·实验技术·

具有缺失值及异常值的时间序列处理与再筛选机制

李逸君[1]，王思淼[1]，赵沐歌[1]，吴然超[1, 2*]

（1. 安徽大学 纽约石溪学院，合肥 230039；2. 安徽大学 数学科学学院，合肥 230039）

摘要：多维时间序列数据应用广泛，但会因缺失或异常值的出现，导致数据不可靠。该文提出了多维时间序列数据处理与再筛选机制 (MTSM) 方法。该方法基于缺失值的 Transformer 填补，结合 3σ法与箱型图检测、分层修正异常值，并依据数据类型应用多尺度模糊熵、边界混合重采样及高斯混合聚类采样，对填补和修正后的数据进行再筛选。基于世界卫生组织的 COVID-19 数据进行了对比分析，结果表明 MTSM 方法在不同缺失率及异常率下均优于 GRU、RNN 和 LATC，在精度与鲁棒性方面也表现突出。

关　键　词：时间序列；缺失值填补；异常值处理；再筛选机制；MTSM 方法

中图分类号：TP312　　**文献标志码：**A　　DOI: 10.12179/1672-4550.20250301

Processing and Reselection Mechanism for Time Series with Missing Values and Outliers

LI Yijun[1], WANG Simiao[1], ZHAO Muge[1], WU Ranchao[1, 2*]

（1. Stony Brook Institute, Anhui University, Hefei 230039, China; 2. School of Mathematial Sciences, Anhui University, Hefei 230039, China）

Abstract: Multidimensional time series data are widely used, but can be rendered unreliable due to missing values or outliers. A multidimensional time series data processing and reselection mechanism (MTSM) method is proposed in this paper. This method is based on Transformer-based imputation for missing values, combined with the 3σ rule and box plots for outlier detection and hierarchical correction. Multi-scale fuzzy entropy, boundary mixture resampling and Gaussian mixture clustering sampling are applied according to data types to re-screen the imputed and corrected data. A comparative analysis was conducted based on the COVID-19 data from the World Health Organization, and the results show that the MTSM method outperforms GRU, RNN, and LATC at different missing and outlier rates, and also demonstrates outstanding accuracy and robustness.

Key words: time series; missing value imputation; outlier handling; reselection mechanism; MTSM method

随着信息技术的发展，多维时间序列数据在金融、医疗和环境监测等领域得到广泛应用。然而，受传感器故障、通信中断或系统误差等现实因素影响，数据常出现缺失值或异常值，削弱其表达能力并影响模型对动态关系的捕捉，从而降低后续分析与预测的准确性。因此，准确识别与填补缺失值、检测与修正异常值，对提升数据质量和建模精度至关重要。

目前，缺失值处理方法大致可分为删除法、统计插补法及机器学习法，其中前者可能造成信息损失，后两者虽能保留样本规模，但部分建模忽略时间序列内部及变量间的依赖结构，难以满足高维、多变量序列的建模需求。传统机器学习方法，如 k-最近邻（k-nearest neighbors, KNN）、期望最大化（expectation-maximization, EM）、核方法（kernel methods, KM）与矩阵分解（matrix factorization, MF），虽具备一定的建模能力，但在多变量依赖捕捉和时序一致性保持方面仍显不足。基于张量分解的补全方法通过挖掘数据的高维结构信息，为处理多维时间序列缺失值提供了重要思路[1−2]，张量理论也为相关研究提供了框架支持[3]。然而，这类方法往往依赖于数据的低秩等

收稿日期：2025−05−20

作者简介：李逸君，本科生，应用统计学专业。E-mail: r22214080@stu.ahu.edu.cn

*** 通信作者：**吴然超，博士，教授，主要从事应用数学方面的研究。E-mail: rcwu@ahu.edu.cn

假设，对复杂非线性时序依赖的刻画能力有限，且在高维场景下常面临计算复杂度的挑战。

异常值检测方法亦存在类似局限。传统阈值判断易受极端样本影响，而基于统计特征或密度估计的模型对结构性变动的适应性有限，难以刻画复杂的时序关联。近年来，深度学习方法在多维时间序列建模及异常识别中展现出潜力，如卷积神经网络（convolutional neural network, CNN）、递归神经网络（recurrent neural network, RNN）以及基于自注意力机制的 Transformer 框架，能够通过端到端学习捕获跨特征与跨时间的依赖关系，从而一定程度上缓解传统检测方法对人工阈值依赖强、难以适应非平稳结构等问题。然而，这类方法仍面临高模型复杂度、高算力需求及可解释性不足的限制，其在资源受限或对可解释性要求较高的领域中仍难以大规模应用。

在缺失值填补方面，文献 [4] 提出的 Transformer 在序列处理中优势显著，其自注意力机制克服了 RNN 训练速度慢和长期依赖下的记忆约束的缺点；文献 [5] 基于 Transformer 提出卷积 Transformer 推断网络（convolutional transformer inference network, CTIN），用于处理高缺失率多维时间序列。在异常值检测方面，文献 [6] 提出利用奇异谱分析和 3σ准则实现数据异常值自动识别；文献 [7] 通过改进箱型图检测法在桥梁异常值检测中取得成效，表明 3σ准则和箱型图可用于异常值检测；文献 [8] 指出可将异常值剔除后当作缺失值处理，并提出均值替换的可能性；文献 [9] 运用箱型图筛选异常值，并用中位数暂代，经实验验证了局部中位数替换法的可行性。

