DOI:10.11737/j.issn.1003-0239.2024.05.008

基于RGCN-SA算法的海上浮标观测数据插补

彭德东[1,2]，梁建峰[1*]，崔学荣[2]，岳心阳[1]

（1.国家海洋信息中心，天津 300171；2.中国石油大学（华东）海洋与空间信息学院，山东 青岛 266580）

摘 要： 针对海洋观测数据的缺失问题，提出一种基于图卷积（GCN）和自注意力机制（SA）的残差网络插补模型（RGCN-SA），该模型由自注意力机制与图卷积构建，利用自注意力机制提取观测数据的时间依赖特征，通过图卷积获取不同位置浮标的空间依赖特征，并添加残差结构提高模型学习能力，结合自监督训练方式对模型进行训练，得到最终的海洋浮标数据插补模型。通过对比实验，证明该模型通过训练后能够有效获取浮标观测数据的时间与空间的关联特征，取得了比其他方法更好的插补效果。通过消融实验，证明模型的各个模块的有效性。

关键词： 自注意力机制；图卷积网络；插补；浮标数据

中图分类号：P731.31 **文献标识码：**A **文章编号：**1003-0239（2024）05-0077-12

0 引言

随着海洋自动化观测仪器增多，海洋观测数据量呈几何级上涨[1]，大量的海洋观测数据对于研究和预测海洋环境变化、气候变化、海洋灾害以及海洋资源开发等具有非常重要的作用[2]。例如，通过浮标数据可以获得海水温度、盐度、流速等水文学数据，这对于预测海洋环境变化和气候变化趋势非常关键；同时，还可以获得风向、风速、气温等气象学数据，这可以帮助船只进行航行安全管理和港口规划；此外，观测数据还可以为海洋生态环境建模提供相关数据支持。总之，海洋数据在海洋科学和气象学领域中扮演着重要的角色，也为人类近海活动提供重要信息，促进了海洋学的研究和应用。

在海上浮标采集海洋数据的过程中，受海洋环境干扰和传感器故障影响，数据无法完整采集，或在采集后受网络波动影响，部分数据丢失。海洋数据缺失成为相对普遍的问题，给数据挖掘、数据分析以及应用造成困难。因此，如何对缺失值进行处理是一个重要的研究内容。

缺失值处理最简单的方法是直接删除法，即直接将存在缺失值的行或列删除。这种方法处理较为简单，但缺点是如果缺失数据较多，不仅会造成数据集规模减小，而且会对下游的应用造成较大障碍[3]，因此，采用可能的值来对缺失值进行填充会减少这种影响。当前国内外研究人员提出了一些有效的缺失数据填充方法，包括传统插补方法、机器学习插补算法和深度学习插补算法。

传统插补方法包括KANTARDZIC等[4]采用的均值填充法（MEAN）、零值填充法、众数填充法和最后观察值前向填补法（Last Observation Carried Forward，LOCF）[5]，其中均值填充法使用未缺失数据行或列的平均值来做填充值。这些传统插补算法虽然可以在某些情况下取得较好的结果，但大多会造成较大误差。YANG等[6]采用回归模型对缺失值进行插补，根据数据间的关联构建回归模型，并根据数据集来确定模型参数，这个方法能够利用原始数据集的一些隐含特征，在一定程度上提高填充值的准确度。在机器学习插补算法方面，LI等[7]采

收稿日期：2023-09-18。
基金项目：国家重点研发计划（2021YFC3101600）。
作者简介：彭德东（1998-），男，硕士在读，主要从事海洋环境预报工作。E-mail: 908869457@qq.com
*通信作者：梁建峰（1983-），男，正高级工程师，硕士，主要从事海洋资料处理工作。E-mail: liangjianfeng@nmdis.org.cn

用K邻近插补法(K-Nearest Neighbors,KNN)对缺失数据进行补全,即利用缺失值最近的K个邻近值的平均值来填充缺失值。为了提高填充效果,KNN也经常和其他算法结合。例如,AL-HELALI等[8]采用一种新的遗传规划算法与KNN算法结合对缺失数据进行补全,提高了填充效果;HASTIE等[9]基于矩阵分解的填充法,将数据样本当作矩阵,采用两个矩阵的乘积表示原始数据样本的矩阵,即将原始数据矩阵分解成两个低秩矩阵的乘积,得到一个近似矩阵,用于填充缺失值;JING等[10]采用随机森林填充算法并与链式方程相结合对水文气象数据进行填充,随机森林通过回归操作可对包含缺失值的节点进行预测填充,而链式方程是多重回归插补方法(Multiple Imputation by Chained Equations,MICE),两者的结合取得了较好的插补效果。随着神经网络的快速发展,大量的深度学习插补算法被应用于缺失值填充上。PAN等[11]结合多层感知机与动力梯度下降对缺失数据进行填补。CAO等[12]采用具有双向循环神经网络的时间序列插补算法(Bidirectional Recurrent Imputation for Time Series,BRITS)并加入衰减结果对数据进行插补,能够进一步提取存在缺失值的时间序列数据间的依赖关系,提高对时间序列的插补效果。对抗生成网络作为一种数据生成网络,近年来被用于数据填充,YOON等[13]提出了一种改进的对抗生成网络对缺失数据进行填充。DU等[14]采用基于自注意力机制(Self-Attention Mechanism,SA)[15]的时间序列插补模型SAITS(Self-Attention-based Imputation for Time Series)并结合联合优化的训练方式,使模型能够捕捉时间依赖性和特征相关性,提高了插补精度和训练速度。

海上观测数据具有多要素的特点[16],即不仅呈现出时间序列特性,也包含空间上的关联性。因此,本文提出一种基于自注意力机制与图卷积(Graph Convolution Network,GCN)的海上浮标观测数据插补模型,两者分别用于提取时间与空间依赖特征,结合残差结构并采用自监督的方式进行训练,得到最终的自注意力机制的残差网络插补模型RGCN - SA(Residual Network Imputation Model Based on Graph Convolution Network and Self - attention Mechanism)。

1 研究方法与模型构建

1.1 方法介绍

1.1.1 自注意力机制

自注意力机制是一种用于序列数据处理的机制,它能够在同一序列中不同位置的元素之间建立联系。它能将输入的序列数据中的每个元素与其他元素进行比较,从而为每个元素赋予一个权重,表示它在序列中的重要性或者与其他元素的相关程度。自注意力层包含3个权重矩阵,可将输入序列$\boldsymbol{X}$映射为3个不同的向量查询向量、键向量和数值向量。3个向量经过自注意力机制运算后得到输出矩阵。运算过程公式为:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q \\ \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K \\ \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V \\ \mathrm{SA} = \mathrm{Softmax}\left(\dfrac{\boldsymbol{Q}\boldsymbol{K}^T}{d_k}\right)\boldsymbol{V} \end{cases} \tag{1}$$

