BP 神经网络和 ARIMA 模型对污水处理厂出水总氮浓度的模拟预测

林佳敏[1]，陈金良[1]，林晶晶[1]，李宣辑[1]，马聪[2*]，张志强[3]，沈亮[1*]

1.厦门大学化学化工学院，福建 厦门 361005

2.厦门水务环境科技股份有限公司，福建 厦门 361009

3.辽宁北方环境检测技术有限公司，辽宁 沈阳 110161

摘要 污水处理厂出水总氮（TN）浓度是评价水处理效果的关键指标之一，利用 BP 神经网络建立模型对污水处理厂脱氮工艺进行模拟，引入自回归移动平均模型（ARIMA）对污水处理厂未来短期出水 TN 浓度进行预测。结果表明：BP 神经网络模型在训练集和测试集模拟结果的平均相对误差分别为 15.9%和 16.5%，模型预测结果的平稳性较差；ARIMA 模型对未来 7 d 出水 TN 浓度的时序预测平均误差为 4.41%，预测精度较高；2 个模型相结合有助于实现污水处理厂快捷和高效的在线检测。

关键词 污水处理；总氮；BP 神经网络；ARIMA 模型

中图分类号:X703 文章编号:1674-991X(2019)04-0-0 doi:10.12153/j.issn.1674-991X.2019.03.260

The simulation and prediction of TN in wastewater treatment effluent using BP neural network and ARIMA model

LIN Jiaming[1], CHEN Jinliang[1], LIN Jingjing[1], LI Xuanji[1], MA Cong[2], ZHANG Zhiqiang[3], SHEN Liang[1]

1.College of Chemistry, Xiamen University, Xiamen 361005, China

2.Xiamen Water Environment Technology Co., Ltd, Xiamen 361009, China

3.Liaoning Northern Environmental Testing Technology Co., Ltd, Shenyang 110161, China

Abstract Total nitrogen in effluent is one of the critical indicators for evaluating the performance of wastewater treatment plants. In this study, a BP neural network model was developed to simulate the present nitrogen removal system for wastewater treatment, and an autoregressive moving average (ARIMA) model was creatively applied to realize the short-term prediction of future effluent. The results showed that the simulation error of BP model on training set was 15.9%, and that on test set was 16.5%，which revealed that the stability of model prediction is poor. The average error of the ARIMA model for predicting the total nitrogen value in the coming week was around 4.41%, which showed high prediction accuracy. The combination of the two models can help fast and efficient on-line detection, establishing accurate aeration system, reducing power cost, and providing technical

收稿日期：2018-11-16

基金项目：福建省自然科学基金项目（2018J01016）；福建省高校青年自然基金重点项目（JZ160461）；厦门市科技计划项目（3502Z20173018）

作者简介：林佳敏（1996—），女，主要从事工业数据处理研究，13276023278@163.com

*责任作者：1.马聪（1988—），女，工程师，博士，主要从事污水处理技术研究，mc@xmwaterenv.com

2.沈亮（1977—），女，副教授，博士，主要从事污水处理技术研究，shenliang@xmu.edu.cn

reference for the operation and regulation of wastewater treatment plants.

Key words wastewater treatment; TN; BP neural network; ARIMA model

污水处理厂出水的总氮（TN）浓度是评价其性能的关键指标之一[1]。污水处理厂脱氮常采用生物脱氮工艺，该过程综合了生物反应、化学反应和物理反应，机理错综复杂，仅基于数学模型和专家经验很难满足污水处理厂实际运营中对出水调控的需求。近年来，基于数据挖掘的建模方法已广泛应用于污水处理中，应用实例主要包括预测模拟[2]、异常情况预警[3]及污水处理系统优化[4]三大类，而涉及的算法也较多，其中BP 神经网络因其强大的非线性自适应能力备受青睐。BP 神经网络是一种基于误差反向传播算法训练的多层感知器前馈网络[5]，通过大量已有的样本学习训练，找到最适的输入-输出之间的非线性映射关系。利用BP 神经网络建立的模型可对污水处理厂脱氮系统进行模拟仿真[6]，但其最大缺陷在于没有引入时间变量，无法满足污水处理厂运行过程中对未来出水变化预测的需求。为解决该问题，引入了经济管理领域的自回归移动平均模型（autoregressive integrated moving average，ARIMA），该模型基于污水处理厂出水 TN 浓度随时间变化的历史数据，可实现对未来出水 TN 浓度的短期预测，适用于运行稳定的污水处理厂。