对于时间序列数据类型，文献 [10] 明确了时期与时点的概念，为本文中不同数据处理策略的选择提供了依据。文献 [11] 提出的边境混合重采样分类处理方法，对有明显决策边界的数据分类效果良好。文献 [12] 提出了高斯混合聚类采样模型，文献 [13] 对其步骤进行了完善，该方法基于后验概率进行数据分类，通过参数迭代优化，并结合欠采样与过采样策略处理数据不平衡问题。文献 [14] 提出基于多尺度模糊熵的时间序列特征提取算法，经实验验证该算法能有效提取特征，在分类任务中表现良好。

综合来看，现有方法在兼顾缺失处理准确性、异常检测精度与序列结构保持性方面尚显不足，亟需构建一种具备鲁棒性与通用性的时间序列数据质量提升机制，以支持后续高质量建模与科学决策。

1　方法

1.1　基础框架

受大规模时间序列分析中“分治”策略的启发[15]，本文将“划分−治理−合并”的思想融入模型设计，提出了多维时间序列数据处理与再筛选机制（multidimensional time series data processing and rescreening mechanism, MTSM）方法，如图 1 所示。首先，通过将每个时间点的多变量特征联合输入 Transformer 模型，利用自注意力机制捕捉变量之间的时序依赖与交互信息，对数据中的缺失值进行 Transformer 填补。其次，在异常值处理阶段，针对不同变量的局部扰动特性，初步采用变量级独立处理方式，使用 3σ与箱型图检测，对数据中的异常值进行分层修正。最后，依据数据类型应用多尺度模糊熵、边界混合重采样及高斯混合聚类采样，对填补和修正后的数据进行再筛选，从而得到具有时间趋势代表性的优化序列。MTSM 方法在保持初步处理高效性的同时，通过建模阶段与特征组合实现了多维结构信息的融合与利用。该方法不仅能有效简化数据，大幅减少噪声和冗余信息，从而降低后续预测任务的复杂性与计算成本，而且能够帮助识别和突出数据中的重要特征，为下游任务提供更有效的特征选择和数据优化方案。

1.2　缺失值填补

基于 MTSM 方法，对多维时间序列数据中因传感器故障、传输错误、录入失误等导致的缺失值按照数据预处理与检测、数据填补分步进行处理。

1.2.1　数据预处理和缺失值检测

首先对时间序列数据进行格式化、规范化处理，完成对数据的初步检测，然后采用基于掩码矩阵的方法进行缺失值检测。在数据初始化阶段，生成一个与原始数据同维度的掩码矩阵，用二进制值表示数据的缺失情况。如对于多维时间序列中的某个时间点，如果该点的数据缺失，则在掩码矩阵中对应位置标记为 0，反之则为 1。

1.2.2　数据的 Transformer 填补

首先通过卷积操作结合输入特征和缺失信息，对缺失值初始化。将输入特征、缺失信息和

时间间隔信息同时输入到卷积核大小为K的卷积层中。通过卷积层学习获得混合信息，提取时间维度信息和特征之间的关系信息，从而初始化缺失值[16]。其中，卷积核大小K的确定采用数据驱动的方法：首先预设一组候选取值范围[1,3,5,7,9]，针对每个K值，在训练集上训练基于 Transformer 的填补模型，并在验证集上评估模型性能。评估指标选用平均绝对误差(mean absolute error, MAE)、平均绝对百分比误差(mean absolute percentage error, MAPE)及均方根误差(root mean square error, RMSE) 3 项指标，最终选择在验证集上使评估指标达到最优的K作为卷积核大小。

