基于多重相关性学习的风电场SCADA数据修复及其功率预测应用

郑李梦千[1],朱利鹏[1],文唯嘉[2],李佳勇[1],张 聪[1]
(1. 湖南大学 电气与信息工程学院,湖南 长沙 410082;2. 国网湖南省电力有限公司信息通信分公司,湖南 长沙 410004)

摘要:风电场数据采集与监视控制(SCADA)系统实测数据中的数据缺失、噪声等非理想测量工况给短期风电功率的可靠预测带来严峻挑战。为解决这个问题,提出了一种基于多重相关性学习的SCADA数据修复方案。对于SCADA实测数据中存在的数据缺失问题,提出综合挖掘多维时序数据多重相关性的数据修复方法,对缺失数据进行初步修复;设计适用于多种复杂工况的残差神经网络,对初步修复结果进行进一步精细化处理,实现精细的缺失值修复和数据去噪;以修复后的数据为输入,通过基于多头注意力机制的卷积神经-长短期记忆深度学习网络构建高可靠的短期风电功率预测模型。华中地区2座风电场实测SCADA数据的算例分析结果验证了所提方法的有效性及其在提升短期风电功率预测性能方面的应用价值。

关键词:SCADA数据修复;多重相关性;短期风电功率预测;深度学习;残差神经网络
中图分类号:TM614;TP18 **文献标志码**:A DOI:10.16081/j.epae.202412026

0 引言

近年来,我国风电的开发和利用呈大规模、高质量发展态势,截至2024年6月,我国累计风电装机容量达4.67×10^8 kW[1-2]。然而,由于风电的波动性和随机性,大规模并网给电力系统的安全稳定运行带来严峻挑战。因此,需进行高效、可靠的风电功率预测,以减小电力系统所面临的风险,提高整体稳定性。

风电功率预测是利用风电场历史输出功率及气象数据建立模型,预测未来风电功率的技术手段[3]。所需数据主要来源于风电场的数据采集与监视控制(supervisory control and data acquisition,SCADA)系统和数值天气预报(numerical weather prediction,NWP)系统,其中前者持续采集、存储与分析风电场的运行数据,后者用来预测未来的气象状况。在实际运行中,由于传感器故障、数据传输异常、测量偏差等原因,SCADA系统采集到的运行数据中不可避免地存在大量异常值、缺失数据、噪声,NWP数据中也存在不符合实际的异常值。若直接使用含异常值、缺失值和噪声的数据进行风电功率预测,则结果可能出现较大偏差,难以满足系统调度和控制等高级应用需求。目前异常值的辨识技术已趋于成熟[4],一经辨识后,因其无法提供有用数值信息,可将其视为伪缺失值,而传统意义上的缺失数据一般用0或其他特殊符号进行标记,无须另作辨识。因此,本文主要聚焦于含缺失值(包括以异常数据形式呈现的伪缺失值)和噪声的SCADA数据修复处理。

对于SCADA数据缺失问题,目前的数据修复方案主要分为基于统计理论的方法和基于机器学习的方法。基于统计理论的方法主要利用SCADA数据的统计特性来估计缺失值,如均值填补法、插值法、回归模型填补法[5]等。这类方法计算简单、易于理解,但难以应对复杂的数据缺失模式。基于机器学习的方法一般利用K-最近邻(K-nearest neighbor,KNN)[6]、缺失森林(missing forest,MF)[7]、支持向量机(support vector machine,SVM)[8]等算法,通过构建预测缺失值的模型修复缺失数据。此类方法可适应不同的数据类型,挖掘数据中复杂的非线性关系,但对参数敏感性较高,实际应用中亦难以适用于未出现在训练数据中的未知缺失模式。

在特定观测时间窗下,风电场SCADA多维数据中存在着复杂的相关性,包括单个变量时序演变过程中不同时间断面测量值的自相关性和多个变量之间相同时间断面测量值的互相关性。文献[5-8]均在一定程度上间接利用了这些相关性,但并未系统、全面地挖掘。为充分挖掘数据内部相关性,研究人员将多种算法结合,从多方面对数据进行修复。文献[9]将SVM、插值法和线性回归相结合,对风电场的缺失数据进行修复。但其仅考虑了数据内部的自相关性,修复效果仍有一定的提升空间。文献[10]针对传感器网络提出一种数据修复方法,通过线性插值估计短时间内平稳变化的缺失数据,再结合近邻节点的互相关性,利用多元回归模型修复缺失值。该方法能同时处理变化平稳与剧烈的缺失数据,但选择合适的回归模型仍是一大问题。文献[11]判断缺失数据是否连续及与何种相关性(自、互相关性)

收稿日期:2024-04-02;**修回日期**:2024-08-26
在线出版日期:2024-12-17
基金项目:国家自然科学基金资助项目(52207094,52377095)
Project supported by the National Natural Science Foundation of China(52207094,52377095)

更显著相关，再根据判别结果从4种回归算法中选择最合适的算法估计缺失值。文献[12]提出一种简化的计及时空相关性的修复算法，采用缺失时刻前的数据和最近邻空间的数据对缺失数据进行修复。上述方法均表现出良好的修复精度，但在面对风电场SCADA数据缺失模式复杂、非线性特征突出等问题时，难以充分挖掘SCADA数据中的相关性，亦未考虑实际工况下不规律噪声对数据修复效果造成的不良影响。

针对现有研究的不足，本文提出一种综合考虑数据内部多重相关性的风电场SCADA数据修复方案。首先，从全局和局部同时分析多维数据内部的自相关性和互相关性，从多个视角对缺失数据进行初步修复；然后，构建残差网络去噪模型，对初步修复后的数据综合进行去噪和精细修复，由此实现SCADA多维数据的完整修复；最后，在此基础上，以实际风电场SCADA数据为例，对本文所提数据修复方案自身的有效性及其对于改善短期风电功率预测精度的效果进行验证。

1　SCADA数据修复方案

本文所提SCADA数据修复方案由两部分构成，整体架构如图1所示，包括基于多重相关性学习（multiple correlation learning，MCL）的两阶段缺失数据初步修复和基于残差神经网络的数据精细去噪，对数据进行全方位综合修复，其中两阶段缺失数据初步修复包括多视角缺失数据填补和多视角填补结果加权回归。