近年来，随着污水处理厂监测水平的提高，污水处理数据的采集更加便捷和快速[7]。数据的积累和更新为数据深度挖掘和模型预测的准确度提升提供了可能。随着互联网、大数据、人工智能的兴起，智慧水务将成为未来污水处理发展的趋势[8]，而大数据分析是建立智慧水务的核心环节。笔者以实际污水处理厂出水 TN 浓度为核心指标，基于对出水 TN 浓度影响因素的分析，简化模型输入变量，集成 BP 神经网络模型和 ARIMA 模型的互补关系，形成智慧水务的数字化模型矩阵，以期为污水处理厂从空间到时间的双维度调控提供依据。

1 BP 神经网络模型

1.1 输入变量的确定

污水处理厂采用的生物脱氮工艺主要是利用硝化和反硝化细菌的协同作用完成的。进行硝化作用的亚硝酸菌和硝酸菌均为好氧、自养型生物，硝化过程消耗氧气；而反硝化过程涉及的反硝化细菌为厌氧、异养型生物，该过程在缺氧条件下发生，需提供碳源作为电子供体，并为反硝化细菌提供能量。因此，影响污水处理厂脱氮效果的因素包括污水的氨氮（NH_4^+-N）浓度、溶解氧（DO）浓度、生化需氧量（BOD_5）、化学需氧量（COD_{Cr}）、碳氮比（BOD_5/TN）；温度（T）和 pH 也是关键因素，二者均会影响相关细菌的活性及其反应速率；污水处理的生物除磷过程会与生物脱氮过程竞争氧气与碳源，因此污水中总磷（TP）浓度在一定程度上也会影响污水的脱氮效果。理论上而言，上述指标都应该列入模型的输入变量，但污水厂实际运行过程中，往往不针对 T 和 pH 作任何调控，二者只作为指示参数进行常规记录，因此 T 与 pH 不纳入模型输入变量。

1.2 数据来源与处理

BP 神经网络建模的数据源自福建省某污水处理厂 2016—2017 年运行报表。该污水处理厂采用卡鲁塞尔 2000 氧化沟，设计污水处理规模为 2.5 万 t/d，出水执行 GB 18918—2002《城镇污水处理厂污染物排放标准》一级 B 标准。由于该污水处理厂没有直接的 DO 浓度数据，考虑到 DO 浓度与曝气系统的耗电量直接相关，将综合电单耗用作反映 DO 浓度的间接指标纳入模型输入变量。

为确定模型的输入变量，把综合电单耗和进水 COD_{Cr}、BOD_5、TN 浓度、NH_4^+-N 浓度、TP 浓度、BOD_5/TN 分别用符号 Elec、COD_{in}、BOD_{in}、TN_{in}、NH_4^+-N_{in}、TP_{in}、BOD_5/TN_{in} 表示，输出变量即出水 TN 浓度用 TN_{out} 表示，按下式分别计算各输入变量与 TN_{out} 的皮尔逊等级相关系数〔r（X，Y)〕。

$$r(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \quad (1)$$

式中：Cov（X，Y）为变量 X、Y 的协方差；Var[X]为变量 X 的方差；Var[Y]为变量 Y 的方差。

各输入变量与 TN_{out} 的 r（X，Y）见表 1。由表 1 可知，BOD_5/TN_{in} 与 TN_{out} 几乎无相关性，这是因为在污水处理过程中相较 TN_{out} 的变化，BOD_5/TN_{in} 波动范围很小，可以看作是常数；其余 6 个输入变量与输出变量 TN_{out} 均有一定的相关性。因此，把 Elec、COD_{in}、BOD_{in}、TN_{in}、NH_4^+-N_{in}、TP_{in} 作为 BP 神经网络输入的 6 个节点，即 X=[X_1，X_2，…，X_6]，而输出节点只有一个，即 Y=TN_{out}。

表 1　各输入变量与 TN_{out} 的 r（X，Y）

Table 1　Pearson correlation coefficient between the indicators and TN_{out}

输入变量	与 TN_{out} 的 r（X，Y）
Elec	0.509 253
COD_{in}	0.536 536
BOD_{in}	0.427 874
TN_{in}	0.649 341
NH_4^+-N_{in}	0.589 526
TP_{in}	0.565 149
BOD_5/TN_{in}	-0.034 380

为提高训练过程中迭代求解的收敛速度和迭代求解的精度，对原始数据进行最大值归一化处理[9]：

$$X_{\mathrm{norm}}^{i,j}=\frac{X^{i,j}}{\max(X^j)};i=1,2,...m;j=1,2,...n \quad (2)$$

式中：$X_{\mathrm{norm}}^{i,j}$ 为第 j 个特征中第 i 个变量归一化处理后的值；$X^{i,j}$ 为第 j 个特征中第 i 个变量的原始值；max（X^j）为所有变量中第 j 个特征的最大值。