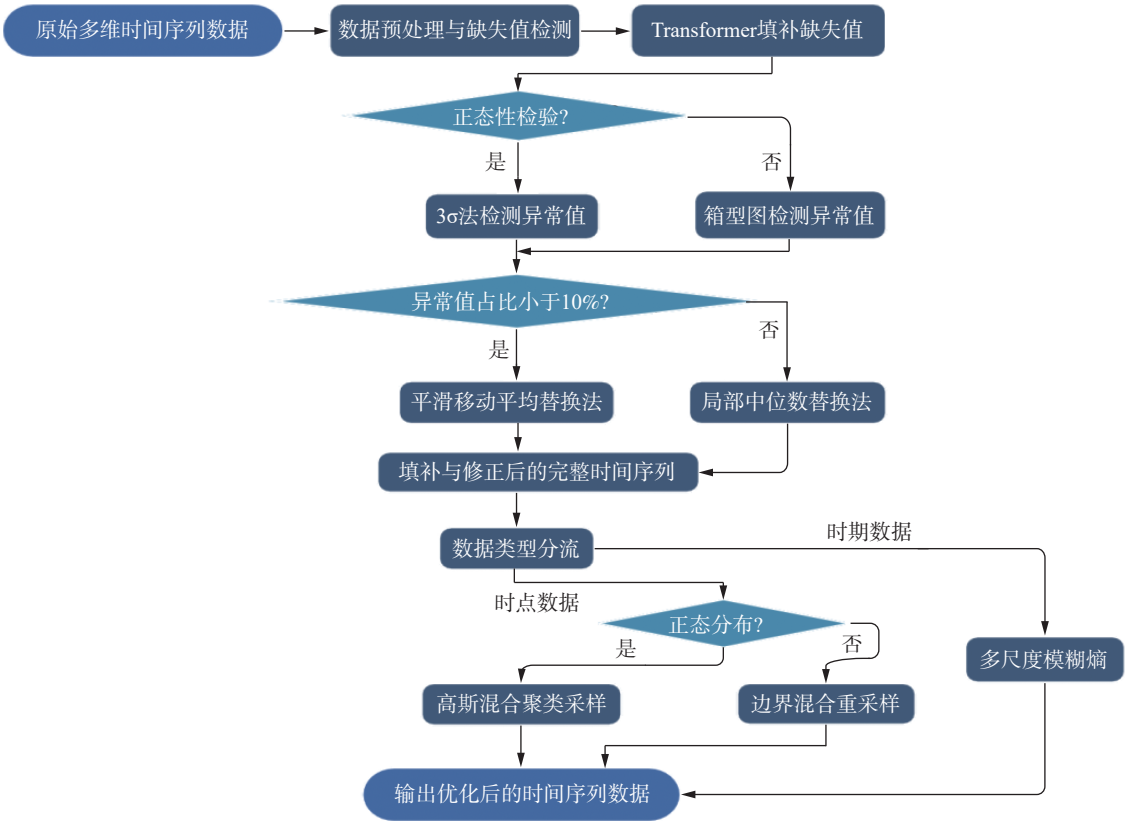

图 1　MTSM 方法流程图

具体卷积操作可表示为：

$$p_{\mathrm{n}} = \mathrm{Conv}(\mathrm{Concat}[x_{\mathrm{n}}, I_{\mathrm{n}}, \Delta_{\mathrm{n}}]) \tag{1}$$

式中：p_{n}是卷积后的输出，x_{n}是输入特征，I_{n}是缺失信息，Δ_{n}是时间间隔信息。

然后利用 Transformer 的自注意力机制从而有效捕捉时间序列数据中的长短期依赖关系。对初始化后的时间序列进行特征提取和关系建模：

$$Z = \mathrm{selfAttention}(x_{\mathrm{c}}, I_{\mathrm{n}}) \tag{2}$$

式中：Z是通过自注意力机制得到的特征表示，x_{c}是卷积后的特征，I_{n}是缺失信息。

最后通过线性层对缺失值进行推理，并将估算的缺失值与原始观测值结合，形成最终填补后的完整时间序列为：

$$\hat{x}_{\mathrm{n}} = I_{\mathrm{n}} \cdot x_{\mathrm{n}} + (1 - I_{\mathrm{n}}) \cdot x_{\mathrm{c}} \tag{3}$$

1.3　异常值检测和处理

异常值通常由数据采集过程中出现的错误、设备故障或极端情况引起。在 MTSM 方法的异常值检测与处理阶段中，依据数据集分布及异常值数量择取恰当修正方式，实现了数据的两次分流处理。

1.3.1　异常值检测

正态分布检验步骤使用 Shapiro-Wilk 测试。其统计量公式如下：

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2} \tag{4}$$

式中：$x_{(i)}$是按照从小到大排好序的样本值；$\bar{x}$是样本均值；a_i是常数系数，根据标准正态分布下样本

秩次的期望值和协方差矩阵计算得到，反映不同秩次样本对正态性判断的权重。

对于服从正态分布的数据集，采用3σ异常值检测法，用于确定正态分布下的异常判别阈值，从而实现对异常数据的初步识别。其公式如下：

$$P(|x-\mu|>3\sigma)\leqslant 0.002\,6 \tag{5}$$

式中：x是数据点，μ是均值，σ是标准差。当数据点的值超过$\mu\pm3\sigma$时，即认为该数据点为异常值。

对于服从非正态分布的数据集，采用箱型图（Boxplot）法，即通过计算数据的四分位距（inter quartile range, IQR），并将超过四分位数 1.5 倍范围以外的值认定为异常值，公式如下：

$$Q=Q_3-Q_1 \tag{6}$$

式中：Q_1和Q_3分别表示数据的第 1 四分位数和第 3 四分位数。

下界和上界分别为：

$$L_1=Q_1-1.5Q \tag{7}$$

$$L_2=Q_3+1.5Q \tag{8}$$

当数据点小于L_1或大于L_2时，认为该数据点为异常值。

对正态分布数据，3σ检测基于其集中趋势和已知概率分布范围，具有显著理论依据；对非正态分布数据，箱型图方法以分位数为基础，更能应对偏态和重尾情形，适用于结构复杂或偏移严重的时间序列异常识别。

