t)))$$

收集影响因素数据，如气象因素、站点周围污染源实时的排放信息等，从而更加全面地了解引起水质时空变化规律的原因，提高预测的精度。

参考文献

[1]付景智. 基于深度卷积神经网络的肺结节 CT 病理图像分类研究[D].中国科学技术大学,2020.

[2]李炳臻,刘克,顾佼佼,姜文志.卷积神经网络研究综述[J].计算机时代,2021(04):8-12+17.

[3]刘东君,邹志红.灰色和神经网络组合模型在水质预测中的应用[J]. 系统工程,2011(09):105-109.

[4]李如忠,汪家权,钱家忠.基于灰色动态模型群法的河流水质预测研究[J].水土保持通报,2002(04):10-12.

[5]罗学科,何云霄,刘鹏,李文.ARIMA-SVR 组合方法在水质预测中的应用[J]. 长江科学院院报, 2020, v.37;No.264(10):25-31.
[6]莫慧芳,谷爱昱,张新政,等.基于 BP 神经网络的水质预测方法的研究[J]. 控制工程, 2004(S1):9-10.
[7]唐宗鑫,简文彬.闽江下游水质预测的时间序列模型[J].水利科技,2002(02):7-9+50.
[8]吴涛,颜辉武,唐桂刚.三峡库区水质数据时间序列分析预测研究[J]. 武汉大学学报:信息科学版, 2006(06):500-502.
[9]王晓峰,周建,邹乐.基于支持向量分类与 GRU 神经网络联合的处理污水水质预测方法, CN111291937A[P].2020.
[10]许佳辉,王敬昌,陈岭,吴勇.基于图神经网络的地表水水质预测模型[J].浙江大学学报(工学版),2021,55(04):601-607.
[11]薛薇.SPSS 统计分析方法及应用(第 4 版)[M].北京:电子工业出版社,2017.
[12]于慧,孙宝盛,李亚楠,张燕,齐庚申.应用灰色模糊马尔科夫链预测海河水质变化趋势[J].中国环境科学,2014,34(03):810-816.
[13]朱建平.应用多元统计分析(第 4 版)[M].北京:科学出版社,2021.
[14]周志青,邹国防,王磊,王磊.基于 ARIMA/RBF-NN 的时间序列水质预测模型研究[J].科技通报,2017,33(09):236-240.
[15]Daniel Meier，Mario V. Wüthrich.Convolutional neural network case studies: anomalies in mortality rates; image recognition[J]. Fachgruppe “Data Science” Swiss Association of Actuaries SAV, Version of July 19, 2020
[16]Robert I.Kabacoff 著,王小宁,刘撷芯,黄俊文译. R 语言实战(第 2 版) [M].北京:人民邮电出版社, 2016.
[17]Wang, Chou-Wen, Jinggong, Wenjun. Neighbouring Prediction for Mortality[J]. ASTIN Bulletin - The Journal of the International Actuarial Association, 11-Feb-2021.

致谢

在此对我们的指导老师 XXX 老师，在建立、优化模型全程中，及时纠正错误，认真且耐心给予意见，提供参考资料，为我们指引模型应用方向；以及对 XXX 老师，耐心且及时地回答我们在建模过程中遇到的各种问题，并且给予专业的建议表示深深的感谢！