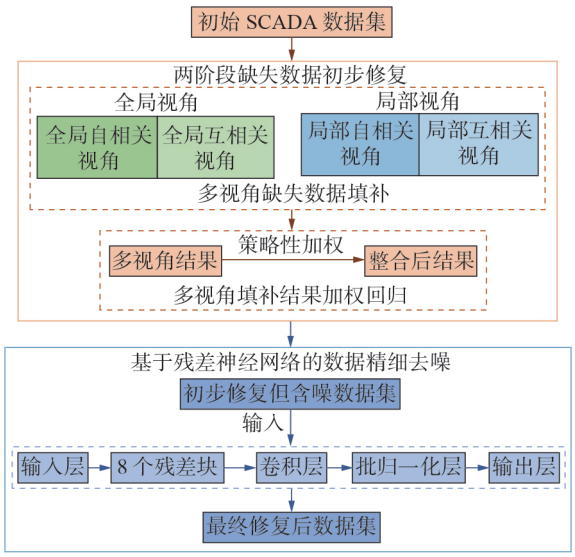


图1　SCADA数据修复整体框架

Fig.1　Overall framework of SCADA data correction

1.1　两阶段SCADA缺失数据初步修复

1.1.1　多视角缺失数据填补

本文所提的多视角缺失数据填补分别涉及全局互相关视角、全局自相关视角、局部互相关视角、局部自相关视角，在不同视角下充分利用SCADA缺失数据点的横向近邻和纵向近邻进行数据填补。不失一般性，SCADA数据典型缺失情况如图2所示。对于某一缺失数据点x，其横向近邻为不同变量在同一时间断面的测量数据点；纵向近邻为同一变量在不同时间断面的测量数据点；局部数据矩阵为在给定时间窗下同时包含x横向近邻和纵向近邻的局部的测量数据点集合。

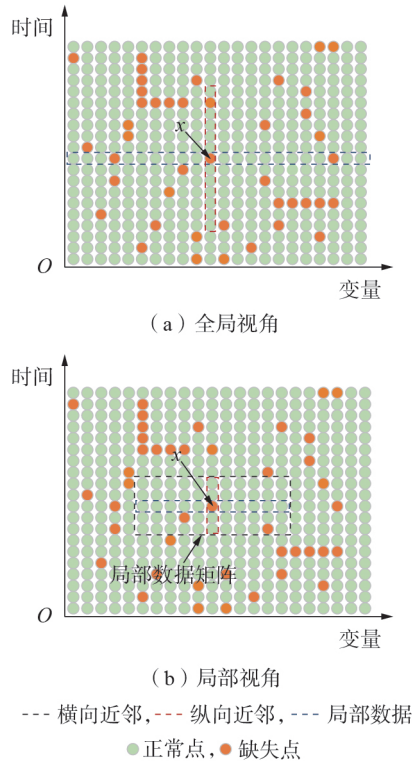


图2　SCADA缺失数据示意图

Fig.2　Schematic diagram of missing data in SCADA

1）全局互相关视角。全局互相关视角下的缺失数据填补利用反距离加权（inverse distance weighting，IDW）方法[13]对缺失数据进行基于邻域的插值。如图2（a）全局视角所示，以多维变量间的欧氏距离为权重，对于图中的某个缺失点x，利用其横向近邻数据点的值对其进行填补，如式（1）所示。

$$\hat{V}_{gc}=\frac{\sum_{i=1}^{m_1} v_i d_i^{-\alpha}}{\sum_{i=1}^{m_1} d_i^{-\alpha}} \tag{1}$$

式中：$\hat{V}_{gc}$为在全局互相关视角下得到的结果；m_1为目标缺失点x的横向近邻中无缺失数据点的近邻数；v_i为横向近邻中第i个变量的测量值；α为控制权值的衰减系数；d_i为候选数据点所在变量与目标缺失点所在变量之间的欧氏距离。设目标缺失点所在变量为第p个变量，则d_i可表示为：

$$d_i=\sqrt{\sum_{j=1}^{h}(v_{ji}-v_{jp})^2} \tag{2}$$

式中：v_{ji}和v_{jp}分别为变量i和p在第j个时间断面的测量值；h为在给定观测时间窗内2个变量均无缺失的时间断面总数。

2）全局自相关视角。考虑到全局视角下的自相关性，基于简单指数平滑（simple exponential smoothing，SES）的思想[14]，利用缺失点所在变量在其他时间断面的值来填补缺失值，如图2（a）中纵向近邻所示。SES本质上是对历史数据的加权平均，如式（3）所示。

$$y'_{t+1}=\delta y_t+\delta(1-\delta)y_{t-1}+\cdots+\delta(1-\delta)^{t-1}y_1 \tag{3}$$

式中：y'_{t+1}为t+1时刻的预测值；y_t为t时刻的测量值；δ为平滑系数，取值范围为0~1，δ越接近于0，表示历史数据衰减越平缓。

传统的SES仅利用了发生缺失前的历史数据，未考虑缺失点所在时间断面后的数据中所蕴含的信息。为充分计及全局自相关性，该视角同时利用发生缺失时刻前、后的数据进行插值计算，如式（4）所示。

$$\hat{V}_{\mathrm{gs}}=\frac{\sum_{j=1}^{m_2}v_j\beta(1-\beta)^{t_j-1}}{\sum_{j=1}^{m_2}\beta(1-\beta)^{t_j-1}} \tag{4}$$

式中：$\hat{V}_{\mathrm{gs}}$为在全局自相关视角下得到的结果；m_2为目标缺失点x的纵向近邻中无缺失数据点的近邻数；t_j为候选数据点与目标缺失点所在时间断面之间的差值；v_j为纵向近邻中第j个时间断面的测量值；β为控制权重随时间差异衰减的衰减系数，取值范围为0~1。

3）局部互相关视角。局部互相关视角的思想源于商业用户推荐系统中基于用户的协同过滤推荐（user-based collaborative filtering，user-based CF）方法[15]，通过捕捉各变量间局部互相关性对缺失数据进行填补。具体而言，根据图2（b）中的局部数据矩阵中的测量值计算各个变量与目标缺失点所在变量间的相似度，设目标缺失点所在变量为变量p，变量i、p的相似度为：

$$\Phi_{i,p}=1/\sqrt{\frac{1}{N_{\mathrm{T}}}\sum_{j=1}^{N_{\mathrm{T}}}(v_{ji}-v_{jp})^2} \tag{5}$$

式中：N_{T}为局部数据矩阵中2个变量均无缺失值的时间断面总数。以相似度为权重对待选变量的测量值进行加权平均计算，如式（6）所示。

$$\hat{V}_{\mathrm{lc}}=\frac{\sum_{i=1}^{m_3}v_i\Phi_{i,p}}{\sum_{i=1}^{m_3}\Phi_{i,p}} \tag{6}$$

式中：$\hat{V}_{\mathrm{lc}}$为在局部互相关视角下得到的结果；m_3为局部数据矩阵中目标缺失点x的横向近邻中无缺失数据点的近邻数。

4）局部自相关视角。与局部互相关视角相似，局部自相关视角的思想源于商业推荐系统中基于推荐内容的协同过滤（item-based collaborative filtering，item-based CF）方法[16]，利用缺失点各个相邻时间断面之间的相关性对缺失数据进行填补修复。首先利用局部数据矩阵中的测量值计算不同时间段断面之间的相关性，设目标缺失点所在时间断面为第q个时间断面，记第j、q个时间断面的相似度如式（7）所示。

$$\Phi_{j,q}=1/\sqrt{\frac{1}{N_{\mathrm{S}}}\sum_{i=1}^{N_{\mathrm{S}}}(v_{ji}-v_{qi})^2} \tag{7}$$