1.3.2　异常值修正

为合理区分不同异常值修正策略，本文设置异常值比例 10% 为处理方法的划分阈值。该阈值参考相关研究中的经验设置，并适配本文模拟实验中设计的缺失/异常比例。10% 被广泛认为是轻度扰动的上限，适用于使用移动平滑等算法进行局部修正，此时异常点对整体分布影响较小；而当异常值比例超过 10% 时，移动平均易受极端值影响，为增强稳健性，需采用对离群点更具鲁棒性的局部中位数替换法。

若检测出的异常值数量小于总数据量的 10%，采用移动平滑替换法，此时少量异常值替换对整体数据分布影响小，平均值可较好代表数据中心趋势。具体步骤如下。

1）定义窗口期大小2k。

2）对于每个异常数据点，计算其前后共2k个相邻数据点的平均值并替换。对应公式为:

$$\hat{x}_i=\frac{1}{2k}\sum_{j=i-k}^{i+k}x_j \tag{9}$$

式中：$\hat{x}_i$为替换后的值，x_j是第 j 个数据点的观测值。

3）若异常值靠近序列边界导致前或后数据不足，则采用边界补偿策略，即从另一侧补足所缺数据，以保证替换值基于完整的2k个样本点计算，确保替换的一致性与有效性。对应公式为：

$$\hat{x}_i=\frac{1}{2k}\left(\sum_{j=1}^{i-1}x_j+\sum_{j=i+1}^{2k}x_{j-(2k-(i-1))}\right) \tag{10}$$

若异常值数量大于总数据量的 10%，采用局部中位数替换法，此时数据集分布偏斜或含离群点，中位数更能反映数据中心位置。具体步骤如下。

1）定义窗口期大小2k（如 10 个数据点）。

2）对于每个异常数据点，计算其及其前后数据点构成的局部数据集合，进而进行替换。若前侧数据不足，则从后侧补足，确保窗口长度为2k；反之，则用前侧数据补足。如若某异常数据是某数据集中的第 3 个数据点，则取其前 2 个、它本身及其后 7 个数据点计算中位数。

3）为避免在趋势性数据中因中位数替换而引入新的异常点，本文在替换前引入趋势性检测机制，即采用局部线性拟合方法对滑动窗口内数据进行线性回归建模，提取回归斜率β_1以表征趋势强度。若斜率绝对值超过设定阈值θ，则判定该窗口存在显著线性趋势，此时不采用中位数替换，而改用线性拟合的预测值进行替换，以兼顾异常修正的稳健性与时间序列趋势的一致性。阈值θ可通过经验设定（如$\theta=0.05$），也可基于训练数据中斜率分布的分位值（如 95% 分位）自适应设定。

1.4　数据再筛选

对时间序列数据进行缺失值填补及异常值处理之后，数据仍可能存在不平衡与噪声问题。为了进一步提升数据质量和分析精度，需要对数据进行再筛选，实现对数据再次进行特征选择和优化，使其更适配后续的分析和预测任务。对时点数据，在适用条件下选取边界混合重采样（boundary synthetic minority over-sampling technique，

BSMOTE)或高斯混合聚类采样(gaussian mixture model sampling, GMMS)进行数据再筛选处理[17−18]。对时期数据采用多尺度模糊熵进行数据再筛选处理。

对于时点数据，当数据分布与正态分布类似时，采取 GMMS；若数据有明显的决策边界，则采用 BSMOTE。GMMS 处理数据时，先对符合高斯分布的数据用高斯滤波器去噪，确定标准差平衡去噪与细节保留[19]。依据概率分布和后验概率判断数据点聚类，迭代更新参数，当极大似然估计函数增长小于阈值时停止。首次聚类前，随机设定聚类数并依原始数据计算参数。采样处理时，用支持向量机(support vector machine，SVM)算法确定超平面，计算聚类中心到超平面的距离和权重，生成新的少数类样本，据此计算多数类样本删除数量。少数类样本按到超平面距离分入不同集合，靠近超平面和距离适中的样本分别结合算法生成新样本，远的不处理。通过这些操作实现重采样和平衡处理，增加少数类样本权重，提升其在训练中的关注度。采用 BSMOTE 通过定义边界样本和非边界样本，对边界样本进行重采样，生成新的样本点[20]。对于少数类样本，通过随机选取两个样本点，并在它们的连线上生成新的样本点，以增加少数类样本的数量[21]。对于多数类样本，则进行欠采样，删除部分样本以减少数据不平衡。

对于时期数据，使用多尺度模糊熵方法对数据进行处理，将模糊熵特征作为多维度组合输入分类器，在多个变量层面上协同判断数据片段的波动性与有效性，进而消除异常波动和噪声，提高数据的稳定性和可靠性[22]。对于一个长度为n的时间序列数据$M=\{m_1,m_2,\cdots,m_n\}$，采用滑动窗口技术进行数据分割，将原始数据重塑为指定长度的样本。窗口大小为s，数据备份为$k=n-s+1$个区间。从第i个时间点开始，构建长度为s的滑动窗口向量，并对每个窗口进行均值归一化处理，得到如下去中心化的时间序列片段M_i^s：

$$M_i^s=[m_i-\overline{m}_i,m_{i+1}-\overline{m}_i,\cdots,m_{i+s-1}-\overline{m}_i] \tag{11}$$