式中:$\boldsymbol{X}$为输入矩阵;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为查询向量、键向量、数值向量;$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$为其对应的权重系数矩阵;SA为输出;Softmax为归一化激活函数,即将实数向量归一化为概率分布向量;d_k为$\boldsymbol{K}$的维度代表缩放注意力分数。自注意力机制结构见图1。

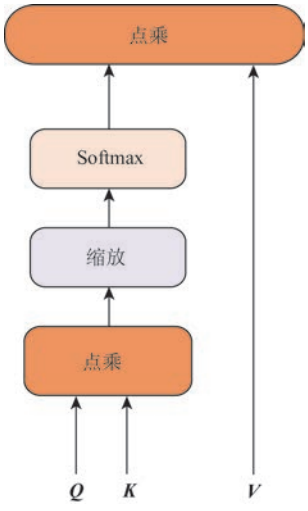


图1 自注意力机制

Fig.1 Self-attention mechanism

多头自注意力机制对自注意力机制进行了改进,从而产生多个不同的“头”,每个“头”都可以关注不同的部分。每个“头”在计算注意力权重的过程中学习一个表示特征的映射空间,并在最终输出时对这些映射空间进行拼接。计算公式为:

$$\begin{cases} h_i = SA_i \\ \mathrm{MHA} = \mathrm{Concat}(h_1, h_2, \cdots h_i \cdots, h_n)\boldsymbol{W} \end{cases} \tag{2}$$

式中:h_i、SA_i为第i个自注意力机制的输出;MHA为多头自注意力机制的输出;Concat为拼接函数,即将不同序列合并为同一个序列;n代表头的数量;$\boldsymbol{W}$为权重。多头自注意力机制结构见图2。

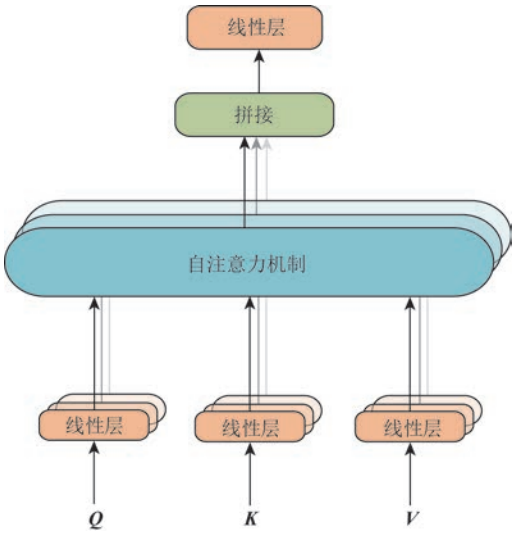


图2　多头自注意力机制

Fig.2　Multi-head self-attention mechanism

1.1.2　位置编码

自注意力机制没有序列先后顺序的结构,自身无法确定输入数据的位置信息。因此模型(model)需要使用位置编码来对确定序列的顺序。位置编码计算公式为:

$$\begin{cases} \mathrm{Pos\_E}_{(\mathrm{pos},2i)} = \sin\left(\dfrac{\mathrm{pos}}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \\ \mathrm{Pos\_E}_{(\mathrm{pos},2i+1)} = \cos\left(\dfrac{\mathrm{pos}}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \end{cases} \tag{3}$$

式中:Pos_E为编码输出;pos代表时间序列中的元素位置;i表示位置编码向量中的维度编号;d_{model}代表model输入的维度。

1.1.3　前馈神经网络

前馈神经网络通过引入非线性映射和扩展隐藏层维度,可以使模型更好地建立复杂的非线性关系。其网络结构为先通过一个线性变换,将序列映射到更高维度的中间表示,通过非线性激活函数Relu得到非线性关系,再将激活后的结果映射回原始维度。结构公式为:

$$\begin{cases} X = \mathrm{Relu}(\boldsymbol{x}\boldsymbol{W}_1 + b_1) \\ \mathrm{FFN} = X\boldsymbol{W}_2 + b_2 \end{cases} \tag{4}$$

式中:X为第一个非线性输出;FFN为输出;Relu为非线性激活函数,即函数将所有负值置为0,正值不变;$\boldsymbol{x}$为输入矩阵;$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为线性变换的权重;b_1、b_2为偏置参数。

1.2　基于RGCN-SA的海上浮标观测数据的缺失数据插补模型

1.2.1　RGCN-SA插补模型

RGCN-SA插补模型由RGCN-SA模块组成,其模块包含Transformer[15]编码器和GCN网络。模型的输入结构经过Transformer结构提取观测数据时间依赖特征,经过GCN结构提取空间依赖特征,再将两者进行特征融合。由于这里空间结构相对简单,得到的信息有限,并且时间依赖特征更为突出,因此在时空依赖特征融合之后再添加一个Transformer模块,进一步提取时间依赖特征。为了促进信息的流动和保持模型的表达能力,在输入和特征融合之间中添加残差结构,它是一种在模型中引入跳跃连接的技术,允许信息直接从一个层级传递到后续层级,有助于解决梯度消失问题和加速模型收敛,确保时间和空间依赖特征的有效传递和整合。最终,模型经过线性层得到完整的插补结果。模型结构见图3。

1.2.2　Transformer编码器时间依赖特征提取

Transformer编码器是一种用于处理序列数据的神经网络结构,广泛用于序列建模任务。其每一层采用多头自注意力机制与前馈神经网络组成的残差结构,Transformer编码器通过将海上浮标观测数据转化为固定维度的向量,并添加位置编码,然后经过多头自注意力机制对观测数据转化的向量进行全局性的建模,使模型能够捕捉海洋数据时间