式中：N_{S}为在第j、q个时间断面均无缺失值的变量数；v_{qi}为第i个变量在第q个时间断面的测量值。在此基础上，以相关性为权重，利用局部数据矩阵中相邻时间断面的测量值对目标缺失点进行插值，如式（8）所示。

$$\hat{V}_{\mathrm{ls}}=\frac{\sum_{j=1}^{m_4}v_j\Phi_{j,q}}{\sum_{j=1}^{m_4}\Phi_{j,q}} \tag{8}$$

式中：$\hat{V}_{\mathrm{ls}}$为在局部自相关视角下得到的结果；m_4为局部数据矩阵中目标缺失点x的纵向近邻中无缺失数据点的近邻数。

1.1.2 多视角填补结果加权回归

上述4种视角相辅相成，对缺失数据点进行全方位的填补修复。最终通过多视角学习将4个视角的计算结果整合，得到最终结果$\hat{V}_{\mathrm{mcl}}$为：

$$\hat{V}_{\mathrm{mcl}}=\omega_1\hat{V}_{\mathrm{gc}}+\omega_2\hat{V}_{\mathrm{gs}}+\omega_3\hat{V}_{\mathrm{lc}}+\omega_4\hat{V}_{\mathrm{ls}}+b \tag{9}$$

式中：ω_1—ω_4为给每个视角分配的权重；b为残差。通过最小化预测结果与真实值之间的二乘误差对每个SCADA变量进行训练，得到最优的权重。

1.2 基于残差神经网络的数据精细去噪

为清除数据集内原本的噪声及由MCL初步填补修复所带来的误差，本文采用一种基于残差神经网络的数据精细去噪模型，进一步减小数据集与真实值之间的偏差，提高数据质量。

在传统的神经网络中，随着网络层数的叠加，误差在反向传播过程中可能会出现梯度爆炸和梯度消失的问题，导致模型训练难以完成，且随着网络深度的增加，模型的训练误差可能表现出先降低而后升高的“性能退化”现象[17-18]。为有效解决上述问题，文献[17]率先提出残差神经网络，通过残差连接的引入，直接将输入链接到输出中，引导神经网络学习从输入到输出的恒等映射，亦使梯度在网络中更高

效地传播，从而有效缓解梯度爆炸或消失的问题。残差连接结构如附录A图A1所示。

在数据去噪的应用场景中，与传统卷积神经网络（convolutional neural network，CNN）模型直接学习从噪声输入到无噪声输出的映射不同，残差网络模型通过学习输入数据中噪声的残差来还原无噪声数据，残差连接的存在使得网络能更好地学习数据集中复杂的噪声分布，具有更好的去噪性能。对残差网络进行训练，整体结构如附录A图A2所示，其中输入和输出分别对应含噪数据集和无噪声数据集。该网络由输入层、8个残差块、卷积层、批归一化层、输出层构成，其中每个残差块都包含2个卷积层、2个批归一化层和1个残差连接，以学习包括噪声在内的序列复杂特征和模式。每个残差块的输出经过Relu激活函数后，作为下一个残差块的输入或输出层的输入。经多个残差块的堆叠处理后，网络通过1个卷积层和1个批归一化层关联至输出层，以输出去噪后的数据。在整体训练过程中，网络通过残差连接学习如何利用残差去除输入数据中的噪声，总的学习目标是最小化损失函数E_{Loss}，如式（10）所示。

$$E_{\text{Loss}}=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i-y_i) \tag{10}$$

式中：N为去噪模型的样本数；y_i、$\hat{y}_i$分别为去噪模型样本的真实值、模型的预测值。在对残差网络训练学习后，模型能输出更接近不含噪声的原始序列，并尽量保留有用的数据信息，更大程度地还原真实数据。

2　风电功率预测方案

2.1　短期风电功率预测基本概念

受地形地貌、气象变化等多方面因素影响，风电具有显著的随机性和波动性。随着风电的高速发展，新能源高效消纳与电力系统安全稳定运行的矛盾日益突出。准确的风电功率预测有助于减小系统备用容量，提高电力系统的安全性与稳定性。风电功率预测按时间尺度可将其划分为超短期、短期、中长期、长期风电功率预测，其中短期风电功率预测是制定日发电计划、电网调度运行的基础，主要任务是预测从次日00:00时起未来3天的风电功率，其时间分辨率为15 min。

本文以基于机器学习的风电功率预测方法为例，对风电场未来3天输出风电功率曲线进行预测，并根据评估结果和实际预测性能验证本文所提数据修复方法在应用于风电功率预测领域时的有效性和可行性。

2.2　短期风电功率预测模型构建

以深度神经网络为代表的先进机器学习方法在风电功率预测中展现出巨大的应用潜力[19]，因此，本文利用深度神经网络算法构建短期风电功率预测模型。在现有深度神经网络算法中，CNN通过卷积层和池化层的交替堆叠，能识别并学习数据中的全局与局部趋势，有效捕捉风电数据内的空间特征，识别风速、风向等局部模型，具有计算效率高、鲁棒性强等优势。此外，由于风电的各变量间存在着复杂的时序依赖性，长短期记忆网络（long short-term memory，LSTM）凭借门控机制和记忆单元的存在，可以有效处理数据中的长序依赖关系，具有强大的时序建模能力，在捕捉风电系统动态变化方面具有其独特的优势。因此，本文将CNN与LSTM相结合，在对风电场SCADA数据时空特性进行深度学习的基础上，提高模型对风场空间结构和时序动态的理解，并引入多头注意力机制，进一步提高短期风电功率预测精度。本文所构建的基于多头注意力机制的CNN-LSTM联合驱动型风电功率预测模型（记为CNN-LSTM-A）如附录A图A3所示，实施流程具体描述如下。

将修复后的风电场SCADA数据和NWP数据作为功率预测的信息源，以与短期风电功率预测时段相对应的未来3天的NWP数据（包括风速、风向数据）、过去3天的历史气象数据（包括风速、风向数据）、过去3天的历史功率数据为主要输入，温度、湿度等其余的NWP和历史气象数据为辅助输入，以预测时段的功率数据为输出，对预测模型进行训练。首先，经卷积层内的卷积核和激活函数对输入进行卷积操作，提取数据内的空间特征；再通过池化层对数据进行降维，保留数据重要特征，并将其作为LSTM_1层的输入；接着，利用LSTM_2层充分捕捉并挖掘序列的时序特征，将每个时间步的输出都传递到下一步，使网络能学习到更长的时间依赖性；再利用多头注意力机制增强模型对不同位置与特征的关注，提高模型对输入数据的表达能力，使模型能在不同的子空间中学习到更丰富的信息表示，以提高风电功率预测模型的准确性；最后，加入Dropout层以防止过拟合，并通过Flatten层和全连接层调整网络输出，得到未来3天的风电功率预测曲线。

3　算例分析

为验证本文所提方法的可行性与有效性，采用我国华中地区2座实际风电场的SCADA运行数据进行算例分析。其中1号风电场的装机容量为95.5 MW，2号风电场的装机容量为150 MW。选取2021年9月至2022年9月的数据进行分析，采样间隔为15 min，每天共96个采样点。2座风电场的SCADA数据均包含场站实发功率及不同高度（例如10、30 m等）的风速、风向、温度、湿度等气象