式中：$i=1,2,\cdots,k$，$\overline{m}_i$为窗口均值，计算公式为：

$$\overline{m}_i=\frac{1}{s}\sum_{t=0}^{s-1}m_{i+t} \tag{12}$$

在计算两个s维矢量M_i^s和M_j^s的距离$d_{i,j}^s$后，采用如下模糊隶属度函数计算相似度：

$$X_{i,j}^s=\begin{cases}1, & d_{i,j}^s=0\\ \mathrm{e}^{\left[-\ln(2)\left(\frac{d_{i,j}^s}{v}\right)^2\right]}, & d_{i,j}^s\neq 0\end{cases} \tag{13}$$

式中：$i,j=1,2,\cdots,k$，且$i\neq j$，v为相似容限参数。本文采用数据驱动的经验设定方法，根据标准化后数据的尺度特性，取$v=0.2$作为默认值，以兼顾系统对微小波动的鲁棒性与对局部结构的敏感性。对于特征波动显著的数据序列，也可根据数据标准差动态设定$v=\alpha\cdot\sigma$，其中$\alpha\in[0.1,0.3]$，以增强模糊熵对不同序列结构的适应能力，进而得到模糊隶属函数如下式：

$$A_i^s(v)=\frac{1}{k}\sum_{j=1,j\neq i}^{k}X_{i,j}^s \tag{14}$$

该模糊隶属度函数能够提高原始时间序列的抗干扰能力，达到去噪声的效果。

计算s维度下的关系维度：

$$\varphi^s(v)=\frac{1}{k}\sum_{i=1}^{k}A_i^s(v) \tag{15}$$

对数据进行粗粒度划分从而得到新的向量：

$$C_j^\tau=\frac{1}{\tau}\sum_{i=(j-1)\tau+1}^{\tau}m_i \tag{16}$$

式中：τ表示尺度因子以确定粗粒划分割数量，取值为$\tau=1,2,\cdots,1\leqslant j\leqslant\dfrac{n}{\tau}$。

根据原始时间序列的模糊熵公式：

$$\mathrm{FuEn}(s,v,n)=\ln\varphi^s(v)-\ln\varphi^{s+1}(v) \tag{17}$$

得到带有尺度因子τ的模糊熵公式：

$$\mathrm{MFE}(\tau,s,v)=\mathrm{FuEn}(C_j^\tau,s,v) \tag{18}$$

上述模糊熵公式通过计算各时间段在多个尺度下的模糊熵值，构建特征向量并输入分类器进行训练，实现对高熵(波动大、噪声强)与低熵(稳定、结构清晰)序列片段的有效识别与筛选。最终保留熵值较低，预测价值更高的时间片段用于后续建模，提升模型对趋势性信息的提取能力与整体预测性能。

2 实验及结果分析

2.1 实验数据概述

世界卫生组织 COVID-19 大流行数据集由世

界卫生组织提供，专门用于记录和分析 COVID-19 大流行期间的相关数据。数据结构按区域、年份和月份统计，形成一个 7×2×12×6 的张量时期序列数据。具体来说，该数据集涵盖了 7 个大区域（如美洲、欧洲等）、从 COVID-19 开始的 24 个月的时间跨度，以及 6 种不同的统计指标。统计指标包括预期死亡人数均值，估计死亡人数均值，与 COVID-19 相关的超额死亡人数均值，以及与其相关的累计超额死亡人数的均值、95% 置信区间下限和上限。

COVID-19 大流行数据集提供的按月统计的死亡数据具有较高的时间分辨率，这使得研究者可以精准地捕捉疫情的变化轨迹，评估不同防控措施的短期和长期效果。数据中的预期死亡人数和估计死亡人数之间的对比，能够揭示出疫情对各国卫生系统的真实压力和挑战。

2.2　实验设计及评价指标

本文采用世界卫生组织提供的 COVID-19 大流行数据集来评估 MTSM 方法的性能。

为确保实验具有良好的时序一致性与现实应用参考价值，本文在实验过程中对原始时间序列数据进行了预先划分：将每个变量对应的序列数据按时间顺序划分为前 80% 与后 20% 两部分。其中，前 80% 的数据被用作处理集，用于模拟缺失值与异常值情形，并分别由 MTSM 方法及各基准模型进行数据修复与处理；后 20% 的数据被设定为预测集，用于验证处理结果在后续预测任务中的有效性。

在实验中，为模拟实际数据处理中常见的缺失值与异常值问题，本文对处理集中的观测记录进行局部扰动。具体地，随机选择一定比例的时间点，并在其包含的多个变量中随机删除（设为空）或扰动其中部分变量值，而非整行删除或替换。以表 1 数据结构为例，每条记录包含 6 个统计指标，本文仅对其中 1~3 个指标进行缺失或异常模拟，以更真实地还原实际数据采集过程中传感器误差或记录缺失的局部性特征，并确保模拟后的数据仍具有统计完整性与时序连续性。