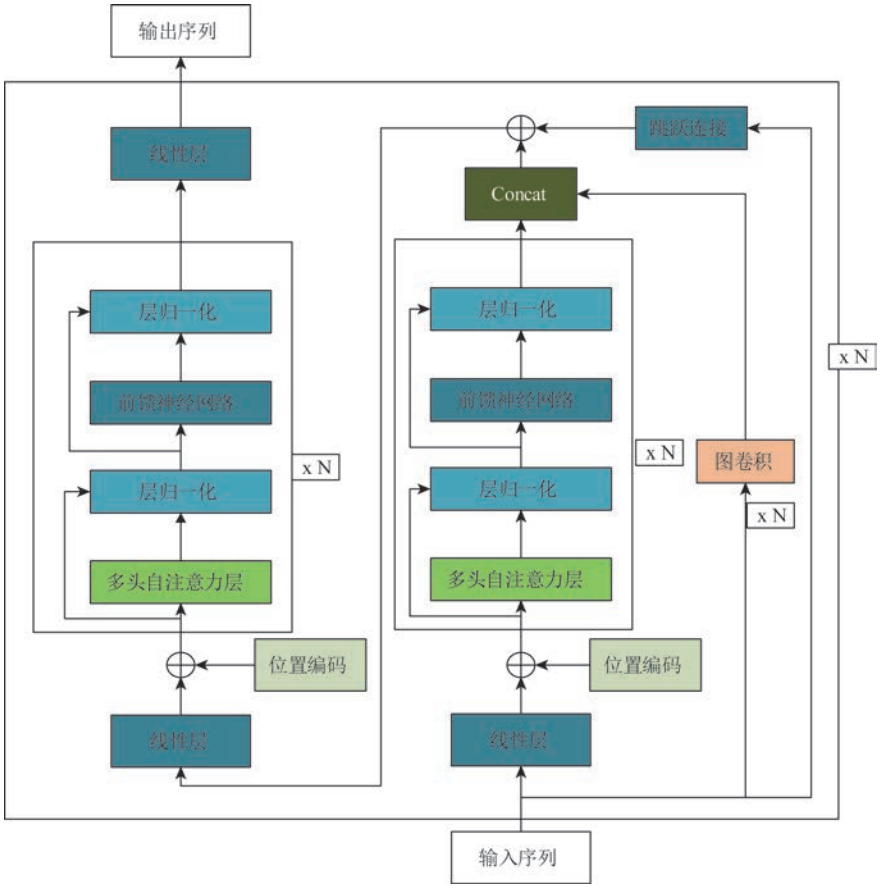


图3 RGCN-SA模型

Fig.3 RGCN-SA model

序列中的长距离依赖关系，在自注意力机制之后，经过前馈神经网络进行非线性变换，帮助模型引入更强的非线性建模关系。另外，Transformer编码器结构引入残差结构和层归一化，即在自注意力机制和前馈神经网络之后，将其输出进行层归一化并与其输入相加。

1.2.3 GCN空间依赖特征提取

GCN是一种图神经网络，其核心是图卷积层。它通过邻接矩阵对海上各个浮标节点进行聚合和更新，每个浮标节点的特征由浮标节点自身特征和相邻浮标节点的特征组成，通过对相邻浮标节点的特征进行加权聚合来更新节点特征。通过这种机制，GCN能够有效地表达节点与其相邻点之间的空间依赖关系，实现对局部特征模式的建模。同时，GCN采用参数共享的方式使节点共享相同的卷积核权重，可以降低模型的复杂性和过拟合的风险。GCN通过堆叠多个图卷积层来构建深层网络，每个图卷积层都可以进一步丰富节点的特征表示，并通过多层的非线性变换提取更高级别的图特征。计算公式为：

$$H=\sigma\left(D^{-\frac{1}{2}}\boldsymbol{A}D^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}\right) \tag{5}$$

式中：H为输出；$\boldsymbol{A}$是包含自环的邻接矩阵；$\boldsymbol{D}$是度矩阵；$\boldsymbol{W}$为卷积核的权重；σ为激活函数，本文使用Relu激活函数；$\boldsymbol{X}$为特征矩阵。

本文的邻接矩阵$\boldsymbol{A}$的权重按照海上浮标之间的距离来进行计算，计算公式为：

$$A(u,v)=\begin{cases}\exp\left(-\dfrac{(\mathrm{dis}(u,v))^2}{\sigma^2}\right) & \mathrm{dis}(u,v)\leqslant \mathrm{th}\\ 0 & \text{其他情况}\end{cases} \tag{6}$$

式中：A为输出；u、v代表节点；$\mathrm{dis}(u,v)$代表节点u、v之间的距离；σ为标准差；th为阈值。

2　数据预处理与模型训练

2.1　数据介绍

本文采用国家海洋信息中心提供的5个海上浮标观测数据,这些浮标位于中国南海北部(112.3°～117.3°E,18.8°～24.8°N)。本文截取每个浮标2021年4月—2022年12月的历史观测数据。观测数据的采样间隔为30 min,每个浮标的数据(包含缺失值在内)共计29 504条,缺失率在4%～36%。每条数据包含了多种要素,本文对平均波高、平均波高周期、平均风速、有效波高、有效波高周期和最大风速进行模型训练并对其缺失数据进行插补和评估。浮标之间的距离见表1。研究区域与浮标位置分布情况见图4。

表1　各个浮标之间的距离(单位:km)

Tab.1　Distances between various buoys (unit:km)

浮标名称	浮标1	浮标2	浮标3	浮标4	浮标5
浮标1	—	148.364	334.887	153.937	504.669
浮标2	148.364	—	246.101	101.117	408.200
浮标3	334.887	246.101	—	181.391	169.816
浮标4	153.937	101.117	181.391	—	351.185
浮标5	504.669	408.202	169.816	351.185	—

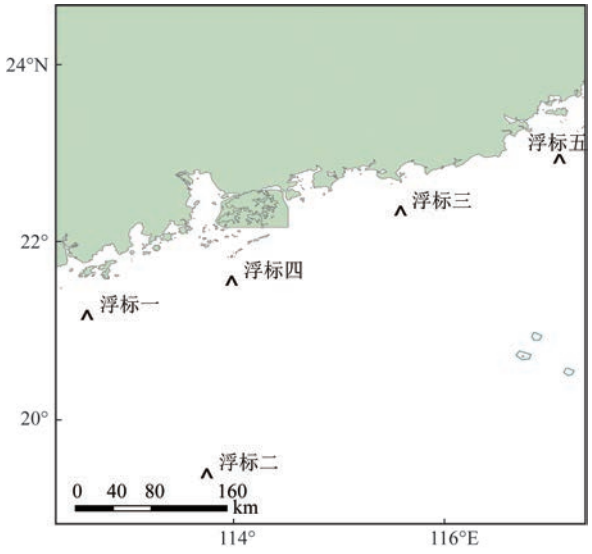


图4　研究区域

Fig.4　Research area

2.2　掩码矩阵

海上浮标采集的数据通常为间隔固定的时间序列数据。对于一段海洋观测数据S,假设数据的时间步长为t,对应的要素维度为d,则$S=\{s_1,s_2,s_3,\cdots,s_i,\cdots,s_{t-1},s_t\}$,第$i$时刻对应的观察要素$s_i=\{s_{i1},s_{i2},s_{i3},\cdots,s_{ij},\cdots,s_{i(d-1)},s_{id}\}$,海洋观测数据存在缺失值,构建一个与数据矩阵相对应掩码矩阵$\boldsymbol{M}_{td}$,s_{ij}为缺失值时对应的M_{ij}为0,否则为1。