信息。

3.1 算例测试设置

3.1.1 数据处理

为提高模型训练效率，增强模型的鲁棒性，首先对不同维度的数据进行归一化处理，计算公式如附录A式(A1)所示。

为充分模拟实际复杂工况，本文通过完全随机缺失的方式构造数据缺失，当缺失率较高时，可覆盖各种可能的分散型缺失和整段连续缺失情况。进一步，向归一化的数据集中添加不同强度的高斯噪声，以充分模拟实际测量环境中各种可能的噪声水平[20]。

3.1.2 参数设置

在基于MCL的两阶段缺失数据填补方案中，以多次经验性测试为依据，设置权值衰减系数$\alpha=9$，$\beta=0.5$，局部数据矩阵的宽度$w=5$，长度$l=11$。在基于残差神经网络的数据精细去噪模型中，共设置8个残差块，统一采用Relu激活函数。主要超参数设置如附录A表A1所示。选择Adam优化器，均方误差(mean squared error，MSE)作为损失函数对模型进行训练，批次为32，迭代次数为200。

在基于多头注意力机制的CNN-LSTM短期风电功率预测模型中：CNN模块由2个卷积层和2个池化层组成，第1层卷积层有64个卷积核，大小设置为4，第2层卷积层有32个卷积核，大小设置为3，采用最大池化方式；LSTM模块由2个LSTM层组成，神经元个数均为64，设置注意力头的数量为4，每个注意力的维度为64。训练过程采用Adam优化器，选择MSE为损失函数，批次设为32，迭代次数为200。

3.1.3 评价指标

为对数据修复效果进行评估，选择平均绝对误差(mean absolute error，MAE)、均方根误差(root mean squared error，RMSE)作为评价指标，分别表示为E_{MAE}、E_{RMSE}，计算公式如附录A式(A2)、(A3)所示。

为有效量化风电功率预测模型的精度，结合风电场规模大小，采用经风电场开机容量规范化后的E'_{MAE}和E'_{RMSE}作为评价指标[21]，计算公式如附录A式(A4)、(A5)所示。

3.2 数据修复性能测试

为充分测试不同场景下的数据修复效果，在2座风电场的SCADA数据集中设置不同严重程度的数据缺失，缺失率分别为10%、30%、50%，并在0.5%~2.5%的标准差范围(代表不同噪声强度)内添加高斯噪声。在此基础上，利用本文的数据修复方案对含缺失值和噪声的SCADA数据进行处理。为直观展示数据修复效果，以1号风电场为例，不失一般性地从中抽取某一天的风速数据，给出其修复前、后的风速曲线，如图3所示(图中20 m、50 m、

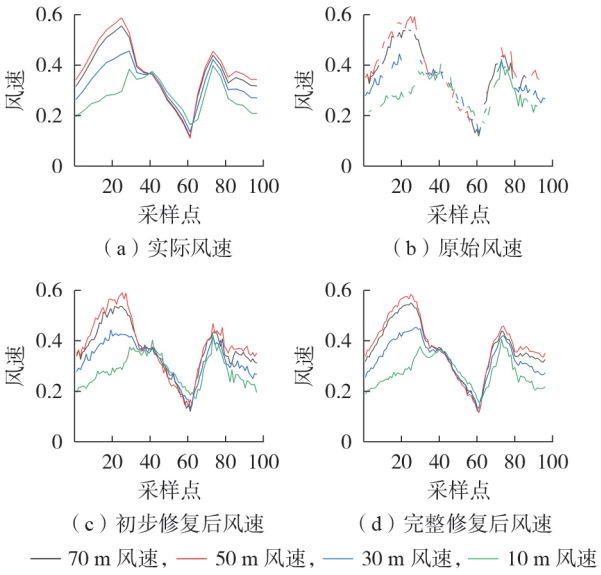


图3 1号风电场风速数据修复前、后对比

Fig.3 Comparison of wind speed data between before and after restoration for wind farm No.1

30 m、10 m为高度，风速为标幺值)，其数据缺失率为30%，添加噪声强度为1.5%。由图可知，对数据进行完整填补及去噪修复后，所得到的风速曲线与原始风速曲线整体变化趋势保持一致，仅在局部数据点上存在细微偏差。相比于图3(b)中含缺失值和噪声的风速曲线，修复后的数据质量得到大幅提升。以此处修复得到的高质量风速数据为输入，将有望进一步提升后续风电功率预测精度。

进一步地，对本文数据修复方案的整体修复效果进行统计分析。2座风电场经MCL初步修复后的SCADA数据集在去噪修复前后的误差统计结果如附录A图A4、A5所示。由图可知，经过残差神经网络模型去噪修复后的数据集与去噪前相比，与真实数据集间的偏差能减小40%以上；且模型能适应不同强度的噪声，在高噪声下仍能保持良好的去噪性能，具有较强的鲁棒性。这是因为残差网络能直接学习输入与期望输出之间的残差，捕捉噪声的非线性特征，集中学习噪声成分，同时更深的网络结构也使得模型在处理复杂数据等方面具有更优越的性能。

为进一步验证本文所提数据修复方法的优越性，将本文所提方法与均值填补法、KNN、加权KNN(weighted K-nearest neighbor，WKNN)、MF、文献[12]中所提的简化时空相关性方法(下文简称为LM方法)、MCL与传统CNN模型相结合的方法(不包含残差连接，记为MCL-CNN)进行对比测试。在缺失率为30%、噪声强度为1.5%的情况下对不同方法的修复性能进行测试，相应误差统计结果如表1所示。

均值填补法是用整体的平均值来填补缺失值，其计算简单、易实施，但误差很大。KNN和WKNN

表1　不同缺失数据修复方法误差统计

Table 1　Statistics of error for different missing data correction methods

风电场	方法	E_{MAE}	E_{RMSE}
1号	均值填补	0.162	0.317
	KNN	0.012	0.023
	WKNN	0.011	0.020
	MF	0.009	0.018
	LM	0.010	0.018
	MCL	0.008	0.016
	MCL-CNN	0.008	0.014
	本文方法	0.007	0.010
2号	均值填补	0.081	0.191
	KNN	0.029	0.071
	WKNN	0.026	0.062
	MF	0.017	0.040
	LM	0.015	0.039
	MCL	0.013	0.032
	MCL-CNN	0.011	0.027
	本文方法	0.008	0.016

都是基于相似性原理，利用邻近的观测值对缺失数据进行估计。由表1可知，KNN和WKNN具有一定的修复效果，但对噪声较为敏感，导致修复性能下降。事实上，由于在高缺失率下数据集将变得十分稀疏，难以找到足够数量的邻近值进行估计，当缺失率更高时，此类方法的数据修复误差将大幅增加。MF是通过建立多个决策树模型来预测缺失值，每个决策树独立训练、集成学习，适合处理复杂的数据结构和高维数据，其计算复杂，且需调整参数以获得最佳的性能，但精度仍优于基于KNN和WKNN的修复方法。与其他方法相比，未进行去噪修复的MCL能从4个视角充分挖掘多维数据内部相关性，与同样考虑相关性的LM相比，误差减小约15%以上，具有优越的缺失数据修复性能，且能适用于不同的数据集，具有很好的稳定性。在MCL的基础上，进一步利用残差神经网络精细去噪后，本文方法能进一步减小数据中的噪声和初步修复误差，大幅提升修复效果。且相比于不含残差连接的MCL-CNN方法，本文方法将数据修复误差降低了1/3以上，由此说明残差神经网络的引入可进一步提升数据修复性能。