表 1　COVID-19 大流行期间 2020 年 1、2 月不同区域的死亡人数数据示例

地区	年份	月份	预期死亡数均值/人	估计死亡数均值/人	超额死亡数均值/人	累计超额死亡数均值/人	累计超额死亡数95%置信区间下限/人	累计超额死亡数95%置信区间上限/人
美洲	2020	1	619691	634106	12691	12691	2703	23011
美洲	2020	2	595470	582995	−14197	−1506	−15722	13557
欧洲	2020	1	831448	856113	23255	23255	15774	31478
欧洲	2020	2	820874	775606	−46680	−23425	−36361	−10667
亚洲	2020	1	1154702	1296477	142631	142631	35727	266577
亚洲	2020	2	1136399	1109634	−27139	115492	−22718	275253
非洲	2020	1	657358	676786	19403	19403	−33118	79515
非洲	2020	2	647915	649830	1890	21294	−53876	106902

为了全面评估 MTSM 方法的性能，本文设计了两个主要实验。

1）综合对比实验。将 MTSM 方法与门控循环神经网络（gated recurrent unit, GRU）模型、RNN 模型和低秩自回归张量补全（low-rank autoregressive tensor completion, LATC）算法等基准模型进行对比。通过随机删除或改变处理集中 10%、20%、30% 的观测数据，模拟不同程度的数据缺失和异常情况，评估各模型在处理缺失值和异常值时的表现，对比它们在不同数据完整性条件下的性能，全面验证 MTSM 方法的有效性和优势。

2）预测比对实验。选取 SVM 预测和长短期记忆网络（long short-term memory, LSTM）这 2 种预测方法进行实验。将所有模型处理后处理集的数据作为输入，分别应用这两种预测方法，对预测集时间段内的观测值进行预测。随后，将各模型在不同数据处理质量下的预测结果与真实的预测集数据进行比较，从而评估 MTSM 方法在提升后续时间序列预测准确性方面的优势与稳定性。

2.3　实验结果及分析

为评估 MTSM 方法在多维时间序列缺失值与异常值处理中的性能表现，本文选取 MAE、MAPE 与 RMSE 3 项指标，分别对其与 GRU、RNN、LATC 模型在不同缺失/异变率条件下的数据处理前后效果进行对比。实验 1 中各模型在 10% 缺失/异变率下处理前后 MAE、MAPE、RMSE 对比如表 2 所示。

表 2　10% 缺失/异变率下数据处理前后模型性能对比

模型	MAE 前	MAE 后	MAPE 前/%	MAPE 后/%	RMSE 前	RMSE 后
MTSM	0.225	0.124	5.21	2.53	0.368	0.192
GRU	0.225	0.164	5.21	3.46	0.368	0.242
RNN	0.225	0.181	5.21	3.67	0.368	0.267
LATC	0.225	0.169	5.21	3.29	0.368	0.238

经过各模型处理后的 MAE、MAPE 和 RMSE 3 项指标在不同缺失/异变率(10%、20%、30%)下的柱状对比如图 2~图 4 所示。为便于对比分析，各指标变化以表 3 中“不同缺失/异变率下数据处理前的指标值”为基准参照。

从上述处理前的指标数据来看，随着缺失率和异常率的上升，各模型的误差指标均呈现出明显上升趋势，说明原始数据质量下降对模型性能产生了显著影响。经过 MTSM 方法处理后，其在 3 项指标上的改善最为显著。如在缺失率为 30% 的情况下，MTSM 方法的 MAE 由 0.401 降至 0.194，MAPE 由 8.43% 降至 4.18%，RMSE 由 0.585 降至 0.292，降幅均超过 50%。相比之下，其他基准模型虽也在处理后取得了一定程度的误差降低，但整体改善幅度相对较小，且在高缺失率下鲁棒性不足。从对比来看，MTSM 方法在所有缺失/异变率水平下均展现出更强的误差抑制能力和更好的指标稳定性，尤其在高缺失率下依然保持较优性能，反映出其在数据完整性恢复、异常值修正及优化重构方面的综合优势。