将污水处理厂运行报表中 731 组样本按 7:3 进行数据切割，划分后洗牌得到的训练集样本为 511 组，测试集样本为 220 组。

1.3 数据挖掘

基于 Python3.5 的 Tensorflow 框架进行神经网络模型的构建。通过文献[10]和多次试验，确定输入层为 6，隐藏层为 5，输出层为 1 的神经网络结构，利用控制变量法逐一设置神经网络的参数组合。采用指数下降的自适应学习率[11]，使损失函数在训练前期较快下降以接近最优解，在后期随着学习率减小缓慢下降而不至于在最优解附近出现震荡。为避免过拟合，采用 L2 正则化[12]。最终确定的神经网络参数见表 2。

表 2　神经网络参数

Table 2　The neural network parameters

调试参数任务	参数值
批处理参数（batch size）	64
初始学习率	0.2
学习率的衰减率	0.93
正则化系数	0.005
激活函数	relu-identity
迭代次数	2 000

采用均方误差作为神经网络回归模型的损失函数（loss）[13]：

$$\text{loss} = \text{MSE}(y, y) = \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n} \quad (3)$$

式中：y_i 为第 i 个变量的观测值；y_i' 为第 i 个变量的神经网络模拟值。

loss 随着训练次数增加不断下降，最终收敛。当 loss 收敛时，即可认为神经网络已完成对训练数据的学习，可以停止迭代。

1.4 结果分析

采用平均绝对误差（ε_{MAE}）和平均相对误差（ε_{MRE}）来对模型的性能进行评估，其计算公式如下：

$$\varepsilon_{MAE} = \frac{\sum_{i=1}^{N}|y_{\_} - y|}{N} \quad (4)$$

$$\varepsilon_{MRE} = \frac{\sum_{i=1}^{N}|(y_{\_} - y)/y|}{N} \times 100\% \quad (5)$$

式中：$y\_$ 为模拟值；y 为观测值。

利用 Google 开源 Tensorflow 可视化软件，对程序运行过程中和运行结果的相关指标和计算图进行展示。将每一次神经网络的迭代以日志文件的形式保存到给定的路径下，训练结束后启动 Tensorboard，由 Tensorboard 得到训练完成后 loss 随时间的变化趋势，如图 1 所示。由图 1 可知，loss 趋近于 0.01。

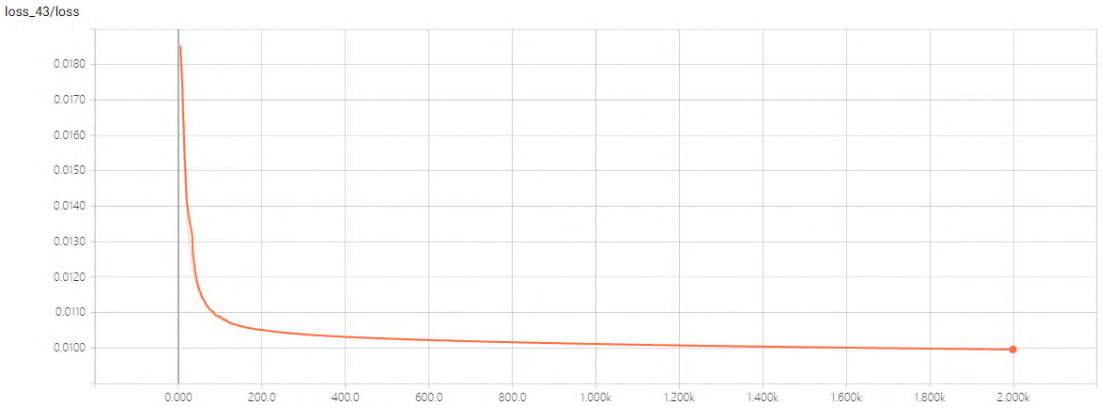
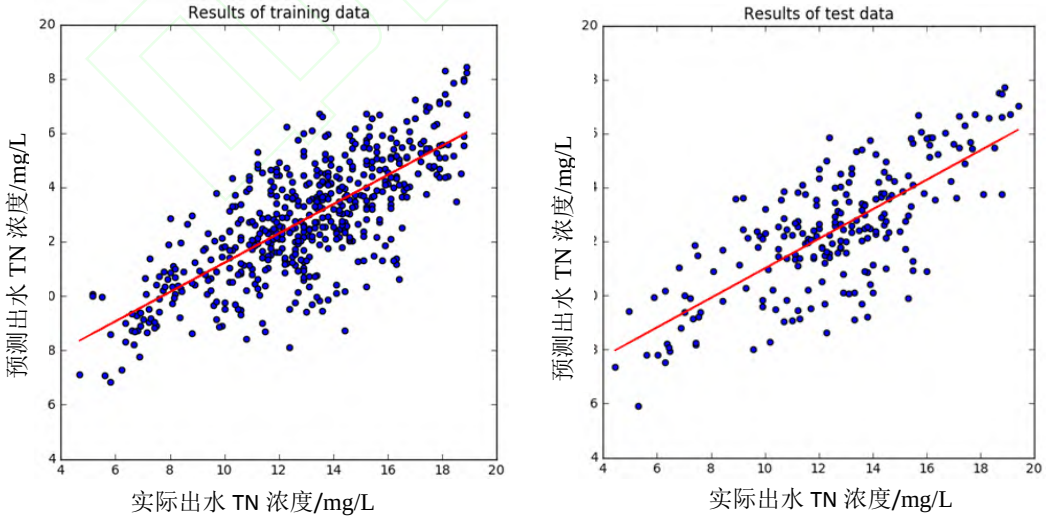