上述结果表明，本文所提方法能实现准确可靠的数据修复，即使在高缺失率、高噪声情况下也能充分挖掘数据内部信息，保证数据集的完整性与可靠性，实现高精度的数据修复。

3.3　风电功率预测性能验证

数据修复的目的是为后续风电功率预测提供有力的数据支撑，因此需进一步验证使用修复后的数据进行短期风电功率预测时的预测性能。

分别对添加不同噪声、不同缺失率的2座风电场SCADA实测数据和NWP数据进行修复后，利用2.2节构建的CNN-LSTM-A风电功率预测模型预测未来3天的风电功率。选择数据集中前80%的数据进行训练，后20%的数据进行测试。为充分验证本文数据修复方法对风电功率预测模型性能的影响，首先利用未添加噪声、无缺失的理想工况测量数据对本文所提预测模型进行测试，以此作为模型预测性能的比较基准。在相同测试数据集下，对场站预测模型（黑箱形式，模型类型为BP神经网络）、未添加多头注意力机制的CNN-LSTM模型（记为CNN-LSTM）及CNN、LSTM、SVM等其他典型机器学习算法的预测结果进行统计分析，结果如表2所示。由表可知，在理想测量工况下，CNN-LSTM-A相较于场站实际预测模型及其他方法可显著提升功率预测性能，其E'_{MAE}和E'_{RMAE}均能维持在3%以内。与CNN-LSTM模型相比，本文模型可将E'_{MAE}和E'_{RMSE}降低0.25%左右，由此说明多头注意力机制有助于进一步引导风电功率预测模型对关键位置和特征的关注，有效提升功率预测精度。

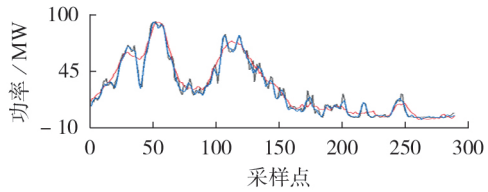
表2　风电功率预测误差统计

Table 2　Statistics of wind power prediction errors

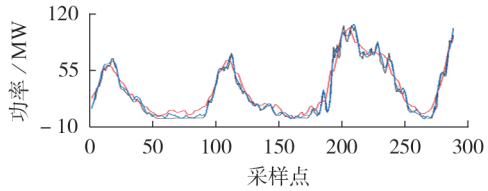
风电场	预测方法	E'_{MAE}	E'_{RMAE}
1号	场站预测	5.54	8.81
	CNN-LSTM-A	1.63	2.36
	CNN-LSTM	1.89	2.60
	CNN	2.68	3.79
	LSTM	2.97	4.04
	SVM	3.15	4.31
2号	场站预测	21.3	27.3
	CNN-LSTM-A	1.79	2.64
	CNN-LSTM	1.93	3.06
	CNN	3.41	4.89
	LSTM	2.54	3.77
	SVM	3.37	4.70

在上述测试基础上，利用添加噪声和数据缺失的测试集对非理想工况下的风电功率预测性能进行进一步测试。为直观展示本文所提模型的功率预测效果，在噪声强度为2.5%、缺失率为50%的极端工况下，从2座风电场的测试集中随机抽取单个测试样本，对未来3天内的风电功率预测结果进行分析，并与使用修复前的数据预测得到的风功率曲线对比，结果如图4所示。由图可见，在高噪声水平、高缺失率的情况下，若直接使用修复前的数据（默认将缺失值置0处理）进行风电功率预测，则所得到的预测曲线整体上可跟随场站实际功率曲线的变化趋势，但局部时段偏差较大。经本文数据修复方法处理后，CNN-LSTM-A预测模型展现出更优的预测性能，其功率预测曲线高度接近场站实发功率曲线，仅在个别数据点存在细微偏差。

对分别使用修复前后的数据所得到的风电功率预测误差进行统计分析，相应结果如表3所示。由



（a）1 号风电场风电功率预测曲线



（b）2 号风电场风电功率预测曲线

—— 实发功率，—— 修复前预测功率，—— 修复后预测功率

图 4　2 座风电场某测试样本的风电功率预测曲线

Fig.4　Wind power prediction curves for a test sample from two wind farms

表 3　数据修复前、后风电功率预测误差对比

Table 3　Comparison of wind power prediction errors between before and after data correction

风电场	数据类别	E'_{MAE}	E'_{RMAE}
1 号	修复前	3.13	4.45
	修复后	1.69	2.46
2 号	修复前	3.83	5.23
	修复后	2.00	2.88

表可知，使用修复后的数据进行风电功率预测可大幅降低预测误差，由此说明本文数据修复方法在保障和提升风电功率预测精度方面的潜在价值。

进一步，在噪声强度为 0.5 %~2.5 %、缺失率为 50 % 的工况下，利用测试集中所有样本对 2 座风电场的风电功率预测误差进行更为广泛的统计分析，结果如附录 A 图 A6 所示。由图可见，在不同噪声水平下，经数据修复后，CNN-LSTM-A 预测模型可将整体功率预测误差保持在较低水平。与表 3 中未添加噪声、无数据缺失的理想工况相比，虽然噪声、数据缺失工况下的预测误差有所增大，但整体误差增长不超过 6 %。由此说明，本文所提数据修复方法可为风电功率预测提供高质量数据源，有效减轻数据缺失、噪声等不良工况对功率预测模型的影响，提升功率预测模型对实际非理想测量工况的适应性。

4　结论

针对风电场 SCADA 实测数据中的数据缺失、噪声等非理想测量工况给短期风电功率预测等数据驱动的应用所带来的不良影响，本文提出一种综合挖掘数据内部自相关性和互相关性的 SCADA 数据高精度修复方案。实测数据的算例分析结果表明，相比传统数据修复方法，本文所提方法可深度挖掘学习数据内部相关性，有效提高缺失数据填补修复精度，并实现有效的噪声滤除，进一步提高整体数据质量。以修复后的 SCADA 数据为输入，所构建的短期风电功率预测模型可实现更加可靠的功率预测，相比于实际场站的预测结果，预测精度得到显著提升。

本文所提的风电场 SCADA 数据修复方法具有较强的通用性，在用于提升短期风电功率预测精度的同时，还可在数据层面为风机故障诊断、风电场经济运行与稳定控制等应用提供重要保障。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] 王海鑫，刘铭崎，董鹤楠，等. 含高比例新能源的电力系统低频振荡分析与抑制综述[J]. 电力自动化设备，2023，43(9)：152-163.
WANG Haixin，LIU Mingqi，DONG Henan，et al. Review on analysis and suppression of low-frequency oscillation in power system with high penetration of renewable energy sources[J]. Electric Power Automation Equipment，2023，43(9)：152-163.