2.3　自监督训练

由于自注意力机制并不是自回归结构,它可以同时访问输入序列中的所有位置,而不需要按顺序结构对序列进行处理。这在许多任务中是非常有用的,但也给训练带来了一些困难。在自注意力机制中,模型在处理每个位置时都可以看到整个输入序列,这使得模型难以学习到未观察到的值或缺失的数据。

为了使模型能够成功训练,模型采用自监督的方式引入人工掩码。该方法通过模拟自然缺失的方式,对输入数据中的未缺失值(即观察值)进行屏蔽,具体地,一定百分比的观察值被随机选择并屏蔽,这些被屏蔽的观察值在训练过程中对模型是不可见的。然而,这些被屏蔽的观察值作为标签来对模型的参数进行优化,从而使模型能够学会对未观察到的值进行推断和填充。采用掩码矩阵$\boldsymbol{I}$表示人工缺失值,即人为缺失值对应的掩码值为1,其他值对应的掩码值为0。针对海洋观测数据,引入20%的人工缺失率,在原始数据的基础上,将训练集中20%的观察值按照一定的随机方式进行屏蔽,其结构见图5。

模型的损失函数由观察值损失函数和人工缺失损失函数组成。公式为:

$$\begin{cases}\text{Loss1}=\dfrac{\sum_{k=1}^{n}|X-Y|\times\boldsymbol{M}}{\sum_{k=1}^{n}\boldsymbol{M}}\\[2ex]\text{Loss2}=\dfrac{\sum_{k=1}^{n}|X-Y|\times\boldsymbol{I}}{\sum_{k=1}^{n}\boldsymbol{I}}\\[2ex]\text{Loss}=\text{Loss1}+\text{Loss2}\end{cases}\tag{7}$$

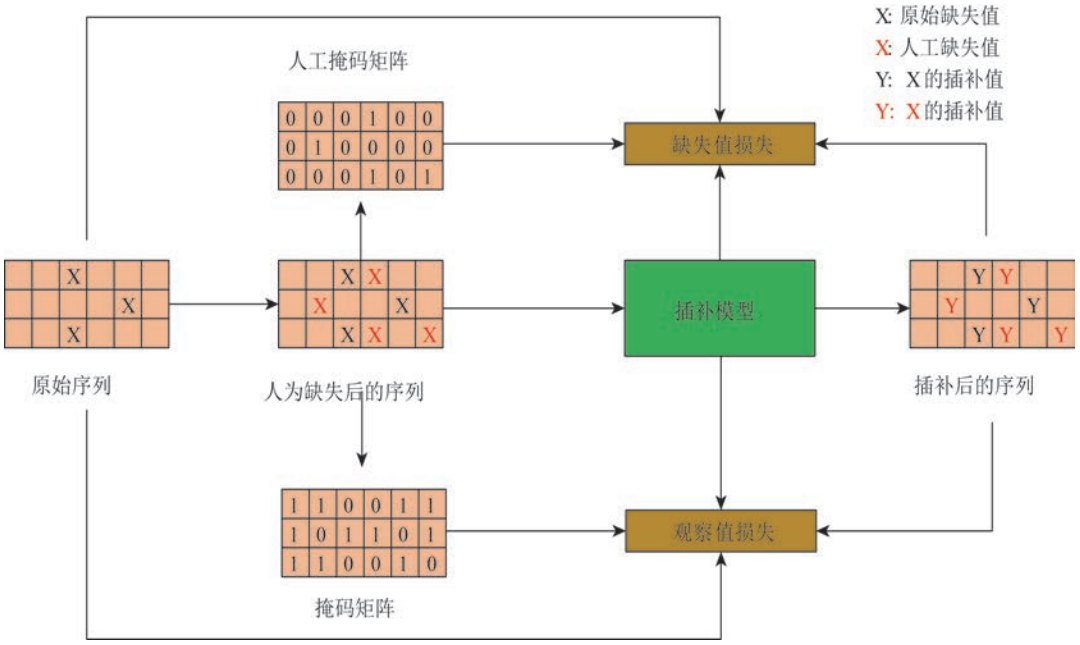


图5 训练方式

Fig.5 Training mode

式中:Loss1为观察值损失输出;X为原始序列;Y为插补后序列;M为掩码矩阵;Loss2为人工损失输出;I为人工掩码矩阵;n为数据的个数;Loss为模型的损失输出。

3 实验与分析

3.1 评价标准

基于RGCN-SA的海洋观测数据插补模型的精度,本文采用平均绝对误差(Mean Absolute Error,MAE)和均方根误差(Root Mean Squared Error,RMSE)对每个要素进行评估,其结果为实验数据观测设备每个要素的平均值。公式为:

$$E_{\mathrm{MAE}}=\frac{\sum_{i=1}^{n}\left|p-\mathrm{true}\right|\times\mathrm{mask}}{\sum_{i=1}^{n}\mathrm{mask}} \tag{8}$$

$$E_{\mathrm{RMSE}}=\sqrt{\frac{\sum_{i=1}^{n}\left((p-\mathrm{true})^{2}\times\mathrm{mask}\right)}{\sum_{i=1}^{n}\mathrm{mask}}} \tag{9}$$

式中:E_{MAE}、E_{RMSE}分别为MAE、RMSE的值;p为预估值;true为真实值;mask为掩码;n为数据个数。

3.2 实验设置

实验算法在CentOS 7.9操作系统上运行,编程采用python语言,模型搭建采用pytorch深度学习框架。实验数据经预处理后分为训练集、验证集和测试集,比例为7∶1∶2,验证集与测试集随机缺失一定比例,对结果进行评估。