图 1　Tensorboard 中 loss 随时间的变化趋势

Fig.1　The change trend of loss function over time in Tensorboard

训练集和测试集的拟合结果如图 2 所示。由图 2 可知，训练集和测试集的结果表现相近，学习结果没有出现拟合，训练集和测试集模拟结果的平均相对误差分别为 15.9%和 16.5%。结合误差分布柱形图（图 3）可知，虽然模型最终收敛，但整体模型的稳定性表现不佳，相对误差（MRE）的分布很宽，甚至有误差超过 50%的模拟点，侧面反映了在污水处理过程中存在着大量不确定的因素，这也是 BP 神经网络无法规避的问题。BP 神经网络预测精度不理想的原因：该模型是简化的间接模型，实际上影响 TN_{out} 的输入变量并非只有 6 个指标，且综合电单耗并不能直接反映 DO 浓度；BP 神经网络难以表达污水环境中存在的不确定性因素，抗干扰能力较差，因此预测结果的平稳性较差。

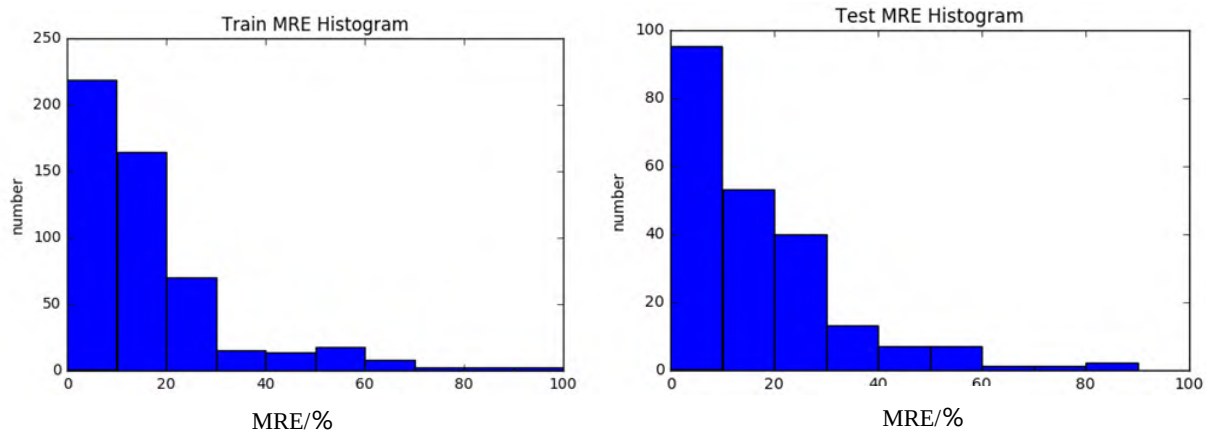
虽然 BP 神经网络模型的精确度和稳定性有待提高，但该模型为污水处理厂从进水到出水的工艺运行模拟提供了可能。将基于 BP 神经网络建立的非线性映射关系推广到实际应用中，理论上可解决传统实验室测量耗时费力的问题，建立出水 TN 浓度的软测量模型[14]，实现出水 TN 浓度的在线、实时监测，还可以利用神经网络对生物除氮的工艺参数和运行参数进行仿真分析，以此优化脱氮效果。



（a）训练集　　（b）测试集

图 2　训练集与测试集的拟合结果

Fig.2　Simulation results of training data and test data



（a）训练集　　（b）测试集

图 3　训练集和测试集模拟结果的相对误差分布

Fig.3　MRE of train data and test data simulation

2　ARIMA 预测模型

2.1 模型定阶

利用基于 Python3.5 环境的 statsmodels 库建立 ARMIA 模型。选取 2016 年 1 月—2017 年 11 月出水 TN 浓度的时间序列作为模型的训练样本（图 4），将 2017 年 12 月第 1 个 7 d 的数据作为模型的验证样本。进行 ARIMA（p，d，q）模型定阶，即确定模型的 3 个参数 p、d、q[15]。对原始数据进行一阶差分处理，以达到 ARIMA 模型平稳性要求，结果如图 5 所示。