[2] 彭光博，向月，陈文淑乐，等. “双碳”目标下电力系统风电装机与投资发展动力学推演及分析[J]. 电力自动化设备，2022，42(11)：70-77.
PENG Guangbo，XIANG Yue，CHEN Wenxule，et al. Kinetic deduction and analysis of installed capacity and investment development for wind power in power system under “dual carbon” target[J]. Electric Power Automation Equipment，2022，42(11)：70-77.

[3] 杨京渝，罗隆福，阳同光，等. 基于气象特征挖掘和改进深度学习模型的风电功率短期预测[J]. 电力自动化设备，2023，43(3)：110-116.
YANG Jingyu，LUO Longfu，YANG Tongguang，et al. Wind power short-term forecasting based on meteorological feature exploring and improved deep learning model[J]. Electric Power Automation Equipment，2023，43(3)：110-116.

[4] 吴永斌，张建忠，袁正舸，等. 风电场风功率异常数据识别与清洗研究综述[J]. 电网技术，2023，47(6)：2367-2380.
WU Yongbin，ZHANG Jianzhong，YUAN Zhengxi，et al. Review on identification and cleaning of abnormal wind power data for wind farms[J]. Power System Technology，2023，47(6)：2367-2380.

[5] 王一棠，庞勇，张立勇，等. 面向盾构机不完整数据的模糊聚类与非线性回归填补[J]. 机械工程学报，2023，59(12)：28-37.
WANG Yitang，PANG Yong，ZHANG Liyong，et al. Fuzzy clustering and nonlinear regression imputation for incomplete data of tunnel boring machine[J]. Journal of Mechanical Engineering，2023，59(12)：28-37.

[6] 冯磊，王石刚，梁庆华. 基于 GAKNN 方法的配电站时间序列缺失数据补全方法[J]. 电力自动化设备，2021，41(12)：187-192.
FENG Lei，WANG Shigang，LIANG Qinghua. Completion method for missing time series data of distribution station based on GAKNN method[J]. Electric Power Automation Equipment，2021，41(12)：187-192.

[7] TANG F，ISHWARAN H. Random forest missing data algorithms[J]. Statistical Analysis and Data Mining：the ASA Data Science Journal，2017，10(6)：363-377.

[8] PISNER D A，SCHNYER D M. Machine learning[M]. Amsterdam，Netherlands：Elsevier，2020：101-121.

[9] SÁNCHEZ C N，ENRÍQUEZ-ZÁRATE J，VELÁZQUEZ R，et al. Analysis of wind missing data for wind farms in Isthmus of Tehuantepec[C]//2018 IEEE International Autumn Meeting

on Power, Electronics and Computing. Ixtapa, Mexico: IEEE, 2018: 1-6.
[10] 潘立强，李建中，骆吉洲. 传感器网络中一种基于时空相关性的缺失值估计算法[J]. 计算机学报，2010，33(1)：1-11.
PAN Liqiang, LI Jianzhong, LUO Jizhou. A temporal and spatial correlation based missing values imputation algorithm in wireless sensor networks[J]. Chinese Journal of Computers, 2010, 33(1): 1-11.
[11] REN X J, SUG H, LEE H. A new estimation model for wireless sensor networks based on the spatial-temporal correlation analysis[J]. Journal of Information and Communication Convergence Engineering, 2015, 13(2): 105-112.
[12] ATWA W, BAHGAT A, REFAIE M. Simple missing data estimation algorithm in WSN based on spatial and temporal correlation[J]. International Journal of Computers and Information, 2020, 7(1): 42-54.
[13] YANG W J, ZHAO Y, WANG D, et al. Using principal components analysis and IDW interpolation to determine spatial and temporal changes of surface water quality of Xin'anjiang river in Huangshan, China[J]. International Journal of Environmental Research and Public Health, 2020, 17(8): 2942.
[14] SMYL S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting[J]. International Journal of Forecasting, 2020, 36(1): 75-85.
[15] JAIN A, NAGAR S, SINGH P K, et al. EMUCF: enhanced multistage user-based collaborative filtering through non-linear similarity for recommendation systems[J]. Expert Systems with Applications, 2020, 161: 113724.
[16] AJAEGBU C. An optimized item-based collaborative filtering algorithm[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(12): 10629-10636.
[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.
[18] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1026-1034.
[19] 徐钽，谢开贵，王宇，等. 基于TCN-Wpsformer混合模型的超短期风电功率预测[J]. 电力自动化设备，2024，44(8)：54-61.
XU Tan, XIE Kaigui, WANG Yu, et al. Ultra-short-term wind power forecasting based on TCN-Wpsformer hybrid model[J]. Electric Power Automation Equipment, 2024, 44(8): 54-61.
[20] 林大略. 噪声环境下的数据驱动进化算法[D]. 广州：华南理工大学，2022.
LIN Dalue. Data-driven evolutionary algorithms in noisy environments[D]. Guangzhou: South China University of Technology, 2022.
[21] 国家市场监督管理总局，中国国家标准化管理委员会. 调度侧风电或光伏功率预测系统技术要求：GB/T 40607—2021[S]. 北京：中国标准出版社，2021.

作者简介：

郑李梦千(2000—)，男，硕士研究生，主要研究方向为风电功率预测(**E-mail**：a1479010379@163.com)；

朱利鹏(1990—)，男，教授，博士，主要研究方向为电力系统稳定分析和控制等(**E-mail**：zhulpwhu@126.com)。

(实习编辑　丁欣欣)

Multiple correlation learning-based wind farm SCADA data correction and its application in wind power prediction

ZHENG Limengqian[1], ZHU Lipeng[1], WEN Weijia[2], LI Jiayong[1], ZHANG Cong[1]

(1. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. Information and Telecommunication Branch of State Grid Hunan Electric Power Co., Ltd., Changsha 410004, China)

Abstract: The non-ideal measurement conditions such as data missing and noise in the measurement data of wind farm supervisory control and data acquisition (SCADA) system bring serious challenges to the reliable short-term prediction of wind power. To address this problem, a SCADA data correction scheme based on multiple correlation learning is proposed. Aiming at the problem of missing data issue in the measured SCADA data, a data recovery me-thod of comprehensively mining multi-correlation for multi-dimensional time series data is proposed to preliminarily correct the missing data. A residual neural network adapted to variety complicated operating conditions is designed to further refine the preliminary recovery results, thereby realizing fine missing value correction and data denoising. With the corrected SCADA data taken as inputs, a highly reliable short-term wind power prediction model is constructed via convolutional neural network-long short-term memory deep learning network based on multi-head attention mechanism. Numerical analysis results with field SCADA data obtained from two real-world wind farms in Central China verify the effectiveness of proposed method and its application value in enhancing the performance of short-term wind power prediction.