3.3 对比试验

本文设置的对比算法包括7种。

MEAN:使用缺失数据所在特征的平均值来填补缺失值。

LOCF:缺失值使用最后一个可用的观察值来填补。

KNN:根据缺失数据点附近的K个已知数据点来估计缺失值。

MICE:使用统计方法生成多个填补的数据集,在这些数据集上进行分析并汇总结果。

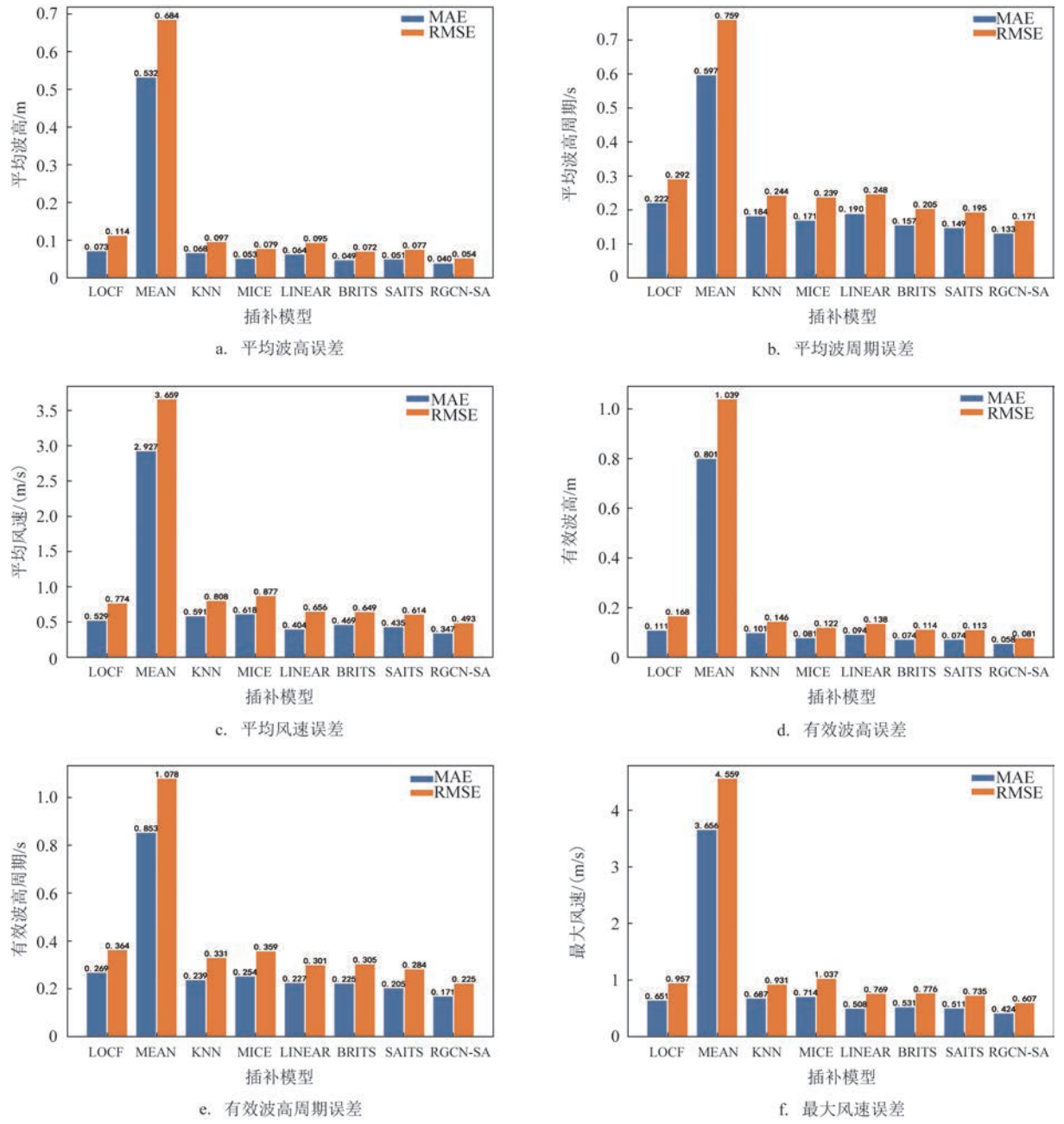
LINEAR(线性插值):根据已知数据点的线性关系来估计缺失数据点。

BRITS:一种带有衰减结构的双向循环神经网络插补算法。

SAITS:一种基于自注意力机制的插补算法。

将本文的算法RGCN-SA分别与上述7种算法进行对比。

在20%缺失率条件下(见图6),MEAN插补方法的MAE和RMSE都相对较高,无法取得能够使用的插补结果;LOCF插补法对平均波高、平均波高周期、平均波高与有效波高周期的插补效果低于机器学习模型与深度学习模型,对平均风速与最大风速上的插补效果强于部分机器学习模型而弱于LINEAR插补法和深度学习模型。LINEAR插补法对平均风速与最大风速的插补效果较强,而对其他参数的插补效果相对较弱。总的来说,深度学习模型在所有要素上的插补结果都具有较好的效果。与其他几种模型相比,本文提出的RGCN-SA具有最小的MAE与RMSE,插补结果相较于其他方法都



a. 平均波高误差　b. 平均波周期误差

c. 平均风速误差　d. 有效波高误差

e. 有效波高周期误差　f. 最大风速误差

图6　20%缺失率下的插补结果精度对比

Fig.6　Precision comparison of imputation results under 20% missing rate

有明显的提高，相较于SAITS，RGCN-SA平均波高、平均波高周期、平均风速、有效波高、有效波高周期和最大风速的MAE分别降低21.6%、10.7%、20.2%、21.6%、16.6%和17.0%，而RMSE分别降低29.9%、12.3%、19.7%、28.3%、20.8%和17.4%。

在30%缺失率条件下（见图7），KNN与MICE的插补结果相对较差，LINEAR、BRITS和SAITS依然保持相对较好的插补水平。在所有要素中，RGCN-SA依然保持最好的插补效果，其中，相较于SAITS，RGCN-SA平均波高、平均波高周期、平均风速、有效波高、有效波高周期和最大风速的MAE分别降低20.6%、11.1%、17.9%、16.6%、14.6%和16.8%，而RMSE分别降低25.8%、13.2%、17.0%、23.1%、15.6%和16.8%。