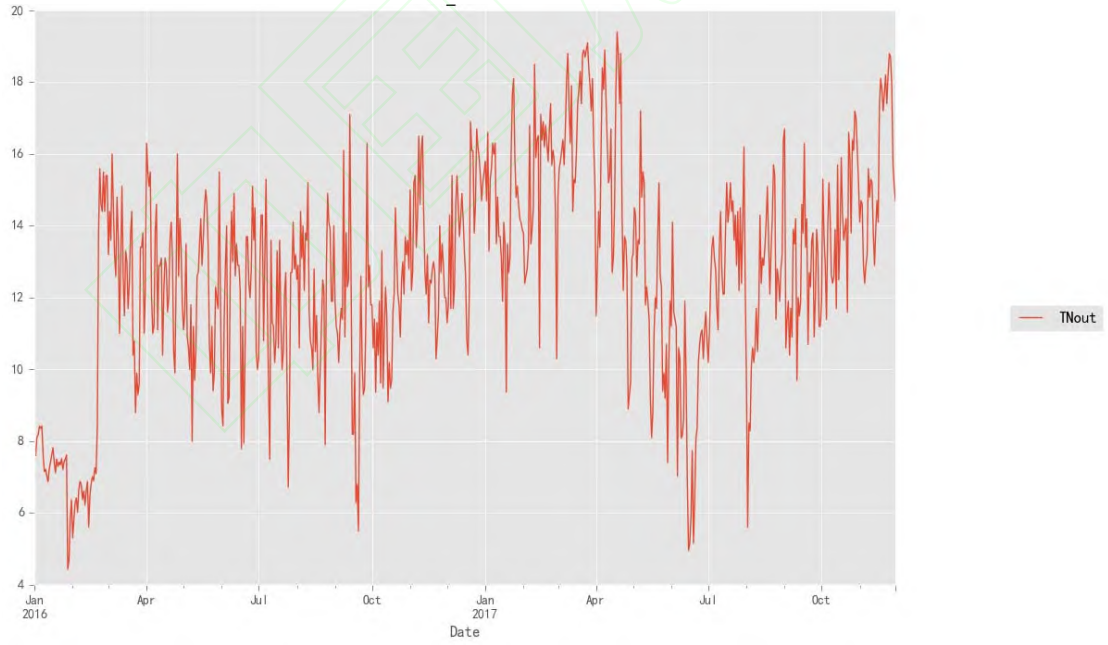


图 4　出水 TN 浓度时间序列

Fig.4　The line chart of change trend of TN_{out} over time

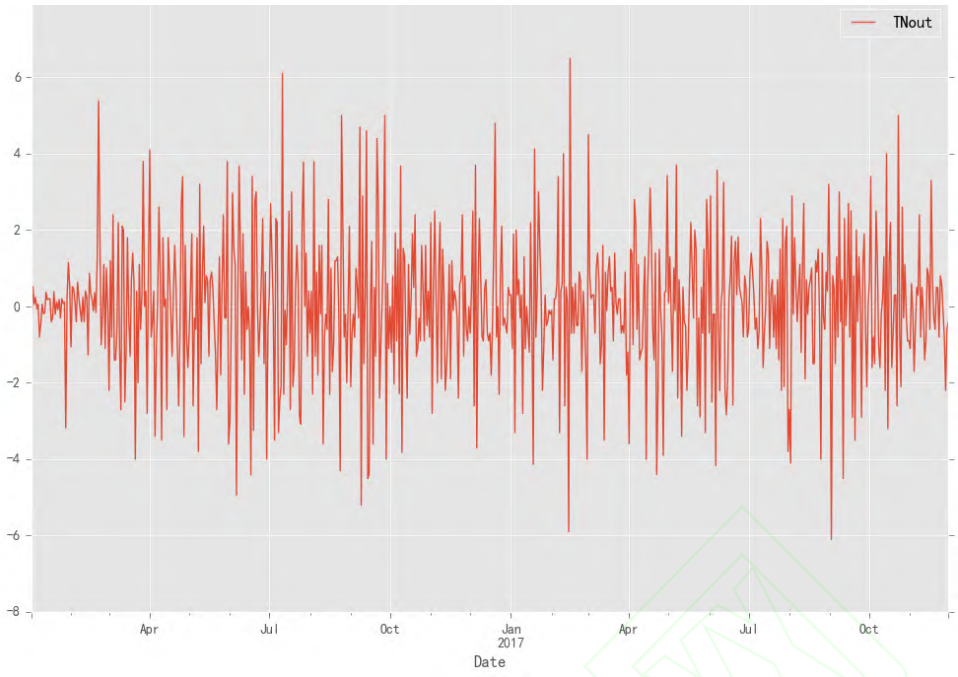


图 5　出水 TN 浓度时间序列的一阶差分结果

Fig.5　The change trend of TN_{out} over time (results from first order difference)