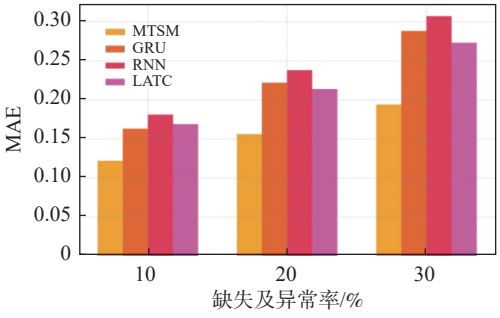

图 2　数据处理后的 MAE 柱状图

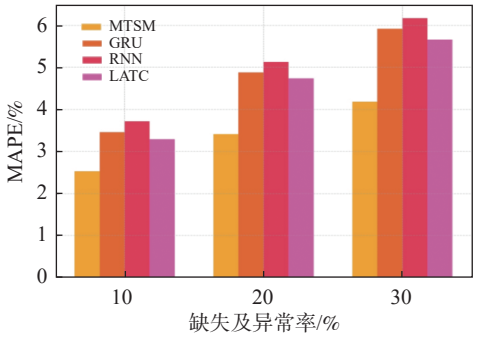

图 3　数据处理后的 MAPE 柱状图

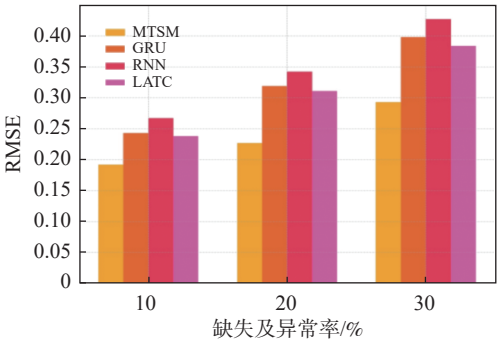

图 4　数据处理后的 RMSE 柱状图

表 3　不同缺失/异变率下数据处理前的 MAE、MAPE、RMSE

缺失/异变率/%	MAE	MAPE/%	RMSE
10	0.225	5.21	0.368
20	0.327	6.92	0.487
30	0.401	8.43	0.585

为进一步对比不同数据处理模型在提升时间序列预测精度方面的效果，实验 2 将 MTSM、GRU、RNN 与 LATC 模型在不同缺失/异常率(10%、20%、30%)下的处理结果，分别输入 SVM 与 LSTM 模型进行预测，并通过 MAE、MAPE 与 RMSE 3 项指标对预测结果进行评估。在 3 种缺失/异常率条件下，各数据处理模型结合不同预测方法所得的误差变化趋势如图 5~图 7 所示。

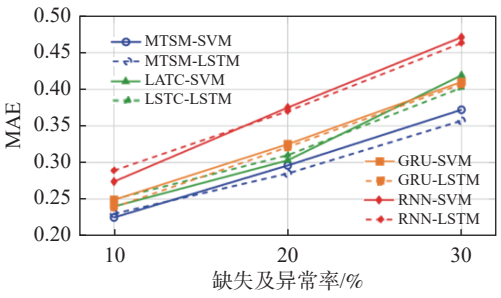

图 5　不同缺失/异常率与预测模型下各处理方法的预测性能对比(MAE)图

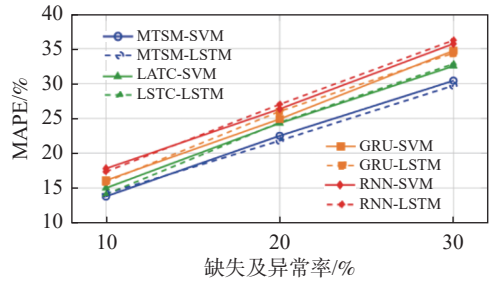

图 6　不同缺失/异常率与预测模型下各处理方法的预测性能对比(MAPE)线性图

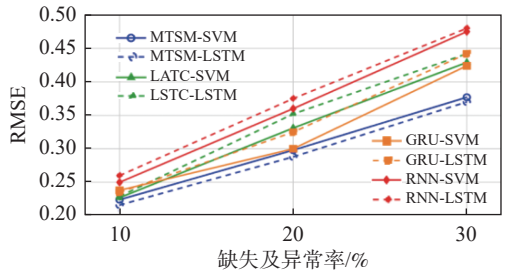

图 7　不同缺失/异常率与预测模型下各处理方法的预测性能对比(RMSE)线性图

整体来看，MTSM 方法在所有条件下均实现了最优预测精度，表现出较强的稳定性与泛化能力。值得注意的是，在缺失/异常率较低的 10% 场景中，MTSM 方法与部分基准模型的预测结果相对接近，表现差距较小，但随着缺失/异常程度加剧，MTSM 方法在各项评估指标(MAE、MAPE、RMSE)上的优势逐渐扩大，进一步验证了其在复杂数据场景下的鲁棒性。

3　结束语

本文针对多维时间序列中常见的缺失值与异常值问题，构建了集成缺失值填补、异常值检测与修正、再筛选的 MTSM 方法。该机制是基于 Transformer 的时序填补方法、分流数据的异常值处理策略以及适应数据类型的再筛选方法的有机结合，显著提升了数据的完整性、稳定性与结构表达能力。通过在 COVID-19 实际数据集上的对比实验验证，MTSM 方法在 MAE、MAPE 及 RMSE 等评价指标上相较于现有主流方法表现出更优性能，体现出较高的精度与鲁棒性。研究结果表明，MTSM 方法在复杂、不完整时间序列的预处理与建模支持中具有广泛的应用潜力与现实意义，可为公共卫生等领域的高质量数据分析与科学决策提供可靠支撑。需要指出的是，本文实验部分采用的是时期数据，暂未涉及对时点数据处理机制的实证验证。尽管如此，MTSM 方法中针对时点数据所提出的再筛选机制具备完整的理论基础与可操作性，方法逻辑与流程设计合理，后续研究可基于金融监测、物联网采集等高频时点型时间序列数据进一步拓展其实验验证与应用场景。

参 考 文 献

[1] SONG Q Q, GE H C, CAVERLEE J, et al. Tensor completion algorithms in big data analytics[J]. ACM Transactions on Knowledge Discovery from Data, 2019, 13(1): 1−48.