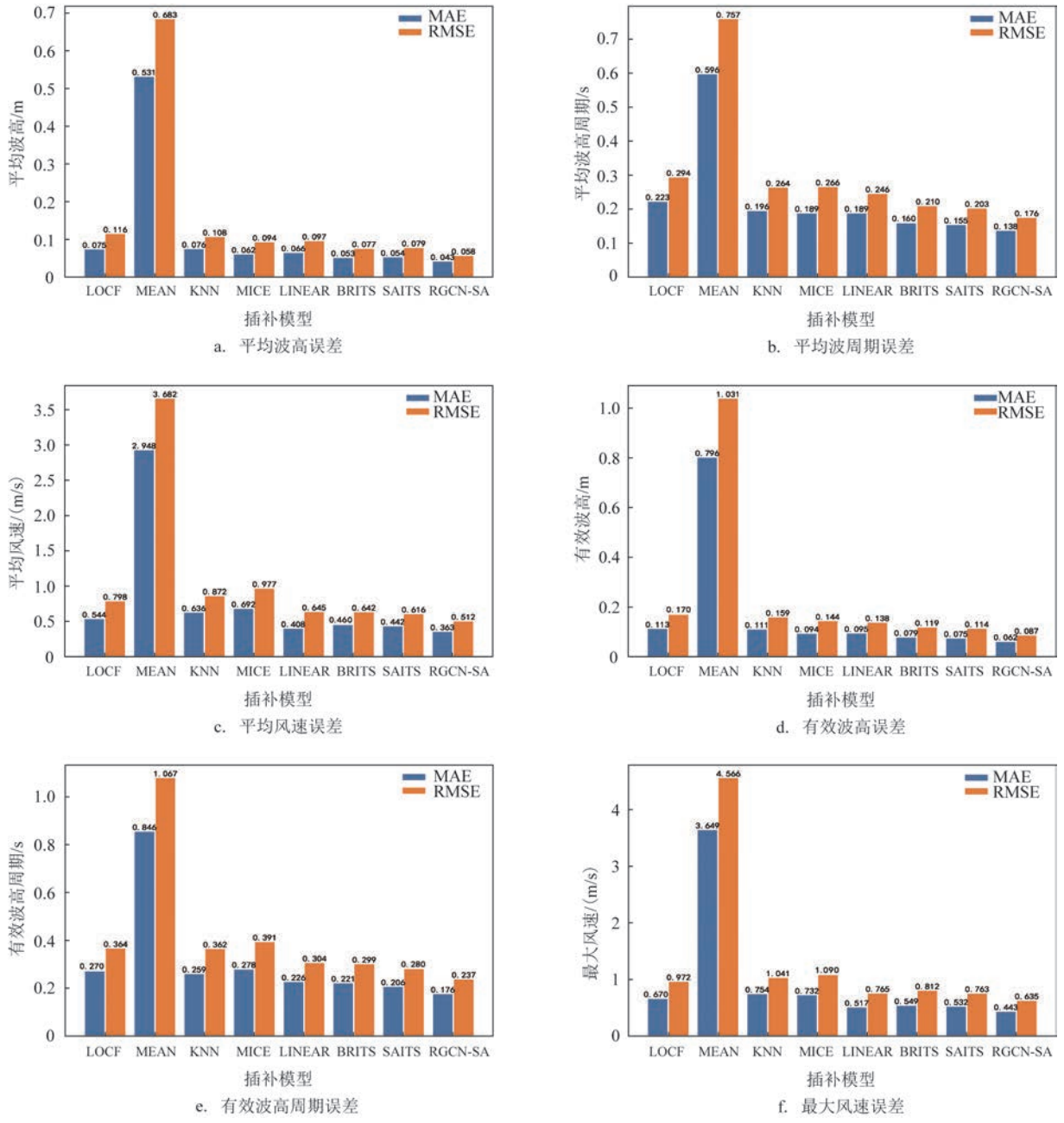


图7 30%缺失率下的插补结果精度对比

Fig.7 Precision comparison of imputation results under 30% missing rate

在50%缺失率条件下(见图8),KNN与MICE的插补误差很大,BRITS和SAITS虽然也有较好的效果,但部分要素的误差明显变大。RGCN-SA在所有要素中仍然具有较好的插补效果,其中,相较于SAITS,RGCN-SA平均波高、平均波高周期、平均风速、有效波高、有效波高周期和最大风速的MAE分别降低22.1%、11.8%、21.5%、19.3%、16.3%和18.0%,而RMSE分别降低24.8%、13.5%、21.7%、23.7%、18.1%和18.0%。

有效波高与最大风速是常用的海洋数据,故本文以浮标一的有效波高与最大风速为例,在50%的缺失率下,选取128步时的不同模型的插补数据与原始数据进行可视化对比。由于MEAN方法误差较大,这里不作对比,可视化结果见图9。从图中可