为了消除人为的判断误差，进一步确定（p，d）值，设置不同的（p，q）组合，按下式比较赤池信息准则（AIC）大小。

$$AIC=2k-2\ln L \qquad (6)$$

式中：k 为模型参数的个数；L 为似然函数。L 越大，模型的拟合能力越好，同时 k 越小，参数越少，能够有效避免过拟合。当 AIC 最小时可获得最优组合参数。

参数 p、q 不同组合的 AIC 热度图如图 6 所示。图 6 纵坐标中的数字表示 p 值，“AR3”表示在 AR 模型中 p=3；横坐标中的数字表示 q 值，“MA3”表示在 MA 模型中 q=3。图中每一个方格内的数字代表对应（p，q）组合生成的 AIC，AIC 越小，颜色越深。由图 6 可以找出，横坐标 AR5 与纵坐标 MA4 对应的方格颜色最深，其 AIC 最小。即当 p=5，q=4 时，$(AIC)_{min}$=2 723.40。综上，p=5，q=4，d=1，则确定移动平均模型 ARIMA 的参数为（5，1，4）。

注：横坐标中的数字代表 p 值；纵坐标中的数字代表 q 值；方块中的数字代表 ACI 值。

图 6　参数 p，q 不同组合的 AIC 热度图

Fig.6　AIC heat map of different combinations of p and q

2.2 模型检验

模型的残差折线图、残差 QQ 图、自相关函数（ACF）与偏相关函数（PACF）图分别如图 7~图 9 所示。图 7~图 9 均表示模型残差是服从均值为 0、方差为常数的正态分布的白噪声序列，即认为该模型是恰当的。

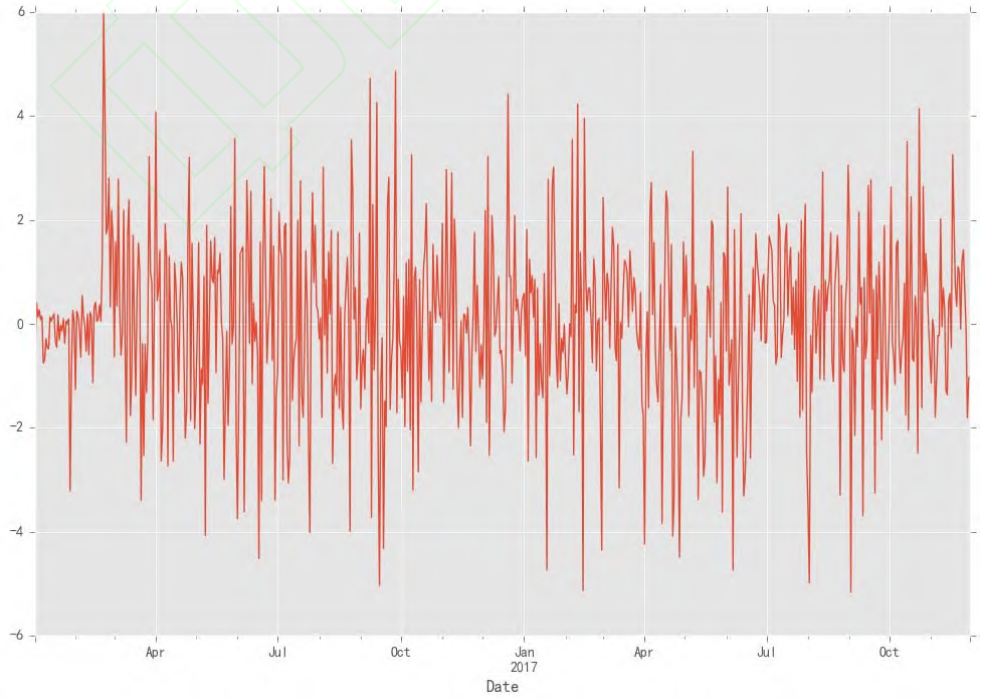


图 7　ARIMA 模型的残差折线图

Fig.7　Residual line chart of ARIMA model

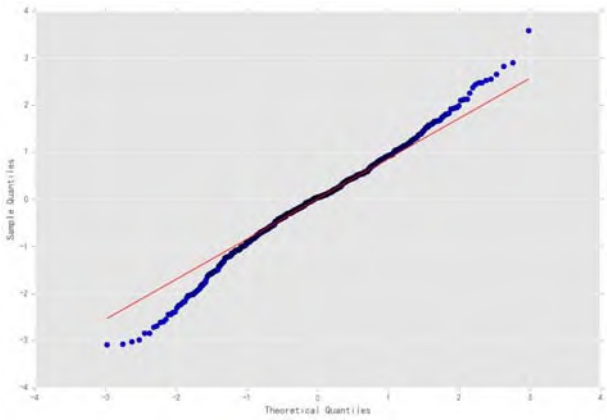


图 8　ARIMA 模型的残差 QQ 图

Fig.8　Residual plot between sample quantiles and theoretical quantiles of ARIMA model