[2] LIU R X, LI W, LIU F, et al. Prediction of time series with missing value based on tensor autoregressive completion[J]. Computer and Modernization, 2023(9): 51−58.

[3] 别金金. 张量分解在高维时间序列中的应用[D]. 上海: 华东师范大学, 2021.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS' 17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Asscociates Inc., 2017: 5998–6008.

[5] 蒋雪琳. 基于 Transformer 模型的多元时间序列填补和预测研究[D]. 西安: 西安理工大学, 2023.

[6] 陈伟楠, 杜国志, 张锏, 等. 基于 3σ-SSA 的数据清洗方法在大坝智慧安全监测系统中的应用[J]. 水利规划与设计, 2024(1): 113−116.

[7] 徐凯, 袁蒋鹏. 粗泛化箱型图法在大中跨径桥梁监测异常值剔除的应用[J]. 广东公路交通, 2024, 50(1): 66−71.

[8] 徐昊, 王永生, 许志伟, 等. 基于生成对抗网络多变量风电时间序列异常值处理[J]. 太阳能学报, 2022, 43(12): 300−311.

[9] 何高清, 肖健. 轴承尺寸检测数据的异常值检测与数据处理研究[J]. 机电工程, 2021, 38(2): 198−203.

[10] 邹文慧. 识别时间序列中的时期与时点概念[J]. 山东纺织经济, 2023, 40(12): 10−13.

[11] 侯贝贝, 刘三阳, 普事业. 基于边界混合重采样的非平衡数据分类方法[J]. 计算机工程与应用, 2020, 56(1): 46−52.

[12] 王宏伟, 柴秀俊. 基于高斯混合模型聚类的非均匀采样系统的多模型切换辨识[J]. 控制与决策, 2021, 36(12): 2946−2954.

[13] 严涛, 江开忠, 姜新盈, 等. 基于高斯混合聚类采样的不平衡数据处理方法[J]. 混合采样, 2023, 40(12):

305−311.
[14] 宋春雷, 路晓亚, 何笑笑. 基于多尺度模糊熵的时间序列特征提取算法[J]. 无线互联科技, 2023, 20(23): 111−114.
[15] 滕飞, 黄齐川, 李天瑞, 等. 大规模时间序列分析框架的研究与实现[J]. 计算机学报, 2020, 43(7): 1279−1292.
[16] 魏延杰. 基于深度学习的复杂时间序列分析[D]. 哈尔滨：哈尔滨工业大学, 2018.
[17] 刘焱昕. 基于数据筛选的不平衡数据重采样方法研究[D]. 太原: 山西财经大学, 2019.
[18] 李欣欣. 基于机器学习和重采样的个人贷款违约预测分析[D]. 上海: 上海师范大学, 2024.
[19] 王俊霞, 张岩波, 余红梅, 等. 基于高斯混合模型双向聚类重采样和随机森林构建 DLBC 早期复发预测模型[J]. 中国卫生统计, 2025, 42(1): 7−11.
[20] 侯贝贝. 基于数据重采样的非平衡数据分类方法研究[D]. 西安: 西安电子科技大学, 2020.
[21] 郑重. 类不平衡数据的选择性混合采样方法研究[D]. 合肥: 安徽大学, 2023.
[22] 张勇飞, 陈涛. 基于时间序列模糊分割的通信数据分类算法设计[J]. 吉林大学学报 (工学版), 2023, 53(11): 3268−3273.

编辑 王燕

（上接第 26 页）

[7] PARK K W, SHIN D W, RHEE Y C. Implementation of MultiBand-digital passive intermodulation distortion measurement system[J]. The Journal of the Korea Institute of Electronic Communication Sciences, 2016, 11(12): 1193−1200.
[8] ZHANG J, CHEN Q, ZHOU N G. Correlation analysis with additive distortion measurement errors[J]. Journal of Statistical Computation and Simulation, 2017, 87(4): 664−688.
[9] 赵二刚, 王艳芳, 张维. 低频信号失真度测量系统设计[J]. 自动化与仪表, 2019, 34(12): 53−56.
[10] 窦如凤, 井娥林. 基于 AD620 的微弱信号放大器设计[J]. 仪表技术与传感器, 2021(3): 45−47.
[11] 洪加强, 刘灵箫, 魏瑞丹, 等. 一种基于自激推挽电路的隔离供电电路: CN115347800A[P]. 2022-11-15.
[12] 张滔. 基于 STM32 单片机 DMA 机制的多通道数据采集[J]. 黑龙江科技信息, 2013(30): 27.

编辑 王燕