Key words: SCADA data correction; multiple correlation; short-term wind power prediction; deep learning; residual neural network

附录 A

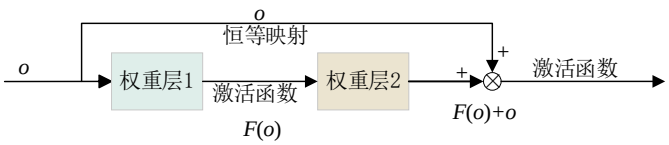
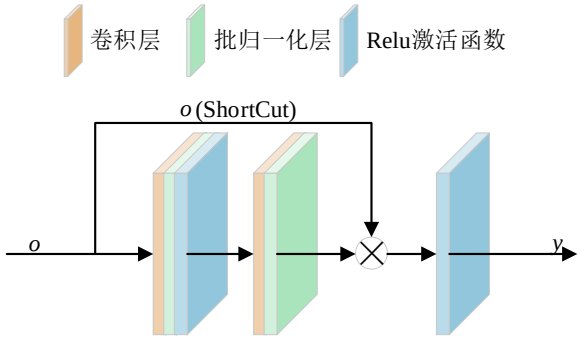
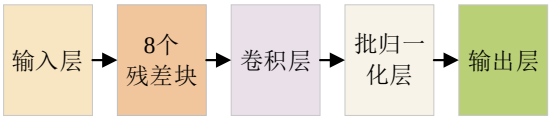


图 A1　残差连接示意图

Fig.A1　Schematic diagram of residual connection



(a) 残差块示意图



(b) 网络整体结构图

图 A2　残差神经网络示意图

Fig.A2　Schematic diagram of residual deep neural network

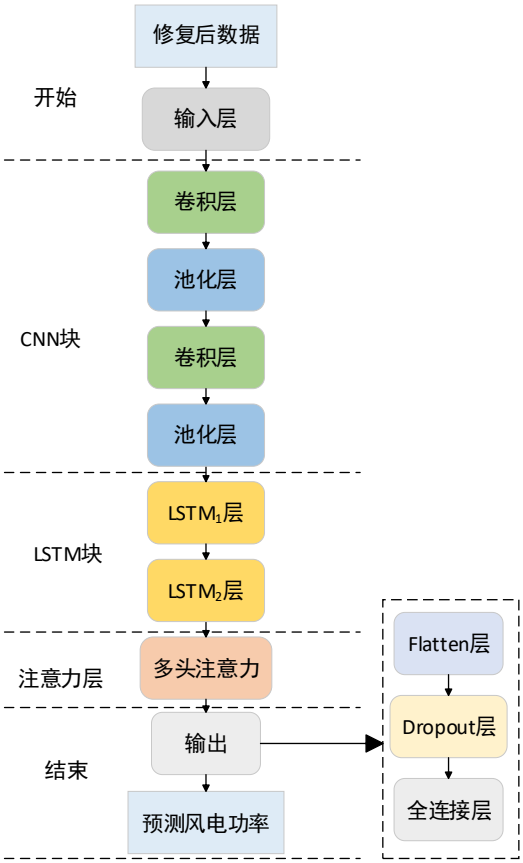


图 A3　CNN-LSTM-A 风电功率预测模型

Fig.A3 Wind power prediction model of CNN-LSTM-A

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{A1}$$

式中：X 为原始数据，$X_{\max}$ 和 $X_{\min}$ 是原始数据中的最大值和最小值，X_{scaled} 为归一化后的数据。

表 A1　残差神经网络超参数设置

Table A1　Hyperparameter setting of proposed residual neural network

层类型	超参数
残差块内卷积层	输出通道数：64 卷积核大小：3 步幅：1
批归一化层	稳定性常数：1e-3 中心化：True 缩放：True
残差块外卷积层	输出通道数：1 卷积核大小：3 步幅：1

$$E_{\mathrm{MAE}}=\frac{1}{n}\sum_{s=1}^{n}\left|y_s-\hat{y}_s\right| \tag{A2}$$

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{s=1}^{n}\left(y_s-\hat{y}_s\right)^2} \tag{A3}$$

式中：y_s 和 $\hat{y}_s$ 分别为第 s 个样本的真实值和修复后的预测值。E_{MAE} 和 E_{RMSE} 越小表示数据修复效果越好，数据集与真实值之间的误差越小。

$$E_{\mathrm{MAE}}^{'}=\frac{1}{N}\sum_{o=1}^{N}\left|\frac{p_o-\hat{p}_o}{C_s}\right|\times 100\% \tag{A4}$$

$$E_{\mathrm{RMSE}}^{'}=\sqrt{\frac{1}{N}\sum_{o=1}^{N}\left(\frac{p_o-\hat{p}_o}{C_o}\right)^2}\times 100\% \tag{A5}$$

式中：N 为用于风电功率预测的样本数；p_o 和 $\hat{p}_o$ 分别为第 o 个样本的真实值和预测值；$E_{\mathrm{MAE}}^{'}$ 和 $E_{\mathrm{RMSE}}^{'}$ 分别为经风电场开机容量规范化后的平均绝对误差和均方根误差；C_o 为第 o 个样本的开机容量。