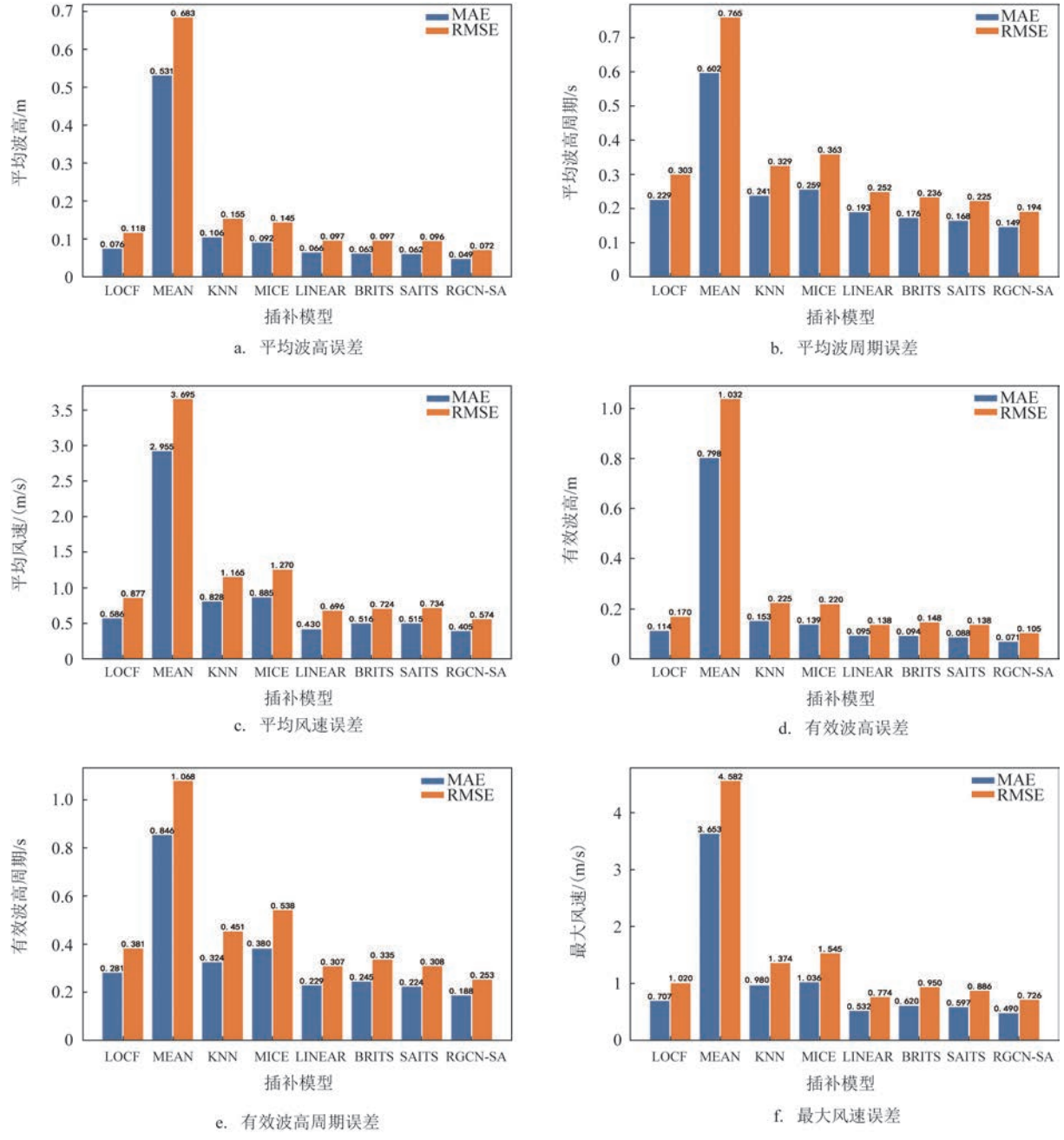


图8　50%缺失率下的插补结果精度对比

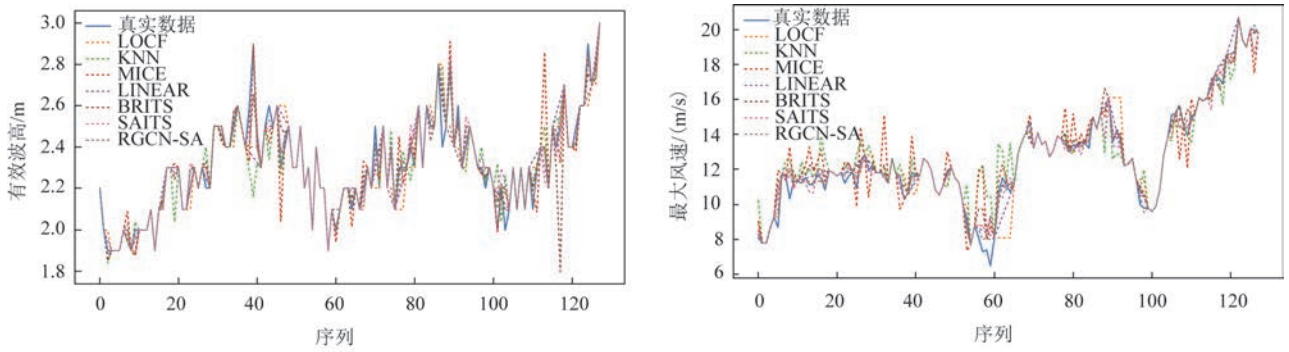
Fig.8　Precision comparison of imputation results under 50% missing rate

以看出，在个别位置点，RGCN-SA的插补误差较某些方法稍大，但整体而言，RGCN-SA插补效果强于其他模型。

综上所述，本文提出的RGCN-SA算法在20%、30%、50%缺失率条件下具有较好的插补效果，能够很好地解决海洋观测数据缺失问题。

针对长序列缺失的情况，本文在随机缺失连续24步长的情况下，以浮标一的有效波高与最大风速为例进行可视化对比，以说明算法的适用性，结果见图10。从结果可以看出，大部分传统算法表现较差，无法拟合出真实数据，而深度学习算法大都保持较好的水平，能够很好地拟合真实数据，整体来说本文提出的算法对原始数据依然具有较好的拟合性。

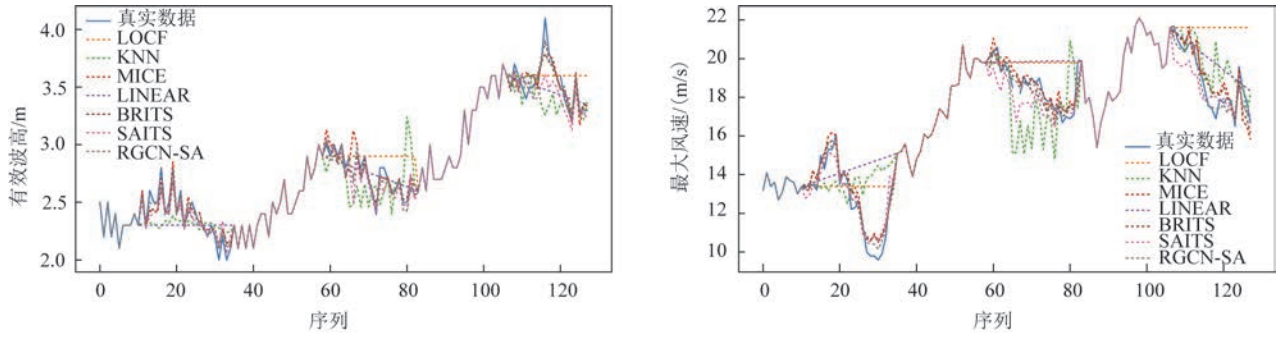
但在长序列整块缺失的情况下，即整个浮标要素都缺失时，例如图11为连续缺失36个步长，此时经过插补得到的结果往往存在较大误差。



a. 有效波高缺失数据插补　　b. 最大风速缺失数据插补

图9 50%缺失率下不同模型对数据插补效果对比

Fig.9 Comparison of data imputation effect of different models under 50% missing rate



a. 有效波高缺失数据插补　　b. 最大风速缺失数据插补

图10 不同模型对连续缺失数据插补效果对比

Fig.10 Comparison of imputation effect of different models on continuous missing data