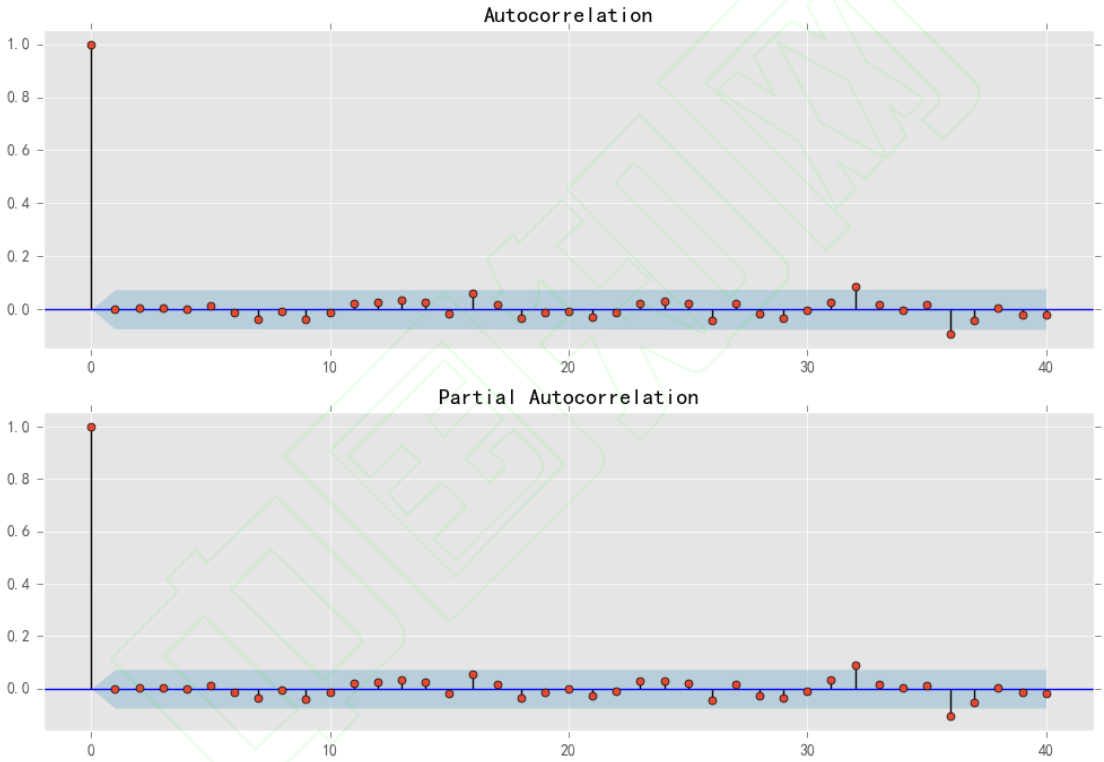


图 9　ARIMA 模型残差的 ACF 和 PACF 图

Fig.9　Residual ACF chart and PACF chart of ARIMA model

2.3 模型预测

利用通过检验的模型对污水处理厂 2017 年 12 月 1—7 日（7 d）的 TN_{out} 进行预测，结果如表 3 所示。由表 3 可知，预测值与实际值比较接近，相对误差均小于 10%，平均相对误差为 4.41%，最小误差为 1.19%，最大误差为 8.26%。可见，ARIMA 模型的预测精度较高，可用来预测未来 7 d 的 TN_{out} 时间序列值。比较预测值与实际值的变化发现，实际值波动比较大，而预测值基本不变，预测值难以敏感地反映 TN_{out} 的变化。这是因为 TN_{out} 除了受各污染物的进水浓度影响之外，还与污水量大小、*T*、pH、运行控制参数等有关，模型模拟时只考虑了部分参数，使模拟精度尚不足，后续 ARIMA 模型尚需进一步完善。

表 3　ARIMA 模型的预测值与实际值

Table 3　Predictive value from ARIMA model and actual value

日期	实际值/(mg/L)	预测值/(mg/L)	MAE	MRE/%
2017-12-01	15.8	16.513 318	0.713 318	4.514 668
2017-12-02	15.2	16.456 224	1.256 224	8.264 631
2017-12-03	16.8	16.450 698	-0.349 302	-2.079 180
2017-12-04	17.5	16.416 359	-1.083 641	-6.192 232
2017-12-05	17.6	16.416 794	-1.183 206	-6.722 761
2017-12-06	16.2	16.394 186	0.194 186	1.198 677
2017-12-07	16.1	16.408 835	0.308 835	1.918 230