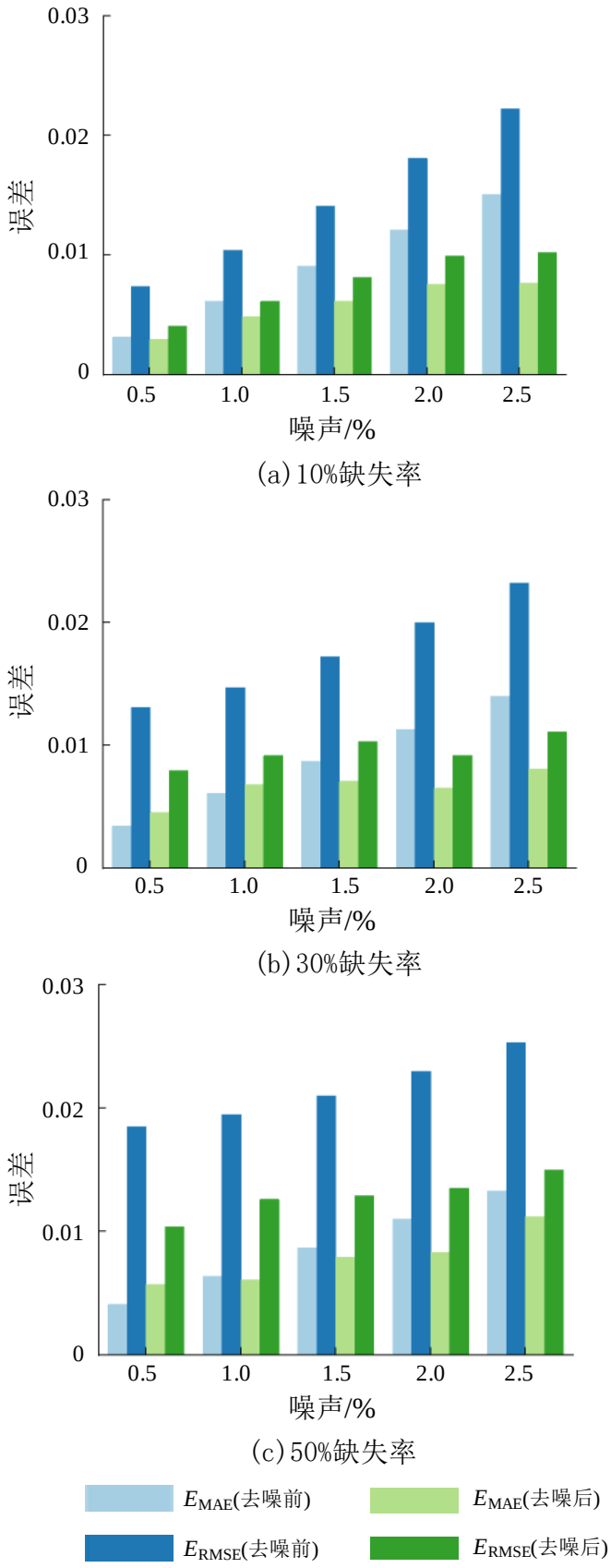


图 A4　1 号风电场整体 SCADA 数据修复误差统计

Fig.A4　Error statistics of overall SCADA data correction of No.1 wind farm

图中：噪声与误差均为标幺值，下同。

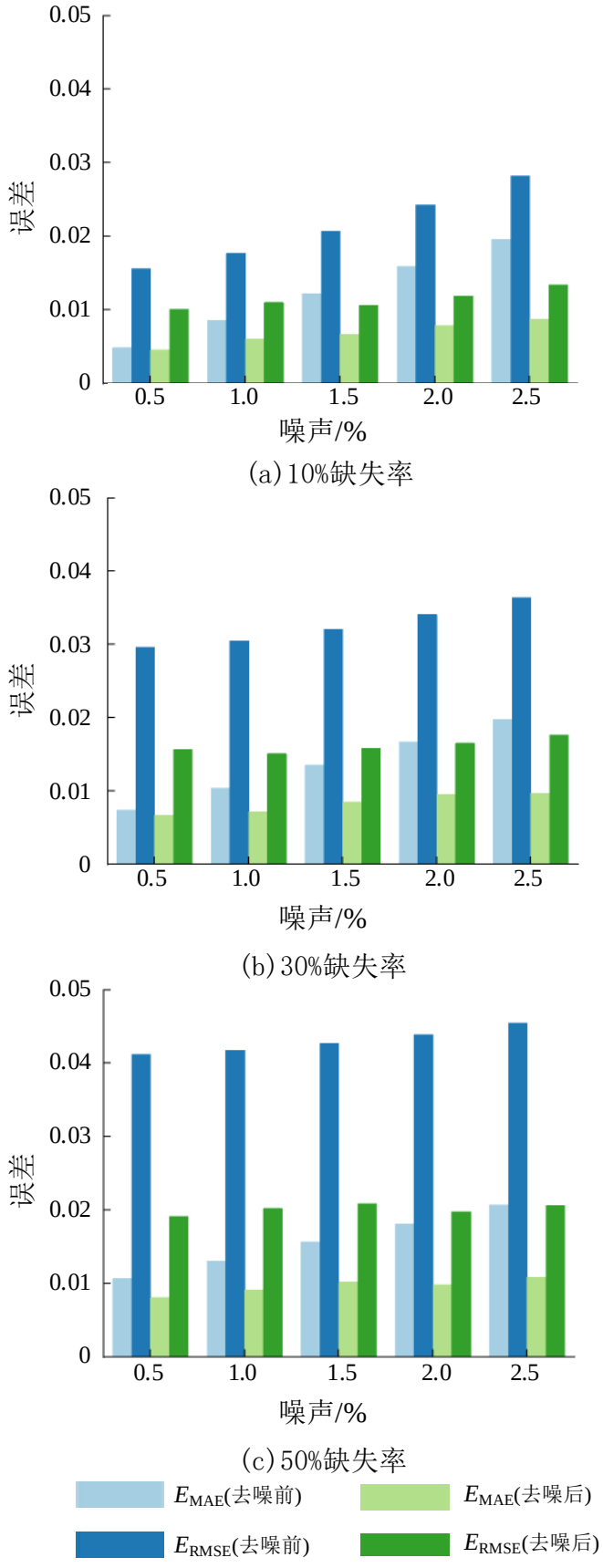
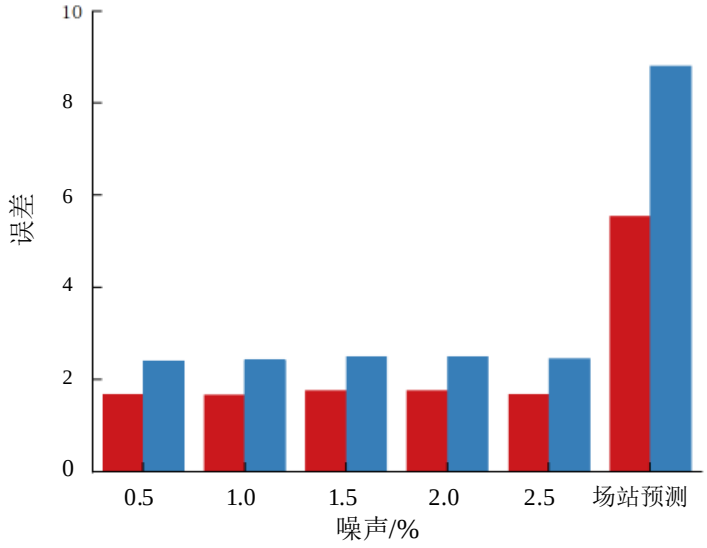
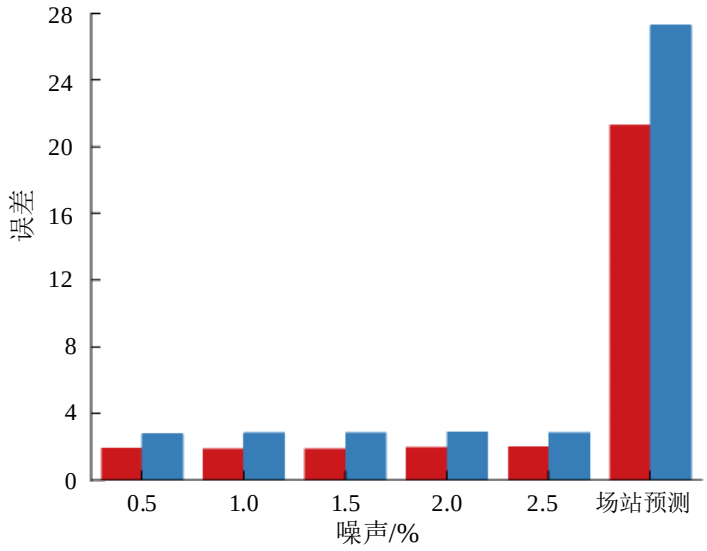


图 A5　2 号风电场整体 SCADA 数据修复误差统计

Fig.A5　Error statistics of overall SCADA data correction of No.2 wind farm



(a) 1号风电场预测误差统计图



(b) 2号风电场预测误差统计图

图 A6　两所风电场的风电功率预测误差统计

Fig.A6　Statistics of wind power prediction errors from two wind farms