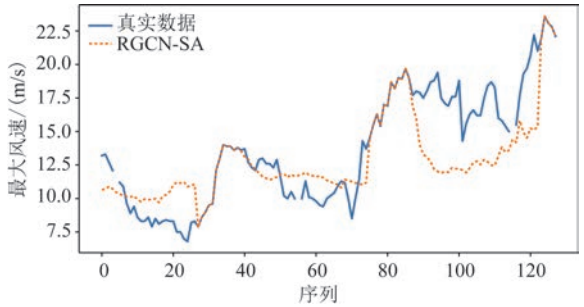


图11 整块要素缺失最大风速插补情况

Fig.11 Imputation of maximum wind speed with missing elements in the whole block

3.4 消融实验

为了验证RGCN-SA模型各个因素的有效性，本文在缺失率为20%的情况下，消除各个因素来验证每个因素的性能。分别采用Transformer模型、RTransformer模型即RGCN-SA去除GCN后的模型与本文模型进行对比，结果见表2。

从表2看出，各个要素在消融实验中都表现了一致的结果，Transformer由于采用了残差结构，性能明显提升，再添加GCN模块提取空间结构，MAE

表2　20%缺失率下经消融插补后的MAE对比

Tab.2　Comparison of MAE after ablation imputation result under 20% missing rate

模型	平均波高/m	平均波高周期/s	平均风速/(m/s)	有效波高/m	有效波高周期/s	最大风速/(m/s)
Transformer	0.046	0.155	0.419	0.068	0.194	0.482
RTransformer	0.042	0.138	0.353	0.061	0.172	0.433
RGCN-SA	0.040	0.133	0.347	0.058	0.171	0.424

进一步减小,其插补性能也进一步提高。因此,RGCN-SA中的各个因素对插补性能的提高都具有一定的效果。

4　结论

受网络波动和传感器故障等因素影响,海洋观测站的数据缺失是一个重要的问题。针对这一问题,本文提出了一种海洋观测数据插补模型RGCN-SA。由于海洋观测数据有明显的时间相关性,而且不同位置的数据也存在不同程度的空间关联性,因此,采用Transformer结构提取时间依赖特征,采用GCN提取空间依赖特征,进行时空依赖特征融合后通过Transformer进一步提取时间依赖特征,并引入残差结构充分提取特征。实验结果表明,RGCN-SA在真实数据上的插补效果优于基线模型,在缺失单个要素长序列插补情况下能够保持较好的效果,但在长序列整块缺失的情况下表现相对较差。因此,下一步将考虑针对长序列数据的缺失问题开展进一步探索和研究。

参考文献:

[1] 刘玉龙, 王国松, 侯敏, 等. 基于深度学习的海温观测数据质量控制应用研究[J]. 海洋通报, 2021, 40(3): 283-291.
LIU Y L, WANG G S, HOU M, et al. Quality control of sea temperature observation data using deep learning neural networks [J]. Marine Science Bulletin, 2021, 40(3): 283-291.

[2] 郑婷婷, 于小宁, 陈旖旎. 基于海洋科技发展的观测监测档案资源动态整合与利用探究[J]. 北京档案, 2023(3): 17-20.
ZHENG T T, YU X N, CHEN Y N. Dynamic integration and utilization of observation and monitoring archive resources based on the development of marine science and technology[J]. Beijing Archives, 2023(3): 17-20.

[3] ZHOU Y Z, SHI J S, STEIN R, et al. Missing data matter: an empirical evaluation of the impacts of missing EHR data in comparative effectiveness research[J]. Journal of the American Medical Informatics Association, 2023, 30(7): 1246-1256.

[4] KANTARDZIC M. Data mining: concepts, models, methods, and algorithms[M]. Hoboken: John Wiley, 2011.

[5] AMIRI M, JENSEN R. Missing data imputation using fuzzy-rough methods[J]. Neurocomputing, 2016, 205: 152-164.

[6] YANG K, LI J Z, WANG C K. Missing values estimation in microarray data with partial least squares regression[C]//6th International Conference on Computational Science. Reading: Springer, 2006: 662-669.

[7] LI Y Y, PARKER L E. Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks[J]. Information Fusion, 2014, 15: 64-79.

[8] AL-HELALI B, CHEN Q, XUE B, et al. A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data[J]. Soft Computing, 2021, 25(8): 5993-6012.

[9] HASTIE T, MAZUMDER R, LEE J D, et al. Matrix completion and low-rank SVD via fast alternating least squares[J]. The Journal of Machine Learning Research, 2015, 16(1): 3367-3402.

[10] JING X, LUO J G, WANG J M, et al. A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest[J]. Water Resources Management, 2022, 36(4): 1159-1173.

[11] PAN H, YE Z W, HE Q Y, et al. Discrete missing data imputation using multilayer perceptron and momentum gradient descent[J]. Sensors, 2022, 22(15): 5645.

[12] CAO W, WANG D, LI J, et al. BRITS: bidirectional recurrent imputation for time series[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 6776-6786.

[13] YOON J, JORDON J, VAN DER SCHAAR M. GAIN: missing data imputation using generative adversarial nets[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 5675-5684.

[14] DU W J, CÔTÉ D, LIU Y. SAITS: self-attention-based imputation for time series[J]. Expert Systems with Applications, 2023, 219: 119619.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010.

[16] 贺琪, 曹万万, 黄冬梅, 等. 面向海洋时序数据异常模式发现的多视图协同可视分析[J]. 海洋通报, 2022, 41(6): 619-629.
HE Q, CAO W W, HUANG D M, et al. Multi-view collaborative visual analysis for anomaly detection of marine environment time series data[J]. Marine Science Bulletin, 2022, 41(6): 619-629.

Maritime buoy observation data Imputation based on RGCN-SA algorithm

PENG Dedong[1,2], LIANG Jianfeng[1*], CUI Xuerong[2], YUE Xinyang[1]

(1. National Marine Data and Information Service, Tianjin 300171, China; 2. College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China)

Abstract: In this paper, a residual network imputation model based on graph convolution network (GCN) and self-attention mechanism (RGCN-SA) is proposed to solve the observational data missing problem. The model is constructed on self-attention mechanism and graph convolution. The self-attention mechanism is used to extract the time -dependent features of observational data, and the space-dependent features of buoys at different positions are obtained through graph convolution. Combined with the self-supervised training method, the model is trained and the final ocean data imputation model is obtained. Through comparative experiments, it is proved that the model can effectively obtain the temporal and spatial correlation features of buoy observations after training, and obtaina better imputation effect than other methods. The effectiveness of each module of the model is proved by the ablation experiment.

Key words: self-attention mechanism; graph convolutional network; imputation; buoy data