3　结论

（1）基于污水处理过程中 6 项进水指标与出水 TN 浓度，建立基于非线性映射关系的 BP 神经网络模型，该模型在训练集和测试集模拟结果的平均相对误差分别为 15.9%和 16.5%，BP 神经网络模型预测结果的平稳性较差，但该模型可对污水处理厂从进水到出水的工艺运行进行模拟。

（2）建立基于 ARIMA 的预测模型，对污水处理厂未来 7 d 出水 TN 浓度的时序预测精度较高，预测值与实际值比较接近，平均相对误差为 4.41%。

基于 ARIMA 的预测模型在污水处理厂优化运行决策、工艺运行模拟、异常情况警报等方面具有一定的意义。ARIMA 模型的预测功能理论上可以实现智能预警，当发现异常可能问题时能够提前通知相关工作人员进行诊断并进行有针对性的运营调控，为污水处理厂的无人值守提供可能。但 ARIMA 模型还需要不断补充新的出水 TN 浓度的时间序列数据，使其得到进一步修正或重新拟合，以得到更好的预测结果。

参考文献

[1] 李佟,李军.基于 BP 神经网络与马尔可夫链的污水处理厂脱氮效果模拟预测[J].环境科学学报,2016,36(2):576-581.

LI D,LI J.The prediction of denitrification efficiency of a wastewater treatment plant by using BP neural network and Markov chain method[J].Acta Scientiae Circumstantiae,2016,36(2):576-581.

[2] WEI X,KUSIAK A,SADAT H R.Prediction of influent flow rate: data-mining approach[J].Journal of Energy Engineering,2013,139(2):118-123.

[3] DOGAN S,DURSUN S.Error checking of input data for web based design calculations of wastewater treatment plant[C]//Albena:7th international scientific conference on modern management of mine producing,geology and environmental protection,2007.

[4] VERMA A,WEI X,KUSIAK A.Predicting the total suspended solids in wastewater:a data-mining approach[J].Engineering Applications of Artificial Intelligence,2013,26(4):1366-1372.

[5] ZHEN L,XU L Y.Research on overcoming the local optimum of BPNN[C]//IEEE Computer Society.Proceedings of the world congress on intelligent control and automation (WCICA).Dalian:2006 6th World Congress on Intelligent Control and Automation,2006:2681-2685.

[6] SHI Y,ZHAO X T,ZHANG Y M,et al.Back propagation neural network (BPNN) prediction model and control strategies of methanogen phase reactor treating traditional Chinese medicine wastewater (TCMW)[J].Journal of Biotechnology,2009,144(1):70-74.

[7] RODRIGUEZ-JEANGROS N,RODRIGUEZ J P,CAMACHO L A,et al.Integrated urban water resources model to improve water quality management in data-limited cities with application to Bogota,Colombia[J].Journal of Sustainable

Water in the Built Environment,2018,4(2):04017019.
[8] 孙国庆.智慧水务关键技术研究及应用[J].水利信息化,2018,142(1):46-49.
SUN G Q.Research and application on key technologies of smart water[J].Water Resources Informatization,2018,142(1):46-49.
[9] 滕明鑫.回归神经网络预测模型归一化方法分析[J].电脑知识与技术,2014,10(7):1508-1510.
TENG M X.The analysis of normalization method of recurrent neural network prediction model[J].Computer Knowledge and Technology,2014,10(7):1508-1510.
[10] CHEN K,YANG S J,BATUR C.Effect of multi-hidden-layer structure on performance of BP neural network:probe[C]//Chongqing:2012 eigth international conference on natural computation.2012:1-5.
[11] DAI H,MACBETH C.Effects of learning parameters on learning procedure and performance of a BPNN[J].Neural Networks,1997,10(8):1505-1521.
[12] YANG Q,HU Y J,XUE L.Back-propagation model for nanofiltration process simulation in pesticide wastewater treatment[J].Advanced Materials Research,2010,168/169/170:404-407.
[13] 辛大欣,王长元,肖峰.BP 神经网络在回归分析中的应用研究[J].西安工业学院学报,2002,22(2):129-135.
XIN D X,WANG C Y,XIAO F.A study on the BP neural network applied to regression analysis[J].Journal of Xi'an Institute of Technology,2002,22(2):129-135.
[14] 韩红桂,陈治远,乔俊飞,等.基于区间二型模糊神经网络的出水氨氮软测量[J].化工学报,2017,68(3):1032-1040.
HAN H G,CHEN Z Y,QIAO J F,et al.Soft-sensor method for effluent ammonia nitrogen based on interval type-2 fuzzy neural networks[J].CIESC Journal,2017,68(3):1032-1040.
[15] IHUEZE C C,ONWURAH U O.Road traffic accidents prediction modelling:an analysis of Anambra State,Nigeria[J].Accident Analysis & Prevention,2018,112:21-29.